The Little Handbook of Statistical Practice

Gerard E. Dallal, Ph.D

Chief, Biostatistics Unit
Jean Mayer USDA Human Nutrition Research Center on Aging
at Tufts University
711 Washington Street
Boston, MA 02111
Gerard.Dallal@tufts.edu

A good case can be made that the best set of articles about statistical practice written for the practitioner is the series of Statistics Notes appearing in the British Medical Journal.

There have been many attempts at online statistics instruction. HyperStat is one of the better ones, not only for the content but also for the additional links.

[back to home page]

Permissions

Gerard E. Dallal, Ph.D.

[There have been no changes of substance since this page was first posted. There have been some changes in language and some corrections to spelling and grammar. However, if you notice the page has been updated since your last visit, you need not worry that the policy might have changed in ways that are not immediately obvious.]

I have received some very kind words from readers who have discovered these pages and wish to use parts of them for their own teaching. There are many excellent discussions of statistical methods on the World Wide Web, so I'm both pleased and flattered that people find these pages useful.

I'm still trying to figure out the implications and ramifications of posting these notes. I'm happy to have anyone link to them. While there are many ways to link to the Table of Contents, perhaps the most straightforward method is to link to http://www.StatisticalPractice.com. Deep linking to individual notes, without going through the Table of Contents, is permitted, too.

These notes should not be copied in electronic form or modified in any way. There are two reasons for this. First, they are my particular view of statistical practice. They are also a work in progress. While it's unlikely I'll find myself reversing my views on a subject, I may add material or change emphasis. I don't want older version perpetuating themselves. I'd like people to have access to the latest versions only. Second, I don't want to end up in competition with myself! Anyone looking for my notes or being referred to them through a search engine should be sent to my web pages rather than someone else's. I mention these issues so that anyone wishing to propose a use of these notes that I might not have already considered will know what my concerns are.

Instructors are permitted to make paper copies for instructional purposes provided

- there is no charge to students,
- the content is not modified, and
- students are provided with the URL to the individual note (many browsers automatically print it at the top of the page) or to http://www.StatisticalPractice.com.

Please check each academic year to be sure the policy has not changed.

[back to LHSP]

Last modified: undefined.

Introductory Remarks

Gerard E. Dallal, Ph.D.

I am not trying to write a statistics textbook! There are already lots of them and many are very good. I am not even trying to write an online statistics textbook. There are a bunch of them, too. Besides, writing comes hard to me, as those notes will demonstrate.

My aim is to describe, for better or worse, what I do rather than simply present theory and methods as they appear in standard textbooks. This is about statistical practice--what happens when a statistician (me) deals with data on a daily basis. There will be topics where some readers may disagree with my approach. This is true of any type of practice. I welcome all comments, advice, brickbats, love letters, and complaints.

This is very much a work in progress. It seems to take 3 or 4 drafts to figure out how to eliminate unnecessary technical detail. Perhaps it's more correct to say it takes 3 or 4 drafts to figure out what's unnecessary!

Not all of the examples are ideal, I hope to find better examples for those cases where where a dataset doesn't illustrate a technique as well as I or a reader might like. Time constraints sometimes cause me to adapt what I have at hand rather than search for something more suitable. Once the drafts are posted, I can fill them in with new examples as they cross my path.

I'm typing as fast as I can!

These notes are being prepared as I teach Nutrition 209, Statistical Methods for Nutrition Research, a one-year first and last course in statistical methods required of most bench science and policy students in the Gerald J. & Dorothy R. Friedman School of Nutrition Science and Policy at Tufts Univeristy. I plan to have them reflect my lectures as closely as possible and, perhaps, use them in place of a textbook.

I intend to stay current with my lectures, so these notes will be sketchy and incomplete in some places depending on how much time I have to devote to them. With any luck, there will be no errors of fact. All corrections and requests for clarification are appreciated. I plan to fill in gaps and polish things up with successive iterations.

Last modified: undefined.

Is Statistics Hard?

Gerard E. Dallal, Ph.D.

No! Questions like this invariably lead to self-fulfilling prophecies. Tell yourself statistics is hard, and it's hard. Tell yourself statistics is easy, and *it's easy*! As with most activities rich enough to demand formal study, there are traps for the unwary that must be avoided. Fall into them at the beginning, and statistics is hard Avoid them from the outset, and you'll wonder what the fuss is all about. The amount of success and the speed with which you'll achieved it depends in large part on how quickly these particular lessons are learned.

1. *Statistics is backwards!* One thing most people (even statisticians!) would like to do is describe how likely a theory or hypothesis might be in light of a particular set of data. *This is not possible* in the commonly used classical/frequentist approach to statistics. Instead, statistics talks about the probability of observing particular sets of data, assuming a theory holds. We are not allowed to say, "Because I've seen these data, there is only a small probability that this theory is true." Instead, we say, "The probability of seeing data like these is very small if the theory is true." This means we need methods for translating this latter type of statement into a declaration that a theory is true or false.

2. *Statistical methods are convoluted!* In order to show an effect exists, statistics begins by assuming there is no effect. It then checks to see whether the data are consistent with the assumption of no effect. If the data are found to be inconsistent with the assumption, the assumption must be false and there is, in fact, an effect! Simple? Maybe. Intuitive? Certainly not!

3. *Failing to find an effect is different from showing there is no effect!* In the convoluted way of showing an effect exists, a statistician draws up a list of all of the possibilities that are consistent with the data. IF one of the possibilities is zero, or no effect, it is said that the statistical test fails to demonstrate an effect. That is, an effect has been demonstrated only when the possibility of "no effect" has been ruled out. When an effect has not been demonstrated, it is sometimes misinterpreted as showing there is no effect. These are two different things. "Failing to show an effect" means just that--"no effect" is among the list of possibilities, which might also include possible effects of great importance. However, because "no effect" has not been ruled out, it cannot be said that an effect has been demonstrated, regardless of what the other possibilities might be! "Showing there is no effect" means something stronger--it says not only that "no effect" is a possibility, but also that the other possibilities are of *no* practical importance.

   A typical misstatement is, "There is no effect," when the analyst should be saying, "The data failed to demonstrate an effect." The distinction is critical. If there is no effect, there is no more work to be done. We know something--no effect. The line of inquiry can be abandoned. On the other hand, it is possible to fail to demonstrate an effect without showing that there is no effect. This usually happens with small samples.

> This is best illustrated by an example. Suppose a researcher decides to compare the effectiveness of two diet programs (W and J) over a six-month period and the best she is able to conclude is that, on average, people on diet W might lose anywhere from 15 pounds more to 20 pounds less than those on diet J. The researcher has failed to show a difference between the diets because "no difference" is among the list of possibilities. However, it would be a mistake to say the data show conclusively that there is no difference between the diets. It is still possible that diet W might be much better or much worse than diet J. Suppose another researcher is able to conclude that, on average, people on diet W might lose anywhere from a pound more to a half pound less than those on diet J. This researcher, too, has failed to show a difference between the diets. However, this researcher is entitled to say there is no difference between the diets because here the difference, whatever it might actually be, is of no practical importance.
>
> This example demonstrates why it is essential that the analyst report all effects that are consistent with the data when no effect has been shown. Only if none of the possibilities are of any practical importance may the analyst claim "no effect" has been demonstrated.

If these hints to the inner workings of statistics can be kept in mind, the rest really *is* easy!

As with any skill, practice makes perfect. The reason seasoned analysts can easily dismiss a data set that might confound novices is that the experienced analysts have seen it all before...many times! This excerpt from *The Learning Curve* by Atul Gawande (The New Yorker, January 28, 2002, pp 52-61) speaks directly to the importance of practice.

> There have now been many studies of elite performers--concert violinists, chess grandmasters, professional ice-skaters, mathematicians, and so forth--and the biggest difference researchers find between them and lesser performers is the amount of deliberate practice they've accumulated. Indeed, the most important talent may be the talent for practice itself. K.Anders Ericsson, a cognitive psychologist and expert on performance, notes that the most important role that innate factors play may be in a person's *willingness* to engage in sustained training. He has found, for example, that top performers dislike practicing just as much as others do. (That's why, for example, athletes and musicians usually quite practicing when they retire.) But, more than others, they have the will to keep at it anyway.

I and others are good at what we do because we keep doing it over and over (and over and over until we get it right!). Persevere and you will succeed. For students, this means working every problem and dataset at their disposal. For those who have completed enough coursework to let them work with data, this means analyzing data every time the opportunity presents itself.

[back to LHSP]

Last modified: undefined.

Cause & Effect

"Cause and Effect"! You almost never hear these words in an introductory statistics course. The subject is commonly ignored. Even on this site, all it gets is this one web page. If cause and effect is addressed at all, it is usually by giving the (proper) warning *"Association does not imply causation!"* along with a few illustrations. For example, in the early part of the twentieth century, it was noticed that, when viewed over time, the number of crimes increased with membership in the Church of England. This had nothing to do with criminals finding religion. Rather, both crimes and Church membership increased as the population increased. *Association does not imply causation!* During WWII it was noticed that bombers were *less* accurate when the weather was *more* clear. The reason was that when the weather was clear there was also more opposition from enemy fighter planes. *Association does not imply causation*, at least not necessarily in the way it appears on the surface!

We laugh at obvious mistakes but often forget how easy it is to make subtle errors *any* time an attempt is made to use statistics to prove causality. This could have disastrous consequences if the errors form the basis of public policy. This is nothing new. David Freedman ("From Association to Causation: Some Remarks on the History of Statistics," Statistical Science, 14 (1999),243-258) describes one of the earliest attempts to use regression models in the social sciences. In 1899, the statistician G. Udny Yule investigated the causes of pauperism in England. Depending on local custom, paupers were supported inside local poor-houses or outside. Yule used a regression model to analyze his data and found that the change in pauperism was positively related to the change in the proportion treated outside of poor-houses. He then reported that welfare provided outside of poor-houses created paupers. A contemporary of Yule's suggested that what Yule was seeing was instead an example of confounding--those areas with more efficient administrations were better at both building poor-houses and reducing poverty. That is, if efficiency could be accounted for, there would be no association between pauperism and the way aid was provided. Freedman notes that after spending much of the paper assigning parts of the change in pauperism to various causes, Yule left himself an out with his footnote 25: "Strictly speaking, for 'due to' read 'associated with'."

Discussions of cause & effect are typically left to courses in study design, while courses in statistics and data analysis have focused on statistical techniques. There are valid historical reasons for this. Many courses are required to earn a degree in statistics. As long as all of the principles were covered, it didn't matter (and was only natural) that some courses focused solely on the theory and application of the techniques. When the first introductory statistics courses were taught, they either focused on elementary mathematical theory or were "cookbook" courses that showed students how to perform by hand the calculations that were involved in the more commonly used techniques. There wasn't time for much else.

Today, the statistical community generally recognizes that these approaches are inappropriate in an era when anyone with a computer and a statistical software package can attempt to be his/her own statistician. "Cause & effect" must be among the first things that are addressed because this is what most people will use statistics for! Newspapers, radio, television, and the Internet are filled with claims based on some form of statistical analysis. Calcium is good for strong bones. Watching TV is a major cause of childhood and adolescent obesity. Food stamps and WIC improve nutritional status. Coffee consumption is responsible for heavens knows what! All because someone got hold of a dataset from somewhere and looked for associations. Which claims should be believed? Only by understanding what it takes to establish causality do we have any chance of being intelligent consumers of the "truths" the world throws at us.

Freedman points out that statistical demonstrations of causality are based on assumptions that often are not checked adequately. "If maintained hypotheses A,B,C,... hold, then H can be tested against the data. However, if A,B,C,... remain in doubt, so must inferences about H. Careful scrutiny of maintained hypotheses should therefore be a critical part of empirical work--a principle honored more often in the breach than in the observance." That is, an analysis could be exquisite and the logic could be flawless **provided A,B,C hold** but the same attention is rarely paid to checking A,B,C as goes into the analysis that assumes A,B,C hold.

The rules for claiming causality vary from field to field. The physical sciences seem to have the easiest time of it because it is easy to design experiments in which a single component can be isolated and studied. Fields like history have the hardest time of it. Not only are experiments all but impossible, but observations often play out over generations, making it difficult to collect new data, while much of the existing data is often suspect. In a workshop on causality that I attended, a historian stated that many outrageous claims were made because people often do not have the proper foundations in logic (as well as in the subject matter) for making defensible claims of causality. Two examples that were offered, (1) Two countries that have McDonalds restaurants have never gone to war. [except for the England and Venezuela!] (2) Before television, two World Wars; after television, no World Wars. In similar fashion, one of my friends recently pointed out to his girlfriend that he didn't have any grey hairs until after he started going out with her...which is true but he's in his late 30s and they've been seeing each other for 3 years. I suppose it *could* be the relationship...

Statisticians have it easy, which perhaps is why statistics courses don't dwell on causality. Cause and effect is established through the intervention trial in which two groups undergo the same experience except for a single facet. Any difference in outcome is then attributed to that single facet.

In epidemiology, which relies heavily on observational studies (that is, taking people as you find them), cause and effect is established by observing the same thing in a wide variety of settings until all but the suspected cause can be ruled out. The traditional approach is that given by Bradford Hill in his Principles of Medical Statistics (first published in 1937; 8th edition 1966). He would have us consider the strength of the association, consistency (observed repeatedly by different persons, in different circumstances and times), specificity (limited to specific sets of characteristics), relationship in time, biological gradient (dose response), biological plausibility (which is the weak link because it depends on the current state of knowledge), and coherence of the evidence.

The classic example of an epidemiological investigation is John Snow's determination that cholera is a waterborne infectious disease. This is discussed in detail in every introductory epidemiology course as a standard against which all other investigations are measured. It is also described in Freedman's article.

A modern example is the link between smoking and lung cancer. Because is it impossible to conduct randomized smoking experiments in human populations, it took many decades to collect enough observational data (some free of one types of bias, others free of another) to establish the connection. Much of the observational evidence is compelling. Studies of death rates show lung cancer increasing and lagging behind smoking rates by 20-25 years while other forms of cancer stay flat. Smokers have lung cancer and heart disease at rates greater than the nonsmoking population even after adjusting for whatever potential confounder the tobacco industry might propose. However, when smoking was first suspected of causing lung cancer and heart disease, Sir Ronald Fisher, then the world's greatest living statistician and a smoker, offered the "constitution hypothesis" that people might be genetically disposed to develop the diseases and to smoke, that is, that genetics was confounding the association. This was not an easy claim to put to an experiment. However, the hypothesis was put to rest in a 1989 Finnish study of 22 smoking-discordant monozygotic twins where at least one twin died. There, the smoker died first in 17 cases. In the nine pairs where death was due to coronary heart disease, the smoker died first in every case.

As connections become more subtle and entangled, researchers tend to rely on complicated models to sort them out. Freedman wrote, "Modern epidemiology has come to rely more heavily on statistical models, which seem to have spread from the physical to the social sciences and then to epidemiology." When I first picked up the article and glanced quickly at this sentence, I misread it as, "Modern epidemiology has come to rely more heavily on statistical models than on epidemiology!" I may have misread it, but I don't think I got it entirely wrong.

As a group, those trained in epidemiology are among the most scrupulous about guarding against false claims of causality. Perhaps I can be forgiven my mistake in an era when much epidemiology is practiced by people without proper training who focus on model fitting, ignore the quality of the data going into their models, and rely on computers and complex techniques to ennoble their results. When we use statistical models, it is essential to heed Freedman's warning about verifying assumptions. It is especially important that investigators become aware of the assumptions made by their analyses. Some approaches to causality are so elaborate that basic assumptions about the subject matter may be hidden to all but those intimately familiar with the underlying mathematics, but this is NEVER a valid excuse for assuming that what we don't understand is unimportant.

A good statistician will point out that causality can be proven only by demonstrating a mechanism. Statistics alone can never prove causality, but it can show you where to look. Perhaps no example better illustrates this than smoking and cancer/heart disease. Despite all of the statistical evidence, the causal relationship between smoking and disease will not be nailed down by the numbers but by the identification of the substance in tobacco that trigger the diseases.

Last modified: undefined.

[Back to *The Little Handbook of Statistical Practice*]

Some Aspects of Study Design

Gerard E. Dallal, Ph.D.

Introduction

100% of all disasters are failures of design, not analysis.
-- Ron Marks, Toronto, August 16, 1994

To propose that poor design can be corrected by subtle analysis techniques is contrary to good scientific thinking.
-- Stuart Pocock (Controlled Clinical Trials, p 58) regarding the use of retrospective adjustment for trials with historical controls.

Issues of design always trump issues of analysis.
-- GE Dallal, 1999, explaining why it would be wasted effort to focus on the analysis of data from a study whose design was fatally flawed.

Bias dominates variability.
-- John C. Bailler, III, Indianapolis, August 14, 2000

Statistics is not just a collection of computational techniques. It is a way of thinking about the world. Anyone can take a set of numbers and apply some formulas to them. There are many computer programs that will do the calculations for you. But there is no point to analyzing data from a study that was not properly designed to answer the research question under investigation. In fact, there's a real point in *refusing* to analyze such data lest faulty results be responsible for implementing a program or policy contrary to what's really needed.

Two of the most valuable things a researcher can possess are knowledge of the principles of good study design and the courage to refuse to cut corners (to make a study more attractive to a funding agency or less inconvenient to the researcher, for example).

The Basics of Study Design

Prologue, Part 1: Statistics is about ***a whole lot of nothing!***

The older I get and the more I analyze data, the more I appreciate that some of the most serious mistakes are made because many researchers failure to understand that classical or frequentist statistics is based on ***"A whole lot of nothing"***.

To repeat the first point of "Is Statistics Hard?", *Statistics is backwards!* To show that an effect or difference exists, classical or frequentist statistics begins by asking what would happen if there were no effect. That is, statistics studies *a whole lot of nothing*! The analyst compares study data to what is expected when there is *nothing*. If the data are not typical of what is seen when there is *nothing*, there **must** be *something*!

Usually "not typical" means that some summary of the data is so extreme that it is seen less than 5% of the time. This is where the problem creeps in.

Statistical methods protect the researcher with a carefully crafted research question and a clearly specified response measure. The chance of seeing an atypical outcome is small when there's *nothing*. However, when the question is vague or there are many ways to evaluate it, statistical methods work against the researcher who uses the same criteria as the researcher with a well-defined study. When the research question is vague or there are many possible response measures, researchers invariably "look around" and perform **many** evaluations of the data. The same statistical methods now guarantee that, when there is no effect, 5% of such investigations will suggest that there *is* an effect.

To put it another way:

- The researcher with a well specified question and outcome measure has only a 5% chance of claiming an effect when there isn't any. (THIS IS **GOOD!**)
- The researcher with a vague question or many outcome measures will certainly find a measure that suggests some kind of an effect when there is none, if s/he continues to come up with different ways of checking the data. (THIS IS **BAD!**)

Prologue, Part 2: There are some general principles of study design that can be offered. However, the specifics can only be learned anecdotally. Every field of study has its peculiarlites. Some things that are major issues in one field may never be encountered in another. Many of the illustiations in these notes are nutrition related because that's where I've done most of my work. If you ask around, others will be only too happy to share horror stories from their own fields.

Now, onto the basics of study design.

(1) There must be a fully formed, **clearly stated research question** and **primary outcome measure**.

- What do we wish to learn?
- Why are we conducting this research?

Focus is critical. There must be a clear goal in mind. Otherwise, time, energy, and money will be invested only to find that nothing has been accomplished. A useful approach is to ask early on if, at the end of the project, only one question could be answered, what would that question be? (Other variations are, "If the results were summarized at the top of the evening news in a phrase or two spoken in a few seconds, what would the reporter say?" or "If the results were written up in the local newspaper, what would the headline be?") Not only does this help a statistician better understand an investigator's goals, but sometimes it forces the investigator to do some serious soul-searching.

Be skeptical of reported results that were not part a study's original goals. Sometimes they are important, but often they are an attempt to justify a study that did not work out as hoped or intended. When many responses and subgroups are examined, statistical theory guarantees that some of them will appear to be statistically significant on the surface, These results should not be given the same status as the results from analyses directed toward the primary research question.

> Suppose I see someone with a coin and a pair of pliers. The coin doesn't look quite right. When it is subsequently flipped there are 65 heads out of 100 tosses, I suspect the coin is no longer fair. It's not impossible for a fair coin to show 65 heads in 100 tosses. However, statistical theory says that only 2 fair coins in 1,000 will show 65 or more heads in 100 tosses. So, why now, if the coin is fairl? Therefore, the results are suspect. On the other hand, if I go to the bank and get $250 in quarters that just arrived from the mint, and flip each coin 100 times, I'm not surprised if one or two coins shows 65 or more heads. Probability theory says that, on average, 2 out of 1,000 coins will show 65 or more heads in 100 tosses. If it **never** happened, **then** I should worry! This illustrates that **a result that is too extreme to be typical behavior in one case *is* typical behavior in another set of circumstances**.
>
> Another (almost certainly apocryphal) example involves a student accused of cheating on a standardized test. Her scores had increased 200 points between two successive tests over a short period of time. The testing organization withheld the new grade and the student took them to court to force them to release it. The hearing is supposed to have gone something like this:
>
>> "A jump like that occurs only in 1 out of 50,000 retests."
>> "Was there any reason to question her performance other than the rise in score? Did the proctor accuse her of cheating during the exam?"
>> "No."
>> "How many took a retest?"
>> "About 50,000."
>> "Then, release the score."
>
> Had the proctor suspecteed cheating, then it would have been quite a coincidence for that student to be the 1 out of 50,000 to have such a rise, but it is NOT surprising that it happened to someone, somewhere when 50,000 took a retest. Once again, **a result that is too extreme to be typical behavior in one case *is* typical behavior in another set of circumstances** The chances of winning a lottery are small, yet there's always a winner.
>
> The same thing applies to research questions. Suppose an investigator has a theory about a specific causal agent for a disease. If the disease shows an association with the causal

> agent, her theory is supported. However, if the same degree of association is found only by sifting through dozens of possible agents, the amount of support for that agent is greatly diminished. Once again, statistical theory says that if enough potential causal agents are examined, a certain proportion of those unrelated to the disease will seem to be associated with the disease if one applies the criterion appropriate for a fully-formed research question regarding a single specified causal agent.

As a member of review committee, I could not approve your proposal if it did not contain both a fully formed research question and a clearly stated outcome measure.

(2) The project must be feasible. This refers not only to resources (time and money), but also to whether there is agreement on the meaning of the research question and to whether everything that needs to be measured can be measured.

If the study involves some condition, can we define it? Can we be sure we'll recognize it when we see it?

- What is an unhealthy eating behavior?
- What is crop yield--what gets harvested or what makes it to market?
- What's the difference between a cold and the flu? For that matter, what is a cold? In your spare time, try to come up with a method that easily distinguishes a cold from an allergy.
- How many methods are there for measuring things like usual dietary intake or lean body mass? Do they agree? Does it matter which one we use?
- What do we mean by *family income* or *improved nutritional status*?

How accurate are the measurements? How accurate do they need to be? What causes them to be inaccurate?

- Calcium intake is easy to measure because there are only a few major sources of calcium. Salt intake is hard because salt is everywhere.
- With dietary intake we've our choice of food diaries where food are recored as they are eaten, 24 hour recalls, and food frequencies (which ask about typical behavior over some previous time period, typically one year), each with different strengths and weaknesses.
- HDL-cholesterol (the "good cholesterol") must be measured almost immediately after blood is drawn because cholesterol migrates between the various lipoproteins. Also, the lab isn't being inefficient when it has you waiting to have your blood drawn. HDL levels are affected by activity immediatly prior to taking blood
- Total body potassium is the best measurement of lean body mass at the moment. Is it good enough?
- Will respondents reveal their income? Will they admit to being in want? Or, will they shade the truth either to please the interviewer, for reasons of pride, or for fear that revealing their true income might somehow lead to the loss of benefits that depend on not exceeding a certain threshhold?

How do we choose among different measurement techniques?

- Is a mechanical blood pressure cuff better than using a stethoscope? What if the cuff breaks? What if the technician quits?
- Is bioelectric impedance (BIA) as good as underwater weighing for determining body fat? Why might we use BIA regardless?
- Should dietary intake be measured by food frequency questionnaire, weighed diet record, or dietary recall?
- Is there a gold standard? Do we need all that gold; is it worth paying for?

Are we measuring what we think we're measuring?

- Parents might take food assistance intended for them and give it to their children. Food assistance intended for children might instead be given to the bread-winner. Food consumption measurements in such cases will not be what they might seem to be.
- Calcification of the abdominal aorta can be misinterpreted as higher bone densities depending on the measurement technique.

Can measurements be made consistently, that is, if a measurement is made twice will we get the same number? Can others get the same value (inter-laboratory, inter-technician variability)? What happens if different measurement techniques are used within a particular study (the x-ray tube breaks, the radioactive source degrades, supplies of a particular batch of reagent are exhausted)?

Sometimes merely measuring something changes it in unexpected ways.

- Does asking people to keep records of dietary intake cause them to change their intake? A few years ago, I got into an elevator to hear one sudent invite another to dinner. The invitation was declined because, the student being invited explained, she was keeping a food diary and it would be too much trouble!
- Are shut-ins more likely to provide responses that they believe will prolong the interview or result in return visits in order to have outside contact? This was encountered in one study when a nurse became suspicious of a subject who was giving what felt like deliberately vague responses when asked whether she had experienced certain symptoms since his last visit. When the nurse assured the subject that he would be visting her every week for the duration of the study regardless of her answers, she immediately reported being symptom free.

Sometimes the answers to these questions say that a study should not be attempted. Other times the issues are found to be unimportant. Many of these questions can be answered only by a subject matter specialist. A common mistake is to go with one's instinct, if only to keep personnel costs down. However, it is essential to assemble a team with the appropriate skills if only to convince funding agencies that their money will be well spent. Even more important, the study will lack credibilty if people with critical skills were not involved in its planning and execution.

A note about **bias**: Bias is an amount by which all measurements are deflected from their true value. For example, a particular technician might produce blood pressure readings that are consistently 5 mm higher than they should be. If this technician makes all of the measurements, then changes over time and differences between groups can be estimated without error because the bias cancels out. In similar fashion, food frequency questionnaires might underestimate total energy intake, but if they underestimate everyone in the same way (whatever that means!), comparisons between groups of subjects will still be valid.

If a technician or method is replaced during a study, some estimates become impossible while others are unaffected. Suppose one technician takes all of the baseline measurements and a second takes all of the followup measurements. If one technician produces biased measurements, we cannot produce valid estimates of a group's change over time. However, we can reliably compare the change over time between two groups, again because the bias cancels out. (If it were important to estimate the individual changes over time--that is, account for possible bias between technicians or measuring devices--the two technicians might be asked to analyze sets of split samples in order to estimate any bias that might be present.)

(3) Every data item and every facet of the protocol must be carefully considered.

All of the relevant data must be collected. If a critical piece of data cannot be obtained, perhaps the study should not be undertaken.

It is equally important to guard against collecting data unrelated to the research question. It is too easy to overlook how quickly data can multiply and bog down a study, if not destroy it. Many years ago, an outstanding doctoral student spent nearly six months developing and implementing a coding scheme for the many herbal teas in a set of food records she had collected. This was done without any sense of what would be done with the data. In the end, they were never used.

Be sure the cost in time and effort of each item is clearly understood. I suspect the urge to collect marginally related data comes from a fear that something might be overlooked ("Let's get it now, while we can!"), but another study can usually be conducted to tie up promising loose ends if they're really that promising. In general, if there is no solid analysis plan for a particular piece of data, it should not be collected.

Treatments must be clearly identified.

- a drug in syrup includes the syrup
- every drug brings a placebo effect with it
- animals know they've been stuck with a needle
- cells know they've been bathed in something and disturbed

Is the active ingredient what you think/hope it is or was the infusing instrument contaminated? Was there something in the water? In essence, anything that is done to subjects may be responsible for any observed effects, and something must be done to rule out those possibilities that aren't of interest (for example, socialization and good feelings or heightened sensitivity to the issue being studied that come from participation). Things aren't always what they appear to be. It is not unheard of for pills to be mislabeled, for purported placebos to contain substances that affect the primary outcome measurement (always assay every lot of every treatment), or for subjects to borrow from each other.

Sometimes convenience can make us lose sight of the bigger picture. Is there any point in studying the effects of a nutritional intervention or *any* health intervention in a population that starts out healthy? Is it ethical to study an unhealthy population? (There's a whole course here!)

(4) Keep it simple!

With the advances in personal computing and statistical program packages, it often seems that no experiment or data set is too complicated to analyze. Sometimes researchers design experiments with dozens of treatment combinations. Other times they attempt to control for dozens of key varibles. Sometimes attempts are made to do both--study dozens of treatment combinations **and** adjust for scores of key variables!

It's not that it can't be done. Statistical theory doesn't care how many treatments or adjustments are involved. The issue is a practical one. I rarely see studies with enough underlying knowledge or data to pull it off.

The aim of these complicated studies is a noble one--to maximize the use of resources--but it is usually misguided. I encourge researchers to study only two groups at once, if at all possible. When there are only two groups, the research question is sharply focused. When many factors are studied simultaneously, it's often difficult to sort things out, especially when the factors are chosen haphazardly. (Why should this treatment be less effective for left-handed blondes?) Just as it's important to learn to crawl before learning to walk, the joint behaviour of multiple factors should be tackled only after gaining a sense of how they behave individually. Besides, once the basics are known, it's usually a lot easier *to get funding* to go after the fine details!

That's not to say that studies involving many factors should be never be attempted. It may be critically important, for example, to learn whether a treatment is less effective for females than for males. However, there should be a strong, sound, overarching theoretical basis if a study of the joint behavior of multiple factors is among the first investigations proposed in a new area of study.

By way of example: Despite all of the advice you see today about the importance of calcium, back in the 1980s there was still some question about the reason older women had brittle bones. Many thought it was due to inadquate calcium intake, but others suggested that older women's bodies had lost the ability to use dietary calcium to maintain bone health. Studies up to that time had been contradictory. Some showed an effect of supplementation; others did not. Dr. Bess Dawson-Hughes and her colleagues decided to help settle the issue by keeping it simple. They looked only at women with intakes of less than half of the recommended daily allowance of calcium. The thought was that if calcium supplementation was of any benefit, this group would be most likely to show it. If calcium supplements didn't help these women, they probably wouldn't help anyone. They found a treatment effect and went on to study other determinants of bone health, such as vitamin D. However, they didn't try to do it all at once.

(5) **Research has consequences!**

Research is usually conducted with a view toward publication and dissemination. When results are reported, not only will they be of interest to other researchers, but it is likely that they will be noticed by the popular press, professionals who deal with the general public, and legislative bodies--in short, anyone who might use them to further his/her personal interests.

You ***must*** be aware of the possible consequences of your work. Public policy may be changed. Lines of inquiry may be pursued or abandoned. If a program evaluation is attempted without the ability to detect the type of effect the program is likely to produce, the program could become targeted for termination as a cost-savings measure when the study fails to detect an effect. If, for expediency, a treatment is evaluated in an inappropriate population, research on that treatment may improperly come to a halt or receive undeserved further funding when the results are reported.

One might seek comfort from the knowledge that the scientific method is based on replication. Faulty results will not replicate and they'll be found out. However, the first report in any area often receives special attention. If its results are incorrect because of faulty study design, many further studies will be required before the original study is adequately refuted. If the data are expensive to obtain, or if the original report satisfies a particular political agenda, replication may never take place.

> The great enemy of the truth is very often not the lie--deliberate, contrived and dishonest--but the myth--persistent, persuasive and unrealistic. --*John F. Kennedy*

These basic aspects of study design are well-known, but often their importance is driven home only after first-hand experience with the consequences of ignoring them. Computer programers say that you never learn from the programs that run, but only from the ones that fail. The Federal Aviation Administration studies aircraft "incidents" in minute detail to learn how to prevent their recurrence. Learn from others. One question you should always ask is how a study could have been improved, regardless of whether it was a success or a failure.

Three useful references that continue this discussion are

> *Feynman R (1986), "Cargo Cult Science,"* reprinted in *"Surely You're Joking, Mr. Feynman!" New York: Bantam Books.*

Moses L (1985), "Statistical Concepts Fundamental to Investigations," The New England Journal of Medicine, 312, 890- 897.
Pocock S (1983), Clinical Trials, New York: John Wiley & Sons.

Types of Studies

Different professions classify studies in different ways. Statisticians tend to think of studies as being of two types: *observational studies* and *intervention trials*. The distinction is whether an intervention is involved, that is, whether the investigator changes some aspect of subjects' behavior. If there's an intervention--assigning subjects to different treatments, for example--it's an intervention trial (sometimes, depending on the setting, called a controlled clinical trial). If there's no intervention--that is, if subjects are merely observed--it's an observational study.

Observational Studies

Take a good epidemiology course! Statisticians tend to worry about issues surrounding observational studies in general terms. Epidemiologists deal with them systematically and have a name for everything! There are *prospective studies*, *retrospective studies*, *cohort studies*, *nested case-control studies*, among others.

Epidemiologists also have good terminology for describing what can go wrong with studies.

- The statistician will say that there's a problem because women who have breast cancer are more likely to remember their history of x-ray exposure while the epidemiologist will say that there's a problem with *recall bias*.
- Statisticians will say that there's a problem comparing treatments in a clinical setting because everyone is given the treatment that is optimal for him or her. The fact that everyone turns out the same doesn't mean that there's no difference between treatments. Rather, it means that physicians are very good at selecting their patients' treatments. An epidemiologist would say that there's a problem with *confounding by indication*.

Statisticians tend to spell out issues in all their glory while epidemiologists capture them in a single phrase. Learn the terminology. Take the course.

Surveys

The survey is a kind of observational study because no intervention is involved. The goal of most surveys is to draw a sample of units from a larger population, measure them, and make statements about the population from which the sample was drawn. For example,

- we might interview a sample of female Boston area high school students about their eating habits in the hope of describing the eating disorders of all female Boston area high school students

- we might interview a sample of families on welfare to learn about the extent of hunger in all families on welfare
- we might survey farmers to learn about awareness of integrated pest management

The analysis of survey data relies on samples being *random samples* from the population. The methods discussed in an introductory statistics course are appropriate when the sample is a *simple random sample*. The formal definition says a sample is a simple random sample if every possible sample had the same chance of being drawn. A less formal-sounding but equally rigorous definition says to draw a simple random sample, write the names of everyone in the population on separate slips of paper, mix them thoroughly in a big box, close your eyes, and draw slips from the box to determine whom to interview. It's a *sample* because it's drawn from the larger population. It's *random* because you've mixed thoroughly and closed your eyes. It's *simple* because there's just one container.

When isn't a random sample *simple*? Imagine having two boxes--one for the names of public high school students, the other for the names of private high school students. If we take separate random samples from each box, it is a *stratified random sample*, where the strata are the types of high schools. In order to use these samples to make a statement about all Boston area students, we'd have to take the total numbers of public and private school students into account, but that's a course in survey sampling and we won't pursue it here.

Sometimes the pedigree of a sample is uncertain, yet standard statistical techniques for simple random samples are used regardless. The rationale behind such analyses is best expressed in a reworking of a quotation from Stephen Fienberg (in which the phrases *contingency table* and *multinomial* have been replaced by *survey* and *simple random*):

> "It is often true that data in a [survey] have not been produced by a [simple random] sampling procedure, and that the statistician is unable to determine the exact sampling scheme which was used. In such situations the best we can do, usually, is to assume a [simple random] situation and hope that it is not *very* unreasonable."

This does not mean that sampling issues can be ignored. It says that in some instances we may decide to treat data as though they came from a simple random sample as long as there's no evidence that such an approach is inappropriate.

Why is there such concern about the way the sample was obtained? With only slight exageration, **if a sample isn't random, statistics can't help you!** We want samples from which we can generalize to the larger population. Some samples have obvious problems and won't generalize to the population of interest. If we were interested in the strength of our favorite candidate in the local election we wouldn't solicit opinions outside her local headquarters. If we were interested in general trends in obesity, we wouldn't survey just health club members. But, why can't we just measure people who seem, well... reasonable?

We often think of statistical analysis as a way to estimate something about a large population. It certainly does this. However, the real value of statistical methods is their ability to describe the uncertainty in the estimates, that is, the extent to which samples can differ from the populations from which they are drawn. For example, suppose in random samples of female public and public high school students 10% more private school students have eating disorders. What does this say about *all* public and private female high school students? Could the difference be as high as 20%? Could it be 0, with the observed difference being "just one of those things"? If the samples were drawn by using probability-based methods, statistics can answer these questions. If the samples were drawn in a haphazard fashion or as a matter of convenience (the members of the high school classes of two acquaintances, for example, or the swim teams) statistics can't say much about the extent to which the sample and population values differ.

The convenience sample--members of the population that are easily available to us--is the antithesis of the simple random sample. Statistics can't do much beyond providing descriptive summaries of the data because the probability models relating samples to populations do not apply. It may still be possible to obtain useful information from such samples, but great care must be exercised when interpreting the results. One cannot simply apply standard statistical methods as though simple random samples had been used. You often see comments along the lines of "These results may be due to the particular type of patients seen in this setting." Just look at the letters section of any issue of *The New England Journal of Medicine*.

Epidemiologists worry less about random samples and more about the comparability of subjects with respect to an enrollment procedure. For example, suppose a group of college students was recruited for a study and classified as omnivores or vegans. There is no reason to expect any statement about *these* omnivores to apply to *all* omnivores or that a statement about *these* vegans to apply to *all* vegans. However, if subjects were recruited in a way that would not cause omnivores and vegans to respond to the invitation differently, we might have some confidence in statistical analyses that compare the two groups, especially if differences between omnivores and vegans in this college setting were seen in other settings, such as working adults, retirees, athletes, and specific ethnic groups.

Cross-sectional studies vs longitudinal studies

> A **cross-sectional study**involves a group of people observed at a single point in time. (Imagine a lot of lines on a plot with time as the horizontal axis and each subject's values as a different line and taking a slice or *cross-section*at a particular point in time.)
>
> A **longitudinal study** involves the same individuals measured over time (or a*long* the time line).

It is often tempting to interpret the results of a cross-sectional study as though they came from a longitudinal study. Cross-sectional studies are faster and cheaper than longitudinal studies, so there's little wonder that this approach is attractive. Sometimes it works; sometimes it doesn't. But, there's no way to know whether it will work simply by looking at the data.

- If you take a sample of people at the same point in time and plot their cholesterol levels against their % of calories from saturated fat, the values will tend to go up and down together and it's probably safe to assume that the relationship is true for individuals over time. What we have observed is that those individuals with higher fat intakes tend to have higher cholesterol levels, but our knowledge of nutrition suggests that an individual's cholesterol level would tend to go up and down with fat intake.
- On the other hand, consider a plot of height against weight for a group of adults. Even though height and weight tend to go up and down together (taller people tend to be heavier and vice-versa) the message to be learned is NOT "eat to grow tall"!

When faced with a new situation, it may not be obvious whether cross-sectional data can be treated as though they were longitudinal. In cross-sectional studies of many different populations, those with higher vitamin C levels tend to have higher HDL-cholesterol levels. Is this an anti-oxidant effect, suggesting that an increase in vitamin C will raise HDL-cholesterol levels and that the data can be interpreted longitudinally, or are both measurements connected through a non-causal mechanism? Perhaps those who lead a generally healthy life style have both higher HDL-cholesterol and vitamin C levels. **The only way to answer a longitudinal question is by collecting longitudinal data.**

Even longitudinal studies must be interpreted with caution. Effects seen over the short term may not continue over the long term. This is the case with bone remodeling where gains in bone density over one year are lost over a second year, despite no obvious change in behavior.

Cohort Studies / Case-Control Studies

[This discussion is just the tip of the iceberg. To examine these issues in depth, find a good epidemiology course. Take it!]

In **cohort studies**, a well-defined group of subjects is followed. Two well-known examples of cohort studies are the Framingham Heart Study, which follows generations of residents of Framingham, Massachusetts, and the Nurses Health Study in which a national sample of nursing professionals is followed through yearly questionnaires. However, cohort studes need not be as large as these. Many cohort studies involve only a few hundred or even a few dozen individuals. Because the group is well-defined, it is easy to study associations within the group, such as between exposure and disease. However, cohort studies are not always an effective way to study associations, particularly when an outcome such as disease is rare or takes a long time to develop.

Case-control studies were born out of the best of intentions. However, they prove once again the maxim that *the road to Hell is paved with good intentions*. In a case-control study, the exposure status of a set of cases is compared to the exposure status of a set of controls. For example, we might look at the smoking habits of those with and without lung cancer. Since we start out with a predetermined number of cases, the rarity of the disease is no longer an issue.

Case-control studies are fine from a mathematical standpoint. However, they present a nearly insurmountable practical problem--the choice of controls. For example, suppose a study will involve cases of stomach cancer drawn from a hospital's gastrointestinal service. Should the controls be healthy individuals from the community served by the hospital, or should they be hospital patients without stomach cancer? What about using only patients of hospital's GI service with complaints other than stomach cancer? There is no satisfactory answer because no matter what group is used, the cases and controls do not represent random samples from any identifiable population. While it might be tempting to "assume a [simple random] situation and hope that it is not *very* unreasonable" there are too many instances where series of case-control studies have failed to provide similar results. Because of this inability to identify a population from which the subjects were drawn, many epidemiologists and statisticians have declared the case-control study to be inherently flawed.

There is one type of case-control study that everone finds acceptable--the ***nested*** **case-control study**. A *nested case-control study* is a case-control study that is nested (or embedded) within a cohort study. The cases are usually all of the cases in the cohort while the controls are selected at random from the non-cases. Since the cohort is well-defined, it is appropriate to compare the rates of exposure among the cases and controls.

It is natural to ask why all non-cases are not examined, which would allow the data to be analyzed as coming from a cohort study. The answer is "resources". Consider a large cohort of 10,000 people that contains 500 cases. If the data are already collected, it's little more work for a computer to analyze 10,000 cases than 1,000, so that data should be analyzed as coming from a cohort study. However, the nested case-control study was developed for those situations where new data would have to be generated. Perhaps blood samples would have to be taken from storage and analyzed. If 500 controls would provide almost as much information as 9,500, it would be wasteful to analyze the additional 9,000. Not only would time and money be lost, but blood samples that could be used for other nested case-control studies would have been destroyed needlessly.

Intervention trials/Controlled clinical trials

Randomized! Double-blind! Controlled!

When the results of an important intervention trial are reported in a highly-regarded, peer-reviewed journal, you will invariably see the trial described as *randomized*, *double-blind,* and (possibly *placebo) controlled*.

Suppose you have two treatments to compare. They might be two diets, two forms of exercise, two methods of pest control, or two ways to deliver prenatal care. How should you design your study to obtain a valid comparison of the two treatments? Common sense probably tells you, correctly, to give the treatments to two groups of comparable subjects and see which group does better.

How do we make our groups of subjects comparable? Who should get what treatment, or, as a trained researcher would put it, how should subjects be assigned to treatment? It would be dangerous to allow treatments to be assigned in a deliberate fashion, that is, by letting an investigator choose a subject's treatment. If the investigator were free to choose, any observed differences in outcomes might be due to the conscious or unconscious way treatments were assigned. Unscrupulous individuals might deliberately assign healthier subjects to a treatment in which they had a financial interest while giving the other treatment to subjects whom nothing could help. Scrupulous investigators, eager to see their theories proven, might make similar decisions unconsciously.

Randomized means that subjects should be assigned to treatment at random, so that each subject's treatment is a matter of chance, like flipping a coin. If nothing else, this provides **insurance** against both conscious and unconscious bias. Not only does it insure that the two groups will be similar with respect to factors that are known to effect the outcome, but also it balances the groups with respect to unanticipated or even unknown factors that might influence the outcome had purposeful assignments been used.

Sometimes randomization is unethical. For example, subjects cannot be randomized to a group that would undergo a potentially harmful experience. In this case, the best we can do is compare groups that choose the behavior (such as smoking) with those who choose not to adopt the behavior, but these groups will often differ in other ways that may be related to health outcomes. When subjects cannot be randomized, studies are viewed with the same skepticism accorded to surveys based on nonrandom samples.

> Resolution of the relation between lung cancer and cigarette smoking was achieved after a host of studies stretching over many years. For each study suggesting an adverse effect of smoking, it was possible to suggest some possible biases that were not controlled, thus casting doubt on the indicated effect. By 1964, so many different kinds of studies had been performed--some free of one type of bias--some of another--that a consensus was reached that heavy cigarette smoking elevated the risk of lung cancer. (Ultimately, the effect was recognized to be about a 10-fold increase in risk.) The story shows that induction in the absence of an applicable probability model is possible, but that induction in those circumstances can be difficult and slow.

According to The Tobacco Institute, the question of second hand smoke is still *far* from settled. Compare the long struggle over lung cancer and smoking to the one summer it took to establish the efficacy of the Salk polio vaccine.

On the other hand, it may not be unreasonable to randomize subjects away from potentially unhealthful behavior. If coffee drinking were thought to have a negative impact on some measure of health status, it would be unethical for a study to have coffee consumed by those who did not normally drink it. However, it might be ethical to ask coffee drinkers to give it up for the duration of the study. (The "might" here refers not so much to this hypothetical study but to others that might seem similar on the surface. For example, to study the cholesterol lowering property of some substance, it seems reasonable to work with subjects who already have elevated cholesterol levels. However, ethical behavior dictates that once subjects are identified as having elevated levels they should be treated according to standard medical practice and not studied!)

Randomization sounds more mysterious than it really is in practice, It can be as simple as assigning treatment by flipping a coin. You can generate a randomization plan automatically at Randomization.com. Specify the names of the treatments and the number of subjects and the script will produce a randomized list of treatments. As subjects are enrolled into the study, they are given the next treatment on the list.

Blinded means blind with respect to treatment. In a *single blind study*, subjects do not know what treatment they are receiving. This insures their responses will not be affected by prior expectations and that subsequent behavior will not be affected by knowledge of the treatment. In some rare instances, *single blind* refers to situations where subjects know their treatments and only the person evaluating them is blinded. In a *double blind study*, subjects and anyone who has contact with them or makes judgments about them is blinded to the assignment of treatments. This insures subjective judgments will not be affected by knowledge of a subject's treatment. In one bone density study, for example, an investigator blind to treatment had to decide whether to exclude subjects because of x-rays suggesting that calcification of the aorta might make bone density measurements unreliable. Had the investigator been aware of a subject's treatment, there would always be a question of whether a decision to exclude certain subjects was based, however unconsciously, on what it might do to the final treatment comparison. (Then, there's *triple blinding*, in which the analyst is given the data with uninformative treatment labels such as A and B and their identities are not revealed until the analyses are completed!)

It's easy to come up with reasonable-sounding arguments for not enforcing blinding ("*I* won't be influenced by the knowledge. Not me!" "My contact is so minimal it can't matter."), but **EVERY ONE IS SPECIOUS.** The following example illustrates how fragile things are:

> Patients (unilateral extraction of an upper and lower third molar) were told that they might receive a placebo (saline), a narcotic analgesic (fentanyl), or a narcotic antagonist (naloxone) and that these medications might increase the pain, decrease it, or have no effect. The clinicians knew that one group (PN) of patients would receive only placebo or naloxone and not fentanyl and that the second group (PNF) would receive fentanyl, placebo, or naloxone. All drugs were administered double blind. Pain after placebo administration in group PNF was significantly less than after placebo in group PN! The two placebo groups differed only in the clinicians' knowledge of the range of possible double blind treatments. (Gracely et al., Lancet, 1/5/85, p 43)

When a new drug/technique is introduced, it is almost always the case that treatment effects diminish as the studies go from unblinded to blind to double-blind.

Sometimes blinding is impossible. Women treated for breast cancer knew whether they had a lumpectomy, simple mastectomy, or radical mastectomy; subjects know whether they are performing stretching exercises or strength training exercises; there is *no* placebo control that is indistinguishable from cranberry juice. (Recently, we were looking for a control for black tea. It had to contain everything in black tea except for a particular class of chemicals. Our dieticians came up with something that tastes like tea, but it is water soluble and doesn't leave any leaves behind.) In each case, we must do the best we can to make treatments as close as possible remembering that the differences we observe reflect any and all differences in the two treatments, not just the ones we think are important.

No matter how hard it might seem to achieve blinding in practice, barriers usually turn out to be nothing more than matters of inconvenience. There are invariably ways to work around them. Often, it is as simple as having a colleague or research assistant randomize treatments, prepare and analyze samples, and make measurements.

Evaluating A Single Treatment

Often, intervention trials are used to evaluate a single treatment. One might think that the way to conduct such studies is to apply the treatment to a group of subjects and see whether there is any change, but that approach makes it difficult to draw conclusions about a treatment's effectiveness.

Things change even when no specific intervention occurs. For example, cholesterol levels probably peak in the winter around Thanksgiving, Christmas, and New Year's when people are eating heavier meals and are lower in the summer when fresh fruits and vegetables are in plentiful supply. In the Northeast US, women see a decline in bone density during the winter and increase during the summer because of swings in vitamin D production from sunlight exposure. (Imagine an effective treatment studied over the winter months and called ineffective because there was no change in bone density! Or, an *in*effective treatment studied over the summer and called effective because there *was* a change!)

When a treatment is described as effective, the question to keep in mind is, "Compared to what?" In order to convince a skeptical world that a certain treatment has produced a particular effect, it must be compared to a regimen that differs only in the facet of the treatment suspected of producing the effect.

Placebo controlled means that the study involves two treatments--the treatment under investigation and an ineffective control (placebo) to which the new treatment can be compared. At first, such a control group seems like a waste of resources--as one investigator describes them, a group of subjects that is "doing nothing". However, a treatment is not just the taking some substance or following a particular exercise program, but all of the ways in which it differs from "doing nothing". That includes contact with health care professionals, heightened awareness of the problem, and any changes they might produce. In order to be sure that measured effects are the result of a particular facet of a regimen, there must be a control group whose experience differs from the treatment group's by that facet only. If the two groups have different outcomes, then there is strong evidence that it is due to the single facet by which the two groups differ.

One study that was saved by having a reliable group of controls is the Multiple Risk Factors Intervention Trial (MRFIT), in which participants were assigned at random to be counseled about minimizing coronary risk factors, or not. Those who were counseled had their risk of heart attack drop. However, the same thing happened in the control group! The reason was that the trial took place at a time when the entire country was becoming sensitive to and educated about the benefits of exercise, a low fat diet, and not smoking. The intervention added nothing to what was already going on in the population. Had there be no control group--that is, had historical controls been used--it is likely that the intervention would have been declared to be of great benefit. Millions of dollars and other resources would have been diverted to an ineffective program and wasted.

What about comparing a treatment group to a convenience sample--a sample chosen not through a formal sampling procedure, but because it is convenient? Perhaps it would allow us to avoid having to enroll and randomize subjects who are "doing nothing" throughout the protocol. The Salk vaccine trials have something to say about that.

Group	**cases per 100,000**
placebo	71
innoculated	28
refused	46

When the trial was proposed, it was suggested that everyone whose parents agreed to participate should be inoculated while all others should be used as the control group. Fortunately, cooler heads prevailed and placebo controls were included. It turned out that those parents who refused were more likely to have lower incomes and only one child. The choice to participate was clearly related to susceptibility to polio. (income=hygiene, contagious disease).

- In the trial as conducted with placebo controls, the proper comparison was *placebo* to *innoculated*, that is, 71/100,000 to 28/100,000.
- Had all participants been inoculated, the placebo group would have behaved like the innocuated group. The rate for those innoculated would continue to be 28/100,000.
- Had the refusals used for comparison purposes, the comparison would have been 46/100,000 to 28/100,000, which is much less dramatic.

Imagine what the results would have been if there were *no* control group and the previous year's rates were used for comparison...and this turned out to be a particularly virulent year.

Even "internal controls" can fool you.

> Studies with clofibrate showed that subjects who took 80% or more of their drug had substantially lower mortality than subjects who took less; this would seem to indicate that

> the drug was beneficial. But the same difference in mortality was observed between subjects with high and subjects with low compliance whose medication was the placebo. Drug compliance, a matter depending on personal choice, was for some reason related to mortality in the patients in this study. Were it not for the control group, the confounding between the quantity of drug actually taken (a personal choice) and other factors related to survival might have gone unnoticed, and the theory "more drug, lower mortality: therefore, the drug is beneficial" might have stood--falsely. (Moses, 1985, p.893)

It is important to check the treatment and placebo to make certain that they are what they claim to be. Errors in manufacturing are not unheard of. I've been involved in studies where the placebo contained a small amount of the active treatment. I've also been involved in a study where the packaging facility reversed the treatment labels so that, prior to having their labels removed before they were given to subject, the active treatment was identified as placebo and placebo as active treatment! Be knowledgeable about the consequences of departures from a study protocol. An unreported communion wafer might not seem like a problem in a diet study--unless it's a study about gluten-free diets.

Placebo controls are unnecessary when comparing two active treatments. I have seen investigators include placebo controls in such studies. When pressed, they sometimes claim that they do so in order to monitor temporal changes, but I try to dissuade them if that is the only purpose.

Subjects As Their Own Controls

Despite the almost universally recognized importance of a control group, it is not uncommon to see attempts to drop it from study in the name of cost or convenience. A telltale phrase that should put you on alert is that "Subjects were used as their own controls."

Subjects can and *should* be used as their own controls if all treatments can be administered simultaneously (e.g., creams A and B randomly assigned to the right and left arms). But in common parlance, "using subjects as their own controls" refers to the practice of measuring a subject, administering a treatment, measuring the subject again, and calling the difference in measurements the treatment effect. This can have disastrous results. Any changes due to the treatment are confounded with changes that would have occurred over time had no intervention taken place. We may observe what looks like a striking treatment effect, but how do we know that a control group would not have responded the same way?

MRFIT is a classic example where the controls showed the same "effect" as the treatment group. Individuals who were counseled in ways to minimize the risk of heart attack did no better than a control group who received no such counseling simply because the population as a whole was becoming more aware of the benefits of exercise. When subjects are enrolled in a study, they often increase their awareness of the topic being studied and choose behaviors that they might not have considered otherwise. Dietary intake and biochemical measures that are affected by diet often change with season. It would not be surprising to see a group of subjects have their cholesterol levels decrease after listening to classical music for six months--if the study started in January after a holiday season of relatively fatty meals and ended in July after the greater availability of fresh produce added more fruits and salads to the diet.

The best one can do with data from such studies is argue the observed change was too great for coincidence. ("While there was no formal control group, it is biologically implausible that a group of subject such as these would see their total cholesterol levels drop an average of 50 mg/dl in only 4 weeks. It would be too unlikely a coincidence for a drop this large to be the result of to some uncontrolled factor.") The judgment that other potential explanations are inconsequential or insubstantial is one that must be made by researchers and their audience. It can't be made by statistical theory. The justification must come from outside the data. It could prove embarrassing, for example, if it turned out that something happened to the measuring instrument to cause the drop. (A control group would have revealed this immediately.)

In the search of ways to minimize the cost of research and to mitigate the effects of temporal effects, some studies adopt a three-phase protocol: measurement before treatment, measurement after treatment, measurement after treatment ceases and a suitable washout period has expired, after which time subjects should have returned to baseline. In theory, if a jump and a return to baseline were observed, it would require the most remarkable of coincidences for the jump to be due to some outside factor. There are many reasons to question the validity of this approach.

- If life were that simple--baseline, jump, baseline all flat--then coincidence would be unlikely but the pattern might, indeed, be temporally induced. Imagine a regime where the treatment portion of the protocol induces tension and anxiety in the volunteers or, due to contact with health care professionals, a feeling of well-being. If the non-treatment portions of the protocol remove these stimuli, any response may well be a product of the stimuli rather than the "treatment".
- Life is hardly ever that simple. One rarely sees a flat response followed by a discrete jump. Often measurements in the presence of an effective treatment drift around before treatment, show an accelerated change during treatment, and continue to drift around after treatment. This drift often shows some non-random trend. Assessing a treatment effect, then, is often not looking for a discrete jump in response but looking for a trend in the treatment group that is different from the trend in the control group.
- What should be done when subjects return to a plateau different from the baseline they left?

Despite this indictment against this use of subjects as their own controls, cost and convenience continue to make it tempting. In order to begin appreciating the danger of this practice, you should **look at the behavior of the control group whenever you read about placebo-controlled trials**. You will be amazed by the kinds of effects they exhibit.

The Ethics of Randomized Trials

When a trial involves a health outcome, investigators should be truly indifferent to the treatments under investigation. That is, if investigators were free to prescribe treatments to subjects, they would be willing to choose by flipping a coin. In the case of a placebo controlled trial, investigators must be sufficiently unsure of whether the "active" treatment is truly effective.

Ethical considerations forbid the use of a placebo control if it would withhold (a standard) treatment known to be effective. In such cases, the control must the treatment dictated by standard medical practice and the research question should be rephrased to ask how the new treatment compares to standard practice. For example, in evaluating treatments for high cholesterol levels, it would be unethical to do nothing for subjects known to have high levels. Instead, they would receive the treatment dictated by standard medical practice, such as dietary consultation and a recommendation to follow the AHA Step 1 diet or even treatment with cholesterol lowering drugs.

It is impossible, in a few short paragraphs, to summarize or even list the ethical issues surrounding controlled clinical trials. Two excellent reference is *Levin RJ (1986), Ethics and Regulation of Clinical Research, 2nd ed. New Haven: Yale University Press* and *Dunn CM & Chadwick G (1999), Protecting Study Volunteers in Research. Boston: CenterWatch, Inc*.

An online short course on the protection of human subjects can be found at http://cme.nci.nih.gov. It's interesting, well-designed, and includes many informative links worth bookmarking. It even offers a certificate for completion.

Sample Size Calculations

A properly designed study will include a justification for the number of experimental units (subjects/animals) being examined. No one would propose using only one or two subjects per drug to compare two drugs, because it's unlikely that enough information could be obtained from such a small sample. On the other hand, applying each treatment to millions of subjects is impractical, unnecessary, and unethical. Sample size calculations are necessary to design experiments that are large enough to produce useful information and small enough to be practical. When health outcomes are being studied, experiments larger than necessary are unethical because some subjects will be given an inferior treatment unnecessarily.

One vs Many

Many measurements on one subject are not the same thing as one measurement on many subjects. With many measurements on one subject, you get to know the one subject quite well but you learn nothing about how the response varies across subjects. With one measurement on many subjects, you learn less about each individual, but you get a good sense of how the response varies across subjects. A common mistake is to treat many measurements on one subject as though they were single measurements from different subjects. Valid estimates of treatment effects can sometimes be obtained this way, but the uncertainty in these estimates is greatly underestimated. This leads investigators to think they have found an effect when the evidence is, in fact, insufficient.

The same ideas apply to *community intervention studies*, also called *group-randomized trials*. Here, entire villages, for example, are assigned to the same treatment. When the data are analyzed rigorously, the sample size is the number of villages, not the number of individuals. This is discussed further under units of analysis.

Paired vs Unpaired Data

Data are paired when two or more measurements are made on the same observational unit. The observational unit is usually a single subject who is measured under two treatment conditions. However, data from units such as couples (husband and wife), twins, and mother-daughter pairs are considered to be paired, too. They differ from unpaired (or, more properly, independent) samples, where only one type of measurement is made on each unit. They require special handling because the accuracy of estimates based on paired data generally differs from the accuracy of estimates based on the same number of unpaired measurements.

Parallel Groups vs Cross-Over Studies

In a parallel groups study, subjects are divided into as many groups as there are treatments. Each subject receives one treatment. In a cross-over study, all subjects are given all treatments. In the case of two treatmetns, half are given A followed by B; the other half are given B followed by A. Cross-over studies are about as close you can come to the savings investigators would like to realize by using subjects as their own controls, but they contain two major drawbacks. The first is carryover--B after A may behave differently from B alone. The second is the problem of missing data; subjects who complete only one of the two treatments complicate the analysis. It's for good reason, then, that the US Food & Drug Administration looks askance at almost anything other than a parallel groups analysis.

Repeated Measures Designs

In repeated measures designs, many measurements are made on the same individual. Repeated measures can be thought of as a generalization of paired data to allow for more than two measurements. The analysis of paired data will be identical to an analysis of repeated measures with two measurements. Some statisticians maintain a distinction between **serial measurements** and **repeated measures** *Serial measurement* are used to describe repeatedly measuring the same thing in the same way over time. *Repeated measures* is reserved for different types of measurements, usually different ways of measuring the same thing. For example, the term *serial measurements* would be used when a subject's blood pressures is measured in the same way over time. *Repeated measures* would be used to describe a study in which subjects' blood pressure was measured many different ways (sitting, lying, manually, automated cuff) at once. I, myself, do not object to the use of the term *repeated measures* in conjunction with serial measurements.

Often the analysis of serial measurements can be greatly simplified by reducing each set of measurements to a single number (such as a regression coefficient, peak value, time to peak, or area under the curve) and then using standard techniques for single measurements.

ITT & Meta Analysis

Among the topics that should be included in these notes are the highly controversial Intention-To-Treat analyses and Meta analysis. Like many statisticians, I have strong feelings about them. Because these are highly charged issues, I have placed them in their own Web pages to give them some distance from the more generally accepted principles presented in this note.

Intention-To-Treat Analysis
Meta Analysis

The Bottom Line

We all want to do research that produces valid results, is worthy of publication, and meets with the approval of our peers. This begins with a carefully crafted research question and an appropriate study design. Sometimes all of the criteria for a perfect study are not met, but this does not necessarily mean that the work is without merit. What *is* critical is that the design be described in sufficient detail that it can be properly evaluated. (The connection between Reye's syndrome and aspirin was established in a case-control pilot study which was meant to try out the machinery before embarking on the real study. An observed odds ratio of 25 led the researchers to publish the results of the pilot.) Any study that is deficient in its design will rarely be able to *settle* the question that prompted the research, but it may be able to provide valuable information nonetheless.

[back to The Little Handbook of Statistical Practice]

Last modified: undefined.

Intention-To-Treat Analysis

Gerard E. Dallal, Ph.D.

Chu-chih [Gutei] Raises One Finger : The Gateless Barrier Case 3

> Whenever Chu-chih was asked a question, he simply raised one finger. One day a visitor asked Chu-chih's attendant what his master preached. The boy raised a finger. Hearing of this, Chu-chih cut off the boy's finger with a knife. As he ran from the room, screaming with pain, Chu-chih called to him. When he turned his head, Chu-chih raised a finger. The boy was suddenly enlightened.
>
> When Chu-chih was about to die, he said to his assembled monks: "I received this one-finger Zen from T'ien-lung. I used it all my life but never used it up." With this he entered his eternal rest.

It is now commonplace, if not standard practice, to see Requests For Proposals specify that study data be subjected to an intention-to-treat analysis (ITT) with "followup and case ascertainment continued regardless of whether participants continued in the trial". *Regardless* means regardless of adherence, change in regimens, reason for outcome [accidental death is death]... A popular phrase used to describe ITT analyses is **"Analyze as randomized!"** Once subjects are randomized, their data **must be** used for the ITT analysis! This sounds...well, the polite word is *counter-intuitive*. *Bizarre* may be closer to the mark.

When Richard Peto first introduced the idea of ITT, the cause was taken up by many prominent statisticians, including Paul Meier, then of the University of Chicago and, later, Columbia University, whom I have heard speak eloquently in its defense. Others thought that Peto's suggestion was a sophisticated joke and awaited the followup article, which never came, to reveal the prank. Initially, I sympathized strongly with this latter camp, but I have become more accepting over the years as ITT has begun to be used intelligently rather than as a totem to make problems vanish magically.

There are four major lines of justification for intention-to-treat analysis.

1. Intention-to-treat simplifies the task of dealing with suspicious outcomes, that is, it guards against conscious or unconscious attempts to influence the results of the study by excluding odd outcomes.
2. Intention-to-treat guards against bias introduced when dropping out is related to the outcome.
3. Intention-to-treat preserves the baseline comparability between treatment groups achieved by randomization.

4. Intention-to-treat reflects the way treatments will perform in the population by ignoring adherence when the data are analyzed.

Dealing with questionable outcomes and guarding against conscious or unconscious introductions of bias

One of Meier's examples involves a subject in a heart study where there is a question of whether his death should be counted against his treatment or set aside. He subject died from falling off his boat after having been observed carrying a few six-packs of beer on board for his solo sail. Meier argues that most researchers would set this event aside as probably unrelated to the treatment, while intention-to-treat would require the death be counted against the treatment. But suppose, Meier continues, that the beer is eventually recovered and every can is unopened. Intention-to-treat does the right thing in any case. By treating all treatments the same way, deaths unrelated to treatment should be equally likely to occur in all groups and the worst that can happen is that the treatment effects will be watered down by the occasional, randomly occurring outcome unrelated to treatment. If we pick and choose which events should count, we risk introducing bias into our estimates of treatment effects.

Guarding against informative dropouts

Imagine two weight loss diets, one of which is effective while the other isn't. People on the effective diet will lose weight and stay in the study. Some of those on the ineffective diet will lose weight anyway and will stay in the study. Those who do poorly are more likely to drop out, if only to try something else. This will make the ineffective diet look better than it really is.

Preserving baseline comparability between treatment groups achieved by randomization.

There have been studies where outcome was unrelated to treatment but *was* related to adherence. In many cases, potentially nonadherent subjects may be more likely to quit a particular treatment. For example, a nonadherent subject might be more likely to quit when assigned to strenuous exercise than to stretching exercises. In an *on treatment* analysis, the balance in adherence achieved at baseline will be lost and the resulting bias might make one of two equivalent treatments appear to be better than it truly is simply because one group of subject, on the whole, are more adherent.

As a more extreme case of Paul Meier's example, consider a study in which severely ill subjects are randomly assigned to surgery or drug therapy. There will be early deaths in both groups. It would be tempting to exclude the early deaths of those in the surgery group who died before getting the surgery on the grounds that *they never got the surgery*. However, this has the effect of making the drug therapy group much less healthy on average at baseline.

Reflecting performance in the population

Intention-to-treat analysis is said to be more realistic because it reflects what might be observed in actual clinical practice. In practice, patients may not adhere, they may change treatments, they may accidentally die. ITT factors this into its analysis. It answers the public health question of what happens when a recommendation is made to the general public and the public decides how to implement it. The results of an intention-to-treat analysis can be quite different from the treatment effect observed when adherence is perfect.

My own views

What troubles me most about intention-to-treat analyses is that the phrase *intention-to-treat* is sometimes used as an incantation to avoid thinking about research issues. Its use often seems to be divorced from any research question. It is easy to imagine circumstances where researchers might argue that the actual research question demands an intention-to-treat analysis to evaluate the results--for example, "For these reasons, we should be following everyone who enters the study regardless of adherence". What worried me most in the past was hearing ITT recommended for its own sake without any reference to the specific questions it might answer. This seems to occur less frequently as awareness of ITT's strengths and weaknesses have become better known. However, whenever I hear that an ITT analysis has been performed, I invariably examine the study and its goals to assure myself that the ITT analysis was appropriate.

Intention-to-treat analysis answers a certain kind of research question. *On treatment* analysis answers a different kind of research question. My own view is to ignore labels, understand the research question, and perform the proper analysis *whatever it's called*. In some cases it may even be ITT! Usually, I perform both an intention-to-treat analysis and an on treatment analysis, using the results from the different analyses to answer different research questions.

If the purpose of a study is to answer "the public health question", then an ITT analysis should be performed. An ITT analysis should not be performed simply to perform an ITT analysis. An ITT analysis should be performed because the researchers are interested in answering the public health question and they have determined that an ITT analysis will answer it.

There are two components to how a treatment will behave in the population at large: efficacy and adherence. However, these are separate issues that should not be routinely combined in a single intention-to-treat analysis. A treatment's efficacy is often of great scientific importance (all exaggeration aside) regardless of adherence issues. Adherence during a trial might be quite different from adherence once a treatment has been proven efficacious. One can imagine, for example, what adherence might be like during a vitamin E trial and then what they would be like if Vitamin E were shown to prevent most forms of cancer! Should a treatment be found to be highly effective but unpalatable, future research might focus on ways to make it more palatable while other research, exploiting the active components of the treatment, might come up with new, more effective treatments. There may be cases, such as the treatment of mental disease, where an intention-to-treat analysis will truly reflect the way the treatments will behave in practice. In the fields in which I work, these situations tend to be exceptions rather than the rule.

Meier's example does not strike me as a compelling reason for ITT. The subject is on treatment. What is unclear is the way the outcome should be classified. This can be an issue even for ITT analyses. In the example, we don't know whether the subject suffered a heart attack. The beer might change the likelihood of various possibilities but the cause of death is still a guess whether the bottles were opened or unopened. In cases like this it makes sense to perform the analysis in many ways--for those outcomes where the cause of death is certain and then for all outcomes.

Informative stopping *is* a serious problem that can distort the results of an on treatment analysis. However, it can also distort an ITT analysis because subjects who quit may have final values--regardless of the method for obtaining them--very different from what they would have had, had they stayed on study. If stopping is noninformative, ITT will tend to attenuate a treatment effect because it will be adding subjects whose outcomes are similar (by virtue of having dropped out before the treatment could take effect) to both groups.

ITT does preserve the comparability at baseline achieved by randomization, but it is not the only way to do so. There might be a run- in period before subjects are randomized in order to identify nonadherent subjects and exclude them before they are assigned to treatment. A different approach is to use adherence as a covariate so that it is not confounded with treatment. In cases such as the surgery/drug therapy example, all deaths within a certain number of days of assignment might be excluded regardless of treatment.

David Salsburg once asked what to do about an intention-to-treat analysis if at the end of a trial it was learned that everyone assigned treatment A was given treatment B and vice-versa. I am living his joke. In a placebo-controlled vitamin E study, the packager delivered the pills just as the trial was scheduled to start. Treatments were given to the first few dozen subjects. As part of the protocol, random samples of the packaged pills were analyzed to insure the vitamin E did not lose potency during packaging. We discovered the pills were mislabeled--E as placebo and placebo as E. Since this was discovered a few weeks into the trial, no one had received refills, which might have been different from what was originally dispensed. We relabeled existing stores properly and I switched the assignment codes for those who had already been given pills to reflect what they actually received. How shall I handle the intention-to-treat analysis?

This slip-up aside, this is an interesting study because it argues both for an against an ITT analysis. Because the study pill is administered along by a nurse along with a subject's medications, it's hard to imagine how adherence might change, even if the results of the trial were overwhelmingly positive. This makes an ITT analysis attractive. However, it is likely that there will be many drop outs unrelated to treatment in any study of a frail population. Should they be allowed to water down any treatment effect? The key issue in answering this question is whether the dropouts are noninformative.

Also, some subjects will leave the study because they cannot tolerate taking the pill, irrespective of whether it is active or inactive, or because their physicians decide, after enrollment, that they should not be in a study in which they might receive a vitamin E supplement. If a recommendation for supplements were made, such subjects would not be able to follow it, so perhaps it is inappropriate to include their data in the analyses of vitamin E's efficacy.

In summary, I will recant a bit of my opening paragraph. ITT is not bizarre. In some circumstances, it may be the right thing to do. A slavish devotion to ITT is worse than bizarre. It could be harmful. The proper approach is to ignore labels, understand the research question, and perform the proper analysis *whatever it's called*!

[back to The Little Handbook of Statistical Practice]

Meta Analysis

Gerard E. Dallal, Ph.D.

It is not every question that deserves an answer.
-- Publilius Syrus

Sometimes there are mixed reports about a treatment's effectiveness. Some studies may show an effect while others do not. Meta analysis is a set of statistical techniques for combining information from different studies to derive an overall estimate of a treatment's effect. The underlying idea is attractive. Just as the response to a treatment will vary among individuals, it will also vary among studies. Some studies will show a greater effect, some will show a lesser effect--perhaps not even statistically significant. There ought to be a way to combine data from different studies, just as we can combine data from different individuals within a single study. That's *Meta Analysis*.

Meta analysis always struggles with two issues:

1. publication bias (also known as *the file drawer problem*) and
2. the varying quality of the studies.

Publication bias is "the systematic error introduced in a statistical inference by conditioning on publication status." For example, studies showing an effect may be more likely to be published than studies showing no effect. (Studies showing no effect are often considered unpublishable and are just filed away, hence the name *file drawer problem*.) Publication bias can lead to misleading results when a statistical analysis is performed after assembling all of the published literature on some subject.

When assembling the available literature, it can be difficult to determine the amount of care that went into each study. Thus, poorly designed studies end up being given the same weight as well designed studies. This, too, can lead to misleading results when the data are summarized.

When large, high quality randomized, double-blind, controlled trials are available, they are the gold standard basis for action. Publication bias and the varying quality of other studies are not issues because there is no need to assemble the research in the area. So, to the two primary concerns about meta-analysis--publication bias and varying quality of the studies--I have added a third:

> (3) Meta analysis is used only when problems (1) and (2) are all but certain to cause the most trouble! That is, meta-analysis is employed only when no large-scale, high quality trials are available and the problems of publication bias and the varying quality and outcomes of available studies all but guarantee it will be impossible to draw a clear conclusion!

Those who perform meta analyses are aware of these problems and have proposed a number of guidelines to minimize their impact.

- A formal protocol should be written specifying the exact question under investigation and describing the studies that will be included in the analysis.
- All research, not just published research, should be included.
- Registries should be established so that studies can be tracked from their inception and not just on publication. This idea has been given a push so that drug companies would not be able to publish trials showing benefit from their products while suppressing those that do not.
- Many meta analytic techniques should be used and all results should be reported. A result would be considered reliable only if all of the techniques give the same result.

I continue to be skeptical and remain unconvinced that these procedures are sufficient to overcome the problems they seek to address.

It is difficult to improve upon the remarks of John C. Bailar, III, taken from his letter to *The New England Journal of Medicine*, 338 (1998), 62, in response to letters regarding LeLorier et al. (1997), "Discrepancies Between Meta-Analyses and Subsequent Large Randomized, Controlled Trials", *NEJM*, 337, 536-542 and his (Bailar's) accompanying editorial, 559-561:

> My objections to meta-analysis are purely pragmatic. It does not work nearly as well as we might want it to work. The problems are so deep and so numerous that the results are simply not reliable. The work of LeLorier et al. adds to the evidence that meta-analysis simply does not work very well in practice.
>
> As it is practiced and as it is reported in our leading journals, meta-analysis is often deeply flawed. Many people cite high-sounding guidelines, and I am sure that all truly want to do a superior analysis, but meta-analysis often fails in ways

that seem to be invisible to the analyst.

The advocates of meta-analysis and evidence-based medicine should undertake research that might demonstrate that meta-analyses in the real world--not just in theory--improve health outcomes in patients. Review of the long history of randomized, controlled trials, individually weak for this specific purpose, has led to overwhelming evidence of efficacy. I am not willing to abandon that history to join those now promoting meta analysis as the answer, no matter how pretty the underlying theory, until its defects are honestly exposed and corrected. The knowledgeable, thoughtful, traditional review of the original literature remains the closest thing we have to a gold standard for summarizing disparate evidence in medicine.

[back to The Little Handbook of Statistical Practice]

Last modified: undefined.

Random Samples / Randomization

Random Samples and **Randomization** are two different things! However, they have something in common, which is what somtimes leads to confusion. As the presence of **random** in both names suggests, both involve the use of a probability device.

- With **random samples**, chance determines who will be in the sample.
- With **randomization**, chance determines the assignment of treatments.

A **random sample** is drawn from a population by using a probability device. We might put everyone's name on a slip of paper, mix thoroughly, and select the number of names we need, or we might have a computer generate random numbers and use them to select our sample. If you don't trust the computer to generate your random numbers, there are always http://random.org, http://www.fourmilab.ch/hotbits/, and even http://www.lavarnd.org/. **The use of a probability device to select the subjects allows us to make valid generalizations from the sample to the population.**

In an intervention trial, **randomization** refers to the use of a probability device to assign subjects to treatment. This allows us to use statistical methods to make valid statements about the difference between treatments for this set of subjects. The subjects who are randomized may or may not be a random sample from some larger population. Typically, when human subjects are involved, they are volunteers. If they *are* a random sample, then statistical theory lets us generalize from this trial to the population from which the sample was drawn. If they are not a random sample from some larger population, then generalizing beyond the trial is a matter of nonstatistical judgement.

Randomization models: Why should statistical methods work for intervention trials involving volunteers?

Intervention trials are typically analyzed by using the same statistical methods for the analyzing random samples. Almost all intervention trials involve volunteers, usually recruited locally. If convenience samples are inappropriate for surveys, how can they be appropriate for intervention trials? There are two distinct issues to address--**validity** and **generalizability**.

- **Validity** is concerned with whether the experiment valid, that is, whether observed differences in this group indicate a real difference in treatments insofar as these subjects are concerned.
- **Generalizability** is concerned with whether the results can be generalized to any other group of individuals.

The reason volunteers can be used to make valid comparisons comes from the use of randomization in the assignment of treatments. It is beyond the scope of these notes to give mathematical proofs, but the common statistical methods that are appropriate to compare simple random samples are also valid for deciding whether the observed difference between the two treatments is greater than would be expected when subjects are assigned to treatments at random and the treatments are equivalent. The probability models for *random sampling* and the probability models for *randomization* lead to the same statistical methods.

Within broad limits, theresults of intervention trials can be genralized because all human beings are made out of the same stuff. While this justification cannot be applied blindly, it may be comforting to know that many of the surgical advances of the mid 20-th century were developed in VA hospitals on middle-age white males. However, the ability to generalize results does not immediately follow from the use of particular numerical methods. Rather, it comes from the subject matter specialist's knowledge of those who were studied and the group to whom the generalization will be made.

It is worth noting here that the quality of evidence about factors under the control of the investigator is different from that for factors that cannot be subjected to randomization. For example, consider an intervention trial that compares the effects of two diets in smoking and nonsmoking pregnant women. The use of statistical methods to compare diets can be justified by the random assignment of subjects to treatment. However, the comparison between smokers and nonsmokers depends on an enrollment procedure that would not recruit smokers who differ from nonsmokers in ways that are associated with response to the diets.

Last modified: undefined.

Welcome to Randomization.com!!!

(where it's never the same thing twice)

There are now three randomization plan generators.

The first (and original) generator randomizes each subject to a single treatment by using the method of *randomly permuted blocks*.

The second generator creates random permutations of treatments for situations where subjects are to receive all of the treatments in random order.

The third generator generates a random permutation of integers. This is particularly useful for selecting a sample without replacement.

New features will be added as the occasion demands. This page will undergo updates and revisions, but links to the randomization plan generators will always be available here, The generators may undergo some cosmetic changes, but the algorithms will not be changed. This will insure that an old plan can always be reconstructed. In the event it becomes necessary to change a generator, a notice will be posted on the website and the program will be updated, but every version of the generators will continue to be available. When a randomization plan is created, the date is now printed on the plan to document the generator that was used.

***Newly added*! Citing Randomization.com**

Please send all comments to HelpDesk@randomization.com

Fc 59178

FastCounter by bCentral

Last modified: undefined.

Units of Analysis

As David Murray points out in his book *Group-Randomized Trials* (Oxford University Press, 1998), there's plenty of opportunity for confusion about **units**. There are *units*, *observational units*, *assignment units*, and *units of analysis*, among other terms. Often, these terms are used interchangeably (I do so myself), but not always.

In any study involving people, the individual is commonly thought of as the unit of analysis because we study people. However, the unit of analysis and the corresponding sample size are determined by the way the study is conducted.

Determining units of analysis and their number recalls the discussion of why measuring a single mouse 100 times is different from measuring 100 mice once each. Measurements on the same mouse are likely to be more similar than measurements made on different mice. If there is something about the way an experiment is conducted that makes it likely that some observations will be more similar than others, this must be reflected in the analysis. This is true whether the study involves diets, drugs, nutritional supplements, methods of planting, social policies, or ways of delivering service. (However, if two measurements on the same mouse were **not** likely to be more similar than two measurements made on different mice, then measuring a single mouse 100 times is **no** different from measuring 100 mice once each!)

Consider a study of 800 10th grade high school students receiving one of two treatments, A & B. Experience has shown that two students selected at random from the same class are likely to be more similar than two students selected at random from the entire school who, in turn, are likely to be more similar than two students selected at random from the entire city, who are likely to be more similar than two students selected at random from the entire state, and so on.

Here are three of the many ways to carry out the study in a particular state.

1. Take a random sample of 800 10th grade students from all school students in the state. Randomize 400 to A and 400 to B.
2. Take a random sample of 40 10th grade classes of 20 students each from the set of all 10th grade classes. Randomize 20 classes to A and 20 classes to B.
3. Take a random sample of 20 schools. From each school, randomly select two classes of 20 students each. Randomize the schools into two groups with classes from the same school receiving the same treatment.

Each study involves 800 students--400 receive A and 400 receive B. However, the units of analysis are different. In the first study, the unit of analysis is the individual student. The sample size is 800. In the second study, the unit of analysis is the class. The sample size is 40. In the third study, the unit of analysis is the school. The sample size is 20.

Murray (p. 105) has an elegant way to identify the units of analysis. It is not a definition a novice can use, but it is rigorous and is the way a trained statistician decides what the units should be. *A unit is the **unit of analysis** for an effect if and only if that effect is assessed against the variation among those units.*

It's not easy to come up with a less technical definition, but most of the time the units of analysis are *the smallest units that are independent of each other* or *the smallest units for which all possible sets are equally likely to be in the sample*. In the examples presented above

1. A random sample of students is studied. The students are independent of each other, so the student is the unit of analysis.
2. Here, students are not independent. Students in the same class are likely to be more similar than students from different classes. Classes *are* independent of each other since we have a simple random sample of them, so class is the unit of analysis.
3. Here, neither students nor classes are independent. Classes from the same school are likely to be more similar than classes from different schools. Schools are selected at random, so school is the unit of analysis.

You might think of causing trouble by asking what the unit of analysis would be in case 3 if, in each school, one class received A and the other received by, with the treatments assigned at random. The unit of analysis would still be the school, but the analysis is now effectively one of paired data because both treatments are observed in each school. In similar fashion

- In a twins study, where the members of each twin pair are purposefully randomized purposefully so that the two twins receive different treatments, the unit of analysis is the twin pair.
- In a study of two types of exercise, where each subject uses a different form of exercise for each arm with treatments assigned at random, the unit of analysis is the pair of arms, that is, the individual subject, not the individual arm.
- In an agricultural study, where each farm has plots devoted to all of the methods under investigation, the unit of analysis is the farm, not the plot.
- In a study of husbands and wives, the unit of analysis is the couple.

Pairing cuts the sample size to half of what it would have been otherwise. However, you have to measure the units twice as long/much and the analysis becomes complicated if one of the two measurements ends up missing.

- Group-randomization (without Pairing) estimates effects with less precision than had individuals been randomized because similar individuals receive the same treatment and tend to behave similarly.
- Pairing usually leads to greater precision because comparisons within pairs are generally more precise that comparisons between unpaired units. In theory, pairing buys back some of the precision lost through group randomization. It would be interesting to do some calculations to

find out how much precision is recovered through pairing.

Returning to study 3: If the two classes within each school were randomized to different treatments, the unit of analysis would be the school, not the class. However, the treatment effect would be compared to the variability in differences between classes within in each school. Therefore, this version of the study would probably be better able to detect a real difference than study 2, which was based on a (simple) random sample of classes.

At first glance, it seems unfair that the third study, involving hundreds of students, has a sample size equal to the number of schools. However, if there were no differences between schools or classes, the two analyses--individuals (incorrect) and schools (correct)--would give the essentially the same result. (This is the same thing as saying 100 measurements on one mouse **are** the same as 1 measurement on each of 100 mice if all mice react the same way.) The sample size may be small but the school means will show much less variability than class means or individual students, so the small sample size is made up for by the increase in precision.

Groups rarely respond exactly the same way, so treating the group as the unit of analysis saves us from making claims that are not justified by the data. The precision of a properly analyzed group-randomized study involving a grand total of N subjects spread out among K groups will be equal to a that of a simple random sample of somewhere between K and N individuals.

- If the groups respond differently but there is no variation within each group, we've essentially got K measurements, which we see over and over again.
- At the other extreme, if there's no difference between groups beyond the fact that they are composed of different individuals, we've essentially got N observations.

The analysis will take care of this automatically if the proper methods are used.

Last modified: undefined.

Creating Data Files

In 1994, this is how I began a review of of the program Epi Info version 5.01 for *Chance* magazine:

> One semester, the data sets I wanted to use in class were already in machine-readable form. I placed them on the [then mainframe!] computer for my students as a matter of convenience to spare them the chore of data entry. Around the middle of the term, this prompted a student to ask, "Where do data files come from?"
>
> Where *do* data files come from? With any luck, they will come from someone else, but eventually we all end up helping others deal with putting data files together.
>
> Methods for recording data often appear to be selected as afterthoughts, and they often use tools designed for other purposes.

In 2000, this note began

> Things haven't changed much over the last six year. Data continue to arrive from a variety of sources, entered in ways suggesting they weren't necessarily recorded with an eye toward future analysis.

Today, it's fair to say that things *are* different. The world needed a standard, so it chose the Excel spreadsheet. With Microsoft's dominance of the spreadsheet market, data are often collected directly into Excel spreadsheets. However, this is only part of the reason why Excel has become the standard. Users demand the ability to transfer data between programs. Software companies cannot ignore this and remain in business. Every program has its "Save As" options. Different programs may provide different options, but Excel is always on the list because it can be found on so many computers. Those who design program that import data, such as statistical program packages, recognize this reality and see to it that their programs can read Excel spreadsheets.

Excel has become analogous to an airline's hub. Just as getting from point A to point B involves passing through the hub, getting data from program A to program B often involves passing through Excel. There are still many cases where one program can read another program's data files directly, but if I were asked today "Where do data files come from?" my response would be, "Probably from someone's Excel spreadsheet."

Still, it is not good enough merely that data be entered into a spreadsheet. There are ways to enter data so that they are nearly unusable. Many spreadsheets require considerable massaging before they are suitable for analysis. This note provides some guidelines to minimize the work of importing Excel spreadsheets into standard statistical packages.

The standard data structure is a table of numbers in which each row corresponds to an individual subject and each column corresponds to a different variable or measurement. If for example, a study recorded the identification number, age, sex, height, and weight of 10 subjects, The resulting dataset would be composed of 10 rows and 5 columns.

Some recent programs use a different structure for repeated measurements on the same individual. The number of rows is determined by the total number of *measurements* rather than the number of subjects. Consider a study where 5 weekly blood pressure readings are made on each of 20 subjects. Until recently, the dataset would be composed of 20 rows and 6 columns (5 blood pressures and an id). Today, some programs want the data entered as 100 rows (5 rows for each of the 20 subjects) of 3 columns (id, week number(=1,2,3,4,5), blood pressure). The process of converting a dataset from one format to the other should be straightforward. The details will depend on the statistical program package.

Once the data structure is understood, the following rules should be observed to achieve a smooth transfer between a spreadsheet and a statistical program package.

- The first row of the spreadsheet should contain only legal variable names. The definition of legal will vary with the target program. All programs will accept names that are no more than eight characters long, are composed only of letters, numbers, and underscores, and begin with a letter. The limit of 8 characters has being increased in the latest round of program updates. This is helpful in cases with large numbers of variables because it is difficult to come up with a set of hundreds of names that are easily distinguished from each other and are easily understood. However, when variable names can be as long as 32 characters, it is easy to get carried away and end up with names that are tedious to type and manage.
- All rows but the first should contain only data. There should be no embedded formulas. The statistical programs may not be able to handle them. There are two ways to deal with formulas. One is to rewrite the formulas in the target package so the statistics package can generate the values. The other is to use the spreadsheet's cut and special paste capabilities to store the derived values as actual data values in the spreadsheet.
- No text should be entered in a column intended for numbers. This includes notations such as "missing", "lost", "N/A", and my personal favorite, "<20". If character strings are present, the statistical package may consider all of the data to be character strings rather than numbers. Numerical data may be mistakenly identified as character strings when one or more spaces are typed into an otherwise empty cell.
- When a study will generate multiple data files

    - Every record in every data file must contain a subject identifier that is consistent across files. This variable should have the same name in each data file. SUBJ, ID, STUDYID, and their variants are often used.
    - Data files that are likely to be merged should not use the same variable names (other than for the common ID varible). For example, if files of baseline and followup data both contain total cholesterol values, it would be better to call them BCHOL in the baseline file and FCHOL in the followup file rather than CHOL in both so that they will be distinct in

  - the merged file. If the same variable name is used in both files, it will be necessary to rename the variables as part of the file merging process. Otherwise, only a single column of values will appear in the merged dataset.
  - In order to protect subjects' privacy and confidentiality, only id numbers assigned for the study should be used as identifiers. Names and social security numbers should not be linked directly to raw data. While there are some arguments for allowing files to contain subjects' names as long as access is restricted to investigators and files never leave a research center, I find it more trouble than it's worth, if only because all printout containing names or social security numbers must be shredded.

The importance of including subject identifiers cannot be overstated. In some cases, files might contain no identifiers because investigators thought a study would generate only one file. In other studies, records might be identified inconsistently. For example, subjects' names might be used with the spelling or capitalization varying from file to file. In still other studies, some files might use subjects' names as identifiers while others might use sample or recruitment id numbers.

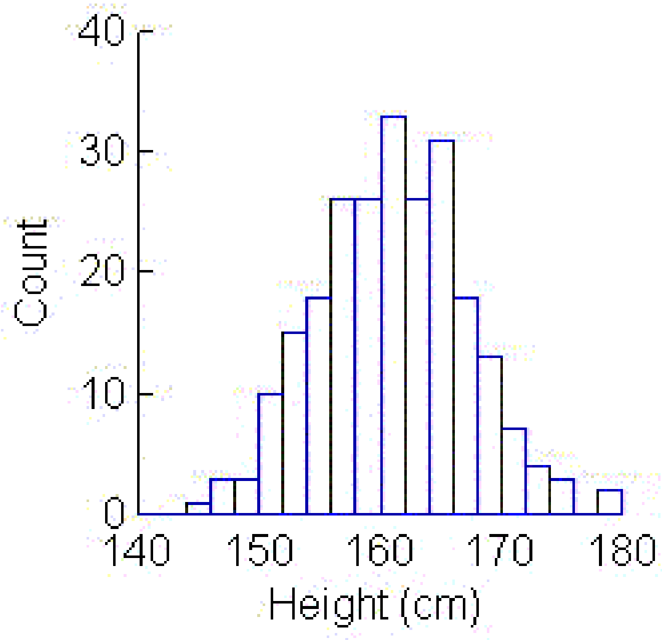
Inadequate identifiers make it difficult to merge files for analysis. It is highly probable that errors will result due to mismatching. The mismatching can be of two forms--either data from two subjects is combined or one subject's data gets split into multiple incomplete records. Regardless of the cause or the form of the mismatching, weeks, months, or even years of otherwise solid work will be compromised.

Last modified: undefined.

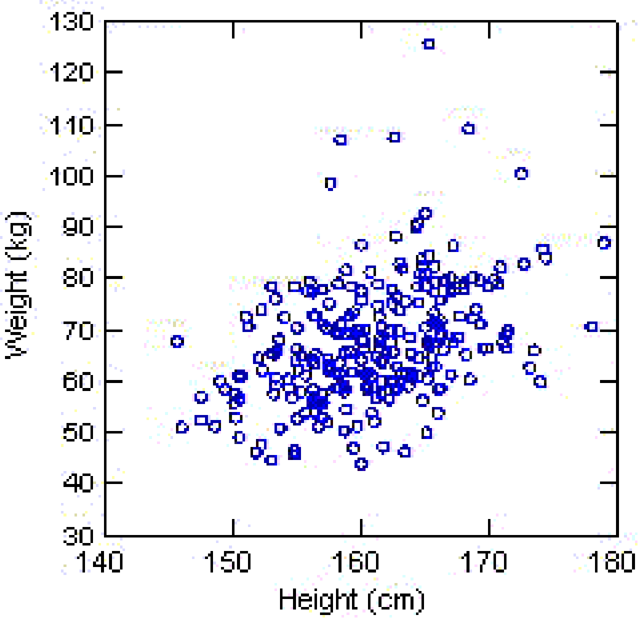
Look At The Data!

"You can observe a lot by watching." --Yogi Berra

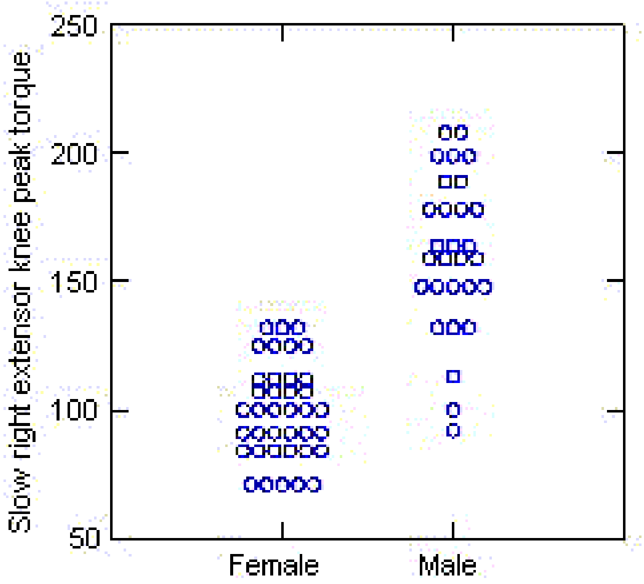
The first thing to do with any data set is look at it. If it fits on a single page, look at the raw data values. Plot them: histograms, dot plots, box plots, schematic plots, scatterplots, scatterplot matrices, parallel plots, line plots. Brush the data, lasso them. Use all of your software's capabilities. If there are too many observations to display, work with a random subset.



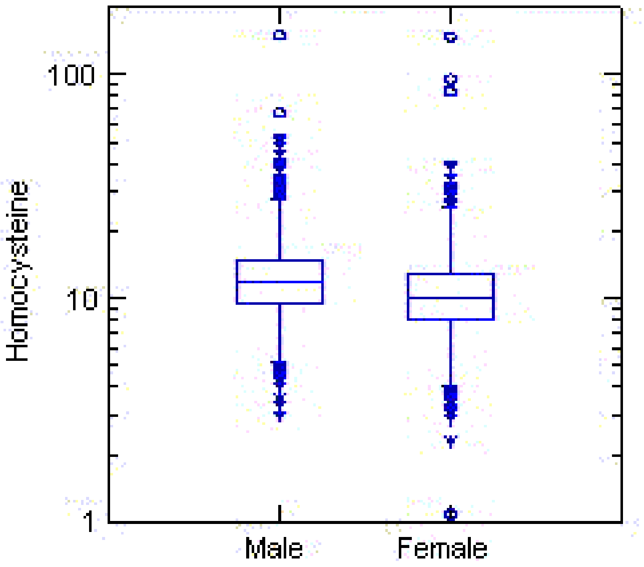
The most familiar and, therefore, most commonly used displays are histograms and scatterplots. With histograms, a single response (measurement, variable) is divided into a series of intervals, usually of equal length. The data are displayed as a series of vertical bars whose heights indicate the number of data values in each interval.



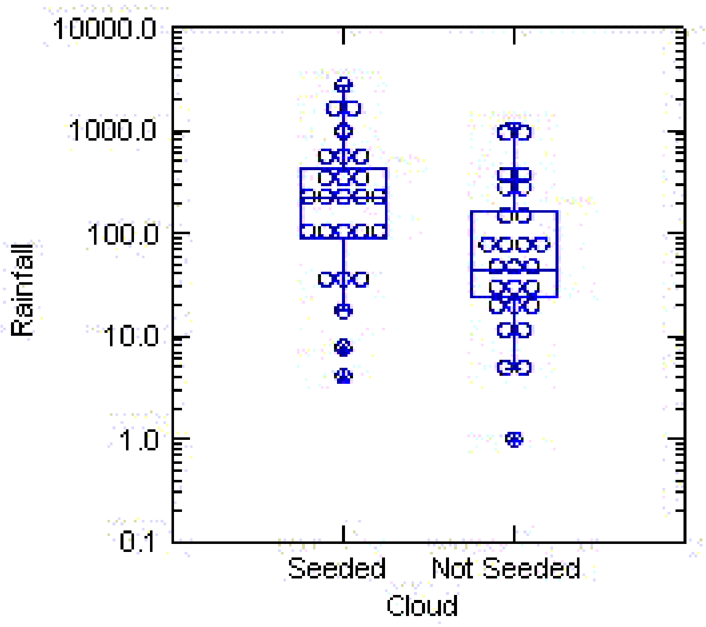
With scatterplots, the value of one variable is plotted against the value of another. Each subject is represented by a point in the display.



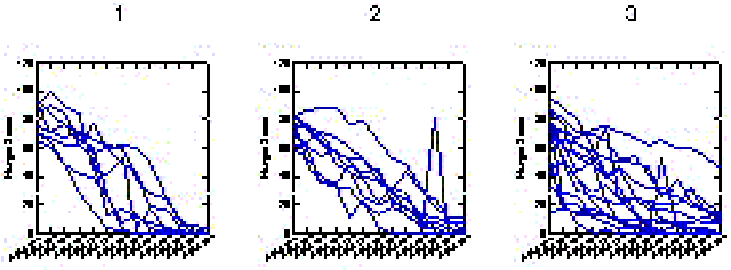
Dot plots (dot density displays) of a single response show each data value individually. They are most effective for small to medium sized data sets, that is, any data set where there aren't too many values to display. They are particularly effective at showing how one group's values compare to another's.



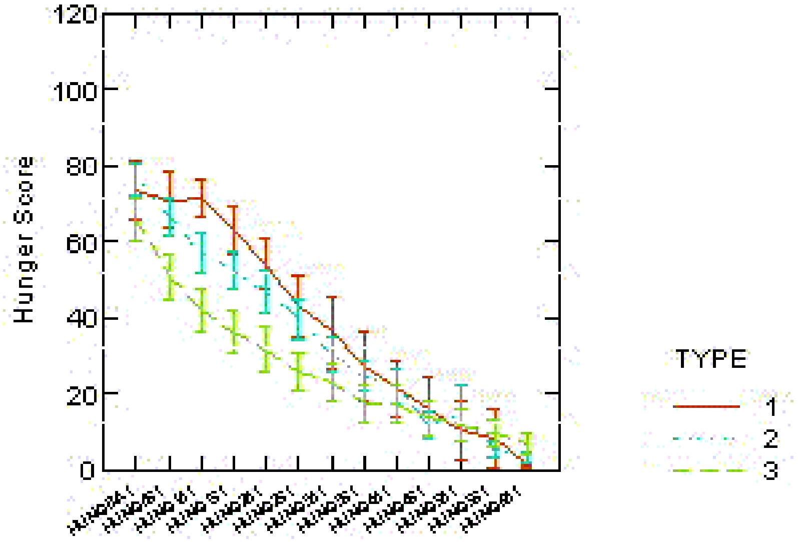
When there are too many values to show in a dotplot, a box plot can be used instead. The top and bottom of the box are defined by the 75-th and 25-th percentiles of the data. A line through the middle of the box denotes the 50-th percentile (median). Box plots have never caught on the way many thought they would. It may depend on the area of application. When data sets contain hundreds of observations at most, it is easy to display them in dot plots, making graphical summaries largely necessary. However, the box plots make it easy to compare medians and quartiles, and they are indispensible when displaying large data sets.

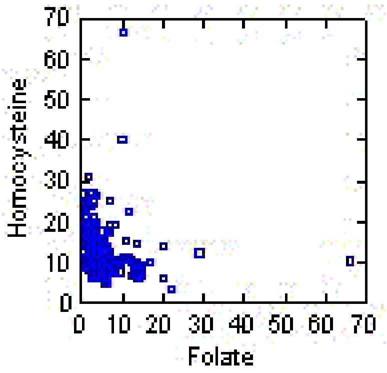


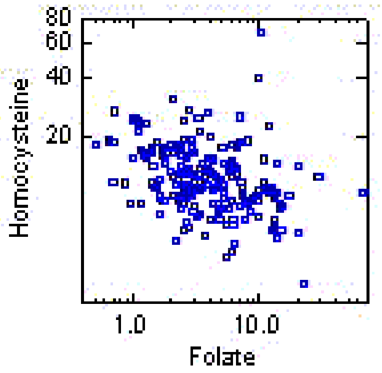
Printing a box plot on top of a dot plot has the potential to give the benefits of both displays. While I've been flattered to have some authors attribute these displays to me, I find them not to be as visually appealing as the dot and box plots by themselves...unless the line thicknesses and symbol sizes are *just right*, which they aren't in the diagram to the left in order to illustrate what I mean.



Parallel coordinate plots and line plots (also known as profile plots) are ways of following individual subjects and groups of subjects over time.







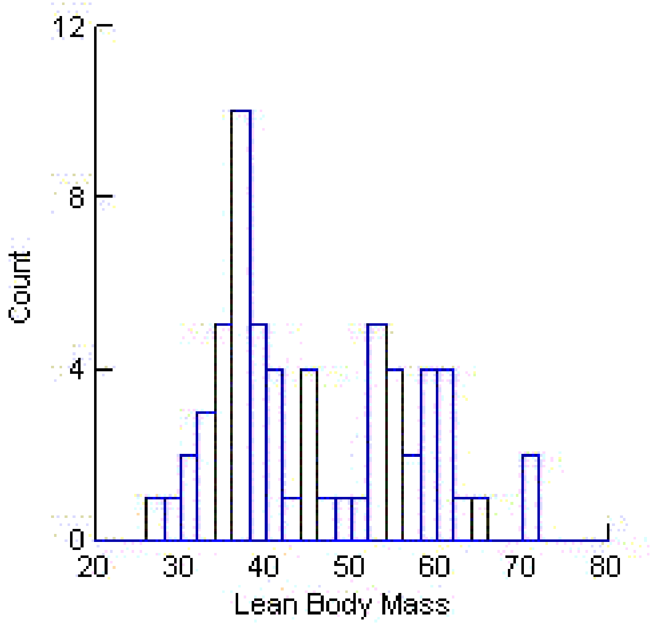
Most numerical techniques make assumptions about the data. Often, these conditions are not satisfied and the numerical results may be misleading. Plotting the data can reveal deficiencies at the outset and suggest ways to analyze the data properly. Often a simple transformation such as a log, square root, or square can make problems disappear.

The diagrams to the left display the relationship between homocysteine (thought to be a risk factor for heart disease) and the amount of folate in the blood. A straight line is often used to describe the general association between two measurements. The relationship in the diagram to the far left looks decidedly *non*linear. However, when a logarithmic transformation is applied to both variables, a straight line does a reasonable job of describing the decrease of homocysteine with increasing folate.

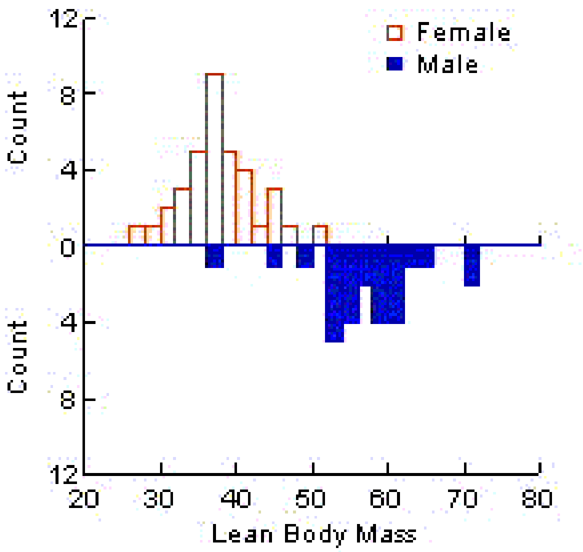
What To Look For:
A Single Response

The ideal shape for the distribution of a single response variable is symmetric (you can fold it in half and have the two halves match) with a single peak in the middle. Such a shape is called *normal* or a *bell-shaped curve*. One looks for ways in which the data depart from this ideal.

- Are there outliers--one or two observations that are far removed from the rest? Are there clusters as evidenced by multiple peaks?
- Are the data skewed? Data are said to be skewed if one of the tails of a histogram (the part that stretches out from the peak) is longer than the other. Data are skewed to the right if the right tail is longer; data are skewed to the left if the left tail is longer. The former is common, the latter is rare. (Can you think of anything that is skewed to the left?) Data are long-tailed if both tails are longer than those of the ideal normal distribution; data are short-tailed if the tails are shorter. Usually, a normal probability plot is needed to assess whether data are short or long tailed.



Is there more than one peak?



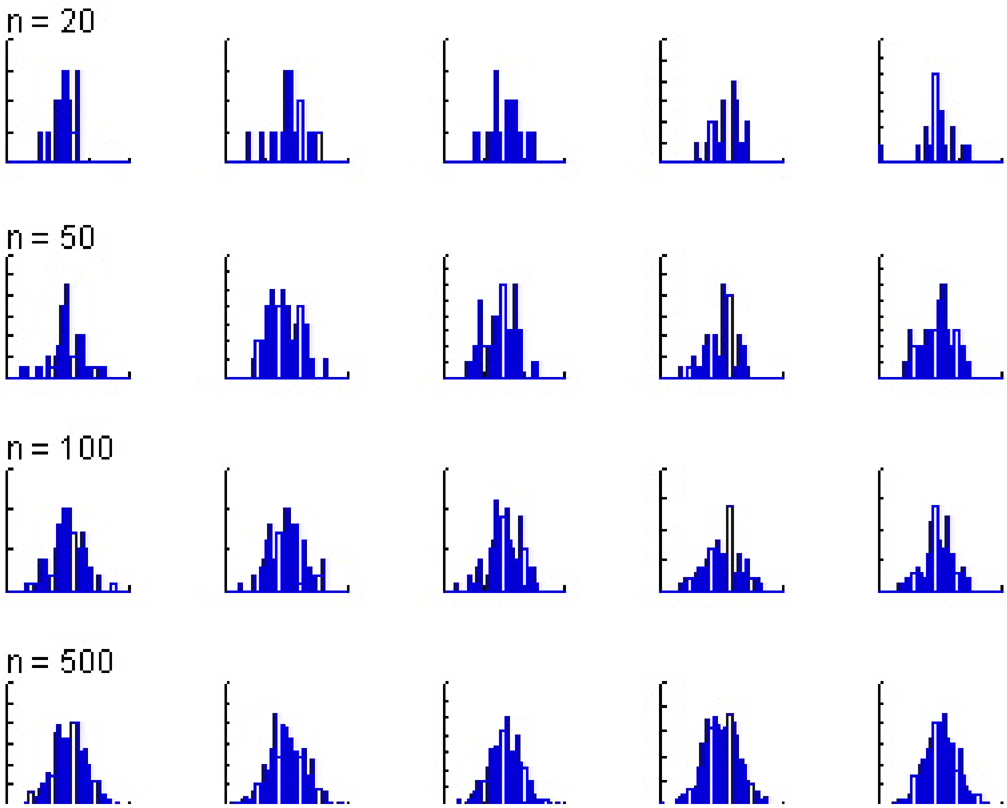
If data can be divided into categories that affect a particular response, the response should be examined within each category. For example, if a measurement is affected by the sex of a subject, or whether a subject is employed or receiving public assistance, or whether a farm is owner-operated, the response should be plotted for men/women, employed/ assistance, owner-operated/not separately. The data should be described according to the way they vary from the ideal *within each category*. It is helpful to notice whether the variability in the data increases as the typical response increases.

Many Responses

The ideal scatterplot shows a cloud of points in the outline of an ellipse. One looks for ways in which the data depart from this ideal.

- Are there outliers, that is, one or two observations that are removed from the rest? It is possible to have observations that are distinct from the overall cloud but are *not* outliers when the variables are viewed one at a time! Are there clusters, that is, many distinct clouds of points?
- Do the data seem to be more spread out as the variables increase in value?
- Do two variables tend to go up and down togther or in opposition (that is, one increasing while the other decreases)? Is the association roughly linear or is it demonstrably nonlinear?

Comment

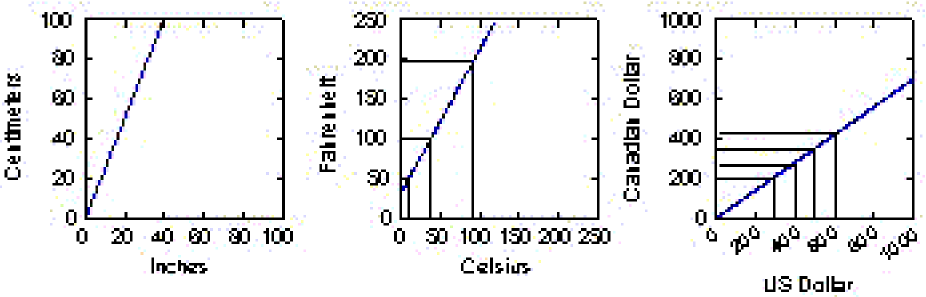


If the departure from the ideal is not clear cut (or, fails to pass what L.J. Savage called the "Inter-Ocular Traumatic Test"--It hits you between the eyes!), it's not worth worrying about. For example, consider this display which shows histograms of five different random samples of size 20, 50, 100, and 500 from a normal distribution. By 500, the histogram looks like the stereotypical bell-shaped curve, but even samples of size 100 look a little rough while samples of size 20 look nothing like what one might expect. The moral of the story is that **if it doesn't look worse than this, don't worry about it!**

Last modified: undefined.

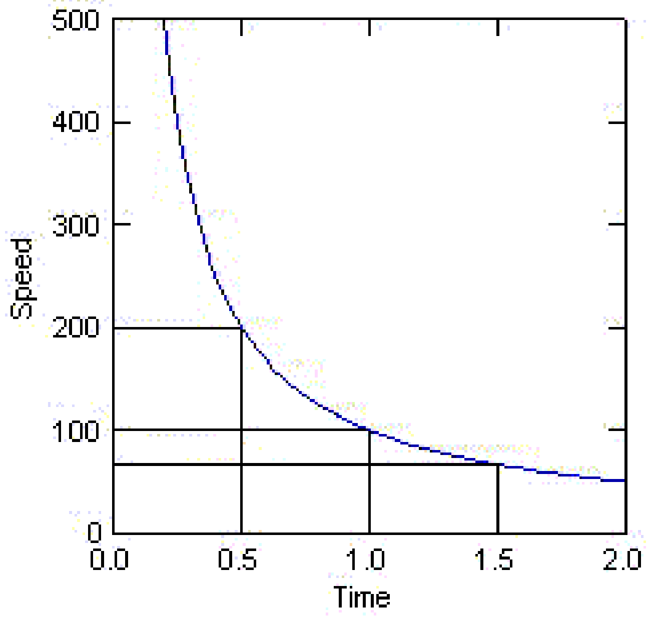
Logarithms

[When I started to write this note, I thought, "Why reinvent the wheel?" so I searched the World Wide Web for *logarithm*. I found some nice web pages--at Oak Road Systems and SCT BOCES, for example. However, they tend to be variations of the standard presentations found in most textbooks. If that type of presentation was sufficient for a general adult audience, then there wouldn't be so many people who were uncomfortable with logarithms! Here's my attempt to approach the subject in a different way. Call it *Logarithms: Part One*. For *Part Two*, search the World Wide Web. There are some excellent discussions out there!]

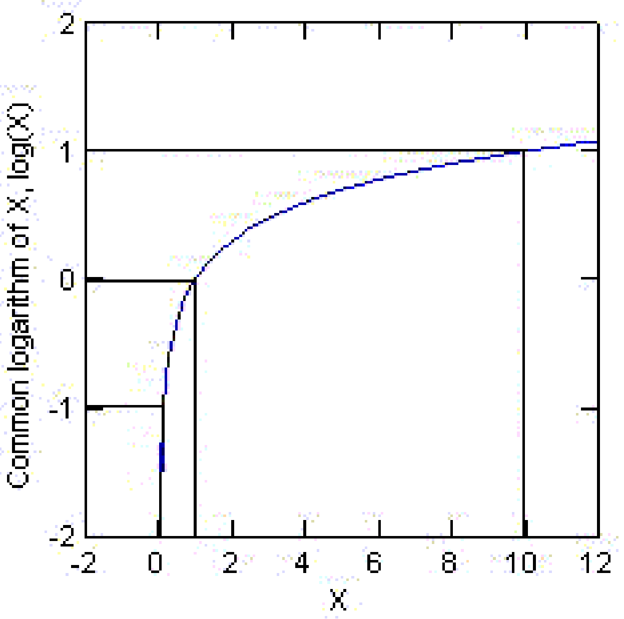


Let's talk about transformations. Some transformations are so commonplace it seems strange to give them a name as formidable as *transformations*--things like centimeters to inches, pounds to kilograms, Fahrenheit to Celsius, and currency conversions. Transformations like these are *linear transformations*. If you take a set of data, transform them, and plot the transformed values against the originals, the points will lie **exactly** on a straight line.

One characteristic of linear transformations is that they preserve relative spacings. Values that are evenly spaced before transformation remain evenly spaced after transformation. Values that are spaced twice as far apart as other values before transformation remain twice as far apart after transformation.



There are common transformations that are not linear. For example, a 100-mile journey can be described by the time it takes (duration) or by the speed of the trip. Since speed is defined as distance divided by duration (or, speed = distance / time), a 1 hour trip is a 100 mph trip, a 2 hour trip is a 50 mph trip, and so on. A plot of speed against gives a curve that is demonstrably nonlinear, but this is a transformation nonetheless. Each speed corresponds to a particular duration, and vice-versa. Nonlinear transformations do not preserve relative spacings. For example, consider the equally spaced durations of 0.5 hours, 1 hour, and 1.5 hours. When expressed as speeds, they are 200 mph, 100 mph, and 66.7 mph.



The logarithm is another nonlinear transformation. Got it? In the spirit of the late Lenny Bruce, lets repeat it so that the word *logarithm* loses some of its shock value.

- The logarithm is just a transformation!
- The logarithm is just a transformation!
- The logarithm is just a transformation!

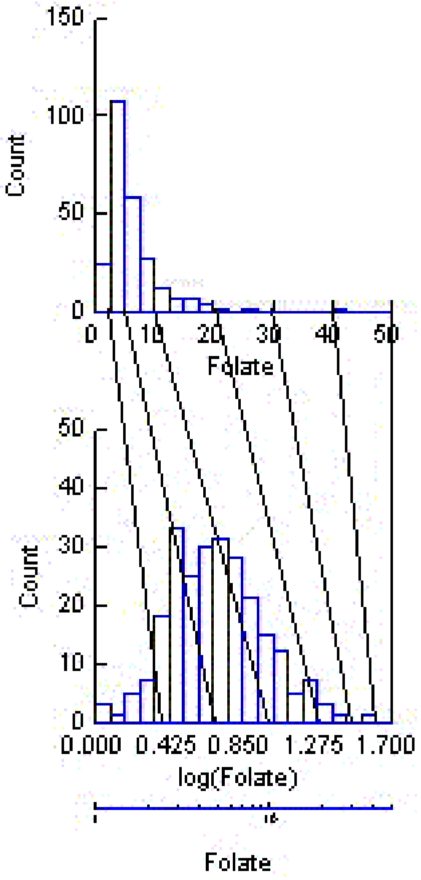
To keep things simple, we'll stick with the kind called *common logarithms* and use the informal name *common logs*. Common logs have the following fascinating property--if you **multiply** something by 10 in the original scale, you **add** 1 unit to its value the log scale. If you **divide** something by 10 in the original scale, you **subtract** 1 unit from its value in the log scale. As we move from 0.1 to 1 to 10 on the original scale, we move from -1 to 0 to 1 on the logarithmic scale,

There are three reasons why logarithms should interest us.

- First, many statistical techniques work best with data that are single-peaked and symmetric (*symmetry*).
- Second, when comparing different groups of subjects, many techniques work best when the variability is roughly the same within each group (*homoscedasticity*).
- Third, it is easier to describe the relationship between variables when it's approximately linear (*linearity*).

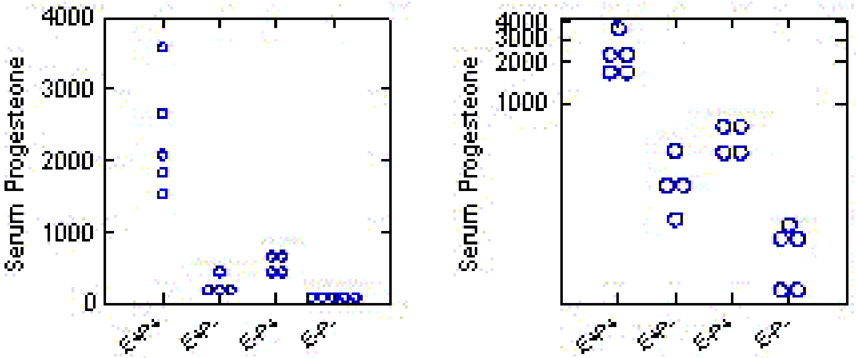
When these conditions are not true in the original data, they can often be achieved by applying a logarithmic transformation.

Symmetry



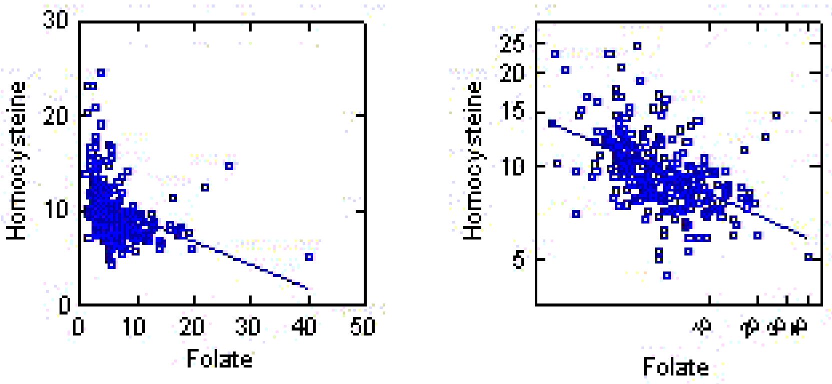
A logarithmic transformation will reduce positive skewness because it compresses the upper end (tail) of the distribution while stretching out the lower end. This is because the distances between 0.1 and 1, 1 and 10, 10 and 100, and 100 and 1000 are the same in the logarithmic scale. This is illustrated by the histogram of folate levels in a sample of healthy adults. In the original scale, the data are long-tailed to the right, but after a logarithmic transformation is applied, the distribution is symmetric. The lines between the two histograms connect original values with their logarithms to demonstrate the compression of the upper tail and stretching of the lower tail.

Homoscedasticity



Often groups that tend to have larger values also tend to have greater within-group variability. A logarithmic transformation will often make the within-group variability more similar across groups. The figure shows the serum progesterone levels in subjects randomly assigned to receive estrogen and (separately) progesterone. In the original scale, variability increases dramatically with the typical response. However, the within-group variability is nearly constant after a logarithmic transformation is applied. Also, in the logarithmic scale, the data tell a simpler story, In the log scale, the effect of progesterone is the same whether or not a subject is taking estrogen. Also, the effect of estrogen is the same whether or not a subject is taking progesterone.

Linearity



Logarithmic transformations are sometimes used when constructing statistical models to describe the relationship between two measurements. Consider homocysteine. It's bad stuff, a sulphur based amino acid that indicates risk of heart disease. Lately, it's been hard to escape advertising that tells you to drink your orange juice because orange juice is a good source of folate, which lowers your homocysteine.

A plot of homocysteine against folate shows a nonlinear relationship with most of the data bunched in the lower left hand portion of the display. When logarithmic transformations are applied to both variables, the association appears to be linear. The fitted equation is

$$\log(\text{homocysteine}) = 1.14 - 0.23 \log(\text{folate}) .$$

If someone has folate levels of 20, her logged folate levels are log(20) or 1.301. Her logged homocysteine value will be estimated to be 1.14 - 0.23 * 1.301 or 0.8408 units. If logged homocysteine is 0.8408, homocysteine itself is $10^{0.8408}$ or 6.93 units.

Some things to notice and/or do

- The common log of a number is the power to which 10 is raised in order to obtain the number. This makes the logs of some numbers easy to calculate. For example,
  1og(1) = 0, because $10^0 = 1$
  1og(10) = 1, because $10^1 = 10$
  1og(100) = 2, because $10^2 = 100$
- Every positive number has a logarithm. You can get the logarithms of numbers that aren't integer powers of 10 from tables or a calculator. For example, log(48) = 1.6812 and log(123) = 2.0899. What is log(480)?.
- The larger the number, the larger its logarithm. Thus, 123 > 48, so log(123) > log(48).
- Only positive numbers can have logarithms. Why? Hint: Think about powers of 10.
- Can logarithms themselves be negative? Yes. Give an example of a number whose logarithm is negative. Hints: What number has a logarithm of 0? The smaller the number, the smaller its logarithm.
- Use a calculator to obtain the common log of some number. Use the calculator to transform back from the logarithm to the original number. "Transforming back" is known as taking the antilogarithm.
- Use a calculator to obtain the common log of some number. Add 1 to the logarithm. Take the antilog of the result. What do you get? How is it related to the number you started with?
- Use a calculator to obtain the common log of some number. Add 0.3010 to the logarithm. Take the antilog. What number do you get? How is it related to the number you started with. Think about it. If your head hurts, try the next exercise!
- Use the calculator to get the antilogarithm of 0.3010. Hmmm . . . The previous paragraph demonstrates that the sum of two logarithms is equal to the logarithm of their product. We took the log of a number, added the log of 2, and obtained the log of twice the original number! But this is getting way too technical.

Ratios

There are two commonly used ways to summarize a difference between two groups. The first is the algebraic difference--for example, changing to this diet will lower your blood pressure 20 mm. The second is the relative change--for example, this diet will lower your cholesterol by 15%. Relative changes are often expressed in terms of ratios, one treatment's response divided by another.

One problem with ratios is that their lack of symmetry. Consider the ratio of A to B, that is, A/B. If A produces values greater than B, the ratio can take theoretically take any value greater than 1. However, if A produces values less than B, the ratio is restricted to the range of 0 to 1. To put it another way, if we change the way we define our ratio--switching to B/A--values in the range 1 to infinity move into the range 0 to 1 while values in the range 0 to 1 get switched into the range 1 to infinity.

Logged ratios solve this problem. Again consider the ratio, A/B. When their effects are the same, their ratio is 1 and the log of the ratio is 0. Also, log(A/B) = -log(B/A), so symmetry is restored. That is, when B is greater A, the log of the ratio has the same magnitude as when A is the same number of multiples of B except that the sign is different. You can use your calculator to check this for various choices of A and B. This is why I rarely analyze ratios but often analyze logged ratios. I might analyze the ratios directly if they are tightly grouped around 1, say, 0.9 to 1.1. There may still be some assymetry, but it will be minor (1/1.1 = 0.0909), and a fair cost for sparing the audience from dealijng with logarithms.

The last thing to bring into the discussion is the logarithm's property that log of a ratio is the difference of the logs, that is, log(A/B) = log(A) - log(B). Many statistical techniques work best when they are describing the algebraic difference between two quantities. Therefore, when it is natural to think of some quantity in terms of ratios rather than simple differences. it is common for analysts to begin with a logarithmic transformation of the data and perform a formal analysis on the logarithms.

Logarithms also play an important role in analyzing probabilities. Statisticians have developed many techniques for fitting straight-line models to predict a variety of outcomes. There is a problem when using these methods to model probabilities. The estimated probabilities can be less than 0 or greater than 1, which are impossible values. Logistic regression models the log odds (odds = probability/(1-probability)) instead. While probabilities must lie between 0 and 1 (with a neutral value of 1/2), odds are ratios that lie between 0 and infinity (with a neutral value of 1). It follows from the discussion two paragraphs above, that log odds can take on any value, with a neutral value of 0 and the log odds in favor of an event being equal in magnitude and opposite in sign to the same odds against the event. Whew!

Different types of logarithms

Just as length can be measured in feet, inches, meters, kilometers, centimeters, or whatever, logarithms can be defined in may ways according to what happens in the original scale when there is a one unit change in the log scale. Common logs are defined so that a 1 unit increase in the log scale is equivalent to multiplying by 10 in the original scale. One could define logarithms so that a one unit increase in the log scale is equivalent to multiplying by 2 in the original scale. These would be called *logs to base 2*. The value by which a number is multiplied in the original scale when its logarithm is increased by 1 is known as the *base* of the logarithm. Any positive number different from 1 can be used as a base.

Mathematicians are fond of *natural logarithms*. A 1 unit increase in this log scale is equivalent to multiplying in the original scale by a factor known as Euler's constant, *e* (approximately 2.71828). Mathematicians like natural logs because they have properties that are not shared by other types of logarithms. For example, if you apply any logarithmic transformation to a set of data, the mean (average) of the logs is approximately equal to the log of the original mean, whatever type of logarithms you use. However, only for natural logs is the measure of spread called the standard deviation (SD) approximately equal to the coefficient of variation (the ratio of the SD to the mean) in the original scale.

However, as already mentioned, the different tyupes of logs are like different units for measuring height. You can't report a height of 71. Is it the 71 inches that might be appropiate for an adult male, or is it the 71 cm that might be appropriate for a toddler? Similarly, you can't report a logarithm of 2. Is it the common log corresponding to a value of 100 in the original scale or a natural log corresponding to a value of 7.39?

Pick a number and write it down. A one or two digit number will do.

- Enter the number into your calculator.
  - Press the LOG key. Note the result.
  - Press the 10^x key. Note the result.
- Enter the number into your calculator.
  - Press the LN key. Note the result.
  - Press the e^x key. Note the result.
- Enter the number into your calculator.
  - Press the LOG key. Note the result.
  - Press the e^x key. Note the result.
- Enter the number into your calculator.
  - Press the LN key. Note the result.

- Press the 10^x key. Note the result.

It really doesn't matter what kind of logarithms you use. It's like choosing units of length. If you measure something in feet and later want it in inches, just multiply by 12. You can also switch between different kind of logarithm by multiplying by the proper constant*.

A comment about notation. For nonmathematicians there are *at most* two kinds of logarithms--common logs, denoted by log(x), and maybe Natural logs, denoted ln(x). Mathematicians like general notation and write logarithms as $\log_b(x)$, where *b* denotes the base. Thus, common logs are written $\log_{10}$. If mathematicians were entirely consistent, natural logs would be written $\log_e$. However, mathematicians use natural logs almost to the exclusion of all others, so 'log' written without a base is understood to stand for natural logs. I do this myself when I am writing mathematics.

It is quite different when I am describing the results of an analysis to a general audience. I go out of my way to avoid using the word *logarithm*. If an analysis demands logs, I often write, "Data in tables and graphs are displayed in the original scale of measurement. However, <for these stated reasons> a logarithmic transformation was applied to the data prior to formal analysis." If I had to discuss logged data formally with a general audience, I would write log(x) for common log and ln(x) for natural log, but I'd do my best to avoid using natural logs. Logarithms of any kind place huge demands on a general audience. They risk confusion and greatly increase the chance the audience will tune out and the message will get lost. If logarithms *must* be used, it is essential to do it in a way that causes the least amount of discomfort for the audience--common logs denoted by 'log'.

Summary

Logarithms are just another transformation. We use them because sometimes it's easier to analyze or describe something in terms of log transformed data than in terms of the original values.

----

*To transform common logs (base 10) to natural logs (base *e*), multiply the common logs by 2.3026, the natural log of 10. Try it with your calculator. Take the common log of 253. It is 2.4031. Multiply it by 2.3026 to get 5.5334. Now press the e^x key to get 253! To transform natural logs (base *e*) to common logs (base 10), the constant is 0.4343, the common log of *e*. In general,

$$\log_b(x) = \log_b(a)\,\log_a(x)\text{ , and}$$
$$\log_b(x) = \log_a(x)\ /\ \log_a(b)$$

[back to The Little Handbook of Statistical Practice]

Last modified: undefined.

Summary Statistics:
Location & Spread

Prologue: Terminology

A sample is a set of observations drawn from a larger **population**. A sample is usually drawn to make a statement about the larger population from which it was taken. *Sample* and *population* are two different things and it is essential to maintain the distinction between them so that we can express ourselves clearly and be understood. John Tukey has suggested using the word **batch**, as in "a batch of numbers", to describe a set of numbers when it doesn't matter whether they are a sample, a population, or of unknown origin. While I'm sympathetic to his suggestion, it does not seem to have been adopted widely.

Descriptive Statistics

After constructing graphical displays of a batch of numbers, the next thing to do is summarize the data numerically. **Statistics** are summaries derived from the data. The two important statistics that describe a single response are measures of location (on the number line) and spread. The number of observations (the sample size, *n*) is important, too, but it is generally considered a "given". It is not counted as one of the summary statistics.

Mean and Standard Deviation

There are many reasonable single number summaries that describe where a set of values is located. Any statistic that describes a typical value will do. Statisticians refer to these measures, as a group, as averages.

The most commonly reported average is the **mean**--the sum of the observations divided by the sample size. The mean of the values 5, 6, 9, 13, 17 is (5+6+9+13+17)/5 or 50/5 = 10.

The mean is invariably what people intend when they say **average**. *Mean* is a more precise term than *average* because the *mean* can *only* be the sum divided by the sample size. There are other quantities that are sometimes called averages. These include the **median** (or middle value), the mode (most commonly occurring value), and even the **midrange** (mean of minumum and maximum values). Statisticians prefer means because they understand them better, that is, they understand the relation between sample and population means better than the relation between the sample and population value of other averages.

The most commonly reported measure of variability or spread is the **standard deviation** (SD). The SD might also be called the "root-mean-square deviation", which describes the way it is calculated. The operations root, mean, and square are applied in reverse order to the *deviations*--the individual differences between the observations and the mean. First, the deviations are squared. Next, the mean of the deviations is calculated. Finally, the square root of the mean is taken to obtain the SD. To be precise, when the mean is taken, the sum of the squared deviations is divided by *one less than the sample size* rather than the sample size itself. There's no reason why it *must* be done this way, but this is the modern convention. It's not important that this seem the most natural measure of spread. It's the way it's done. You can just accept it (which I recommend) or you'll have to study the mathematics behind it (but that's another course).

To see how the SD works, consider the values 5,6,9,13,17, whose mean as we've already seen is 10. The deviations are {(5-10), (6-10), (9-10), (13-10), (17-105)} or -5, -6, -1, 3, 7. (It is not an accident that the deviations sum to 0, but I digress.) The squared deviations are 25, 16, 1, 9, 49 and the standard deviation is the square root of (25+16+1+9+49)/(5-1), that is, the square root of (100/4) or $\sqrt{25} = 5$.

Why do we use something that might seem so complicated? Why not the range (difference between the highest and lowest observations) or the mean of the absolute values of the deviations? Without going into details, the SD has some attractive mathematical properties that make it the measure of choice. It's easy for statisticians to develop statistical techniques around it. So we use it. In any case, the SD satisfies the most important requirement of a measure of variability--the more spread out the data, the larger the SD. And the best part is, we have computers to calculate the SD for us. We don't have to compute it. We just have to know how to use it...properly!

Some Mathematical Notation

Back in olden times, mathematics papers contained straightforward notation like

$$a+b+c+d+...$$

It was awkward having all of those symbols, especially if you wanted to be adding up heights, weights, incomes, and so on. So, someone suggested using subscripts and writing sums in the form

$$x_1+x_2+x_3+x_4+\ldots+ x_n,$$

where 'n' is the sample size or number of observations, and using different letters for each quantity ('h' for heights, 'w' for weights, and so on).

The plus signs could be eliminated by writing the expression as

$$\mathrm{Sum}(x_1,x_2,x_3,x_4,\ldots,x_n)$$

and once people were used to, Sum could be abbreviated to just S, as in

$$S(x_1,x_2,x_3,x_4,\ldots, x_n).$$

The notion of limits of notation was then introduced so the expression could be reduced to $S(x_i,:i=1,\ldots, n)$ and the limits of notation were moved to decorate the "S"

$$\overset{n}{\underset{i=1}{S}} x_i$$

Now all that was left was to replace the S by its Greek equivalent, sigma, and here we are in modern times!

$$\sum_{i=1}^{n} x_i$$

Because we almost always sum from 1 to 'n', the limits of summation are often left off unless the sum is *not* from 1 to 'n'.

Now that we have this nice notation, let's use it to come up with expressions for the sample mean, which we'll write as the letter 'x' with a bar over it, and the standard deviation, s. The mean is easy. It's the sum of the observations (which we've already done) divided by the sample size

$$\overline{x} = \frac{\sum x_i}{n} \, .$$

The standard deviation isn't much more difficult. Recall "root-mean-square". Begin with the deviations

$$(x_i - \overline{x})$$

then square them

$$(x_i - \overline{x})^2$$

then take their "mean"

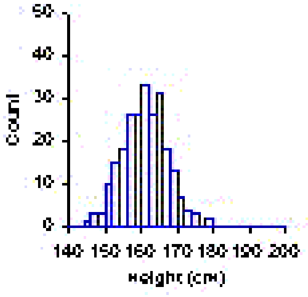
$$\frac{\sum (x_i - \overline{x})^2}{n-1}$$

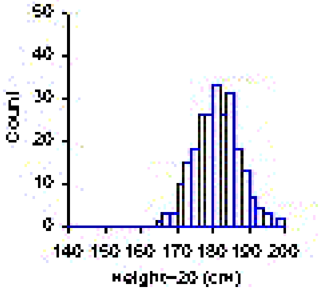
then take a square root

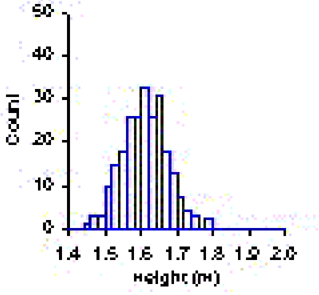
$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}} \, .$$

All done.

Some facts about the mean and standard deviation







If you're the mathematical type, you can prove these statements for yourself by using the formulas just developed for the mean and standard deviation. If you're the visual type, you should be able to see why these results are so by looking at the pictures to the left.

- When a constant is added to every observation, the new sample mean is equal to original mean plus the constant.
- When a constant is added to every observation, the standard deviation is unaffected.
- When every observation is multiplied by the same constant, the new sample mean is equal to original mean multiplied by the constant.
- When every observation is multiplied by the same constant, the new sample standard deviation is equal to original standard deviation multiplied by the magnitude of the constant. (The reason for including the phrase "the magnitude of" is that if the constant is negative, the sign is dropped when the new SD is calculated.)

Mental Pictures

The mean and SD are a particularly appropriate summary for data whose histogram approximates a normal distribution (the bell-shaped curve). If you say that a set of data has a mean of 220, the typical listener will picture a bell-shaped curve centered with its peak at 220.

What information does the SD convey? When data are approximately normally distributed,

- approximately 68% of the data lie within one SD of the mean.
- approximately 95% of the data lie within two SDs of the mean.
- approximately 99.7% of the data lie within three SDs of the mean.

For example, if a set of total cholesterol levels has a mean of 220 mg/dl and a SD of 20 mg/dl and its histogram looks like a normal distribution, then about 68% of the cholesterol values will be in the range 200 to 240 (200 = 220 - 20 and 240 = 220 + 20). Similarly, about 95% of the values will be in the range 180 to 260 (180 = 220 - 2*20 and 280 = 220 + 2*20) and 99.7% of the values will be in the range 160 to 280 (160 = 220 - 3*20 and 280 = 220 + 3*20). Why do we add relatively fewer observations as the number of SDs increases? Because of the bell-shaped curve with its peak in the middle.

Percentiles

When the histogram of the data does not look approximately normal, the mean and SD can be misleading because of the mental picture they paint. Give people a mean and standard deviation and they think of a bell-shaped curve with observations equally likely to be a certain distance above the mean as below. But, there's no guarantee that the data aren't really skewed or that outliers aren't distorting the mean and SD, making the rules stated earlier invalid for that particular data set.

One way to describe such data in a way that does not give a misleading impression of where they lie is to report some percentiles. The p-th percentile is the value that p-% of the data lie are less than or equal to. If p-% of the data lie below the p-th percentile, it follows that (100-p)-% of the data lie above it. For example, if the 85-% percentile of household income is $60,000, then 85% of households have incomes of $60,000 or less and the top 15% of households have incomes of $60,000 or more.

The most famous of all percentiles--the 50-th percentile--has a special name: the **median**. Think of the median as the value that splits the data in half--half of the data are above the median; half of the data are below the median[*]. Two other percentiles with special names are the **quartiles**: the lower quartile (the 25-th percentile) and the upper quartile (the 75-th percentile). The median and the quartiles divide the data into quarters. One-quarter of the data is less than the lower quartile; one-quarter of the data falls between the lower quartile and the median; one-quarter of the data falls between the median and the upper quartile; one-quarter of the data is greater than the upper quartile.

Sometimes the minimum and maximum are presented along with the median and the quartiles to provide a five number summary of the data. Unlike a mean and SD, this five number summary can be used to identify skewed data. When there are many observations (hundreds or thousands), some investigators report the 5-th and 95-th percentiles (or the 10th and 90- th or the 2.5-th and the 97.5-th percentiles) instead of the minimum and maximum to establish so-called **normal ranges**.

You'll sometimes see the recommendation that the **Inter-Quartile Range** (the difference between the upper and lower quartiles) be reported as a measure of spread. It's certainly a measure of spread--it measures the spread of the middle half of the data. But as a pair, the median and IQR have the same deficiency as the mean and the SD. There's no way a two number summary can describe the skewness of the data. When one sees a median and an IQR, one suspects they are being reported because the data are skewed, but one has no sense of how skewed! It would be much better to report the median and the quartiles.

In practice, you'll almost always see means and SDs. If your goal is to give a simple numerical summary of the distribution of your data, look at graphical summaries of your data to get a sense of whether the mean and SD might produce the wrong mental picture. If they might, consider reporting percentiles instead.

[I'm getting a bit ahead of myself with this paragraph, but in many cases researchers are not concerned with describing the distribution of individual responses. Instead, they focus on how well a sample mean might be estimating a population mean. In these cases, they report the mean and a measure of uncertainty called the **standard error of the mean**. This type of summary is appropriate even when the histogram of the data is not normal. We will discuss this in detail later. I mention it now because I it can be confusing to spend time discussing something that never seems to arise in practice. Percentiles are given the same amount of discussion here as means and SDs, even though they are used much less often. But that's because it takes a minimum amount of time to discuss *anything*! Even though percentiles aren't used as often as means and SDs, it's important to know why they are sometimes necessary. To prepare you for what goes on in practice, I wanted to say that you will rarely see percentiles reported. That's true, but I couldn't write that you'll almost always see means and *SD*s! I would guess that most papers report means and *SEM*s. Among the rest, the vast majority report means and SDs and a few report percentiles.]

Mean versus Median

The mean is the sum of the data divided by the sample size. If a histogram could be placed on a weightless bar and the bar on a fulcrum, the histogram would balance perfectly when the fulcrum is directly under the mean. The median is the value in the middle of the histogram. If the histogram is symmetric, the mean and the median are the same. If the histogram is not symmetric, the mean and median can be quite different. Take a data set whose histogram is symmetric. Balance it on the fulcrum. Now take the largest observation and start moving it to the right. The fulcrum must move to the right with the mean, too, if the histogram is to stay balanced. You can make the mean as large as you want by moving this one observation farther and farther to the right, but all this time the median stays the same!

A point of statistical trivia: If a histogram with a single peak is skewed to the right, the order of the three averages lie along the measurement scale in reverse alphabetical order--mode, median, mean.

Geometric Mean

When data do not follow a normal distribution, reports sometimes contain a statement such as, "Because the data were not normally distributed, {some transformation} was applied to the data before formal analyses were performed. Tables and graphs are presented in the original scale."

When data are skewed to the right, it often happens that the histogram looks normal, or at least symmetric, after the data are logged. The transformation would be applied prior to formal analysis and this would be reported in the Statistical Methods section of the manuscript. In summary tables, it is common for researchers to report **geometric means**. The geometric mean is the antilog of the mean of the logged data--that is, the data are logged, the mean of the logs is calculated, and the anti-log of the mean is obtained. The presence of geometric means indicates the analysis was done in the log scale, but the results were transformed back to the original scale for the convenience of the reader.

If the histogram of the log-transformed data is approximately symmetric, the geometric mean of the original data is approximately equal to the median of the original data. The logarithmic transformation is monotone, that is, if a<b, then log(a)<log(b) and vice-versa. The logs of observations are ordered the same way the original observations are ordered. Therefore, the log of the median is the median of the logs[**]. The reverse is true, too. The anti-log of the median of the logs is the median of the original values. In the log scale where the histogram is symmetric, the mean and the median are about the same. Therefore, the geometric mean (anti-log of the mean) will be approximately equal to the anti-log of the median, which is the median in the original scale.

Is there a geometric SD? Yes. It's the antilog of the SD of the log transformed values. The interpretation is similar to the SD. If GBAR is the geometric mean and GSD is the geometric standard deviation, 95% of the data lie in the range from $\mathrm{GBAR}/(\mathrm{GSD}^2)$ to $\mathrm{GBAR}*(\mathrm{GSD}^2)$, that is, instead of adding and subtracting 2 SDs we multiply and divided by the square of the SD.

These differences follow from properties of the logarithm, namely,

$$\log(ab) = \log(a) + \log(b) \text{ and}$$

$$\log(a/b) = \log(a) - \log(b)$$

that is, the log of a product is the sum of the logs, while the log of a ratio is the difference of the logs.

Since the data are approximately normally distributed in the log scale, it follows that 95% of the data lie in the range mean-2SD to mean+2SD. But this is

$$\log(\mathrm{GBAR}) + \log(\mathrm{GSD}) + \log(\mathrm{GSD}) = \log(\mathrm{GBAR}*\mathrm{GSD}^2) \text{ and}$$

$$\log(\mathrm{GBAR}) - \log(\mathrm{GSD}) - \log(\mathrm{GSD}) = \log(\mathrm{GBAR}/\mathrm{GSD}^2)$$

----

[*] This definition isn't rigorous for two reasons. First, the median may not be unique. If there is an even number of of observations, then any number between the two middle values qualifies as a median. Standard practice is to report the mean of the two middle values. Second, if there is an odd number of observations or if the two middle values are tied, no value has half of the data greater than it and half less. A rigorous definition of the median is that it is *a* value such that at least half of the data are less

than or equal to it and half of the data are greater than or equal to it. Consider the data set 0,0,0,0,1,7. The median is 0 since 4/6 of the data are less than or equal to 0, while all of the data are greater than or equal to 0. Similar remarks apply to all other percentiles. However, so we don't get bogged down in details, let's think of the p-th percentile as the value that has "p-% of the data below it; (100-p)-% of the data above it".

**If the number of observations is even, it is more correct to say that the log of *a* median in the original scale is *a* median in the log scale. That's because when the number of observations is even, *any* value between the two middle values satisfies the definition of a median. Standard practice is to report the mean of the two middle values, but that's just a convention.
Consider a data set with two observations--10 and 100, with a median of 55. Their common logarithms are 1 and 2, with a median of 1.5. Now, log(55)=1.74 which is not 1.5. Nevertheless, 1.74 is *a* median in the log scale since it lies between the two middle values.

Last modified: undefined.

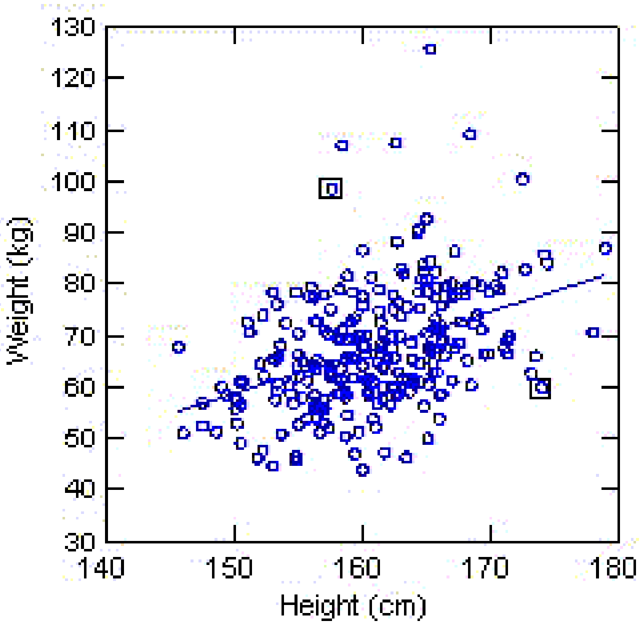
CORRELATION COEFFICIENTS

We've discussed how to summarize a single variable. The next question is how to summarize a pair of variables measured on the same observational unit--(percent of calories from saturated fat, cholesterol level), (amount of fertilizer, crop yield), (mother's weight gain during pregnancy, child's birth weight). How do we describe their joint behavior?

Scatterplots! Scatterplots! Scatterplots!

The first thing to do is construct a scatterplot, a graphical display of the data. There are too many ways to be fooled by numerical summaries, as we shall see!

The numerical summary includes the mean and standard deviation of each variable separately plus a measure known as the *correlation coefficient* (also the *Pearson correlation coefficient,* after Karl Pearson), a summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.



"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.

Correlations tend to be positive. Pick any two variables at random and they'll almost certainly be positively correlated, if they're correlated at all--height and weight; saturated fat in the diet and cholesterol levels; amount of fertilizer and crop yield; education and income. Negative correlations tend to be rare--automobile weight and fuel economy; folate intake and homocysteine; number of cigarettes smoked and child's birth weight.

The correlation coefficient of a set of observations {(x_i,y_i): i=1,..,n} is given by the formula

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The key to the formula is its numerator, the sum of the products of the deviations.

[Scatterplot of typical data set with axes drawn through (Xbar,Ybar)]

Quadrant	x(i)-xbar	y(i)-ybar	(x(i)-xbar)*(y(i)-ybar)
I	+	+	+
II	-	+	-
III	-	-	+
IV	+	-	-

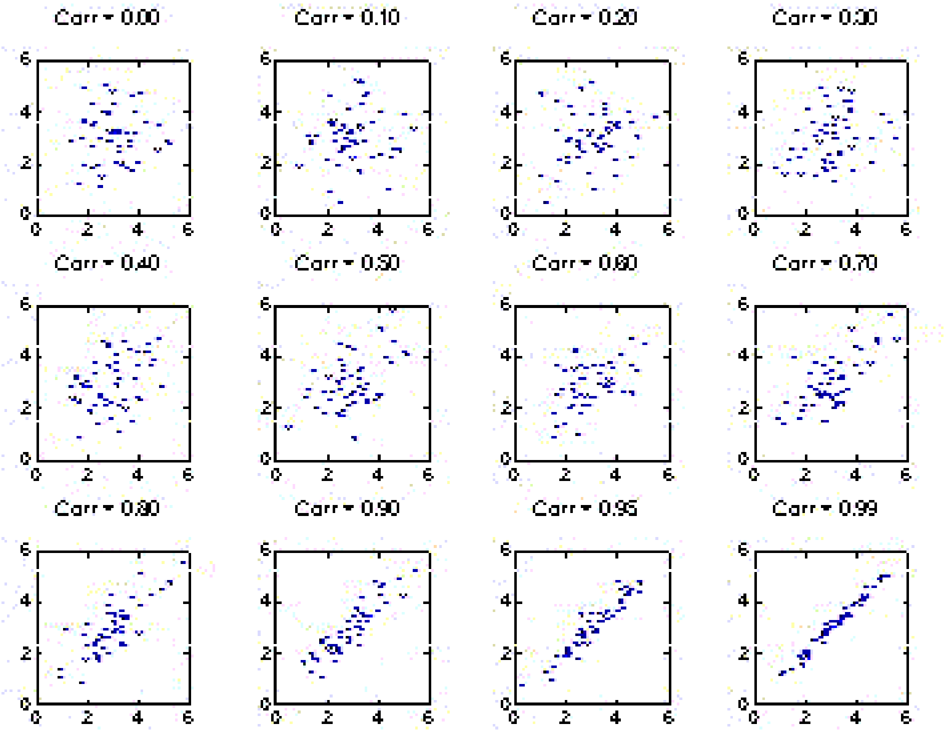
If the data lie predominantly in quadrants I and III, the correlation coefficient will be positive. If the data lie predominantly in quadrants II and IV the correlation coefficient will be negative.

The denominator will always be positive (unless all of the x's or all of the y's are equal) and is there only to force the correlation coefficient to be in the range [-1,1].

Properties of the correlation coefficient, r:

- -l <= r <= +1
- |r| = 1 if and only if the points lie exactly on a straight line.
- If the same constant is added to all of the Xs, the correlation coefficient is unchanged. Similarly for the Ys
- If all of the Xs are multiplied by a constant, the correlation coefficient is unchanged, except that the sign of the correlation coefficient is changed if the constant is negative. Similarly for the Ys.

The last two properties mean the correlation coefficient doesn't change as the result a linear transformation, aX+b, where 'a' and 'b' are constants, except for a change of sign if 'a' is negative. Hence, when investigating height and weight, the correlation coefficient will be the same whether height is measured in inches or centimeters and the weight is measured in pounds or kilograms.

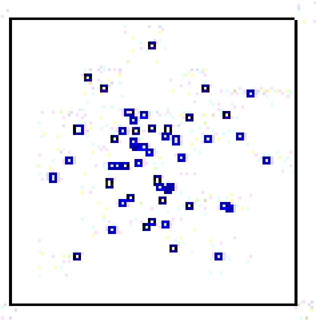


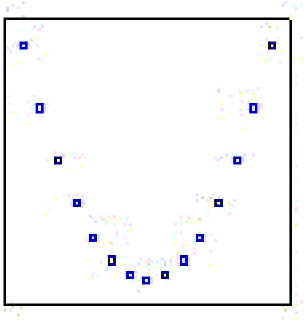
How do values of the correlation coefficient correspond to different data sets? As the correlation coefficient increases in magnitude, the points become more tightly concentrated about a straight line through the data. Two things should be noted. First, correlations even as high as 0.6 don't look that different from correlations of 0. I want to say that correlations of 0.6 and less don't mean much if the goal is to predict individual values of one variable from the other. The prediction error is nearly as great as we'd get by ignoring the second variable and saying that everyone had a value of the first variable equal to the overall mean! However, I'm afraid that this might be misinterpreted as suggesting that all such associations are worthless. They have important uses that we will discuss in detail when we consider linear regression. Second, although the correlation can't exceed 1 in magnitude, there is still a lot of variability left when the correlation is as high as 0.99.

[(American Statistician article) conducted an experiment in which people were asked to assign numbers between 0 and 1 to scatterplots showing varying degrees of association. They discovered that people perceived association not as proportional to the correlation coefficient, but as proportional to $1 - \sqrt{(1-r^2)}$.

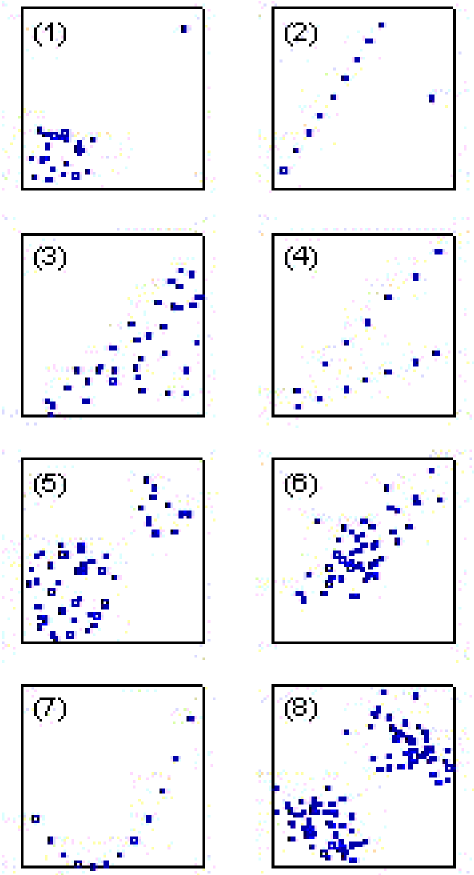
r	$1-\sqrt{(1-r^2)}$
0.5	0.13
0.7	0.29
0.8	0.40
0.9	0.56
0.99	0.86
0.999	0.96

Trouble!



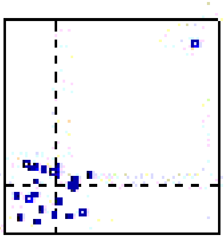


The pictures like those in the earlier displays are what one usually thinks of when a correlation coefficient is presented. But the correlation coefficient is a single number summary, a measure of linear association, and like all single number summaries, it can give misleading results if not used with supplementary information such as scatterplots. For example, data that are uniformly spread throughout a circle will have a correlation coefficient of 0, but so, too, will data that is symmetrically placed on the curve Y = X^2! The reason the correlation is zero is that high values of Y are associated with both high and low values of X. Thus, here is an example of a correlation of zero even where there is Y can be predicted perfectly from X!



To further illustrate the problems of attempting to interpret a correlation coefficient without looking at the corresponding scatterplot, consider this set of scatterplots, which duplicates most of the examples from pages 78-79 of *Graphical Methods for Data Analysis* by Chambers, Cleveland, Kleiner, and Tukey. **Each data set has a correlation coefficient of 0.7.**

What to do:



The correlation is 0 within the bulk of the data in the lower left-hand corner. The outlier in the upper right hand corner increases both means and makes the data lie predominantly in quadrants I and III. Check with the source of the data to see if the outlier might be in error. Errors like these often occur when a decimal point in both measurements is accidentally shifted to the right. Even if there is no explanation for the outlier, it should be set aside and the correlation coefficient or the remaining data should be calculated. The report must include a statement of the outlier's existence. It would be misleading to report the correlation based on all of the data because it wouldn't represent the behavior of the bulk of the data.

As discussed below, correlation coefficients are appropriate only when data are obtained by drawing a random sample from a larger population. However, sometimes correlation coeficients are mistakenly calculated when the values one of the variables--X, say--are determined or constrained in advance by the investigator. In such cases, the message or the outlier may be real, namely, that over the full range of values, the two variables tend to increase and decrease together. It's poor study design to have the answer determined by a single observation and it places the analyst in an uncomfortable position. It demands that we assume thr association is roughly linear over the entire range and that the variability in Y will be no different for large X from what it is for small X. Unfortunately, once the study is performed, there isn't much that can be done about it. The outcome hinges on a single obsrevation.

2. Similar to 1. Check the outlier to see if it is in error. If not, report the correlation coefficient for all points except the outlier along with the warning that the outlier occurred. Unlike case 1 where the outlier is an outlier in both dimensions, here the outlier has a reasonable Y value and only a slightly unreasonable X value. It often happens that observations are *two-dimensional outliers*. They are unremarkable when each response is viewed individually in its histogram and do not show any aberrant behavior until they are viewed in two dimensions. Also, unlike case 1 where the outlier increases the magnitude of correlation coefficient, here the magnitude is decreased.
3. This sort of picture results when one variable is a component of the other, as in the case of (total energy intake, energy from fat). The correlation coefficient almost always has to be positive since increasing the total will tend to increase each component. In such cases, correlation coefficients are probably the wrong summaries to be using. The underlying research question should be reviewed
4. The two nearly straight lines in the display may be the result of plotting the combined data from two identifiable groups. In might be as simple as one line corresponding to men, the other to women. It would be misleading to report the single correlation coefficient without comment, even if no explanation manifests itself.
5. The correlation is zero within the two groups; the overall correlation of 0.7 is due to the differences between groups. Report that there are two groups and that the within group correlation is zero. In cases where the separation between the groups is greater, the comments from case 1 apply as well. It may be that the data are not a simple random sample from a larger popultion and the division between the two groups may be due to a conscious decision to exclude values in the middle of the range of X or Y. The correlation coefficient is an inappropriate

summary of such data because its value is affected by the choice of X or Y values.

6. What most researcher think of when a correlation of 0.7 is reported.
7. A problem mentioned earlier. The correlation is not 1, yet the observations lie on a smooth curve. The correlation coefficient is 0.70 rather than 0 because here the curve is not symmetric. Higher values of Y tend to go with higher values of X. A correlation coefficient is an inappropriate numerical summary of this data. Either (i) derive an expression for the curve, (ii) transform the data so that the new variables have a linear relationship, or (iii) rethink the problem.
8. This is similar to case 5, but with a twist. Again, there are two groups, and the separation between them produces the postive overall correlation. But, here, the within-group correlation is negative! I would do my best to find out why there are two groups and report the within group correlations.

The moral of these displays is clear: **ALWAYS LOOK AT THE SCATTERPLOTS**!

The correlation coefficient is a numerical summary and, as such, it can be reported as a measure of association for any batch of numbers, no matter how they are obtained. Like any other statistic, its proper interpretation hinges on the sampling scheme used to generate the data.

The correlation coefficient is most appropriate when both measurements are made from a simple random sample from some population. The sample correlation then estimates a corresponding quantity in the population. It is then possible to compare sample correlation coefficients for samples from different populations to see if the association is different within the populations, as in comparing the association between calcium intake and bone density for white and black postmenopausal females.

If the data do not constitute a simple random sample from some population, it is not clear how to interpret the correlation coefficient. If, for example, we decide to measure bone density a certain number of women at each of many levels of calcium intake, the correlation coefficient will change depending on the choice of intake levels.

This distortion most commonly occurs in practice when the range of one of the variables has been restricted. How strong is the association between MCAT scores and medical school performance? Even if a simple random sample of medical students is chosen, the question is all but impossible to answer because applicants with low MCAT scores are less likely to be admitted to medical school. We can talk about the relationship between MCAT score and performance only within a narrow range of high MCAT scores.

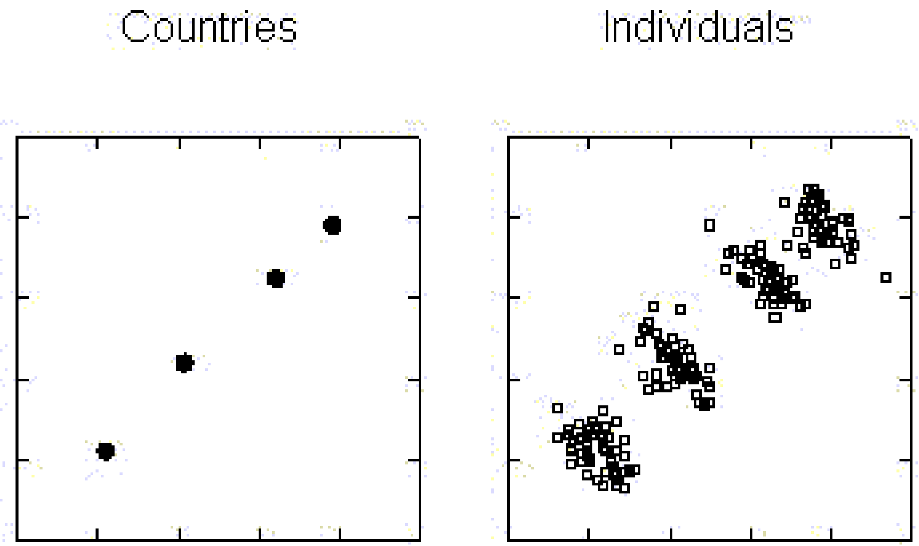
[One major New York university with a known admissions policy that prohibited penalizing an applicant for low SAT scores investigated the relationship between SAT scores and freshman year grade point average. The study was necessarily non-scientific because many students with low SAT scores realized that while the scores wouldn't hurt, they wouldn't help, either, and decided to forego the expense of having the scores reported. The relationship turned out to be non-linear. Students with very low SAT Verbal scores (350 or less) had low grade point averages. For them, grade point average increased with SAT score. Students with high SAT Verbal scores (700 and above) had high grade point averages. For them, too, grade point average increased with SAT score. But in the middle (SAT Verbal score between 350 and 700), there was almost no relationship between SAT Verbal score and grade point average.

```
  |                                                 *
  |                                                *
  |                                               *
G |                                             *
P |                                          *
A |                 *        *       *
  |          *
  |      *
  |    *
  |   *
   --------------------------------------------------
                       SAT Verbal
```

Suppose these students are representative of all college students. What if this study were performed at another college where, due to admissions policies, the students had SAT scores only within a restricted range?

- How would the results of that study differ from the results here?
- What would be the effect on the correlation coefficient?
- Could a valid comparison of the relationship between SAT scores and grade point average in the two schools be made by comparing correlation coefficients? If not, then how?]

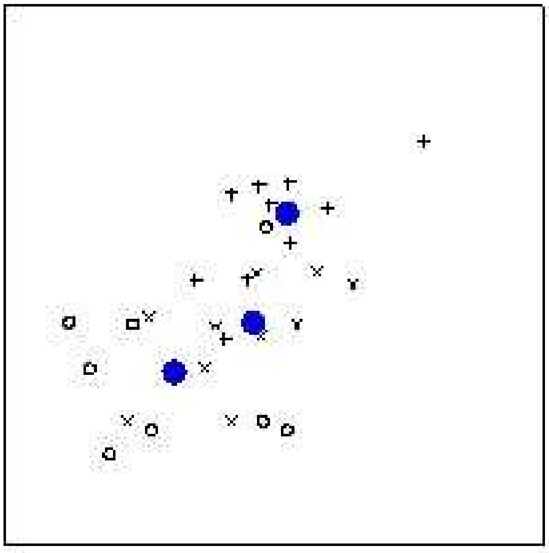
Ecological Fallacy



Another source of misleading correlation coefficients is *the ecological fallacy*. It occurs when correlations based on grouped data are incorrectly assumed to hold for individuals.

Imagine investigating the relationship between food consumption and cancer risk. One way to begin such an investigation would be to look at data on the country level and construct a plot of overall cancer risk against per capita daily caloric intake. The display shows cancer increasing with food consumption. But it is people, not countries, who get cancer. It could very well be that within countries those who eat more are less likely to develop cancer. On the country level, per capita food intake may just be an indicator of overall wealth and industrialization.

The ecological fallacy was in studying countries when one should have been studying people.



When the association is in the same direction for both individuals and groups, the ecological correlation, based on averages, will typically overstate the strength of the association in individuals. That's because the variablity within the groups will be eliminated. In the picture to the left, the correlation between the two variables is 0.572 for the set of 30 individual observations. The large blue dots represent the means of the crosses, plus signs, and circles. The correlation for the set of three dots is 0.902

Spurious Correlations

Correlation is not causation. The observed correlation between two variables might be due to the action of a third, unobserved variable. Yule (1926) gave an example of high positive correlation between yearly number of suicides and membership in the Church of England due not to cause and effect, but to other variables that also varied over time. (Can you suggest some?) Mosteller and Tukey (1977, p. 318) give an example of aiming errors made during bomber flights in Europe. Bombing accuracy had a high positive correlation with amount of fighter opposition, that is, the more enemy fighters sent up to distract and shoot down the bombers, the more accurate the bombing run! The reason being that lack of fighter opposition meant lots of cloud cover obscuring bombers from the fighters and the target from the bombers, hence, low accuracy.

[back to LHSP]

Last modified: undefined.

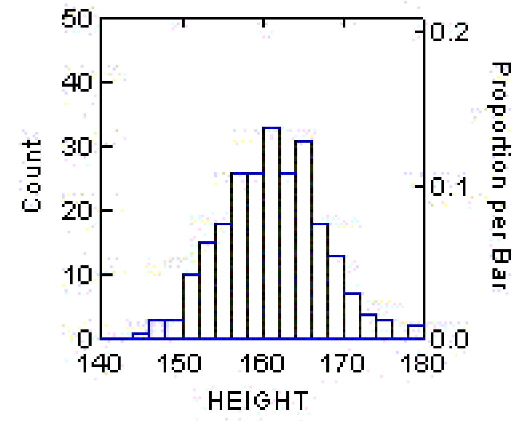
Probability Theory

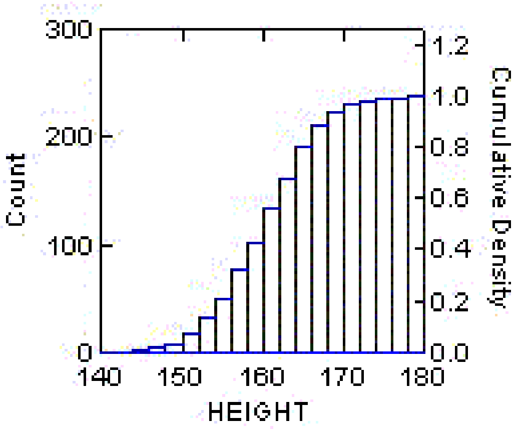
There's a lot that could be said about probability theory. Probability theory is what makes statistical methods work. Without probability theory, there would be no way to describe the way samples might differ from the populations from which they were drawn. While it is important for mathematical statisticians to understand all of the details, all that is necessary for most analysts is to insure that random sampling is involved in observational studies and randomization is involved in intervention trials. Beyond that, there are just four things the analyst needs to know about probability.

1. The probability of an event E, P(E), is the proportion of times the event occurs in a long series of experiments.
2. $0 \leq P(E) \leq 1$, where P(E) = 0 if E is an impossible event and P(E) = 1 if E is a sure thing.
3. If ~E is the opposite or complement of E, then P(~E) = 1 - P(E). Thus, the probability that (an individual has high blood pressure) is 1 - the probability that (an individual does not have high blood pressure).
4. The probability that something is true for an individual selected at random from a population is equal to the fraction of the population for whom it is true. For example, if 10% of a population is left-handed, the probability is 0.10 or 10% that an individual chosen at random will be left-handed.

Probability, Histograms, Distributions

We've seen histograms--bar charts in which the area of the bar is proportional to the number of observations having values in the range defining the bar. Just as we can construct histograms of samples, we can construct histograms of populations. The population histogram describes the proportion of the population that lies between various limits. It also describes the behavior of individual observations drawn at random from the population, that is, it gives the probability that an individual selected at random from the population will have a value between specified limits. It is critical that you understand that population histograms describe the way individual observations behave. You should not go on unless you do!





When we're talking about populations and probability, we don't use the words "population histogram". Instead, we refer to *probability densities* and *distribution functions*. (However, it will sometimes suit my purposes to refer to "population histograms" to remind you what a density is.) When the area of a histogram is standardized to 1, the histogram becomes a probability density function. The area of any portion of the histogram (the area under any part of the curve) is the proportion of the population in the designated region. It is also the probability that an individual selected at random will have a value in the designated region. For example, if 40% of a population have cholesterol values between 200 and 230 mg/dl, 40% of the area of the histogram will be between 200 and 230 mg/dl. The probability that a randomly selected individual will have a cholesterol level in the range 200 to 230 mg/dl is 0.40 or 40%.

Strictly speaking, the histogram is properly a *density*, which tells you the proportion that lies between specified values. A *(cumulative) distribution function* is something else. It is a curve whose value is the proportion with values less than or equal to the value on the horizontal axis, as the example to the left illustrates. Densities have the same name as their distribution functions. For example, a bell-shaped curve is a normal density. Observations that can be described by a normal density are said to follow a normal distribution.

If you understand that population histograms describe the way individual observations behave, you're well on your way to understanding what statistical methods are all about. One of the jobs of the mathematical statistician is to describe the behavior of things other than individual observations. If we can describe the behavior of an individual observation, then perhaps we can describe the behavior of a sample mean, or a sample proportion, or even the difference between two sample means. We can! Here is the one sentence condensation of an entire course in distribution theory: Starting with a distribution function that describes the behavior of individual observations, it is possible to use mathematics to find the distribution functions that describe the behavior of a wide variety of statistics, including, means, proportions, standard deviations, variances, percentiles, and regression coefficients.

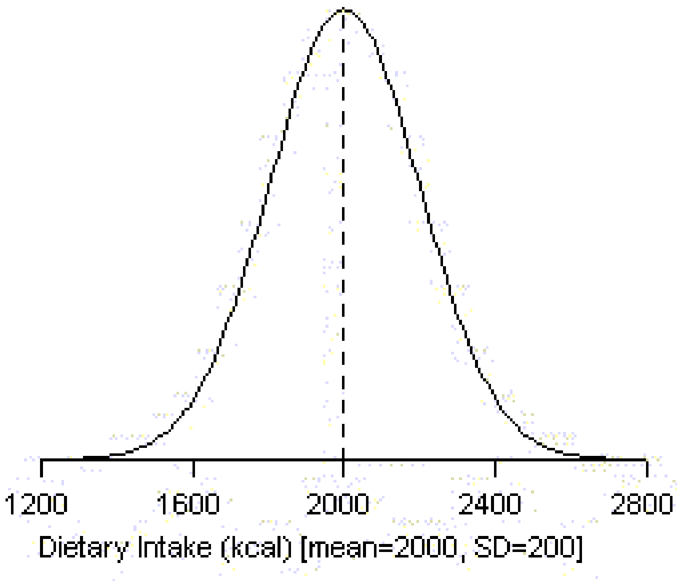
If you ever take a mathematical statistics course, you'll go through a large number of examples to learn how the mathematics works. You'll gain the skills to extend statistical theory to derive distributions for statistics that have not previously been studied. However, the basic idea will be the same. Given a distribution function that describes the behavior of individual observations, you'll derive distribution functions that describe the behavior of a wide variety of statistics, In these notes, we will accept the fact that this can be done and we will use the results obtained by others to describe the behavior of statistics that interest us. We will not bother to derive them ourselves.

This is the most important idea, after study design, that we've discussed so far--that distributions describe the behavior of things. They tell us how likely it is that the quantity being described will take on particular values. So far, we've talked about individual observations only. That is, all of the densities we've seen so far describe the behavior of individual observations, such as the individual heights displayed above.

We will soon be seeing distributions that describe the behavior of things such as sample means, sample proportions, and the difference between two sample means and two sample proportions. These distributions are all used the same way. For example, the distribution of the difference between two sample means describes what is likely to happen when two samples are drawn and the difference in their means is calculated. If you ever wanted to verify this, you could repeat the study over and over and construct a histogram of mean differences. You would find that it looks the same as the density function predicted by probability theory.

Last modified: undefined.

The Normal Distribution



When people first began constructing histograms, a particular shape occurred so often that people began to expect it. Hence, it was given the name *normal* distribution. The normal distribution is symmetric (you can fold it in half and the two halves will match) and unimodal (single peaked). It is what psychologists call the *bell-shaped curve*. Every statistics course spends lots of time discussing it. Text books have long chapters about it and many exercises involving it. The normal distribution is important. Statistics couldn't function without it!

Some statistical techniques demand that individual data items follow a normal distribution, that is, that the population histogram from which the sample is drawn have a normal shape. When the data are not normal, the results from these techniques will be unreliable. We've already seen cases where reporting a mean and SD can give a misleading mental picture and it would be better to replace or supplement them with percentiles. Not every distribution is normal. Some are far from it. We'll discuss the importance of normality on a technique-by-technique basis.

When normality of the individual observations is essential, transformations such as logarithms can sometimes be used to produce a set of transformed data that is better described by a normal distribution that the original data. Transformations aren't applied to achieve a specific outcome but rather to put the data in a form where the outcome can be relied upon. If you ran up the stairs to be on time for a doctor's appointment, you wouldn't mind waiting to have your blood pressure measured if a high reading might result in treatment for hypertension. Values obtained after running up a flight of stairs are unsuitable for detecting hypertension. The reason for waiting isn't to avoid a diagnosis of high blood pressure and necessary treatment, but to insure that a high reading can be believed. It's the same with transformations.

Comment: Professor Herman Rubin of Purdue University provides an additional insight into the statistician's fascination with the normal distribution (posted to the Usenet group sci.stat.edu, 3 Oct 1998 08:07:23 -0500; Message-ID: 6v57ib$s5q@b.stat.purdue.edu):

> Normality is rarely a tenable hypothesis. Its usefulness as a means of deriving procedures is that it is often the case, as in regression, that the resulting procedure is robust in the sense of having desirable properties without it, while nothing better can be done uniformly.

Comment: The probability that a normally distributed quantity will be within a specified multiple of standard deviations of its mean is the same for all normal distributions. For example, the probability of being within 1.96 SDs of the mean is 95%. [More often than not, people make casual use of 2 in place of the exact 1.96.] This is true whatever the mean and SD.

Comment: In the old days (B.C.: before computers) it was important to learn how to read tables of the normal distribution. Almost every analysis required the analyst to translate normal values into probabilities and probabilities into normal values. Now that computers do all the work, it is possible to be a good data analyst without ever having looked at a such a table.

OUTLIERS

Forty years later, it's still hard to improve on William Kruskal's February1960 Technometrics paper, "Some Remarks on Wild Observations". Since permission was granted to reproduce it in whole or in part, here it is in its entirety.

Some Remarks on Wild Observations [*]

William H. Kruskal[**]
The University of Chicago

[*] This work was sponsored by the Army, Navy and Air Force through the Joint Services Advisory Committee for Research Groups in Applied Mathematics and Statistics by Contract No. N6qri4)2035. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[**] With generous suggestions from LJ Savage, HV Roberts, KA Browalee, and F Mosteller.

> Editor's Note: At the 1959 meetings of the American Statistical Association held in Washington D.C., Messrs. F. J. Anscombe and C. Daniel presented papers on the detection and rejection of 'outliers', that is, observations thought to be maverick or unusual. These papers and their discussion will appear in the next issue of *Technometrics*. The following comments of Dr. Kruskal are another indication of the present interest of statisticians in this important problem.

The purpose of these remarks is to set down some non-technical thoughts on apparently wild or outlying observations. These thoughts are by no means novel, but do not seem to have been gathered in one convenient place.

1. Whatever use is or is not made of apparently wild observations in a statistical analysis, it is very important to say something about such observations in any but the most summary report. At least a statement of how many observations were excluded from the formal analysis, and why, should be given. It is much better to state their values and to do alternative analyses using all or some of them.

2. However, it is a dangerous oversimplification to discuss apparently wild observations in terms of inclusion in, or exclusion from, a more or less conventional formal analysis. An apparently wild (or otherwise anomalous) observation is a signal that says: "Here is something from which we may learn a lesson, perhaps of a kind not anticipated beforehand, and perhaps more important than the main object of the study." Examples of such serendipity have been frequently discussed--one of the most popular is Fleming's recognition of the virtue of penicillium.

3. Suppose that an apparently wild observation is *really known* to have come from an anomalous (and perhaps infrequent) causal pattern. Should we include or exclude it in our formal statistics? Should we perhaps change the structure of our formal statistics?

Much depends on what we are after and the nature of our material. For example, suppose that the observations are five determinations of the percent of chemical A in a mixture, and that one of the observations is badly out of line. A check of equipment shows that the out of line observation stemmed from an equipment miscalibration that was present only for the one observation.

If the magnitude of the miscalibration is known, we can probably correct for it; but suppose it is not known? If the goal of the experiment is only that of estimating the per cent of A in the mixture, it would be very natural simply to omit the wild observation. If the goal of the experiment is mainly, or even partly, that of investigating the *method* of measuring the per cent of A (say in anticipation of setting up a routine procedure to be based on one measurement per batch), then it may be very important to keep the wild observation in. Clearly, in this latter instance, the wild observation tells us something about the frequency and magnitude of serious errors in the method. The kind of lesson mentioned in 2 above often refers to methods of sampling, measurement, and data reduction, instead of to the underlying physical phenomenon.

The mode of formal analysis, with a known anomalous observation kept in, should often be different from a traditional means-and-standard deviations analysis, and it might well be divided into several parts. In the above very simple example, we might come out with at least two summaries: (1) the mean of the four good observations, perhaps with a plus-or-minus attached, as an estimate of the per cent of A in the particular batch of mixture at hand, and (2) a statement that serious calibration shifts are not unlikely and should be investigated further. In other situations, nonparametric methods might be useful. In still others, analyses that suppose the observations come from a mixture of two populations may be appropriate.

The sort of distinction mentioned above has arisen in connection with military equipment. Suppose that 50 bombs are dropped at a target, that a few go wildly astray, that the fins of these wild bombs are observed to have come loose in flight, and that their wildness is unquestionably the result of loose fins. If we are concerned with the accuracy of the whole bombing system, we certainly should not forget these wild bombs. But if our interest is in the accuracy of the bombsight, the wild bombs are irrelevant.

4. It may be useful to classify different degrees of knowledge about an apparently wild observation in the following way:

a. We may know, even *before* an observation, that it is likely to be wild, or at any rate that it will be the consequence of a variant causal pattern. For example, we may see the bomb's fins tear loose before it has fallen very far from the plane. Or we may know that a delicate measuring instrument has been jarred during its use.

b. We may be able to know, *after* an observation is observed to be apparently outlying, that it was the result of a variant causal pattern. For example, we may check a laboratory notebook and see that some procedure was poorly carried out, or we may ask the bombardier whether he remembers a particular bomb's wobbling badly in flight. The great danger here, of course, is that it is easy after the fact to bias one's memory or approach, knowing that the observation seemed wild. In complex measurement situations we may often find something a bit out of line for almost any observation.

c. There may be *no evidence* of a variant causal pattern aside from the observations themselves. This is perhaps the most difficult case, and the one that has given rise to various rules of thumb for rejecting observations.

Like most empirical classifications, this one is not perfectly sharp. Some cases, for example, may lie between b and c. Nevertheless, I feel that it is a useful trichotomy.

5. In case c above, I know of no satisfactory approaches. The classical approach is to create a test statistic, chosen so as to be sensitive to the kind of wildness envisaged, to generate its distribution under some sort of hypothesis of nonwildness, and then to 'reject' (or treat differently) an observation if the test statistic for it comes out improbably large under the hypothesis of nonwildness. A more detailed approach that has sometimes been used is to suppose that wildness is a consequence of some definite kind of statistical structure--usually a mixture of normal distributions--and to try to find a mode of analysis well articulated with this structure.

My own practice in this sort of situation is to carry out an analysis both with and without the suspect observations. If the broad conclusions of the two analyses are quite different, I should view any conclusions from the experiment with very great caution.

6. The following references form a selected brief list that can, I hope, lead the interested reader to most of the relevant literature.

References

1. CI Bliss, WO Cochran, and JW Tukey, "A rejection criterion based upon the range," *Biometrika*, 43 (1956), 41822.
2. WJ Dixon, "Analysis of extreme values," *Ann. Math. Stat.*, 21(1950), 488-506.
3. WJ Dixon, "Processing data for outliers," *Biometrics*, 9 (1953), 74-89.
4. Frank E. Grubbs, "Sample criteria for testing outlying observations," *Ann. Math. Stat.*, 21 (1950), 27-58.
5. EP King, "On some procedures for the rejection of suspected data," *Jour. Amer. Stat. Assoc.*, 48 (1953), 531-3.
6. Julius Lieblein, "Properties of certain statistics involving the closest pair in a sample of three observations," *Jour. of Research of the Nat. Bureau of Standards*, 48 (1952), 25548.
7. ES Pearson and C Chandra Sekar, "The efficiency of statistical tools and a criterion for the rejection of outlying observations," *Biometrika*, 28 (1936), 308-320.
8. Paul R. Rider, "Criteria for rejection of observations," *Washington University Studies, New Series.*

Science and Technology, 8 (1933).

Gerard E. Dallal
Last modified: undefined.

The Behavior of the Sample Mean
(or *Why Confidence Intervals Always Seem to be Based On the Normal Distribution*)

[Many of the figures in this note are screen shots from a simulation at the Rice Virtual Lab in Statistics. You might enjoy trying the simulation yourself after (or even while) reading this note. Java must be enabled in your browser for this simulation to run.]

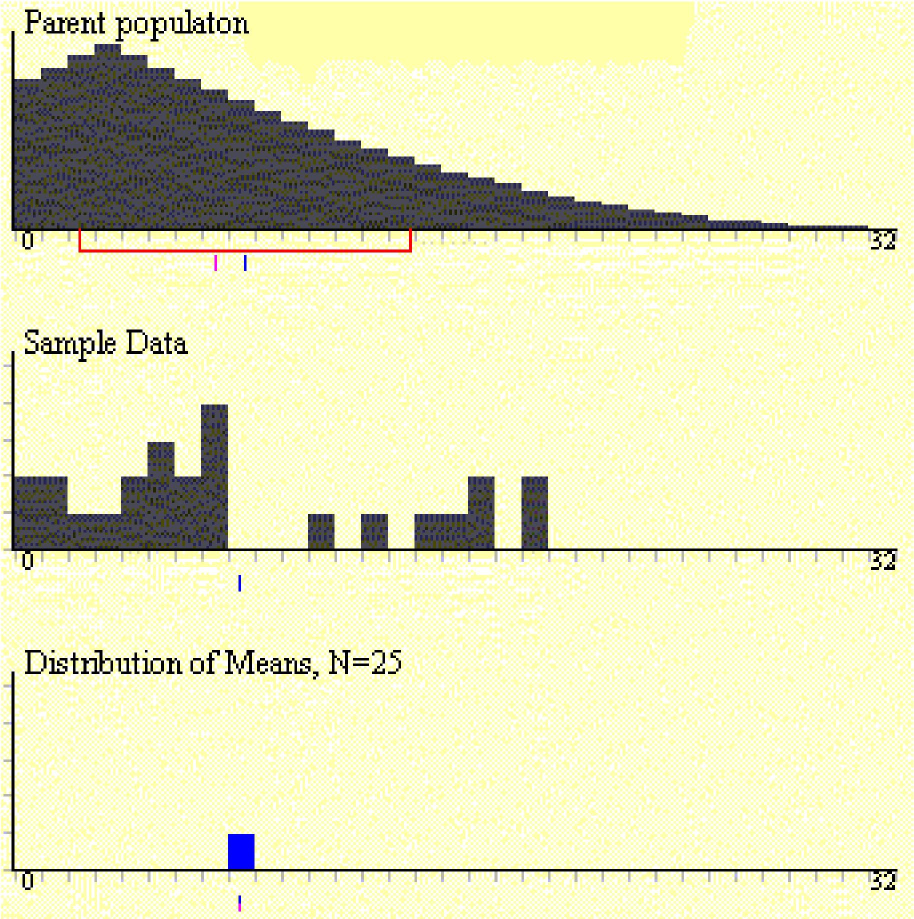
There is arguably no more important lesson to be learned in statistics than **how sample means behave**. **It explains *why* statistical methods work.** The vast majority of the things people do with statistics is compare populations, and most of the time populations are compared by comparing their means.

The way **individual observations behave** depends on the population from which they are drawn. If we draw a sample of individuals from a normally distributed population, the sample will follow a normal distribution. If we draw a sample of individuals from a population with a skewed distribution, the sample values will display the same skewness. **Whatever the population looks like--normal, skewed, bimodal, whatever--a sample of individual values will display the same characteristics.** This should be no surprise. Something would be very wrong if the sample of individual observations *didn't* share the characteristics of the parent population.

We are now going to see a truly wondrous result. Statisticians refer to it as ***The Central Limit Theorem***. It says that if you draw a large enough sample, the way the sample mean varies around the population mean can be described by **a normal distribution**, **NO MATTER WHAT THE POPULATION HISTOGRAM LOOKS LIKE!**

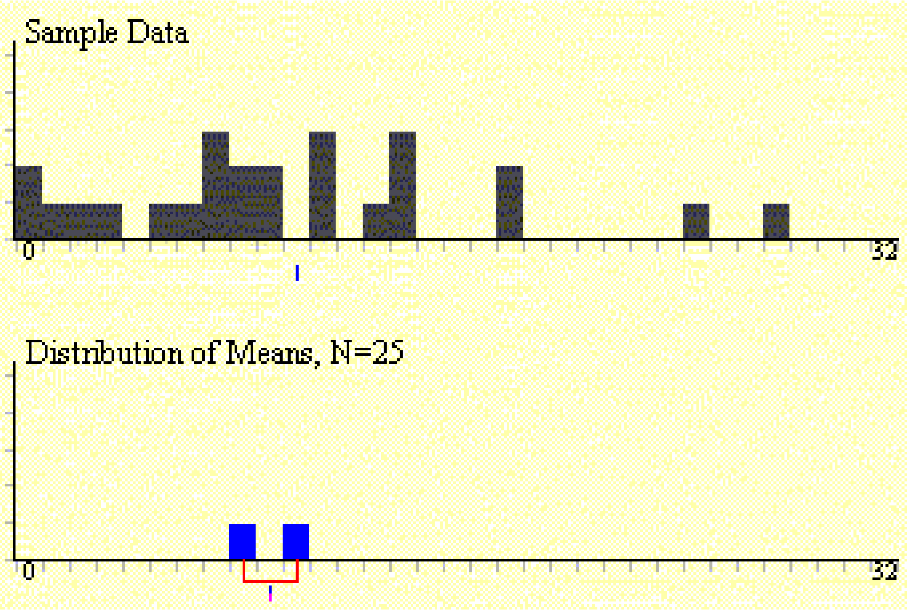
I'll repeat and summarize because this result is so important. **If you draw a large sample, the histogram of the individual observations will look like the population histogram from which the observations were drawn. However, the way the sample mean varies around the population mean can be described by the normal distribution.** This makes it very easy to describe the way population means behave. The way they vary about the population mean, for large samples, is unrelated to the shape of the population histogram.

Let's look at an example. In the picture to the left,

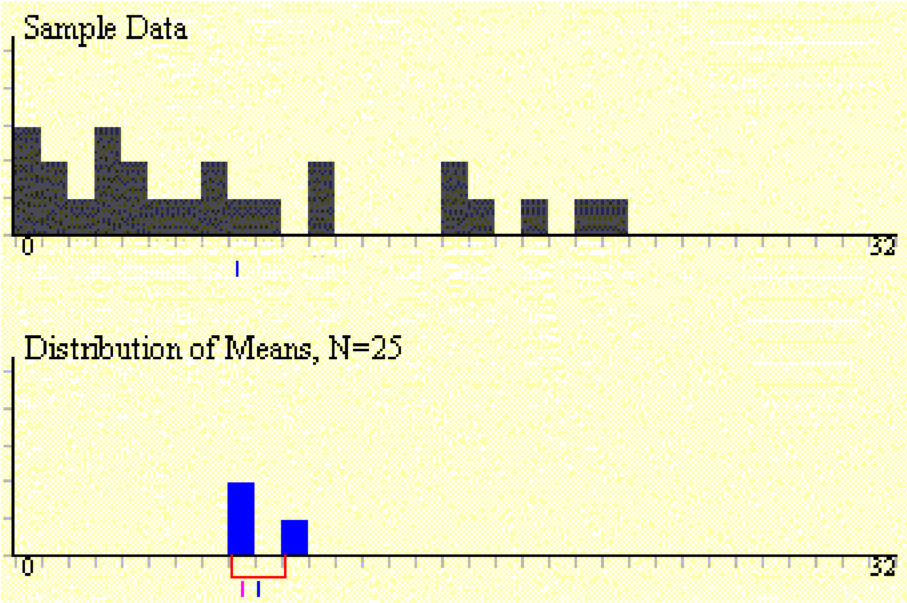


- the top panel shows a population skewed to the right
- the middle panel shows a sample of 25 observations drawn from that population
- the bottom panel shows the sample mean.

The 25 observations show the kind of skewness to be expected from a sample of 25 from this population.



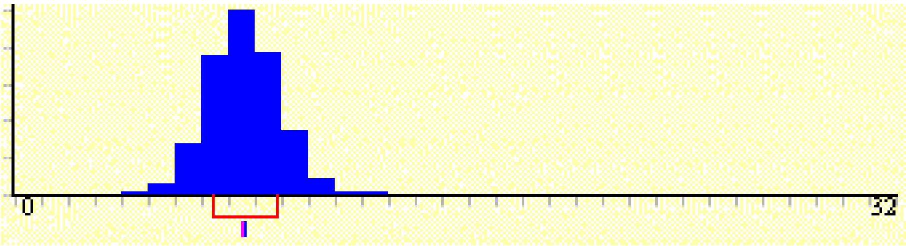
Let's do it again and keep collecting sample means.



And one more time.

In each case, the individual observations are spread out in a manner reminiscent of the population histogram. The sample means, however, are tightly grouped. This is not unexpected. In each sample, we get observations from throughout the distribution. The larger values keep the mean from being very small while the smaller values keep the mean from being very large. There are so many observations, some large, some small, that the mean ends up being "average". If the sample contained only a few observations, the sample mean might jump around considerably from sample to sample, but with lots of observations the sample mean doesn't get a chance to change very much.

Since the computer is doing all the work, let's go hog wild and do it **10,000 times**!



Here's how those means from the 10,000 samples of 25 observations each, behave. They behave like things drawn from a normal distribution centered about the mean of the original population!

At this point, the most common question is, "What's with the 10,000 means?" and it's a good question. Once this is sorted out, everything will fall into place.

- We do the experiment only once, that is, we get to see only one sample of 25 observations and one sample mean.
- The reason we draw the sample is to say something about the population mean.
- In order to use the sample mean to say something about the population mean, we have to know something about how different the two means can be.
- This simulation tells us. The sample mean varies around the population mean as though
  - it came from a normal distribution
  - whose standard deviation is estimated by the Standard Error of the Mean, SEM = s/ √n. (More about the SEM below.)
- All of the properties of the Normal Distribution apply:
  - 68% of the time, the sample mean and population mean will be within 1 SEM of each other.
  - 95% of the time, the sample mean and population mean will be within 2 SEMs of each other.

- 99% of the time, the sample mean and population mean will be within 2.57 SEMs of each other, and so on.

We will make formal use of this result in the note on Confidence Intervals.

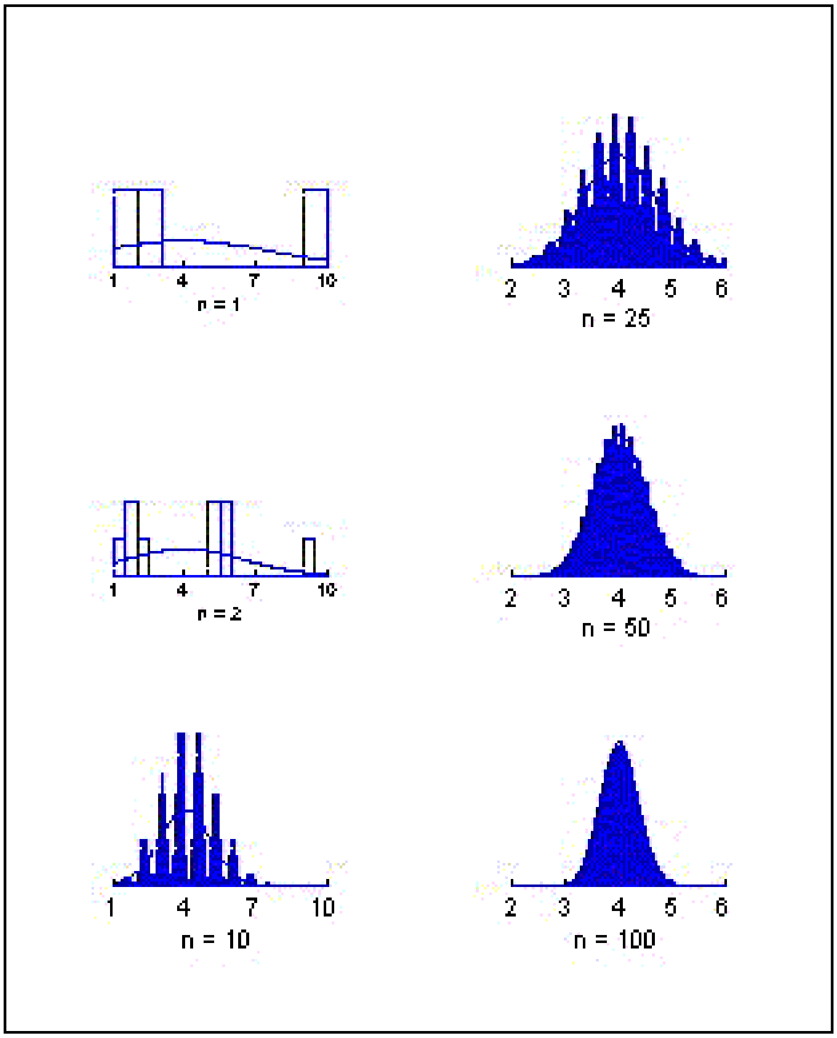
This result is *so* important that statisticians have given it a special name. It is called **The Central Limit Theorem**. It is a *limit theorem* because it describes the behavior of the sample mean in the limit as the sample size grows large. It is called the *Central* limit theorem not because there's any central limit, but because it's a limit theorem that is *central* to the practice of statistics!

The key to the Central limit Theorem is *large* sample size. The closer the histogram of the individual data values is to normal, the smaller *large* can be.

- If individual observations follow a normal distribution exactly, the behavior of sample means can be described by the normal distribution for *any* sample size, even 1.
- If the departure from normality is mild, *large* could be as few as 10. For biological units measured on a continuous scale (food intake, weight) it's hard to come up with a measurement for which a sample of 100 observations is not sufficient.
- One can always be perverse. If a variable is equal to 1 if "struck by lightning" and 0 otherwise, it might take many millions of observations before the normal distribution can be used to describe the behavior of the sample mean.

For variables like birth weight, caloric intake, cholesterol level, and crop yield measured on a continuous underlying scale, large is somewhere between 30 and 100. Having said this, it's only fair that I try to convince you that it's true.

The vast majority of the measurements we deal with are made on biological units on a continuous scale (cholesterol, birth weight, crop yield, vitamin intakes or levels, income). Most of the rest are indicators of some characteristic (0/1 for absence/presence of premature birth, disease). Very few individual measurements have population histograms that look less normal than one with three bars of equal height at 1,2, and 9, that is, a population that is one-third *1*s, one- third *2*s, and one-third *9*s. It's not symmetric. One-third of the population is markedly different from the other two-thirds. If the claim is true for this population, then perhaps it's true for population histograms closer to the normal distribution.



The distribution of the sample mean for various sample sizes is shown at the left. When the sample size is 1, the sample mean is just the individual observation. As the number of samples of a single observation increases, the histogram of sample means gets closer and closer to three bars of equal height at 1,2,9--the population histogram for individual values. The histogram of sample individual values always looks like the population histogram of individual values as you take more samples of individual values. It does NOT look more and more normal unless the population from which the data are drawn is normal.

When samples of size two are taken, the first observation is equally likely to be 1, 2 or 9, as is the second observation.

Obs 1	Obs 2	Mean
1	1	1.0
1	2	1.5
1	9	5.0
2	1	1.5
2	2	2.0
2	9	5.5
9	1	5.0
9	2	5.5
9	9	9.0

The sample mean can take on the values 1, 1.5, 2, 5, 5.5, and 9.

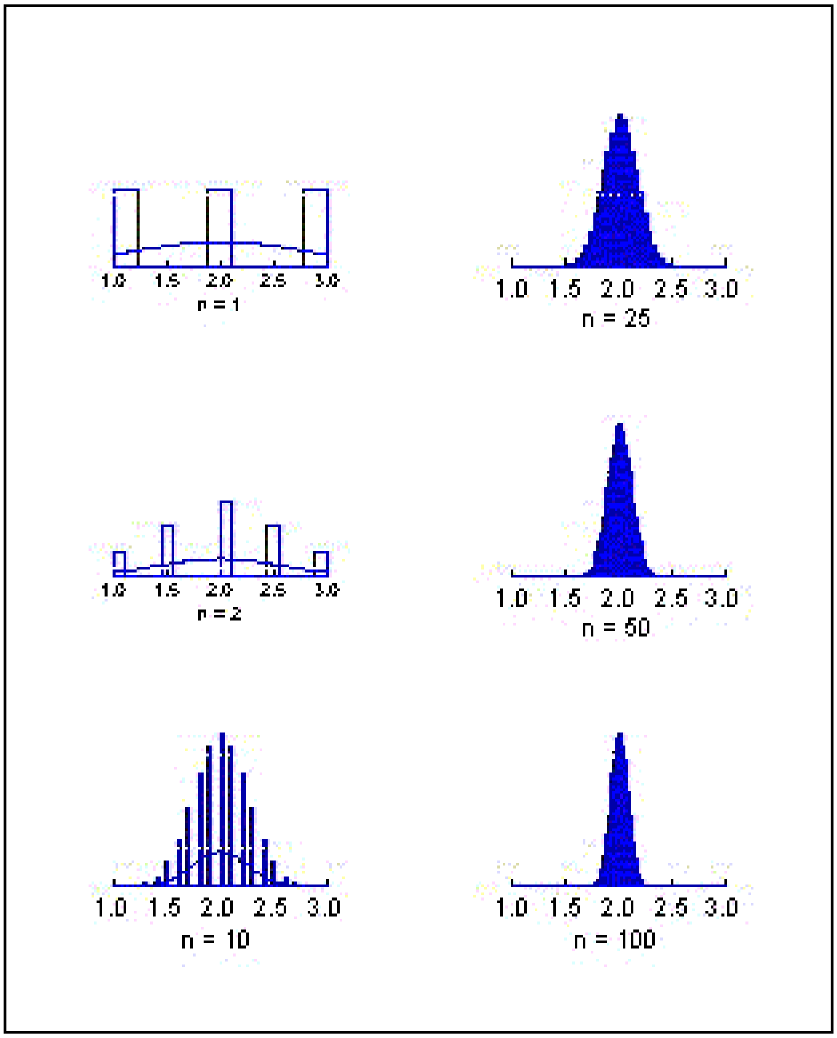
- There is only one way for the mean to be 1 (both observations are 1), but
- there are two ways to get a mean of 1.5 (the first can be 1 and the second 2, or the first can be 2 and the second 1).
- There is one way to get a mean of 2,
- two ways to get a mean of 5,
- two ways to get a mean of 5.5, and
- one way to get a mean of 9.

Therefore, when many samples of size 2 are taken and their means calculated, 1, 2, and 9 will each occur

1/9 of the time, while 1.5, 5, and 5.5 will each occur 2/9 of the time, as shown in the picture.

And so it goes for all sample sizes. Leave that to the mathematicians. The pictures are correct. Trust me. However, you are welcome to try to construct them for yourself, if you wish.

When n=10, the histogram of the sample means is very bumpy, but is becoming symmetric. When n=25, the histogram looks like a stegosaurus, but the bumpiness is starting to smooth out. When n=50, the bumpiness is reduced and the normal distribution is a good description of the behavior of the sample mean. The behavior (distribution) of the mean of samples of 100 individual values is nearly indistinguishable from the normal distribution to the resolution of the display. If the mean of 100 observations from this population of 1s, 2s, and 9s can be described by a normal distribution, then perhaps the mean of our data can be described by a normal distribution, too.



When the distribution of the individual observations is symmetric, the convergence to normal is even faster. In the diagrams to the left, one-third of the individual observations are 1s, one-third are 2s, and one-third are 3s. The normal approximation is quite good, even for samples as small as 10. In fact, even n=2 isn't too bad!

To summarize once again, **the behavior of sample means of large samples can be described by a normal distribution even when individual observations are not normally distributed**.

This is about as far as we can go without introducing some notation to maintain rigor. Otherwise, we'll sink into a sea of confusion over samples and populations or between the standard deviation and the (about-to-be-defined) standard error of the mean.

Sample		Population
$\bar{x}$	mean	μ
s	standard deviation	σ
n	sample size	

The **sample** has mean $\bar{x}$ and standard deviation *s*. The sample comes from a population of individual values with mean μ and standard deviation σ.

The behavior of sample means of large samples can be described by a normal distribution, but which normal distribution? If you took a course in distribution theory, you could prove the following results: The mean of the normal distribution that describes the behavior of a sample mean is equal to μ, the mean of the distribution of the individual observations. For example, if individual daily caloric intakes have a population mean μ = 1800 kcal, then the mean of 50 of them, say, is described by a normal distribution with a mean also equal to 1800 kcal.

The standard deviation of the normal distribution that describes the behavior of the sample mean is equal to the standard deviation of the individual observations divided by the square root of the sample size, that is, $\sigma/\sqrt{n}$. Our estimate of this quantity, $s/\sqrt{n}$, is called the *Standard Error of the Mean* (SEM), that is,

$$\text{SEM} = s/\sqrt{n}.$$

I don't have a nonmathematical answer for the presence of the square root. Intuition says the mean should vary less from sample-to-sample as the sample sizes grow larger. This is reflected in the SEM, which decreases as the sample size increases, but it drops like the *square root* of the sample size, rather than the sample size itself.

To recap...

1. There are probability distributions. They do two things.
   - They describe the population, that is, they say what proportion of the population can be found between any specified limits.
   - They describe the behavior of individual members of the population, that is, they give the probability that an individual selected at random from the population will lie between any specified limits.
2. When single observations are being described, the "population" is obvious. It is the population of individuals from which the sample is drawn. When probability distributions are used to describe statistics such as sample means, there is a population, too. It is the (hypothetical) collection of

values of the statistic should the experiment or sampling procedure be repeated over and over.

3. (**Most important and often ignored!**) The common statistical procedures we will be discussing are based on the probabilistic behavior of statistical measures. They are *guaranteed* to work as advertised, but only if the data arise from a probability based sampling scheme or from randomizing subjects to treatments. If the data do not result from random sampling or randomization, there is no way to judge the reliability of statistical procedures based on random sampling or randomization.

The Sample Mean As an Estimate of The Population Mean

These results say that for large sample sizes the behavior of sample means can be described by a normal distribution whose mean is equal to the population mean of the individual values, μ, and whose standard deviation is equal to $\sigma / \sqrt{n}$, which is estimated by the SEM. In a course in probability theory, we use this result to make statements about the a yet-to-be-obtained sample mean when the population mean is known. In statistics, we use this result to make statements about an unknown population mean when the sample mean is known.

Preview: Let's suppose we are talking about 100 dietary intakes and the SEM is 40 kcal. The results of this note say the behavior of the sample mean can be described by a normal distribution whose SD is 40 kcal. We know that when things follow a normal distribution, they will be within 2 SDs of the population mean 95% of the time. In this case, 2 SDs is 80 kcal. Thus, the sample mean and population mean will be within 80 kcal of each other 95% of the time.

- If we were told the population mean were 2000 kcal and were asked to predict the sample mean, we would say there's a 95% chance that our sample mean would be in the range (1920[=2000-80], 2080[-2000+80]) kcal.
- It works the other way, too. If the population mean is unknown, but the sample mean is 1980 kcal, we would say we were 95% confident that the population mean was in the range (1900 [=1980-80], 2060[=1980+80]) kcal.

Note: The use of the word *confident* in the previous sentence was not accidental. *Confident* and *confidence* are the technical words used to describe this type of estimation activity. Further discussion occurs in the notes on Confidence Intervals

The decrease of SEM with sample size reflects the common sense idea that the more data you have, the better you can estimate something. Since the SEM goes down like the square root of the sample size, the bad news is that to cut the uncertainty in half, the sample size would have to quadrupled. The good news is that if you can gather only half of the planned data, the uncertainty is only 40% larger than what it would have been with all of the data, not twice as large.

Potential source of confusion: How can the SEM be an SD? Probability distributions have means and standard deviations. This is true of the probability distribution that describes individual observations and the probability distribution that describes the behavior of sample means drawn from that population Both of these distributions have the same mean, denoted μ here. If the standard deviation of the distribution that describes the individual observations is σ, then the standard deviation of the distribution that describes the sample mean is $\sigma/\sqrt{n}$, which is estimated by the SEM.

When you write your manuscripts, you'll talk about the SD of individual observations and the SEM as a measure of uncertainty of the sample mean as an estimate of the population mean. You'll never see anyone describing the SEM as estimating the SD of the sample mean. However, we have to be aware of this role for the SEM if we are to be able to understand and discuss statistical methods clearly.

[back to LHSP]

Last modified: undefined.

Coming Attractions: Where Are We Going?

Our goal is to get to the point were we can read, understand, and write statements like

- A 95% confidence interval for the mean yield of corn from farms using integrated pest management is 142.7 to 153.9 bushels per acre. **OR** Farms that practice integrated pest management while growing corn have a mean yield of 142.7 to 153.9 bushels per acre (95% confidence interval).
- We are 95% confident that mean caloric intake of infants of low-income mothers receiving WIC assistance is 80 to 200 kcal per day greater than that of infants of low-income mothers who do not receive assistance. **OR** Infants of low-income mothers receiving WIC assistance have a greater mean daily caloric intake than infants of low-income mothers not receiving assistance (95%CI: 80 to 200 kcal).
- We are 95% confident that the mean total cholesterol level resulting from a canola oil diet is between 9.3 mg/dl less and 17.2 mg/dl more than the mean cholesterol level resulting from a rice oil diet. **OR** Our study was unable to distinguish between rice and canola oil. Based on our data, the effect of canola oil could do anything from reducing the mean cholesterol level 9.3 mg/dl to increasing it 17.2 mg/dl relative to a rice oil diet.

Confidence Intervals
Part I

Does the mean vitamin C blood level of smokers differ from that of nonsmokers? Let's suppose for a moment they do, with smokers tending to have lower levels. Nevertheless, we wouldn't expect every smoker to have levels lower than those of every nonsmoker. There would be some overlap in the two distributions. This is one reason why questions like this are usually answered in terms of population means, namely, how the mean level of all smokers compares to that of all nonsmokers.

The statistical tool used to answer such questions is the *confidence interval* (CI) for the difference between the two population means. But let's forget the formal study of statistics for the moment. What might you do to answer the question if you were on your own? You might get a random sample of smokers and nonsmokers, measure their vitamin C levels, and see how they compare. Suppose we've done it. In a sample of 40 Boston male smokers, vitamin C levels had a mean of 0.60 mg/dl and an SD of 0.32 mg/dl while in a sample of 40 Boston male nonsmokers (Strictly speaking, we can only talk about Boston area males rather than all smokers and nonsmokers. No one ever said research was easy.), the levels had a mean of 0.90 mg/dl and an SD of 0.35 mg/dl. The difference in means between nonsmokers and smokers is 0.30 mg/dl!

The difference of 0.30 looks impressive compared to means of 0.60 and 0.90, but we know that if we were to take another random sample, the difference wouldn't be exactly the same. It might be greater, it might be less. What kind of population difference is consistent with this observed value of 0.30 mg/dl?

How much larger or smaller might the difference in population means be if we could measure all smokers and nonsmokers? In particular, is 0.30 mg/dl the sort of sample difference that might be observed if there were **no** difference in the population mean vitamin C levels? We estimate the difference in mean vitamin C levels at 0.30 mg/dl, but 0.30 mg/dl "give-or-take what"? This is where statistical theory comes in.

One way to answer these questions is by reporting a 95% *confidence interval*. A 95% confidence interval is an interval generated by a process that's right 95% of the time. Similarly, a 90% confidence interval is an interval generated by a process that's right 90% of the time and a 99% confidence interval is an interval generated by a process that's right 99% of the time. If we were to replicate our study many times, each time reporting a 95% confidence interval, then 95% of the intervals would contain the population mean difference. In practice, we perform our study only once. We have no way of knowing whether our particular interval is correct, but we behave as though it is. Here, the 95% confidence interval for the difference in mean vitamin C levels between nonsmokers and smokers is 0.15 to 0.45 mg/dl. Thus, not only do we estimate the difference to be 0.30 mg/dl, but we are 95% confident it is no less than 0.15 mg/dl or greater than 0.45 mg/dl.

In theory, we can construct intervals of any level of confidence from 0 to 100%. There is a tradeoff between the amount of confidence we have in an interval and its length. A 95% confidence interval for a population mean difference is constructed by taking the sample mean difference and adding and subtracting 1.96 standard errors of the mean difference. A 90% CI adds and subtracts 1.645 standard errors of the mean difference, while a 99% CI adds and subtracts 2.57 standard errors of the mean difference. The shorter the confidence interval, the less likely it is to contain the quantity being estimated. The longer the interval, the more likely to contain the quantity being estimated. Ninety-five percent has been found to be a convenient level for conducting scientific research, so it is used almost universally. Intervals of lesser confidence would lead to too many misstatements. Greater confidence would require more data to generate intervals of usable lengths.

Part II

[Zero is a special value. If a difference between two means is 0, then the two means are equal!]

Confidence intervals contain population values found to be consistent with the data. If a confidence interval for a mean difference *includes* 0, the data are consistent with a population mean difference of 0. If the difference is 0, the population means are equal. If the confidence interval for a difference *excludes* 0, the data are not consistent with equal population means. Therefore, one of the first things to look at is whether a confidence interval for a difference contains 0. If 0 is not in the interval, a difference has been established. If a CI contains 0, then a difference has not been established. When we start talking about significance tests, we'll refer to differences that exclude 0 as a possibility as *statistically significant*. For the moment, we'll use the term sparingly.

A statistically significant difference may or may not be of practical importance. ***Statistical significance* and *practical importance* are separate concepts.** Some authors confuse the issues by taking about *statistical significance* and *practical significance* or by talking about, simply, *significance*. In these notes, there will be no mixing and matching. It's either *statistically significant* or *practically important* any other combination should be consciously avoided.

Serum cholesterol values (mg/dl) in a free-living population tend to be between the mid 100s and the high 200s. It is recommended that individuals have serum cholesterols of 200 or less. A change of 1 or 2 mg/dl is of no importance. Changes of 10-20 mg/dl and more can be expected to have a clinical impact on the individual subject. Consider an investigation to compare mean serum cholesterol levels produced by two diets by looking at confidence intervals for $\mu_1 - \mu_2$ based on $\bar{x}_1 - \bar{x}_2$. High cholesterol levels are bad. If $\bar{x}_1 - \bar{x}_2$ is positive, the mean from diet 1 is greater than the mean from diet 2, and diet 2 is favored. If $\bar{x}_1 - \bar{x}_2$ is negative, the mean from diet 1 is less than the mean from diet 2, and diet 1 is favored. Here are six possible outcomes of experiment.

	$\bar{x}_1 - \bar{x}_2$	95% CI for $\mu_1 - \mu_2$
	(what was observed)	(what the truth might be)
Case 1	2	(1,3)
Case 2	30	(20,40)
Case 3	30	(2,58)
Case 4	1	(-1,3)
Case 5	2	(-58,62)
Case 6	30	(-2,62)

For each case, let's consider, first, whether a difference between population means has been demonstrated and then what the clinical implications might be.

In cases 1-3, the data are judged inconsistent with a population mean difference of 0. In cases 4-6, the data are consistent with a population mean difference of 0.

- Case 1: There is a difference between the diets, but it is of no practical importance.
- Case 2: The difference is of practical importance even though the confidence interval is 20 mg/dl wide.
- Case 3: The difference may or may not be of practical importance. The interval is too wide to say for sure. More study is needed.
- Case 4: We cannot claim to have demonstrated a difference. We are confident that if there is a real difference it is of no practical importance.

- Cases 5 and 6: We cannot claim to have demonstrated a difference. The population mean difference is not well enough determined to rule out all cases of practical importance.

Cases 5 and 6 require careful handling. While neither interval formally demonstrates a difference between diets, case 6 is certainly more suggestive of something than Case 5. Both cases are consistent with differences of practical importance and differences of no importance at all. However, Case 6, unlike Case 5, seems to rule out any advantage of practical importance for Diet 1, so it might be argued that Case 6 is like Case 3 in that both are consistent with important and unimportant advantages to Diet 2 while neither suggests any advantage to Diet 1.

It is common to find reports stating that there was no difference between two treatment. As Douglas Altman and Martin Bland emphasize, absence of evidence is not evidence of absence, that is, failure to show a difference is not the same thing as showing two treatments are the same. Only Case 4 allows the investigators to say there is no difference between the diets. The observed difference is not statistically significant and, if it should turn out there really is a difference (no two population means are *exactly* equal to an infinite number of decimal places), it would not be of any practical importance.

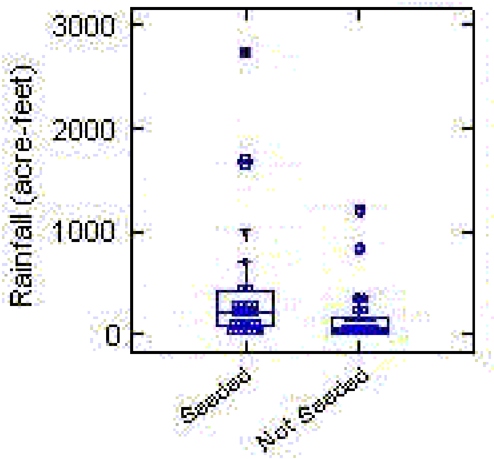
Many writers make the mistake of interpreting cases 5 and 6 to say there is no difference between the treatments or that the treatments are the same. This is an error. It is not supported by the data. All we can say in cases 5 and 6 is that we have been unable to demonstrate a difference between the diets. We cannot say they are the same. The data say they *may* be the same, but they *may* be quite different. Studies like this--that cannot distinguish between situations that have very different implications--are said to be underpowered, that is, they lack the power to answer the question definitively one way or the other.

In some situations, it's important to know if there is an effect no matter how small, but in most cases it's hard to rationalize saying whether or not a confidence interval contains 0 without reporting the CI, and saying something about the magnitude of the values it contains and their practical importance. If a CI does not include 0, are all of the values in the interval of practical importance? If the CI includes 0, have effects of practical importance been ruled out? If the CI includes 0 AND values of practical importance, YOU HAVEN'T LEARNED ANYTHING!

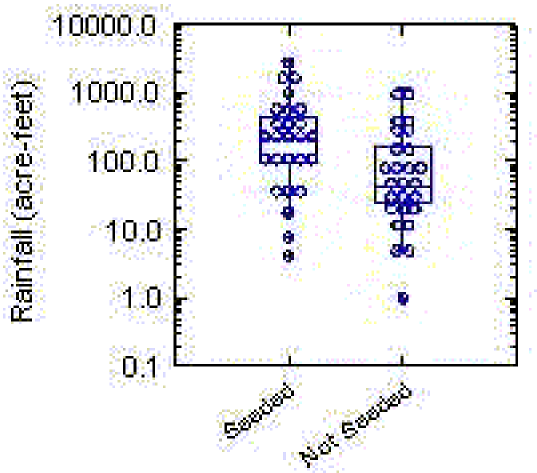
[back to LHSP]

Last modified: undefined.

Confidence Intervals Involving Data to Which a Logarithmic Transformation Has Been Applied



These data were originally presented in Simpson J, Olsen A, and Eden J (1975), "A Bayesian Analysis of a Multiplicative Treatment effect in Weather Modification," Technometrics, 17, 161-166, and subsequently reported and analyzed by Ramsey FL and Schafer DW (1997), *The Statistical Sleuth: A Course in Methods of Data Analysis*. Belmont, CA: Duxbury Press. They involve an experiment performed in southern Florida between 1968 and 1972. An aircraft was flown through a series of cloud and, at random, seeded some of them with massive amounts of silver iodide. Precipitation after the aircraft passed through was measured in acre-feet.

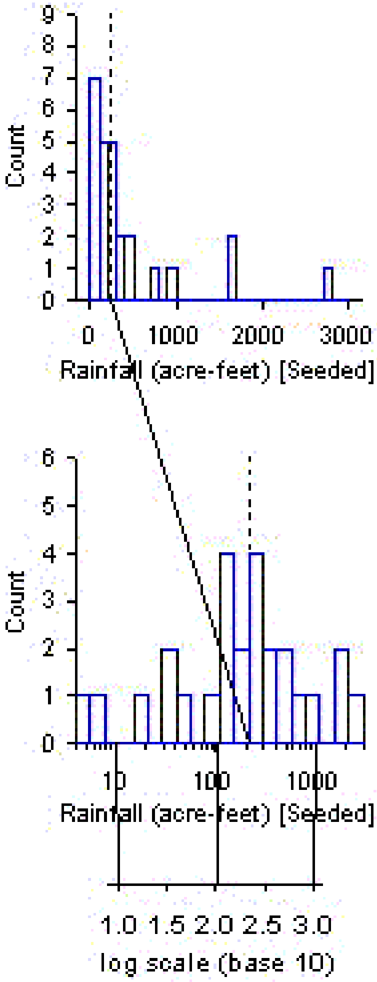


The distribution of precipitation within group (seeded or not) is positively skewed (long-tailed to the right). The group with the higher mean has a proportionally larger standard deviation as well. Both characteristics suggest that a logarithmic transformation be used to make the data more symmetric and homoscedastic (more equal spread). The second pair of box plots bears this out. This transformation will tend to make CIs more reliable, that is, the level of confidence is more likely to be what is claimed.

		N	Mean	Std. Deviation	Median
Rainfall	**Not Seeded**	26	164.6	278.4	44.2
	Seeded	26	442.0	650.8	221.6

		N	Mean	Std. Deviation	Geometric Mean
LOG_RAIN	**Not Seeded**	26	1.7330	.7130	54.08
	Seeded	26	2.2297	.6947	169.71

	95% Confidence Interval for the Mean Difference Seeded - Not Seeded (logged data)	
	Lower	**Upper**
Equal variances assumed	0.1046	0.8889
Equal variances not assumed	0.1046	0.8889



Researchers often transform data back to the original scale when a logarithmic transformation is applied to a set of data. Tables might include *Geometric Means*, which are the anti-logs of the mean of the logged data. When data are positively skewed, the geometric mean is invariably less than the arithmetic mean. This leads to questions of whether the geometric mean has any interpretation other than as the anti-log of the mean of the log transformed data.

The geometric mean is often a good estimate of the original median. The logarithmic transformation is monotonic, that is, data are ordered the same way in the log scale as in the original scale. If a is greater than b, then log (a) is greater than log(b). Since the observations are ordered the same way in both the original and log scales, the observation in the middle in the original scale is also the observation in the middle in the log scale, that is,

the log of the median = the median of the logs

If the log transformation makes the population symmetric, then the population mean and median are the same in the log scale. Whatever estimates the mean also estimates the median, and vice-versa. The mean of the logs estimates both the population mean and median in the log transformed scale. If the mean of the logs estimates the median of the logs, its anti-log--the geometric mean--estimates the median in the original scale!

The median rainfall for the seeded clouds is 221.6 acre-feet. In the picture, the solid line between the two histograms connects the median in the original scale to the mean in the log-transformed scale.

One property of the logarithm is that "the difference between logs is the log of the ratio", that is, log(x)-log(y)=log(x/y). The confidence interval from the logged data estimates the difference between the population means of log transformed data, that is, it estimates the difference between the logs of the geometric means. However, the difference between the logs of the geometric means is the log of the ratio of the geometric means. The anti-logarithms of the end points of this confidence interval give a confidence interval for the ratio of geometric means itself. Since the geometric mean is sometime an estimate of the median in the original scale, it follows that a confidence interval for the geometric means is approximately a confidence interval for the ratio of the medians in the original scale.

In the (common) log scale, the mean difference between seeded and unseeded clouds is 0.4967. Our best estimate of the ratio of the median rainfall of seeded clouds to that of unseeded clouds is $10^{0.4967}$ [= 3.14]. Our best estimate of the effect of cloud seeding is that it produces 3.14 times as much rain on average as not seeding.

Even when the calculations are done properly, the conclusion is often misstated.

> The difference 0.4967 does **not**mean seeded clouds produce 0.4967 acre-feet more rain that unseeded clouts. It is also improper to say that seeded clouds produce 0.4967 log-acre-feet more than unseeded clouds.
>
> The 3.14 means 3.14 times as much. It does **not**mean 3.14 times more (which would be 4.14 times as much). It does **not**mean 3.14 acre-feet more. It is a ratio and has to be described that way.

The a 95% CI for the population mean difference (Seeded - Not Seeded) is (0.1046, 0.8889). For reporting purposes, this CI should be transformed back to the original scale. A CI for a **difference** in the log scale becomes a CI for a **ratio** in the original scale.

The antilogarithms of the endpoints of the confidence interval are $10^{0.1046}$ = 1.27, and $10^{0.8889}$ = 7.74. Thus, the report would read: "The geometric mean of the amount of rain produced by a seeded cloud is 3.14 times as much as that produced by an unseeded cloud (95% CI: 1.27 to 7.74 times as much)." If the logged data have a roughly symmetric distribution, you might go so far as to say,"The median amount of rain...is approximately..."

Comment: The logarithm is the only transformation that produces results that can be cleanly expressed in terms of the original data. Other transformations, such as the square root, are sometimes used, but it is difficult to restate their results in terms of the original data.

Last modified: undefined.

LARGE SAMPLE Formulas for Confidence Intervals Involving Population Means

All of these 95% confidence intervals will be of the form *point estimate* plus and minus *1.96* times *the appropriate measure of uncertainty for the point estimate*.

A 95% confidence interval for a single population mean is

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} \quad \text{or} \quad \bar{x} \pm 1.96 \sqrt{\frac{s^2}{n}} \quad \text{or} \quad \bar{x} \pm 1.96\, SEM\ .$$

A 95% confidence interval for the difference between two population means, $\mu_x - \mu_y$, is

$$(\bar{x} - \bar{y}) \pm 1.96 \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

When **population** standard deviations are equal, a 95% confidence interval for the difference between two population means is

$$(\bar{x} - \bar{y}) \pm 1.96 \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} \quad \text{or} \quad (\bar{x} - \bar{y}) \pm 1.96 \sqrt{s_p^2 \left(\frac{1}{n_x} + \frac{1}{n_y}\right)} \quad \text{or} \quad (\bar{x} - \bar{y}) \pm 1.96 s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}\ ,$$

where s_p is the pooled sample standard deviation, so called because it combines or pools the information from both samples to estimate their common population variance

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_x^2}{n_x + n_y - 2} \quad \text{or} \quad s_p^2 = \frac{\sum_{i=1}^{n_x} (x_i - \bar{x})^2 + \sum_{i=1}^{n_y} (y_i - \bar{y})^2}{n_x + n_y - 2}\ .$$

Both expressions are informative. The first shows that $s_p^2$is a weighted combination of the individual sample variances, with weights equal to one less than the sample sizes. The second shows that it is calculated by summing up the squared deviations from each sample and dividing by 2 less than the combined sample size. It's worth noting that

$$n_x + n_y - 2 = (n_x - 1) + (n_y - 1).$$

The right hand side is the sum of the denominators that are used when calculating the individual SDs.

In general, when there are many ways to answer a question, the approach that makes assumptions is better in some sense when the assumptions are met. The 95% CIs that assume equal population variances will have true coverage closer to 95% for smaller sample sizes if the population variances are, in fact, equal. The downside is that the population variances have to be equal (or not so different that it matters).

Many argue that the interval that makes no assumptions should be used routinely for large samples because it will be approximately correct whether or not the assumptions are met. However, methods (yet to be seen) that adjust for the effects of other variables often make assumptions similar to the equality of population SDs. It seems strange to say that the SDs should be treated as unequal unless adjustments are being made! For this reason, I tend to use the common variances version of CIs, transforming the data if necessary to better satisfy requirement for equal population variances. That said, it is important to add that assumptions should not be made when they are known from the start to be false.

Last modified: undefined.

Paired Data / Paired Analyses
Gerard E. Dallal, Ph.D.

Introduction

Two measurements are **paired** when they come from the same observational unit: before and after, twins, husbands and wives, brothers and sisters, matched cases and controls. Pairing is determined by a study's design. It has nothing to do with the actual data values but, rather, with the way the data values are obtained. Observations are paired rather than independent when there is a natural link between an observation in one set of measurements and a particular observation in the other set of measurements, irrespective of their actual values.

The best way to determine whether data are paired is to identify the natural link between the two measurements. (*Look for the link!*) For example,

- when husbands and wives are studied, there is a natural correspondence between a man and his wife.
- When independent samples of men and women are studied, there's no particular female we associate with a particular male.

When measurements are paired, *the pairing must be reflected in the analysis*. The data **cannot** be analyzed as independent samples.

Why pair?

Pairing seeks to reduce variability in order to make more precise comparisons with fewer subjects. When independent samples are used, the difference between treatment means is compared to the variability of individual responses within each treatment group. This variability has two components:

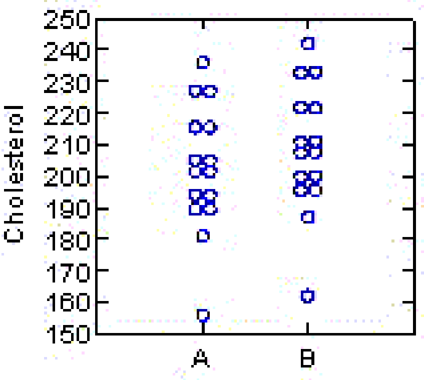
- The larger component is usually the variability between subjects (*between-subject variability*). It's there because not every subject will respond the same way to a particular treatment. There will be variability between subjects.

- The other component is *within-subject variability*. This variability is present because even the same subject doesn't give exactly the same response each time s/he is measured. There will be variability within subjects.

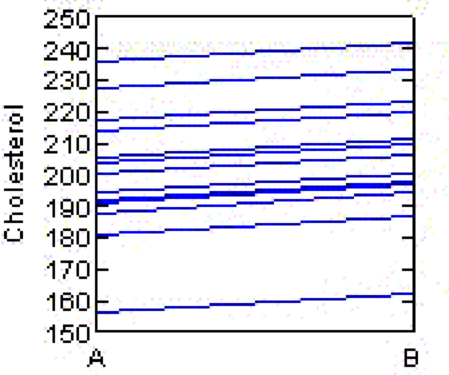
When both measurements are made on the same subject, the between-subjects variability is eliminated from the comparison. The difference between treatments is compared to the way the difference changes from subject to subject. If this difference is roughly the same for each subject, small treatment effects can be detected even if different subjects respond quite differently.

If measurements are made on paired or matched samples, the between-subject variability will be reduced according to the effectiveness of the pairings. The pairing or matching need not be perfect. The hope is that it will reduce the between-subject variability enough to justify the effort involved in obtained paired data. If we are interested in the difference in dairy intake of younger and older women, we could take random samples of young women and older women (independent samples). However, we might interview mother/daughter pairs (paired samples), in the hope of removing some of the lifestyle and socioeconomic differences from the age group comparison. Sometimes pairing turns out to have been a good idea because variability is greatly reduced. Other times it turns out to be have been a bad idea, as is often the case with matched samples.

Pairing has no effect on the way the difference between two treatments is estimated. The estimate is the difference between the sample means, whether the data are paired or not. What changes is the uncertainty in the estimate.



Consider these data from an experiment in which subjects are assigned at random to one of two diets and their cholesterol levels are measured. Do the data suggest a real difference in the effect of the two diets? The values from Diet A look like they might be a bit lower, but this difference must be judged relative to the variability within each sample. One of your first reactions to looking at these data should be, "Wow! Look at how different the values are. There is so much variability in the cholesterol levels that these data don't provide much evidence for a real difference between the diets." And that response would be correct. With P = 0.47 and a 95% CI for A-B of (-21.3, 9.3) mg/dl), we could say only that diet A produces a mean cholesterol level that could be anywhere from 21 mg/dL ***less*** than that from diet B to 9 mg/dL ***more***.



However, suppose you are now told that a mistake had been made. The numbers are correct, but the study was performed by having every subject consume both diets. The order of the diets was selected at random for each subject with a suitable washout period between diets. Each subject's cholesterol values are connected by a straight line in the diagram to the left.

Even though the mean difference is the same (6 mg/dl) we conclude the diets are certainly different because we now compare the mean difference of 6 to how much the individual differences vary. Each subject's cholesterol level on diet A is *exactly* 6 mg/dl less than on diet B! There is no question that there is an effect and that it is 6 mg/dl!

Paired data do not always result in a paired analysis

Paired analyses are required when *the outcome variable* is measured on the same or matched units. If there is an opportunity for confusion, it is because paired data do not always result in paired outcomes, as the following example illustrates. Suppose an investigator compares the effects of two diets on cholesterol levels by randomizing subjects to one of the two diets and measuring their cholesterol levels at the start and end of the study. The primary outcome will be the **change** in cholesterol levels. Each subject's before and after measurements are paired because they are made on the same subject. However, the diets will be compared by looking at **two independent samples of changes**. If, instead, each subject had eaten both diets--that is, if there were two diet periods with a suitable washout between them and the order of diets randomized--a paired analysis would be required because both diets would have been studied on the same people.

The need for a paired analysis is established by the study design. If an investigator chooses to study husbands and wives rather than random samples of men and women, the data must be analyzed as paired outcomes regardless of whether the pairing was effective. Whenever outcome measures are paired or matched, they cannot be analyzed as independent samples.

Paired analyses comparing two population means are straightforward. Differences are calculated within each observational unit and the single sample of differences is examined. If the sample size is large, normal theory applies and the sample mean difference and population mean difference will be within two standard errors of the mean difference 95% of the time. If, by mistake, the data were treated as independent samples, the mean difference will be estimated properly but the amount of uncertainty against which it must be judged will be wrong. The uncertainty will usually be overstated, causing some real differences to be missed. However, although it is unlikely, it is possible for uncertainty to be *under*stated, causing things to appear to be different even though the evidence is inadequate. Thus, criticism of an improper analysis cannot be dismissed by claiming that because an unpaired analysis shows a difference, the paired analysis will show a difference, too.

Pairing is usually optional. In most cases an investigator can choose to design a study that leads to a paired analysis or one that uses independent samples. The choice is a matter of tradeoffs between cost, convenience, and likely benefit. A paired study requires fewer subjects, but the subjects have to experience both treatments, which might prove a major inconvenience. Subjects with partial data usually do not contribute to the analysis. Also, when treatments must be administered in sequence rather than simultaneously, there are questions about whether the first treatment will affect the response to the second treatment (*carry-over effect*). In most cases, a research question will not require the investigator to take paired samples, but if a paired study is undertaken, a paired analysis **must** be used. That is, **the analysis must *always* reflect the design that generated the data**.

It is possible for pairing to be ineffective, that is, the variability of the difference between sample means can be about the same as what would have been obtained from independent samples. In general, matched studies in human subjects with matching by sex, age, BMI and the like are almost always a disaster. The matching is almost always impossible to achieve in practice (the subjects needed for the last few matches never seem to volunteer) and the efficiencies are rarely better than could be achieved by using statistical adjustment instead.

Examples -- Paired or Independent Analysis?

1. A hypothesis of ongoing clinical interest is that vitamin C prevents the common cold. In a study involving 20 volunteers, 10 are randomly assigned to receive vitamin C capsules and 10 are randomly assigned to receive placebo capsules. The number of colds over a 12 month period is recorded.

2. A topic of current interest in ophthalmology is whether or not spherical refraction is different between the left and right eyes. To examine this issue, refraction is measured in both eyes of 17 people.

3. In order to compare the working environment in offices where smoking is permitted with that in offices where smoking was not permitted, measurements were made at 2 p. m. in 40 work areas where smoking was permitted and 40 work areas was not permitted.

4. A question in nutrition research is whether male and female college students undergo different mean weight changes during their freshman year. A data file contains the September 1994 weight (lbs), May 1995 weight (lbs), and sex (1=male/2=female) of students from the class of 1998. The file is set up so that each record contains the data for one student. The first 3 records, for example, might be

120	126	2
118	116	2
160	149	1

5. To determine whether cardiologists and pharmacists are equally knowledgeable about how nutrition and vitamin K affect anticoagulation therapy (to prevent clotting), an investigator has 10 cardiologists and 10 pharmacists complete a questionnaire to

measure what they know. She contacts the administrators at 10 hospitals and asks the administrator to select a cardiologist and pharmacist at random from the hospital's staff to complete the questionnaire.

6. To determine whether the meals served on the meal plans of public and private colleges are equally healthful, an investigator chooses 7 public colleges and 7 private colleges at random from a list of all colleges in Massachusetts. On each day of the week, she visits one public college and one private college. She calculates the mean amount of saturated fat in the dinner entrees at each school.

Last modified: undefined.

What does pairing *really* do?

Whether data are independent samples or paired, the best estimate of the difference between population means is the difference between sample means. When the data are two independent samples of size *n* with approximately equal sample standard deviations ($s_x \approx s_y \approx s$), a 95% confidence interval for the population mean difference, $\mu_x - \mu_y$, is

$$(\bar{x} - \bar{y}) \pm 1.96 s \sqrt{\frac{2}{n}}$$

Now suppose the data are *n* paired samples ((X_i,Y_i): i=1,..,n) where the sample standard deviations of the Xs and Ys are roughly equal ($s_x \approx s_y \approx s$) and the correlation between X & Y is *r*. A 95% confidence interval for the population mean difference, $\mu_x - \mu_y$, is

$$(\bar{x} - \bar{y}) \pm 1.96 s \sqrt{\frac{2(1-r)}{n}}$$

If the two responses are uncorrelated--that is, if the correlation coefficient is 0--the pairing is ineffective. The confidence interval is no shorter than it would have been had the investigators not taken the trouble to collect paired data. On the other hand, the stronger the correlation, the narrower the confidence interval and the more effective was the pairing. This formula also illustrates that pairing can be worse than ineffective. Had the correlation been negative, the confidence interval would have been longer than it would have been with independent samples.

[back to LHSP]

Last modified: undefined.

The Ubiquitous Sample Mean!

The sample mean plays many distinct roles.

- It is the best estimate of an individual value in the sample. ("If I were to select one observation at random from the sample, what would you guess that value to be?" "<The sample mean>.")
- It is the best estimate of an individual value drawn from the population. ("If I were to select one observation from the population, what would you guess that value to be?" "<The sample mean>." Or, "If we were to collect one more observation, what would you guess its value to be?" "<The sample mean>.") Notice that collecting one more observation is the same thing as drawing an observation at random from the population.
- It is the best estimate of the mean of the population from which the sample was drawn. ("What would you guess the mean of all values in the population to be?" "<The sample mean>.")
- Whatever else it is, it *is* the mean of the sample.

The differences between these roles must be appreciated and understood. Failing to distinguish between them is a common cause of confusion about many basic statistical techniques.

The sample mean and standard deviation ($\bar{x}$, s) together summarize individual data values when the data follow a normal distribution or something not too far from it. The sample mean describes a typical value. The sample standard deviation (SD) measures the spread of individual values about the sample mean. The SD also estimates the spread of individual values about the population mean teh extent to which a single value chosen at random might differ from the population mean.

Just as the sample standard deviation measures the uncertainty with which the sample mean estimates individual measurements, a quantity called *the Standard Error of the Mean* (SEM = $s/\sqrt{n}$) measures the uncertainty with which the sample mean estimates a population mean. Read the last sentence again... and again.

- The sample mean estimates individual values.
  - The uncertainty with which $\bar{x}$ estimates individual values is given by the SD.
- The sample mean estimates the population mean.
  - The uncertainty with which $\bar{x}$ estimates the population mean is given by the SEM.

Intuition says the more data there are, the more accurately we can estimate a population mean. With more data, the sample and population means are likely to be closer. The SEM expresses this numerically. The SEM says the likely difference between the sample and population means, $\bar{x} - \mu$, decreases as the sample size increases, but the decrease is proportional to the *square root* of the sample size. To decrease the uncertainty by a factor of 2, the sample size must be increased by a factor of 4; to cut the uncertainty by a factor of 10, a sample 100 times larger is required.

We have already noted that when individual data items follow something not very far from a normal distribution, 68% of the data will be within one standard deviation of the mean, 95% will be within two standard deviations of the mean, and so on. But, this is true only when the individual data values are roughly normally distributed.

There is an elegant statistical limit theorem that describes the likely difference between sample and population means, $\bar{x} - \mu$, when sample sizes are large. It is so central to statistical practice that is is called the Central Limit Theorem. It says that, for large samples, the normal distribution can be used to describe the likely difference between the sample and population means *regardless* of the distribution of the individual data items! In particular, 68% of the time the difference between the sample and population means will be less than 1 SEM, 95% of the time the difference will be less than 2 SEMs, and so on. You can see why the result is central to statistical practice. It lets us ignore the distribution of individual data values when talking about the behavior of sample means in large samples. The distribution of individual data values becomes irrelevant when making statements about the difference between sample and population means. From a statistical standpoint, sample means obtained by replicating a study can be thought of as individual observations whose standard deviation is equal to the SEM.

Let's stop and summarize: When describing the behavior of individual values, the normal distribution can be used only when the data themselves follow something close to a normal histogram. When describing the difference between sample and population means based on large enough samples, the normal distribution can be used whatever the histogram of the individual observations. Let's continue…

Anyone familiar with mathematics and limit theorems knows that limit theorems begin, "As the sample size approaches infinity . . ." No one has infinite amounts of data. The question naturally arises about the sample size at which the result can be used in practice. Mathematical analysis, simulation, and empirical study have demonstrated that for the types of data encountered in the natural and social sciences (and certainly almost any response measured on a continuous scale) sample sizes as small as 30 to 100 (!) will be adequate.

To reinforce these ideas, consider dietary intake, which tends to follow a normal distribution. Suppose we find that daily caloric intakes in a random sample of 100 undergraduate women have a mean of 1800 kcal and a standard deviation of 200 kcal. Because the individual values follow a normal distribution, approximately 95% of them will be in the range (1400, 2200) kcal $(= \bar{x} \pm 2\ SD = 1800 \pm 2 \times 200)$. The Central Limit theorem lets us do the same thing to estimate the (population) mean daily caloric intake of all undergraduate women. The SEM is 20 (=200/$\sqrt{}$100). A 95% confidence interval for the mean daily caloric intake of all undergraduate women is (1760, 1840) kcal $(= \bar{x} \pm 2\ SEM = 1800 \pm 2 \times 20)$. That is, we are 95% confident the mean daily caloric intake of all undergraduate women falls in the range (1760, 1840) kcal.

Consider household income, which invariably is skewed to the right. Most households have low incomes while a few have very large incomes. Suppose household incomes measured in a random sample of 400 households have a mean of $10,000 and a SD of $3000. The SEM is $150 (= $3000/\sqrt{400}$). Because the data do not follow a normal distribution, there is no simple rule involving the sample mean and SD that can be used to describe the location of the bulk of the individual values. However, we can still construct a 95% confidence interval for the population mean income as $\bar{x} \pm 2\,SEM\ (= \$10{,}00 \pm 2 \times \$150)$ or $(9700, 10300). Because the sample size is large, the distribution of individual incomes is irrelevant to constructing confidence intervals for the population mean.

Comments

- In most textbooks, the discussion of confidence intervals begins by assuming the population standard deviation, σ, is known. The sample and population means will be within $2\sigma/\sqrt{n}$ of each other, 95% of the time. The reason the textbooks take this approach is that the mathematics is easier when σ is known. In practice, the population standard deviation is never known. However, statistical theory shows that the results remain true when the sample SD, s, is used in place of the population SD, σ .
- There is a direct link between *having 95% confidence* and *adding and subtracting 2 SEMs*. If more confidence is desired, the interval must be made larger/longer/wider. For less confidence, the interval can be smaller/shorter/narrower. In practice, only 95% confidence intervals are reported, although on rare occasions, 90% (± 1.645 SEM) or 99% (± 2.58 SEM) confidence intervals may appear. The reason ± 2 SEMs gives a 95% CI and ± 2.58 SEMs gives a 99% CI has to do with the shape of the normal distribution. You can study the distribution in detail, but in practice, it's always going to be 95% confidence and ± 2 SEM.
- ±2 SEM is a commonly used approximation. The exact value for a 95% confidence interval based on the Normal distribution is ± 1.96 SEM rather than 2, but 2 is used for hand calculation as a matter of convenience. Computer programs use a value that is close to 2, but the actual value depends on the sample size, as we shall see.

SD or SEM?

A question commonly asked is whether summary tables should include mean ± SD or mean ± SEM. In many ways, it hardly matters. Anyone wanting the SEM merely has to divide the SD by $\sqrt{n}$. Similarly, anyone wanting the SD merely has to multiply the SEM by $\sqrt{n}$.

The sample mean describes both the population mean and an individual value drawn from the population. The sample mean and SD together describe individual observations. The sample mean and SEM together describe what is known about the population mean. If the goal is to focus the reader's attention on the distribution of individual values, report the mean ± SD. If the goal is to focus on the precision with which population means are known, report the mean ± SEM.

Last modified: undefined.

What Student Did

With large samples, confidence intervals for population means can be constructed by using only the sample mean, sample standard deviation, sample size, and the properties of the normal distribution. This is true regardless of the distribution of the individual observations.

Early in the history of statistical practice, it was recognized that there was no similar result for small samples. Even when the individual observations themselves follow a normal distribution exactly, the difference between the sample and population means tends to be greater than the normal distribution predicts. For small samples, confidence intervals for the population mean constructed by using the normal distribution are too short (they contain the population mean less often than expected) and statistical tests (to be discussed) based on the normal distribution reject a true null hypothesis more often than expected. Analysts constructed these intervals and performed these tests for lack of anything better to do, but they were aware of the deficiencies and treated the results as descriptive rather than inferential.

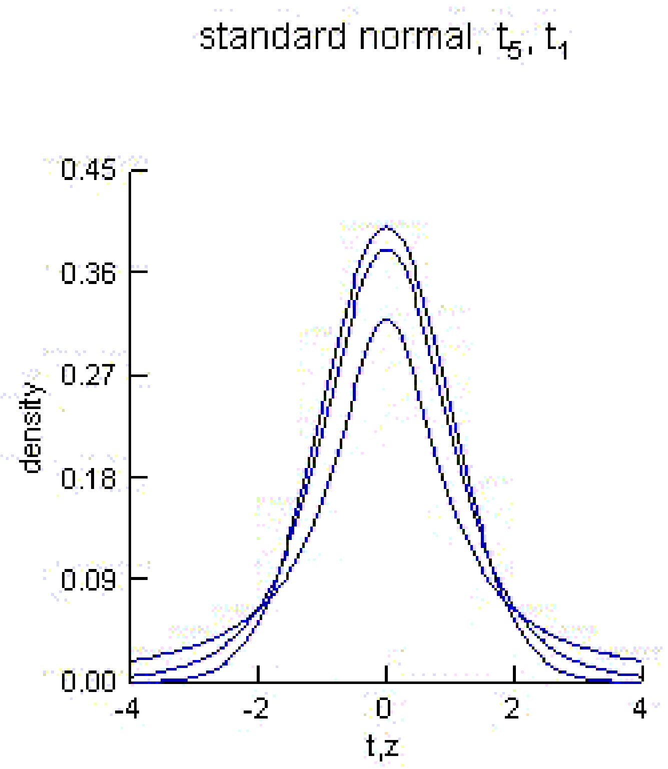
William Sealey Gosset, who published under the pseudonym 'A Student of Statistics', discovered that when individual observations follow a normal distribution, confidence intervals for population means could be constructed in a manner similar to that for large samples. The only difference was that the usual multiplier was replaced by one that grew larger as the sample size became smaller. He also discovered that a similar method could be used to compare two population means provided individual observations in both populations follow normal distributions and the **population** standard deviations were equal (sample standard deviations are never equal)--the 1.96 is replaced by a multiplier that depends on the combined sample size. Also, the two sample standard deviations were combined (or *pooled*) to give a best estimate of the common population standard deviation. If the samples have standard deviations s_1 and s_2, and sample sizes n_1 and n_2, then the pooled standard deviation is

$$s_p = \surd\ (\ [(n_1-1)\ s_1^2 + (n_2-1)\ s_2^2]\ /\ [n_1 + n_2 - 2]\ )$$

and the standard deviation of the difference between the sample means is

$$s_p\ \surd\ (1/n_1 + 1/n_2)$$

It was now possible to perform exact significance tests and construct exact confidence intervals based on small samples in many common situations. Just as the multipliers in the case of large samples came from the normal distribution, the multipliers in the case of small samples came from a distribution which Student named the t distribution. Today, it is known as Student's t distribution.



There isn't just one t distribution. There is an infinite number of them, indexed (numbered) 1, 2, 3, and so on. The index, called "degrees of freedom," allows us to refer easily to any particular t distribution. ("Degrees of freedom" is not hyphenated. The only terms in statistics that are routinely hyphenated are "chi-square" and "goodness-of-fit.") The t distributions are like the normal distribution--unimodal and symmetric about 0--but they are spread out a bit more (heavier in the tails). As the degrees of freedom get larger, the t distribution gets closer to the standard normal distribution. A normal distribution is a t distribution with infinite degrees of freedom.

Each analysis has a particular number of degrees of freedom associated with it. Virtually all computer programs calculate the degrees of freedom automatically, but knowing how to calculate degrees of freedom by hand makes it easy to quickly check that the proper analysis is being performed and the proper data are being used.

When estimating a single population mean, the number of degrees of freedom is n - 1. When estimating the difference between two population means, the number of degrees of freedom is $n_1 + n_2 - 2$.

The only change in tests and confidence intervals from those based on large sample theory is the value obtained from the normal distribution, such as 1.96, is replaced by a value from a t distribution.

In the old days (B.C: before computers) when calculations were done by hand, analysts would use the normal distribution if the degrees of freedom were greater than 30 (for 30 df, the proper multiplier is 2.04; for 60 df, it's 2.00). Otherwise, the t distribution was used. This says as much about the availability of tables of the t distribution as anything else.

Today, tables of distributions have been replaced by computer programs. The computer thinks nothing about looking up the t distribution with 2351 degrees of freedom, even if it is almost identical to the standard normal distribution. There is no magic number of degrees of freedom above which the computer switches over to the standard normal distribution. Computer programs that compare sample means use Student's t distribution for every sample size and the standard normal distribution never comes into play.

We find ourselves in a peculiar position. Before computers, analysts used the standard normal distribution to analyze every large data set. It was an approximation, but a good one. After computers, we use t distributions to analyze every large data set. It works for large non-normal samples because a t distribution with a large number of degrees of freedom is essentially the standard normal distribution. The output may say *t test*, but it's the large sample theory that makes the test valid and large sample theory says that the distribution of a sample mean is approximately normal, not t!

Last modified: undefined.

What Did Student *Really* Do?
(What Student Did, Part II)

When Student's t test for independent samples is run, every statistics package reports two results. They may be labeled

equal variances assumed	equal variances not assumed
common variance	separate variances
pooled variance	separate variances

Which results should be used?

The variances mentioned in the table and the output are **population** variances. One thing Student did was to say that if the population variances were known to be equal or could be assumed to be equal, exact tests and confidence intervals could be obtained by using his t distribution. This is the test labeled *equal variances assumed*, *common variance* or *pooled variance*. The term *pooled variance* refers to way the estimate of the common variance is obtained by pooling the data from both samples.

The test labeled *equal variances not assumed* or *separate variances* is appropriate for normally distributed individual values when the population variances are known to be unequal or cannot be assumed to be equal. This is not an exact test. It is approximate. The approximation involves t distributions with non-integer degrees of freedom. Before the ready availability of computers, the number of degrees of freedom was awkward to calculate and the critical values were not easy to obtain, so statisticians worried about how much the data could depart from the ideal of equal variances without affecting the validity of Student's test. It turned out the t test was extremely robust to departures from normality and equal variances.

Some analysts recommended performing preliminary statistical tests to decide whether the data were normally distributed and whether population variances were equal. If the hypothesis of equal population variances was rejected, the *equal variances not assumed* form of the test would be used, otherwise *equal variances assumed* version would be used. However, it was discovered that Students t test for independent samples was so robust that the preliminary tests would have analysts avoiding the *equal variances assumed* form when it was in no danger of it giving misleading results. These preliminary tests often detect differences too small to affect Student's t test. The analogy most often given is that using preliminary tests of normality and equality of variances to decided whether it was safe to use the *equal variances assumed* version of the t test was like sending out a rowboat to see whether it was safe for the ocean liner. Today, common practice is to avoid preliminary tests. Important violations of the requirements will be detectable to the naked eye without a formal significance test.

Rupert Miller, Jr., in his 1986 book *Beyond ANOVA, Basics of Applied Statistics*, {New York: John Wiley & Sons] summarizes the extent to which the assumptions of normality and equal population variances can be violated without affecting the validity of Student's test.

- If sample sizes are equal, (a) nonnormality is not a problem and (b) the t test can tolerate population standard deviation ratios of 2 without showing any major ill effect. The worst situation occurs when one sample has a much larger variance and a much smaller sample size than the other. For example, if the variance ratio is 5 and the sample size ratio is 1/5, a nominal P value of 0.05 is actually 0.22.

- Serious distortion of the P value can occur when the skewness of the two populations is different.
- Outliers can distort the mean difference and the t statistic. They tend to inflate the variance and depress the value and corresponding statistical significance of the t statistic.

Still, which test should be used?

Frederick Mosteller and John Tukey, on pages 5-7 of *Data Analysis and Regression* [Reading, MA: Addison-Wesley Publishing Company, Inc., 1997] provide insight into what Student really did and how it should affect our choice of test.

> The value of Student's work lay not in great numerical change, but in:
>
> - recognition that one could, if appropriate assumptions held, make allowances fo the "uncertainties" of small samples, not only in Student's original problem, but in others as well;
> - provision of a numerical assessment of how small the necessary numerical adjustment of confidence points were in Student's problem...
> - presentation of tables that could be used--in setting confidence limits, in making significance tests--to assess the uncertainty associated with even very small samples.
>
> Besides its values, Student's contribution had its drawbacks, notably:
>
> - it made it too easy to neglect the proviso "if appropriate assumptions held";
> - it overemphasized the "exactness of Student's solution for his idealized problem";
> - it helped to divert the attention of theoretical statisticians to the development of "exact" ways of treating other problems; and
> - it failed to attack the "problem of multiplicity": the difficulties and temptation associated with the application of large numbers of tests to the same data.
>
> The great importance given to exactness of treatment is even more surprising when we consider how much the small differences between the critical values of the normal approximation and Student's *t* disappears, especially at and near the much-used two-sided 5% point, when, as suggested by Burrau (1943), we multiply *t* by the constant required to bring its variance to 1, namely, [√ {(f - 2) /f}].

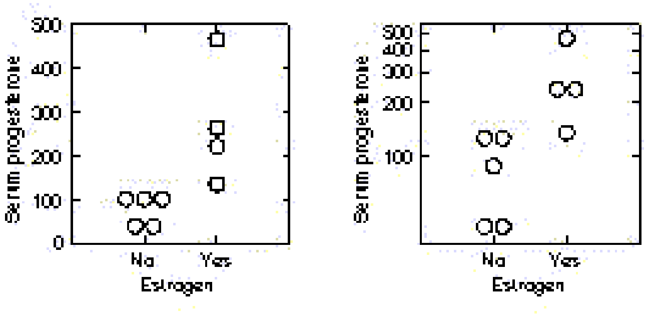
The separate variances version rarely differs from common variance. When it does, there's usually a problem with the common variances version.

When sample sizes are large, the Central Limit Theorem takes over. The behavior of the separate variances t statistic is described by the normal distribution regardless of the distribution of the individual observations. The two populations of individual observations need not have the same variances. They need not even be normal.

If the separate variances t test is always valid for large samples and if the common variances test is probably invalid when the two tests disagree in small samples, why not use the separate variances version exclusively? Some statisticians seem to advocate this approach. The primary advantage of the common variances test is that it generalizes to more than two groups (analysis of variance).

When within group variances are unequal, it often happens that the standard deviation is proportional to the mean. For example, instead of the within group standard deviation being a fixed value such as 5 mg/dl, it is often a fixed percentage. If the standard deviation were 20% of the mean, one group might have values of 10 give or take 2, while the other might have values of 100 give or take 20. In such cases, a logarithmic transformation will produce groups with the same standard deviation. If natural logarithms are use, the common within group standard deviation of the transformed data will be equal to the ratio of the within group standard deviation to the mean (also known as the *coefficient of variation*).

The purpose of transforming data is not to achieve a particular result. Transformations are not performed to make differences achieve statistical significance. Transformations allow us to use standard statistical techniques confident they are appropriate for the data to which they're applied. That is, transformations are applied not to achieve a particular result, but to insure the results we obtain will be reliable.



The following small dataset illustrates some of the issues. Serum progesterone levels were measured in subjects randomized to receive estrogen or not. The group with the higher serum progesterone levels also has the greater spread. The equal variances assumed t test has an observed significance level of 0.022; the unequal variances assumed t test has an observed significance level of 0.069. When a logarithmic transformation is applied to the data, the within group standard deviations are both around 0.50, which is approximately the ratio of the SD to the mean, and both P values are 0.012. The conclusion is that the geometric mean of progesterone levels of those given this dose of progesterone is between 1.4 and 7.9 times the levels of those on placebo (95% CI). Insofar as the geometric mean is a good approximation to the median, the previous sentence might be restated in terms of medians.

This is a very small dataset, so small that tests for the inequality of population variances do not achieve statistical significance, even though one SD is three times larger than the other. Still, it possesses the essential features of one group having a much larger standard deviation and the standard deviations being proportional to mean response, so it is worthy of consideration.

	Estrogen	N	Mean	SD	SEM
SPROG	No	5	81.8000	40.4747	18.1008
	Yes	4	271.5000	139.5337	69.7669
ln(SPROG)	No	5	4.2875	.5617	.2512
	Yes	4	5.5072	.5076	.2538

	Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper

SPROG	Equal variances assumed	2.843	.136	-2.935	7	.022	-189.7000	64.6229	-342.5088	-36.8912
	Equal variances not assumed			-2.632	3.406	.069	-189.7000	72.0767	-404.3650	24.9650
ln(SPROG)	Equal variances assumed	.639	.450	-3.372	7	.012	-1.2197	.3617	-2.0750	-.3644
	Equal variances not assumed			-3.416	6.836	.012	-1.2197	.3571	-2.0682	-.3712

Gerard E. Dallal

Last modified: undefined.

Significance Tests / Hypothesis Testing

Statistical theory says that in many situations where a population quantity is estimated by drawing random samples, the sample and population values will be within two standard errors of each other 95% of the time. That is, 95% of the time,

$$-1.96\ SE \leq \textit{population quantity} - \textit{sample value} \leq 1.96\ SE \quad [*]$$

This is the case for means, differences between means, proportions, and differences between proportions,

We have used this fact to construct 95% confidence intervals by restating the result as

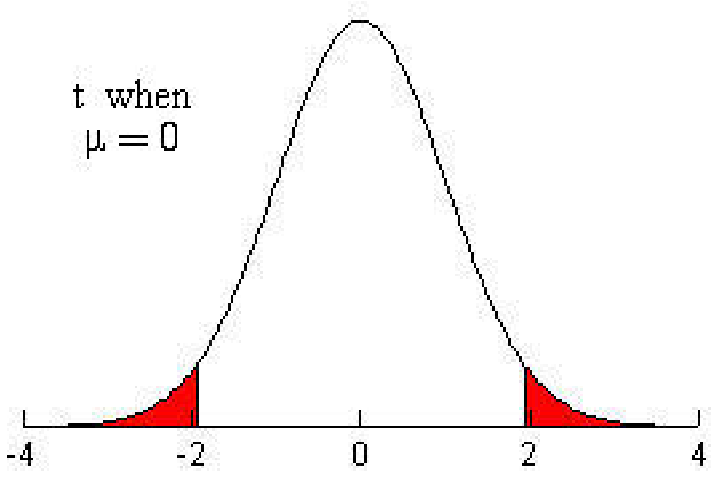
$$\textit{sample value} - 1.96\ SE \leq \textit{population quantity} \leq \textit{sample value} + 1.96\ SE$$

95% of intervals constructed in this way will contain the population quantity of interest. For example, a 95% CI for the difference between two population means, μ_x-μ_y, is given by

$$(\bar{x}-\bar{y}) - 1.96\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}} \leq \mu_x - \mu_y \leq (\bar{x}-\bar{y}) + 1.96\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}\ .$$

Here's another way to look at [*]: 95% of the time

$$-1.96 \leq \frac{\textit{sample value} - \textit{population quantity}}{SE} \leq 1.96$$



Suppose you wanted to test whether a population quantity were equal to 0. You could calculate the value of

$$t = \frac{(\textit{sample value}) - 0}{SE}$$

which we get by inserting the hypothesized value of the population mean difference (0) for the population_quantity. If *t<-1.96* or *t>1.96* (that is, |*t*|*>1.96*), we say the data are not consistent with a population mean difference of 0 (because *t* does not have the sort of value we expect to see when the population value is 0) or "we **reject the hypothesis**

that the population mean difference is 0". If t were -3.7 or 2.6, we would reject the hypothesis that the population mean difference is 0 because we've observed a value of t that is unusual if the hypothesis were true.

If $-1.96 \le t \le 1.96$ (that is, $|t| \le 1.96$), we say the data are consistent with a population mean difference of 0 (because *t* has the sort of value we expect to see when the population value is 0) or "we **fail to reject the hypothesis that the population mean difference is 0**". For example, if t were 0.76, we would fail reject the hypothesis that the population mean difference is 0 because we've observed a value of t that is unremarkable if the hypothesis were true.

This is called "fixed level testing" (at the 0.05 level). We hypothesize a value for a population quantity and determine values of t that would cause us to reject the hypothesis. We then collect the data and reject the hypothesis or not depending on the observed value of t. For example, if H_0: $\mu_x = \mu_y$ (which can be rewritten H_0: $\mu_x - \mu_y = 0$), the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

If $|t| > 1.96$, reject H_0: $\mu_x = \mu_y$ at the 0.05 level of significance.

When we were constructing confidence intervals, it mattered whether the data were drawn from normally distributed populations, whether the population standard deviations were equal, and whether the sample sizes were large or small, The answers to these questions helped us determine the proper multiplier for the standard error. The same considerations apply to significance tests. The answers determine the critical value of *t* for a result to be declared statistically significant.

When populations are normally distributed with unequal standard deviations and the sample size is small, the multiplier used to construct CIs is based on the t distribution with noninteger degrees of freedom. The same noninteger degrees of freedom appear when performing significance tests. Many ways to calculate the degrees of freedom have been proposed. The statistical program package SPSS, for example, uses the Satterthwaite formula

$$\frac{(d_x + d_y)^2}{\frac{d_x^2}{n_x - 1} + \frac{d_y^2}{n_y - 1}} \text{, where } d_i = \frac{s_i^2}{n_i} \text{ .}$$

Terminology

- **Null hypothesis--**the hypothesis under test, denoted H_0. The null hypothesis is usually stated as the absence of a difference or an effect. It functions as what debaters call a "straw man", something set up solely to be knocked down. It is usually the investigator's intention to demonstrate an effect is present. The null hypothesis says there is no effect. The null hypothesis is rejected if the significance test shows the data are inconsistent with the null hypothesis. Null hypothesis are *never* accepted. We either reject them or fail to reject them. The distinction between "acceptance" and "failure to reject" is best understood in terms of confidence intervals. Failing to reject a hypothesis means a confidence interval contains a value of "no difference". However, the data may also be consistent with differences of practical importance. Hence, failing to reject H_0 does not mean that we have shown that there is no difference (accept H_0).
- **Alternative Hypothesis--**the alternative to the null hypothesis. It is denoted H', H_1, or H_A. It is usually the complement of the null hypothesis. Many authors talk about rejecting the null hypothesis in favor of the alternative. If, for example, the null hypothesis says two population means are equal, the alternative says the means are unequal. The amount of emphasis on the alternative hypothesis will depend on whom you read. It is central to the Neyman-Pearson school of frequentist statistics. Yet, R.A. Fisher, who first introduced the notion of significance tests in a formal systematic way, never considered alternative hypotheses. He focused entirely on the null.
- **Critical Region (Rejection Region)--**the set of values of the test statistic that cause the null hypothesis to be reject. If the test statistic falls in the rejection region--that is, if the statistic is a value that is in the rejection region--the null hypothesis is rejected. In the picture above, the critical region is the area filled in with red.
- **Critical Values--**the values that mark the boundaries of the critical region. For example, if a critical region is $\{t \leq -1.96, t \geq 1.96\}$, the critical values are ± 1.96 as in the picture above.
- **Power** is the probability of rejecting the null hypothesis. It is not a single value. It varies according to the underlying truth. For example, the probability of rejecting the hypothesis of equal population means depends on the actual difference in population means. The probability of the rejecting the null hypothesis increases with the difference between population means.
- The **level** (or **size**) of a test is the probability of rejecting the null hypothesis when it is true. It is denoted by the Greek letter α (*alpha*). Rejecting the null hypothesis, H_0, when it is true is called a **Type I Error**. Therefore, if the null hypothesis is true α, the level of the test, is the probability of a type I error. α is also the power of the test when the null hypothesis, H_0, is true. In the picture above, α is the proportion of the distribution colored in red. The choice of α determines the critical values. The tails of the distribution of *t* are colored in until the proportion filled in is α, which determines the critical values.
- A **Type II Error** occurs when we fail to reject the null hypothesis when it is false. The probability of a type II error depends on the way the null hypothesis is false. For example, for a fixed sample size, the probability of failing to reject a null hypothesis of equal population means decreases as the difference between population means increases. The probability of a type II error is denoted by the Greek letter β (*beta*). By definition, power = 1 - β when the null hypothesis is false.

The difference between type I & type II errors is illustrated by the following legal analogy. Under United States law, defendants are presumed innocent until proven guilty. The purpose of a trial is to see whether a null hypothesis of innocence is rejected by the weight of the data (evidence). A type I error (rejecting the null hypothesis when it is true) is "convicting the innocent." A type II error (failing to reject the null hypothesis when it is false) is "letting the guilty go free."

A common mistake is to confuse a type I or II error with its probability. α is not a type I error. It is the *probability* of a type I error. Similarly, β is not a type II error. It is the *probability* of a type II error.

There's a trade-off between α and β. Both are probabilities of making an error. With a fixed sample size, the only way to reduce the probability of making one type of error is to increase the other. For the problem of comparing population means, consider the rejection region whose critical values are $\pm\infty$. This excludes every possible difference in sample means. H_0 will never be rejected. Since the null hypothesis will never be rejected, the probability of rejecting the null hypothesis when it is true is 0. So, α=0. However, since the null hypothesis will never be rejected, the probability of failing to reject the null hypothesis when it is false is 1, that is, β=1.

Now consider the opposite extreme--a rejection region whose critical values are 0,0. The rejection region includes every possible difference in sample means. This test always rejects H_0. Since the null hypothesis is always rejected, the probability of rejecting H_0 when it is true is 1, that is, α=1. On the other hand, since the null hypothesis is always rejected, the probability of failing to reject it when it is false is 0, that is, β=0.

To recap, the test with a critical region bounded by $\pm\infty$ has α=0 and β=1, while the test with a critical region bounded by 0,0 has α=1 and β=0. Now consider tests with intermediate critical regions bounded by $\pm k$. As k increases from 0 to ∞, α decreases from 1 to 0 while β increases from 0 to 1.

Every statistics textbook contains discussions of α, β, type I error, type II error, and power. Analysts should be familiar with all of them. However, α is the only one that is encountered regularly in reports and published papers. That's because standard statistical practice is to carry out significance tests at the 0.05 level. As we've just seen, choosing a particular value for α determines the value of β.

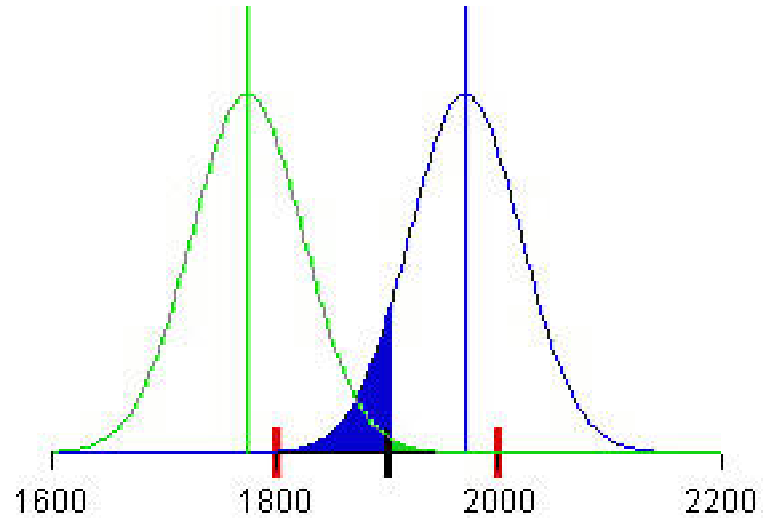
The one place where β figures prominently in statistical practice is in determining sample size. When a study is being planned, it is possible to choose the sample size to set the power to any desired value for some particular alternative to the null hypothesis. To illustrate this, suppose we are testing the hypothesis that two population means are equal at the 0.05 level of significance by selecting equal sample sizes from the two populations. Suppose the common population standard deviation is 12. Then, if the population mean difference is 10, a sample of 24 subjects per group gives an 81% chance of rejecting the null hypothesis of no difference (power=0.81, β=0.19). A sample of 32 subjects per group gives an 91% chance of rejecting the null hypothesis of no difference (power=0.91, β=0.09). This is discussed in detail in the section on sample size determination.

[back to LHSP]

Last modified: undefined.

Significance Tests -- Take 2

Here's another way to look at significance tests as a reformulation of confidence intervals. A data set has a mean and standard deviation. The standard deviation and the sample size determine the standard error of the mean (SEM=SD/ √n). The standard error of the mean tells us how far apart the sample mean and population mean are likely to be: They will be within 2 SEMs of each other 95% of the time.



In this example, a sample of 36 students on a particular meal plan is studied to determine their mean daily caloric intake. The sample has a mean of 1900 kcal and a SD of 300 kcal. The SEM is 50 [= 300 / √36]. The 95% confidence interval for the population mean is illustrated in the picture to the left. The CI is centered at 1900, the sample mean, which is indicated by the **thick black mark** on the horizontal axis. The ends of the confidence interval are located at 1800 [= 1900 - 2(50)]1800 and 2000 [= 1900 + 2(50)] and are indicated by the **thick red marks**. This CI says that we are 95% confident that the population mean, whatever it is, is somewhere between 1800 and 2000 kcal.

Significance testing does the same thing in a different way. Consider the hypothesis H_0: μ=1970. If the hypothesis were true, our sample mean would have come from the distribution to the right, where the curve is drawn in blue. The population mean 1970 is within the 95% CI and the thick black mark at 1800 is well within the distribution as indicated by the area shaded blue. Now consider the hypothesis H_0: μ=1775. If *this* hypothesis were true, our sample mean would have come from the distribution indicated by the green curve to the left. The population mean 1775 lies outside the 95% CI and the thick black mark at 1800 is out in the tail of the green distribution as indicated by the area shaded green. In the language of significance tests, a value of 1900 is relatively rare coming from a normal distribution with a mean of 1775 and a SD of 50, so we reject the hypothesis that the population mean is 1775. The same argument holds for any hypothesized value of the population mean that is outside the confidence interval. On the other hand, values like 1900 are typical from a normal distribution with a mean of 1970 and a SD of 50, so we lack statistical sufficient statistical evidence to reject the hypothesis that the population mean is 1970. This argument holds for any hypothesized value that is *inside* the confidence interval.

[back to LHSP]

Last modified: undefined.

Significance Tests Simplified

If an observed significance level (P value) is less than 0.05, reject the hypothesis under test; if it is greater than 0.05, fail to reject the null hypothesis.

That was the most difficult sentence I've had to write for these notes, not because it's wrong (indeed, it's what good analysts often do when they conduct a statistical test) but because its indiscriminate and blind use is at the root of much bad statistical practice. So, I hate to just come out and say it.

A good analyst knows this prescription should be applied only when all of the principles of good study design and execution have been followed. A well-posed research question would have been developed and the proper data collected. The significance test (and appropriate confidence intervals) for judging the data are just one of many carefully thought-out steps.

A good analyst also knows that 0.05 is not a magic number. There is little difference between P=0.04 and P=0.06. Construct a few 94 and 96% CIs from the same data set and see how little they differ (one is about 10% longer than the other.) Is the question only about whether there is a difference no matter how small, or does practical importance play a role?

The danger is that any set of numbers can be fed into a statistical program package to produce a P value. The proper procedure is to

1. State the hypothesis to be tested.
2. Select a test procedure.
3. Collect relevant data.
4. Obtain a P value, which is a measure of whether the data are consistent with the hypothesis.
5. If the P value is small enough (usually, less than 0.05) the data are not consistent with the hypothesis and the hypothesis is rejected. If the P value is large (usually, greater than 0.05), the data are not judged inconsistent with the hypothesis and we fail to reject it.

Item (5) is just a piece of the puzzle. It is not the whole answer. Statistical significance is irrelevant if the effect is of no practical importance. That said, significance tests are an important and useful piece of the puzzle. Every so often, a cry is raised that P values should no longer be used because of the way they can be abused. Those who would abandon significance tests entirely because of the potential of misuse make an even greater mistake that those who abuse them.

Good Analyst, Bad Analyst

The difference between a good analyst and a bad analyst is that the bad analyst follows each step without any idea of what is necessary to insure the validity and generalizability of the results. The bad analyst sees only P<0.05 or P>0.05 with no regard for confidence intervals or for the context in which the data were collected.

The good analyst knows that the first two steps require that all of the principles of good study design have been followed. The good analyst knows what a test procedure requires for the resulting P value to be valid. The good analyst treats the P value as an important part of the analysis, but not as the whole answer.

Last modified: undefined.

Student's t Test for Independent Samples

Student's t test for independent samples is used to determine whether two samples were drawn from populations with different means. If both samples are large, the *separate* or *unequal variances* version of the t test has many attractive features. The denominator of the test statistic correctly estimates the standard deviation of the numerator, while the Central Limit Theorem guarantees the validity of the test even if the populations are nonnormal. "Large" sample sizes can be as small as 30 per group if the two populations are roughly normally distributed. The more the populations depart from normality, the larger the sample size needed for the Central Limit Theorem to weave its magic, but we've seen examples to suggest that 100 observations per group is often quite sufficient.

For small and moderate sample sizes, the *equal variances* version of the test provides an exact test of the equality of the two population means. The validity of the test demands that the samples be drawn from normally distributed populations with equal (population) standard deviations. Just as one reflexively asks about randomization, blinding, and controls when evaluating a study design, it should become second-nature to ask about normality and equal variances when preparing to use Student's t test.

Formal analysis and simulations offer the following guidelines describing extent to which the assumptions of normality and equal population variances be violated without affecting the validity of Student's test for independent samples. [see Rupert Miller, Jr., (1986) *Beyond ANOVA, Basics of Applied Statistics*, New York: John Wiley & Sons]

- If sample sizes are equal, (a) nonnormality is not a problem and (b) the t test can tolerate population standard deviation ratios of 2 without showing any major ill effect. (For equal sample sizes, the two test statistics are equal.) The worst situation occurs when one sample has both a much larger variance and a much smaller sample size than the other. For example, if the variance ratio is 5 and the sample size ratio is 1/5, a nominal P value of 0.05 is actually 0.22.
- Serious distortion of the P value can occur when the skewness of the two populations is different.
- Outliers can distort the mean difference and the t statistic. They tend to inflate the variance and depress the value and corresponding statistical significance of the t statistic.

Preliminary tests for normality and equality of variances--using Student's t test only if these preliminary tests fail to achieve statistical significance--should be avoided. These preliminary tests often detect differences too small to affect Student's t test. Since the test is such a convenient way to compare two populations, it should not be abandoned without good cause. Important violations of the requirements will be detectable to the naked eye without a formal significance test.

What should be done if the conditions for the validity of Student's t test are violated? The best approach is to transform the data to a scale in which the conditions are satisfied. This will almost always involve a logarithmic transformation. On rare occasions, a square root, inverse, or inverse square root might be used. For proportions, arcsin(sqrt(p)) or log(p/(1-p)) might be used. If no satisfactory transformation can be found, a nonparametric test such as the median test or the Wilcoxon-Mann-Whitney test might be used.

The major advantage of transformations is that they make it possible to use standard techniques to construct confidence intervals for estimating between-group differences. In theory, it is possible to construct confidence intervals (for the diffference in medians, say) when rank tests are used. However, we are prisoners of our software. Programs that construct these confidence intervals are not readily available.

Gerard E. Dallal
Last modified: undefined.

P Values

To understand **P values**, you have to understand **fixed level testing**. With fixed level testing, a null hypothesis is proposed (usually, specifying no treatment effect) along with a level for the test, usually 0.05. All possible outcomes of the experiment are listed in order to identify extreme outcomes that would occur less than 5% of the time in aggregate if the null hypothesis were true. This set of values is known as the **critical region**. They are *critical* because if any of them are observed, something extreme has occurred. Data are now collected and if any one of those extreme outcomes occur the results are said to be *significant at the 0.05 level*. The null hypothesis is rejected at the 0.05 level of significance and one star (*) is printed somewhere in a table. Some investigators note extreme outcomes that would occur less than 1% of the time and print two stars (**) if any of those are observed.

The procedure is known as fixed level testing because the level of the test is fixed prior to data collection. In theory if not in practice, the procedure begins by the specifying the hypothesis to be tested and the test statistic to be used along with the set of outcomes that will cause the hypothesis to be rejected. Only then are data collected to see whether they lead to rejection of the null hypothesis.

Many researchers quickly realized the limitations of reporting only whether a result achieved the 0.05 level of significance. Was a result just barely significant or wildly so? Would data that were significant at the 0.05 level be significant at the 0.01 level? At the 0.001 level? Even if the result are wildly statistically significant, is the effect large enough to be of any *practical* importance?

As computers became readily available, it became common practice to report the **observed significance level** (or **P value**)**--the smallest fixed level at which the the null hypothesis can be rejected**. If your personal fixed level is greater than or equal to the P value, you would reject the null hypothesis. If your personal fixed level is less than to the P value, you would fail to reject the null hypothesis. For example, if a P value is 0.027, the results are significant for all fixed levels greater than 0.027 (such as 0.05) and not significant for all fixed levels less than 0.027 (such as 0.01). A person who uses the 0.05 level would reject the null hypothesis while a person who uses the 0.01 level would fail to reject it.

A P value is often described as the probability of seeing results as or more extreme as those actually observed if the null hypothesis were true. While this description is correct, it invites the question of why we should be concerned with the probability of events that have not occurred! (As Harold Jeffreys quipped, "What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.") In fact, we care because the P value is just another way to describe the results of fixed level tests.

Every so often, a call is made for a ban on significance tests. Papers and books are written, conferences are held, and proceedings are published. The main reason behind these movements this is that P values tell us nothing about the magnitudes of the effects that might lead to us to reject or fail to reject the null hypothesis. Significance tests blur the distinction between *statistical significance* and *practical importance*. It is possible for a difference of little practical importance to achieve a high degree of statistical significance. It is also possible for clinically important differences to be missed because an experiment lacks the power to detect them. However, significance tests provide a useful summary of the data and these concerns are easily remedied by supplementing significance tests with the appropriate confidence intervals for the effects of interest.

When hypotheses of equal population means are tested, determining whether P is less than 0.05 is just another way of examining a confidence interval for the mean difference to see whether it excludes 0. The hypothesis of equality will be rejected at level α if and only if a 100 (1-α)% confidence interval for the mean difference fails to contain 0. For example, the hypothesis of equality of population means will be rejected at the 0.05 level if and only if a 95% CI for the mean difference does not contain 0. The hypothesis will be rejected at the 0.01 level if and only if a 99% CI does not contain 0, and so on.

Case	$\bar{x}_1 - \bar{x}_2$	SE	t	P	P<0.05	95% CI	Practical Importance
1	2	0.5	4	<0.0001	Y	(1,3)	N
2	30	5	6	<0.0001	Y	(20,40)	Y
3	30	14	2.1	0.032	Y	(2,58)	?
4	1	1	1	0.317	N	(-1,3)	N
5	2	30	0.1	0.947	N	(-58,62)	?
6	30	16	1.9	0.061	N	(-2,62)	?

This is a good time to revist the cholesterol studies presented during the discussion of confidence intervals. We assumed a treatment mean difference of a couple of units (mg/dl) was of no consequence, but differences of 10 mg/dl and up had important public health and policy implications. The discussion and interpretation of the 6 cases remains the same, except that we can add the phrase *statistically significant* to describe the cases where the P values are less than 0.05.

Significance tests can tell us whether a difference between sample means is statistically significant, that is, whether the observed difference is larger than would be due to random variation if the underlying population difference were 0. But significance tests do not tell us whether the difference is of practical importance. *Statistical significance* and *practical importance* are distinct concepts.

In cases 1-3, the data are judged inconsistent with a population mean difference of 0. The P values are less than 0.05 and the 95% confidence intervals do not contain 0. The sample mean difference is much larger than can be explained by random variability about a population mean difference of 0. In cases 4-6, the data are consistent with a population mean difference of 0. The P values are greater than 0.05 and the 95% confidence intervals contain 0. The observed difference is consistent with random variability about 0.

Case 1: There is a statistically significant difference between the diets, but the difference is of no practical importance, being no greater than 3 mg/dl.

Case 2: The difference is statistically significant and is of practical importance even though the confidence interval is 20 mg/dl wide. This case illustrates that a wide confidence interval is not necessarily a bad thing, if all of the values point to the same conclusion. Diet 2 is clearly superior to diet 1, even though we the likely benefit can't be specified to within a range of 20 mg/dl.

Case 3: The difference is statistically significant but it may or may not be of practical importance. The confidence interval is too wide to say for sure. The difference may be as little as 2 mg/dl, but could be as great as 58 mg/dl. More study may be needed. However, knowledge of a difference between the diets, regardless of its magnitude, may lead to research that exploits and enhances the beneficial effects of the more healthful diet.

Case 4: The difference is not statistically significant **and** we are confident that if there is a real difference it is of no practical importance.

Cases 5 and 6: The difference is not statistically significant, so we cannot claim to have demonstrated a difference. However, the population mean difference is not well enough determined to rule out all differences of practical importance.

Cases 5 and 6 require careful handling. Case 6, unlike Case 5, seems to rule out any advantage of practical importance for Diet 1, so it might be argued that Case 6 is like Case 3 in that both of them are consistent with important and unimportant advantages for Diet 2 while neither suggests any advantage to Diet 1.

Many analysts accept illustrations such as these as a blanket indictment of significance tests. I prefer to see them as a warning to continue beyond significance tests to see what other information is contained in the data. In some situations, it's important to know if there is an effect no matter how slight, but in most cases it's hard to justify publishing the results of a significance test without saying something about the magnitude of the effect*. If a result is statistically significant, is it of practical importance? If the result is not statistically significant, have effects of practical importance been ruled out? If a result is not statistically significant but has not ruled out effects of practical importance, YOU HAVEN'T LEARNED ANYTHING!

Case 5 deserves another visit in order to underscore an important lesson that is usually not appreciated the first time 'round: "Absence of evidence is not evidence of absence!" In case 5, the observed difference is 2 mg/dl, the value 0 is nearly at the center of the confidence interval, and the P value for testing the equality of means is 0.947. It is correct to say that the difference between the two diets did not reach statistical significance or that no statistically significant difference was shown. Some researchers refer to such findings as "negative", yet, it would be incorrect to say that the diets are the same. The absence of evidence for a difference is not the same thing as evidence of absence of an effect. In BMJ,290(1985),1002, Chalmers proposed outlawing the term "negative trial" for just this reason.

When the investigator would like to conjecture about the absence of an effect, the most effective procedure is to report confidence intervals so that readers have a feel for the sensitivity of the experiment. In cases 4 and 5, the researchers are entitled to say that there was no significant finding. Both have P values much larger than 0.05. However, only in case 4 is the researcher entitled to say that the two diets are equivalent: the best available evidence is that they produce mean cholesterol values within 3 mg/dl of each other, which is probably too small to worry about. One can only hope that a claim of no difference based on data such as in case 5 would never see publication.

Should P values be eliminated from the research literature in favor of confidence intervals? This discussion provides some support for this proposal, but there are many situations were the magnitude of an effect is not as important as whether or not an effect is present. I have no objection to using P values to focus on the presence or absence of an effect, provided the confidence intervals are available for those who want them, statistical significance is not mistaken for practical importance, and absence of evidence is not mistaken for evidence of absence.

As useful as confidence intervals are, they are not a cure-all. They offer estimates of the effects they measure, but only in the context in which the data were collected. It would not be surprising to see confidence intervals vary between studies much more than any one interval would suggest. This can be the result of the technician, measurement technique, or the particular group of subjects being measured, among other causes. This is one of the things that plagues meta-analysis, even in medicine where the outcomes are supposedly well-defined. This is yet another reason why significance tests are useful. There are many situations where the most useful piece of information that a confidence interval provides is simply that there is an effect or treatment difference.

What P values are not!

A P value is the probability of observing data as or more extreme as the actual outcome when the null hypothesis is true. A small P value means that data as extreme as these are unlikely under the null hypothesis. The P value is NOT the probability that the null hypothesis is true. A small P value makes us reject the null hypothesis because an event has occurred that is unlikely if H_0 were true.

Classical (or frequentist) statistics does not allow us to talk about the probability that a hypothesis is true. Statements such as, "There's a 5 percent chance that these two diets are equally effective at lowering cholesterol" have no meaning in classical statistics. Either they are equally effective or they aren't. All we can talk about is the probability of seeing certain outcomes *if* the hypothesis were true**.

The reason these methods work regardless is that, although we haven't said so explicitly, there is a tacit presumption that the alternative hypothesis provides a more reasonable explanation for the data. However, it's not built into the methods, and need not be true. It is possible to reject a hypothesis even though it is the best explanation for the data, as the following two examples illustrate.

Example 1: A single value is observed from a normal distribution with a standard deviation of 1. Suppose there are only two possibilities: Either the population mean is 0 or it is 100. Let H_0 be μ =0 and H_1 be μ=100. Suppose a value of 3.8 is observed. The P value is 0.0001 because, if the population mean is 0, the probability of observing an observation as or more extreme than 3.8 is 0.0001. We have every right to reject H_0 at the 0.05, 0.01, or even the 0.001 level of significance. However, the probability of observing 3.8 is even less under the alternative hypothesis! Even though we can reject H_0 at the usual levels of significance, common sense says that the null hypothesis is more likely to be true than the alternative hypothesis.

Example 2: Suppose only 35 heads occur in 100 flips of a coin. The P value for testing the null hypothesis that the coin is fair (equally likely to come up heads or tails) versus the alternative that is it unfair is 0.0035. We can reject the hypothesis that the coin is fair at the 0.01 level of significance, but does this mean that there is less than a 1-% chance that the coin is fair? It depends on things other than the number of heads and tails. If the coin were a gambling device belonging to someone else and it was causing you to lose money, you might think it highly unlikely that the coin was fair. However, if the coin was taken from a roll of newly minted coins just delivered to your bank and you did the flipping yourself by letting the coin bounce off some soft surface (to foil any possible regularity in your flipping motion), you might still find it quite likely that the coin is fair. Standard statistical theory cannot answer this question.

[*] I was asked recently why confidence intervals were common in the medical literature but not in other fields. My immediate, tongue-partially-in-cheek response was that for a confidence interval to be useful, you had to have some idea of what it meant! Many areas of investigation summarize their experiments in scales and indices that often lack an operational interpretation. Some scales are the sum of positive responses to items on a questionnaire. Others are composites of related but different components. Those scoring higher are different from those scoring lower, but it's often not clear what a 1 or 10 unit difference means in *any* sense, let alone in terms of practical importance.

[**] While standard frequentist methods cannot answer the question, another approach to statistics--*Bayesian methods*--attempts to provide an answer. If prior to flipping the coin, you could quantify the probability that the coin is fair, Bayesian methods provide a way to update this probability after the coin is flipped. The trick is in coming up with the inital probability. For example, before flipping the coin, what *is* the probability that the coin is fair?

[back to The Little Handbook of Statistical Practice]

Last modified: undefined.

Why P=0.05?

The standard level of significance used to justify a claim of a statistically significant effect is 0.05. For better or worse, the term *statistically significant* has become synonymous with $P \leq 0.05$.

There are many theories and stories to account for the use of P=0.05 to denote statistical significance. All of them trace the practice back to the influence of R.A. Fisher. In 1914, Karl Pearson published his *Tables for Statisticians & Biometricians*. For each distribution, Pearson gave the value of P for a series of values of the random variable. When Fisher published *Statistical Methods for Research Workers* (SMRW) in 1925, he included tables that gave the value of the random variable for specially selected values of P. SMRW was a major influence through the 1950s. The same approach was taken for Fisher's *Statistical Tables for Biological, Agricultural, and Medical Research*, published in 1938 with Frank Yates. Even today, Fisher's tables are widely reproduced in standard statistical texts.

Fisher's tables were compact. Where Pearson described a distribution in detail, Fisher summarized it in a single line in one of his tables making them more suitable for inclusion in standard reference works[*]. However, Fisher's tables would change the way the information could be used. While Pearson's tables provide probabilities for a wide range of values of a statistic, Fisher's tables only bracket the probabilities between coarse bounds.

The impact of Fisher's tables was profound. Through the 1960s, it was standard practice in many fields to report summaries with one star attached to indicate $P \leq 0.05$ and two stars to indicate $P \leq 0.01$, Occasionally, three starts were used to indicate $P \leq 0.001$.

Still, why should the value 0.05 be adopted as the universally accepted value for statistical significance? Why has this approach to hypothesis testing not been supplanted in the intervening three-quarters of a century?

It was Fisher who suggested giving 0.05 its special status. Page 44 of the 13th edition of SMRW, describing the standard normal distribution, states

> The value for which P=0.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.

Similar remarks can be found in Fisher (1926, 504).

> ... it is convenient to draw the line at about the level at which we can say: "Either there is

> something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials."...
>
> If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.

However, Fisher's writings might be described as inconsistent. On page 80 of SMRW, he offers a more flexible approach

> In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. Belief in the hypothesis as an accurate representation of the population sampled is confronted by the logical disjunction: *Either* the hypothesis is untrue, *or* the value of χ^2 has attained by chance an exceptionally high value. The actual value of P obtainable from the table by interpolation indicates the strength of the evidence against the hypothesis. A value of χ^2 exceeding the 5 per cent. point is seldom to be disregarded.

These apparent inconsistencies persist when Fisher dealt with specific examples. On page 137 of SMRW, Fisher suggests that values of P slightly less than 0.05 are are not conclusive.

> [T]he results of *t* shows that P is between .02 and .05.
>
> The result must be judged significant, though barely so; in view of the data we cannot ignore the possibility that on this field, and in conjunction with the other manures used, nitrate of soda has conserved the fertility better than sulphate of ammonia; the data do not, however, demonstrate this point beyond the possibility of doubt.

On pages 139-140 of SMRW, Fisher dismisses a value greater than 0.05 but less than 0.10.

> [W]e find...*t*=1.844 [with 13 df, P = 0.088]. The difference between the regression coefficients, though relatively large, cannot be regarded as significant. There is not sufficient evidence to assert that culture B was growing more rapidly than culture A.

while in Fisher [19xx, p 516] he is willing pay attention to a value not much different.

> ...P=.089. Thus a larger value of χ^2 would be obtained by chance only 8.9 times in a hundred, from a series of values in random order. There is thus some reason to suspect that the distribution of rainfall in successive years is not wholly fortuitous, but that some slowly changing cause is liable to affect in the same direction the rainfall of a number of consecutive years.

Yet *in the same paper* another such value is dismissed!

> [paper 37, p 535] ...P=.093 from Elderton's Table, showing that although there are signs of association among the rainfall distribution values, such association, if it exists, is not strong enough to show up significantly in a series of about 60 values.

Part of the reason for the apparent inconsistency is the way Fisher viewed P values. When Neyman and Pearson proposed using P values as absolute cutoffs in their style of fixed-level testing, Fisher disagreed strenuously. Fisher viewed P values more as measures of the evidence against a hypotheses, as reflected in the quotation from page 80 of SMRW above and this one from Fisher (1956, p 41-42)

> The attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to hypothetical frequencies of possible statements, based on them, being right or wrong, thus seem to miss the essential nature of such tests. A man who "rejects" a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection. This inequality statement can therefore be made. However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. Further, the calculation is based solely on a hypothesis, which, in the light of the evidence, is often not believed to be true at all, so that the actual probability of erroneous decision, supposing such a phrase to have any meaning, may be much less than the frequency specifying the level of significance.

Still, we continue to use P values nearly as absolute cutoffs but with an eye on rethinking our position for values close to 0.05**. Why have we continued doing things this way? A procedure such as this has an important function as a gatekeeper and filter--it lets signals pass while keeping the noise down. The 0.05 level guarantees the literature will be spared 95% of potential reports of effects where there are none.

For such procedures to be effective, it is essential ther be a tacit agreement among researchers to use them in the same way. Otherwise, individuals would modify the procedure to suit their own purposes until the procedure became valueless. As Bross (1971) remarks,

> Anyone familiar with certain areas of the scientific literature will be well aware of the need for curtailing language-games. Thus if there were no 5% level firmly established, then some persons would stretch the level to 6% or 7% to prove their point. Soon others would be stretching to 10% and 15% and the jargon would become meaningless. Whereas nowadays a phrase such as *statistically significant difference* provides some assurance that the results are not merely a manifestation of sampling variation, the phrase would mean very little if everyone played language-games. To be sure, there are always a few folks who fiddle with significance levels--who will switch from two-tailed to one-tailed tests or from one significance test to another in an effort to get positive results. However such gamesmanship is severely frowned upon and is rarely practiced by persons who are *native speakers* of fact-limited scientific languages--it is the mark of an amateur.

Bross points out that the continued use of P=0.05 as a convention tells us a good deal about its practical value.

> The continuing usage of the 5% level is indicative of another important practical point: it is a feasible level at which to do research work. In other words, if the 5% level is used, then in most experimental situations it is feasible (though not necessarily easy) to set up a study which will have a fair chance of picking up those effects which are large enough to be of scientific interest. If past experience in actual applications had not shown this feasibility, the convention would not have been useful to scientists and it would not have stayed in their languages. For suppose that the 0.1% level had been proposed. This level is rarely attainable in biomedical experimentation. If it were made a prerequisite for reporting positive results, there would be very little to report. Hence from the standpoint of communication the level would have been of little value and the evolutionary process would have eliminated it.

The fact that many aspects of statistical practice in this regard *have* changed gives Bross's argument additional weight. Once (mainframe) computers became available and it was possible to calculate precise P values on demand, standard practice quickly shifted to reporting the P values themselves rather than merely whether or not they were less than 0.05. The value of 0.02 suggested by Fisher as a *strong indication that the hypothesis fails to account for the whole of the facts has been replaced by 0.01. However, science has seen fit to continue letting 0.05 retain its special status denoting statistical significance.*

[*] Fisher may have had additional reasons for developing a new way to table commonly used distribution functions. Jack Good, on page 513 of the discussion section of Bross (1971), says, "Kendall mentioned that Fisher produced the tables of significance levels to save space and to avoid copyright problems with Karl Pearson, whom he disliked."

[**] It is worth noting that when researchers worry about P values close to 0.05, they worry about values slightly greater than 0.05 and why they deserve attention nonetheless. I cannot recall published research downplaying P values less than 0.05. Fisher's comment cited above from page 137 of SMRW is a rare exception.

References

- *Bross IDJ (1971), "Critical Levels, Statistical Language and Scientific Inference," in Godambe VP and Sprott (eds) Foundations of Statistical Inference. Toronto: Holt, Rinehart & Winston of Canada, Ltd.*
- Fisher RA (1956), *Statistical Methods and Scientific Inference* New York: Hafner
- Fisher RA (1926), "The Arrangement of Field Experiments," Journal of the Ministry of Agriculture of Great Britain, 33, 503-513.
- Fisher RA (19xx), "On the Influence of Rainfall on the Yield of Wheat at Rothamstead,"

Gerard E. Dallal
Last modified: undefined.

20 Independent 0.05 Level Tests For An Effect Where None Is Present

Test! / Clear

The chance that nothing is significant is only 0.3585,
so don't give up hope!

Last modified: undefined.

One Sided Tests

When comparing population means or proportions, tests of the null hypothesis that two population means or proportions are equal

$$H_0\text{: } \mu_1 = \mu_2 \text{ (that is, } \mu_1 - \mu_2 = 0)$$

are almost always two-sided (or two-tailed*). That is, the alternative hypothesis is

$$H_1\text{: } \mu_1 \neq \mu_2 \text{ (that is, } \mu_1 - \mu_2 \neq 0)$$

To make the discussion more concrete, suppose the comparison involves two treatment means. While we're at it, let's call the treatments *N* and *S* and let's suppose that small values are good. The null hypothesis of equal effectiveness is then

$$H_0\text{: } \mu_N = \mu_S \text{ (that is, } \mu_N - \mu_S = 0)$$

and the alternative is

$$H_1\text{: } \mu_N \neq \mu_S \text{ (that is, } \mu_N - \mu_S \neq 0)$$

One criticism of significance tests is that no null hypothesis is ever true. For example, the claim is made that two population means are always unequal as long as our measurements have enough decimal places. Then why should we bother testing whether two population means are equal? While there may be some truth to the criticism, one interpretation of the alternative hypothesis is that it is says we are unsure of the *direction* of the difference.

Every so often, someone claims that if there is a difference it can be in only one direction. For example, an investigator might claim that a newly proposed treatment *N* might possibly prove no more effecive than standard treatment *S* but it cannot be harmful. One-sided tests have been proposed for such circumstances. The alternative hypothsis states that the difference, if any, can be in only one direction

$$H_{1b}\text{: } \mu_N < \mu_S \text{ (that is, } \mu_N - \mu_S < 0)$$

For example, an investigator might propose using a one-tailed test to test the efficacy of a cholesterol lowering drug because the drug cannot raise cholesterol. Under a one-tailed test, the hypothesis of no difference is rejected if and only if the subjects taking the drug have cholesterol levels significantly lower than those of controls. All other outcomes are treated as failing to show a difference.

One-tailed tests make it easier to reject the null hypothesis when the alternative is true. A large sample, two-sided, 0.05 level t test needs a t statistic less than -1.96 to reject the null hypothesis of no difference in means. A one-sided test rejects the hypothesis for values of t less than -1.645. Therefore, a one-sided test is more likely likely to reject the null hypothesis when the difference is in the expected direction. This makes one-sided tests very attractive to those definition of success is having a statistically significant result.

What damns one-tailed tests in the eyes of most statisticians is the demand that *all* differences in the unexpected direction--large and small--be treated the same as simply nonsignificant. I have never seen a situation where the researchers were willing to do this in practice. In practice, things can *always* get worse! Suppose subjects taking the new cholesterol lowering drug ended up with levels 50 ± 10 mg/dl *higher* than those of the control group. The use of a one-tailed test implies that the researchers would pursue this no further. However, we know they would immediately begin looking for an underlying cause and question why the drug was considered for human intervention trials.

A case in point is the Finnish Alpha-Tocopherol, Beta-Carotene Cancer Prevention Trial ("The Effect Of Vitamin E and Beta-Carotene on the Incidence of Lung Cancer and other Cancers in Male Smokers" N Engl J Med 1994;330:1029-35). 18% more lung cancers were diagnosed and 8% more overall deaths occurred in study participants taking beta carotene. If a one-sided analysis had been proposed for the trial, these results would have been ignored on the grounds that they were the result of unlikely random variability under a hypothesis of no difference between beta- carotene and placebo. When the results of the trial were first reported, this was suggested as one of the many possible reasons for the anomolous outcome. However, after these results were reported, investigators conducting the Beta Carotene and Retinol Efficacy Trial (CARET), a large study of the combination of beta carotene and vitamin A as preventive agents for lung cancer in high-risk men and women, terminated the intervention after an average of four years of treatment and told the 18,314 participants to stop taking their vitamins. Interim study results indicate that the supplements provide no benefit and may be causing harm. 28% more lung cancers were diagnosed and 17% more deaths occurred in participants taking beta carotene and vitamin A than in those taking placebos. Thus, the CARET study replicated the ATBC findings. More details can be found at this NIH fact sheet and this one, too.

The usual 0.05 level two-tailed test puts half of the probabilty (2.5%) in each tail of the reference distribution, that is, the cutoff points for the t statistic are ±1.96. Some analysts have proposed two-sided tests with unequal tail areas. Instead of having 2.5% in each tail, there might be 4% in the expected direction and 1% in the other tail (for example, cutoffs of -1.75 and 2.33) as insurance against extreme results in the unexpected direction. However, there is no consensus or obvious choice for the way to divide the probability (e.g., 0.005/0.045, 0.01/0.04, 0.02/0.03) and some outcomes might give the false impression that the split was chosen after the fact to insure statistical signifcance. This leads us back to the usual two-tailed test (0.025, 0.025).

Marvin Zelen dismisses one-sided tests in another way--he finds them unethical! His argument is as simple as it is elegant. Put in terms of comparing a new treatment to standard, anyone who insists on a one- tailed test is saying the new treatment can't do worse than the standard. If the new treament has any effect, it can only do better. The method of analysis should be part of a study's design. If investigators are prepared to justify the use of a one-tailed test at the start of the study, then it is unethical not to give the new treatment to everyone!

----

*Some statisticians find the word *tails* to be ambiguous and use *sided* instead. *Tails* refers to the distribution of the test statistic and there can be many test statistics. While the most familiar test statistic might lead to a two-tailed test, other statistics might not. When the hypothesis H_0: $\mu_1 = \mu_2$ is tested against the alternative of inequality, it is rejected for large positive values of t (which lie in the upper tail) and large negative values of t (which lie in the lower tail). However, this test can also be performed by using the square of the t or z statistics ($t^2 = F_{1,n}$; $z^2 = \chi^2_1$). Then only large values of the test statistic will lead to rejecting the null hypothesis. Since only one tail of the reference distribution leads to rejection, it is a one-*tailed* test.

Side refers to the hypothesis, namely, on which the side of 0 the difference $\mu_1 - \mu_2$ lies (positive or negative). Since this is a statement about the hypothesis, it is independent of the choice of test statistic. Nevertheless, the terms *two-tailed* and *two-sided* are often used interchangeably.

[back to The Little Handbook of Statistical Practice]

Last modified: undefined.

Contingency Tables

Gerard E. Dallal, Ph.D.

A **contingency table** is a table of counts. A two-dimensional contingency table is formed by classifying subjects by two variables. One variable determines the row categories; the other variable defines the column categories. The combinations of row and column categories are called *cells*. Examples include classifying subjects by sex (male/female) and smoking status (current/former/never) or by "type of prenatal care" and "whether the birth required a neonatal ICU" (yes/no). For the mathematician, a two-dimensional contingency table with *r* rows and *c* columns is the set $\{x_{ij}: i=1,...,r; j=1,...,c\}$.

In order to use the statistical methods usually applied to such tables, subjects must fall into one and only one row and column categories. Such categories are said to be **exclusive** and **exhaustive**. **Exclusive** means the categories don't overlap, so a subject falls into only one category. **Exhaustive** means that the categories include all possibilities, so there's a category for everyone. Often, categories can be made exhaustive by creating a catch-all such as "Other" or by changing the definition of those being studied to include only the available categories.

Also, the observations must be independent. This can be a problem when, for example, families are studied, because members of the same family are more similar than individuals from different families. The analysis of such data is beyond the current scope of these notes.

Textbooks often devoting a chapter or two to the comparison of two proportions (the percentage of high school males and females with eating disorders, for example) by using techniques that are similar to those for comparing two means. However, two proportions can be represented by a 2-by-2 contingency table in which one of the classification variables defines the groups (male/female) and the other is the presence or absence of the characteristic (eating disorder), so standard contingency table analyses can be used, instead.

When plots are made from two continuous variables where one is an obvious response to the other (for example, cholesterol level as a response to saturated fat intake), standard practice is to put the response (cholesterol) on the vertical (Y) axis and the carrier (fat intake) on the horizontal (X) axis. For tables of counts, it is becoming common practice for the row categories to specify the populations or groups and the column categories to specify the responses. For example, in studying the association between smoking and disease, the rows categories would be the categories of smoking status while the columns would denote the presence or absence of disease. This is in keeping with A.S.C. Ehernberg's observation that it is easier to make a visual comparison of values in the same column than in the same row. Consider

		Disease		Disease		
		Yes	No	Yes	No	
Smoke	Yes	13	37	26%	74%	100%
	No	6	144	4%	96%	100%
		(A)		(B)		

(In table A the entries are counts;
in table B the entries are percentages within each row.)

The 26 and 4% are easy to compare because they are lined up in the same column.

Sampling Schemes

There are many ways to generate tables of counts. Three of the most common sampling schemes are

Unrestricted (Poisson) sampling: Collect data until the sun sets, the money runs out, fatigue sets in,...

Sampling with the grand total fixed (multinomial sampling): Collect data on a predetermined number of individuals and classify them according to the two classification variables.

Sampling with one set of marginal totals fixed (compound multinomial sampling): Collect data on a predetermined number of individuals from each category of one of the variables and classify them according to the other variable. This approach is useful when some of the categories are rare and might not be adequately represented if the sampling were unrestricted or only the grand total were fixed. For example, suppose you wished to assess the association between tobacco use and a rare disease. It would be better to take fixed numbers of subjects with and without the disease and examine them for tobacco use. If you sampled a large number of individuals and classified them with respect to smoking and disease, there might be too few subjects with the disease to draw any meaningful conclusions*.

Each sampling scheme results in a table of counts. It is impossible to determine which sampling scheme was used merely by looking at the data. Yet, the sampling scheme is important because some things easily estimated from one scheme are impossible to estimate from the others. The more that is specified by the sampling scheme, the fewer things that can be estimated from the data. For example, consider the 2 by 2 table

		Eating Disorder	
		Yes	No
College:	Public		
	Private		

If sampling occurs with only the grand total fixed, then any population proportion of interest can be estimated. For example, we can estimate the population proportion of individuals with eating disorders, the proportion attending public colleges, the proportion attending public college and are without eating disorder, and so on.

Suppose, due to the rarity of eating disorders, 50 individuals with eating disorders and 50 individuals without eating disorders are studied. Many population proportions can no longer be estimated from the data. It's hardly surprising we can't estimate the proportion of the population with eating disorders. If we choose to look at 50 individuals with eating disorders and 50 without, we obviously shouldn't be able to estimate the population proportion that suffers from eating disorders. The proportion with eating disorders in our sample will be 50%, not because 50% of the population have eating disorders but because we specifically chose a sample in which 50% have eating disorders.

Is it as obvious that we cannot estimate the proportion of the population that attends private colleges? We cannot if there is an association between eating disorder and type of college. Suppose students with eating disorders are more likely to attend private colleges than those without eating disorders. Then, the proportion of students attending a private college in the combined sample will change according to the way the sampling scheme fixes the proportions of students with and without an eating disorder.

Even though the sampling scheme affects what we can estimate, all three sampling schemes use the same test statistic and reference distribution to decide whether there is an association between the row and column variables. However, the name of the problem changes according to the sampling scheme.

When the sampling is unrestricted or when only the grand total is fixed, the hypothesis of no association is called **independence** (of the row and column variables)--the probability of falling into a particular column is independent of the row. It does not change with the row a subject is in. Also, the probability of falling into a particular row does not depend on the column the subject is in.

If the row and column variables are independent, the probability of falling into a particular *cell* is the product of *the probability of being in a particular row* and *the probability of being in a particular column*. For example, if 2/5 of the population attends private colleges and, independently, 1/10 of the population has an eating disorder, then 1/10 of the 2/5 of the population that attends private colleges should suffer from eating disorders, that is, 2/50 (= 1/10 × 2/5) attend private college and suffer from eating disorders.

When one set of marginal totals--the rows, say--is fixed by the sampling scheme, the hypothesis of no association is called **homogeneity of proportions**. It says the proportion of individuals in a particular column the same for all rows.

The *chi-square statistic*, χ^2, is used to test both null hypotheses ("independence" and "homogeneity of proportions"). It is also known as the *goodness-of-fit statistic* or *Pearson's goodness-of-fit statistic*. The test is known as the *chi-square test* or the *goodness-of-fit test*.

Let the observed cell counts be denoted by {x_{ij}: i=1,...,r; j=1,...,c} and the expected cell counts under a model of independence or homogeneity of proportions be denoted by {e_{ij}: i=1,...,r; j=1,...,c}. The test statistic is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(x_{ij}-e_{ij})^2}{e_{ij}} = \sum_{\mathit{all\ cells}}\frac{(\text{observed-expected})^2}{\text{expected}},$$

where the expected cell counts are given by

$$\frac{(\mathit{row\ total})\times(\mathit{column\ total})}{\mathit{grand\ total}}$$

The derivation of the expression for expected values is straightforward. Consider the cell at row 1, column 1. Under a null hypothesis of homogeneity of proportions, say, both rows have the same probability that an observation falls in column 1. The best estimate of this common probability is

$$\frac{(column\ 1\ total)}{grand\ total}$$

Then, the expected number of observations in the cell at row 1, column 1 is the number of observations in the first row (row 1 total) multiplied by this probability, that is,

$$\frac{(row\ 1\ total) \times (column\ 1\ total)}{grand\ total}$$

In the chi-square statistic, the square of the difference between observed and expected cell counts is divided by the expected cell count. This is because probability theory shows that cells with large expected counts vary more than cells with small expected cell counts. Hence, a difference in a cell with a larger expected cell count should be downweighted to account for this.

The chi-square statistic is compared to the percentiles of a chi-square distribution. The chi-square distributions are like the t distributions in that there are many of them, indexed by their degrees of freedom. For the goodness-of-fit statistics, the degrees of freedom equal the product of (the number of rows - 1) and (the number of columns - 1), or **(r-1)(c-1)**. When there are two rows and two columns, the degrees of freedom is 1. Any disagreement between the observed and expected values will result in a large value of the chi-square statistic, because the test statistic is the sum of the squared differences. The null hypothesis of independence or homogeneity of proportions is rejected for large values of the test statistic.

Tests of Significance

Three tests have been suggested for testing the null hypotheses of independence or homogeneity of proportions. Pearson's goodness-of-fit test, the goodness-of-fit test with Yates's continuity correction, and Fisher's exact test.

Pearson's Goodness-of-Fit Test

We just discussed Pearson's goodness of fit statistic.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(x_{ij}-e_{ij})^2}{e_{ij}} = \sum_{all\ cells}\frac{(\text{observed-expected})^2}{\text{expected}}$$

The way it is typically used--compared to percentiles of the chi-square distribution with (r-1)(c-1) degrees of freedom--is based on large sample theory. Many recommendations for what constitutes a large sample can be found in the statistical literature. The most conservative recommendation says all expected cell counts should be 5 or more. Cochran recommends that at least 80% of the expected cell count be 5 or more and than no expected cell count be less than 1. For a two-by-two table, which has only four cells, Cochran's recommendation is the same as the "all expected cell counts should be 5 or more" rule.

Fisher's Exact Test

Throughout the 20th century, statisticians argued over the best way to analyze contingency tables. As with other test procedures, mathematics is use to decide whether the observed contingency table is in some sense extreme. The debate, which is still not fully resolved, has to do with what set of tables to use. For example, when multinomial sampling is used, it might seem obvious that the set should include all possible tables with the same total sample size. However, today most statisticians agree that the set should include only those tables with the same row and column totals as the observed table, regardless of the sampling scheme that was used. (Mathematical statisticians refer to this as performing the test conditional on the margins, that is, the table's marginal totals.)

This procedure is known as Fisher's Exact Test. All tables with the same row and column totals have their probability of occurrence calculated according to a probability distribution known as *the hypergeometric distribution*. For example, if the table

```
1  3 | 4
4  3 | 7
------
5  6
```

were observed, Fisher's exact test would look at the set of all tables that have row totals of (4,7) and column totals of (5,6). They are

```
          0 4    |    1 3    |    2 2    |    3 1
|    4 0
```

```
                5 2    |    4 3    |    3 4    |    2 5
|    1 6

probability  21/462     140/462     210/462      84/462
7/462
```

While it would not be profitable to go into a detailed explanation of the hypergeometric distribution, it's useful to remove some of the mystery surrounding it. That's more easily done when the table has labels, so lets recast the table in the context of discrimination in the workplace. Suppose there are 11 candidates for 5 partnerships in a law firm. The results of the selection are

	Partner	
	Yes	No
Female	1	3
Male	4	3

Five partners were selected out of 11 candidates--4 of 7 men, but only 1 of 4 women.

The hypergeometric distribution models the partnership process this way. Imagine a box with 11 slips of paper, one for each candidate. *Male* is written on 7 of them while *female* is written on the other 4. If the partnership process is sex-blind, the number of men and women among the new partners should be similar to what would result from drawing 5 slips at random from the box. The hypergeometric distribution gives the probability of drawing specific numbers of males and females when 5 slips are drawn at random from a box containing 7 slips marked *males* and 4 slips marked *females*. Those are the values in the line above labeled "probability".

The calculation of a one-tailed P value begins by ordering the set of all tables with the same margins (according to the value of the cell in the upper right hand corner, say). The probability of observing each table is calculated by using the hypergeometric distribution. Then the probabilities are summed from each end of the list to the observed table. The smaller sum is the one-tailed P value. In this example, the two sums are 21/462+140/462 (=161/462) and 7/462+84/462+210/462+140/462 (=441/462), so the one-tailed P value is 161/462. Yates (1984) argues that a two-tailed P value should be obtained by doubling the one-tailed P value, but most statisticians would compute the two tailed P value as the sum of the probabilities, under the null hypothesis, of all tables having a probability of occurrence no greater than that of the observed table. In this case it is 21/462+140/462+84/462+7/462

(=252/462). And, yes, if the observed table had been (4,0,1,6) the one-sided and two-sided P values would be the same (=7/462).

The Yates Continuity Correction

The Yates continuity correction was designed to make the Pearson chi- square statistic have better agreement with Fisher's Exact test when the sample size is small. The corrected goodness-of-fit statistic is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(\left|x_{ij}-e_{ij}\right|-\frac{1}{2}\right)^2}{e_{ij}} = \sum_{\text{all cells}}\frac{\left(\left|\text{observed-expected}\right|-\frac{1}{2}\right)^2}{\text{expected}}$$

While Pearson's goodness-of-fit test can be applied to tables with any number of rows or columns, the Yates correction applies only to 2 by 2 tables.

There were compelling arguments for using the Yates correction when Fisher's exact test was tedious to do by hand and computer software was unavailable. Today, it is a trivial matter to write a computer program to perform Fisher's exact test for any 2 by 2 table, and there no longer a reason to use the Yates correction.

Advances in statistical theory and computer software (Cytel Software's StatXact, in particular, and versions of their algorithms incorporated into major statistical packages) make it possible to use Fisher's exact test to analyze tables larger than 2 by 2. This was unthinkable 15 years ago. In theory, an exact test could be constructed for any contingency table. In practice, the number of tables that have a given set of margins is so large that the problem would be insoluble for all but smaller sample sizes and the fastest computers. Cyrus Mehta and Nitin Patel, then the Harvard School of Public Health, devised what they called a *network algorithm*, which performs Fisher's exact test on tables larger than 2 by 2. Their technique identifies large sets of tables which will be negligible in the final tally and skips over them during the evaluation process. Thus, they are able to effectively examine all tables when computing their P values by identifying large sets of tables that don't have to be evaluated.

At one time, I almost always used the Yates correction. Many statisticians did not, but the arguments for its use were compelling (Yates, 1984). Today, most computer programs report Fisher's exact test for every 2x2 table, so I use that. For larger tables, I follow Cochran's rule. I use the uncorrected test statistic (Pearson's) for large samples and Fisher's exact test whenever the size of the sample is called into question and available software will allow it.

Example

```
          8  5
          3 10
```

X^2=3.94 (P=0.047). X_c^2=2.52 (P=0.112). Fisher's exact test gives P=0.111.

Summary

Use Fisher's Exact Test whenever the software provides it. Otherwise, follow Cochrans rule. If Cochran's rule is satisfied (no expected cell count is less than 1 and no more than 20% are less than 5), use the uncorrected Pearson goodness-of-fit statistic. If the sample size is called into question, use Fisher's exact test if your software can provide it.

It is straightforward mathematically to show for large samples that P values based on Pearson's goodness-of-fit test and Fisher's exact test are virtually identical. I do not recall a single case where a table satisfied Cochran's rule and the two P values differed in any manner of consequence.

----

[*] This is true from a statistical standpoint, but it is overly simplistic from a practical standpoint. Case-control studies involve sampling fixed numbers of those with and without a disease. The cases (those with the disease) are compared to those without the disease (controls) for the presence of some potential causal factor (exposure). However, it is often the case that there are no sampling frames (lists of individuals) for drawing random samples of those with and without the disease. It has been argued that case-control studies are inherently flawed because of biases between the case and control groups. In order to meet this criticism, it has become common to conduct nested case-control studies in which the cases and controls are extracted truly at random from an identifiable group being studied over time for some other purpose, such as Framingham or the Nurses Health Study. While the generalizability of nested case-control studies might be questioned, they are internally valid because cases and controls were recruited in the same way.

Proportions

This is another way of looking at the content of the *Contingency Tables* page when two-by-two contingency tables are used to compare two proportions. This approach appears in almost every introductory statistics text. It's easily understood, and it shows how the analysis of proportions is nearly the same as the analysis of means, despite the difference in appearance.

[Notation: The obvious notational choice for proportion or probability is *p*. The standard convention is to use Roman letters for sample quantities and the corresponding Greek letter for population quantities. Some books do just that. However, the Greek letter π has its own special place in mathematics. Therefore, instead of using *p*for sample proportion and π for population proportion, many authors use *p* for population proportion and *p* with a hat (caret) on it, $\hat{p}$ (called p-hat), as the sample proportion. The use of "hat" notation for differentiating between sample and population quantities is quite common.]

Confidence Intervals

There's really nothing new to learn to compare two proportions because we know how to compare means. Proportions are just means! The proportion having a particular characteristic is the number of individuals with the characteristic divided by total number of individuals. Suppose we create a variable that equals 1 if the subject has the characteristic and 0 if not. The proportion of individuals with the characteristic is the mean of this variable because the sum of these 0s and 1s is the number of individuals with the characteristic.

While it's never done this way (I don't know why not*), two proportions could be compared by using Student's t test for independent samples with the new 0/1 variable as the response.

An approximate 95-% confidence interval for the difference between two population proportions (p_1-p_2) based on two independent samples of size n_1 and n_2 with sample proportions $\hat{p}_1$ and $\hat{p}_2$ is given by

$$(\hat{p}_1-\hat{p}_2)\pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Even though this looks different from other formulas we've seen, it's nearly identical to the formula for the "equal variances not assumed" version of Student's t test for independent samples. The only difference is that the SDs are calculated with *n* in the denominator instead of *n-1*.

An approximate 95-% confidence interval for a single population proportion based on a sample of size n with sample proportion $\hat{p}$ is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Significance Tests

Comparing Two Proportions

There is a choice of test statistics for testing the null hypothesis H_0: $p_1=p_2$ (the population proportions are equal) against H_1: $p_1 \neq p_2$ (the population proportions are not equal). The test is performed by calculating one of these statistics and comparing its value to the percentiles of the standard normal distribution to obtain the observed significance level. If this P value is sufficiently small, the null hypothesis is rejected.

Which statistic should be used? Many statisticians have offered arguments for preferring one statistic over the others but, in practice, most researchers use the one that is provided by their statistical software or that is easiest to calculate by hand.

All of the statistics can be justified by large sample statistical theory. They all reject H_0 100(1-α)% of the time when $H_0$is true. (However, they don't always agree on the same set of data.) Since they all reject H_0 with the same frequency when it is true, you might think of using the test that is more likely to reject H_0 when it is false, but none has been shown to be more likely than the others to reject H_0 when it is false for all alternatives to H_0.

The first statistic is

$$z_1 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}},$$

The second is

$$z_2 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}},$$

where $\hat{p}$ is the proportion of individuals having the characteristic when the two samples are lumped together.

A third statistic is

$$z_3 = \frac{|\hat{p}_1 - \hat{p}_2| - \frac{1}{2}(\frac{1}{n_1} + \frac{1}{n_2})}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

The test statistic $\mathbf{z_1}$ is consistent with the corresponding confidence interval, that is, z_1 rejects H_0 at level α if and only if the 100(1-α)% confidence interval does not contain 0.

The test statistic $\mathbf{z_2}$ is equivalent to the chi- square goodness-of-fit test, also called (correctly) a test of homogeneity of proportions and (incorrectly, for this application) a test of independence.

The test statistic $\mathbf{z_3}$ is equivalent to the chi- square test with Yates's continuity correction. It was developed to approximate another test statistic (Fisher's exact test) that was difficult to compute by hand. Computers easily perform this calculation, so this statistic is now obsolete. Nevertheless, most statistical program packages continue to report it as part of their analysis of proportions.

Examples

1. $\hat{p}_1$=8/13 and $\hat{p}_2$=3/13. Then, z_1=2.155 (P=0.031), z_2=1.985 (p=0.047), and z_3=1.588 (P=0.112). Fisher's exact test gives P=0.111.
2. $\hat{p}_1$=16/34 and $\hat{p}_2$=6/26. Then, z_1=2.016 (P=0.044) and z_2=1.910 (p=0.056). A 95% CI for p1-p2 is 0.2398 ± 0.2332=(0.0066,0.4730). The confidence interval agrees with z_1. The CI does not contain 0, while z_1 rejects H_0: p_1=p_2. However, z_1 and the CI disagree with z_2 which fails to reject H_0.

Common sense suggests using z_1 because it avoids conflicts with the corresponding confidence interval. However, in practice, the chi-square test for homogeneity of proportions (equivalent to z_2) is used because that's what statistical software packages report. I don't know any that report z_1. However, z_2 (in the form of the chi-square test) has the advantage of generalizing to tests of the equality of more than two proportions.

Tests Involving a Single Population Proportion

When testing the null hypothesis H_0: the population proportion equals some specified value p_0 against H_1: the population proportion does not equal p_0, there is, once again, a choice of test statistics.

$$z_1 = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}, \quad z_2 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \quad z_3 = \frac{|\hat{p} - p_0| + \frac{1}{2n}}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

all of which are compared to the percentiles of the standard normal distribution.

Again, z_1 gives tests that are consistent with the corresponding confidence intervals, z_2 is equivalent to the chi-square goodness-of-fit test, and z_3 gives one-sided P- values that usually have better agreement with exact P-values obtained, in this case, by using the binomial distribution.

Comment

These techniques are based on large sample theory. Rough rules of thumb say they may be applied when there are at least five occurrences of each outcome in each sample and, in the case of a single sample, provided confidence intervals lie entirely in the range (0,1).

Summary

1. We can construct confidence intervals for population proportions and for the difference between population proportions just as we did for population means.
2. We can test the hypothesis that two population proportions are equal just as we did for population means.
3. The formulas for constructing confidence intervals and for testing the hypothesis of equal proportions are slightly different, unlike the case of means where the two formulas are the same.
4. As a consequence of (3), it is possible (although uncommon) for the test to reject the hypothesis of equal proportions while the CI for their difference contains 0, or for the test to fail to reject while the CI does not contain 0!
5. The formula for CIs can be adapted for significance testing. However, the formula for significance tests cannot be adapted for constructing CIs.
6. Which test statistic should be used? All are equally valid. Almost every statistical program provides a test procedure that is equivalent to z_2 for comparing proportions, so that's what people use.
7. Why is the test statistic based on the CI for population differences not widely available in statistical software? Because the chi-square test is easily generalized to classifications with more than two categories. The other test statistic is not.
8. This is just the tip of the iceberg. When the response is counts, there can be dozens of valid test statistics and methods for constructing confidence intervals, all giving slightly different results. The good news is that they tend to give the same inference (lead to the same conclusion).

Odds

Gerard E. Dallal, Ph.D.

The odds *o(E)* of an event *E* is the ratio of the probability that the event will occur, *P(E)*, to the probability that it won't, *1-P(E)*, that is,

$$o(E) = \frac{P(E)}{1 - P(E)}$$

For example, if the probability of an event is 0.20, the odds are 0.20/0.80 = 0.25.

In epidemiology, odds are usually expressed as a single number, such as the 0.25 of the last paragraph. Outside of epidemiology, odds are often expressed as the ratio of two integers--2:8 (read "2 to 8") or 1:4. If the event is less likely to occur than not, it is common to hear the odds stated with the larger number first and the word "against" appended, as in "4 to 1 against".

When the odds of an event are expressed as X:Y, an individual should be willing to lose X if the event does not occur in order to win Y if it does. When the odds are 1:4, an individual should be willing to lose $1 if the event does not occur in order to win $4 if it does.

The Fascination With Odds

A common research question is whether two groups have the same probability of contracting a diseaase. One way to summarize the information is the **relative risk**--the ratio of the two probabilities. For example, in 1999, He and his colleagues reported in the New England Journal of Medicine that the realative risk of coronary heart disease for those exposed to second hand smoke is 1.25--those exposed to second hand smoke are 25% more likely to develop CHD.

As we've already seen, when sampling is performed with the totals of the disease categories fixed, we can't estimate the probability that either exposure category gets the disease. Yet, the medical literature is filled with reports of case-sontrol studies where the investigators do just that--examine a specified number of subjects with the diesease and a number without it. In the case of rare diseases this is about all you can do. Otherwise, thousands of individuals would have to be studied to find that one rare case. The reason for the popularity of the case control study is that, thanks to a little bit of algbera, odds ratios give us something almost as good as the relative risk. allow us to obtain something almost as good as a relative risk.

There are two odds ratios. The **disease odds ratio** (or **risk odds ratio**) is the ratio of (the odds of disease for those with some exposure) to (the odds of disease for those without the exposure). The **exposure**

odds ratio is ratio of (the odds of exposure for those with disease) to (the odds of exposure for those without disease).

When sampling is performed with the totals of the disease categories fixed, we can *always* estimate the exposure odds ratio. Simple algebra shows that the exposure odds ratio is equal to the disease odds ratio! Therefore, sampling with the totals of the disease categories fixed allows us to determine whether two groups have different probabilities of having a disease.

1. We sample with disease category totals fixed.
2. We estimate the exposure odds ratio.
3. The exposure odds ratio is equal to the disease odds ratio. Therefore, if the exposure odds ratio is different from 1, the disease odds ratio is different from 1.

A bit more simple algebra shows that if the disease is rare (<5%), then the odds of contracting the disease is almost equal to the probability of contracting it. For example, for *p*=0.05, $\frac{p}{1-p} = 0.0526$, which is not much different from 0.05. Therefore, **when a disease is rare, the exposure odds ratio is equal to the disease odds ratio, which, in turn, is approximately equal to the relative risk!**

Hence, the fascination with odds!

Last modified: undefined.

Paired Counts

Gerard E. Dallal, Ph.D.

There are as many ways to collect paired counts as there are reasons for measuring something twice. Subjects might be classified into one of two categories according to two different measuring devices. Opinions (pro and con) might be assessed before and after some intervention. Just as it was necessary account for pairing when analyzing continuous data--choosing Student's t test for paired samples rather than the test for independent samples--it is equally important to take account of pairing when analyzing counts.

Consider a study to examine whether food frequency questionnaires and three-day food diaries are equally likely to label a women as consuming less than the RDA of calcium. One way to conduct this study is to take a sample of women and assign them at random to having their calcium intake measured by food frequency or diary. However, calcium intake can vary considerably from person to person, so a better approach might be to use both instruments to evaluate a single sample of women.

Suppose this latter approach is taken with a sample of 117 women and the results are as follows:

Diet Record	Food Frequency Questionnaire	Count
<RDA	<RDA	33
<RDA	≥RDA	27
≥RDA	<RDA	13
≥RDA	≥RDA	44

How should the data be analyzed? Pearson's test for homogeneity of proportions comes to mind and it is tempting to construct the table

	Calcium Intake	
	<RDA	**≥RDA**
Food Frequency Questionnaire	46	71
Diet Record	60	57

Pearson's chi-square statistic is 3.38 and the corresponding P value is 0.066.

However, there are some **problems** with this approach.

- **The table contains 234 observations, not 117**. One of the requirements of the standard chi-square test is that each subject appear in one and only one cell.
- Further, there's no obvious way in which the paired nature of the data was used in the analysis.

Here's another way to represent the data in which each subject appears once and only once.

Diet Records	**Food Frequency Questionnaire**	
	<RDA	**≥RDA**
<RDA	33	27
≥RDA	13	44

However, even though each person appears only once, *you have to resist the urge to use Pearson's goodness-of-fit test* because **it tests the wrong hypothesis**!

The question is still whether the two instruments identify the same proportion of women as having calcium intakes below the RDA. The Pearson goodness-of-fit statistic does not test this. It tests **whether the classification by food frequency is independent of the classification by Diet Record!**

> [These are two different things! Consider the following table.
>
> | **Diet Records** | **Food Frequency Questionnaire** | |
> |---|---:|---:|
> | | **<RDA** | **≥RDA** |
> | **<RDA** | 20 | 20 |
> | **≥RDA** | 10 | 10 |
>
> The two instruments are independent because half of the subjects' intakes are declared inadequate by the FFQ regardless of what the Diet Record says. Yet, while **the FFQ says half** of the subjects (30 out of 60) have inadequate intake, **the Diet Record says two-thirds** (40 out of 60) of the intakes are inadequate.]

There may be cases where you want to test for independence of the two instruments. Those who have little faith in either the Diet Record or Food Frequency Questionnaire might claim that the test is appropriate in this case! But, usually you already know that the methods agree to some extent. This makes a test of independence pointless.

The appropriate test in this situation is known as **McNemar's test**. It is based on the observation that if the two proportions are equal, then discordant observations (where the methods disagree) should be equally divided between (low on frequency, high on diary) and (high on frequency, low on diary). Some commonly used test statistics are

$$X_1 = (b - c)^2 / (b + c)$$

and

$$X_2 = (|b - c| - 1)^2 / (b + c)$$

where *b* and *c* are the discordant cells in the 2 by 2 table. Both of statistics are referred to the chi-square distribution with 1 degree of freedom. Since the test statistics involve the square of the difference between the counts, they are necessarily two-sided tests (For these data: $X_1 = 4.90$, $P = 0.0269$; $X_2 = 4.23$, $P = 0.0397$.)

While it may seem strange, counter-intuitive, and even *wrong* when the realization first hits, the only relevant data are the numbers in the discordant cells, here the 27 and the 13. The information about how diet records and FFQs disagree is the same whether the cell counts showing agreement are 33 and 44 or 33,000,000 and 44,000,000. The distinction is that in this lattercase a statistically significant difference may be of no practical importance.

Other situations in which McNemar's test is appropriate include measuring change (status before and after an intervention) and case- control studies in which everyone is measured for the presence/absence of a characteristic. The feature that should sensitize you to McNemar's test is that both measurements are made on the same observational unit, whether it be an individual subject or case-control pair.

[back to The Little Handbook of Statistical Practice]

Last modified: undefined.

What Underlies Sample Size Calculations

Gerard E. Dallal, Ph.D.

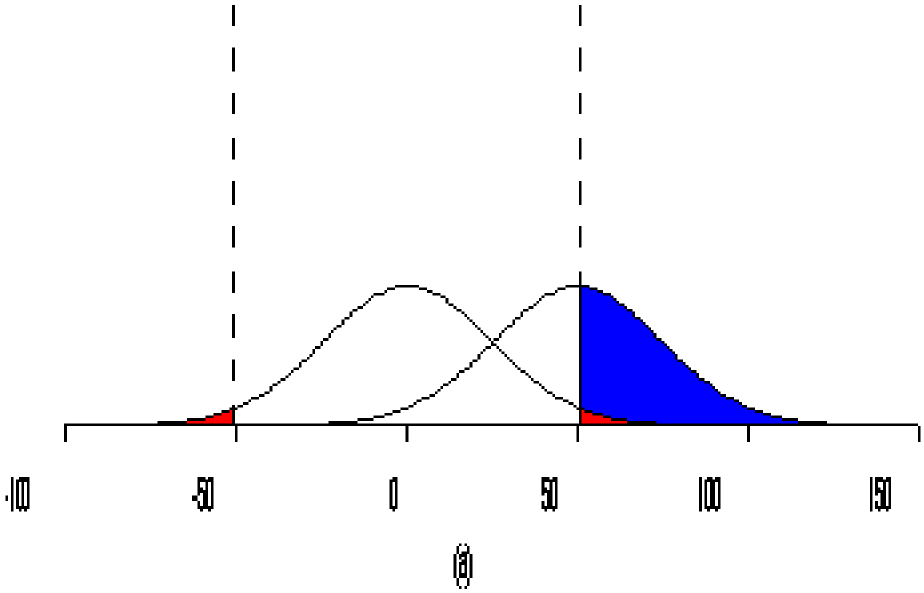
Prologue

Just as the analysis of a set of data is determined by the research question and the study design, the way the sample size is estimated is determined by the way the data will be analyzed. This note (at least until the next draft!) is concerned with comparing population means. There are similar methods for comparing proportions and different methods for assessing correlation coefficients. Unfortunately, it is not uncommon to see sample size calculations that are totally divorced from the study for which they are being constructed because the sample sizes are calculated for analyses that will never be used to answer the question prompting the research. The way to begin, then, is by thinking of the analysis that will ultimately be performed to insure that the corresponding sample size calculations have been used. This applies even to comparing two population means. If experience suggests a logarithmic transformation will be applied to the data prior to formal analysis, then the sample size calculations should be performed in the log scale.

Comparing Population Means

Studies are generally conducted because an investigator expects to see a specific treatment effect.* Critical regions and tests of significance are determined by the way data should behave if there is no treatment effect. Sample sizes are determined by the way data should behave if the investigator has estimated the treatment effect correctly.**

Comparing Two Population Means: Independent Samples

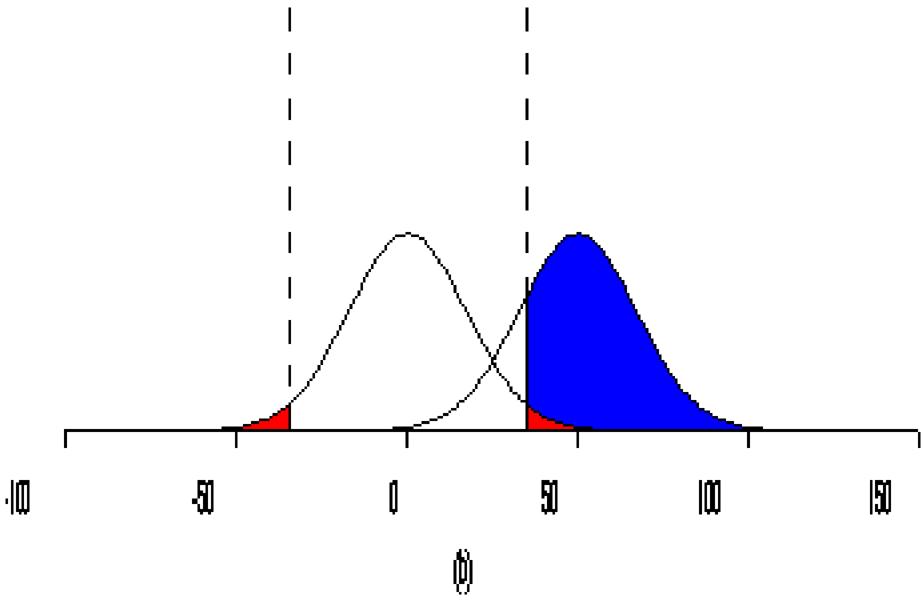


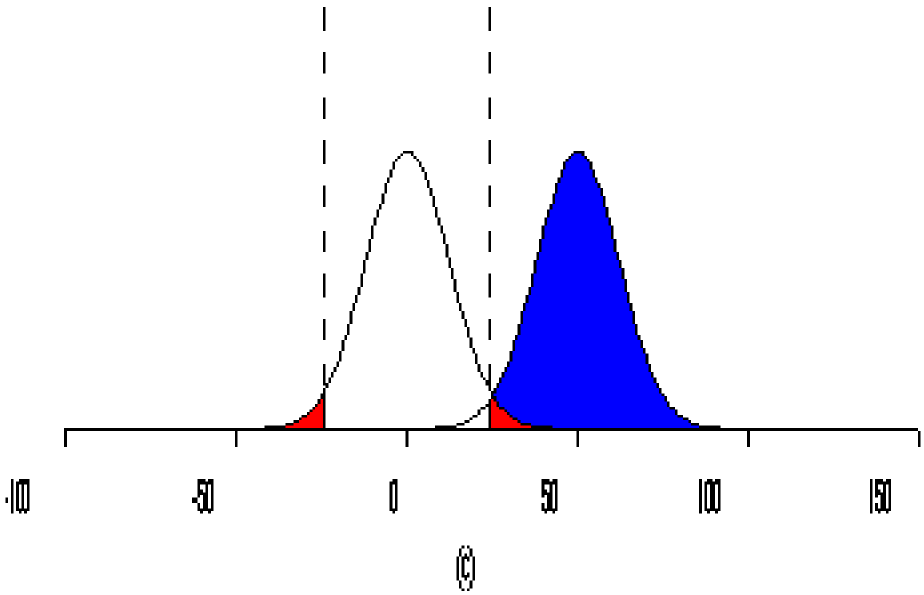
Consider a study using two independent samples to compare their population means. Let the common population standard deviation be 60. The behavior of the difference in sample means under the null hypothesis of equal population means is illustrated by the normal distributions on the left-hand side of displays (a) through (d) for sample sizes of 12, 24, 48, and 96 per group, respectively.

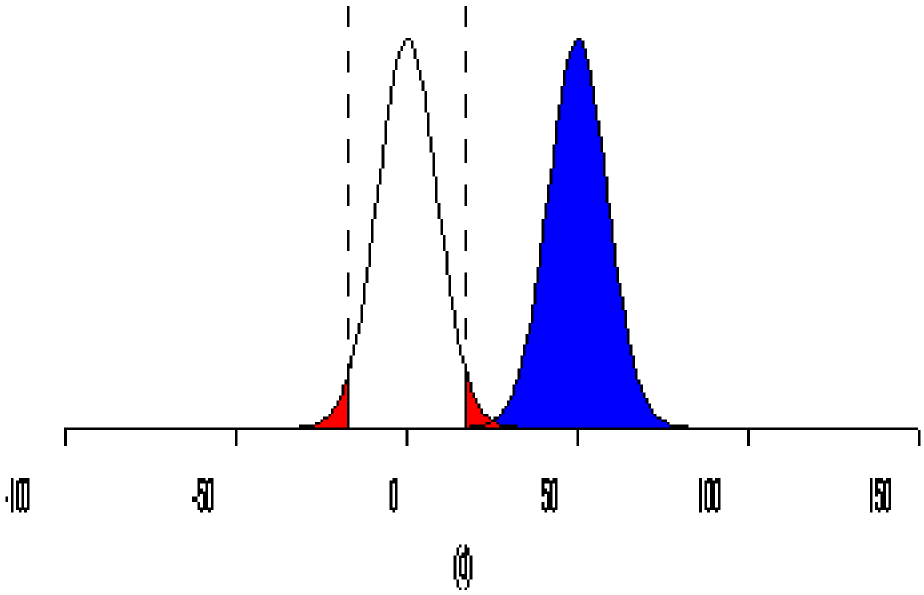
Suppose the investigator expects the difference in population means to be 50 units. Then, the behavior of the difference in sample means is described by the curves on the right-hand side of the displays.

Things to notice about (a)--(d):

- The horizontal scales are the same.
- The normal curves on the left-hand side of the display are centered at 0.
- As the sample size increases, the distribution of the difference in sample means as given by the normal curves on the left-hand side of the display are more tightly concentrated about 0.
- The critical values for an 0.05 level test--sample mean differences that will lead to rejecting the hypothesis of equal population means--are given by the vertical dashed lines. The critical region is shaded red. If the mean difference falls outside the vertical lines (in the critical region), the hypothesis of equal population means is rejected.
- As the sample size increases, the critical values move closer to 0. This reflects the common sense notion that the larger the sample size, the harder it is (less likely) for the sample mean difference to be at any distance from 0.







Other things to notice about (a)--(d):

- The normal curves on the right-hand side of the display are centered at 50.
- As the sample size increases, the distribution of the difference in sample means as given by the normal curves on the right-hand side of the display are more tightly concentrated about 50.
- As the sample size increases, more of the curve on the right-hand side of the displays falls into the critical region. The portion of the distribution on the right-

hand side of the displays that falls into the critical region is shaded blue.

The region shaded blue gives the power of the test. It is 0.497, 0.807, 0.981, and 1.000 for panels (a) through (d), respectively.

Choosing a sample size is just a matter of getting the picture "just right", that is, seeing to it that there's just the right amount of blue.

It seems clear that a sample size of 12 is too small because there's a large chance that the expected effect will not be detected even if it is true. At the other extreme, a sample size of 96 is unnecessarily large. Standard practice is to choose a sample size such that the power of the test is no less than 80% when the effect is as expected. In this case, the sample size would be 24 per group. Whether a sample size larger than 24 should be used is a matter of balancing cost, convenience, and concern the effect not be missed.

The pictures show how the sample size is a function of four quantities.

- the presumed underlying difference (Δ), that is, that is, the *expected difference* between the two populations means should they be unequal. In each of the displays, changing the expected difference moves the two distributions further apart or closer together. This will affect the amount of area that is shaded blue. Move them farther apart and the area increases. Move them closer together and the area decreases.
- the *within group standard deviation* (σ), which is a measure of the variability of the response. The width of the curves in the displays is determined by the with group standard deviation and the sample size. If the sample size is fixed, then the greater/smaller the standard deviation, the wider/narrower the curves. If the standard deviation is fixed, then the larger/smaller the sample size, the narrower/wider the curves. Changing width of the curves will move the critical values, too. Displays (a)--(d) were constructed for different sample sizes with the population standard deviation fixed. However, the same pictures could have been obtained by holding the sample size fixed but changing the population standard deviation.
- the size or *level of the* statistical *test* (α). Decreasing the level of the test--from 0.05 to 0.01, say--moves the critical valued further away from 0, reducing the amount of area that is shaded red. It also reduces the amount of area shaded blue. This represents a trade off. Reducing the amount of area shaded red reduces the probability of making an error when there is no difference. This is good. Reducing the amount of area shaded blue reduces the probability of making the correct decision when the difference is as expected. This is bad.
- the probability of rejecting the hypothesis of equal means if the difference is as specified, that is, the *power of the test* (π) when the difference in means is as expected. This is the area that is shaded blue.

The sample size is determined by the values of these four quantities. Denoting the expected mean difference locates the centers of the distributions on the number line. Picking the size of the test determines the amount of area that will be shaded red. For a fixed sample size, it also determines the critical values and the amount of area that will be shaded blue. Increasing the sample size makes the distributions narrower which moves the critical values closer to the mean of the distribution of the test statistic under the null hypothesis. This increases the amount of area shaded blue.

In practice, we don't draw lots of diagrams. Instead, there is a formula that yields the per group sample size when the four quantities are specified. For large samples, the per group sample size is given by

$$\frac{2(z_{1-\alpha/2}+z_{\pi})^2\sigma^2}{\Delta^2},$$

where $z_{(1-\alpha/2)}$ (>0) is the percentile of the normal distribution used as the critical value in a two-tailed test of size α (1.96 for an 0.05 level test) and $z\pi$ is the $100\times\pi$-th percentile of the normal distribution (0.84 for the 80-th percentile).

> **Technical detail:** For small sample sizes, percentiles of the t distribution replace the percentiles of the normal distribution. Since the particular t distribution depends on the sample size, the equation must be solved iteratively (trial-and-error). There are computer programs that do this with little effort.

The sample size increases with the **square** of the within group standard deviation and decreases with the **square** of the expected mean difference. If, for example, when testing a new treatment a population can be found where the standard deviation is half that of other populations, the sample size will be cut by a factor of 4.

Points To Keep In Mind

The alternative to equality must be realistic. The larger the expected difference, the smaller the required sample size. It can be QUITE TEMPTING to overstate the expected difference to lower the sample size and convince one's self or a funding agency of the feasibility of the study. All this strategy will do, however, is cause a research team to spend months or years engaged in a hopeless investigation--an underpowered study that cannot meet its goals. A good rule is to ask whether the estimated difference would still seem reasonable if the study were being proposed by someone else.

The power, π--that is, probability of rejecting H0 when the alternative holds--can, in theory, be made as large or small as desired. Larger values of π require larger sample sizes, so the experiment might prove too costly. Smaller values of π require smaller sample sizes, but only by reducing the chances of observing a significant difference if the alternative holds. Most funding agencies look for studies with at least 80-% power. In general, they do not question the study design if the power is 80-% or greater. Experiments with less power are considered too chancy to fund.

Estimating the within group standard deviation, σ, When The Response Is a Single Measurement

The estimate of the within group standard deviation often comes from similar studies, sometimes even 50 years old. If previous human studies are not available to estimate the variability in a proposed human study, animal studies might be used, but animals in captivity usually show much less variability than do humans. Sometimes it is necessary to guess or run a pilot study solely to get some idea of the inherent variability.

Many investigators have difficulty estimating standard deviations simply because it is not something they do on a regular basis. However, standard deviations can often be obtained in terms of other measures that are more familiar to researchers. For example, a researcher might specify a range of values that contains most of the observations. If the data are roughly normally distributed, this range could be treated as an interval that contains 95% of the observations, that is, as an interval of length 4σ. The standard deviation, then, is taken to be one-fourth of this range. If the range were such that it contains virtually all of the population, it might be treated as an interval of length 6σ. The standard deviation, then, is taken to be one-sixth of this range.

Underestimating the standard deviation to make a study seem more feasible is as foolhardy as overestimating an expected difference. Such estimates result in the investment of up resources in studies that should never have been performed. Conservative estimates (estimates that lead to a slightly larger sample size) are preferable. If a study is feasible when conservative estimates are used, then it is well worth doing.

Estimating the within group standard deviation, σ, When the Response Is a Difference

When the response being studied is change or a difference, the sample size formulas require the standard deviation of the difference between measurements, not the standard deviation of the individual measurements. It is one thing to estimate the standard deviation of total cholesterol when many individuals are measure once; it is quite another to estimate the standard deviation of the change in cholesterol levels when changes are measured.

One trick that might help: Often a good estimate of the standard deviation of the differences is unavailable, but we have reasonable estimates of the standard deviation of a single measurement. The standard deviations of the individual measurements will often be roughly equal. Call that standard deviation σ. Then, the standard deviation of the paired differences is equal to

$$\sigma \ \sqrt{}\ (2[1-\rho]),$$

where ρ is the correlation coefficient when the two measurements are plotted against each other. If the correlation coefficient is a not terribly strong 0.50, the standard deviation of the differences will be equal to σ and gets smaller as the correlation increases.

Many Means

Sometimes a study involves the comparison of many treatments. The statistical methods are discussed in detail under *Analysis of Variance (ANOVA)*. Historically, the analysis of many groups begins by asking whether all means are the same. There are formulas for calculating the sample size necessary to reject this hypothesis according to the particular configuration of population means the researchers expect to encounter. These formulas are usually a bad way to choose a sample size because the purpose of the experiment is rarely (never?) to see whether all means are the same. Rather, it is to catalogue the differences. The sample size that may be adequate to demonstrate that the population means are not all the same may be inadequate to demonstrate exactly where the differences occur.

When many means are compared, statisticians worry about the problem of multiple comparisons, that is, the possibility that some comparison may be call statistically significant simply because so many comparisons were performed. Common sense says that if there are no differences among the treatments but six comparisons are performed, then the chance that something reaches the level of statistical significance is a lot greater than 0.05. There are special statistical techniques such as *Tukey's Honestly Significant Differences (HSD)* that adjust for multiple comparisons, but there are no easily accessible formulas or computer programs for basing sample size calculations on them. Instead, sample sizes are calculated by using a Bonferroni adjustment to the size of the test, that is, the nominal size of the test is divided by the number of comparisons that will be performed. When there are three means, there are three possible comparisons (AB,AC,BC). When there are four means, there are six possible comparisons (AB,AC,AD,BC,BD,CD), and so on. Thus, when three means are to be compared at the 0.05 level, the two-group sample size formula is used, but the size of each individual comparison is taken to be 0.05/3 (=0.0167). When four means are compared, the size of the test is 0.05/6 (=0.0083).

The Log Scale

Sometimes experience suggests a logarithmic transformation will be applied to the data prior to formal analysis. This corresponds to looking at ratios of population parameters rather than differences. When the analysis will be performed in the log scale, the sample size calculations should be performed in the log scale, too. If only summary data are available for sample size calculations and they are in the original scale, the behavior in the log scale can be readily approximated. The expected difference in means in the log scale is approximately equal to the log of the ratio of means in the original scale. The common within group standard deviation in the natural log scale (base *e*) is approximately equal to the coefficient of variation in the original scale (the roughly constant ratio of the within standard deviation to the mean). If the calculations are being performed in the common log scale (base 10), divide the cv by 2.3026 to estimate the common within group standard deviation.

Example: (α=0.05, π=0.80) Suppose a response will be analyzed in the log scale and that in the original scale, the population means are expected to be 40 and 50 mg/dl and the common coefficient of variation (σ/μ) is estimated to be 0.30. Then, in the (natural) log scale the estimated effect is ln(50/40)

= ln(1.25) = 0.2231 and common within group standard deviation is estimated to be 0.30 (the cv). The per group sample size is approximately 1+16(0.30/0.2231)^2 or 30. In the common log scale, the estimated effect is log(50/40) = 0.0969 and the estimated common within group standard deviation is estimated to be 0.30/2.3026 = 0.1303. The per group sample size is approximately 1+16(0.1301/0.0969)^2 or 30. It is not an accident that the sample sizes are the same. The choice of a particular base for logarithms is like choosing to measure height in cm or in. It doesn't matter which you use **as long as you are consistent!** No mixing allowed! A few things worth noting:

- log(40/50) = -0.0969, that is, -log(50/40). Since this quantity is squared when sample sizes are being estimated, it doesn't matter which way the ratio is calculated.
- The cv estimates the common within group SD for log transformed data works only for natural logs. When you take the log of the ratio to estimate the treatment effect in the log scale, you pick the particular type of log you prefer. Since cv estimates the common within group SD for natural-log transformed data, you have to adjust it accordingly if you calculate the treatment effect in logs of a different base.
- 2.3026--the factor which, when divided into natural logs, converts *ln*s to *log*s-- = ln(10).

A potential **gotcha!**: When calculating the treatment effect in the log scale, you can never go wrong calculating the log of the ratio of the means in the original scale. However, you have to be careful if the effect is stated in terms of a percent increase or decrease. Increases and decreases are not equivalent. Suppose the standard treatment yields a mean of 100. A 50% increase gives a mean of 150. The ratio of the means is 150/100(=3/2) or 100/150(=2/3), Now consider a 50% decrease from standard. This leads to a mean of 50. The ratio is now 100/50(=2) or 50/100(=1/2). There's no trick here. The mathematics is correct. The message is that you have to be careful when you translate statements about expected effects into numbers needed for the formal calculations.

Comparing Two Population Means:
Dealing With Paired Responses

Sometimes responses are truly paired. Two treatments are applied to the same individual or the study involves matched or paired subjects. In the case of paired samples, the formula for the total number of pairs is the same as for the number of independent samples except that the factor of 2 is dropped, that is,

$$\frac{(z_{1-\alpha/2}+z_{\pi})^2\,\sigma^2}{\Delta^2},$$

where σ is now the standard deviation of the differences between the paired measurements. In many (most?) cases, especially where a study involves paired changes, σ is *not* easy to estimate. You're on your own!

It is clear from the formulas why paired studies are so attractive. First, is the factor of 2. All other things being equal, a study of independent samples that requires, say, 100 subjects per group or a total of 200 subjects, requires only 50 pairs for a total of 100 subjects. Also, if the pairing is highly effective, the standard deviation of the differences within pair can be quite small, thereby reducing the sample size even further. However, these saving occur because elements within the same pair are expected to behave somewhat the same. If the pairing is ineffective, that is, if the elements within each pair are independent of each other, the standard deviation of the difference will be such that the number of pairs for the paired study turns out to be equal to the number of subjects per group for the independent samples study so that the total sample size is the same.

There is a more important concern than ineffective pairing. When some investigators see how the sample sizes required for paired studies compared to those involving independent samples, their first thought is to drop any control group in favor of "using subjects as their own control". Who wouldn't prefer to recruit 50 subjects and look at whether their cholesterol levels change over time rather than 200 subjects (100 on treatment; 100 on placebo) to see if the mean change in the treatment group is different from that in the control group? However, this is not an issue of sample size. It is an issue of study design. An investigator who measured only the 50 subjects at two time points would be able to determine whether there was a change over time, but s/he would not be able to say how it compared to what would have happened over the same time period in the absence of any intervention.

----

[*] There are exceptions such as equivalence trials where the goal is to show that two population means are the same, but they will not concern us here.

[**] It may sound counter-intuitive for the investigator to have to estimate the difference when the purpose of the study is to determine the difference. However, it can't be any other way. Common sense suggests it takes only a small number of observations to detect a large difference while it takes a much larger sample size to detect a small difference. Without some estimate of the likely effect, the sample size cannot be determined. Sometimes there will be no basis for estimating the likely effect. The best that can be done in such circumstances is a pilot study to generate some preliminary data and estimates.

[back to LHSP]

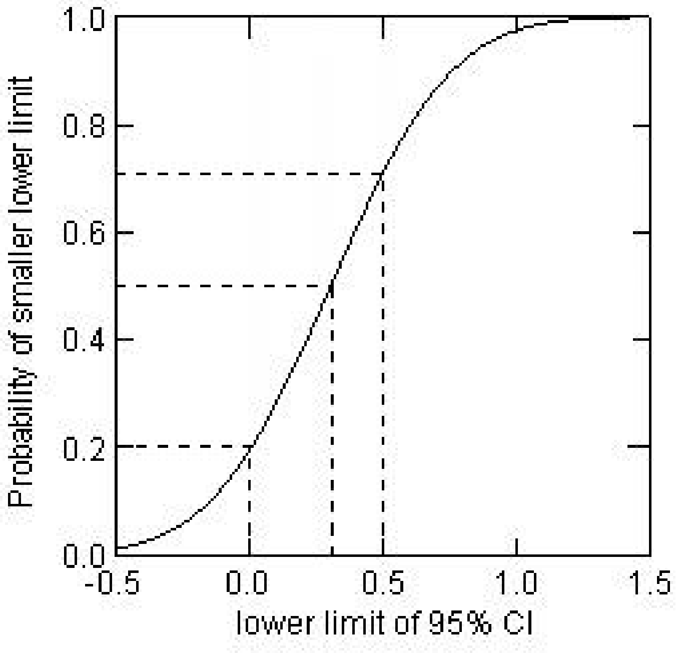
Last modified: undefined.

An Underappreciated Consequence of Sample Size Calculations
As They Are Usually Performed

[Someday, this will get integrated into the main sample size note. At the moment, I don't know how to do it without the message getting lost, so I've created this separate note with a provocative title so that it will be noticed.]

My concern with sample size calculations is related to the distinction between significance tests and confidence intervals. Sample size calculations as described in these notes and most textbooks are designed to answer the very simple question, *"How many observations are needed so that an effect can be detected?"* While this question is often of great interest, the *magnitude* of the effect is often equally important. While it is possible to sidestep the issue by asserting that the magnitude of the effect can be assess after it's determined that there's something to assess, the question must be addressed at some point.

As with significance tests, knowing whether there is an effect tells you something, but leaves a lot unsaid. There are statistically significant effects of great practical importance and effects of no practical importance. The problem with sample size calculations as they are ususally performed is that there is a substantial chance that one end of the confidence will include values of no practical importance. Thus, while an experiment has a large chance of demonstrating the effect if it is what the investigators expect, there is a good chance that the corresponding confidence interval might leave open the possibility that the effect is quite small.



For example, consider a comparison of two population means where the expectd mean difference and *known* within group standard deviation are both equal to 1. The standard deviation is treated as known for this example to keep the mathematics manageable. A sample of 16 subjects per group gives an 81% chance that the hypothesis of no difference will be rejected by Student's t test at the 0.05 level of significance.

The picture at the left shows what happens with the lower limit of the 95% confidence interval for the population mean difference when the underlying mean difference and within group standard deviation are both 1. There is a 20% chance that the lower confidence limit will be less than 0, in keeping with the 20% chance that the expereiment will fail to show a statistically significant difference. As the curve demonstrates, there is also a 50% chance that the lower limit will be less than 0.31 and a

70% chance that it will be less than 0.50, that is, there is a 70% chance that the lower limit of the 95% CI will be less than half of the expected effect!

This is not a problem if the goal of a study is merely to demonstrate a difference in population means. If the goal is to estimate the difference accurately, the sample size calculations must take this into account, perhaps by using a method such as the one presented by Kupper and Hafner in their 1989 article "How Appropriate Are Popular Sample Size Formulas?" (*The American Statistician*, vol 43, pp 101-105).

[back to The Little Handbook of Statistical Practice]

Gerard E. Dallal
Last modified: undefined.

SAMPLE SIZE CALCULATIONS SIMPLIFIED
Controlled Trials

Most sample size calculations involve estimating the number of observations needed to compare two means by using Student's t test for independent samples or two proportions by using Pearson's chi-square test. Standard practice is to determine the sample size that gives an 80% chance of rejecting the hypothesis of no difference at the 0.05 level of significance.

Two Means

The sample size estimate depends on the difference between means and the within-group variability of individual measurements. A formula for the approximate per group sample size is

$$16\ s^2/d^2 + 1,$$

where 'd' is the expected difference between means and 's' is the within-group standard deviation of the individual measurements. For example, if the difference between means is expected to be 18 mg/dl and the within-group standard deviation is 30 mg/dl, the required sample size is approximately 46 (= 16 $30^2/18^2 + 1$) per group. (The exact answer is 45.)

Many Means

Sometimes a study involves the comparison of many treatments. The statistical methods are discussed in detail under *Analysis of Variance (ANOVA)*. Historically, the analysis of many groups begins by asking whether all means are the same. There are formulas for calculating the sample size necessary to reject this hypothesis according to the particular configuration of population means the researchers expect to encounter. These formulas are usually a bad way to choose a sample size because the purpose of the experiment is rarely (never?) to see whether all means are the same. Rather, it is to catalogue the differences. The sample size that may be adequate to demonstrate that the population means are not all the same may be inadequate to demonstrate exactly where the differences occur.

When many means are compared, statisticians worry about the problem of multiple comparisions, that is, the possiblity that some comparison may be call statistically significant simply because so many comparisons were performed. Common sense says that if there are no differences among the treatments but six comparisons are performed, then the chance that something reaches the level of statistical significance is a lot greater than 0.05. There are special statistical techniques such as *Tukey's Honestly Significant Differences (HSD)* that adjust for multiple comparisons, but there are no easily accessbile formulas or computer programs for basing sample size calculations on them. Instead, sample sizes are calculated by using a Bonferroni adjustment to the size of the test, that is, the nominal size of the test is divided by the number of comparisons that will be performed. When there are three means, there are three possible comparisons (AB,AC,BC). When there are four means, there are six possible comparisons

(AB,AC,AD,BC,BD,CD), and so on. Thus, when three means are to be compared at the 0.05 level, the two-group sample size formula is used, but the size of each individual comparison is taken to be 0.05/3 (=0.0167). When four means are compared, the size of the test is 0.05/6 (=0.0083). The approximate per group sample size when three means are compared at the 0.05 level is

$$22\ s^2/d^2 + 1,$$

while for four means it is

$$26\ s^2/d^2 + 1.$$

Comparing Changes

Often, the measurement is change (change in cholesterol level, for example). Estimating the difference in mean change is usually not a problem. Typically, one group has an expected change of 0 while the other has an expected change determined by the expected effectiveness of the treatment.

When the measurement is change, sample size formulas require the within-group standard deviation of individual changes. Often, it is often unavailable. However, the within-group standard deviation of a set of individual measurements at one time point is usually larger than the standard deviation of change and, if used in its place, will produce a conservative (larger than necessary) sample size estimate. The major drawback is that the cross-sectional standard deviation may be so much larger than the standard deviation of change that the resulting estimate my be useless for planning purposes. The hope is that the study is will prove feasible even with this inflated sample size estimate.

For example, suppose the primary response in a comparative trial is change in ADL score (activities of daily living). It is expected that one group will show no change while another group will show an increase of 0.6 units. There are no data reporting the standard deviation of change in ADL score over a period comparable to the length of the study, but it has been reported in a cross-sectional study that ADL scores had a standard deviation of 1.5 units. Using the standard deviation of the cross-section in place of the unknown standard deviation of change gives a sample size of 101 ($=1.5^2/0.6^2 + 1$) per group.

Two Proportions

The appended chart gives the per group sample size needed to compare proportions. The expected proportions for the two groups are located on the row and column margins of the table and the sample size is obtained from corresponding table entry. For example, if it is felt that the proportion will be 0.15 in one group and 0.25 in the other, 270 subjects per group are needed to have an 80% chance of rejecting the hypothesis of no difference at the 0.05 level.

Points to Consider

The calculations themselves are straightforward. A statistician reviewing sample size estimates will have two concerns: (1) Are the estimates of within-group variability valid, and (2) are the anticipated effects biologically plausible? If I were the reviewer, I would seek the opinion of a subject matter specialist. As long as you can say to yourself that you would not question the estimates had they been presented to you by someone else, outside reviewers will probably not find fault with them, either.

Per group sample size required for an 80% chance of rejecting the hypothesis of equal proportions at the 0.05 level of significance when the true proportions are as specified by the row and column labels

	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.05	0	474	160	88	59	43	34	27	22	19
0.10	474	0	726	219	113	72	51	38	30	25
0.15	160	726	0	945	270	134	83	57	42	33
0.20	88	219	945	0	1134	313	151	91	62	45
0.25	59	113	270	1134	0	1291	349	165	98	66
0.30	43	72	134	313	1291	0	1417	376	176	103
0.35	34	51	83	151	349	1417	0	1511	396	183
0.40	27	38	57	91	165	376	1511	0	1574	408
0.45	22	30	42	62	98	176	396	1574	0	1605
0.50	19	25	33	45	66	103	183	408	1605	0
0.55	16	20	26	35	48	68	106	186	412	1605
0.60	14	17	22	28	36	49	70	107	186	408
0.65	12	15	18	22	28	37	49	70	106	183
0.70	11	13	15	19	23	29	37	49	68	103
0.75	9	11	13	16	19	23	28	36	48	66
0.80	8	10	11	13	16	19	22	28	35	45
0.85	7	9	10	11	13	15	18	22	26	33
0.90	7	8	9	10	11	13	15	17	20	25
0.95	6	7	7	8	9	11	12	14	16	19

	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.05	19	16	14	12	11	9	8	7	7	6
0.10	25	20	17	15	13	11	10	9	8	7
0.15	33	26	22	18	15	13	11	10	9	7
0.20	45	35	28	22	19	16	13	11	10	8
0.25	66	48	36	28	23	19	16	13	11	9
0.30	103	68	49	37	29	23	19	15	13	11
0.35	183	106	70	49	37	28	22	18	15	12
0.40	408	186	107	70	49	36	28	22	17	14

0.45	1605	412	186	106	68	48	35	26	20	16
0.50	0	1605	408	183	103	66	45	33	25	19
0.55	1605	0	1574	396	176	98	62	42	30	22
0.60	408	1574	0	1511	376	165	91	57	38	27
0.65	183	396	1511	0	1417	349	151	83	51	34
0.70	103	176	376	1417	0	1291	313	134	72	43
0.75	66	98	165	349	1291	0	1134	270	113	59
0.80	45	62	91	151	313	1134	0	945	219	88
0.85	33	42	57	83	134	270	945	0	726	160
0.90	25	30	38	51	72	113	219	726	0	474
0.95	19	22	27	34	43	59	88	160	474	0

Resources

As of February 15, 2005, some useful sample size calculators for a wide range of situations may be found at

- the UCLA Department of Statistics website
- Russell Lenth's website. Your browser must be enabled to run these Java applets. They may be downloaded to your personal computer for those times when an Internet connection is unavailable.

[back to LHSP]

Last modified: undefined.

Sample Size Calculations
Surveys

In a survey, there's usually no hypothesis being tested. The sample size determines the precision with which population values can be estimated. The usual rules apply--to cut the uncertainty (for example, the length of a confidence interval) in half, quadruple the sample size, and so on. The sample size for a survey, then, is determined by asking the question, "How accurately do you need to know something?" Darned if I know!

Sometimes imprecise estimates are good enough. Suppose in some underdeveloped country a 95% confidence interval for the proportion of children with compromised nutritional status was (20%, 40%). Even though the confidence interval is quite wide, every value in that interval points to a problem that needs to be addressed. Even 20% is too high. Would it help (would it change public policy) to know the true figure more precisely?

In his book *Sampling Techniques, 3rd ed.* (pp 72-74), William Cochran gives the example of an anthropologist who wishes to know the percentage of inhabitants of some island who belong to blood group O. He decides he needs to know this to within 5%. Why 5%? Why not 4% or 6%. I don't know. Neither does Cochran. Cochran asks!

> He strongly suspects that the islanders belong either to a racial type with a P of about 35% or to one with a P of about 50%. An error limit of 5% in the estimate seemed to him small enough to permit classification into one of these types. He would, however, have no violent objection to 4 or 6% limits of error.
>
> Thus the choice of a 5 %limit of error by the anthropologist was to some extent arbitrary. In this respect the example is typical of the way in which a limit of error is often decided on. In fact, the anthropologist was more certain of what he wanted than many other scientists and administrators will be found to be. When the question of desired degree of precision is first raised, such persons may confess that they have never thought about it and have no idea of the answer. My experience has been, however, that after discussion they can frequently indicate at least roughly the size of a limit of error that appears reasonable to them. [Cochran had a lot of experience with sample surveys. I don't. I have yet to have the experience where investigators can "indicate at least roughly the size of a limit of error that appears reasonable to them" with any degree of confidence or enthusiasm. I find the estimate is given more with resignation.]
>
> Further than this we may not be able to go in many practical situations. Part of the difficulty is that not enough is known about the consequences of errors of different sizes as they affect the wisdom of practical decisions that are made from survey results. Even when these consequences are known, however, the results of many important surveys are

used by different people for different purposes, and some of the purposes are not foreseen at the time when the survey is planned.

Thus, the specification of a sample size for a survey invariably contains a large element of guesswork. The more the survey can be made to resemble a controlled trial with comparisons between groups, the easier it is to come up with sample size estimates.

Sampling Schemes

With *simple random samples*, every possible sample has the same probability of being selected. Estimates of population quantities and their uncertainties are relatively straightforward to calculate. Many surveys are conducted by using random samples that are not simple. Two of the most commonly used alternatives to simple random samples are *stratified random samples* and *cluster samples*.

With *stratified random sampling*, the population is divided into strata and a simple random sample is selected from each stratum. This insures it is possible to make reliable estimates for each stratum as well as for the population as a whole. For example, if a population contains a number of ethnic groups, a simple random sample might contain very few of certain ethnicities. If we were to sample equal numbers of each ethnicity, then characteristics of all ethnicities can be estimated with the same precision. Overall population estimates and their standard errors can be obtained by combining the stratum-specific estimates in the proper way.

For example, suppose a population is 90% white and 10 % black, a stratified sample of 500 whites and 500 blacks is interviewed, and the mean time per day spent watching television is 4 hours for whites and 2 hours for black. Then, the estimated mean number of hours spent watching television for the population combines the two stratum-specific estimates by giving 90% of the weight to the mean for whites and 10% of the weight to the mean for blacks, that is 0.90*4 + 0.10*2 = 3.8 hours. Similar calculations are used to calculate the overall SD.

With *cluster sampling*, the population is divided into clusters. A set of clusters is selected at random and individual units are selected within each cluster. Cluster sampling is typically used for convenience. Imagine a country composed of hundreds of villages. Rather than survey a simple random sample of the population (which might have the survey team visiting every village), it is usually more practical to take a simple random sample of villages and then take a random sample of individuals from each village. A cluster sample is always less precise than a simple random sample of the same size, but it is usually a lot less expensive to obtain. To put it another way, to achieve a specified level of precision it is often less expensive and more convenient to use a larger cluster sample than a smaller simple random sample. Once again, there are special formulas that allow analysts to combine the data from the clusters to calculate estimates and of population quantities and their standard errors.

Many of the sample size calculations for complex surveys involve estimates of quantities that are often unavailable. Levy & Lemeshow (*Sampling of Populations*, New York: John Wiley & Sons, 1991) are explicit about what investigators face: *These quantities are population parameters that in general would be unknown and would have to be either estimated from preliminary studies or else guessed by means of intuition or past experience.* (p 198)

A common method for obtaining sample size calculations for cluster sampling is by performing them as though simple random sampling were being used, except that the variances (SD^2) used in the formulas are multiplied by a *Design Effect* which involves intraclass correlations, a measure of hove much of the variability between subjects is due to the variability between clusters. It has never been clear to me how design effects are estimated in practice. The ones I've seen have invariably been 2.

Over the last decade, statistical program packages have been developed for analyzing data from complex sample surveys. The best known of these is SUDAAN (from SUrvey DAta ANalysis), which is available as a stand-alone program or as an add-on to SAS. Lately, SAS has been adding this functionality to its own program with its SURVEYMEANS and SURVEYREG procedures.

Gerard E. Dallal
Last modified: undefined.

Nonparametric Statistics

Gerard E. Dallal, Ph.D.

Before discussing *non*parametric techniques, we should consider why the methods we usually use are called *parametric*. Parameters are indices. They index (or label) individual distributions within a particular family. For example, there are an infinte number of normal distributions, but each normal distribution is uniquely determined by its mean (μ) and standard deviation (σ). If you specify all of the parameters (here, μ and σ), you've specified a unique normal distribution.

Most commonly used statistical techniques are properly called parametric because they involve estimating or testing the value(s) of parameter(s)--usually, population means or proportions. It should come as no suprise, then, that nonparametric methods are procedures that work their magic without reference to specific parameters.

The precise definition of nonparametric varies slightly among authors[1]. You'll see the terms *nonparametric* and *distribution-free*. They have slightly different meanings, but are often used interchangeably--like *arteriosclerosis* and *atherosclerosis*.

Ranks

Many nonparametric procedures are based on ranked data. Data are ranked by ordering them from lowest to highest and assigning them, in order, the integer values from 1 to the sample size. Ties are resolved by assigning tied values the mean of the ranks they would have received if there were no ties, e.g., 117, 119, 119, 125, 128 becomes 1, 2.5, 2.5, 4, 5. (If the two 119s were not tied, they would have been assigned the ranks 2 and 3. The mean of 2 and 3 is 2.5.)

For large samples, many nonparametric techniques can be viewed as the usual normal-theory-based procedures applied to ranks. The following table contains the names of some normal-theory-based procedures and their nonparametric counterparts. For smaller sample sizes, the same statistic (or one mathematically equivalent to it) is used, but decisions regarding its significance are made by comparing the observed value to special tables of critical values[2].

Some Commonly Used Statistical Tests

Normal theory based test	**Corresponding nonparametric test**	**Purpose of test**
t test for independent samples	Mann-Whitney U test; Wilcoxon rank-sum test	Compares two independent samples
Paired *t* test	Wilcoxon matched pairs signed-rank test	Examines a set of differences
Pearson correlation coefficient	Spearman rank correlation coefficient	Assesses the linear association between two variables.

One way analysis of variance (*F* test)	Kruskal-Wallis analysis of variance by ranks	Compares three or more groups
Two way analysis of variance	Friedman Two way analysis of variance	Compares groups classified by two different factors

Some nonparametric procedures

The *Wilcoxon signed rank test* is used to test whether the median of a symmetric population is 0. First, the data are ranked without regard to sign. Second, the signs of the original observations are attached to their corresponding ranks. Finally, the one sample z statistic (mean / standard error of the mean) is calculated from the signed ranks. For large samples, the z statistic is compared to percentiles of the standard normal distribution. For small samples, the statistic is compared to likely results if each rank was equally likely to have a + or - sign affixed.

The *Wilcoxon rank sum test* (also known as *the Mann-Whitney U test* or the *Wilcoxon-Mann-Whitney test*) is used to test whether two samples are drawn from the same population. It is most appropriate when the likely alternative is that the two populations are shifted with respect to each other. The test is performed by ranking the combined data set, dividing the ranks into two sets according the group membership of the original observations, and calculating a two sample z statistic, using the pooled variance estimate. For large samples, the statistic is compared to percentiles of the standard normal distribution. For small samples, the statistic is compared to what would result if the data were combined into a single data set and assigned at random to two groups having the same number of observations as the original samples.

Spearman's rho (*Spearman rank correlation coefficient*) is the nonparametric analog of the usual Pearson product-moment correlation coefficent. It is calculated by converting each variable to ranks and calculating the Pearson correlation coefficient between the two sets of ranks. For small sample sizes, the observed correlation coefficient is compared to what would result if the ranks of the X- and Y-values were random permuations of the integers 1 to *n* (sample size).

Since these nonparametic procedures can be viewed as the usual parametric procedures applied to ranks, it is reasonable to ask what is gained by using ranks in place of the raw data.

Advantages of nonparametric procedures

(1) Nonparametric test make less stringent demands of the data. For standard parametric procedures to be valid, certain underlying conditions or assumptions must be met, particularly for smaller sample sizes. The one-sample t test, for example, requires that the observations be drawn from a normally distributed population. For two independent samples, the t test has the additional requirement that the population standard deviations be equal. If these assumptions/conditions are violated, the resulting P-values and confidence intervals may not be trustworthy[3]. However, normality is not required for the Wilcoxon signed rank or rank sum tests to produce valid inferences about whether the median of a symmetric population is 0 or whether two samples are drawn from the same population.

(2) Nonparametric procedures can sometimes be used to get a quick answer with little calculation.

Two of the simplest nonparametric procedures are the sign test and median test. The *sign test* can be used with paired data to test the hypothesis that differences are equally likely to be positive or negative, (or, equivalently, that the median difference is 0). For small samples, an exact test of whether the proportion of positives is 0.5 can be obtained by using a binomial distribution. For large samples, the test statistic is

$$(\text{plus - minus})^2 / (\text{plus + minus}) ,$$

where *plus* is the number of positive values and *minus* is the number of negative values. Under the null hypothesis that the positive and negative values are equally likely, the test statistic follows the chi-square distribution with 1 degree of freedom. Whether the sample size is small or large, the sign test provides a quick test of whether two paired treatments are equally effective simply by counting the number of times each treatment is better than the other.

Example: 15 patients given both treatments A and B to test the hypothesis that they perform equally well. If 13 patients prefer A to B and 2 patients prefer B to A, the test statistic is $(13 - 2)^2 / (13 + 2)$ [= 8.07] with a corresponding P-value of 0.0045. The null hypothesis is therefore rejected.

The *median test* is used to test whether two samples are drawn from populations with the same median. The median of the combined data set is calculated and each original observation is classified according to its original sample (A or B) and whether it is less than or greater than the overall median. The chi-square test for homogeneity of proportions in the resulting 2-by-2 table tests whether the population medians are equal.

(3) Nonparametric methods provide an air of objectivity when there is no reliable (universally recognized) underlying scale for the original data and there is some concern that the results of standard parametric techniques would be criticized for their dependence on an artificial metric. For example, patients might be asked whether they feel *extremely uncomfortable* / *uncomfortable* / *neutral* / *comfortable* / *very comfortable*. What scores should be assigned to the comfort categories and how do we know whether the outcome would change dramatically with a slight change in scoring? Some of these concerns are blunted when the data are converted to ranks[4].

(4) A historical appeal of rank tests is that it was easy to construct tables of exact critical values, provided there were no ties in the data. The same critical value could be used for all data sets with the same number of observations because every data set is reduced to the ranks 1,...,*n*. However, this advantage has been eliminated by the ready availability of personal computers[5].

(5) Sometimes the data do not constitute a random sample from a larger population. The data in hand are all there are. Standard parametric techniques based on sampling from larger populations are no longer appropriate. Because there are no larger populations, there are no population parameters to estimate. Nevertheless, certain kinds of nonparametric procedures can be applied to such data by using *randomization models*.

From Dallal (1988):

> Consider, for example, a situation in which a company's workers are assigned in haphazard fashion to work in one of two buildings. After yearly physicals are administered, it appears that workers in one building have higher lead levels in their blood. Standard sampling theory techniques are inappropriate because the workers do not represent samples from a large population--there is no large population. The randomization model, however, provides a means for carrying out statistical tests in such circumstances. The model states that if there were no influence exerted by the buildings, the lead levels of the workers in each building should be no different from what one would observe after combining all of the lead values into a single data set and dividing it in two, at random, according to the number of workers in each building. The stochastic component of the model, then, exists only in the analyst's head; it is not the result of some physical process, except insofar as the haphazard assignment of workers to buildings is truly random.
>
> Of course, randomization tests cannot be applied blindly any more than normality can automatically be assumed when performing a t test. (Perhaps, in the lead levels example, one building's workers tend to live in urban settings while the other building's workers live in rural settings. Then the randomization model would be inappropriate.) Nevertheless, there will be many situations where the less stringent requirements of the randomization test will make it the test of choice. In the context of randomization models, randomization tests are the ONLY legitimate tests; standard parametric test are valid only as approximations to randomization tests. [6]

Disadvantages of nonparametric procedures

Such a strong case has been made for the benefits of nonparametric procedures that some might ask why parametric procedures aren't abandoned entirely in favor of nonparametric methods!

The major disadvantage of nonparametric techniques is contained in its name. Because the procedures are *nonparametric,* there are no parameters to describe and it becomes more difficult to make quantitative statements about the actual difference between populations. (For example, when the sign test says two treatments are different, there's no confidence interval and the test doesn't say by how much the treatments differ.) However, it is sometimes possible with the right software to compute estimates (and even confidence intervals!) for medians, differences between medians. However, the calculations are often too tedious for pencil-and-paper. A computer is required. As statistical software goes though its various iterations, such confidence intervals may become readily available, but I'm still waiting![7]

The second disadvantage is that nonparametric procedures throw away information! The sign test, for example, uses only the signs of the observations. Ranks preserve information about the order of the data but discard the actual values. Because information is discarded, nonparametric procedures can never be as powerful (able to detect existing differences) as their parametric counterparts when parametric tests can be used.

How much information is lost? One answer is given by the asymptotic relative efficiency (ARE) which, loosely speaking, describes the ratio of sample sizes required (parametric to nonparametric) for a parametric procedure to have the same ability to reject a null hypothesis as the corresponding nonparametric procedure. When the underlying distributions are normal (with equal population standard deviations for the two-sample case)

Procedure	**ARE**
sign test	$2/\pi = 0.637$
Wilcoxon signed-rank test	$3/\pi = 0.955$
median test	$2/\pi = 0.637$
Wilcoxon-Mann-Whitney U test	$3/\pi = 0.955$
Spearman correlation coefficient	0.91

Thus, if the data come from a normally distributed population, the usual z statistic requires only 637 observations to demonstrate a difference when the sign test requires 1000. Similarly, the t test requires only 955 to the Wilcoxon signed-rank test's 1000. It has been shown that the ARE of the Wilcoxon-Mann-Whitney test is always at least 0.864, regardless of the underlying population. Many say the AREs are so close to 1 for procedures based on ranks that they are the best reason yet for using nonparametric techniques!

Other procedures

Nonparametric statistics is a field of specialization in its own right. Many procedures have not been touched upon here. These include the Kolmogorov-Smirnov test for the equality of two distribution functions, Kruskal-Wallis one-way analysis of variance, Friedman two-way analysis of variance, and the logrank test and Gehan's generalized Wilcoxon test for comparing two survival distributions. It would not be too much of an exaggeration to say that for every parametric test there is a nonparametric analogue that allows some of the assumptions of the parametric test to be relaxed. Many of these procedures are discussed in Siegel (1956), Hollander and Wolfe (1973) and Lee (1992).

Example

Ellis et al. (1986) report in summary form the retinyl ester concentrations (mg/dl) of 9 normal individuals and 9 type V hyperlipoproteinemic individuals. Although all of the normal individuals have higher concentrations than those of the abnormals, these data are not quite barely significant at the 0.05 level according to the t test using Satterthwaite's approximation for unequal variances. But, even the lowly median test points to substantial differences between the two groups.

```
          Type V hyper-                         Normal
         lipoproteinemic

                1.4                              30.9
                2.5                             134.6
                4.6                              13.6
                0.0                              28.9
                0.0                             434.1
```

```
              2.9                        101.7
              1.9                         85.1
              4.0                         26.5
              2.0                         44.8



     H
     H
     H
     H                              X
     H                             XXXXX X            X
  min--------------------max    min--------------------max
      an H =    2 cases             an X =    2 cases

     mean          2.1444          mean        100.0222
     SD            1.5812          SD          131.7142
     SEM            .5271          SEM          43.9048
     sample size        9          sample size        9


                statistics           P-value    df

          t (separate)    -2.23       .0564     8.0
          t (pooled)      -2.23       .0405    16
          F (variances) 6938.69       .0000     8,  8

         < median   > median
Group 1       9          0
Group 2       0          9          P-value (exact) =  .0000

Wilcoxon-Mann-Whitney test:  P-value =  .0000
Pitman randomization  test:  P-value =  .0000   (data * 1E 0)
```

References

- Bradley JV (1968), Distribution Free Statistical Tests. Prentice Hall: Englewood Cliffs, NJ.
- Dallal GE (1988), "PITMAN: A FORTRAN Program for Exact Randomization Tests," Computers and Biomedical Research, 21, 9-15.
- Ellis JK Russell RM Makraurer FL and Schaefer EJ (1986), "Increased Risk for Vitamin A Toxicity in Severe Hypertriglyceridemia," Annals of Internal Medicine, 105, 877-879.
- Fisher LD and van Belle G (1993), Biostatistics: A Methodology for the Health Sciences. New York: John Wiley & Sons, Inc.
- Hollander M and Wolfe DA (1973), Nonparametric Statistical Methods. New York: John Wiley & Sons, Inc.
- Lee ET (1992), Statistical Methods for Survival Data Analysis. New York: John Wiley & Sons, Inc.
- Lehmann EL (1975), Nonparametrics: Statistical Methods Based on Ranks. San Francisco: Holden-Day, Inc.

- Mehta C and Patel N (1992), StatXact-Turbo: Statistical Software for Exact Nonparametric Inference. Cambridge, MA: CYTEL Software Corporation.
- Siegel S (1956), Nonparametric Satistics. New York: Mc Graw- Hill Book Company, Inc.
- Velleman PF and Wilkinson L (1993), "Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading," The American Statistician, 47, 65-72.

Notes

1. For example:

   > Fisher and van Belle (1993, p. 306): A family of probability distributions is nonparametric if the distributions of the family cannot be conveniently characterized by a few parameters. [For example, all possible continuous distributions.] Statistical procedures that hold or are valid for a nonparametric family of distributions, are called nonparametric statistical procedures.
   >
   > Bradley (1968, p. 15): The terms nonparametric and distribution-free are not synonymous . . . Popular usage, however, has equated the terms . . . Roughly speaking, a nonparametric test is test one which makes no hypothesis about the value of a parameter in a statistical density function, whereas a distribution-free test is one which makes no assumptions about the precise form of the sampled population.
   >
   > Lehmann (1975, p. 58): . . . distribution-free or nonparametric, that is, free of the assumption that [the underlying distribution of the data] belongs to some parametric family of distributions.

2. For small samples, the tables are constructed by straightforward enumeration. For Spearman's correlation coefficient, the possible values of the correlation coefficient are enumerated by holding one set of values held fixed at 1,...,*n* and paired with every possible permutation of 1,...,*n*. For the Wilcoxon signed rank test, the values of the test statistic (whether it be the t statistic or, equivalently, the sum of the positive ranks) are enumerated for all 2^n ways of labelling the ranks with + or - signs. Similar calculations underlie the construction of tables of critical values for other procedures. Because the critical values are based on all possible permutations of the ranks, these procedures are sometimes called *permutation tests*.
3. On the other hand, a violation of the standard assumptions can often be handled by analyzing some transformation of the raw data (logarithmic, square root, and so on). For example, when the within-group standard deviation is seen to be roughly proportional to the mean, a logarithmic transformation will produce samples with approximately equal standard deviations. Some researchers are unnecessarily anxious about transforming data because they view it as tampering. However, it is important to keep in mind that the point of the transformation is to insure the validity of the analysis (normal distribution, equal standard deviations) and *not* to insure a certain type of outcome. Given a choice between two transformations, one that produced a statistically significant result and another that produced an insignificant result, I would always believe the result for which the data more closely met the requirments of the procedure being applied. This is no different from trusting the results of a fasting blood sample, if that is what is required, when both fasting and non-fasting samples are

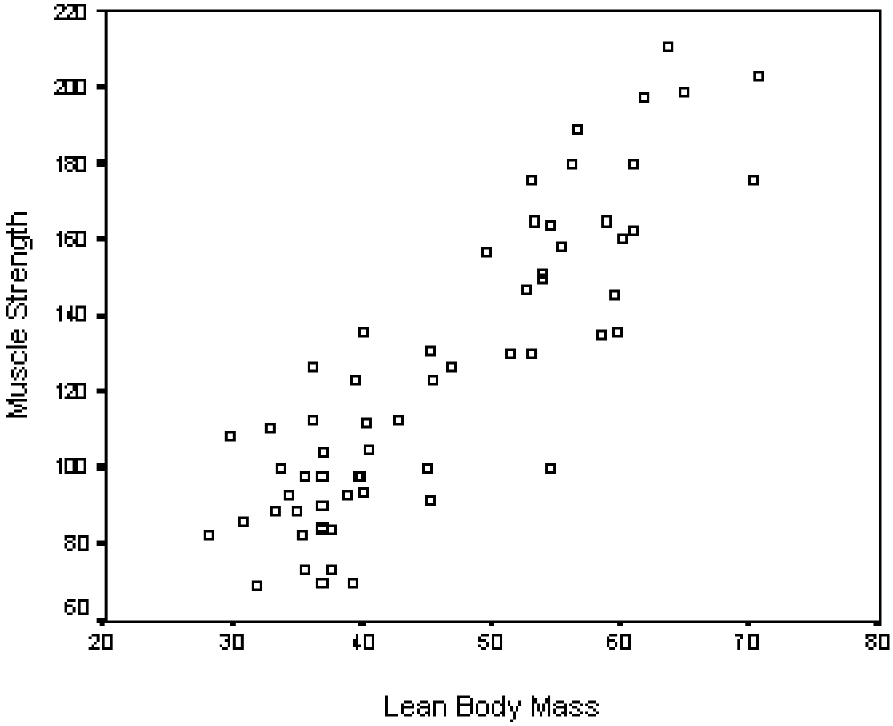
available.

4. Many authors discuss "scales of measurement," using terms such as nominal, ordinal, interval, or ratio data as guides to what statistical procedure can be applied to a data set. The terminology often fails in practice because, as Velleman and Wilkinson (1993) observe, "scale type...is not an attribute of the data, but rather depends upon the questions we intend to ask of the data and upon any additional information we might have." Thus, patient identification number might be ordinarily viewed as a nominal variable (that is, a mere label). However, IDs are often assigned sequentially and in some cases it may prove fruitful to look for relationships between ID and other important variables. While the ideas behind scales of measurement are important, the terminology itself is best ignored. Just be aware that when you score *neutral* as 0, *comfortable* as 1, and *very comfortable* as 2, you should be wary of any procedure that relies heavily on treating "*very comfortable*" as being twice as comfortable as *comfortable*.
5. The ready availability of computers has made much theoretical work concerning approximations and corrections for ties in the data is obsolete, too. Ties were a problem because, with ties, a set of *n* observations does not reduce to the set of ranks 1,...,n. The particular set of ranks depends on the number and pattern of ties. In the past, corrections to the usual z statistic were developed to adjust for tied ranks. Today, critical values for exact nonparametric tests involving data with ties can be calculated on demand by specialized computer programs such as StatXact (Mehta, 1992).
6. The data need not be converted to ranks in order to perform a permutation test. However, if the raw data are used, a critical value must be calculated for the specific data set if the sample size is small or moderate. (The usual t test has been shown to be a large sample approximation to the permutation test!) At one time, the computational complexity of this task for moderate and even small samples was considered a major disadvantage. It has become largely irrelevant due to specialized computer programs that perform the calculations in an efficient manner.
7. This illustrates an often unspoken aspect of statistical computing: **We are prisoners of our software!** Most analysts can do only what their software allows them to do. When techniques become available in standard software packages, they'll be used. Until then, the procedures stay on the curio shelf. The widespread availability of personal computers and statistical program packages have caused a revolution in the way data are analyzed. These changes continue with the release of each new package and update.

Last modified: undefined.

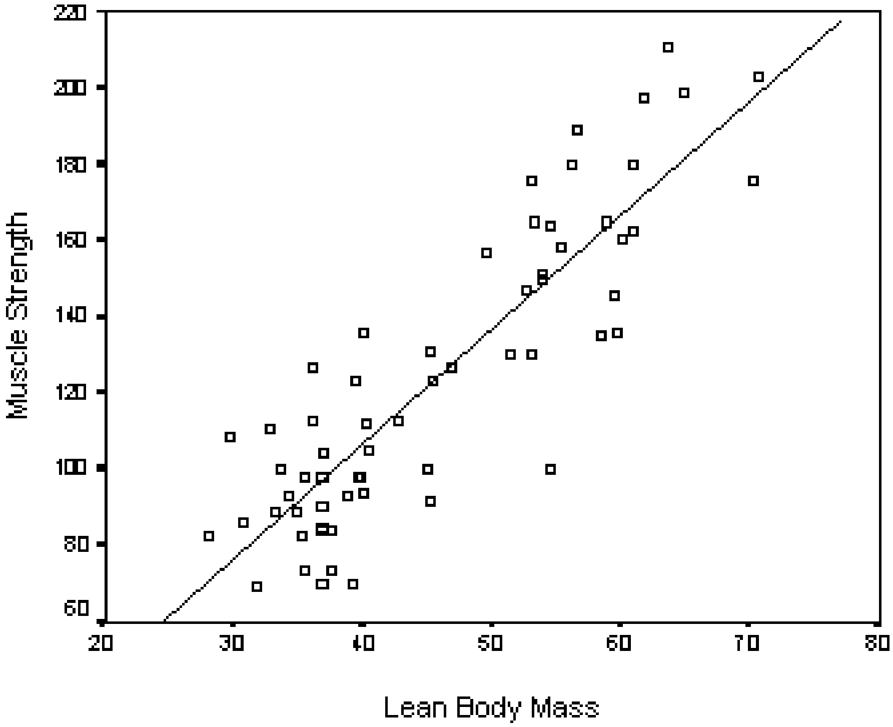
Introduction to Simple Linear Regression

Gerard E. Dallal, Ph.D.



How would you characterize this display of muscle strength[1] against lean body mass? Those who have more lean body mass tend to be stronger. The relationship isn't perfect. It's easy to find two people where the one with more lean body mass is the weaker, but in general strength and lean body mass tend to go up and down together. *Comment*: When two variables are displayed in a scatterplot and one can be thought of as a response to the other (here, muscles produce strength), standard practice is to place the response on the vertical (or Y) axis. The names of the variables on the X and Y axes vary according to the field of application. Some of the more common usages are

X-axis	**Y-axis**
independent	dependent
predictor	predicted
carrier	response
input	output



The association looks like it could be described by a straight line. There are many straight lines that could be drawn through the data. How to choose among them? On the one hand, the choice is not that critical because all of the reasonable candidates would show strength increasing with mass. On the other hand, a standard procedure for fitting a straight line is essential. Otherwise, different analysts working on the same data set would produce different fits and it would make communication difficult. Here, the fitted equation is

Strength = -13.971 + 3.016 LBM .

It says an individual's strength is predicted by multiplying lean body mass by 3.016 and subtracting 13.971. It also says the strength of two individuals is expected to differ by 3.016 times their difference in lean body mass.

The analysis is always described as *the regression of the **response** on the **carrier***. Here, the example involves "the regression of muscle strength on lean body mass", not the other way around.

The Regression Equation

[Standard notation: The data are pairs of independent and dependent variables {(x_i,y_i): i=1,...,n}. The fitted equation is written $\hat{y} = b_0 + b_1 x$, where $\hat{y}$ is the predicted value of the response obtained by using the equation. The *residuals* are the differences between the observed and the predicted values $\{(y_i - \hat{y}_i): i = 1, \ldots, n\}$. They are *always* calculated as (observed-predicted), never the other way 'round.]

There are two primary reasons for fitting a regression equation to a set of data--first, to describe the data; second, to predict the response from the carrier. The rationale behind the way the regression line is calculated is best seen from the point-of-view of prediction. A line gives a good fit to a set of data if the points are close to it. Where the points are not tightly grouped about any line, a line gives a good fit if the points are closer to it than to any other line. For predictive purposes, this means that the predicted values obtained by using the line should be close to the values that were actually observed, that is, that the residuals should be small. Therefore, when assessing the fit of a line, the vertical distances of the points to the line are the only distances that matter. Perpendicular distances are not considered because errors are measured as vertical distances, not perpendicular distances.

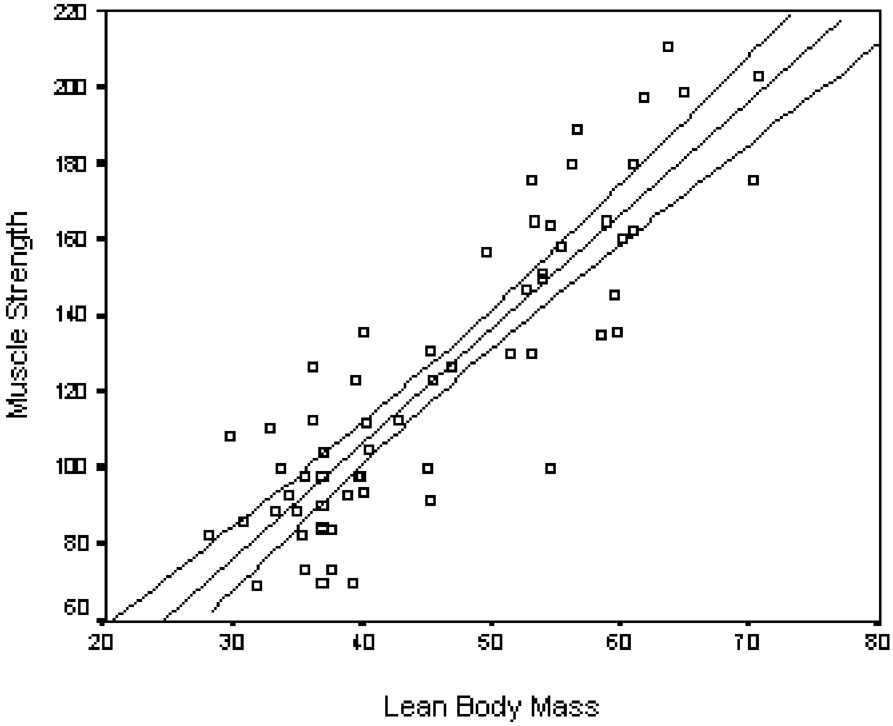
The simple linear regression equation is also called the *least squares* regression equation. Its name tells us the criterion used to select the best fitting line, namely that the sum of the *squares* of the residuals should be *least*. That is, the least squares regression equation is the line for which the sum of squared residuals $\sum (y_i - \hat{y}_i)^2$ is a minimum.

It is not necessary to fit a large number of lines by trial-and-error to find the best fit. Some algebra shows the sum of squared residuals will be minimized by the line for which

$$b_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}, \quad b_0 = \overline{y} - b_1\overline{x}$$

This can even be done by hand if need be.

When the analysis is performed by a statistical program package, the output will look something like this.



A straight line can be fitted to any set of data. The formulas for the coefficients of the least squares fit are the same for a sample, a population, or any arbitrary batch of numbers. However, regression is usually used to let analysts generalize from the sample in hand to the population from which the sample was drawn. There *is* a population regression equation,

$$\beta_0 + \beta_1 X$$

and

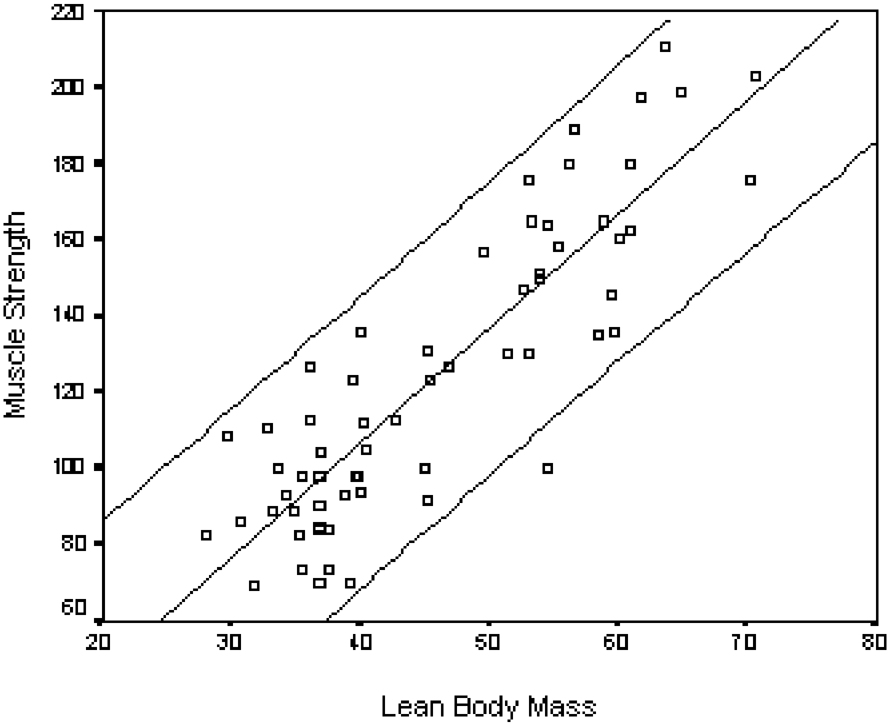
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where β_0 and β_1 are the population regression coefficients and ε_i is a random error peculiar to the i-th observation. Thus, each response is expressed as the sum of a value predicted from the corresponding X, plus a random error.

The sample regression equation is an estimate of the population regression equation. Like any other estimate, there is an uncertainty associated with it. The uncertainty is expressed in confidence bands about the regression line. They have the same interpretation as the standard error of the mean, except that the uncertainty varies according to the location along the line. The uncertainty is least at the sample mean of the Xs and gets larger as the distance from the mean increases. The regression line is like a stick nailed to a wall with some wiggle to it. The ends of the stick will wiggle more than the center. The distance of the confidence bands from the regression line is

$$t\, s_e \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum (x_i - \overline{x})^2}}\ ,$$

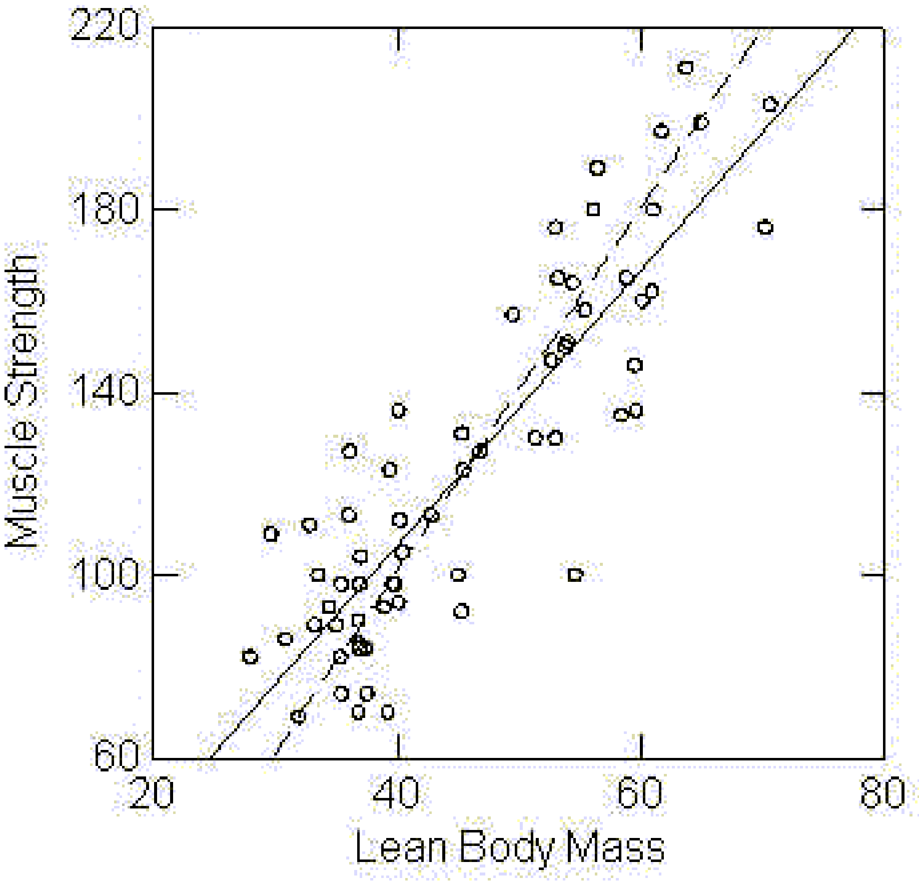
where *t* is the appropriate percentile of the t distribution, s_e is the standard error of the estimate, and x^* is the location along the X-axis where the distance is being calculated. The distance is smallest when $x^* = \overline{x}$. These bands also estimate the population mean value of Y for $X=x^*$.



There are also bands for predicting a single response at a particular value of X. The best estimate is given, once again, by the regression line. The distance of the prediction bands from the regression line is

$$t\, s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum (x_i - \overline{x})^2}}\ .$$

For large samples, this is essentially ts_e, so the standard error of the estimate functions like a standard deviation around the regression line.



The regression of X on Y is different from the regression of Y on X. If one wanted to predict lean body mass from muscle strength, a new model would have to be fitted (dashed line). It could not be obtained by taking the original regression equation and solving for strength. The reason is that in terms of the original scatterplot, the best equation for predicting lean body mass minimizes the errors in the horizontal direction rather than the vertical. For example,

- The regression of Strength on LBM is **Strength = -13.971 + 3.016 LBM** .
- Solving for LBM gives **LBM = 4.632 + 0.332 Strength** .
- However, the regression of LBM on Strength is **LBM = 14.525 + 0.252 Strength** .

Borrowing Strength

Simple linear regression is an example of borrowing strength from some observations to make sharper (that is, more precise) statements about others. If all we wanted to do was make statements about the strength of individuals with specific amounts lean body mass, we could recruit many individuals with that amount of LBM, test them, and report the appropriate summaries (mean, SD, confidence interval,...). We could do this for all of the LBMs of interest. Simple linear regression assumes we don't have to start from scratch for each new amount of LBM. It says that the expected amount of strength is linearly related to LBM. The regression line does two important things. First, it allows us to estimate muscle strength for a particular LBM more accurately than we could with only those subjects with the particular LBM. Second, it allows us to estimate the muscle strength of individuals with amounts of lean body mass that aren't in our sample!

These benefits don't come for free. The method is valid only insofar as the data follow a straight line, which is why it is essential to examine scatterplots.

Interpolation and Extrapolation

Interpolation is making a prediction within the range of values of the predictor in the sample used to generate the model. Interpolation is generally safe. One could imagine odd situations where an investigator collected responses at only two values of the predictor. Then, interpolation might be uncertain since there would be no way to demonstrate the linearity of the relationship between the two variables, but such situations are rarely encountered in practice. *Extrapolation* is making a prediction outside the range of values of the predictor in the sample used to generate the model. The more removed the prediction is from the range of values used to fit the model, the riskier the prediction becomes because there is no way to check that the relationship continues to be linear. For example, an individual with 9 kg of lean body mass would be expected to have a strength of -4.9 units. This is absurd, but it does not invalidate the model because it was based on lean body masses in the range 27 to 71 kg.

----

[1] The particular measure of strength is slow right extensor peak torque in the knee.

How to Read the Output From Simple Linear Regression Analyses

This is the typical output produced from a simple linear regression of muscle strength (STRENGTH) on lean body mass (LBM). That is, lean body mass is being used to predict muscle strength.

Model Summary(b)

R	R Square	Adjusted R Square	Std. Error of the Estimate
.872(a)	.760	.756	19.0481
a Predictors: (Constant), LBM			
b Dependent Variable: STRENGTH			

ANOVA

Source	Sum of Squares	df	Mean Square	F	Sig.
Regression	68788.829	1	68788.829	189.590	.000
Residual	21769.768	60	362.829		
Total	90558.597	61			

Coefficients

Variable	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	-13.971	10.314		-1.355	.181	-34.602	6.660
LBM	3.016	.219	.872	13.769	.000	2.577	3.454

Table of Coefficients

The column labeled **Variable** should be self-explanatory. It contains the names of the items in the equation and labels each row of output.

The **Unstandardized coefficients (B)** are the regression coefficients. The regression equation is

STRENGTH = -13.971 + 3.016 LBM

The predicted muscle strength of someone with 40 kg of lean body mass is

$$-13.971 + 3.016\ (40) = 106.669$$

For cross-sectional data like these, the regression coefficient for the predictor is the difference in response per unit difference in the predictor. For longitudinal data, the regression coefficient is the change in response per unit change in the predictor. Here, strength differs 3.016 units for every unit difference in lean body mass. The distinction between cross-sectional and longitudinal data is still important. These strength data are cross-sectional so differences in LBM and strength refer to differences between people. If we wanted to describe how an individual's muscle strength changes with lean body mass, we would have to measure strength and lean body mass as they change within people.

The **Standard Errors** are the standard errors of the regression coefficients. They can be used for hypothesis testing and constructing confidence intervals. For example, the standard error of the STRENGTH coefficient is 0.219. A 95% confidence interval for the regression coefficient for STRENGTH is constructed as $(3.016 \pm k\ 0.219)$, where *k* is the appropriate percentile of the t distribution with degrees of freedom equal to the Error DF from the ANOVA table. Here, the degrees of freedom is 60 and the multiplier is 2.00. Thus, the confidence interval is given by $(3.016 \pm 2.00\ (0.219))$. If the sample size were huge, the error degress of freedom would be larger and the multiplier would become the familiar 1.96.

The **Standardized coefficients (Beta)** are what the regression coefficients would be if the model were fitted to standardized data, that is, if from each observation we subtracted the sample mean and then divided by the sample SD. People once thought this to be a good idea. It isn't, yet some packages continue to report them. Other packages like SAS do not. We will discuss them later when we discuss multiple regression.

The **t** statistic tests the hypothesis that a population regression coefficient β is 0, that is, H_0: $\beta = 0$. It is the ratio of the sample regression coefficient B to its standard error. The statistic has the form (estimate - hypothesized value) / SE. Since the hypothesized value is 0, the statistic reduces to Estimate/SE. If, for some reason, we wished to test the hypothesis that the coefficient for STRENGTH was 1.7, we could calculate the statistic (3.016-1.700)/0.219.

Sig. labels the **two-sided P values** or **observed significance levels** for the t statistics. The degrees of freedom used to calculate the P values is given by the Error DF from the ANOVA table. The P value for the independent variable tells us whether the independent variable has statistically signifiant predictive capability.

In theory, the P value for the constant could be used to determine whether the constant could be removed from the model. In practice, we do not usually do that. There are two reasons for this.

1. When there is no constant, the model is

$$Y = b_1 X\ ,$$

which forces Y to be 0 when X is 0. Even this is condition is appropriate (for example, no lean body mass means no strength), it is often wrong to place this constraint on the regression line. Most studies are performed with the independent variable far removed from 0. While a straight line may be appropriate for the range of data values studied, the relationship may not be a straight line all the way down to values of 0 for the predictor.

2. Standard practice (hierarchical modeling) is to include all simpler terms when a more complicated term is added to a model. Nothing is simpler than a constant. So if a change of Y with X is to be place in a model, the constant should be included, too. It could be argued this is a variant of (1).

The Analysis of Variance Table

The **Analysis of Variance** table is also known as the **ANOVA table** (for ANalysis Of VAriance). It tells the story of how the regression equation accounts for variablity in the response variable.

The column labeled **Source** has three rows: Regression, Residual, and Total. The column labeled **Sum of Squares** describes the variability in the response variable, Y.

The total amount of variability in the response is the **Total Sum of Squares**, $\sum (y_i - \bar{y})^2$. (The row labeled **Total** is sometimes labeled **Corrected Total**, where *corrected* refers to subtracting the sample mean before squaring and summing.) If a prediction had to be made without any other information, the best that could be done, in a certain sense, is to predict every value to be equal to the sample mean. The error--that is, the amount of variation in the data that can't be accounted for by this simple method--is given by the Total Sum of Squares.

When the regression model is used for prediction, the error (the amount of uncertainty that remains) is the variability about the regression line, $\sum (y_i - \hat{y}_i)^2$. This is the **Residual Sum of Squares** (*residual* for *left over*). It is sometimes called the Error Sum of Squares. The **Regression Sum of Squares** is the difference between the **Total Sum of Squares** and the Residual Sum of Squares. Since the **total sum of squares** is the total amount of variablity in the response and the **residual sum of squares** that still cannot be accounted for after the regression model is fitted, the **regression sum of squares** is the amount of variablity in the response that is accouned for by the regression model.

Each sum of squares has a corresponding degrees of freedom (DF) associated with it. Total df is *n-1*, one less than the number of observations. The Regression df is the number of independent variables in the model. For simple linear regression, the Regression df is 1. The Error df is the difference between the Total df and the Regression df. For simple linear regression, the residual df is *n-2*.

The **Mean Squares** are the Sums of Squares divided by the corresponding degrees of freedom.

The **F** statistic, also known as the **F ratio**, will be described in detail during the discussion of multiple regression. When there is only one predictor, the F statistic will be the square of the predictor variable's t statistic.

R^2 is the squared multiple correlation coefficient. It is also called the **Coefficient of Determination**. R^2 is the Regression sum of squares divided by the Total sum of squares, RegSS/TotSS. It is the fraction of the variability in the response that is accounted for by the model. Since the Total SS is the sum of the Regression and Residual Sums of squares, R^2 can be rewritten as (TotSS-ResSS)/TotSS = 1- ResSS/TotSS. Some call R^2 *the proportion of the variance explained by the model*. I don't like the use of the word *explained* because it implies causality. However, the phrase is firmly entrenched in the literature. Even Fisher used it. If a model has perfect predictability, the Residual Sum of Squares will be 0 and R^2=1. If a model has no predictive capability, R^2=0. In practice, R^2 is never observed to be exactly 0 the same way the difference between the means of two samples drawn from the same population is never exaxctly 0 or a sample correlation coefficient is never exactly 0.

R, the multiple correlation coefficient and square root of R^2, is the correlation between the predicted and observed values. In simple linear regression, R will be equal to the magnitude correlation coefficient between X and Y. This is because the predicted values are b_0+b_1X. Neither multiplying by b_1 or adding b_0 affects the magnitude of the correlation coefficient. Therefore, the correlation between X and Y will be equal to the correlation between b_0+b_1X and Y, except for their sign if b_1 is negative.

Adjusted-R^2 will be described during the discussion of multiple regression.

The **Standard Error of the Estimate** (also known as the **Root Mean Square Error**) is the square root of the Residual Mean Square. It is the standard deviation of the data about the regression line, rather than about the sample mean. That is, it is

$$\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \text{ rather than } \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

Last modified: undefined.

Correlation and Regression

Correlation and regression are intimately related. The sample correlation coefficient between X and Y is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

When Y is regressed on X, the regression coefficient of X is

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Therefore, the regression coefficient is the correlation coefficent multiplied by the ratio of the standard deviations.

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{s_y}{s_x} r$$

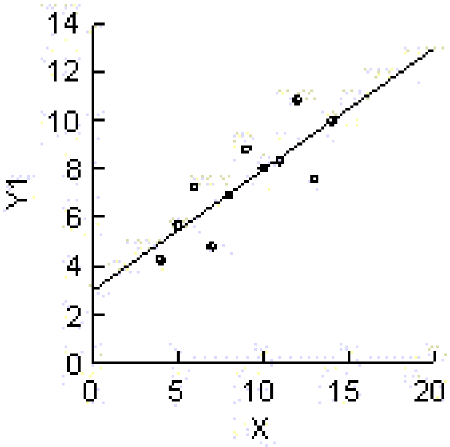
Since the ratio of standard deviatons is always positive, testing whether the population regression coefficient is 0 is equivalent to testing whether the population correlation coefficient is 0. That is, the test of H_0: $\beta_1 = 0$ is equivalent to the test of H_0: $\rho = 0$.

While correlation and regression are intimately related, they are not equivalent. The regression equation can be estimated whenever the Y values result from random sampling. The Xs can result from random sampling or they can be specified by the investigator. For example, crop yield can be regressed on the amount of water crops are given regardless of whether the water is rainfall (random) or the result of turning on an irrigation system (by design). The correlation coefficient is a characteristic of the joint distribution of X and Y. In order to estimate the correlation coefficient, both variables must be the result of random sampling. It makes sense to talk about the correlation between yield and rainfall, but it does not make sense to talk about the correlation between yield and amounts of water under the researcher control. This latter correlation will vary according to the specific amounts used in the study. In general, the correlation coefficient will increase or decrease along with the range of the values of the predictor.

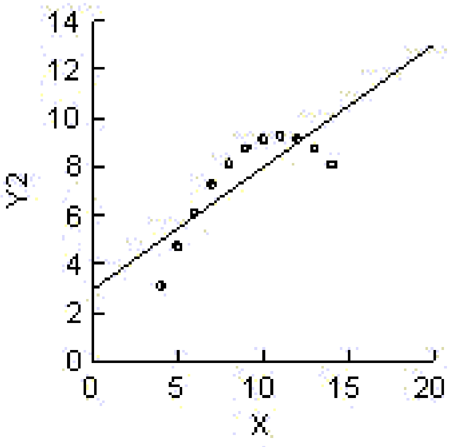
Last modified: undefined.

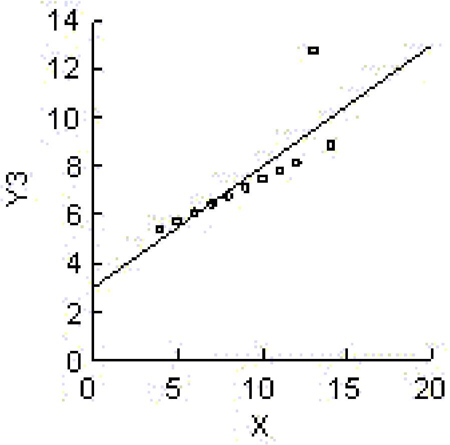
Frank Anscombe's Regression Examples

The intimate relationship between correlation and regression raises the question of whether it is possible for a regression analysis to be misleading in the same sense as the set of scatterplots all of which had a correlation coefficient of 0.70. In 1973, Frank Anscombe published a set of examples showing the answer is a definite yes (Anscombe FJ (1973), "Graphs in Statistical Analysis," The American Statistician, 27, 17-21). Anscombe's examples share not only the same correlation coefficient, but also the same value for any other summary statistic that is usually calculated.



n	11
$\bar{x}$	9.0
$\bar{y}$	7.5
Regression equation of y on x	y = 3 + 0.5 x
$\sum (x_i - \bar{x})^2$	110.0
Regression SS	27.5
Residual SS	13.75 (9 df)
Estimated SE of b_1	0.118
r	0.816
R^2	0.667





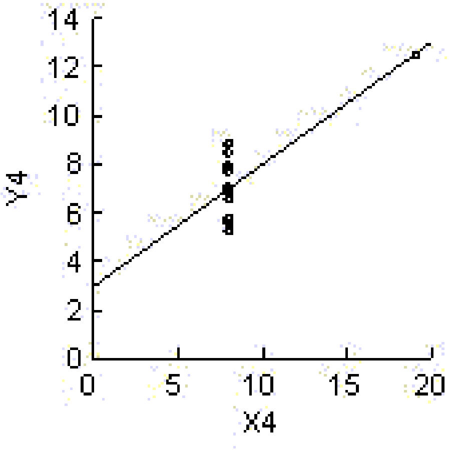


Figure 1 is the picture drawn by the mind's eye when a simple linear regression equation is reported. Yet, the same summary statistics apply to figure 2, which shows a perfect curvilinear relation, and to figure 3, which shows a perfect linear relation except for a single outlier.

The summary statistics also apply to figure 4, which is the most troublesome. Figures 2 and 3 clearly call the straight line relation into question. Figure 4 does not. A straight line may be appropriate in the fourth case. However, the regression equation is determined entirely by the single observation at x=19. Paraphrasing Anscombe, we need to know the relation between y and x *and* the special contribution of the observation at x=19 to that relation.

x	y1	y2	y3	x4	y4
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.10	8.84	8	7.04
6	7.24	6.13	6.08	8	5.25
4	4.26	3.10	5.39	19	12.50
12	10.84	9.13	8.15	8	5.56
7	4.82	7.26	6.42	8	7.91
5	5.68	4.74	5.73	8	6.89

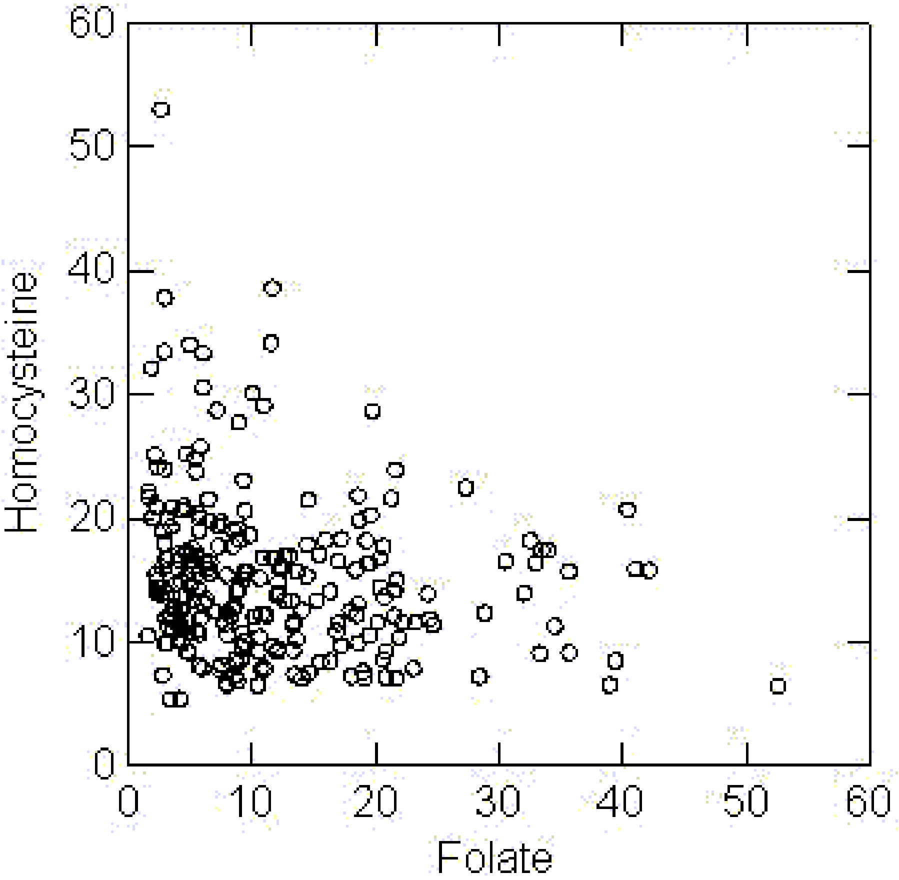
Last modified: undefined.

Transformations In Linear Regression

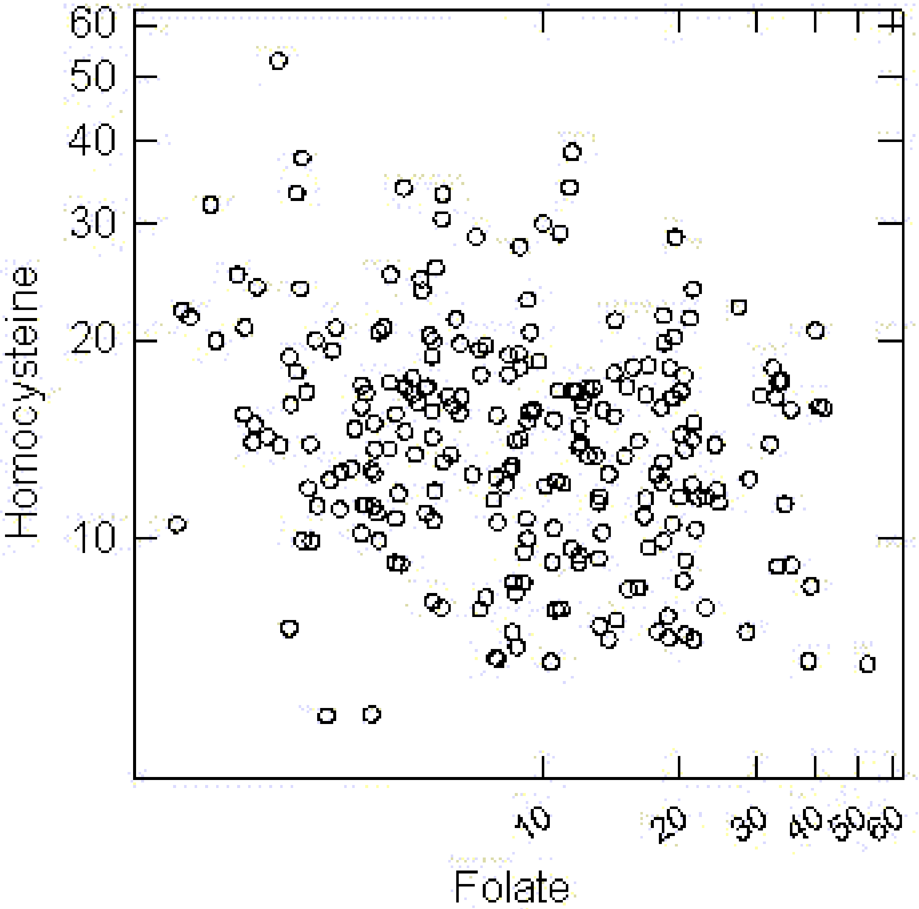
There are many reasons to transform data as part of a regression analysis.

- to achieve linearity.
- to achieve homogeneity of variance, that is, constant variance about the regression equation.
- to achieve normality or, at least, symmetry about the regression equation.

A transformation that achieves one of these goals often ends up achieving all three. This sometimes happens because when data have a multivariate normal distribution, the linearity of the regression and homogeneity follow automatically. So anything that makes a set of data look multivariate normal in one respect often makes it look multivariate normal in other respects. However, it is not necessary that data follow a multivariate normal distribution for multiple linear regression to be valid. For standard tests and confidence intervals to be reliable, the responses should be close to normally distributed with constant variance about their predicted values. The values of the predictors need not be a random sample from *any* distribution. They may have any arbitrary joint distribution without affecting the validity of fitting regression models.



Here are some data where the values of both variables were obtained by sampling. They are the homocysteine and folate (as measured by CLC) levels for a sample of individuals. Both variables are skewed to the right and the joint distribution does not have an elliptical shape. If a straight line was fitted to the data with HCY as a response, the variability about the line would be much greater for smaller values of folate and there is a suggestion that the drop in HCY with increasing vitamin status is greater at lower folate levels.



When logarithmic transformations are applied to both variables, the distributions of the individual variables are less skewed and their joint distributions is roughly ellipsoidal. A straight line seem a like reasonable candidate for describing the association between the variables and the variances appear to be roughly constant about the line.

Often both variables will not need to be transformed and, even when two transformations are necessary, they may not be the same, When only one variable needs to be transformed in a simple linear regression, should it be the response or the predictor? Consider a data set showing a quadratic (parabolic) effect between Y and X. There are two ways to remove the nonlinearity by transforming the data. One is to square the predictor; the other is to take the square root of the response. The rule that is used to determine the approach is, "First, transform the Y variable to achieve homoscedasticity (constant variance). Then, transform the X variable to achieve linearity."

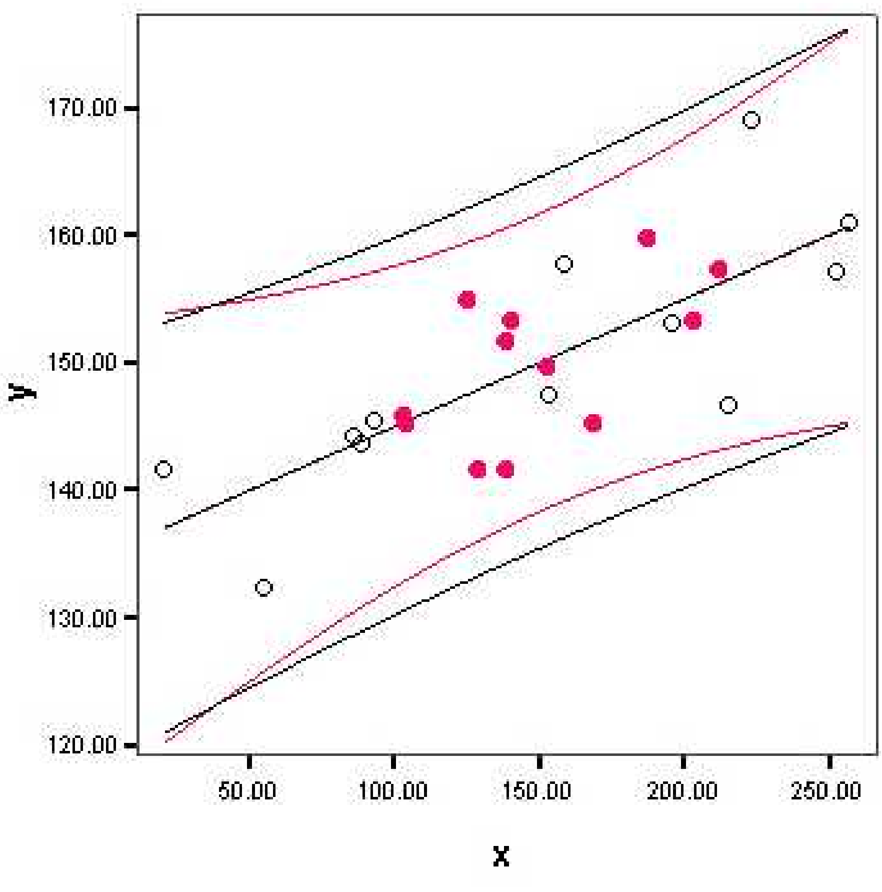
Transforming the X variable does little to change distribution of the data about the (possibly nonlinear) regression line. Transforming X is equivalent to cutting the joint distribution into vertical slices and changing the spacing of the slices. This doesn't do anything to the vertical locations of data within the slices. Transforming the Y variable not only changes the shape of regression line, but it alters the relative vertical spacing of the observations. Therefore, it has been suggested that the Y variable be transformed first to achieve constant variance around a possibly non-linear regression curve and then the X variable be transformed to make things linear.

Last modified: undefined.

Which fit is better?

Sometimes the same model is fitted to two different populations. For example, an researcher might wish to investigate whether weight predicts blood pressure in smokers and nonsmokers and, if so, whether the regression model fits one group better than the other. The problem with questions like this is that the answer depends on what we mean by *better*.

It is common to hear investigators speak of the model with the larger **coefficient of determination**, R^2, as though it fits better because it accounts for more of the variability in the response. However, it is possible for the model with the smaller R^2 to have the smaller standard error of the estimate and make more precise predictions. Here is a small dataset to illustrate this behavior.



X	Y	X	Y
158.2	157.8	140.4	153.2
214.9	146.6	211.9	157.4
153.2	147.5	152.4	149.6
196.0	153.1	124.7	154.9
88.5	143.7	103.9	145.2
55.5	132.3	128.5	141.7
86.4	144.3	187.1	159.7
223.6	169.1	168.5	145.3
256.9	160.9	138.3	151.7
252.4	157.1	137.9	141.7
20.9	141.6	203.3	153.3
92.9	145.4	102.5	145.8

The two data sets need not have the same regression line, but they have been constructed with the same regression line in this example to remove any suspicion that these results might have something to do with the slopes of the regression lines. They don't!

	$Y = 134.9 + 0.100 X$	
	R^2	**s_e**

Red	0.36	**5.04**
Black	**0.64**	6.28

The **black** data set, with open circles and outer prediction bands, has the **larger R^2**. The **red** data set, with filled circles and inner prediction bands, has the **smaller s_e**.

Does the model fit one group better than the other? I try to avoid questions demanding one word answers where the answer depends on the choice of summary measure. However, if pressed, I would argue that the answer is red. R^2 is just a disguised correlation coefficient (the square of the correlation between the observed and predicted values). I have yet to encounter a real research question for which the answer is "correlation coefficient". If I were to use "better" in connection with linear regression it would almost certainly have something to do with prediction. The standard error of the estimate (s_e) estimates the precision of the predictions. The accuracy of the predictions typically determines whether the regression equation will be useful. While the regression equation may account for more variability in the black group, the predictions are more precise in the red group.

Mathematical Details

R^2 can be written as

$$R^2 = 1 - \text{Residual SS}/\text{Total SS} ,$$

while s^2_e can be written as

$$s^2_e = \text{Residual SS} / (n-2)$$

The fit with the larger R^2 is the one that accounts for the greater *proportion* of the variability in the response, that is, the one for which **Residual SS/Total SS is smaller**. The fit with the smaller s_e is the one that leaves the smaller *amount* of variability unaccounted for, that is, the one for which **Residual SS/(n-2) is smaller**. If the sample sizes are equal the model with the smaller s_e is the one for which **Residual SS is smaller**. The model for which the ratio (Residual SS/Total SS) is smaller need not be the same model for which the numerator (Residual SS) is smaller.

Comment

These results apply when the same model is fitted to two different sets of observations. If two models were fitted to the same set of responses--for example, if weight and amount of exercise were used separately to predict blood pressure in the same set of individuals--then the model for which R^2 is larger would necessarily be the model for which s_e is smaller. That's because Total SS would be the same for both, so the model for which Residual SS/Total SS is smaller must also be the one for which Residual SS is smaller.

[back to The Little Handbook of Statistical Practice]

Last modified: undefined.

The Regression Effect / The Regression Fallacy

Gerard E. Dallal, Ph.D.

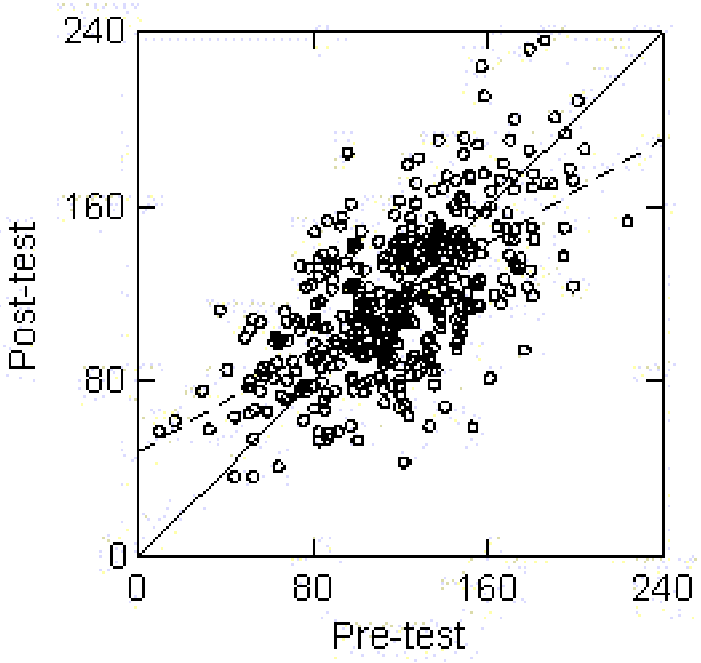
The Regression Effect

Suppose you were told that when any group of subjects with low values on some measurement is later remeasured, their mean value will increase without the use of any treatment or intervention. Would this worry you? *It sure had better!*

If this were true and an ineffective treatment were applied to a such a group, the increase might be interpreted improperly as a treatment effect. This could result in the costly implementation of ineffective programs or faulty public policies that block the development of real solutions for the problem that was meant to be addressed.

The behavior described in the first paragraph is real. It is called **the regression effect**. Unfortunately, the misinterpretation of the regression effect described in the second paragraph is real, too. It is called **the regression fallacy**.

The regression effect is shown graphically and numerically in the following series of plots and computer output.



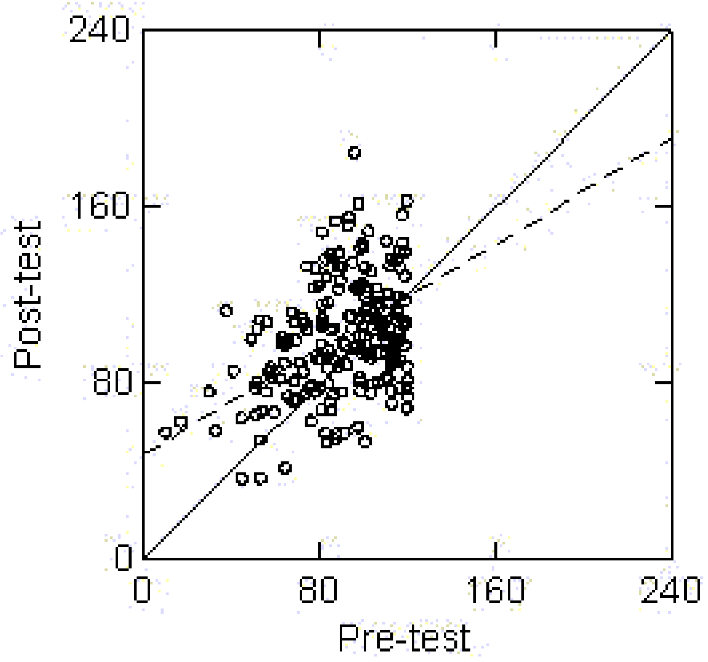
```
                 (Full Data Set)

                    PRE
POST        DIFF
N of cases          400
400         400
Mean            118.420
118.407      -0.012
Std. Error        1.794
1.718       1.526
Standard Dev     35.879
34.364      30.514


      Mean Difference =        -
0.012
```

```
        SD Difference =
30.514
             paired t =       -0.008 (399 df)
                    P =        0.994
```

The first plot and piece of output show a sample of pre-test and post-test measurements taken before and after the administration of an ineffective treatment. The observations don't lie exactly on a straight line because the measurement is not entirely reproducible. In some cases, a person's pre-test measurement will be higher than the post-test measurement; in other cases the post-test measurement will be higher. Here, the pre- and post-test means are 118; the standard deviations are 35. The mean difference is 0 and the t test for equality of population means yields a P value of 0.994. There is no change in the mean or SD over time.

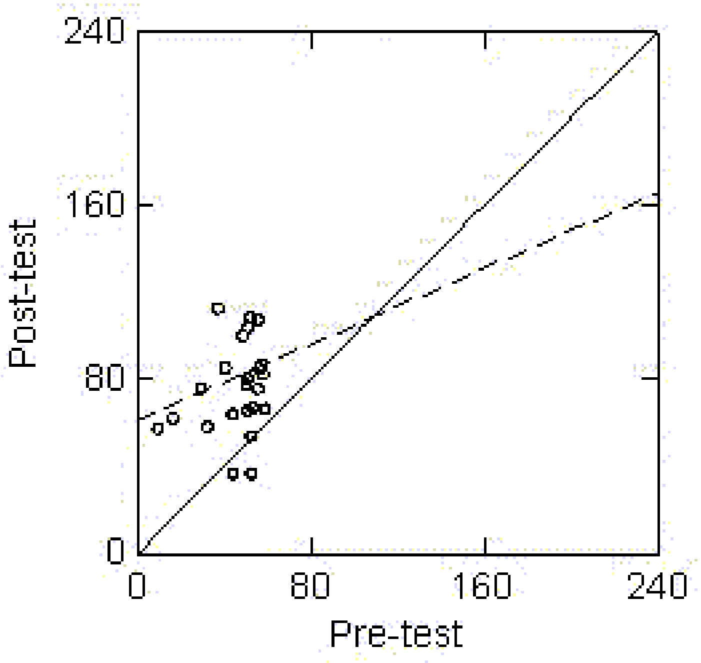


```
          (Observations with PRE
<= 120)

                    PRE
POST        DIFF
N of cases          201
201         201
Mean              90.029
100.503      10.474
Std. Error         1.580
1.844       1.930
Standard Dev      22.394
26.140      27.358
```

```
                                         Mean Difference =
10.474
        SD Difference =       27.358
             paired t =        5.428 (200 df)
                    P =       <0.001
```

The second plot and piece of output show what happens when post-test measurements are made only on those with pre-test measurements less than 120. In the plot, many more observations lie above the line PRE=POST than below it. The output shows that the pre-test mean is 90 while the post-test mean is 100, some 10 units higher (P < 0.001)!



```
                   (Observations with PRE
<= 60)

                          PRE
POST        DIFF
N of cases                 23
23          23
Mean                   46.060
76.111      30.050
Std. Error              2.733
4.441       4.631
Standard Dev       13.107
21.301      22.209
```

```
   Mean Difference =       30.050
     SD Difference =       22.209
          paired t =        6.489  (22 df)
                 P =       <0.001
```

The third plot and final piece of output show what happens when a post-test measurements is taken only for those with pre-test measurements less than 60. In the plot, most observations lie above the line PRE=POST. The output shows that the pre-test mean is 46 while the post-test mean is 76, some *30* units higher (P < 0.001)!

This is how an ineffective treatment behaves. The plots and output clearly demonstrate how an analyst could be misled into interpreting the the regression effect as a treatment effect.

A Closer Look

The regression effect causes an individual's expected post-test measurement to fall somewhere between her pre-test measurement and the mean pre-test measurement. Those with very low pre-test measurements will see their average move up toward the overall mean while those with high pre-test measurements will see them move down. This is how regression got its name--Sir Francis Galton noticed that the sons of tall fathers tended to be shorter than their fathers while sons of short fathers tended to be taller. The sons "regressed to the mean".

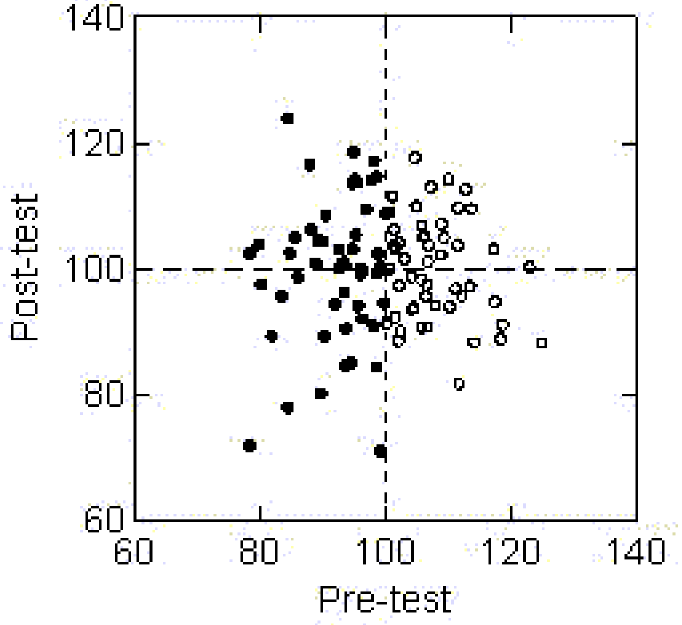
This happens because there are two types of people with very low pre-test measurements: those who are truly low, and those with higher underlying values but appear low due to random variation. When post-test measurements are made, those who are truly low will tend to stay low, but those with higher underlying scores will tend to migrate up toward overall the mean, dragging the group's mean post-test measurement with them. A similar argument applies to those with pre-test measurements greater than the overall mean.

Another Approach

Another way to get a feel for the regression effect is to consider a situation where the pre-test and post-test measurements are completely uncorrelated. If the measurements are uncorrelated, then the best estimate of a subject's post-test measurement is the overall mean of the pretest measurements. Consider those subjects whose pre-test measurements are less than the overall mean (filled circles). The mean of these subjects' pre-test values must be *less than* the overall pre-test mean. Yet, their post-test mean will be *equal to* the overall pre-test mean!

```
                 (Full Data Set)


                    PRE
POST        DIFF
N of cases          100
100         100
Mean            100.000
100.000       0.000
Std. Error        1.000
1.000       1.414
Standard Dev     10.000
10.000      14.142
```



```
     Mean Difference =        0.000
       SD Difference =       14.142
            paired t =        0.000  (99 df)
                   P =        1.000

-----------------------------------------------------

          (Observations with PRE <= 100)

                    PRE        POST        DIFF
N of cases           50          50          50
```

```
Mean                92.030      99.966       7.936
Std. Error           0.859       1.650       1.761
Standard Dev         6.072      11.666      12.453

     Mean Difference =           7.936
       SD Difference =          12.453
                   t =           4.506  (49 df)
                   P =          <0.000
```

A Third Approach

When there is no intervention or treatment effect, a plot of post-test measurements against pre-test measurements reflects only the reproducibility of the measurements. If the measurements are perfectly reproducible, the observations will lie on the line POST = PRE and the best prediction of a subject's post-test measurement will be the pre-test measurement. At the other extreme, if there is no reproducibility, the observations will lie in a circular cloud and the best prediction of a subject's post-test measurement will be the mean of all pre-test measurements. The prediction equation, then, is the line POST = mean (PRE).

In intermediate situations, where there is some reproducibility, the prediction equation given by the linear regression of post-test on pre- test lies between the line POST = PRE and the horizontal. This means an individual's post-test measurement is predicted to be somewhere between his pre-test measurement and the overall mean pre-test measurement. Thus, anyone with a pre-test measurement greater than the pretest mean will be predicted to have a somewhat lower post-test measurement, while anyone with a pre-test measurement less than the pretest mean will be predicted to have a somewhat higher post-test measurement.

None of this speaks against regression analysis or in any way invalidates it. The best estimate of an individual's post-test measurement *is* the mean of the post-test measurements for those with the same pre-test score. When the pre- and post-test measurements are uncorrelated, the best estimate of an individual's post-test measurement *is* the mean of the pre-test measurements, regardless of an individual's pre-test measurement. The purpose of this discussion is to make you aware of the way data behave in the absence of any treatment effect so the regression effect will not be misinterpreted when it is encountered in practice.

Change and The Regression Effect

According to the regression effect, *those who have extremely low pretest values show the greatest increase* while *those who have extremely high pretest values show the greatest decrease*. Change is most positive for those with the lowest pretest values and most negative for those with the largest pretest values, that is, **change is negatively correlated with pretest value**.

The Regression Fallacy

The **regression fallacy** occurs when the regression effect is mistaken for a real treatment effect. The regression fallacy is often observed where there is no overall treatment effect, prompting investigators to conduct extensive subset analyses. A typical misstatement is, "While the education program produced no overall change in calcium intake, those with low initial intakes subsequently increased their intake while those with higher initial intakes subsequently decreased their intake. We recommend that the education program be continued because of its demonstrated benefit to those with low intakes. However, it should not offered to those whose intake is adequate to begin with." Or, in Fleiss's words, "Intervention A failed to effect a strong or significant change on the average value of X from baseline to some specified time after the intervention was applied, but a statistically significant correlation was found between the baseline value of X and the change from baseline. Thus, while the effectiveness of A cannot be claimed for all individuals, it can be claimed for those who were the worst off at the start."

Another popular variant of the regression fallacy occurs when subjects are enrolled into a study on the basis of an extreme value of some measurement and a treatment is declared effective because subsequent measurements are not as extreme. Similarly, it is falacious to take individuals with extreme values from one measuring instrument (a food frequency, say), reevaluate them using a different instrument (a diet record), and declare the instruments to be biased relative to each other because the second instrument's measurements are not as extreme as the first's. The regression effect guarantees that such results must be observed in the absence of any treatment effect or bias between the instruments. To quote Fleiss (p.194), "Studies that seek to establish the effectiveness of a new therapy or intervention by studying one group only, and by analyzing change either in the group as a whole or in a subgroup that was initially extreme, are inherently flawed."

While the regression effect is real and complicates the study of subjects who are initially extreme on the outcome variable, it does not make such studies impossible. Randomization and controls are enough to compensate for it. Consider a study of subjects selected for their initially low single measurement on some measure (such as vitamin A status) who are enrolled in a controlled diet trial to raise it. Regression to the mean says even the controls will show an increase over the course of the study, but if the treatment is effective the increase will be greater in the treated group than in the controls.

Honors question: Suppose a treatment is expected to lower the post-test measurements of those with high pre-test measurements and raise the post-test measurements of those with low pre-test measurements. For example, a broad-based health care program might be expected raise mean birthweight in villages where birthweight was too low and lower mean birthweight in villages where birthweight was too high. How would this be distinguished from regression to the mean?

Answer: If the program were effective, the follow-up SD would be smaller than the initial SD. When a treatment is ineffective, the marginal distributions of the two measurements are identical. If the health care program were making birthweights more homogeneous, the follow-up SD would be smaller than the initial SD.

Because the measurements are paired (made on the same subjects), tests for equal population SDs based on independent samples cannot be used. Here, a test of the equality of the initial and follow-up SDs is equivalent to testing for a correlation of 0 between the sums and differences of the measurement pairs.

Last modified: undefined.

Comparing Two Measurement Devices
Part I

It's rare when a paper says everything that needs to be said in a way that can be readily understood by a nontechnical audience, but this is one of those cases. The paper is "Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement," by JM Bland and DG Altman (The Lancet, February 8, 1986, 307-310). Perhaps it is so approachable because it was written for medical researchers three years after an equally readable version appeared in the applied statistics literature (Altman and Bland, 1983) and about the same time as a heated exchange over another approach to the problem (Kelly, 1985; Altman and Bland, 1987; Kelly, 1987).

This could very well have been a two-sentence note: "Here's the Bland and Altman reference. Please, read it." Still, its message is so elegant by virtue of its simplicity that it's worth the time and space to review the approach and see why it works while other approaches do little more than confuse the issues.

The Problem

Suppose there are two measurement techniques*, both of which have a certain amount of measurement error**, and we widh to know whether they are comparable. (Altman and Bland use the phrasing, "Do the two methods of measurement agree sufficiently closely?") Data are obtained by collecting samples and splitting them in half. One piece is analyzed by each method.

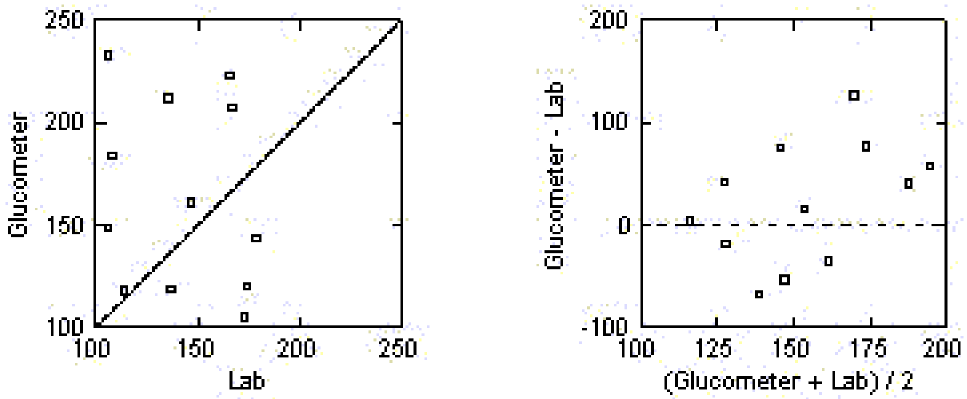
The meaning of "comparable" will vary according to the particular application. For the clinician, it might mean that diagnoses and prescriptions would not change according to the particular technique that generated a particular value. For the researcher, "comparable" might mean being indifferent to (and not even caring to know) the technique used to make a particular measurement--in the extreme case, even if the choice was made purposefully, such as having all of the pre-intervention measurements made using one technique and the post-intervention measurements made with the other. (This would *always* make me nervous, regardless of what had been learned about the comparability of the methods!)

The Bland-Altman approach is so simple because, unlike other methods, it never loses focus of the basic question of whether the two methods of measurement agree sufficiently closely. The quantities that best answer this question are the differences in each split-sample, Bland and Altman focus on the differences exclusively. Other approaches, involving correlation and regression, can never be completely successful because they summarize the data through things other than the differences.

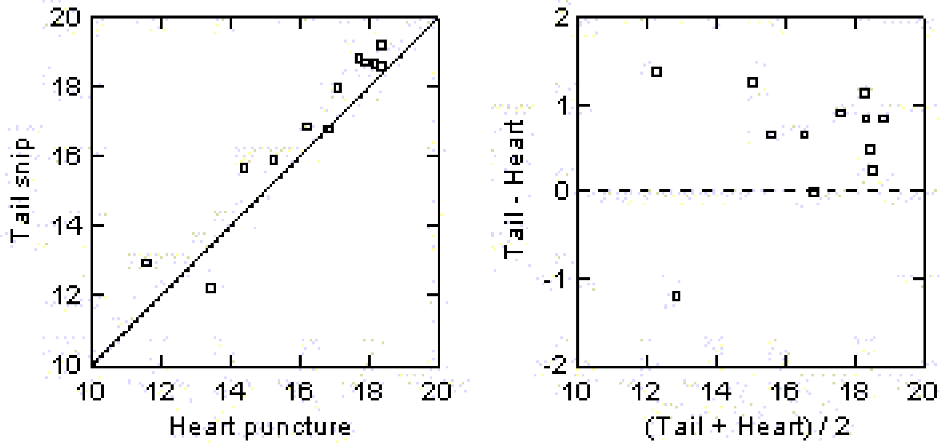
The Bland-Altman papers begin by discussing inappropriate methods and then shows how the comparison can be made properly. This note takes the opposite approach. It first shows the proper analysis and then discuss how other methods fall short. In fact, this note has already presented the Bland-Altman approach in the previous paragraph--do whatever you can to understand the observed differences between the paired measurements:

1. Plot the two sets of measurements along with the line Y=X. If the measurements are comparable, they will be tightly scattered about the line.
2. Because the eye is better at judging departures from a horizontal line than from a tilted line, plot the difference between a pair of measurements against their mean. If the measurements are comparable, the differences should be small, centered around 0, and show no systematic variation with the mean of the measurement pairs. Those who like to supplement plots with formal analysis might a construct confidence interval for the mean difference and test the statistical significance of the correlation coefficient between the sums and differences.
3. Assuming no warning signs are raised by the plot in part (2), (that is, if the observations are centered around 0 and there is no systematic variation of the difference with the mean) the data are best summarized by the standard deviation of the differences. If this number is sufficiently small from a practical (clinical) standpoint, the measurements are comparable.

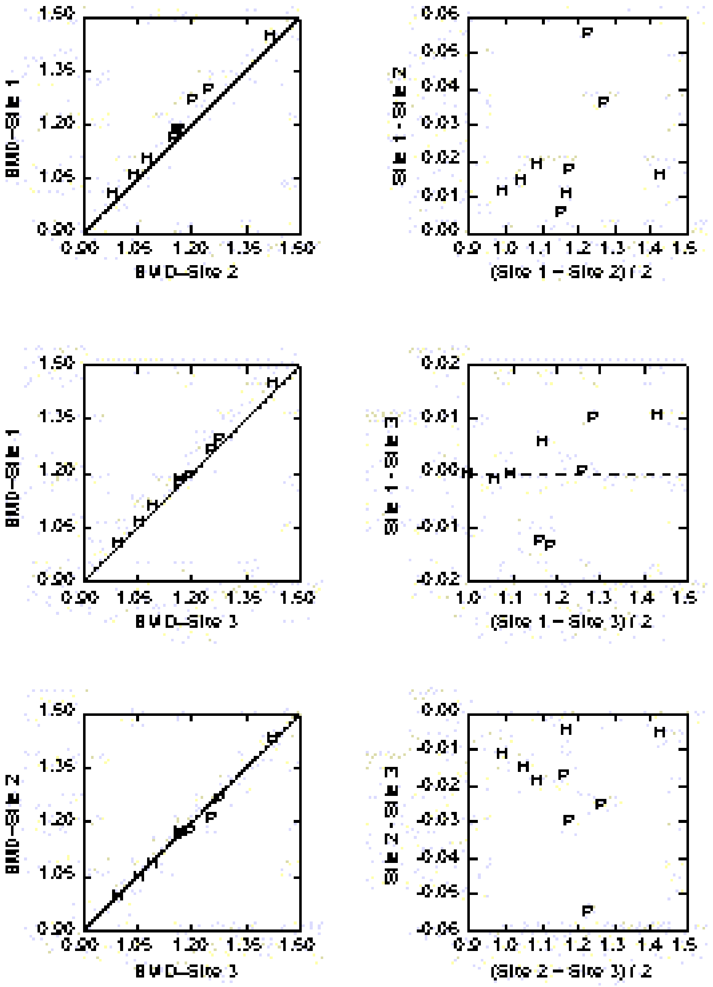
Examples



These data represent an attempt to determine whether glucose levels of mice determined by a simple device such as a Glucometer could be used in place of standard lab techniques. The plots of Glucometer value against lab values and their difference against their mean shows that there is essentially no agreement between the two measurements. Any formal statistical analyses would be icing for a nonexistent cake!



These data represent an attempt to determine whether vitamin C levels obtained from micro-samples of blood from tail snips could be used in place of the standard technique (heart puncture, which sacrifices the animal). The plots clearly demonstrate that the tail snips tend to give values that are 0.60 units higher than the standard technique. With a standard deviation of the differences of 0.69 units, perhaps the tail snip could be of practical use provided a small downward adjustment was applied to the measurements.



These data come from a study of the comparability of three devices for measuring bone density. The observations labelled 'H' are human subjects; those labelled 'P' are measurements made on phantoms. Since there are three devices, there are three pairs of plots: 1/2, 1/3, 2/3. Here we see why the plot of one measurement against another may be inadequate. All three plots look satisfactory. However, when we plot the differences against the mean values, we see that the measurements from site 2 are consistently less than the measurements from the other two sites, which are comparable.

Comment

It may take large samples to determine that there is no statistically significant difference of practical importance, but it often takes only a small sample to show that the two techniques are dramatically different. When it comes to comparability, the standard deviation of the differences is as important as their mean. Even a small sample can demonstrate a large standard deviation.

Other Approaches and Why They Are Deficient

1. *Paired t tests* test only whether the mean responses are the same. Certainly, we want the means to be the same, but this is only a small part of the story. The means can be equal while the (random) differences between measurements can be huge.

2. The *correlation coefficient* measures linear agreement--whether the measurements go up-and-down together. Certainly, we want the measures to go up-and-down together, but the correlation coefficient itself is deficient in at least three ways as a measure of agreement.

    i. The correlation coefficient can be close to 1 (or equal to 1!) even when there is considerable bias between the two methods. For example, if one method gives measurements that are always 10 units higher than the other method, the correlation will be 1 exactly, but the measurements will always be 10 units apart.

ii. The magnitude of the correlation coefficient is affected by the range of subjects/units studied. The correlation coefficient can be made smaller by measuring samples that are similar to each other and larger by measuring samples that are very different from each other. The magnitude of the correlation says nothing about the magnitude of the differences between the paired measurements which, when you get right down to it, is all that really matters.

iii. The usual significance test involving a correlation coefficient-- whether the population value is 0--is irrelevant to the comparability problem. What is important is not merely that the correlation coefficient be different from 0. Rather, it should be close to (ideally, equal to) 1!

3. The *intra-class correlation coefficient* has a name guaranteed to cause the eyes to glaze over and shut the mouth of anyone who isn't an analyst. The ICC, which takes on values between 0 and 1, is based on analysis of variance techniques. It is close to 1 when the differences between paired measurements is very small compared to the differences between subjects. Of these three procedures--t test, correlation coefficient, intra-class correlation coefficient--the ICC is best because it can be large only if there is no bias *and* the paired measurements are in good agreement, but it suffers from the same faults ii and iii as ordinary correlation coefficients. The magnitude of the ICC can be manipulated by the choice of samples to split and says nothing about the magnitude of the paired differences.

4. *Regression analysis* is typically misused by regressing one measurement on the other and declare them equivalent if and only if the confidence interval for the regression coefficient includes 1. Some simple mathematics shows that if the measurements are comparable, the population value of the regression coefficient will be equal to the correlation coefficient between the two methods. The population correlation coefficient may be close to 1, but is never 1 in practice. Thus, the only things that can be indicated by the presence of 1 in the confidence interval for the regression coefficient is (1) that the measurements are comparable but there weren't enough observations to distinguish between 1 and the population regression coefficient, or (2) the population regression coefficient *is* 1 and therefore, the measurements aren't comparable.

5. There is a line whose slope will be 1 if the measurements are comparable. It is known as a *structural equation* and is the method advanced by Kelly (1985). Altman and Bland (1987) criticize it for a reason that should come as no surprise: Knowing the data are consistent with a structural equation with a slope of 1 says something about the absence of bias but *nothing* about the variability about Y = X (the difference between the measurements), which, as has already been stated, is all that really matters.

The Calibration Problem

Calibration and comparability differ in one important respect. In the comparability problem, both methods have about the same amount of error (reproducibility). Neither method is inherently more accurate than the other. In the calibration problem, an inexpensive, convenient, less precise measurement technique (labelled C, for "crude") is compared to an expensive, inconvenient, highly precise technique (labelled P, for "precise"). Considerations of cost and convenience make the crude technique attractive despite the decrease in precision.

The goal of the calibration problem is use the value from the crude method to estimate the value that would have been obtained from the precise method. This sounds like a problem regression in regression, which it is but with a twist!

With ordinary regression, an outcome variable (labelled Y) is regressed on an input (labelled X) to get an equation of the form Y = a + b X. However, the regression model says the response for fixed X varies about the regression line with a small amount of random error. In the calibration problem, the error is attached to the predictor C, while there is no error attached to P. For this reason, many authors recommend the use of inverse regression, in which the crude technique is regressed on the precise technique (in keeping with the standard regression model: response is a linear function of the predictor, plus error) and the equation is inverted in order to make predictions. That is, the equation $C = b_0 + b_1 P$ is obtained by least squares regression and inverted to obtain

$$P = (C - b_0) / b_1$$

for prediction purposes. For further discussion, see Neter, Wasserman, and Kutner (1989, sec 5.6).

The calibration literature can become quite confusing (see Chow and Shao, 1990, for example) because the approach using inverse regression is called the "classical method" while the method of regressing C on P directly is called the "inverse method"!

----

[*] *Device* would be a better work than *technique*. I've seen the Bland-Altman method used in situations where one or both of the "techniques" were prediction equations. This might be appropriate according to the error structure of the data, but it is unlikely that such an error structure can be justified.

[**] Even *gold standards* have measurement error. The Bland-Altman technique assumes the measurement errors of the two devices are comparable. This will be discussed further in Part II.

References

- Altman DG and Bland JM (1983), "Measurement in Medicine: the Analysis of Method Comparison Studies, " The Statistician, 32, 307-317.
- Altman DG and Bland JM (1987), Letter to the Editor. Applied Statistics, 36, 224-225.
- Chow SC and Shao J (1990), "On the Difference Between the Classical and Inverse Methods of Calibration," Applied Statistics, 39, 219-228.
- Kelly GE (1985), "Use of the Structural Equation Model in Assessing the Reliability of a New Measurement Technique," Applied Statistics, 34, 258-263.
- Kelly GE (1987), Letter to the editor. Applied Statistics, 36, 225- 227.
- Neter J, Wasserman W, and Kutner M (1989), Applied Linear Regression Models. Boston, MA: Richard D. Irwin.

Last modified: undefined.

Comparing Two Measurement Devices
Part II

In 1995, Bland and Altman published "Comparing Methods of Measurement: Why Plotting Differences Against Standard Method Is Misleading" (Lancet 1995: 356:1085-1087) as a followup to their original paper on comparing two measuring devices.

When two methods of measurement are compared, it is sometimes common to see the differences between the measurements plotted against the measure that is considered to be the *standard*. This is often the result of a mistaken notion that *standard* is the same thing as *truth*. However, if the standard is subject to measurement error as most standards are, the differences will be correlated with the standard, no matter how *golden* the standard might be.

The paper is correct. However, the mathematical demonstration is presented in a way that masks much of what's going on. This note presents the same material in a different way.

Let each individual be characterized by a true underlying value U_i. Let the Us be distributed with mean θ and variance σ^2_U, that is

$$U \sim D(\theta, \sigma^2_U)$$

Suppose S and T are both unbiased estimates of U, that is,

$$S = U + \delta\text{, with } \delta \sim D(0, \sigma^2 \delta)$$
$$T = U + \varepsilon\text{, with } \varepsilon \sim D(0, \sigma^2 \varepsilon)$$

This says S and T are both unbiased methods of measuring U with their own respective measurment errors, $\sigma^2 \delta$ and $\sigma^2 \varepsilon$. Further, assume that all of the errors are uncorrelated, that is,

$$\text{cov}(U_i,U_j) = \text{cov}(\delta_i, \delta_j) = \text{cov}(\varepsilon_i, \varepsilon_j) = 0 \text{, for all } i \neq j\text{, and}$$
$$\text{cov}(U_i, \delta_j) = \text{cov}(\delta_i, \varepsilon_j) = \text{cov}(\delta_i, \varepsilon_j) = 0 \text{, for all } i,j.$$

Then,

$$\sigma_S^2 = \sigma_U^2 + \sigma_\delta^2, \quad \sigma_T^2 = \sigma_U^2 + \sigma_\varepsilon^2, \text{ and } \rho_{ST} = \frac{\sigma_U^2}{\sigma_S \sigma_T}.$$

and it follows that

$$corr(T-S,\frac{T+S}{2}) = \frac{\sigma_\varepsilon^2 - \sigma_\delta^2}{\sqrt{(\sigma_\varepsilon^2+\sigma_\delta^2)(4\sigma_U^2+\sigma_\varepsilon^2+\sigma_\delta^2)}},$$

$$corr(T-S,S) = \frac{-\sigma_\delta^2}{\sqrt{(\sigma_U^2+\sigma_\delta^2)(\sigma_\varepsilon^2+\sigma_\delta^2)}}, \text{ and}$$

$$corr(T-S,T) = \frac{\sigma_\varepsilon^2}{\sqrt{(\sigma_U^2+\sigma_\varepsilon^2)(\sigma_\varepsilon^2+\sigma_\delta^2)}},$$

This demonstrates that when the two measuring techniques have equal measurement error, their difference is uncorrelated with their mean. If one of the measurment techniques has no measurement error, the differences will be uncorrelated with it. Indeed, when one of the measurment techniques has no measurement error, the differences will be correlated with the means

Terminology: Regression, ANOVA, ANCOVA

In every field, it's essential to use the proper terminology in order to be understood and to inspire confidence in your audience. In statistics, some analyses can be described correctly in different ways. This can be viewed as liberating or as evidence of a sinister plot according to one's general outlook on life.

An example of this overlap in nomenclature occurs with **regression**, **analysis of variance (ANOVA)**, and **analysis of covariance (ANCOVA)**. These methods are used to model a quantitative response variable (such as cholesterol level) in terms of predictor variables. The name of the method depends on whether the predictors are quantitative, qualitative (factors composed of levels, such as Oil [rice/canola/peanut]), or both.

One simple rule states that if all predictors are quantitative, the analysis is called **regression**; if all predictors are qualitative, the analysis is **analysis of variance**; if there are both qualitative and quantitative predictors, the analysis is **analysis of covariance**.

This rule is correct most of the time, but sometimes additional criteria are applied when both qualitative and quantitative predictors are used. If the focus is on the quantitative predictors, the analysis is often called regression. If the focus is on the qualitative predictors, the analysis is almost always called ANCOVA and the quantitative predictors are called **covariates**.

Some say the name ANCOVA should be used only when the model does not include interactions between the covariates and the factor of interest. Thus, a strict ANCOVA model is a "parallel slopes" model, and the regression coefficients for the covariates are the same for all factor levels. When an author says that an ANCOVA model was fitted, assume no allowance was made for an interaction between the factor and covariates unless there is a statement to the contrary.

The name of the analysis is not always the the name of the computer program that performs it. Any ANOVA or ANCOVA can be performed by using an ordinary regression package if one is clever about constructing the proper sets of indicator variables. One can and should report that an ANOVA or ANCOVA was performed even when a regression program is used to do it.

[back to LHSP]

Introduction to Regression Models

Gerard E. Dallal, Ph.D.

[Notation: Upper case Roman letters represent random variables. Lower case Roman letters represent realizations of random variables. For example, if X is WEIGHT, then x is 159 lbs. E(Y) is the population mean value of the random variable Y. E(Y|X) is the population mean value of Y when X is known. E(Y|X=x) is the population mean value of Y when X=x.]

The least squares regression equation

$$\hat{y} = b_0 + b_1 x$$

is an estimate of the population regression equation

$$E(Y|X=x) = \beta_0 + \beta_1 x$$

The response variable, Y, is described by the model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where ε_i is a random error. The usual tests produced by most statisical program packages assume the errors

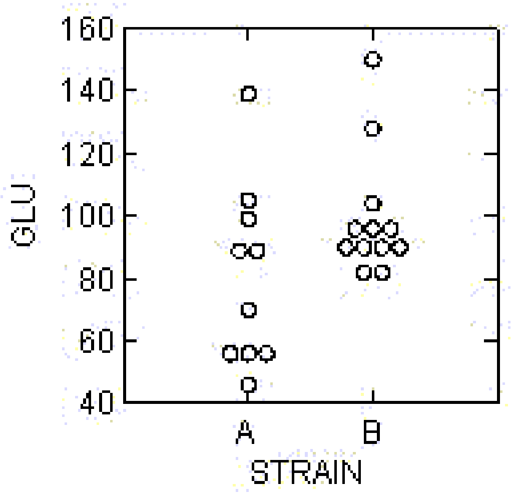
- are independent and
- follow a normal distribution with mean 0 and
- constant variance. This means that the variability of responses for small X values is the same as the variability of responses for large X values.

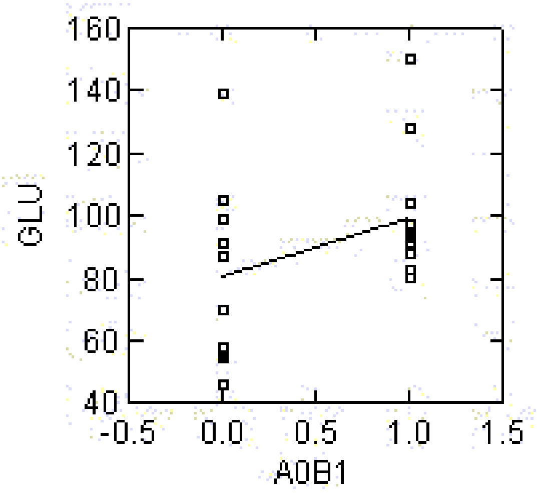
This is usually written $\varepsilon \sim N(0, \sigma^2)$--that is, normally distributed with mean 0 and variance σ^2--where σ is a fixed but unknown constant. (The standard error of the estimate estimates σ.)

Last modified: undefined.

Student's t Test for Independent Samples Is A Special Case of Simple Linear Regression

Student's t test for independent samples is equivalent to a linear regression of the response variable on the grouping variable, where the grouping variable is recoded to have numerical values, if necessary.





Here's an example involving glucose levels in two strains of rats, A and B. First, the data are displayed in a dot plot. Then, Glucose is plotted against A0B1, where **A0B1** is created by setting it equal 0 for strain A and 1 for strain B.

Student's t test for independent samples yields

```
Variable: GLU

STRAIN        N          Mean       Std Dev     Std Error
-----------------------------------------------------
A            10   80.40000000   29.20502240   9.23543899
B            12   99.66666667   19.95601223   5.76080452

Variances        T        DF    Prob>|T|
---------------------------------------
Unequal     -1.7700     15.5      0.0965
Equal       -1.8327     20.0      0.0818
```

The linear regression of glucose on A0B1 gives the equation GLU = $b_0 + b_{A0B1}$ A0B1 .

```
Dependent Variable: GLU
                        Parameter Estimates
```

```
                 Parameter      Standard    T for H0:
Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

INTERCEPT  1     80.400000    7.76436303        10.355        0.0001
A0B1       1     19.266667   10.51299725         1.833        0.0818
```

The P value for the Equal Variances version of the t test is equal to the P value for the regression coefficient of the grouping variable A0B1 (P = 0.0818). The corresponding t statistics are equal in magnitude (|t| = 1.833). This is not a coincidence. Statistical theory says the two P values **must** be equal. The t statistics **must** be equal in magnitude. The signs will be the same if the t statistic is calculated by subtracting the mean of group 0 from the mean of group 1.

The equal variances version of Student's t test is used to test the hypothesis of the equality of $\mu_{\mathbf{A}}$ and $\mu_{\mathbf{B}}$, the means of two normally distributed populations with equal population variances. The hypothesis can be written H_0: $\mu_A = \mu_B$. The population means can be reexpressed as $\mu_A=\mu$ and $\mu_B=\mu + \delta$, where $\delta = \mu_{\mathbf{B}} - \mu_{\mathbf{A}}$ (that is, data from strain A are normally distributed with mean μ and standard deviation σ while data from strain B are normally distributed with mean $\mu + \delta$ and standard deviation σ) and the hypothesis can be rewriten as $\mathbf{H_0}$**:** δ **= 0**.

The linear regression model says data are normally distributed about the regression line with constant standard deviation σ. The predictor variable A0B1(the grouping variable) takes on only two values. Therefore, there are only two locations along the regression line where there are data (see the display). "Homoscedastic (constant spread about the regression line) normally distributed values about the regression line" is equivalent to "two normally distributed populations with equal variances". μ_A is equal to β_0, μ_B is equal to $\beta_0 + \beta_{A0B1}$, and β_{A0B1} is equal to δ. Thus, the hypothesis of equal means (H_0: $\delta = 0$) is equivalent to the hypothesis that the regression coefficient of A0B1 is 0 (H_0: $\beta_{A0B1} = 0$). The population means are equal if and only if the regression line is horizontal. Since the probability structure is the same for the two problems (homoscedastic, normally distributed data), test statistics and P values will be the same, too.

The numbers confirm this. For strain A, the predicted value $b_0+b_{A0B1}*0$, is 80.400000 + 19.266667*0 = 80.40, the mean of strain A. For strain B, $b_0+b_{A0B1}*1$ is 80.400000 + 19.266667*1 = 99.67, the mean of strain B. Had the numerical codes for strain been different from 0 and 1, the intercept and regression coefficient would change so that the two predicted values would continue to be the sample means. The t statistic and P value for the regression coefficient would not change. The best fitting line passes through the two points whose X-values are equal to the coded Strain values and whose Y-values are equal to the corresponding sample means. This minimizes the sum of squared differences between observed and predicted Y-values. Since this involves only two points and two points determine a straight line, the linear regression equation will always have the slope & intercept necessary to make the line pass through the two points. To put it another way, two points define the regression line. The Y-values are the sample means. The X-values are determined by the coding scheme. Whatever the X-values, the slope & intercept of the regression line will be those of the line that passes through the two points.

Last modified: undefined.

Introduction to Multiple Linear Regression

Gerard E. Dallal, Ph.D.

If you are familiar with simple linear regression, then you know the very basics of multiple linear regression. Once again, the goal is to obtain the least squares equation (that is, the equation for which the sum of squared residuals is a minimum) to predict some response. With simple linear regression there was one predictor. The fitted equation was of the form

$$\hat{y} = b_0 + b_1 x \,.$$

With multiple linear regression, there are multiple predictors. The fitted equation is of the form

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_p x_p \,,$$

where *p* is the number of predictors.

The output from a multiple linear analysis will look familiar. Here is an example of cross-sectional data where the log of HDL cholesterol (the so-called good cholesterol) in women is predicted from their age, body mass index, blood vitamin C, systolic and diastolic blood pressures, skinfold thickness, and the log of total cholesterol.

```
                              The REG Procedure
                                Model: MODEL1
                         Dependent Variable: LHCHOL

                             Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square     F
Value    Pr > F

Model                     8        0.54377        0.06797
6.16    <.0001
Error                   147        1.62276
0.01104
Corrected Total         155
2.16652


             Root MSE              0.10507    R-Square      0.2510
             Dependent Mean        1.71090    Adj R-Sq      0.2102
             Coeff Var             6.14105
```

```
                              Parameter Estimates

                                Parameter       Standard
    Variable     DF              Estimate          Error    t Value    Pr >
|t|

    Intercept     1               1.16448        0.28804       4.04
<.0001
    AGE           1           -0.00092863        0.00125      -0.74
0.4602
    BMI           1              -0.01205        0.00295      -4.08
<.0001
    BLC           1               0.05055        0.02215       2.28
0.0239
    PRSSY         1           -0.00041910     0.00044109      -0.95
0.3436
    DIAST         1               0.00255        0.00103       2.47
0.0147
    GLUM          1           -0.00046737     0.00018697      -2.50
0.0135
    SKINF         1               0.00147        0.00183       0.81
0.4221
    LCHOL         1               0.31109        0.10936       2.84
0.0051
```

To predict someone's logged HDL cholesterol, just take the values of the predictors, multiply them by their coefficients, and add them up. Some coefficients are statistically significant; some are not. What we make of this or do about it depends on the particular research question.

Warning

It is reasonable to think that statistical methods appearing in a wide variety of text books have the imprimatur of the statistical community and are meant to be used. However, multiple regression includes many methods that were investigated for the elegance of their mathematics. Some of these methods (such as stepwise regression and principal component regression) should not be used to analyze data. We will discuss these methods in future notes.

The analyst should be mindful from the start that multiple regression techniques should never be studied in isolation from data. What we do and how we do it can only be addressed in the context of a specific research question.

How to Read the Output From Multiple Linear Regression Analyses

Here's a typical piece of output from a multiple linear regression of homocysteine (LHCY) on vitamin B12 (LB12) and folate as measured by the CLC method (LCLC). That is, vitamin B12 and CLC are being used to predict homocysteine. A (common) logarithmic transformation had been applied to all variables prior to formal analysis, hence the initial L in each variable name, but that detail is of no concern here.

```
Dependent Variable: LHCY
                              Analysis of Variance

                                 Sum of         Mean
        Source          DF      Squares       Square      F
Value       Prob>F
        Model            2      0.47066      0.23533
8.205       0.0004
        Error          233      6.68271      0.02868
        C Total        235      7.15337

            Root MSE       0.16936     R-square       0.0658
            Dep Mean       1.14711     Adj R-sq       0.0578
            C.V.          14.76360

                              Parameter Estimates

                       Parameter      Standard    T for H0:
      Variable  DF      Estimate         Error   Parameter=0    Prob
> |T|
      INTERCEP   1      1.570602    0.15467199        10.154
0.0001
      LCLC       1     -0.082103    0.03381570        -2.428
0.0159
      LB12       1     -0.136784    0.06442935        -2.123
0.0348
```

Parameter Estimates.

The column labeled **Variable** should be self-explanatory. It contains the names of the predictor variables which label each row of output.

DF stands for **degrees of freedom**. For the moment, all entries will be 1. Degrees of freedom will be discussed in detail later.

The **Parameter Estimates** are the regression coefficients. The regression equation is

```
LHCY = 1.570602 - 0.082103 LCLC - 0.136784 LB12
```

To find the predicted homocysteine level of someone with a CLC of 12.3 and B12 of 300, we begin by taking logarithms. Log(12.3)=1.0899 and log(300)=2.4771. We then calculate

```
LHCY = 1.570602 - 0.082103 1.0899 - 0.136784 2.4771
     = 1.1423
```

Homocysteine is the anti-logarithm of this value, that is, $10^{1.1423}$ = 13.88.

The **Standard Errors** are the standard errors of the regression coefficients. They can be used for hypothesis testing and constructing confidence intervals. For example, confidence intervals for LCLC are constructed as (-0.082103± k 0.03381570), where k is the appropriate constant depending on the level of confidence desired. For example, for 95% confidence intervals based on large samples, k would be 1.96.

The **T** statistic tests the hypothesis that a population regression coefficient is 0 **WHEN THE OTHER PREDICTORS ARE IN THE MODEL**. It is the ratio of the sample regression coefficient to its standard error. The statistic has the form (estimate - hypothesized value) / SE. Since the hypothesized value is 0, the statistic reduces to Estimate/SE. If, for some reason, we wished to test the hypothesis that the coefficient for LCLC was -0.100, we could calculate the statistic (-0.082103-(-0.10))/0.03381570.

Prob > |T| labels the **P values** or the **observed significance levels** for the t statistics. The degrees of freedom used to calculate the P values is given by the Error DF from the ANOVA table. The P values tell us whether a variable has statistically significant predictive capability in the presence of the other variables, that is, whether it adds something to the equation. In some circumstances, a nonsignificant P value might be used to determine whether to remove a variable from a model without significantly reducing the model's predictive capability. For example, if one variable has a nonsignificant P value, we can say that it does not have predictive capability in the presence of the others,remove it, and refit the model without it. These P values should not be used to eliminate more than one variable at a time, however. A variable that does not have predictive capability in the presence of the other predictors may have predictive capability when some of those predictors are removed from the model.

The Analysis of Variance Table

The **Analysis of Variance** table is also known as the **ANOVA table** (for ANalysis Of VAriance). There is variability in the response variable. It is the uncertainty that would be present if one had to predict individual responses without any other information. The best one could do is predict each observation to be equal to the sample mean. The amount of uncertainty or variability can be measured by the Total Sum of Squares, which is the numerator of the sample variance. The ANOVA table partitions this variability into two parts. One portion is accounted for (some say "explained by") the model. It's the reduction in uncertainty that occurs when the regression model is used to predict the responses. The remaining portion is the uncertainty that remains even after the model is used. The model is considered to be statistically significant if it can account for a large amount of variability in the response.

The column labeled **Source** has three rows, one for total variability and one for each of the two pieces that the total is divided into--**Model,** which is sometimes called **Regression**, and **Error**, sometimes called **Residual**. The **C** in **C Total** stands for **corrected**. Some programs ignore the **C** and label this **Total**. The C Total **Sum of Squares** and **Degrees of Freedom** will be the sum of Model and Error.

Sums of Squares: The total amount of variability in the response can be written $\sum$(y-ybar)², where ybar is the sample mean. (The "Corrected" in "C Total" refers to subtracting the sample mean before squaring.) If we were asked to make a prediction without any other information, the best we can do, in a certain sense, is the sample mean. The amount of variation in the data that can't be accounted for by this simple method of prediction is given by the Total Sum of Squares.

When the regression model is used for prediction, the amount of uncertainty that remains is the variability about the regression line, $\sum$(y-yhat)². This is the Error sum of squares. The difference between the Total sum of squares and the Error sum of squares is the Model Sum of Squares, which happens to be equal to $\sum$(yhat-ybar)².

Each sum of squares has corresponding degrees of freedom (DF) associated with it. Total df is one less than the number of observations, n-1. The Model df is the number of independent variables in the model, p. The Error df is the difference between the Total df (n-1) and the Model df (p), that is, n-p-1.

The **Mean Squares** are the Sums of Squares divided by the corresponding degrees of freedom.

The **F Value** or **F ratio** is the test statistic used to decide whether the model as a whole has statistically significant predictive capability, that is, whether the regression SS is big enough, considering the number of variables needed to achieve it. **F** is the ratio of the Model Mean Square to the Error Mean Square. Under the null hypothesis that the model has no predictive capability--that is, that all population regression coefficients are 0 simultaneously--the F statistic follows an F distribution with *p* numerator degrees of freedom and *n-p-1* denominator degrees of freedom. The null hypothesis is rejected if the F ratio is large. Some analysts recommend ignoring the P values for the individual regression coefficients if the overall F ratio is not statistically significant, because of the problems caused by multiple testing. I tend to agree with this recommendation with one important exception. If the purpose of the analysis is to examine a particular regression coefficient after adjusting for the effects of other variables, I would ignore everything but the regression coefficient under study. For example, if in order to see whether dietary fiber has an effect on cholesterol, a multiple regression equation is fitted to predict cholesterol levels from dietary fiber along with all other known or suspected determinants of cholesterol, I would focus on the regression coefficient for fiber regardless of the overall F ratio. (This isn't quite true. I would certainly wonder why the overall F ratio was not statistically significant if I'm using the known predictors, but I hope you get the idea. If the focus of a study is a particular regression coefficient, it gets most of the attention and everything else is secondary.)

The **Root Mean Square Error** (also known as **the standard error of the estimate**) is the square root of the Residual Mean Square. It is the standard deviation of the data about the regression line, rather than about the sample mean.

R^2 is the squared multiple correlation coefficient. It is also called the **Coefficient of Determination**. R^2 is the ratio of the Regression sum of squares to the Total sum of squares, RegSS/TotSS. It is the proportion of the variability in the response that is accounted for by the model. Since the Total SS is the sum of the Regression and Residual Sums of squares, R^2 can be rewritten as (TotSS-ResSS)/TotSS = 1-ResSS/TotSS. Some call R^2 *the proportion of the variance explained by the model*. I don't like the use of the word *explained* because it implies causality. However, the phrase is firmly entrenched in the literature. If a model has perfect predictability, $R^2=1$. If a model has no predictive capability, $R^2=0$. (In practice, R^2 is never observed to be exactly 0 the same way the difference between the means of two samples drawn from the same population is never exactly 0.) R, the multiple correlation coefficient and square root of R^2, is the correlation between the observed values (y), and the predicted values (yhat).

As additional variables are added to a regression equation, R^2 increases even when the new variables have no real predictive capability. The **adjusted-R^2** is an R^2-like measure that avoids this difficulty. When variables are added to the equation, adj-R^2 doesn't increase unless the new variables have additional predictive capability. Where R^2 is 1 - ResSS/TotSS , we have
adj R^2 = 1 - (ResSS/ResDF)/(TotSS/(n-1)), that is, it is 1 minus the ratio of (the square of the standard error of the estimate) to (the sample variance of the response). Additional variables with no explanatory capability will increase the Regression SS (and reduce the Residual SS) slightly, except in the unlikely event that the sample partial correlation is *exactly* 0. However, they won't tend to decrease the standard error of the estimate because the reduction in Residual SS will be accompanied by a decrease in Residual DF. If the additional variable has no predictive capability, these two reductions will cancel each other out.

Last modified: undefined.

What do the Coefficients in a Multiple Linear Regression Mean?

The regression coefficient for the i-th predictor is the expected change in response per unit change in the i-th predictor, all other things being equal. That is, if the i-th predictor is changed 1 unit while all of the other predictors are held constant, the response is expected to change b_i units. As always, it is important that cross-sectional data not be interpreted as though they were longitudinal.

The regression coefficient and its statistical significance can change according to the other variables in the model. Among postmenopausal women, it has been noted that bone density is related to weight. In this cross-sectional data set, density is regressed on weight, body mass index, and percent ideal body weight*. These are the regression coefficients for the 7 possible regression models predicting bone density from the weight measures.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	0.77555	0.77264	0.77542	0.77065	0.74361	0.77411	0.75635
WEIGHT	0.00642	.	0.00723	0.00682	0.00499	.	.
BMI	-0.00610	-0.04410	.	-0.00579	.	0.01175	.
PCTIDEAL	0.00026	0.01241	-0.00155	.	.	.	0.00277

Not only do the magnitudes of the coefficients change from model to model, but for some variables the sign changes, too.

For each regression coefficient, there is a t statistic. The corresponding P value tells us whether the variable has statistically significant predictive capability in the presence of the other predictors. A common mistake is to assume that when many variables have nonsignificant P values they are all unnecessary and can be removed from the regression equation. This is not necessarily true. When one variable is removed from the equation, the others may become statistically significant. Continuing the bone density example, the P values for the predictors in each model are

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
WEIGHT	0.1733	.	0.0011	<.0001	<.0001	.	.
BMI	0.8466	0.0031	.	0.1960	.	<.0001	.
PCTIDEAL	0.9779	0.0002	0.2619	.	.	.	<.0001

All three predictors are related, so it is not surprising that model (1) shows that all of them are nonsignifcant in the presence of the others. Given WEIGHT and BMI, we don't need PCTIDEAL, and so on. Any one of them is superfluous. However, as models (5), (6),and (7) demonstrate, All of them are highly statistically significant when used alone.

The P value from the ANOVA table tells us whether there is predictive capability in the model as a whole. All four combinations in the following table are possible.

		Overall F	
		Significant	**NS**
Individual t	**Significant**		
	NS		

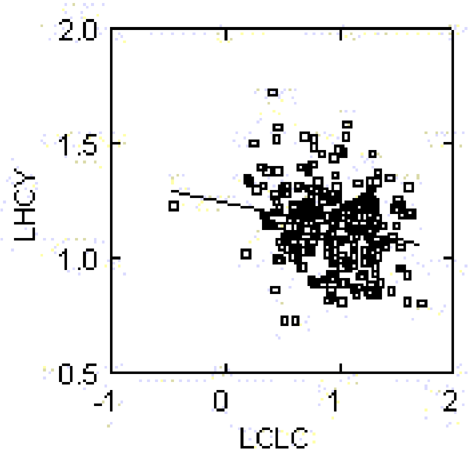
- Cases where the t statistic for every predictor and the F statistic for the overall model are statistically significant are those where every predictor has something to contribute.
- Cases where nothing reaches statistical significance are those where none of the predictors are of any value.
- This note has shown that it is possible to have the overall F ratio statistically significant and all of the t statistics nonsignificant.
- It is also possible to have the overall F ratio nonsignificant and some of the t statistics significant. There are two ways this can happen.
    - First, there may be no predictive capability in the model. However, if there are many predictors, statistical theory guarantees that on average 5% of them will appear to have statistically significant predictive capability when tested individually.
    - Second, the investigator may have chosen the predictors poorly. If one useful predictor is added to many that are unrelated to the outcome, its contribution may not be large enough for the overall model to appear to have statistically significant predictive capability. A contribution that might have reached statistical significant when viewed individually, might not make it out of the noise when viewed as part of the whole.

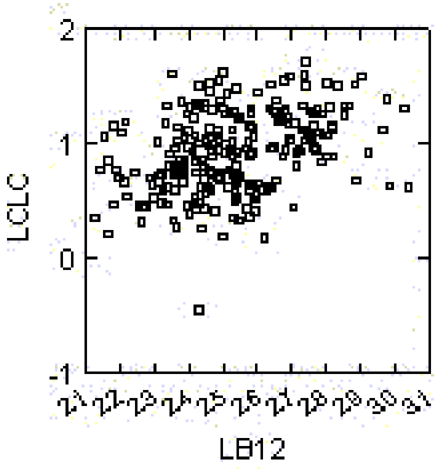
----

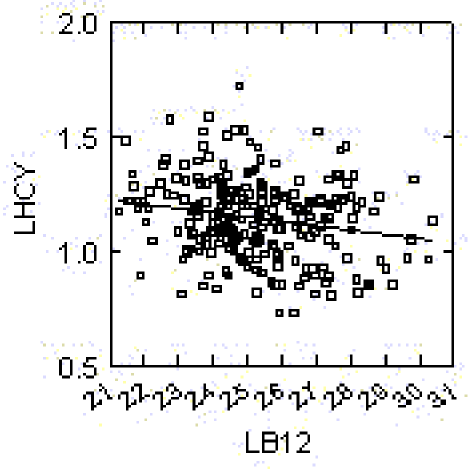
[*] In general, great care must be used when using a predictor such as body mass index or percent ideal body weight that is a ratio of other variables. This will be discussed in detail later.

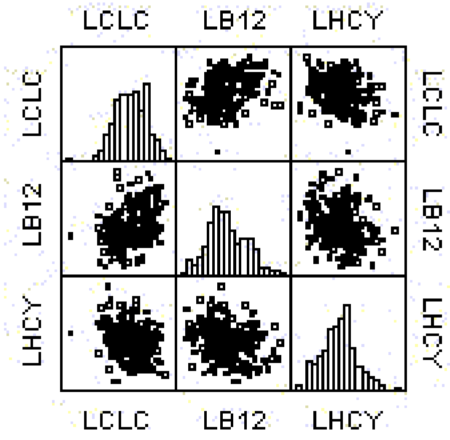
Last modified: undefined.

What Does Multiple Linear Regression Look Like?





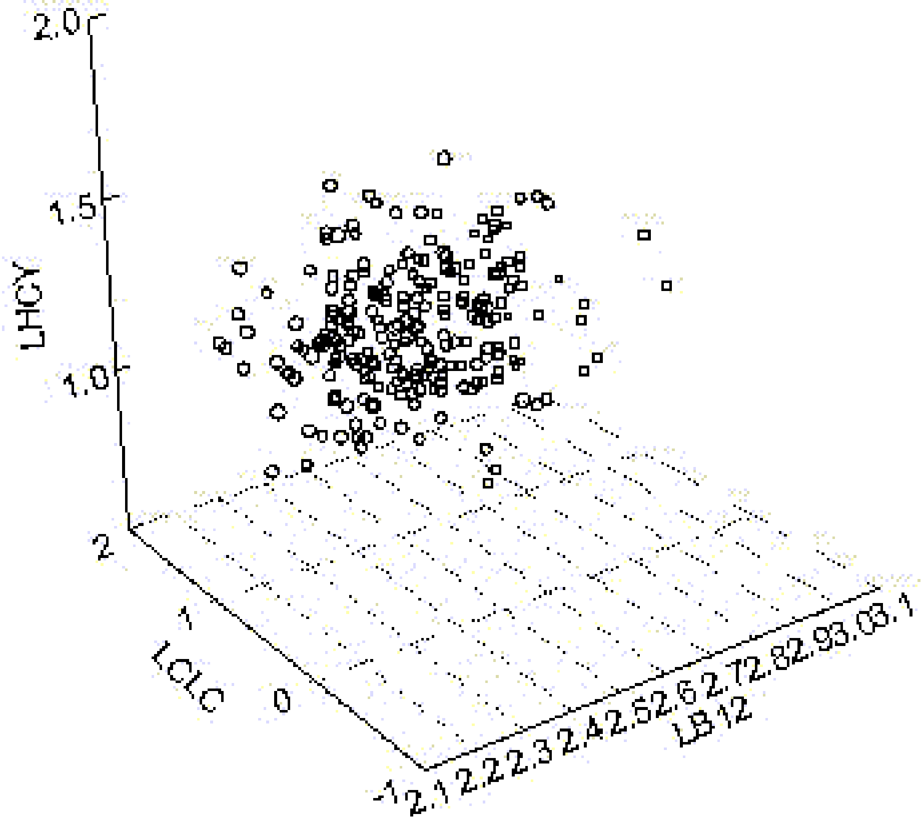




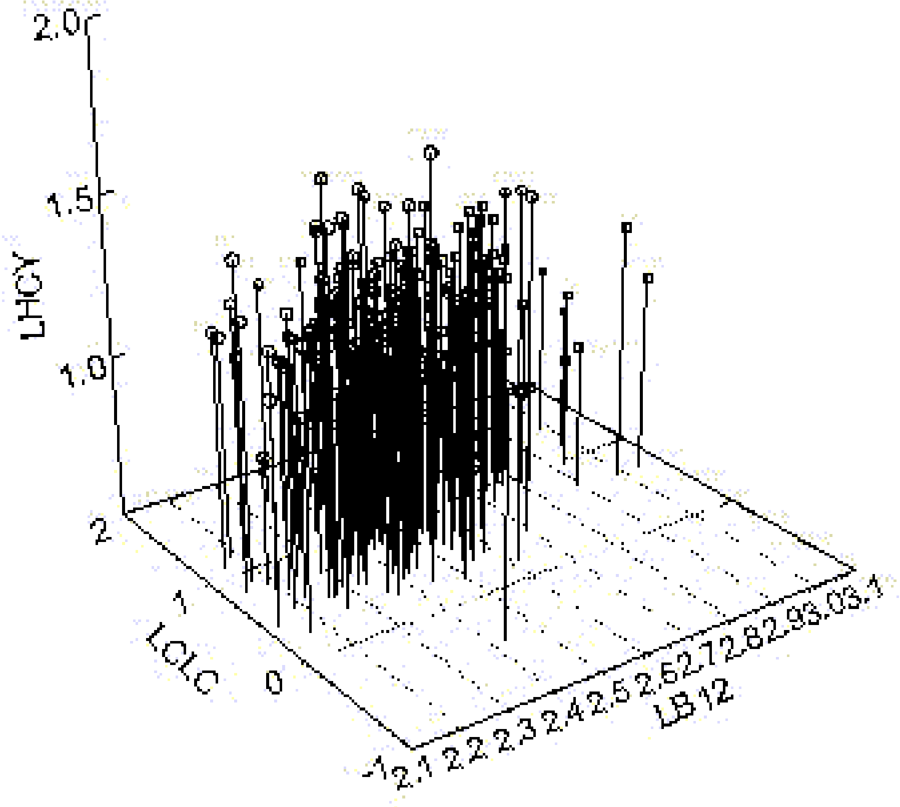
Consider once again the regression of homocysteine on B12 and folate (all logged). It's common to think of the data in terms of pairwise scatterplots. The regression equation

LHCY = 1.570602 - 0.082103 LCLC - 0.136784 LB12

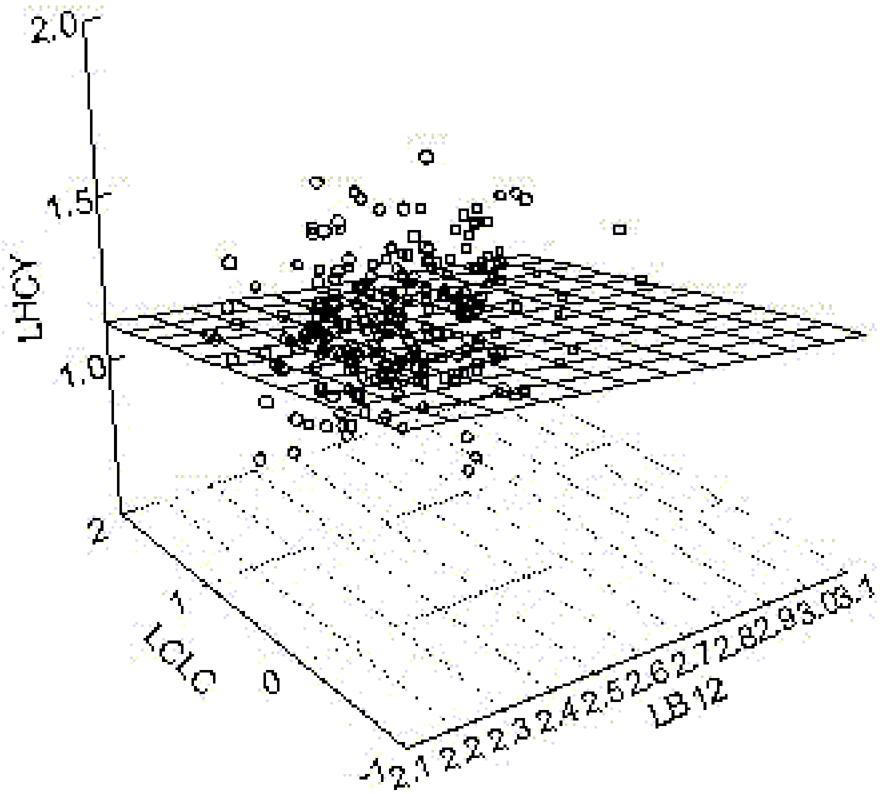
is often mistakenly thought of as a kind of line. However, it is not a line, but a surface.



Each observation is a three-dimensional vector {(x_i, y_i, z_i), i = 1,..n} [here, ($LCLC_i$, $LB12_i$, $LHCY_i$)]. When plotted in a three-dimensional space, the data look like the picture to the left.



It can be difficult to appreciate a two-dimensional representation of three- dimensional data. The picture is redrawn with spikes from each observation to the plane defined by LCLC and LB12 to give a better sense of where the data lie.



The final display shows the regression surface. It is a flat plane. Predicted values are obtained by staring at the intersection of LB12 and LCLC on the LB12-LCLC plane and travelling parallel to the LHCY axis until the plane is reached (in the manner of the spike, but to the plane instead of the observation). Residuals are calculated as the distance from the observation to the plane, again travelling parallel to the LCHY axis.

The same thing happens with more that 2 predictors, but it's hard to draw a two-dimensional representation of it. With *p* predictors, the regression surface is a *p*-dimensional hyperplane in a (*p*+*1*)-dimensional space.

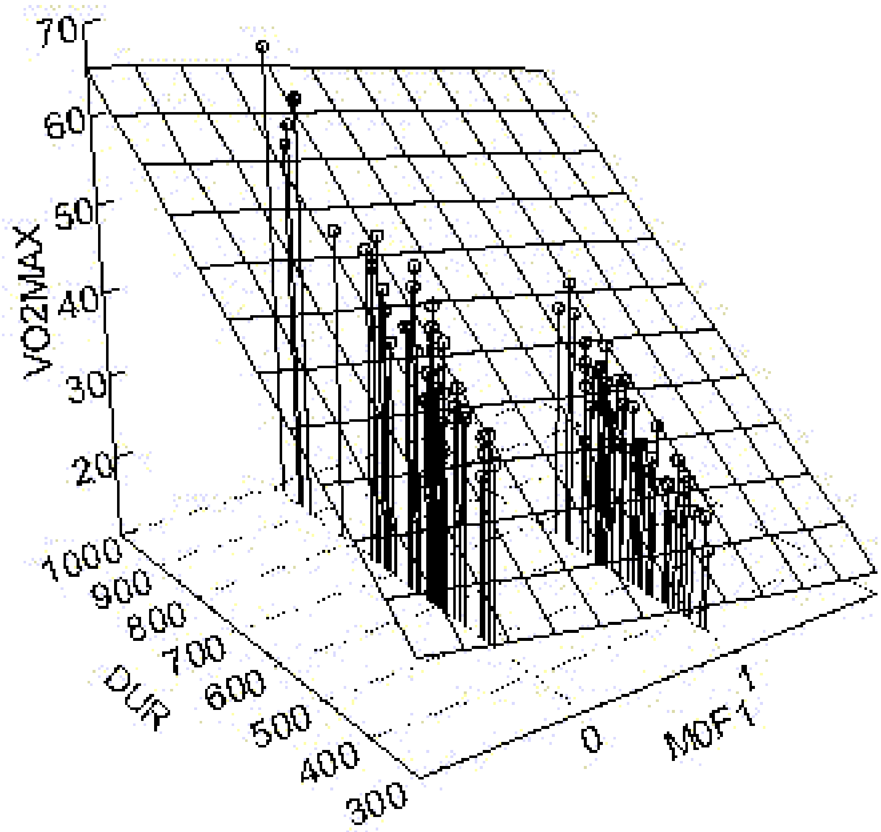
What Does Multiple Linear Regression Look Like? (Part 2)

This note considers the case where one of the predictors is an indicator variable. It will be coded 0/1 here, but these results do not depend on the the two codes used. Here, men and women are placed on a treadmill. When they can no longer continue, duration (DUR) an maximum oxygen usage (VO2MAX) are recorded. The purpose of this analysis is to predict VO2MAX from sex (M0F1 = 0 for males, 1 for females) and DUR. When the model

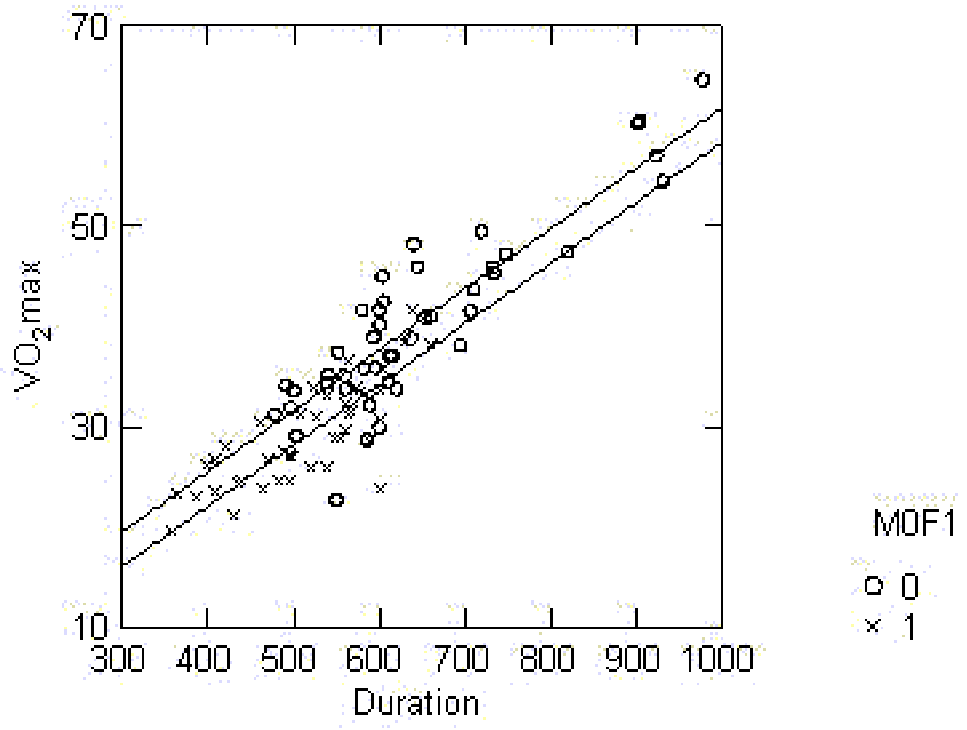
$$\text{VO2MAX} = \beta_0 + \beta_1\,\text{DUR} + \beta_2\,\text{M0F1} + \varepsilon$$

is fitted to the data, the result is

$$\text{VO2MAX} = 1.3138 + 0.0606\ \text{DUR} - 3.4623\ \text{M0F1}$$



When the data are plotted in three dimensions, it is seen that they lie along two slices--one slice for each of the two values of M0F1. The regression surface is once again a flat plane. This follows from our choice of a model.

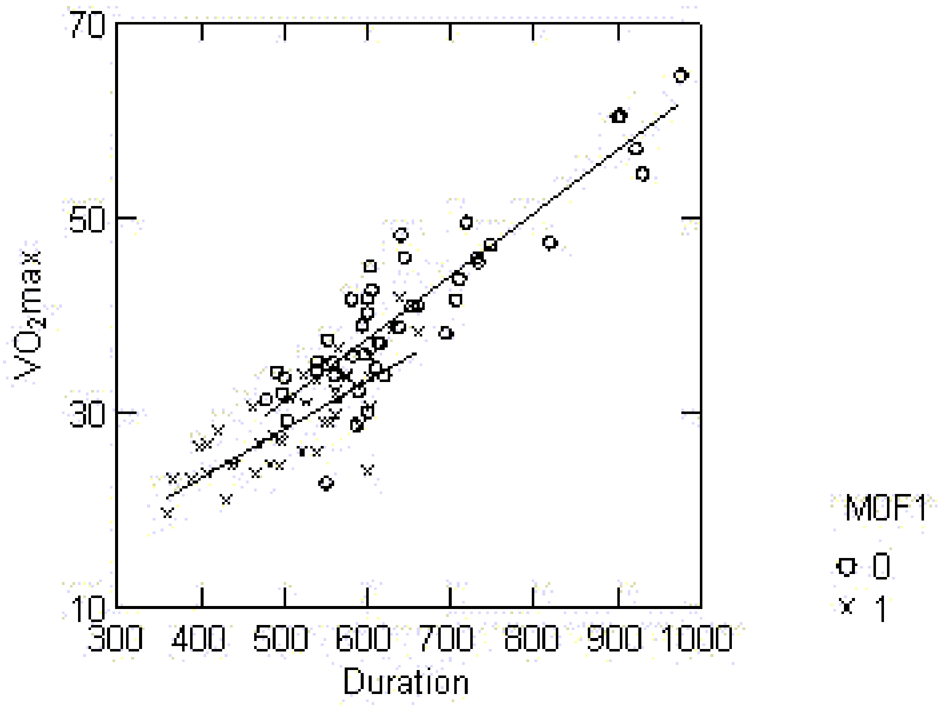


The data in each slice can be plotted as VO2MAX against DUR and the two plots can be superimposed. The two lines are the pieces of the plane corresponding to M0F1=0 and M0F1=1. The lines are parallel because they are parallel strips from the same flat plane. This also follow directly from the model. The fitted equation can be written conditional on the two values of M0F1. When M0F1=0, the model is

YO2MAX = 1.3138 + 0.0606 DUR - 3.4623 * 0, or YO2MAX = 1.3138 + 0.0606 DUR

When M0F1=1, the model is

YO2MAX = 1.3138 + 0.0606 DUR - 3.4623 * 1, or
YO2MAX = -2.1485 + 0.0606 DUR.



A more complicated model can be fitted that does not force the lines to be parallel. This is discussed in the note on interactions. The seaparate lines are fitted in the picture to the left. The test for whether the lines are parallel has an observed significance level of 0.102. Thus, the regression coefficients are within sampling variability of each other and the lines are within sampling variability of what one would expect of parallel lines.

Last modified: undefined.

Why Is a Regression Line Straight?

This could have been part of the "What does multiple linear regression look like?" note. However, I didn't want it to be seen as a footnote to the pretty pictures. This is the more important lesson.

A simple linear regression line is straight because **we fit a straight line to the data**! We could fit something other than a straight line if we want to. For example, instead of fitting

$$\text{BONE DENSITY} = b_0 + b_1 \text{ AGE}$$

we might fit the equation

$$\text{BONE DENSITY} = b_0 + b_1 \text{ AGE} + b_2 \text{ AGE}^2$$

if we felt the relation was quadratic. This is one reason for looking at the data as part of the analysis.

When homocysteine was regressed on CLC-folate and vitamin B12, why was the regression surface flat? The answer here, too, is because we fit a flat surface!

Let's take a closer look at the regession equation

$$\text{LHCY} = 1.570602 - 0.082103 \text{ LCLC} - 0.136784 \text{ LB12}$$

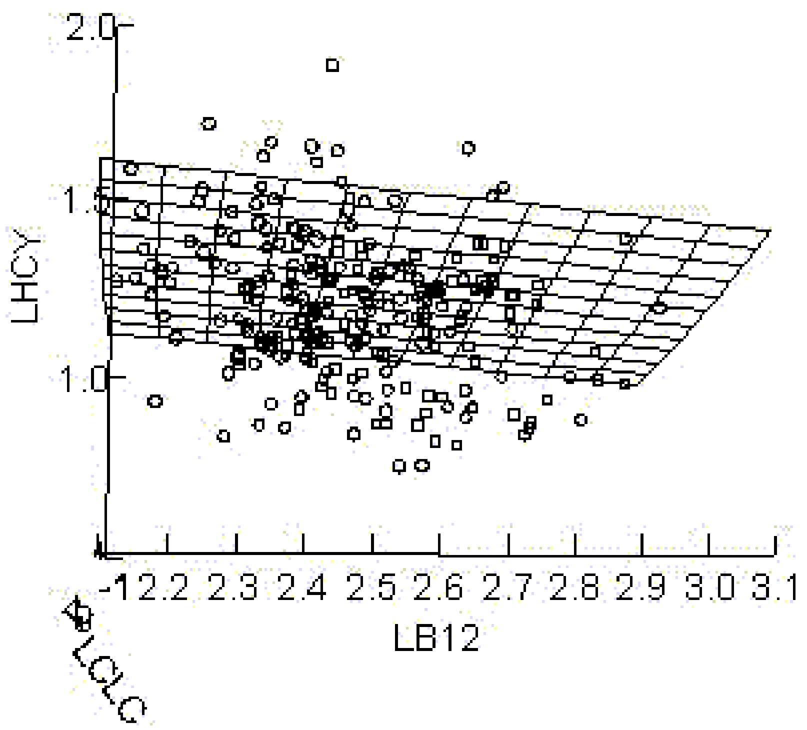
Suppose LCLC is 1.0. Then

$$\text{LHCY} = 1.570602 - 0.082103 * 1 - 0.136784 \text{ LB12}$$

or

$$\text{LHCY} = 1.488499 - 0.136784 \text{ LB12}$$

There is a straight line relation between LHCY and LB12 for any fixed value of LCLC. WHen LCLC changes, the Y intercept of the straight line changes, but the slope remains the same. Since the slope remains the same, the change in LHCY per unit change in LB12 is the same for all values of LCLC.



If you draw the regression lines for various values of LCLC in the scatterplot of LHCY against LB12, you get a series of parallel lines, that is, you get the regression plane viewed by sighting down the LCLC axis.

The same argument applies to the regression surface for fixed LB12.

The first important lesson to be learned is that the shape of the regression surfaces and the properties of the regression equation follow from the model **we choose to fit to the data**. The second is that **we are responsible for the models we fit**. We are obliged to understand the interpretation and consequences of the models we fit. It we don't believe a particular type of model will adequately describe a dataset, we shouldn't be fitting that model! The responsibility is not with the statistical software. It is with the analyst.

Last modified: undefined.

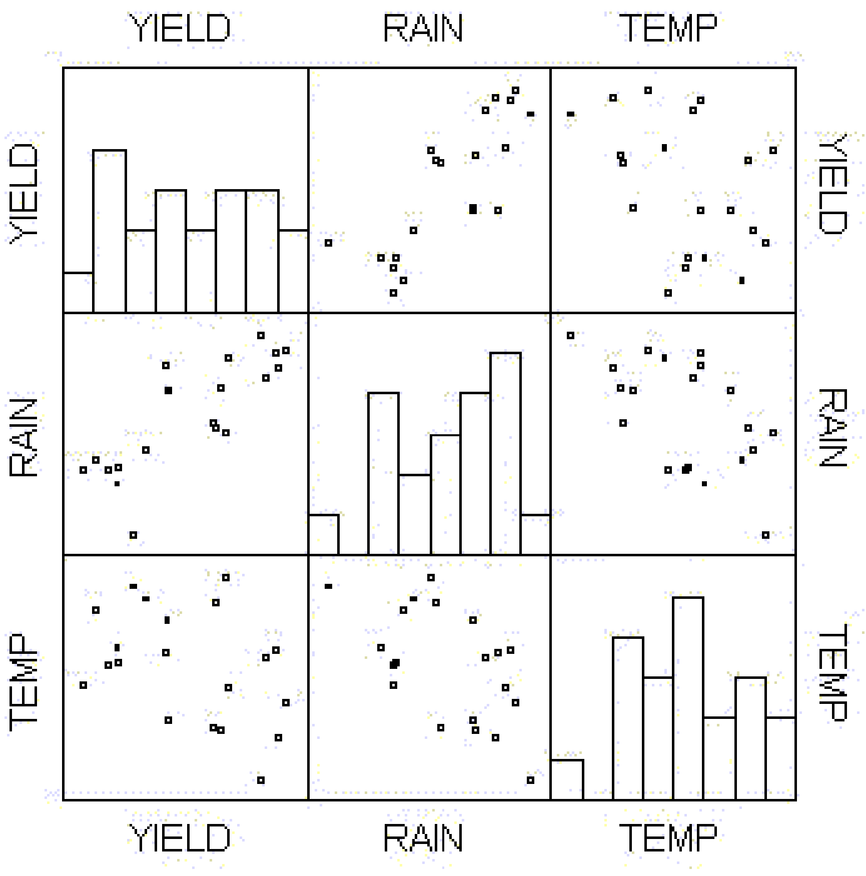
Partial Correlation Coefficients

Gerard E. Dallal, Ph.D.

Scatterplots, correlation coefficients, and simple linear regression coefficients are inter-related. The scattterplot shows the data. The correlation coefficient measures of linear association between the variables. The regression coefficient describes the linear association through a number that gives the expected change in the response per unit change in the predictor.

The coefficients of a multiple regression equation give the change in response per unit change in a predictor when all other predictors are held fixed. This raises the question of whether there are analogues to the correlation coefficient and the scatterplot to summarize the relation and display the data after adjusting for the effects of other variables.

This note answers these questions and illustrates them by using the crop yield example of Hooker reported by Kendall and Stuart in volume 2 of their *Advanced Theory of Statistics, Vol, 2, 3 rd ed.*(example 27.1) Neither Hooker nor Kendall & Stuart provide the raw data, so I have generated a set of random data with means, standard deviations, and correlations identical to those given in K&S. These statistics are sufficient for all of the methods that will be discussed here (*sufficient* is a technical term meaning nothing else to do with the data has any effect on the analysis. All data sets with the same values of the sufficient statistics are equivalent for our purposes), so the random data will be adequate.



The variables are yields of "seeds' hay" in cwt per acre, spring rainfall in inches and the accumulated temperature above 42 F in the spring for an English area over 20 years. The plots suggest yield and rainfall are positively correlated, while yield and temperature are negatively correlated! This is borne out by the correlation matrix itself.

```
   Pearson Correlation
Coefficients, N = 20
             Prob > |r| under
H0: Rho=0

             YIELD
RAIN         TEMP

YIELD    1.00000    0.80031   -
0.39988
                       <.0001
0.0807
```

```
RAIN    0.80031   1.00000  -0.55966
         <.0001              0.0103

TEMP   -0.39988  -0.55966   1.00000
         0.0807    0.0103
```

Just as the simple correlation coefficient between Y and X describes their joint behavior, the partial correlation describes the behavior of Y and X_1 when $X_2..X_p$ are held fixed. The partial correlation between Y and X_1 holding $X_2..X_p$ fixed is denoted $r_{X_1Y\bullet X_2..X_p}$ or $r_{X_1Y|X_2..X_p}$.

A partial correlation coefficient can be written in terms of simple correlation coefficients

$$r_{XY\bullet Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}}$$

Thus, $r_{XY|Z} = r_{XY}$ if X & Y are both uncorrelated with Z.

A partial correlation between two variables can differ substantially from their simple correlation. Sign reversals are possible, too. For example, the partial correlation between YIELD and TEMPERATURE holding RAINFALL fixed is 0.09664. While it does not reach statistical significance (P = 0.694), the sample value is positive nonetheless.

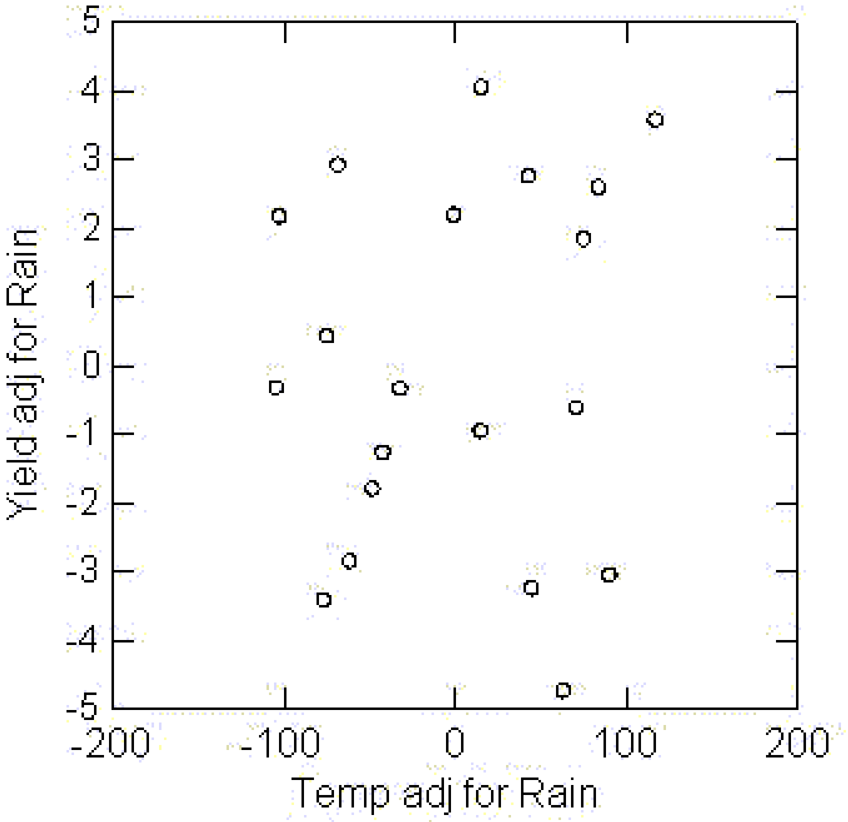
The partial correlation between X & Y holding a set of variables fixed will have the same sign as the multiple regression coefficient of X when Y is regressed on X and the set of variables being held fixed. Also,

$$r_{XY\bullet\text{list}} = \frac{t}{\sqrt{t^2 + \text{Res } df}}$$

where *t* is the t statistic for the coefficient of X in the multiple regression of Y on X and the variables in the list.

Just as the simple correlation coefficient describes the data in an ordinary scatterplot, the partial correlation coefficient describes the data in the partial regression residual plot.

Let Y and X_1 be the variables of primary interest and let $X_2..X_p$ be the variables held fixed. First, calculate the residuals after regressing Y on $X_2..X_p$. These are the parts of the Ys that cannot be predicted by $X_2..X_p$. Then, calculate the residuals after regressing X_1 on $X_2..X_p$. These are the parts of the X_1s that cannot be predicted by $X_2..X_p$. The partial correlation coefficient between Y and X_1 adjusted for $X_2..X_p$ is the correlation between these two sets of residuals. Also, the regression coefficient when the Y residuals are regressed on the X_1 residuals is equal to the regression coefficient of X_1 in the multiple regression equation when Y is regressed on the entire set of predictors.



For example, the partial correlation of YIELD and TEMP adjusted for RAIN is the correlation between the residuals from regressing YIELD on RAIN and the residuals from regressing TEMP on RAIN. In this partial regression residual plot, the correlation is 0.9664. The regression coefficient of TEMP when the YIELD residuals are regessed on the TEMP residuals is 0.003636. The multiple regression equation for the original data set is

YIELD = 9.298850 + 3.373008 RAIN + 0.003636 TEMP

Because the data are residuals, they are centered around zero. The values, then, are not similar to the original values. However, perhaps this is an advantage. It stops them from being misinterpreted as Y or X_1 values "adjusted for $X_2..X_p$".

While the regression of Y on $X_2..X_p$ seems reasonable, it is not uncommon to hear questions about adjusting X_1, that is, some propose comparing the residuals of Y on $X_2..X_p$ with X_1directly.

This approach has been suggested many times over the years. Lately, it has been used in the field of nutrition by Willett and Stampfer (AJE, 124(1986):17-22) to produce "calorie-adjusted nutrient intakes", which are the residuals obtained by regressing nutrient intakes on total energy intake. These adjusted intakes are used as predictors in other regression equations. However, total energy intake does not appear in the equations and the response is not adjusted for total energy intake. Willett and Stampfer recognize this, but propose using calorie-adjusted intakes nonetheless. They suggest "calorie-adjusted values in multivariate models will overcomethe problem of high-collinearity frequently observed between nutritional factors", but this is just an artifact of adjusting only some of the factors. The correlation between an adjusted factor and an unadjusted factor is always smaller in magnitude than the correlation between two adjusted factors.

This method was first proposed before the ready availability of computers as a way to approximate multiple regression with two independent variables (regress Y on X1, regress the residuals on X2) and was given the name two-stage regression. Today, however, it is a mistake to use the approximation when the correct answer is easily obtained. If the goal is to report on two variables after adjusting for the effects of another set of variables, then both variables must be adjusted.

Last modified: undefined.

Which Predictors Are More Important?

Gerard E. Dallal, Ph.D.

When a multiple regression is fitted, it is not uncommon for someone to ask which predictors are more important. This is a reasonable question. There have been some attempts to come up with a purely statistical answer, but they are unsatisfactory. The question can be answered only in the context of a specific research question by using subject matter knowledge.

To focus the discussion, consider the regression equation for predicting HDL cholesterol presented earlier.

```
                          The REG Procedure
                      Dependent Variable: LHCHOL

                         Parameter Estimates

            Parameter     Standard               Pr >    Standardized
Variable     Estimate        Error       T        |t|       Estimate

Intercept     1.16448      0.28804     4.04     <.0001              0
AGE          -0.00092      0.00125    -0.74     0.4602       -0.05735
BMI          -0.01205      0.00295    -4.08     <.0001       -0.35719
BLC           0.05055      0.02215     2.28     0.0239        0.17063
PRSSY        -0.00041      0.00044    -0.95     0.3436       -0.09384
DIAST         0.00255      0.00103     2.47     0.0147        0.23779
GLUM         -0.00046      0.00018    -2.50     0.0135       -0.18691
SKINF         0.00147      0.00183     0.81     0.4221        0.07108
LCHOL         0.31109      0.10936     2.84     0.0051        0.20611
```

The predictors are age, body mass index, blood vitamin C, systolic and diastolic blood pressures, skinfold thickness, and the log of total cholesterol. The regression coefficients range from 0.0004 to 0.3111 in magnitude.

One possibility is to measure the importance of a variable by the magnitude of its regression coefficient. This approach fails because the regression coefficients depend on the underlying scale of measurements. For example, the coefficient for AGE measures the expected difference in response for each year of difference in age. If age were recorded in months instead of years, the regression coefficient would be divided by 12, but surely the change in units does not change a variable's importance.

Another possibility is to measure the importance of a variable by its observed significance level (P value). However, the distinction between statistical significant and practical importance applies here, too. Even if the predictors are measured on the same scale, a small coefficient that can be estimated precisely will have a small P value, while a large coefficient that is not estimate precisely will have a large P value.

In an attempt to solve the problem of units of measurement, many regression programs provide **standardized regression coefficients**. Before fitting the multiple regression equation, all variables--response and predictors--are standardized by subtracting the mean and dividing by the standard deviation. The standardized regression coefficients, then, represent the change in response for a change of one standard deviation in a predictor. Some like SPSS report them automatically, labeling them "Beta" while the ordinary coefficients are labelled "B". Others, like SAS, provide them as an option and label them "Standardized Coefficient".

Advocates of standardized regression coefficients point out that the coefficients are the same regardless of a predictor's underlying scale of units. They also suggest that this removes the problem of comparing *years* with *mm Hg* since each regression coefficient represents the change in response per standard unit (one SD) change in a predictor. However, this is illusory. there is no reason why a change of one SD in one predictor should be equivalent to a change of one SD in another predictor. Some variables are easy to change--the amount of time watching television, for example. Others are more difficult--weight or cholesterol level. Others are impossible--height or age.

The answer to which variable is most important depends on the specific context and why the question is being asked. The investigator and the analyst should consider specific changes in each predictor and the effect they'd have on the response. Some predictors will not be able to be changed, regardless of their coefficients. This is not an issue if the question asks what most determines the response, but it is critical if the point of the exercise is to develop a public policy to effect a change in the response. When predictors can be modified, investigators will have to decide what changes are feasible and what changes are comparable. Cost will also enter into the discussion. For example, suppose a change in response can be obtained by either a large change in one predictor or a small change in another predictor. According to circumstances, it might prove more cost-effective to attempt the large change than the small change.

Last modified: undefined.

The Extra Sum of Squares Principle

Gerard E. Dallal, Ph.D.

The *Extra Sum of Squares Principle* allows us to compare two models for the same response where one model (the full model) contains all of the predictors in the other model (the reduced model) and more. For example, the reduced model might contain *m* predictors while the full model contains *p* predictors, where *p* is greater than *m* and all of the *m* predictors in the reduced model are among the *p* predictors of the full model, that is,

$$\text{full: } Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m + \cdots + \beta_p X_p + \varepsilon$$

$$\text{reduced: } Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m + \varepsilon$$

The extra sum of squares principle allows us to determine whether there is statistically significant predictive capability in the set of additional variables. The specific hypothesis it tests is

$$H_0\text{: } \beta_{m+1} =..= \beta_p = 0$$

The method works by looking at the reduction in the Residual Sum of Squares (or, equivalently, at the increase in Regression Sum of Squares) when the set of additional variables is added to the model. This change is divided by the number of degrees of freedom for the additional variables to produce a mean square. This mean square is compared to the Residual mean square from the full model. Most full featured software packages will handle the arithmetic for you. All the analyst need do is specify the two models.

Example: An investigator wanted to know, in this set of cross-sectional data, whether muscle strength was predictive of bone density after adjusting for age and measures of body composition. She had eight strength measures and no prior hypothesis about which, if any, might be more useful than the others. In such situations, it is common practice to ask whether there is any predictive capability in the set of strength measures.

Two models will be fitted, one containing all of the predictors and the other containing everything but the strength measures. The extra sum of squares principle can then be used to assess whether there is any predictive capability in the set of strength measures.

```
                    ** ** **  Full Model ** ** **

                            Sum of        Mean
Source               DF    Squares      Square     F Value    Pr > F
```

```
Model                13    0.33038    0.02541       4.86    0.0003
Error                26    0.13582    0.00522
Corrected Total      39    0.46620
------------------------------------------------------------------


                  ** ** **  Reduced Model ** ** **

                            Sum of       Mean
Source               DF    Squares     Square    F Value    Pr > F

Model                 5    0.18929    0.03786       4.65    0.0024
Error                34    0.27691    0.00814
Corrected Total      39    0.46620
------------------------------------------------------------------


                 ** ** ** Extra Sum of Squares ** ** **

                                  Mean
          Source         DF     Square    F Value    Pr > F

          Numerator       8    0.01764       3.38    0.0087
          Denominator    26    0.00522
```

Adding the strength measures to the model increases the Regression Sum of Squares by 0.14109 (=0.33038-0.18929). Since there are eight strength measures, the degrees of freedom for the extra sum of squares is 8 and the mean square is 0.01764 (=0.14109/8). The ratio of this means square to the Error mean square from the full model is 3.38. When compared to the percentiles of the F distribution with 8 numerator degrees of freedom and 26 denominator degrees of freedom, the ratio of mean squares gives an observed significance level of 0.0087. From this we conclude that muscle strength is predictive of bone density after adjusting for various measures of body composition.

The next natural question is "which measures are predictive?" This is a difficult question, which we will put off for the moment. There are two issues. The first is the general question of how models might be simplified. This will be discussed in detail, but there is no satisfactory answer. The second is that there are too many predictors in this model--thirteen--to hope to be able to isolate individual effects with only 40 subjects.

Last modified: undefined.

Simplifying a Multiple Regression Equation

Sometimes research questions involve selecting the best predictors from a set of candidates that, at the outset, seem equally likely to prove useful. The question is typically phrased, "Which ones of these predictors do I need in my model?" or "Which predictors really matter?" Although many methods have been proposed, **the standard purely statistical approaches for simplifying a multiple regression equation are unsatisfactory.** The reason is simple. With rare exception, **a hypothesis cannot be validated in the dataset that generated it**

Many multiple regression models contain variables whose t statistics have nonsignificant P values. These variables are judged to have not displayed statistically significant predictive capability in the presence of the other predictors. The question is then whether some variables can be removed from the model. To answer this question, many models are examined to find the one that's *best* in some sense.

The theoretical basis for concern over most simplification methods

The main concern is that many of the measures used to assess the importance of a variable were developed for examining a single variable only. They behave differently when assessing the *best*.

If you take a fair coin and flip it 100 times, common sense as well as probability theory says the chance of getting more heads than tails is 50%. However, suppose a large group of people were to each flip a coin 100 times. Again, both common sense and probability theory say that it is **unlikely** that the coin with the **most heads** has more tails than heads. For the best coin to show more tails than heads, they would *all* have to show more tails than heads. The chance of this becomes smaller the more coins that are flipped.

Most of the time (95%) there will be between 40 to 60 heads when a single fair coin is flipped 100 times. The chances of getting more than 60 heads are small, *for a single coin*. If we were suspicious of a particular coin and noticed that there were 70 heads in the next 100 flips, we'd have some statistical evidence to back up our suspicions.

Have a large group of people flip fair coins 100 times and the chances of *someone* getting more than 60 heads grows. On average, about 2.5% of the participants will get more than 60 heads.

In this situation, we *might* become suspicious of the coin that recorded the most heads but, we'd have to test it again to be sure. If we had no reason other than the number of heads for being suspicious and were flip the coin another 100 times, it wouldn't be surprising to see it behave more typically this time.

To summarize:

- If our attention is drawn to a particular coin and it *subsequently* shows an excess of heads, we have some basis for our suspicion.
- If a large number of coins are flipped, we can't judge the coin with the largest number of heads as though it were the only coin flipped.

The same thing applies to building models. If there is a reason for singling out a predictor before the data are collected, then it is fair to say the variable has predictive value if it achieves statistical significance. However,

- when many predictors are considered and
- there is nothing special about any of them before the data are collected and
- we judge them as though each were the only predictor being considered,

probability theory says that *something* will likely achieve statistical significance due to chance alone. We shouldn't be surprised if 1 of 20 such predictors achieves what would be statistical significance for a single predictor at the 0.05 level. This explains why measures that are unrelated to the response sometimes appear to be statistically significant predictors.

A similar problem arises when many variables predict the response equally well. Statistical theory says that, in any sample, some variables will appear to be better predictors than others. However, since all variables predict equally well, these particular variables are not really better. They appear that way due to chance, showing once again that a hypothesis cannot be validated in the dataset that generated it.

We need a procedure that can distinguish between variables that are truly better predictors and those that appear to be better due to the luck of the draw. Unfortunately, that procedure does not yet exist. Still, many procedures have been tried.

Stepwise Procedures

One approach to simplifying multiple regression equations is the stepwise procedures. These include **forward selection**, **backwards elimination**, and **stepwise regression**. They add or remove variables one-at-a-time until some stopping rule is satisfied. They were developed before there were personal computers, when time on mainframe computers was at a premium and when statisticians were considering the problem of what to do when there might be more predictors than observations.

Forward selection starts with an empty model. The variable that has the smallest P value when it is the only predictor in the regression equation is placed in the model. Each subsequent step adds the variable that has the smallest P value in the presence of the predictors already in the equation. Variables are added one-at-a-time as long as their P values are small enough, typically less than 0.05 or 0.10.

Backward elimination starts with all of the predictors in the model. The variable that is least significant--that is, the one with the largest P value--is removed and the model is refitted. Each subsequent step removes the least significant variable in the model until all remaining variables have individual P values smaller than some value, such as 0.05 or 0.10.

Stepwise regression is similar to forward selection except that variables are removed from the model if they become nonsignificant as other predictors are added.

Backwards elimination has an advantage over forward selection and stepwise regression because it is possible for a set of variables to have considerable predictive capability even though any subset of them does not. Forward selection and stepwise regression will fail to identify them. Because the variables don't predict well individually, they will never get to enter the model to have their joint behavior noticed. Backwards elimination starts with everything in the model, so their joint predictive capability will be seen.

Since variables are chosen because they look like good predictors, estimates of anything associated with prediction can be misleading. Regression coefficients are biased away from 0, that is, their magnitudes often appear to be larger than they really are. (This is like estimating the probability of a head from the fair coin with the most heads as the value that gained it the title of "most heads.") The t statistics tend to be larger in magnitude and the standard errors smaller than what would be observed if the study were replicated. Confidence intervals tend to be too narrow. Individual P values are too small. R^2, and even adjusted R^2, is too large. The overall F ratio is too large and its P value is too small. The standard error of the estimate is too small.

> Nominal Significance: Stepwise procedures are sometimes described as adding variables one-at-a-time as long as they are statistically significant or removing them if they are nonsignificant. This means comparing a variable's P values to some predetermined value, often 0.05. With forward selection, we are looking at the smallest P value. With backwards elimination, we are looking at the largest P value. However, for the reasons already stated, these P values are artificially small or large. It is incorrect to call them *statistically significant* because the reported P values don't take account of the selection procedure. To acknowledge this, many statisticians call them *nominally significant*, that is, significant in name only.

When variables are highly correlated, the ones that appear in the model do so as a matter of chance and can change with the addition of one or two more observations. In general, the idea that our assessment of a particular predictor might change with the addition of one or two observations doesn't bother me. That's part of the game. We choose our test, collect our data, and calculate the results, letting the chips fall where they may. In multiple regression, the worst that can happen is that some coefficients and P values might change a bit. P values might move from one side of 0.05 to the other, but confidence intervals for regression coefficients will be grossly the same. The troublesome feature of *stepwise* procedures is that the characteristics of the report model can change dramatically, with some variables entering and others leaving.

A final condemnation of stepwise procedures is often encountered when missing data are involved. Stepwise procedures must exclude observations that are missing *any* of the *potential* predictors. However, some of these observations will not be missing any of the predictors in the final model. Sometimes one or more of the predictors in the final model are no longer statistically significant when the model is fitted to the data set that includes these observations that had been set aside, even when values are missing at random.

All possible regressions

Other simplification procedures examine all possible models and choose the one with the most favorable value of some summary measure such as adjusted R^2 or Mallows' C(p) statistic. "All Possible Regressions" has a **huge** advantage over stepwise procedures, namely, it can let the analyst see competing models, models that are almost as good as the "best"and possibly more meaningful to a subject matter specialist. However, whenever a model is chosen because of an extreme value of some summary statistic, it suffers from those same problems already mentioned. While I've seen many discussions of examining all possible models, I've never seen a report of anyone doing this in practice.

Some investigators suggest plotting various summary statistics from different models as a function of the number of predictors. When the standard error of the estimate is plotted against the number of predictors, the SEE will typically drop sharply until some floor is reached. The number of predictors needed to adequately describe the data is suggested by where the floor starts. The final model might be chosen from competing models with the same number of predictors, or maybe 1 or 2 more, by our knowledge of the variables under study. In similar fashion, some statisticians recommend using knowledge of the subject matter to select from nearly equivalent models the first time C(p) meets its target of being less than or equal to p+1.

Data Splitting

Another approach to the problem is **data splitting**. The dataset is divided in two, at random. One piece is used to derive a model while the other piece is used to verify it. The method is rarely used. In part, this is due to of the loss in power (ability to detect or verify effects) from working with only a subset of the data. Another reason is a general because that different investigators using the same data could split the data differently and generate different models.

It has always struck me as a peculiar notion that one could use a subset of the data challenge from what was observed in the reaminder or in data as a whole. if the full dataset has some peculiarity, the laws of probability dictate that each of the two halves should share it.

The Bootstrap

Today, some analysts are looking to the **bootstrap** for assistance. Bootstrap samples are obtained by selecting observations with replacement from the original sample. Usually, bootstrap samples are the same size as the original sample. They can be the same size as the original sample because the observations composing a bootstrap sample are chosen independently with replacement (that is, when an observation is chosen, it is thrown back into the pot before another is chosen). The typical bootstrap sample will contain duplicates of some original observations and no occurrences of others. The stepwise procedure is applied to each bootstrap sample to see how the model changes from sample to sample, which, it is hoped, will give some indication of the stability of the model. I am not optimistic about this approach for reasons stated here.

So...what do we do?

Some analysts soldier on regardless and look for consistency among the methods. They gain confidence in a model if most every method leads to the same candidate. Perhaps there is some justification for this belief, but I am inclined to think not. If due to chance a particular set of variables looks better than it really is, it's unlikely that the reason for this excellence will be uncovered, regardless of the lens used to examine the data.

Perhaps this outlook is too pessimistic. In a November, 2000, post to the S-News mailing list for users of S-Plus, Jeff Simonoff presented a cogent argument for using automatic methods. He states that he considers stepwise methods obsolete but does talk about "all subsets regression" in his teaching. He is adamant about validation, but would use a version of data splitting to do it. The central point of his argument is given here in case the link to his post should become inoperative:

> I can't agree, however, with the comments...that state that these problems with inference measures imply "never do it." The arguments that inference methods are based on prespecified hypotheses didn't impress me 25 years ago (when I was learning statistics), and they still don't. Nobody *ever* does statistics this way; if we did, we would never identify outliers, look for transformations, enrich the model in response to patterns in residual plots, and so on (all of which also can increase the apparent strength of a regression). Further, I would argue that with the explosion of methods commonly called "data mining," these pieces of advice are ludicrously anachronistic. All subset regression is nothing compared to those kinds of methods. We are no longer in the era of small data sets isolated from each other in time; we are now in one of large (or even massive) ones that are part of an ongoing continuing process. In that context, I would argue that automatic methods are crucial, and the key for statisticians should be to get people to validate their models and correct for selection effects, not tell them what nobody ever believed anyway.

On one level, I've no argument with this stance. I would *qualify it* by saying that activities such as identifying outlier and searching for transformations within a narrow set of options (original or logarithmic scales) are fundamentally different in nature from automatic model fitting procedures because they are done to improve the validity of our models. No one would argue with stopping an outlier from producing a model that failed to fit the bulk of the data, nor would anyone argue for fitting a linear model to highly nonlinear data. The important point is that automatic methods can be useful as long as the model is tested in other data sets. Unfortunately, too often studies are *not* repeated, if only because there's no glory in it, and the results of automatic model fitting procedures are treated as though they came from validation studies.

I have no doubt that stepwise and "all possible models" procedures can identify gross effects such as the dependence of body weight on caloric intake. However, in practice these procedures are often used to tease out much more complicated and subtle effects. It is these less obvious relationships that, in my experience, are less likely to be reproduced. Saying that these procedures are fine as long as the model is validated may offer false hope in these cases.

Final Comments

It's easy to cheat. When we fit a model to data and report our findings, it is essential to describe how we got the model so that others can judge it properly. It is impossible to determine from the numerical results whether a set of predictors was specified before data collection or was obtained by using a selection procedure for finding the "best" model. The parameter estimates and ANOVA tables don't change according to whether or not a variable selection procedure was used. The results are the same as what would have been obtained if that set of predictor variables had been specified in advance.

Perhaps the fundamental problem with automatic methods is that they often substitute for thinking about the problem. As Shayle Searle wrote in Section 1.1, Statistics and Computers, of his *Linear Models For Unbalanced Data*, published in 1987 by John Wiley & Sons, Inc., of New York:

> Statistical computing packages available today do our arithmetic for us in a way that was totally unthinkable thirty years ago. The capacity of today's computers for voluminous arithmetic, the great speed with which it is accomplished, and the low operating cost per unit of arithmetic--these characteristics are such as were totally unimaginable to most statisticians in the late 1950s. Solving equations for a 40-variable regression analysis could take six working weeks, using (electric) mechanical desk calculators. No wonder that regression analyses then seldom involved many variables. Today that arithmetic takes no more than ten seconds... But the all-important question would then be: Does such an analysis make sense?

> Thinking about such a question is essential to sane usage of statistical computing packages. Indeed, a more fundamental question prior to doing an intended analysis is "Is it sensible to do this analysis?". Consider how the environment in which we contemplate this question has changed as a result of the existence of today's packages. Prior to having high-speed computing, the six weeks that it took for solving the least squares equations for a 40-variable regression analysis had a very salutory effect on planning the analysis. One did not embark on such a task lightly; much forethought would first be given as to whether such voluminous arithmetic would likely be worthwhile or not. Questions about which variables to use would be argued at length: are all forty necessary, or could fewer suffice, and if so, which ones? Thought-provoking questions of this nature were not lightly dismissed. Once the six-week task were to be settled on and begun, there would be no going back; at least not without totally wasting effort up to that point. Inconceivable was any notion of "try these 40 variables, and then a different set of maybe 10, 15 or 20 variables". Yet this is an attitude that can be taken today, because computing facilities (machines and programs) enable the arithmetic to be done in minutes, not weeks, and at very small cost compared to six weeks of human labor. Further; and this is the flash-point for embarking on thoughtless analyses, these computing facilities can be initiated with barely a thought either for the subject-matter of the data being analyzed or for that all-important question "Is this a sensible analysis?"
>
> ...[V]ery minimal (maybe zero) statistical knowledge is needed for getting what can be voluminous and sophisticated arithmetic easily accomplished. But that same minimal knowledge may be woefully inadequate for understanding the computer output, for knowing what it means and how to use it.

If You Need Further Convincing

Everything I've written is true, but I've noticed that many people have trouble fully grasping it. It may seem reasonable that when a program is allowed to pick the best variables, everything will look better than it would if the predictors were picked at random, but the idea often remains an abstraction.

Simulations can make this more concrete. I'm not a Java programmer (I think it's for the best. Otherwise, I'd be doing nothing but programming!), so I don't have any applets to offer. However, I've written some SAS code to illustrate the problem.

The first example looks at whether the intake of various vitamins affects the time it takes to commute to work. One hundred fifty subjects keep a 7 day diary to record their dietary intake and the time it takes to commute to work. In the command language that follows, every pair of variables looks like a sample from a population in which the correlation coefficient is rho. Here, rho = 0, so the data are drawn from a population in which none of the variables are associated with each other.

If you paste the command language into SAS, you'll find that forward selection regression with a significance-level-to-enter of 0.05 will select something 54% of the time. That is, at least one vitamin will appear to be associated with commuting time in more than half of the instances when the program is run, even though these observations are drawn from a population in which no two variables are associated!

[The constants in the variable definitions make the values look more realistic. For example, the commuting times will look like a sample from a normal distribution with a mean of 1 hour and a SD of 15 minutes (= 0.25 hour), the vitamin A values will look like a sample from a normal distribution with a mean of 800 IUs and an SD of 200 IUs, and so on. These adjustments are linear transformation, which have no effect on the correlations between the variables. Someone wanting simpler code and generic variables could change the definitions to

```
   variable_name = rannor(0) + d;
```

to obtain random values from a normal distribution with a mean of 0.]

```
options ls=80 ps=56;

data analysis;
   rho = 0;
   c = (rho/(1-rho))**0.5;
     do i = 1 to 150;
       d = c * rannor(0);
       commute   =   1 + 0.25 * (rannor(0) + d);
       vit_A     = 800 +  200 * (rannor(0) + d);
       vit_B1    = 1.3 +  0.3 * (rannor(0) + d);
       vit_B2    = 1.7 +  0.4 * (rannor(0) + d);
       vit_B6    = 2.0 +  0.4 * (rannor(0) + d);
       vit_B12   = 2.0 + 0.35 * (rannor(0) + d);
       vit_C     =  55 +   14 * (rannor(0) + d);
       vit_D     =   8 +    2 * (rannor(0) + d);
       vit_E     =   9 +  2.2 * (rannor(0) + d);
       vit_K      =  60 +   12 * (rannor(0) + d);
       calcium    = 800 +  200 * (rannor(0) + d);
       folate     = 190 +   30 * (rannor(0) + d);
       iron       =  12 +    4 * (rannor(0) + d);
       niacin     =  15 +    3 * (rannor(0) + d);
       magnesium = 300 +   50 * (rannor(0) + d);
       potassium =  75 +   10 * (rannor(0) + d);
       zinc       =  13 +    3 * (rannor(0) + d);
       output;
     end;
   keep commute vit_A vit_B1 vit_B2 vit_B6 vit_B12 vit_C vit_D
          vit_E vit_K calcium folate iron magnesium niacin potassium
zinc;

proc reg data=analysis;
    model commute = vit_A vit_B1 vit_B2 vit_B6 vit_B12 vit_C vit_D
            vit_E vit_K calcium folate iron magnesium niacin
potassium zinc /
            selection=forward sle=0.05 ;
run;
```

SAS PROCs can be placed after the data step to check on the data, for example, to see that the correlation coefficients behave like a sample from a population in which they are all 0.

The second example is even more troublesome because it has some plausibility to it. Also, as you think about what you might do with the results of any one of these "experiments", you'll probably be reminded of a few published reports you've read.

Let the response be *birth weight* instead of commuting time, and let the vitamin variables measure the nutritional status of the baby's mother. We know nutrients are related to each other and it is likely that they will have an effect on birth weight. To reflect this in the data,

1. change all instances of *commute* to *bwt*,
2. let birth weight (in grams) be defined by

   bwt = 2200 + 225 * (rannor(0) + d);

   and
3. change *rho* to 0.50.

Then, data will be observations drawn from a population in which every pair of variables has a correlation of 0.50. We're no longer upset at seeing the predictors (vitamin levels) related to the response (birth weight) because it's now biologically plausible. However, since every pair of variables has the same correlation, the particular variables that enter the forward selection regression equation will be *a matter of chance alone*. This illustrates the danger of using an automated procedure to decide *which* predictors are important.

[back to The Little Handbook of Statistical Practice]

Gerard E. Dallal

Last modified: undefined.

Using the Bootstrap to Simplify a Multiple Regression Equation

Gerard E. Dallal, Ph.D.

I am unaware of any formal literature about the validity of using the bootstrap to simplify a multiple regression equation, but examples appear in the scientific literature occasionally. One of the earliest comes from Gail Gong's doctoral thesis. It is recounted in her article with Brad Efron in the American Statistician (Vol 37, Feb 1983, 36-48). The model building technique was not exactly a pure forward selection procedure, but was quite close. From a set of 19 predictors, she

1. ran 19 separate single-predictor logistic regressions, noting which variables achieved significance at the 0.05 level.
2. ran a forward selection multiple logistic regression program with an 0.10 level of significance as the stopping criterion, using the statistically significant predictors from step 1.
3. ran a forward selection stepwise (that is, allowing for removals) logistic regression with an 0.05 level of significance as the entry/removal criterion, using the variables that entered the model developed in step 2.

> "Figure 6 illustrates another use of the bootstrap replications. The predictions chosen by the three-step selection procedure, applied to the bootstrap training set X^* are shown for the last 25 of 500 replications. Among all 500 replications, predictor 13 was selected 37 percent of the time, predictor 15 selected 48 percent, predictor 7 selected 35 percent, and predictor 20 selected 59 percent. No other predictor was selected more than 50 percent of the time. No theory exists for interpreting Figure 6, but the results certainly discourage confidence in the casual nature of the predictors 13, 15, 7, 20." (Efron and Gong, p. 48)

Phillip Good in his 2003 text *Common Errors in Statistics: (And How To Avoid Them)* (2003, John Wiley & Sons, pp 147) makes this approach central to his model building strategy.

> We strongly urge you to adopt Dr. Gong's bootstrap approach to validating multi-variable models. Retain only those variables which appear consistently in the bootstrap regression models.

Pointing to such examples as Gong's, I've done something similar a few times when investigators were determined to use stepwise regression. I implemented the bootstrap the hard way--generating individual datasets, analyzing them one-at-a-time (in a batch program) and using a text processor to extract relevant portions of the output.

Recently I decided to automate the procedure by writing a SAS macro that not only generated and analyzed the bootstrap samples, but also used the SAS output delivery system to collect the results. That way, the entire process could be carried out in one step. I analyzed a half-dozen research datasets. I found *no* cases where the bootstrap suggested instability in the model produced by stepwise regression applied to the original dataset. This was not necessarily a problem. It could have been that the signals were so strong that they weren't distorted by stepwise regression.

To test this theory, I generated some random datasets so that I could control their structure. Each consisted of 100 cases containing a response and 10 predictors. The variables were jointly normally distributed with the same underlying correlation between any pair of variables. Therefore, all ten predictors predicted the response equally well. Setting the correlation to something other than 0 insured that some predictors would enter the stepwise regression equation, but the ones that entered would be just a matter of chance.

When I did this, I found the same thing as in the real data. The bootstrap samples pointed to the same model as the stepwise regression on the full dataset. For example, one dataset with a common underlying correlation of 0.50 (Here's the code if you'd like to try it yourself.) led to a forward selection regression model that included X1, X3, and X5. In the 100 bootstrap samples drawn from this dataset, the 10 predictors entered with the following frequencies.

Variable Entered	Number of Appearances
X1	57
X2	4
X3	83
X4	8
X5	76
X6	28
X7	4
X8	6
X9	14
X10	9

And the winners are...X3, X5, and X1! We can quibble over X1, but the frequency with which X3 and X5 appear are impressive. There is nothing to suggest these are random data.

It appears the problem is what's worried me all along about data splitting. Whatever peculiarities in the dataset that led X1, X3, and X5 to be the chosen ones in the stepwise regressions also make them the favorites in the bootstrap samples. In retrospect, it seems obvious that this would happen. Yet, even Gong & Efron considered this approach as a possibility. While Good (page 157, step 4) advises limiting attention of one or two of the most significant predictor variables, the examples here show that such advice is not enough to avoid choosing an improper model. Good warns about the importance of checking the validity of the model in another set of data, but it is not easy to do and seems to happen too seldom in practice.

My hope is that statisticians will discover how to modify the bootstrap to study model stability properly. Until then, I'll no longer be using it to evaluate models generated by stepwise regression, but it would have been nice if it worked.

[back to The Little Handbook of Statistical Practice]

Last modified: undefined.

Simplifying a Multiple Regression Equation: The Real Problem!

Gerard E. Dallal, Ph.D.

[Early draft subject to change.]

When my students and colleagues ask me whether a particular statistical method is appropriate, I invariably tell them to state their research question and the answer will be clear. Applying the same approach to regression models reveals the real barrier to using automatic model fitting procedures to answer the question, "Which variables are important?"

Let's back up. It is well-known that when testing whether the mean change produced by a treatment is different for two groups, it is not appropriate to evaluate the mean change for each group separately. That is, it is not appropriate to say the groups are different if the mean change in one group is statistically significant while the other is not. It may be that the mean changes are nearly identical, with the P value for one group being slightly less than 0.05 and the other slightly more than 0.05. To determine whether the mean changes for the two groups differ, the changes have to be compared directly. perhaps by using Student's t test for independent samples applied to changes for the two groups..

There's a similar problem with simplifying multiple regression models. The automatic techniques find a model that fits the data. However, the question isn't just a matter of what model fits the data, but what model is demonstrably better than all other models in terms of fit or relevance. In order to do this, automatic procedures would have to compare models to each other directly, but they don't! At least the stepwise procedures don't.

The "all possible models" approach may suffer from trying to summarize a model in a single number and it certainly overestimates the utility of the models it identifies as best, However, unlike the stepwise procedures, the "all possible models" approach gives the analyst a feel for competing models. Unlike the automatic stepwise procedures which generate a single sequence of models, the "all possible models" approach forces the analyst to come to grips with the fact that there may be many models that look quite different from each other but fit the data almost equally well. However, because the technique overstates the value of the models it identifies as best, it is still necessary for those models to be evaluated in another dataset.

[back to The Little Handbook of Statistical Practice]

Last modified: undefined.

Which variables go into a multiple regression equation?

Gerard E. Dallal, Ph.D.

Estrogen and the Risk of Heart Disease

[This section talks about the potentially beneficial effect of estrogen on heart disease risk. In July, 2002, the estrogen plus progestin component of the Women's Health Initiative, the largest Hormone Replacement Therapy trial to date, was halted when it was discovered that women receiving HRT experienced heart attack, stroke, blood clots, and breast cancer at a higher rate than those who did not take HRT (Journal of the American Medical Association 2002;288:321-333). This study and others like it have discredited estrogen therapy as a means of lowering heart disease risk. However, these results are in stark contrast to epidemiological studies that show a protective benefit from estrogen therapy.

No one doubts the findings of the randomized trials, but it has been said that if the epidemiology is wrong, this will be the first time the epidemiology has failed so miserably. To date, no one has come up with a satisfactory explanation of the discrepancy. Michels and Manson review many of the proposed explanations in their 2003 editorial in *Circulation*.

I've decided to let the example remain until there is general consensus over the reason why the trials and epidemiology disagree. It should be noted, however, that Michels and Manson end their editorial with the recommendation that "HT should not be initiated or continued for primary or secondary prevention of cardiovascular disease."]

The October 25, 1985 issue of the New England Journal of Medicine is notable for the reason given by John C. Bailar III in his lead editorial: "One rare occasions a journal can publish two research papers back-to-back, each appearing quite sound in itself, that come to conclusions that are incompatible in whole or in part... In this issue we have another such pair."

The two papers were

- Wilson PWF, Garrison RJ, Castelli WP (1985), "Postmenopausal Estrogen Use, Cigarette Smoking, and Cardiovascular Morbidity In Women Over 50: The Framingham Study", New England Journal of Medicine, 313, 1038-1043.
- Stampfer MJ, Willett WC, Colditz GA, Rosner B, Speizer FE, Hennekens CH (1985), "A Prospective Study of Postmenopausal Estrogen Therapy and Coronary Heart Disease", New England Journal of Medicine, 313, 1044-1049.

Both papers were based on epidemiologic studies rather than intervention trials. Wilson et al. studied women participating in the Framingham Heart Study. Stampfer et al. studied women enrolled in Nurses' Health Study. The disagreement is contained in the last sentence of each abstract.

- *Wilson:* No benefits from estrogen use were observed in the study group; in particular, mortality from all causes and from cardiovascular disease did not differ for estrogen users and nonusers.
- *Stampfer:* These data support the hypothesis that the postmenopausal use of estrogen reduces the risk of severe coronary heart disease.

The reports generated an extensive correspondence suggesting reasons for the discrepancy (New England Journal of Medicine, 315 (July 10, 1986), 131-136). A likely explanation for the apparent inconsistency was proposed by Stamper:

> Among the reasons for the apparent discrepancy...may be their [Wilson's]...adjustment for the effects of high-density lipoprotein, which seems unwarranted, since high-density lipoprotein is a likely mediator of the estrogen effect. By adjusting for high-density lipoprotein, one only estimates the effect of estrogen beyond its beneficial impact on lipids.

Stampfer was saying that the way estrogen worked was by raising the levels of HDL-cholesterol, the so-called good cholesterol. When Wilson's group fitted their regression model to predict the risk of heart disease, they included both estrogen and HDL-cholesterol among their predictors. A multiple regression equation gives the effect of each predictor after adjusting for the effects of the other predictors (or, equivalently, with all other predictors held fixed). The Wilson equation estimated the effect of estrogen after adjusting for the effect of HDL cholesterol, that is the effect of estrogen when HDL cholesterol was not allowed to change. To put it another way, it estimated the effect of estrogen after adjusting for the effect of estrogen! This is an example of **over adjustment**--adjusting for the very effect you are trying to estimate.

Added Sugars

The November 14-18, 1999, annual meeting of the North American Association for the Study of Obesity in Charleston, SC, USA, included some presentations discussing the role of added sugar in the diet.

In "Do Added Sugars Affect Overall Diet Quality?", R. Forshee and M Storey developed a multiple regression model to predict the number of food group servings from the amount of added sugar in the diet. If added sugar was displacing important foods and nutrients from the diet, those eating more added sugar would be consuming less of these other important items. The models adjust for age, sex, fat, carbohydrates (less added sugar), protein, and alcohol. The investigators noted the regression coefficient for added sugars, while statistically significant, was always quite small. They interpret this as saying those who eat more added sugar do not have appreciably different predicted numbers of servings of grains, vegetables, fruits, dairy, or lean meat.

The interpretation was correct in its way, but it's hard to imagine how the result could have been otherwise. The result was predetermined! By adding fat, carbohydrates (less added sugar), protein, and alcohol to their statistical model, the researchers were asking the question, "When you take a bunch of people eating the same amount of fat, carbohydrates (less added sugar), protein, and alcohol, does their consumption of specific food groups vary according to the amount of added sugar they eat?" It would be truly astounding if other foods could vary much when fat, carbohydrates (less added sugar), protein, and alcohol were held fixed. By adding all of the components of food to the model, the investigators were asking whether food groups varied with added sugar intake when food was held constant!

They use their regression model as though it were developed on longitudinal data to predict the amount of added sugar it would take to reduce the number of predicted dairy servings by 1. They conclude, "Children would have to consume an additional 15 twelve-ounce cans of carbonated soft drinks to displace one serving of dairy foods." With a regression model in the background, it sounds very impressive but the words make no sense. One doesn't have to be a nutritionist to know an additional 15 twelve-ounce cans of carbonated soft drinks will displace a lot more than one serving of dairy foods! As nonsensical as this claim appears, the report garnered a lot of publicity as can be seen by using the terms "sugar" and "Forshee" in any Internet search engine.

A second presentation, "Energy Intake From Sugars and Fat In Relation to Obesity in U.S. Adults, NHANES III, 1988-94" by DR Keast, AJ Padgitt, and WO Song, shows how one's impression of the data can change with the particular model that is fitted.

Their figure 8 showed those in the highest quarter of sugar intake are least likely to have deficient intakes of selected nutrients, but this is undoubtedly true because those who eat more added sugar are eating more of everything. The researchers also report, "When the data are presented as quartiles of percent kilocalories from total sugars, individuals in the highest quartile of total sugars are more likely to fall below 2/3 of the RDA for all nutrients listed except for vitamin C (Figure 9)." The only way sweeteners can be greater percentage of one's diet is if other things are a lesser percentage. This leaves us with the question of the great American philosopher Johnny Cash who asks, "What is truth?" Is either piece of information relevant to assessing the effect of added sugar on nutritional status. I could argue more strenuously that the calorie-adjusted values are more pertinent.

Dietary Patterns and 20-year mortality

In "Dietary Pattern and 20 Year Mortality In Elderly Men In Finland, Italy, and the Netherlands: Longitudinal Cohort Study" (BMJ,315(1997), 13-17) Huijbregts, Feskens, Rasanen, Fidanza, Nissinen, Menotti, and Kromhout investigated whether healthy dietary patterns were inversely associated with mortality. The data were fitted by a survival model that included an indicator of a healthy diet among the predictors.

The researchers were faced with the thorny question of whether country should be included in the model. If country were included, the coefficient for diet would answer the question of whether diet was predictive of survival after accounting for the participants' country of residence. The authors argue against including country.

> Since dietary patterns are highly determined by cultural influences (for example, the Mediterranean dietary pattern), we did not adjust for country in the pooled population analyses. Country has a strong cultural component which is responsible for (part of) the variation in dietary patterns. Adjustment for this variable would result in an overcorrection and hence an underestimation of the true association between the quality of the diet and mortality.

It is true that the effect of diet will be underestimated to the extent to which diet and culture are correlated and there are other things about culture that predict survival. However, it is equally true that if country is left out of the model the effect of diet will be **overestimated** to the extent to which diet and culture are correlated and things in the culture other than diet affect longevity! For this reason, the conservative approach is to fit all known or suspected predictors of longevity, including country, so that claims for the predictive capability of a healthful diet will be free of counterclaims that a healthful diet is a surrogate for something else. The conservative approach means that we often lack the power to separate out individual effects, which is what happened here. The authors continue

> When the countries were analyzed separately, the associations between the healthy diet indicator and all cause mortality were essentially the same, although they no longer reached significance. This was due to a low statistical power resulting from the smaller numbers of subjects within a country.

When this happens, the investigators have no choice, in my opinion, but to design a better study. There are so many things in a culture other than diet then might influence survival that it seems unwise not to adjust for country. The authors are no doubt correct that "dietary patterns are highly determined by cultural influences", but this strikes me as an insufficient reason for allowing everything else associated with diet and survival to be attributed to diet. Good science is often expensive, inconvenient, and difficult.

The article focuses on the beneficial effects of diet, with the possible effects of not adjusting for country relegated to the Discussion section. A Web search on "Huijbregts" and "dietary patterns" reveals the cautions were lost when the message was transmitted to the general public.

Last modified: undefined.

The Mechanics of Categorical Variables With More Than Two Categories

Gerard E. Dallal, Ph.D.

Categorical variables with only two categories can be included in a multiple regression equation without introducing complications. As already noted, such a predictor specifies a regression surface composed of two parallel hyperplanes. The sign of the regression coefficients determines which plane lies above the other while the magnitude of the coefficient determines the distance between them.

When a categorical variable containing more than two categories is place in a regression model, the coding places specific contstraints on the estimated effects. This can be seen by generalizing the regression model for the t test to three groups. Consider the simple linear regression model

$$Y = b_0 + b_1 X$$

where X is a categorical predictor taking on the values 1,2,3, that is, X is either 1, 2, or 3, but the numbers represent categories, such as country, diet, drug, or type of fertilizer. The model gives the fitted values

- $Y = b_0 + b_1$ for the first category
- $Y = b_0 + 2\ b_1$ for the second category
- $Y = b_0 + 3\ b_1$ for the third category

The model forces a specific ordering on the predicted values. The predicted value for the second category must be exactly half-way between first and third category. However, category labels are usually chosen arbitrarily. There is no reason why the group with the middle code can't be the one with the largest or smallest mean value. If the goal is to decide whether the categories are different, a model that treats a categorical variable as though its numerical codes were really numbers is the wrong model.

One way to decide whether *g* categories are not all the same is to create a set of *g-1* indicator variables. Arbitrarily choose *g-1* categories and, for each category, define one of the indicator variables to be 1 if the observation is from that category and 0 otherwise. For example, suppose X takes on the values *A*, *B*, or *C*. Create the variables X_1 and X_2, where $X_1 = 1$ if the categorical variable is *A* and $X_2 = 1$ if the categorical variable is *B*, as in

X	X1	X2
A	1	0
B	0	1
A	1	0
C	0	0

```
  and so on...
```

The regression model is now

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

and the predicted values are

- Group A: $Y = b_0 + b_1\ 1 + b_2\ 0 = b_0 + b_1$
- Group B: $Y = b_0 + b_1\ 0 + b_2\ 1 = b_0 + b_2$
- Group C: $Y = b_0 + b_1\ 0 + b_2\ 0 = b_0$

The hypothesis of no differences between groups can be tested by applying the extra sum of squares principle to the set (X_1,X_2). This is what ANalysis Of VAriance (ANOVA) routines do automatically.

Last modified: undefined.

Interactions In Multiple Regression Models

Continuous Predictors

[This example involves a cross-sectional study of HDL cholesterol (HCHOL, the so-called good cholesterol) and body mass index (BMI), a measure of obesity. Since both BMI and HDL cholesterol will be related to total cholesterol (CHOL), it would make good sense to adjust for total cholesterol.]

In the multiple regression models we have been considering so far, the effects of the predictors have been **additive**. When HDL cholesterol is regressed on total cholesterol and BMI, the fitted model is

```
Dependent Variable: HCHOL

                       Parameter   Standard    T for H0:
      Variable  DF      Estimate     Error    Parameter=0    Prob > |
T|

      INTERCEPT  1     64.853        8.377        7.742        0.000
      BMI        1     -1.441        0.321       -4.488        0.000
      CHOL       1      0.068        0.027        2.498        0.014
```

says that the expected difference in HCHOL is 0.068 per unit difference in CHOL when BMI is held fixed. This is true whatever the value of BMI. The difference in HCHOL is -1.441 per unit difference in BMI when CHOL is held fixed. This is true whatever the value of CHOL. The effects of CHOL and BMI are additive because the expected difference in HDL cholesterol corresponding to differences in both CHOL and BMI is obtained by adding the differences expected from CHOL and BMI determined without regard to the other's value.

The model that was fitted to the data (HCHOL = b_0 + b_1 CHOL + b_2 BMI) *forces* the effects to be additive, that is, the effect of CHOL is the same for all values of BMI and vice-versa because the model won't let it be anything else. While this condition might seem restrictive, experience shows that it is a satisfactory description of many data sets. (I'd guess it depends on your area of application.)

Even if additivity is appropriate for many situations, there are times when it does not apply.

Sometimes, the purpose of a study is to formally test whether additivity holds. Perhaps the way HDL cholesterol varies with BMI depends on total cholesterol. One way to investigate this is by including an interaction term in the model. Let BMICHOL=BMI*CHOL, the product of BMI and CHOL. The model incorporating the interaction is

```
Dependent Variable: HCHOL

                       Parameter   Standard    T for H0:
      Variable  DF      Estimate      Error    Parameter=0    Prob > |
T|

      INTERCEPT  1    -24.990        38.234        -0.654        0.515
      BMI        1      2.459         1.651         1.489        0.139
      CHOL       1      0.498         0.181         2.753        0.007
      BMI*CHOL   1     -0.019         0.008        -2.406        0.018
```

The general form of the model is

$$Y = b_0 + b_1 X + b_2 Z + b_3 XZ$$

It can be rewritten two ways to show how the change in response with one variable depends on the other.

(1) $Y = b_0 + b_1 X + (b_2 + b_3 X) Z$
(2) $Y = b_0 + b_2 Z + (b_1 + b_3 Z) X$

Expression (1) shows the difference in Y per unit difference in Z when X is held fixed is $(b_2 + b_3X)$. This varies with the value of X. Expression (2) shows the difference in Y per unit difference in X when Z is held fixed is (b_1+b_3Z). This varies with the value of Z. The coefficient b_3 measures the amount by which the change in response with one predictor is affected by the other predictor. If b_3 is not statistically significant, then the data have not demonstrated the change in response with one predictor depends on the value of the other predictor. In the HCHOL, COL, BMI example, the model

$$\text{HCHOL} = -24.990 + 0.498\ \text{CHOL} + 2.459\ \text{BMI} - 0.019\ \text{CHOL} * \text{BMI}$$

can be rewritten

HCHOL = -24.990 + 0.498 CHOL + (2.459 - 0.019 CHOL) BMI or

HCHOL = -24.990 + 2.459 BMI + (0.498 - 0.019 BMI) CHOL

Comment: Great care must be exercised when interpreting the coefficients of individual variables in the presence of interactions. The coefficient of BMI is 2.459. In the absence of an interaction, this would be interpreted as saying that among those with a given total cholesterol level, those with greater BMIs are expected to have greater HDL levels! However, once the interaction is taken into account, the coefficient for BMI is, in fact, (2.459-0.019 CHOL), which is negative provided total cholesterol is greater than 129, which is true of all but 3 subjects.

Comment: The inclusion of interactions when the study was not specifically designed to assess them can make it difficult to estimate the other effects in the model. If

- a study was not specifically designed to assess interactions,
- there is no a priori reason to expect an interaction,
- interactions are being assessed "for insurance" because modern statistical software makes it easy, and
- no interaction is found,

it is best to refit the model without the interaction so other effects might be better assessed.

Indicator Predictor Variables

Interactions have a special interpretation when one of the predictors is a categorical variable with two categories. Consider an example in which the response Y is predicted from a continuous predictor X and indicator of sex (M0F1, =0 for males and 1 for females). The model

$$Y = b_0 + b_1\ X + b_2\ \mathrm{M0F1}$$

specifies two simple linear regression equations. For men, M0F1=0 and

$$Y = b_0 + b_1\ X$$

while, for women, M0F1=1 and

$$Y = (b_0 + b_2) + b_1\ X$$

The change in Y per unit change in X--b_1--is the same for men and women. The model forces the regression lines to be parallel. The difference between men and women is the same for all values of X and is equal to b_2, the difference in Y-intercepts.

Including a sex-by-X interaction term in the model allows the regression lines for men and women to have different slopes.

$Y = b_0 + b_1 X + b_2 \text{ M0F1} + b_3 X * \text{M0F1}$

For men, the model reduces to $Y = b_0 + b_1 X$
while for women, it is $Y = (b_0 + b_2) + (b_1 + b_3) X$

Thus, b_3 is the difference in slopes. The slopes for men and women will have been shown to differ if and only if b_3 is statistically significant.

The individual regression equations for men and women obtained from the multiple regression equation with a sex-by-X interaction are identical to the equations that are obtained by fitting a simple linear regression of Y on X for men and women separately. The advantage of the multiple regression approach is that it simplifies the task of testing whether the regression coefficients for X differ between men and women.

Comment: A common mistake is to compare two groups by fitting separate regression models and declaring them different if the regression coefficient is statistically significant in one group and not the other. it may be the two regression coefficients are similar with P values close to and on either side of 0.05. In order to show men and women response differently to a change in the continuous predictor, the multiple regression approach must be used and the difference in regression coefficients as measured by the sex-by-X interaction must be tested formally.

Centering

Centering refers to the practice of subtracting a constant from predictors before fitting a regression model. Often the constant is a mean, but it can be any value.

There are two reasons to center. One is technical. The numerical routines that fit the model are often more accurate when variables are centered. Some computer programs automatically center variables and transform the model back to the original variables, all without the user's knowledge.

The second reason is practical. The coefficients from a centered model are often easier to interpret. Consider the model that predicts HDL cholesterol from BMI and total cholesterol and a centered version fitted by subtracting 22.5 from each BMI and 215 from each total cholesterol.

Original: HCHOL = -24.990 + 0.498 CHOL + 2.459 BMI - 0.019 CHOL * BMI
Centered: HCHOL = 47.555 + 0.080 (CHOL-215) - 1.537 (BMI-22.5) - 0.019 (CHOL-215) (BMI-22.5)

In the original model

- -24.990 is the expected HDL cholesterol level for someone with total cholesterol and BMI of 0,
- 0.498 is the difference in HDL cholesterol corresponding to a unit difference in total cholesterol for someone with a BMI of 0, and
- 2.459 is the difference in HDL cholesterol corresponding to a unit difference in BMI for someone with a total cholesterol of 0.

Not exactly the most useful values. In the centered model, however,

- 47.555 is the expected HDL cholesterol level for someone with total a cholesterol of 215 and a BMI of 22.5,
- 0.080 is the difference in HDL cholesterol corresponding to a unit difference in total cholesterol for someone with a BMI of 22.5, and
- -1.537 is the difference in HDL cholesterol corresponding to a unit difference in BMI for someone with a total cholesterol of 215.

When there is an interaction in the model,

- the coefficients for the individual *uncentered* variables are the differences in response corresponding to a unit change in the predictor when the other predictors are 0, while
- the coefficients for the individual *centered* variables are the differences in response corresponding to a unit change in the predictor when the other predictors are at their centered values.

Last modified: undefined.

Collinearity

Gerard E. Dallal, Ph.D.

Prolog: Part 1

This message was posted to the Usenet group comp.soft-sys.stat.systat:

> I have run a multiple linear regression model with about 20 independent variables regressed against a dependent variable. I am getting an output I have never seen. In the coefficients, it gives me values for 5 independent variables but all the t-stats are blank and the standard errors are all zeros. My F and SEE are also blank. Also, it excluded 15 of the independent variables. Some of the excluded variables would not surprise me to be insignificant, but many I know are significant.
>
> The only note it gives me is tolerance = 0 limits reached. Can anyone give me some guidance on this output?

Prolog: Part 2

A predictor can't appear in a regression equation more than once. Suppose some response (Y) is regressed on height in inches (HIN) and the resulting equation is

$$Y = 17.38 + 5.08 * \text{HIN} .$$

Now suppose we attempt to fit an equation in which HIN appears twice as a predictor. To do this, let HINCOPY be an exact copy of HIN, that is, HINCOPY=HIN and fit the equation

$$Y = b_0 + b_1 \text{ HIN} + b_2 \text{ HINCOPY} .$$

What is a self-respecting computer program to do? It's supposed to come up with the best solution, but there are many equivalent solutions. All equations for which b0 = 17.38 and b1 +b2=5.08 are equivalent. So, a self-respecting computer program might do you a favor by recognizing the problem, excluding either HIN or HINCOPY, and continuing to fit the model.

Collinearity

The problem described in the Prolog is collinearity, where variables are so highly correlated that it is impossible to come up with reliable estimates of their individual regression coefficients. Collinearity does not affect the ability of a regression equation to predict the response. It poses a real problem if the purpose of the study is to estimate the contributions of individual predictors.

The two variables don't have to be exact copies for problems to arise. If Y is regressed on height in centimeters (HCM), the resulting equation must be

$$Y = 17.38 + 2.00 * \text{HCM} .$$

Otherwise, the the two equations would not give the same predictions. Since 1 inch = 2.54 centimeters,

2 (height in cm) is the same as 5.08 (height in inches).

[Those with a science background might wonder how this works out in terms of "units of measuremnt". This is discussed on its own web page in order to keep the discussion of collinearity flowing smoothly.]

Suppose Y is regressed on both HIN and HCM. What are the resulting coefficients in the regression equation

$$Y = b_0 + b_1 \text{ HIN} + b_2 \text{ HCM} ?$$

Again, there is no unique answer. There are many sets of coefficients that give the same predicted values. Any b_1 and b_2 for which $b_1 + 2.54\ b_2 = 5.08$ is a possibility. Some examples are

- $Y = 17.38 + 5.08 \text{ HIN} + 0.00 \text{ HCM}$
- $Y = 17.38 + 2.54 \text{ HIN} + 1.00 \text{ HCM}$
- $Y = 17.38 + 0.00 \text{ HIN} + 2.00 \text{ HCM}$
- $Y = 17.38 + 6.35 \text{ HIN} - 0.50 \text{ HCM}$

Collinearity (or *multicollinearity* or *ill-conditioning*) occurs when independent variables are so highly correlated that it becomes difficult or impossible to distinguish their individual influences on the response variable. As focus shifted from detecting exact linear relations among variables to detecting situations where things are so close that they cannot be estimated reliably, the meaning of *collinear* in a regression context was altered (some would say "devalued") to the point where it is sometimes used as a synonym for *correlated*, that is, correlated predictors are sometimes called *collinear* even when there isn't an exact linear relation among them.

Strictly speaking, "collinear" means just that--an exact linear relationship between variables. For example, if HIN is height in inches and HCM is height in centimeters, they are collinear because HCM = 2.54 HIN. If TOTAL is total daily caloric intake, and CARB, PROTEIN, FAT, and ALCOHOL are calories from TOTAL = CARB + PROTEIN + FAT + ALCOHOL.

[I prefer to write these linear relations as

$$\text{HCM} - 2.54\ \text{HIN} = 0 \text{ and}$$
$$\text{TOTAL} - \text{CARB} - \text{PROTEIN} - \text{FAT} - \text{ALCOHOL} = 0$$

in keeping with the general form of a linear relation

$$c_1 X_1 + ... + c_m X_m = k ,$$

where $c_1,...,c_m$, and k are constants.

This makes it easier to see that things like percent of calories from carbohydrates, protein, and fat are collinear, because

$$\%\text{CARB} + \%\text{PROTEIN} + \%\text{FAT} + \%\text{ALCOHOL} = 100 ,$$

with $c_1=c_2=c_3=c_4=1$ and k=100.]

Exact linear relationships might not appear exactly linear to a computer, while some relationships that were not collinear appeared to be collinear. This happens because computers store data to between 7 and 15 digits of precision. Roundoff error might mask some exact linear relationships and conceivably make other relationships look like they were collinear. This is reflected in the behavior of inexpensive calculators. When 1 is divided by 3 and the result is multiplied 3, the result is 0.9999999 rather than 1, so that 1 is not equal to the result of dividing 1 by 3 and multiplying it by 3!

For numerical analysts, the problem of collinearity had to do with identifying sets of predictors that were collinear or *appeared to be collinear*. Once "appear to be collinear" was part of the mix, "collinear" began to be used more and more liberally.

There are three different situations where the term "collinearity" is used:

1. where there is an exact linear relationship among the predictors by definition, as in percent of calories from fat, carbohydrate, protein, and alcohol,

2. where an exact or nearly exact linear relationship is forced on the data by the study design (Before the recent focus on vitamin E, supplementary vitamin E and A were almost always obtained through multi- vitamins. While the strength of the multi-vitamins varied among brands, A & E almost always appeared in the same proportion. This forced a linear relationship on the two vitamins and made it impossible to distinguish between their effects in observational studies.), and

3. where correlation among the predictors is *serious enough to matter*, in ways to be defined shortly.

In cases (1) and (2), any competent regression program will not allow all of the predictors to appear in the regression equation. Prolog 1 is the classic manifestation of the effects of collinearity in practice. In case (3), a model may be fitted, but there will be clear indications that something is wrong. If these indicators are present, it is appropriate to say there is a problem with collinearity. Otherwise, there is merely correlation among the predictors. While some authors equate collinearity with any correlation, I do not.

Serious correlations among predictors will have the following effects:

- Regression coefficients will change dramatically according to whether other variables are included or excluded from the model.
- The standard errors of the regression coefficients will be large.
- In the worst cases, regression coefficients for collinear variables will be large in magnitude with signs that seem to be assigned at random.
- Predictors with known, strong relationships to the response will not have their regression coefficients achieve statistical significance.

If variables are perfectly collinear, the coefficient of determination R^2 will be 1 when any one of them is regressed upon the others. This is the motivation behind calculating a variable's **tolerance**, a measure of collinearity reported by most linear regression programs. Each predictor is regressed on the other predictors. Its tolerance is $1-R^2$. A small value of the tolerance indicates that the variable under consideration is almost a perfect linear combination of the independent variables already in the equation and that it should not be added to the regression equation. All variables involved in the linear relationship will have a small tolerance. Some statisticians suggest that a tolerance less than 0.1 deserves attention. If the goal of a study is to determine whether a particular independent variable has predictive capability in the presence of the others, the tolerance can be disregarded if the predictor reaches statistical significance despite being correlated with the other predictors. The confidence interval for the regression coefficient will be wider than if the predictors were uncorrelated, but the predictive capability will have been demonstrated nonetheless. If the low value of tolerance is accompanied by large standard errors and nonsignificance, another study may be necessary to sort things out if subject matter knowledge cannot be used to eliminate from the regression equation some of the variables involved in the linear relation.

The tolerance is sometimes reexpressed as the Variance Inflation Factor (VIF), the inverse of the tolerance (= 1/tolerance). Tolerances of 0.10 or less become VIFs of 10 or more.

Other measures of collinearity, such as condition numbers, have been appeared in the statistical literature and are available in full-featured statistical packages. They have their advantages. When many variables have low tolerances, there is no way to tell how many nearly linear relations there are among the predictors. The condition numbers tell the analyst the number of relations and the associated matrices identify the variables in each one. For routine use, however, the tolerance or VIF is sufficient to determine whether any problems exist.

Some statisticians have proposed techniques--including ridge regression, robust regression, and principal components regression--to fit a multiple linear regression equation despite serious collinearity. I'm uncomfortable with all of them because they are purely mathematical approaches to solving things.

Principal components regression, replaces the original predictor variables with uncorrelated linear combinations of them. (It might help to think of these linear combinations as scales. One might be the sum of the first three predictors, another might be the difference between the second and fourth, and so on.) The scales are constructed to be uncorrelated with each other. If collinearity among the predictors was not an issue, there would be as many scales as predictors. When collinearity is an issue, there are only as many scales as there are nearly noncollinear variables. To illustrate, suppose X_1 and X_2 are correlated but not collinear. The two principal components might be their sum and difference ($X_1 + X_2$ and $X_1 - X_2$). If X_1 and X_2 and nearly collinear, only one principal component ($X_1 + X_2$) would be used in a principal component regression. While the mathematics is elegant and the principal components will not be collinear, there is no guarantee that the best predictor of the response won't be the last principal ($X_1 - X_2$) that never gets used.

When all is said and done, collinearity has been masked rather than removed. Our ability to estimate the effects of individual predictors is still compromised.

Last modified: undefined.

Centering

Gerard E. Dallal, Ph.D.

[Early draft subject to change.]

[Some of these notes must involve more mathematical notation than others. This is one of them. However, the mathematics is nothing more than simple algebra.]

This note was prompted by a student's question about interactions. She noticed that many of the tolerances became low when she had two predictors in a multiple regression equation along with their interaction. She wondered what these low tolerances had to do with the collinearity low tolerances usually signaled.

There are two reasons why tolerances can be small. The first is true collinearity, that is, a linear relation among predictors. The second is high correlated predictors that raise concerns about computational accuracy and whether individual coefficients and estimates of their contributions are numerically stable. In some cases where there is high correlation without a linear relation among the variables, the collinearity is avoidable and can be removed by **centering**, which transforms variables by subtracting a variable's mean (or other typical value) from all of the observations.

Before we go further, let's look at some data where we'll consider the regression of Y on X and X^2 and there can be no linear relation among the predictors.

Y	**X**	**X^2**	**Z** (=X-3)	**Z^2** (=$(X-3)^2$)
18	5	25	2	4
15	4	16	1	1
12	3	9	0	0
3	2	4	-1	1
9	1	1	-2	4

In this example, Z is the centered version of X, that is, Z=X-3, where 3 is the mean of the Xs. We'll be referring to 6 different regressions

1. $Y = b_0 + b_1 X$

2. $Y = c_0 + c_1 Z$
3. $Y = b_0 + b_2 X^2$
4. $Y = c_0 + c_2 Z^2$
5. $Y = b_0 + b_1 X + b_2 X^2$
6. $Y = c_0 + c_1 Z + c_2 Z^2$

They should be thought of as three pairs of equations. A particular coefficient does not have to have the same value in all equations in which it appears. That is, there is not just one b_0, but different b_0s for equations (1), (3), and (5). If this is proves to be confusing, I'll rewrite the note.

So that computer output will not clutter this note, I've placed it in a separate web page.

Things to notice:

- The correlation between X & X^2 is 0.98 while the correlation between Z & Z^2 is 0.00.
- Equations (1) & (2): The regression of Y on X is virtually identical to the regression of Y on Z. R^2 is the same for both equations (0.676). The coefficients for X and Z are the same (-3.00) as are their P values (0.088). This must happen any time Z = X - k, where k is any constant (here, k is 3) because

  $$Y = c_0 + c_1 Z$$
  $$Y = c_0 + c_1 (X - k)$$
  $$Y = (c_0 - c_1 k) + c_1 X$$

  Thus, regressing Y on Z is equivalent to regressing Y on X with

  $$b_1 = c_1$$
  $$b_0 = c_0 - c_1 k$$

  This is why the slopes of the two equations are the same, but their intercepts differ.

  Another way to see why the equations must be so similar is to recognize that because Z is X shifted by a constant, the correlation between Y and Z will be equal to the correlation between Y and X. Further, the SDs of X and Z will be equal and the SD of Y will be common to both equations. Thus, the three quantities that determine the regression coefficient--the SD of the response, the SD of the predictor, and the correlation between response and predictor--are the same for both equations!

- **Equations (3) & (4):** On the other hand, the regression of Y on X^2 is different from the regression of Y on Z^2. There is no reason why they must be the same. Z_2 is not X_2 shifted by a constant. Z_2 is not even a linear function of X_2. Regressing Y on Z_2 is not equivalent to regressing Y on X^2.

$$Y = c_0 + c_2\, Z^2$$
$$Y = c_0 + c_2\, (X - k)^2$$
$$Y = (c_0 + c_2\, k^2) - 2\, c_2\, k\, X + c_2X^2$$

  which includes a term involving X.
- **Equations (5) & (6):** In many ways, the multiple regression of Y on X and X^2 is similar to the regression of Y on Z and Z^2. R^2 is the same for both equations (0.753). The coefficients for X^2 and Z^2 (0.857) along with their P values (0.512).

  The agreement is close because the two regressions are equivalent.

$$Y = c_0 + c_1\, Z + c_2\, Z^2$$
$$Y = c_0 + c_1\, (X - k) + c_2\, (X - k)^2$$
$$Y = (c_0 - c_1\, k + c_2\, k^2) + (c_1 - 2\, c_2\, k)\, X + c_2\, X^2$$

  Thus, the regression of Y on Z and Z^2 is equivalent to the regression Y on X and X^2 with

$$b_2 = c_2$$
$$b_1 = c_1 - 2\, c_2\, k$$
$$b_0 = c_0 - c_1\, k + c_2\, k^2$$

One question remains: **In the regressions of Y on Z & Z^2 and Y on X & X^2, why are the P values for the coefficients of Z^2 and X^2 the same while the P values for Z and X differ?** The answer is supplied by the description of the P value as an indicator of the extent to which the variable adds predictive capability to the other variables in the model. We've already noted that the regression of Y on Z has the same predictive capability (R^2) as the regression of Y on X and the regression of Y on Z and Z^2 has the same predictive capability as the regression of Y on X and X^2. Therefore, adding Z^2 to Z has the same effect as adding X^2 to X. We start from the same place (X and Z) and end at the same place (Z,Z^2 and X,X^2), so the way we get there (Z^2 and X^2) must be the same.

We've also noted that the regression of Y on Z^2 does not have the same predictive capability (R^2) as the regression of Y on X^2. Since we start from different places (Z^2 and X^2) and end at the same place (Z,Z^2 and X,X^2), the way we get there (X and Z) must be different.

Interactions behave in a similar fashion. Consider predicting Y from X, Z, and their interaction XZ and predicting Y from $(X-k_x)$, $(Z-k_z)$ and their interaction $(X-k_x)$ $(Z-k_z)$, where k_x and k_z are constants that center X and Z.

- Regressing Y on $(X-k_x)$ & $(Z-k_z)$ is equivalent to regressing Y on X & Z because
  $Y = c_0 + c_1 (X - k_x) + c_2 (Z - k_z)$
  $Y = (c_0 - c_1 k_x - c_2 k_z) + c_1 X + c_2 Z$
- Regressing Y on $(X-k_x)$ & $(X-k_x)$ $(Z-k_z)$ is different from regressing Y on X & XZ because
  $Y = c_0 + c_1 (X - k_x) + c_3 (X - k_x) (Z - k_z)$
  $Y = (c_0 - c_1 k_x + c_3 k_x k_z) + (c_1 - c_3 k_z) X - c_3 k_x Z + c_3 X Z$
  which includes a term involving Z alone.
- Regressing Y on $(X-k_x)$, $(Z-k_z)$, and $(X-k_x)$ $(Z-k_z)$ is equivalent to regressing Y on X, Z, and XZ because
  $Y = c_0 + c_1 (X - k_x) + c_2 (Z - k_z) + c_3 (X - k_x) (Z - k_z)$
  $Y = (c_0 - c_1 k_x - c_2 k_z + c_3 k_x k_z) + (c_1 - c_3 k_z) X + (c_2 - c_3 k_x) Z + c_3 XZ$

Adding XZ to X and Z will have the same effect as adding $(X-k_x)(Z-k_z)$ to $(X-k_x)$ and $(Z-k_z)$ because the models start from the same place and end at the same place, so the P values for XZ and $(X-k_x)(Z-k_z)$ will be the same. However, adding Z to X and XZ is different from adding $(Z-k_z)$ to $(X-k_x)$ and $(X-k_x)(Z-k_z)$ because the models start from different places and end up at the same place. (In similar fashion, adding adding X to Z and XZ is different from adding $(X-k_x)$ to $(Z-k_z)$ and $(X-k_x)(Z-k_z)$.)

If there is a linear relation among the variables, centering will not remove it. If

$$c_1X_1+c_2X_2+..+c_pX_p = m$$

and X_1 is replaced by $(Z=X_1-k)$, then

$$c_1Z+c_1k +c_2X_2+..+c_pX_p = m$$

or

$$c_1Z+c_2X_2+..+c_pX_p = m - c_1k$$

Because m - c_1k is a(nother) constant, there is still a linear relation among the variables.

Comment: While I might worry about centering when fitting polynomial regressions (if I didn't use software specially designed for the purpose), I tend not to worry about it when fitting interactions. There has been more than a quarter century of research into the problems of numerical accuracy when fitting multiple regression equations. Most statistical software, including all of the software I use personally, makes use of this work and is fairly robust. In addition, I rarely fit anything more complicated than a first-order interaction, which won't grow any faster than a square. If the software shows a signifcant or important interaction, I tend to believe regardless of any collinearity measure because the effect of collinearity is to mask things. I would look more closely if collinearity measures were suspicious and an expected effect were nonsignifcant.

Comment: Centering can make regression coefficients easier to understand. Consider an equation that predicts Systolic Blood Pressure (SBP) from AGE and Physical Activity Level (PAL, which ranges from 1.4 to 2.0 in a typical healthy adult population). An AGE by PAL interaction is included in the model because it is felt that age will have more of an effect at low Physical Activity Levels than at high levels. The resulting equation is

$$\text{SBP} = 78.6 + 2.6\ \text{AGE} + 14\ \text{PAL} - 1.0\ \text{AGE*PAL}$$

Those unaccustomed to dealing with interactions will be surprised to see that the coefficient of PAL is positive. This seems to suggest that exercise raises blood pressure! However, when the data are centered by subtracting the mean age, 34, from AGE and the mean PAL, 1.6, from PAL, the equation becomes

$$\text{SBP} = 135 + 1.0\ (\text{AGE-34}) - 20\ (\text{PAL-1.6}) - 1.0\ (\text{AGE-34})*(\text{PAL-1.6})$$

The two equations are the same, that is, they give the same predicted values and simple algebra can be used to transform one into the oter.. Now, however, the coefficient of PAL is negative and the coefficient of age is less substantial. It's the interaction that's causing all the changes. If there were no interaction, the coefficent of PAL would be the change in SBP with each unit change in PAL. With the interaction in the model, this interpretation is correct only when the interaction term is 0. But that can happen only when age is 0, which is not true for anyone in this population.

To see this another way, rewrite the original equation as

SBP = 78.6 + (14 - 1.0 AGE) PAL + 2.6 AGE

- The change of SBP per unit change in PAL is (14 - 1.0 AGE). This is 14 when age is 0 but is -20 when age is equal to the more typical value of 34.
- As age increases, the effect of PAL (its coefficient) becomes greater.
- For the ages in the sample (20-50), the coefficent of PAL ranges from -6 to -36. Since PAL takes on values between 1.4 and 2.0, the full range of PAL (1.4 to 2.0) accounts for a difference in SBP of 4 mm at the low end of age (20) and 22 mm at the high end of 50.

When the data are centered, the coefficient for AGE is the change in SBP per unit change in age when PAL is equal to its mean value, 1.6. The coefficient for PAL is the change in SBP per unit change in PAL when age is equal to its mean value, 34. In general, when data are centered, the coefficients for each individual variable are the changes in response per unit change in predictor when all other predictors are equal to their sample means. This is usually more informative to the reader than the change in response per unit change in predictor when all other predictors are equal to 0.

[back to LHSP]

Last modified: undefined.

Regression Diagnostics

Gerard E. Dallal, Ph.D.

In the 1970s and 80s, many statisticians developed techniques for assessing multiple regression models. One of the most influential books on the topic was *Regression Diagnostics: Identifyin Influential Data and Sources of Collinearity* by Belsley, Kuh, and Welch. Roy Welch tells of getting interested in regression diagnostics when he was once asked to fit models to some banking data. When he presented his results to his clients, they remarked that the model could not be right because the sign of one of the predictors was different from what they expected. When Welch looked closely at the data, he discovered the sign reversal was due to an outlier in the data. This example motivated him to develop methods to insure it didn't happen again!

Perhaps the best reason for studying regression diagnostics was given by Frank Anscombe when he was discussing outliers.

> We are usually happier about asserting a regression relation if the relation is appropriate after a few observations (any ones) have been deleted--that is, we are happier if the regression relation seems to permeate all the observations and does not derive largely from one or two.

Regression diagnostics were developed to measure various ways in which a regression relation might derive largely from one or two observations. Observations whose inclusion or exclusion result in substnatial changes in the fitted model (coefficients, fitted values) are said to be **influential**. Many of these diagnostics are available from standard statistical program packages.

Scatterplots

Multiple regression models have three primary characteristics: linearity, homogeneity of variance, and normally distributed residuals. Serious departures can be detected by scatterplots of the response against each predictor, residual plots (residuals against predicted values) and normal plots of the residuals.

Detecting Outliers

It is common practice to distinguish between two types of outliers. Outliers in the response variable represent model failure. Such obeservations are called **outliers**. Outliers with respect to the predictors are called **leverage points**. They can affect the regression model, too. Their response variables need not be outliers. However, they may almost uniquely determine regression coefficients. They may also cause the standard errors of regression coefficients to be much smaller than they would be if the obeservation were excluded.

The ordinary or simple residuals (observed - predicted values) are the most commonly used measures for detecting outliers. The ordinary residuals sum to zero but do not have the same standard deviation. Many other measures have been offered to improve on or complement simple residuals. **Standardized Residuals** are the residuals divided by the estimates of their standard errors. They have mean 0 and standard deviation 1. There are two common ways to caculate the standardized residual for the i-th observation. One uses the residual mean square error from the model fitted to the full dataset (internally studentized residuals). The other uses the residual mean square error from the model fitted to the all of the data except the i-th observation (externally studentized residuals). The externally standardized residuals follow a t distribution with n-p-2 df. They can be thought of as testing the hypothesis that the corresponding observation does not follow the regression model that describes the other observations.

In practice, I find ordinary residuals the most useful. While the standard deviations of the residuals are different, they are usually not different enough to matter when looking for outliers. They have the advantage of being in the same scale as the response.

Detecting Influential Observations

- **Cook's Distance** for the i-th observation is based on the differences between the predicted responses from the model constructed from all of the data and the predicted responses from the model constructed by setting the i-th observation aside. For each observation, the sum of squared residuals is divided by (p+1) times the Residual Mean Square from the full model. Some analysts suggest investigating observations for which Cook's distance is greater than 1. Others suggest looking at a dot plot to find extreme values.
- **DFITS$_i$** is the scaled difference between the predicted responses from the model constructed from all of the data and the predicted responses from the model constructed by setting the i-th observation aside. It is similar to Cook's distance. Unlike Cook's distance, it does not look at all of the predicted values with the i-th observation set aside. It looks only at the predicted values for the i- th observation. Also, the scaling factor uses the standard error of the estimate with the i-th observation set aside. To see the effect of this, consider a dataset with one predictor in which all of the observations lie exactly on a straight line. The Residual Mean Square using all of the data will be positive. The standard errors of the estimate obained by setting one observation aside in turn will be positive except for the observation that does not lie on the line. When it is set aside, the standard error of the estimate will be 0 and DFITS$_i$ will be arbitrarily large. Some analysts suggest investigating observations for which $|\text{DFITS}_i|$ is greater than $2 \sqrt{[(p+1)/(n-p-1)]}$. Others suggest looking at a dot plot to find extreme values.
- **DFBETAS$_\mathbf{i}$** are similar to DFITS. Instead of looking at the difference in fitted value when the i-th observation is included or exlcuded, DFBETAS looks at the change in each regression coefficient.

In theory, these can be useful measures. However, I have not found that to be the case in my own practice. It may be the sort of data I analyze. Often, I see people using these measure finding themselves in a vicious cycle. They calculate some measures, remove some observations, and find additional observations have suspicious measures when they recalculate. They remove more observations and the cycle starts all over again. By the time they are done, many observations are set aside, no one is quite sure why, and no one feels very good about the final model.

Leverage points do not necessarily correspond to outliers. There are a few reasons why this is so. First, an observation with sufficiently high leverage might exert enough influence to drag the regression equation close to its response and mask the fact that it might otherwise be an outlier. See the third and fourth Anscombe datasets, for example. When they do not, it's not clear what can be done about the leverage point except, perhaps, to note it. The fourth Anscombe example is as extreme as it gets. The regression coefficient is completely determined by a single obsrvation. Yet, what is one to do? If one believes the model (linearity, homoscedasticity, normal errors), then a regression coefficient determined by one or two observations is the best we can do if that's the way our data come to us or we choose to collect it.

Robust Regression

A least squares model can be distorted by a single observation. The fitted line or surface might be tipped so that it no longer passes through the bulk of the data in order to intricued many small or moderate errors in order to reduce the effect of a very large error. For example, if a large error is reduced from 200 to 50, its square is reducted from 40,000 to 2,500. Increasing an error from 5 to 15 increases its square from 25 to 225. Thus, a least squares fit might introduce many small errors in order to reduce a large one.

Robust regression is a term used to desctribe model fitting procedures that are insensitive to the effects of maverick observations. My personal favorite is least median of squares (LMS) regression, developed by Peter Rousseeuw. LMS regression minimizes the median squared resduals. Since it focuses on the median residual, up to half of the observations can disagree without masking a model that fits the rest of the data.

Fitting an LMS regression model poses some difficulties. The first is computational. Unlike least squares regression, there is no formula that can be used to calculate the coefficents for an LMS regression. Random samples of size *p+1*, are drawn. A regression surface is fitted to each set of observations and the median squared residual is calculated. The model that had the smallest median squared residual is used.

The LMS solution can be found by fitting regression surfaces to all possible subsets of p+1 points, where *p* is the number of predictors . (This is merely a matter of solving set of p+1 linear equations with p+1 unknown parameters.) The LMS regression is given by the parameters, chosen over all possible sets of p+1 observations, that have the minimum median squared residual when applied to the entire data set. Evaluating all possible subsets of p observations can be computationally infeasible for large data sets. When n is large, Rousseeeuw recommends taking random samples of observations and using the best solution obtained from these randomly selected subsets. The second problem is that there is no theory for constructing confidence intervals for LMS regression coefficients or for testing hypotheses about them. Rousseeuw has proposed calculating a distance measure based on LMS regression and using it to identify outliers with respect to the LMS regression. These observations are set aside and least squares regression is fitted to the rest of the data. The result is called reweighted least squares regression.

This approach has some obvious appeal. A method insensitive to maverick observations is used to identify outliers that are set aside so an ordinary multiple regression can be fitted. However, there are no constrains that force the reweighted least squares model to resemble the LMS model. It is even possible for the signs of some regression coefficient to be different in the two models. This places the analyst in the awkward position of explaining how a model different from the final model was used to determine which observations determine the final model.

The real drawback to using these procedures is the lack of readily available software. Rousseeuw distributes FORTRAN source code to fit LMS regression models and Dallal and Rousseeuw added a front-end in the style of a DOS SYSTAT module so that the program can be used to analyze data in SYSTAT system files. The method has yet to be added to any of the major commercial packages.

Last modified: undefined.

Single Factor Analysis of Variance

Gerard E. Dallal, Ph.D.

Terminology

A **factor** is a categorical predictor variable. Factors are composed of **levels**. For example, **treatment** is a factor with the various types of treatments comprising the levels. The levels should be exclusive, that is, a subject should appear under only one level which in this case means given a single type of treatment.

While I've yet to see it stated explicitly in any textbook, it is important to be aware of two different types of factors--those where subjects are randomized to the levels and those where no randomization is involved. The same statistical methods are use for analyzing both types of factors, but the justification for the use of statistical methods differs, just as for intervention trials and observational studies. When subjects are randomized to levels, as in the case of treatments, the validity of the analysis follows the randomization. When subjects are not randomized to levels, as in the case of sex or smoking status, the validity of the analysis follows either from having random samples from each level or, more likely, from having used an enrollment procedure that is believed to treat all levels the same. For example, a door-to-door study of adults with and without children in primary school conducted in the early afternoon is likely to produce very different results from what would be obtained in the early evening.

_The terms **Single Factor Analysis of Variance**, **Single Factor ANOVA**, **One Way Analysis of Variance**, and **One Way ANOVA** are used interchangeably to describe the situation where a contiuouse response is being described in terms of a single categorical variable or factor composed of two or more categories. It is a generalization of Student's t test for independent samples to situations with more that two groups.

I have sometimes been guilty of putting a hyphen in *single-factor analysis of variance*. This was prompted by referee's report on a colleague's paper in which the reviewer had confused the analysis of variance with another statistical technique, factor analysis. The reviewer wanted to know how factor analysis could be performed with a single factor!

Notation

[Do I want to do this? I'm not sure. For years I've tried to avoid it, but I'm afraid too much gets lost when there's not a good way to notate it. The trick is to use it sparingly to get essential points across without making it so complex that it's difficult to follow. So, here it is. Time will tell if it stays.]

Let there be *g* groups. Let y_{ij} be the value for the j-th subject in the i-th group, where i=1,..,g and j=1,.., n_i. That is, the number of subjects in group *i* is n_i. Let N = Σn_i.

Means are denoted by putting a dot in place of the subscripts over which the means are calculated. The mean for the i-th group is denoted $\bar{y}_{i.} = \left(\sum_{j=1}^{n_i} y_{ij}\right)\Big/n_i$ and the overall mean is denoted $\bar{y}_{..} = \left(\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij}\right)\Big/\sum_{i=1}^{g} n_i = \left(\sum_{i=1}^{g} n_i\bar{y}_{i.}\right)\Big/\sum_{i=1}^{g} n_i$.

The Model

The model for one way ANOVA can be written simply as

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where Y_{ij} is the response for the j-th subject in the i-th group, μ_i is the mean of the i-th group, and ε_{ij} is a random error associated with the j-th subject in the i-th group. The model usually specifies the errors to be independent, normally distributed, and with constant variance.

While this model is fine for one way ANOVA, it is usually written in a different way that generalizes more easily when there is more than one factor in the model.

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where Y_{ij} is the response for the j-th subject in the i-th group, μ is an overall effect and α_i is the effect of the i-th group. One problem with this model is that there are more parameters than groups. Some constraint must be placed on the parameters so they can be estimated. This is easily seen with just two groups. The predicted value for group 1 is $\mu + \alpha_1$ while for group 2 it is $\mu + \alpha_2$. Three parameters, μ, α_1, and α_2 are being used to model two values, so there are many ways the parameters can be chosen.

The interpretation of the model's parameters depends on the constraint that is placed upon the them. Let there be *g* groups. If α_g is set to 0 as many software packages do, then μ estimates the mean of group *g* and α_i estimates the mean difference between groups *i* and *g*.

The so-called *usual constraint* has the parameters sum to 0, that is, $\Sigma\alpha_i = 0$. In the case of two groups, $\alpha_1 = -\alpha_2$. In this case, μ is the simple mean of the group means, that is, $\left(\sum \bar{y}_{i.}\right)\Big/g$. The constraint $\Sigma n_i\alpha_i = 0$ is worth noting because μ then estimates the overall mean, $\bar{y}_{..}$, while μ estimates the difference between the mean of the i-th group and the overall mean.

The simple mean of the group means, $\left(\sum \bar{y}_{i.}\right)\Big/g$, looks on when first encounterd but is often more useful than the overall mean. Suppose in order to do background work on a proposed exercise study we take a random cross-section of people who exercise. We classify them according to their form of exercise and measure their blood pressure. The overall mean estimates the mean blood pressure in the population of exercisers. However, there may be many more joggers than anything else and relatively few weightlifters. The overall mean would then be weighted toward the effect of jogging. On the other hand, the mean of the joggers--no matter how few or many--is our best estimate of the mean blood pressure in the population of joggers. The mean of the weight lifters--no matter how few or many--is our best estimate of the mean blood pressure in the population of weight lifters, and similarly for all of the other forms of exercise. The simple mean of the group means represents the mean of the different types of exercise and the αs estimates the difference between the i-th form of exercise and this mean. This seems like a more compelling measure of the effect of a particular form of exercise.

Still, after all the notation has been introduced and formulas have been written, single factor analysis of variance is nothing more than a generalization of Student's t test for independent samples to allow for more than two groups. The new wrinkles involve the issue of multiple comparions and multiple testing made possible by having more than two groups to compare. Two immediate questions are (1) how do we decide whether there are any differences among the groups, that is, how do we test the hypothesis (stated in three equivalent forms)

- H_0: all population means are equal
- H_0: $\mu_1 = .. = \mu_g$
- H_0: $\alpha_1 = .. = \alpha_g = 0$

and (2) if there are differences, how do we decide which groups are different?

[back to LHSP]

Last modified: undefined.

How to Read the Output From One Way Analysis of Variance

Here's a typical piece of output from a single-factor analysis of variance. The response is the two year change in bone density of the spine (final - initial) for postmenopausal women with low daily calcium intakes (≤400 mg) assigned at random to one of three treatments--placebo, calcium carbonate, calcium citrate maleate).

```
                       Class         Levels    Values
                       GROUP              3    CC CCM P

Dependent Variable: DBMD05
                                 Sum of
 Source                 DF      Squares    Mean Square   F Value    Pr
> F
 Model                   2   44.0070120     22.0035060      5.00
0.0090
 Error                  78  343.1110102
4.3988591
 Corrected Total        80
387.1180222


              R-Square     Coeff Var      Root MSE    DBMD05 Mean
              0.113679     -217.3832      2.097346      -0.964815


 Source                 DF     Type I SS   Mean Square   F Value    Pr
> F
 GROUP                   2   44.00701202   22.00350601      5.00
0.0090

 Source                 DF   Type III SS   Mean Square   F Value    Pr
> F
 GROUP                   2   44.00701202   22.00350601      5.00
0.0090


                                        Standard
Parameter             Estimate             Error    t Value    Pr > |
t|
Intercept         -1.520689655 B      0.38946732      -3.90
0.0002
GROUP     CC       0.075889655 B      0.57239773       0.13
0.8949
GROUP     CCM      1.597356322 B      0.56089705       2.85
0.0056
GROUP     P        0.000000000
B       .                ..         .

NOTE: The X'X matrix has been found to be singular, and a generalized
inverse
      was used to solve the normal equations.  Terms whose estimates
are
      followed by the letter 'B' are not uniquely estimable.

                               The GLM Procedure
                              Least Squares Means

                                      DBMD05      LSMEAN
                       GROUP          LSMEAN      Number
                       CC        -1.44480000           1
                       CCM        0.07666667           2
                       P         -1.52068966           3

                      Least Squares Means for effect GROUP
                      Pr > |t| for H0: LSMean(i)=LSMean(j)

                 i/j              1             2             3
                    1                      0.0107        0.8949
                    2        0.0107                      0.0056
                    3        0.8949        0.0056

NOTE: To ensure overall protection level, only probabilities
      associated with pre-planned comparisons should be used.


               Adjustment for Multiple Comparisons: Tukey-Kramer

                      Least Squares Means for effect GROUP
                      Pr > |t| for H0: LSMean(i)=LSMean(j)

                 i/j              1             2             3
                    1                        0.0286           0.9904
                    2          0.0286                         0.0154
                    3          0.9904        0.0154
```

The Analysis of Variance Table

The **Analysis of Variance** table is just like any other ANOVA table. The Total Sum of Squares is the uncertainty that would be present if one had to predict individual responses without any other information. The best one could do is predict each observation to be equal to the overall sample mean. The ANOVA table partitions this variability into two parts. One portion is accounted for (some say "explained by") the model. It's the reduction in uncertainty that occurs when the ANOVA model,

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

is fitted to the data. The remaining portion is the uncertainty that remains even after the model is used. The model is considered to be statistically significant if it can account for a large amount of variability in the response.

Model, Error, **Corrected Total**, **Sum of Squares**, **Degrees of Freedom**, **F Value**, and **Pr F** have the same meanings as for multiple regression. This is to be expected since analysis of variance is nothing more than the regression of the response on a set of indicators definded by the categorical predictor variable.

The degrees of freedom for the model is equal to one less than the number of categories. The F ratio is nothing more than the extra sum of squares principle applied to the full set of indicator variables defined by the categorical predictor variable. The F ratio and its P value are the same regardless of the particular set of indicators (the constraint placed on the α-s) that is used.

Sums of Squares: The total amount of variability in the response can be written $\sum_{i,j}(y_{ij} - \bar{y}_{..})^2$, the sum of the squared differences between each observation and the overall mean. If we were asked to make a prediction without any other information, the best we can do, in a certain sense, is the overall mean. The amount of variation in the data that can't be accounted for by this simple method of prediction is the Total Sum of Squares.

When the Analysis of Variance model is used for prediction, the best that can be done is to predict each observation to be equal to its group's mean. The amount of uncertainty that remains is sum of the squared differences between each observation and its group's mean, $\sum_{i,j}(y_{ij} - \bar{y}_{i.})^2$. This is the Error sum of squares. In this outpur it also appears as the GROUP sum of squares. The difference between the Total sum of squares and the Error sum of squares is the Model Sum of Squares, which happens to be equal to $\sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2$.

Each sum of squares has corresponding degrees of freedom (DF) associated with it. Total df is one less than the number of observations, N-1. The Model df is the one less than the number of levels The Error df is the difference between the Total df (N-1) and the Model df (g-1), that is, N-g. Another way to calculate the error degrees of freedom is by summing up the error degrees of freedom from each group, n_i-1, over all *g* groups.

The **Mean Squares** are the Sums of Squares divided by the corresponding degrees of freedom.

The **F Value** or **F ratio** is the test statistic used to decide whether the sample means are withing sampling variability of each other. That is, it tests the hypothesis H_0: μ_1...μ_g. This is the same thing as asking whether the model as a whole has statistically significant predictive capability in the regression framework. **F** is the ratio of the Model Mean Square to the Error Mean Square. Under the null hypothesis that the model has no predictive capability--that is, that all of thepopulation means are equal--the F statistic follows an F distribution with *p* numerator degrees of freedom and *n-p-1* denominator degrees of freedom. The null hypothesis is rejected if the F ratio is large. This statstic and P value might be ignored depending on the primary research question and whether a multiple comparisons procedure is used. (See the discussion of multiple comparison procedures.)

The **Root Mean Square Error** (also known as **the standard error of the estimate**) is the square root of the Residual Mean Square. It estimates the common within-group standard deviation.

Parameter Estimates

The parameter estimates from a single factor analysis of variance might best be ignored. Different statistical program packages fit different paraametrizations of the one-way ANOVA model to the data. SYSTAT, for example, uses the usual constraint where $\Sigma\alpha_i$=0. SAS, on the other hand, sets α_g to 0. Any version of the model can be used for prediction, but care must be taken with significance tests involving individual terms in the model to make sure they correspond to hypotheses of interest. In the SAS output above, the Intercept tests whether the mean bone density in the Placebo group is 0 (which is, after all, to be expected) while the coefficients for CC and CCM test whether those means are different from placebo. It is usually safer to test hypotheses directly by using the whatever facilities the software provides that by taking a chance on the proper interpretation of the model parametrization the software might have implemented. The possiblity of many different parametrizations is the subject of the warning that *Terms whose estimates are followed by the letter 'B' are not uniquely estimable.*

After the parameter estimates come two examples of multiple comparisons procedures, which are used to determine which groups are different given that they are not all the same. These methods are discussed in detail in the note on multiple comparison procedures. The two methods presented here are

Fisher's Least Significant Differences and *Tukey's Honestly Signficant Differences*. Fisher's Least Significant Differences is essentially all possible t tests. It differs only in that the estimate of the common within group standard deviation is obtained by pooling information from all of the levels of the factor and not just the two being compared at the moment. The values in the matrix of P values comparing groups 1&3 and 2&3 are identical to the values for the CC and CCM parameters in the model.

[back to LHSP]

Last modified: undefined.

Multiple Comparison Procedures

Gerard E. Dallal, PhD
Scientist I, JM USDA HNRC

[Much of this discussion involves tests of significance. Since most tests are performed at the 0.05 level, I will use 0.05 throughout rather than an abstract symbol such as α that might make some readers uncomfortable. Whenever you see "0.05 level", feel free to substitute your own favorite value, such as 0.01, or even a symbol such as α, if you'd like.]

At some point in a career that requires the use of statistical analysis, an investigator will be asked by a statistician or a referee to use a multiple comparison procedure to adjust for having performed many tests or for having constructed many confidence intervals. What, exactly, is the issue being raised, why is it important, and how is it best addressed? We'll start with significance tests and later draw some comparisons with confidence intervals.

Let's start with some "dumb" questions. The answers will be obvious. Yet, they're all one needs to know to understand the issue surrounding multiple comparisons.

When playing the lottery, would you rather have one ticket or many tickets? Many. Lottery numbers are a random phenomenon and having more tickets increases your chances of winning.

There's a severe electrical storm and you have to travel across a large, open field. Your main concern is about being hit by lightning, a somewhat random phenomenon. Would you rather make the trip once or many times? Once. The more trips you make, the more likely it is that you get hit by lightning.

Similar considerations apply to observing statistically significant test results. When there is no underlying effect or difference, we want to keep the chance of obtaining statistically significant results small. Otherwise, it would be difficult to claim that that our observed differences were anything more than the vagaries of sampling and measurement.

For better or worse, much of statistical analysis is driven by significance tests. The scientific community as a whole has decided that the vast majority of those tests will be carried out at the 0.05 level of significance. This level of significance is a value that separates results typically seen when a null hypothesis is true from those that are rare when the null hupothesis is true. The classic frequentist methods do not give a probability that a hypothesis is true or false. Instead, they provided indirect evidence. The rules of the game say that if results are typical of what happens when there is no effect, investigators can't claim evidence of an effect. However, if the observed results occur rarely when there is no effect, investigators may say there *is* evidence of an effect. The level of significance is the probability of those rare events that permit investigators to claim an effect. When we test at the 0.05 level of significance, the probability of observing one of these rare results when there is no effect is 5%.

In summary, a significance test is a is a way of deciding whether something rare has occurred if there is no effect. It may well be that there is no effect and something rare has occurred, but we cannot know that. By the rules of the game, we conclude that there is an effect and *not* that we've observed a "rare event".

When there's no underlying effect or difference, getting statistically significant results is supposed to be like winning the lottery or getting hit by lightning. The probability is supposed to be small (well, 5%, anyway). But just as with the lottery or with lightning--where the probability of winning or getting hit can increase dramatically if you take lots of chances--many tests, increase the chance that at something will be statistically significant at the nominal 5%. In the case of 4 independent tests each at the 0.05 level, the probability that one or more will achieve significance is about 19%. This violates the spirit of the significance test. The chance of a statistically significant result is suppose to be small when there's no underlying effect, but performing lots of tests makes it large.

If the chance of seeing a statistically significant result is large, why should we pay it any attention and why should a journal publish it as though it were small? Well, we shouldn't and they shouldn't. In order to insure that the statistically significant results we observe really are rare when there is no underlying effect, some adjustment is needed to keep the probability of getting *any* statistically significant results small when many tests are performed. This is the issue of multiple comparisons. The way we adjust for multiple tests will depend on the number and type of comparisons that are made. There are common situations that occur so often they merit special attention.

Comparing many groups

Consider an experiment to determine differences among three or more treatment groups (e.g., cholesterol levels resulting from diets rich in different types of of oil: olive, canola, rice bran, peanut). This is a generalization of Student's t test, which compares 2 groups.

How might we proceed? One way is to perform all possible t tests. But this raises the problem we discussed earlier. When there are 4 treatments, there are 6 comparisons and the chance that *some* comparison will be significant (that some pair of treatments will look different from each other) is much greater than 5% if they all have the same effect. (I'd guess it's around 15%.) If we notice a t statistic greater than 1.96 in magnitude, we'd like to say, "Hey, those two diets are different because, if they weren't, there's only a 5% chance of an observed difference this large." However, with that many tests (lottery tickets, trips in the storm) the chance of a significant result (a win, getting hit) is much larger, the t statistic is no longer what it appears to be, and the argument is no longer sound.

Statisticians have developed many "multiple comparison procedures" to let us proceed when there are many tests to be performed or comparisons to be made. Two of the most commonly used procedures are **Fisher's Least Significant Difference (LSD)** and **Tukey's Honestly Significant Difference (HSD)**.

Fisher's LSD: We begin with a one-way analysis of variance. If the overall F-ratio (which tests that hypothesis that all group means are equal) is statistically significant, we can safely conclude that not all of the treatment means are identical. Then, and only then...we carry out all possible t tests! Yes, the same "all possible t tests" that were just soundly criticized. The difference is that the t tests can't be performed unless the overall F-ratio is statistically significant. There is only a 5% chance of that the overall F ratio will reach statistical significance when there are no differences. Therefore, the chance of reporting a significant difference when there are none is held to 5%. Some authors refer to this procedure as Fisher's **Protected** LSD to emphasize the protection that the preliminary F-test provides. It is not uncommon to see the term *Fisher's LSD* used to describe all possible t tests without a preliminary F test, so stay alert and be a careful consumer of statistics.

Tukey's HSD: Tukey attacked the problem a different way by following in Student's (WS Gosset) footsteps. Student discovered the distribution of the t statistic when there were [b]two[/b] groups to be compared and there was no underlying mean difference between them. When there are *g* groups, there are *g(g-1)/2* pairwise comparisons that can be made. Tukey found the distribution of the *largest* of these t statistic when there were no underlying differences. For example, when there are 4 treatements and 6 subjects per treatment, there are 20 degrees of freedom for the various test statistics. For Student's t test, the critcal value is 2.09. To be statistically significant according to Tukey's HSD, a t statistic must exceed 2.80. Because the number of groups is accounted for, there is only a 5% chance that Tukey's HSD will declare something to be statistically significant when all groups have the same population mean. While HSD and LSD are the most commonly used procedures, there are many more in the statistical literature (a dozen are listed in the PROC GLM section of the SAS/STAT manual) and some see frequent use.

Multiple comparison procedures can be compared to buying insurance. Here, the insurance is against making a claim of a statistically significant result when it is just the result of chance variation. Tukey's HSD is the right amount of insurance when all possible pairwise comparisons are being made in a set of *g* groups. However, sometimes not all comparisons will be made and Tukey's HSD buys too much insurance.

In the preliminary stages of development, drug companies are interested in identifing compounds that have some activity relative to placebo, but they are not yet trying to rank the active compounds. When there are g treatments including placebo, only g-1 of the g(g-1)/2 possible pairwise comparisons will be performed. Charles Dunnett determined the behavior of the largest t statistic when comparing all treatments to a control. In the case of 4 groups with 6 subjects per group, the critical value for the three comparions of Dunnett's test is 2.54.

Similar considerations apply to Scheffe's test, which was once one of the most popular procedures but has now fallen into disuse. Scheffe's test is the most flexible of the multiple comparison procedures. It allows analysts to perform any comparison they might think of--not just all pairs, but the mean of the 1st and 2nd with the mean of the 4th and 6th, and so on. However, this flexibility comes with a price. The critical value for the four group, six subjects per group situation we've been considering is 3.05. This makes it harder to detect any differences that might be present. If pairwise comparisons were the only things an investigator wants to do, then it is unnecessary (foolish?) to pay the price of protection that the Scheffe test demand.

The moral of the story is to never take out more insurance than necessary. If you use Scheffe's test so that you're allowed to perform any comparison you can think of when all you really want to do is compare all treatments to a control, you'll be using a critical value of 3.05 instead of 2.54 and may miss some effective treatments.

The Bonferroni Adjustment

The most flexible multiple comparisons procedure is the **Bonferroni adjustment**. In order to insure that the probability is no greater than 5% that something will appear to be statistically significant when there are no underlying differences, each of 'm' individual comparisons is performed at the (0.05/m) level of significance. For example, with 4 treatments, there are m=4(4-1)/2=6 comparisons. In order to insure that the probability of no greater than 5% that something will appear to be statistically significant when there are no underlying differences, each of 'm' individual comparisons is performed at the 0.0083 (=0.05/6) level of significance. An equivalent procedure is to multiply the unadjusted P values by the number of test and compare the results to the nominal significance level--that is, comparing P to 0.05/m is equivalent to comparing mP to 0.05.

The Bonferroni adjustment has the advantage that it can be used in *any* multiple testing situation. For example, when an investigator and I analyzed cataract data at five time points, we were able to assure the paper's reviewers that our results were not merely an artifact of having examined the data at five different points in time because we had used the Bonferroni adjustment and performed each test at the 0.01 (=0.05/5) level of significance.

The major disadvantage to the Bonferroni adjustment is that it is not exact procedure. The Bonferroni adjusted P value is larger than the true P value. Therefore, in order for the Bonferroni adjusted P value to be 0.05, the true P-value must be smaller. No one likes using a smaller P value than necessary because it makes effects harder to detect. An exact procedure will be preferred when one is available. Tukey's HSD will afford the same protection as the Bonferroni adjustment when comparing many treatment groups and the HSD makes it easier to reject the hypothesis of no difference when there are real differences. In our example of four groups with six subjects per group, the critical value for Tukey's HSD is 2.80, while for the Bonferroni adjustment it is 2.93 (the percentile of Student's t distribution with 20 df corrsponding to a two-tail probability of 0.05/6=0.008333).

This might make it seem as though there is no place for the Bonferroni adjustment. However, as already noted, the Bonferroni adjustment can be used in any multiple testing situation. If only 3 comparions are to be carried out, the Bonferroni adjustment would have them performed at the 00.5/3=0.01667 level with a critical value of 2.63, which is less than the critical value for Tukey's HSD.

Summary Table

The critical values a t statistic must achieve to reach statistical significance at the 0.05

level
(4 groups, 6 subjects per group, and 20 degrees of freedom for the error variance).

Test	critical value
t test (LSD)	2.09
Duncan[*]	2.22
Dunnett	2.54
Bonferroni (3)	2.63
Tukey's HSD	2.80
Bonferroni (6)	2.93
Scheffe	3.05

[*] Duncan's New Multiple Range Test is a stepwise procedure. This is the critical value for assessing the homogeneity of all 4 groups.

If you look these values up in a table, Duncan, Dunnett, and Tukey's HSD will be larger by a factor of √2. I have divided them by √2 to make them comparable. The reason for the difference is the tables assume equal sample sizes of *n*, say. In that case, the denominator of the t statistic would contain the factor √[(1/n)+(1/n)] = √(2/n). Instead of referring to the usual t statistic (xbar$_i$-xbar$_j$)/[s$_p$ √(2/n)], the tables refer to the statistic (xbar$_i$öxbar$_j$)/[s$_p$ √(1/n)]. Since this statistic is the ordinary t statistic multiplied by √2, the critical values must be adjusted accordingly. If you should have occasion to use such a table, check the critical value for 2 groups and infinite degrees of freedom. If the critical value is 1.96, the test statistic is the usual t statistic. If the critical value is 2.77, the table expects the √2 to be removed from the denominator of the t statistic.

[Student]-Newman-Keuls Procedure

Most analysts agree that Fisher's LSD is too liberal. Some feel that Tukey's HSD is too conservative. While it is clear that the largest difference between two means should be compared by using Tukey's HSD, it is less obvious why the same criterion should be used to judge the *smallest* difference. The **[Student]-Newman-Keuls Procedure** is a compromise between LSD and HSD. It acknowledges the multiple comparison problem but invokes the following argument: Once we determine that the two extreme treatments are different according to the Tukey HSD criterion, we no longer have a homogeneous set of 'g' groups. At most, 'g-1' of them are the same. Newman and Keuls proposed that these means be compared by using the Tukey criteria to assess homogeneity in 'g-1' groups. The procedure continued in like fashion considering homogeneous groups of 'g-2' groups, 'g-3' groups, and so on, as long as heterogeneity continued to be uncovered. That is, the critical value of the t statistic got smaller (approaching the critical value for Student's t test) as the number of groups that might have the same mean decreased. At one time, the SNK procedure was widely used not only because it provided genuine protection against falsely declaring differences to be real but also because it let researchers have more significant differences than Tukey's HSD would allow. It is now used less often, for two reasons. The first is that, unlike the HSD or even the LSD approach, it cannot be used to construct confidence intervals for differences between means. The second reason is the growing realization that differences that depend strongly on the choice of particular multiple comparison procedure are probably not readily replicated.

Duncan's New Multiple Range Test

[You have two choices. You can promise never to use this test or you can read this section!]

Duncan's New Multiple Range Test is a wolf in sheep's clothing. It looks like the SNK procedure. It has a fancy name suggesting that it adjusts for multiple comparisons. And, to the delight of its advocates, gives many more satistically significant differences. It does this, despite its official sounding name, by failing to give real protection to the significance level. Whenever I am asked to review a paper that uses this procedure, I always ask the investigators to reanalyze their data.

This New Multiple Range Test, despite its suggestive name, does not really adjust for multiple comparisions. It is a stepwise procedure that uses the Studentized range statistic, the same statistic used by Tukey's HSD, but it undoes the adjustment for multiple comparisons!

The logic goes something like this: When there are g groups, there are g(g-1)/2 comparisons that can be made. There is some redundancy here because there are only g-1 independent pieces of information. Use the Studentized range statistic for g groups and the appropriate number of error degrees of freedom. To remove the penalty on the g-1 independent pieces of information, perform the Studentized range test at the $1-(1-\alpha)^{g-1}$ level of significance. In the case of 4 groups (3 independent pieces of information), this corresponds to performing the Studentized range test at the 0.143 level of significance.

When 'm' independent tests of true null hypotheses are carried out at some level α, the probability that none are statistically significant is $(1-\alpha)^m$ and the Type I error is $1-(1-\alpha)^m$. Therefore, to insure that the Studentized range statistic does not penalize me, I use at the level that corresponds to having used α for my individual tests. In the case of 4 groups, there are three independent pieces of information. Testing the three peices at the 0.05 level is like using the Studentized range statistic at the $1-(1-0.05)^3$ (=0.143) level. That is, if I use the Studentized range statistic with α=0.143, it is just as though I performed my 3 independent tests at the 0.05 level.

Additional Topics

Many Response Variables

The problem of multiple tests occurs when two groups are compared with respect to many variables. For example, suppose we have two groups and wish to compare them with respect to three measures of folate status. Once again, the fact that three tests are performed make it much more likely than 5% that something will be statistically significant at a nominal 0.05 level when there is no real underlying difference between the two groups. Hotelling's T^2 statistic could be used to test the hypothesis that the means of all variables are equal. A Bonferroni adjustment could be used, as well.

An Apparent Paradox

An investigator compares three treatments A, B, and C. The only significant difference is between B and C with a nominal P value of 0.04. However, when any multiple comparison procedure is used, the result no longer achieves statistical significance. Across town, three different investigators are conducting three different experiments. One is comparing A with B, the second is comparing A with C, and the third is comparing B with C. Lo and behold, they get the same P values as the investigator running the combined experiment. The investigator comparing B with C gets a P value of 0.04 and has no adjustment to make; thus, the 0.04 stands and the investigator will have an easier time of impressing others with the result.

Why should the investigator who analyzed all three treatments at once be penalized when the investigator who ran a single experiment is not? This is part of Kenneth Rothman's argument that there should be no adjustment for multiple comparisons; that all significant results should be reported and each result will stand or fall depending on whether it is replicated by other scientists.

I find this view shortsighted. The two P-values are quite different, even though they are both 0.04. In the first case (big experiment) the investigator felt it necessary to work with three groups. This suggests a different sort of intuition than that of the scientist who investigated the single comparison. The investigator working with many treatments should recognize that there is a larger chance of achieving nominal significance and ought to be prepared to pay the price to insure that many false leads do not enter the scientific literature. The scientist working with the single comparison, on the other hand, has narrowed down the possibilities from the very start and can correctly have more confidence in the result. For the first scientist, it's, "I made 3 comparisons and just one was barely significant." For the second scientist, it's, "A difference, right where I expected it!"

Planned Comparisons

The discussion of the previous section may be unrealistically tidy. Suppose, for example, the investigator working with three treatments really felt that the only important comparison was between treatments B and C and that treatment A was added only at the request of the funding agency or a fellow investigator. In that case, I would argue that the investigator be allowed to compare B and C without any adjustment for multiple comparisons because the comparison was planned in advance and had special status.

It is difficult to give a firm rule for when multiple comparison procedures are required. The most widely respected statistician in the field was Rupert G. Miller, Jr. who made no pretense of being able to resolve the question but offered some guidelines in his book Simultaneous Statistical Inference, 2nd edition (Chapter 1, section 5, emphasis is his):

> Time has now run out. There is nowhere left for the author to go but to discuss just what constitutes a family [of comparisons to which multiple comparison procedures are applied]. This is the hardest part of the book because this is where statistics takes leave of mathematics and must be guided by subjective judgment. . . .
>
> Provided the nonsimultaneous statistician [one who never adjusts for multiple comparisons] and his client are well aware of their error rates for groups of statements, and feel the group rates are either satisfactory or unimportant, the author has no quarrel with them. Every man should get to pick his own error rates. SImultaneous techniques certainly do not apply, or should not be applied, to every problem.
>
> [I]t is important to distinguish between two types of experiments. The first is the preliminary, search- type experiment concerned with uncovering leads that can be pursued further to determine their relevance to the problem. The second is the final, more definitive experiment from which conclusions will be drawn and reported. Most experiments will involve a little of both, but it is conceptually convenient to being basically distinct. The statistician does not have to be as conservative for the first type as for the second, but simultaneous techniques are still quite useful for keeping the number of leads that must be traced within reasonable bounds. In the latter type multiple comparison techniques are very helpful in avoiding public pronouncements of red herrings simply because the investigation was very large.
>
> The *natural family* for the author *in the majority of instances* is the *individual experiment* of a *single researcher*. . . . The loophole is of course the clause *in the majority of instances*. Whether or not this rule of thumb applies will depend upon the size of the experiment. Large single experiments cannot be treated as a whole without an unjustifiable loss in sensitivity. . . . *There are no hard-and-fast rules for where the family lines should be drawn, and the statistician must rely on his own judgment for the problem at hand*.

Unequal Sample Sizes

If sample sizes are unequal, exact multiple comparison procedures may not be available. In 1984, Hayter showed that the unequal sample size modification of Tukey's HSD is conservative. that is, the true significance level is no greater than the observed significance level. Some computer programs perform multiple comparison procedures for unequal sample sizes by pretending that the sample sizes are equal to their harmonic mean. This is called an *unweighted means analysis*. It was developed before the time of computers when the more precise calculations could not be done by hand. When the first computer programs were written, the procedure was implemented because analysts were used to it and it was easy to program. Thus, we found ourselves using computers to perform an analysis that was developed to be done by hand because there were no computers! The unweighted means analysis is not necessarily a bad thing to do if the sample sizes are all greater than 10, say, and differ by only 1 or 2, but this approximate test is becoming unnecessary as software packages are updated.

What do I do?

My philosophy for handling multiple comparisons is identical to that of Cook RJ and Farewell VT (1996), "Multiplicity Considerations in the Design and Analysis of Clinical Trials," Journal of the Royal Statistical Society, Series A, 159, 93-110. (The link will get you to the paper if you subscribe to JSTOR.) An extreme view that denies the need for multiple comparison procedures is Rothman K (1990), "No Adjustments Are Needed for Multiple Comparisons," Epidemiology, 1, 43-46.

I use Tukey's HSD for the most part, but I'm always willing to use unadjusted t tests for planned comparisons. One general approach is to use both Fisher's LSD and Tukey's HSD. Differences that are significant according to HSD are judged significant; differences that are not significant according to LSD are judged nonsignificant; differences that are judged significant by LSD by not by HSD are judged open to further investigation.

For sample size calculations, I apply the standard formula for the two sample t test to the most important comparisons, with a Bonferroni adjustment of the level of the test. This guarantees me the necessary power for critical pairwise comparisons.

[back to LHSP]

Last modified: undefined.

Obtaining Superscripts to Affix to Means That Are Not Significantly Different From Each Other

Gerard E. Dallal, PhD

[This page started out a few years ago as a technical paper. It's okay as technical papers go. It served me well as a web page for nearly four years. Still, it's technical. Here's the nontechnical version (well, less technical, anyway). Since the original version was never formally published, I'm hesitant to erase it and let it vanish. So, I moved it here.

To explain the concepts in a straightforward manner, I'm being a bit loose with my language. I am using the word *similar* to indicate *not shown to be different statistically* or *within sampling variability of each other*.]

When statistical program packages report the results of a multiple comparisons procedure, the output is usually in the form of a list of pairwise comparisons along with an indication whether each comparison is statistically significant. When these results are summarized for publication, standard practice is to present a table of mean with various superscripts attached and a comment such as, "Means sharing the same superscript are not significantly different from each other (Tukey's HSD, P<0.05)" or "Means that have no superscript in common are significantly different from each other (Tukey's HSD, P<0.05)." This procedure is widely used. Nevertheless, at the time of this writing (November 2003; the last version was written in March 2000!), none of the major statistical packages--SAS, SPSS, SYSTAT--provides the superscripts automatically. The analyst must deduce them from the table of P values. The one exception is the MEANS statement of SAS's GLM procedure, which can be used only when the number of observations is the same for each group or treatment.

The analyst must translate the list of pairwise differences into a set of superscripts so that those not judged different from each other share a superscript while those judged different do not have a superscript in common. By way of example, consider a set of four groups--A,B,C,D--where A was judged different from B and B was judged different from D. A brute force approach might use a different superscript for each possible comparison, eliminating those superscripts where the pair is judged significantly different. There are six possible comparisons--AB, AC, AD, BC, BD, CD--so the brute force approach would start with six superscripts

$$A^{abc}\ B^{ade}\ C^{bdf}\ D^{cef}\ ,$$

where the superscript *a* indicates that A & B are similar, the superscript *b* indicates that A & C are similar, and so on. The superscripts *a* and *e* would be eliminated--*a* because A & B were judged different and *e* because B & D were judged different. This leaves

$$A^{bc}\ B^{ad}\ C^{bdf}\ D^{cf}.$$

This is a true description of the differences between the groups, but it is awkward when you consider that the same set of differences can be written

$A^a\ B^b\ C^{ab}\ D^a$.

In both cases, A & B do not share a superscript, nor do B & D. However, every other combination *does* share a superscript. The second expression is much easier to interpret because it contains only two superscripts rather than the four in the first expression. The second expression makes it much easier to identify sets of 3 or more similar treatments.

There is a straightforward way to obtain the simpler expression. A computer program based on this method is now available.

- **It begins by writing out all possible subsets of treatments including the full set**, excluding the empty set and sets with one only one treatment.

  With four treatments A,B,C,D, the set of all possible subsets, excluding singletons and the empty set, is ABCD, BCD, ACD, ABD, ABC, AB, AC, AD, BC, BD, CD. If there are no statistically significant differences, all of these sets contain treatments that are similar to each other. The singletons (A, B, C, D) and the empty set are not used because they contain no more than one group. Therefore, they can *never* contain treatments that will be judged significantly different from each other.
- **Next, eliminate all sets that contain pairs judged significantly different.** That's because these sets no longs contain treatments that are all similar to each other.

  If the comparisons A & B and B & D are judged statistically significant, any set containing AB or BD is eliminated. The sets that are eliminated are

  ABCD, BCD, ABD, ABC, AB, BD

  leaving the sets

  ACD, AC, AD, BC, CD.
- **Then, eliminate any set that is contained in any other set.** That's because the smaller sets are implied by the sets that contain them. In this example, AC, AD, and CD are dropped because they are implied by ACD. If A, C, and D are similar, then A & C must be similar, and so on. This leaves

  ACD, BC.

  Thus, two marks/superscripts are needed. One is attached to the means of A, C, and D. The other is attached to the means of B and C.

  $A^a\ B^b\ C^{ab}\ D^a$

  This is consistent with the analysis that said the only statistically significant differences among these means were between A & B and B & D. A & B do not share a superscript, nor do B & D . Every other combination, however, *does* share a superscript.

Attaching Superscripts To Singletons

Some researchers have attached unique superscripts to single means that are judged to be different from all other means. For example, suppose when comparing four treatment means, D was judged significantly different from A, B, and C, while A, B, and C showed no statistically significant differences among themselves. Some researchers would attach a superscript to D, expressing the differences as

$$A^a\ B^a\ C^a\ D^b$$

rather than

$$A^a\ B^a\ C^a\ D.$$

I find superscripts affixed to a single mean to be the *worst* kind of visual clutter. They invite the reader to look for matches that don't exist. It's similar to reading an article that includes a symbol indicating a footnote and being unable to find the footnote! Without such superscripts, unique means stand unadorned and the absence of any superscript trumpet a mean's uniqueness. For this reason, I never use superscripts that are attached to only one mean.

Last modified: undefined.

Identifying Similar Groups

Check off all pairs of groups that are significantly different from each other:

#1	---							
#2	□	---						
#3	□	□	---					
#4	□	□	□	---				
#5	□	□	□	□	---			
#6	□	□	□	□	□	---		
#7	□	□	□	□	□	□	---	
#8	□	□	□	□	□	□	□	---
	#1	#2	#3	#4	#5	#6	#7	#8

Number of groups 3

Generate Superscripts

[back to the article describing this program]

[back to LHSP]

Adjusted Means, a.k.a. Least Squares Means

Gerard E. Dallal, Ph.D.

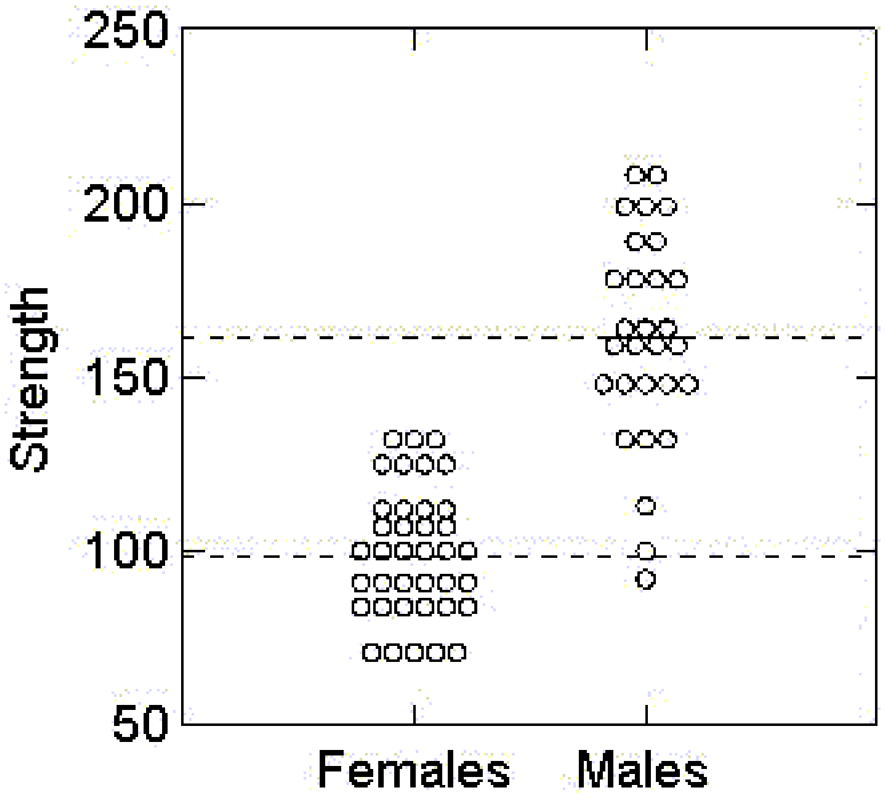
Means that have been corrected for imbalances in other variables are called **adjusted means**. The phrase **least squares means** was used in place of *adjusted means* by the statistical program package SAS. SAS is so widely used and highly respected that *least squares means* has begun to replace *adjusted means* in the applied scientific literature, Call me a traditionalist. I prefer *adjusted means*.

[back to LHSP]

Last modified: undefined.

Adjusted Means: Adjusting For Numerical Variables

Gerard E. Dallal, Ph.D.

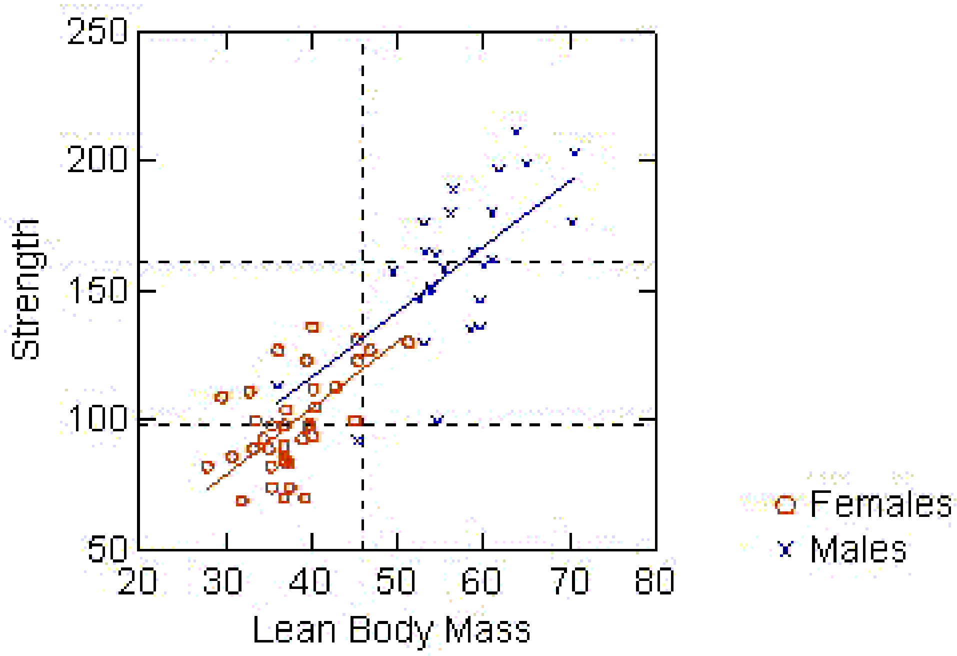


Are men stronger than women? It sure looks like it. Here are some data from a sample of healthy young adults. The measure of strength is something called slow, right extensor, knee peak torque. The dashed lines are drawn through the means of the men and women at 162 and 99 ft-lbs, respectively (P < 0.001, Student's t test for independent samples).

One might then ask the question of whether this is still true after adjusting for the fact that men tend to be bigger than women. In other words, ounce for ounce, are women just as strong as men? One way to answer this question is by fitting the analysis of covariance model

$$\text{strength} = b_0 + b_1 \text{ lean body mass} + b_2 \text{ SEX} ,$$

where SEX is coded, say, 0 for women and 1 for men. [Since sex can take on only two values, many analyst would prefer the variable name MALE, coded 0 for women and 1 for men, in keeping with the convention that when a variable is named after a "condition", 1 denotes the presence of the condition and 0 denotes its absence. I tend to use variable names like F0M1 where the name contains the codes.] We can use the fitted model to estimate the difference in strength between men and women with the same lean body mass. In this way, the model is said to adjust for lean body mass,



The fitted model is

strength = 2.17 + 2.55 lean body mass + 12.28 SEX ,

For a given amount of lean body mass, a male is predicted to be 12.28 ft-lbs stronger than a woman. However, this difference (the coefficient for SEX in the model) is not statistically significant (P = 0.186).

This is illustrated graphically in the scatterplot of strength against lean body mass. Once again, the horizontal dashed lines are drawn through the data at the mean strength values for men and women. We see that men are stronger, but they also have a lot more lean body mass. The red and blue lines are the model fitted to the data. The closeness of the lines suggests that, in this regard at least, men can be thought of as big women or women as small men.

Many investigators like to summarize the data numerically through **adjusted means**, which take the differences in other variables such as lean body mass into account. Adjusted means are nothing more than the predicted muscle strength of men and women with a given amount of lean body mass. Since the data are consistent with parallel lines, the difference between men and women will be the same whatever the amount of lean body mass. We could report predicted stength for any particular amount of lean body mass without distorting the difference between men and women. Standard practice is to predict muscle strength at the mean value of lean body mass in the combined sample. Here, the mean value of lean body mass is 45.8 kg. Thus, the adjusted mean (strength) for men is 131.3 ft-lb and 119.0 ft-lb for women. These values can be read off the vertical axis by following the vertical line at 45.8 kg of LBM to where it intersects the fitted regression lines.

This example illustrates everything I don't like about adjusted means. In essence, individual adjusted means by themselves don't mean anything! Well, they do, but they may be uninteresting or misleading. These adjusted means *are*the estimated strength of a man and woman with 45.8 kg of lean body mass. This is a lot of lean body mass for a women. It's not a lot of lean body mass for a man. It's a value that isn't typical for either group. Those familiar with muscle strength measures might be distracted as they wonder why our men are so weak or women so strong. This gets in the way of the data's message. I find it better to report the constant difference between the two groups.

To be fair, if there is considerable overlap between the groups--that is, if the groups would be expected to be the same with respect to variables being adjusted for if not for the effects of sampling--adjusted means can help bring things back into alignment. However, adjusted means should never be reported without giving critical thought to how they represent the data..

[back to LHSP]

Last modified: undefined.

Adjusted Means: Adjusting For Categorical Variables

Gerard E. Dallal, Ph.D.

In a study of the cholesterol levels of omnivores (meat eaters) and vegans (no animal products), suppose the data are something like

Mean (n)	Omnivore	Vegan
Male	230 (10)	220 (90)
Female	210 (40)	200 (10)

The mean cholesterol level of all omnivores is 214 mg/dl (= [230*10+210*40]/50) while for vegans it is 218 (= [220*90+200*10]/100). Thus, the mean cholesterol level of omnivores is 4 mg/dl **lower** than that of vegans even though both male and female omnivores have mean levels 10 mg/dl **higher** than vegans'! The reason for the discrepancy is a confounding, or mixing up, of sex and diet. Males have mean levels 20 mg/dl higher than females regardless of diet. The vegans are predominantly male and while the omnivores are predominantly female. The benefit of being a vegan is swamped by the deficit of being male while the deficit of being an omnivore is swamped by the benefit of being female.

Means that have been corrected for such imbalances are called **adjusted means** or, lately, **least squares means**. Adjusted means are predicted values from a multiple regression equation (hence, the name *least squares means*). The equation will contain categorical predictors (factors) and numerical predictors (covariates). Standard practice is to estimate adjusted means by plugging in the mean value of any covariate to estimate the mean response for all combinations of the factors and taking simple means of these estimates over factor levels. Those familiar with directly standardized rates will see that this is essentially the same operation.

If SEX is treated as a categorical variable, the adjusted mean cholesterol level for omnivores is calculated by taking the simple mean of the mean cholesterol levels for male omnivores and female omnivores (that is, 220 [= (230+210)/2]) and similarly for vegans (210 [= (220 +200)/2]). The adjusted mean for omnivores is 10 mg/dl higher than the vegans', which is the same as the difference observed in men and women separately. The calculations reflect the notion that, despite the imbalance in sample sizes, the best estimates of the cholesterol levels of male and female omnivores and vegans are given by the four cell means. The adjusted means simply average them.

If SEX is coded 0 for males and 1 for females, say, most statistical programs will evaluate the adjusted means at the mean value for SEX, which is 0.3333, the proportion of females. The adjusted means will be a weighted average of the cell means with the males being given weight 100/150 and the females given weight 50/150. For omnivores the adjusted mean is 223.3 [= 230 (100/150) + 210 (50/150)], while for vegans it is 213.3 [= 220 (100/150) + 200 (50/150)]. While these values differ from the earlier adjusted means, the difference between them is the same.

Choosing whether or not to name a two-level indicator variable as a factor can be thought of as choosing a set of weights to be applied to the individual levels. If the variable is categorical, the weights are equal. If the variable is numerical, the weight are proportional to the number of observations in each level.

In the previous example, the difference between adjusted means is 10 mg/dl, regardless of whether SEX is treated as categorical or numerical, because the difference between omnivores and vegans is the same for men and women. In practice, the differences will never be identical and the differences in adjusted means will depend on the choice of weights.

The following data are from a study of vitamin D-25 levels in healthy New Englanders during the wintertime. The actual analysis was more complicated, but here we will look at the difference between men and women adjusted for vitamin D intake. Vitamin D is manufactured in the body as the result of skin exposure to the sun, so it was decided to include an indicator for travel below 35 degrees north latitude.

Mean (n)	No travel	Traveler
Male	22.8 (47)	33.2 (5)
Female	22.3 (73)	29.5 (10)

The mean levels were 23.8 mg/dl for males and 23.1 for females. Adjusted means calculated by treating TRAVEL as a categorical variable in a model that also included a SEX-by-TRAVEL interaction and vitamin D intake as a covariate are 27.9 for males and 26.0 for females. The difference is 1.9 mg.dl. When TRAVEL is treated as a numerical variable, the adjusted means are 23.9 for males and 23.1 for females with a difference of 0.8 mg/dl. The former is the estimate based on equal numbers of travelers and nontravelers. The latter is the estimate based on mostly nontravelers.

While it is appropriate to compare adjusted means to each other, the individual adjusted means themselves are usually best ignored. They represent the estimated values for a specific set of circumstances that may not be realistic in practice.

In the previous example, the adjusted means calculated by treating TRAVEL as a categorical variable are 27.9 for males and 26.0 for females. These values are much larger than are typically seen in such a population and might be considered suspect. However, the reason they are so large is that they are the simple means of the values for travelers and nontravelers. There are very few travelers, but their vitamin D levels are 50% greater than those of nontravelers!

[back to LHSP]

Last modified: undefined.

Which Variables Should We Adjust For?

Gerard E. Dallal, Ph.D.

I suppose that before talking about what we should adjust for, a few sentence are in order about what we mean by adjusting and why we might want to do it.

How are adjustments made?

Adjustment is often nothing more than a linear adjustment achieved by adding another term to a regression model, as in

$$Y_{ij} = \mu + \alpha_i + \beta\, WT_{ij} + \varepsilon_{ij}$$

Within each group, we fit a linear relation between the response and the covariate(s). More complicated models can be fitted if the need arises, but unless the data are compelling, the linear term is commonly used as an approximation to whatever the relation might be.

When are adjustments made?

There are two chief reasons for adjusting for covariates. The one most people are familiar with is to adjust for imbalances in baseline variables that are related to the outcome. The adjustment helps correct for the groups' predisposition to behave differently from the outset. For example, if body weight was such a variable and one group was much heavier on average than the other, we might adjust for body weight.

The second, which is not fully appreciated, is to reduce the underlying variability in the data so that more precise comparisons can be made. Consider Student's t test for independent samples. There the difference in sample means is compared to the within-group standard deviation. Now, consider a simple analysis of covariance model

$$Y_{ij} = \mu + \alpha_i + \beta\, X_{ij} + \varepsilon_{ij}$$

Here, the difference in intercepts is compared to the variability about the within-group regression lines. If Y and X are highly correlated, the variabilty about the regression line will be much less than the within-group standard deviation. It is very nearly $\sqrt{(1-r^2)}$ times the within-group standard deviation. Thus, even if the groups are not imbalanced with respect to a covariate, it can still be a good idea to adjust for it to enhance the ability to recognize statistically significant effects.

What adjustments should be made?

It is always a good idea to make adjustments that will reduce variability inherent in treatment comparisons. The variables that will reduce variability will be know beforehand. These adjustments will be specified in the protocol before the data are collected. The design of the study--sample size calculations, in particular--will take these variance reductions into account.

Adjustments to correct imbalances are more controversial. We could adjust for everything imaginable. This may not do any harm other than cost us some error degrees of freedom. If there are enough data, it won't be of any real consequence. At the other extreme (for randomized trials), some argue that because of the randomization, it's not necessary to adjust for anything. While this is true from a theoretical perspective, let's not be stupid about it, I have yet to meet the statistician who in practice would fail to adjust once a large imbalance was detected in a baseline variable related to outcome. If no adjustment is made, it is impossible to tell whether any difference (or similarity!) in outcome is due to the treatments or the imbalance at baseline.

The sensible approach is an intermediate path that attempts to avoid adjustment but concedes the need for it when large imbalances are detected in variables that are know to be related to the outcome. Typical practice is to perform t tests or chi-square tests on the baseline variables and adjust for any where the observed significance level reaches a particular value (the ubiquitous 0.05, although some may choose a larger P value just to be safe).

An excellent discussion of these issues can be found in Assmann SF, Pocock SJ, Enos LE, Kasten LE (2000), "Subgroup Analysis and Other (Mis)Uses of Baseline Data in Clinical Trials", Lancet, 355, 1064-1069. I recommend it highly and agree completely, especially with the first two paragraphs of their discussion section, which touch on all of the important topics.

> In general, simple unadjusted analyses that compare treatment groups should be shown. Indeed they should be emphasised, unless the baseline factors for covariate adjustment are predeclared on the basis of their known strong relation to outcome. One notable exception is the baseline value of a quantitative outcome, in which analysis of covariance adjustment is the recommended primary analysis since a strong correlation is expected.
>
> Many trials lack such prior knowledge, requiring any strong predictors of outcome to be identified from the trial data by use of an appropriate variable selection technique. Covariate adjustment should then be a secondary analysis. Adjustment for baseline factors with treatment imbalances is unimportant, unless such factors relate to outcome. Nevertheless, such secondary analyses help achieve peace of mind.

Never underappreciate the value of "peace of mind"!

[back to LHSP]

Last modified: undefined.

Multi-Factor Analysis of Variance

Gerard E. Dallal, Ph.D.

With only a slight exaggeration, if you understand two-factor analysis of variance, you understand all of multi-factor analysis of variance. If you understand the issues raised by analyzing two factors simultaneously, then you'll understand the issues regardless of the number of factors involved.

With two-factor analysis of variance, there are two study factors (we'll call them factor A with *a* levels and factor B with *b* levels) and we study all (*a* times *b*) combinations of levels. For example, in a diet and exercise study, DIET and EXERCISE are the two study factors and we study all combinations of DIET and EXERCISE. The data can be displayed in a two-way table like a contingency table except that each cell might contain a mean, standard deviation, and sample size.

The secret to mastering two-factor analysis of variance is to understand the underlying model. The principal reason why multi-factor analyses are interpreted incorrectly is that users do not understand what is meant by the seductively named **main effect**. A **main effect** is the effect of a particular factor ***on average***. For example, the main effect of diet is the effect of diet averaged over all forms of exercise. Main effects are important, but focusing on them alone makes it possible to relive a series of bad jokes, namely, "The person who had his feet in the icebox and his head in the over but was fine, on average" or "The person who drowned in a pool that was 2 feet deep, on average".

In a multi-factor analysis of variance, we look at **interactions** along with main effects. Interactions are the extent to which the effects of one factor differs according to the levels of another factor. If there is an interaction between DRUG and SEX, say, the drug that is best for men might be different from the one that is best for women. If there is no interaction between the factors, then the effect of one factor is the same for all levels of the other factor. With no interaction, the drug that is best on average is the best for everyone.

When a computer program reports that the main effect of drug is highly statistically significant, it is tempting to stop right there, write it up, and send off a manuscript immediately. As we've just seen, an analysis should begin with an examination of the interactions because the interpretation of the main effects changes according to whether interactions are present. However, every computer package tempts us to look at main effects first by listing them in the output *before* the interactions.

The Model

Let

- the *a* levels of factor A define the rows of a table,
- the *b* levels of factor B define the columns,
- n_{ij} be the number of subjects in the (i,j)-th cell, that is, the number of subjects measured at the

combination of A_i and B_j,

- y_{ijk} be the response of the k-th subject in the (i,j)-th cell, where i=1,..,a; j=1,..,b; k=1,..,n_{ij}, and
- $N = \sum n_{ij}$.

The model could be written as

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

but it is usually written in a different way that takes advantage of the special structure of the study.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

where

- Y_{ijkj} is the response of the k-th subject measured at the combination of the i-th level of factor A and the j-th level of factor B,
- μ is an overall effect,
- α_i is the **main effect** of the i-th level of factor A,
- β_j is the **main effect** of the j-th level of factor B, and
- $(\alpha\beta)_{ij}$ is an **interaction**, an effect unique to the particular combination of levels. The combination $(\alpha\beta)$ to be read as a single symbol. It is called the **two factor interaction** (or **first order interaction**) between A and B. In computer models and output, it is denoted AB or A*B. It is **not** the product of α_i and β_j, which would be written $\alpha_i\beta_j$.

  Using $(\alpha\beta)_{ij}$ rather than a new symbol such as γ_{ij} allows the notation to represent many factors in a convenient manner. In a study involving four factors, there are four main effects, six two-factor interactions, four three-factor interactions, and a four-factor interaction. Sixteen unique symbols would be required to represent all of the effects and the underlying model would be difficult to read. On the other hand $(\alpha\beta\delta)_{ijl}$ is easily understood to be the three-factor interaction between factors A, B, and D.

A model without interactions is simpler to write and easier to explain. That model is said to be **additive** because the individual effects of the two factors are added together to describe their joint effect. The effect of a particular level of factor A is the same whatever the level factor B and vice-versa. The difference between two levels of factor A is the same for all levels of factor B. For example, If we focus on level *i* of factor A, the expected responses at levels 3 and 5 of factor B are

$$\mu + \alpha_i + \beta_3$$

and

$$\mu + \alpha_i + \beta_5$$

The effect of the level of factor B is added on to the effect of level *i* of factor A. The difference between the expected values is

$$\beta_3 - \beta_5$$

which is the same for all values of , that is, the difference is the same for all levels of factor A. This is no different from an ordinary regression model with no interactions. In fact, it *is* an ordinary regression model with no interactions.

When interactions are present, the effect of factor A depends on the level of factor B and the effect of factor B depends on the level of factor A. With interactions, the expected values become

$$\mu + \alpha_i + \beta_3 + (\alpha\,\beta)_{i3}$$

and

$$\mu + \alpha_i + \beta_5 + (\alpha\,\beta)_{i5}$$

The difference between them is

$$[\,\beta_3 + (\alpha\,\beta)_{i3}] - [\,\beta_5 + (\alpha\,\beta)_{i5}]$$

This difference depends on the value of *i*. The difference *changes* according to the level of factor A.

Just as with single factor ANOVA there are more parameters than groups, only more so! Constraints must be placed on the parameters so they can be estimated. The *usual constraints* force the parameters to sum to 0 in various ways.

- $\Sigma\,\alpha_i = 0$
- $\Sigma\,\beta_j = 0$
- $\sum_j (\alpha\beta)_{ij} = 0,$ for all *i*
- $\sum_i (\alpha\beta)_{ij} = 0,$ for all *j*

So, what's the problem?

Virtually every statistical software package displays its output starting with main effects followed successively more complicated interactions, that is, first come the two-factor interactions, then the three-factor interactions, and so on. However, the evaluation of a multi-factor analysis of variance should proceed in the opposite order, that is, by first looking at the most complicated interaction and, if it can be dismissed, by successively less complicated interactions. The underlying principle behind the analysis stated in its most dramatic form is: **Never analyze main effects in the presence of an interaction.** More properly, the principle is "never analyze an effect without regard to the presence of higher order relatives" but this lacks some of the dramatic bite of the first statement.

The reasons for this advice (and an understanding of when it can be safely ignored!) is easily obtained from a close examination of the model. The test for interaction asks whether the row effects are constant across the columns and, equivalently, whether the column effects are constant across the rows. If this is true--that is, if there is no interaction--then the model has been simplified dramatically. It makes sense to talk about *row effects* because they are the same for all columns. A similar argument applies regarding *column effects*.

Regardless of whether interactions are present, the test of row effects tests whether there is a common mean response for each row after averaging across all columns--that is, the test for row effects tests the hypothesis

$$H_0:\ \bar{\mu}_{1.} = .. = \bar{\mu}_{a.}$$

In similar fashion, the test of column effects tests whether there is a common mean response for each column after averaging across all rows--that is, the test for column effects tests the hypothesis

$$H_0:\ \bar{\mu}_{.1} = .. = \bar{\mu}_{.b}$$

If there is no interaction in the model, it makes sense to look for global (or *overall* or *simple*) row effects since they describe the differences between row levels regardless of the column level. Similarly, for column effects.

If interaction is present in the model, it doesn't make sense to talk about simple row effects because the row effects are column specific. For example, suppose the rows represent two drugs (X and Y) and the columns represent the sex of the subject. Suppose X is ineffective for both men and women while Y is ineffective for men but helps women. There is a drug-by-sex interaction since the difference between the drug changes with sex. The simple drug effect says that Y is better than X on average, that is, the hypothesis

$$H_0:\ (\mu_{X,men} + \mu_{X,women})/2 = (\mu_{Y,men} + \mu_{Y,women})/2$$

will be rejected even though both drugs are ineffective for men because Y is effective for women. The main effects look at whether the drugs behave the same when their effect is averaged over both men and women. When averaged over both men and women, the effect is *not* the same. Thus, the result of testing main effect is likely to be irrelevant since it doesn't apply equally to men and women. When an interaction is present, it is usually a mistake to report an analysis of the main effects because the effects will either be irrelevant or be misinterpreted as applying equally to everyone. Hence, the maxim **Never**

analyze main effects in the presence of an interaction.

I would prefer to leave it at that--**Never analyze main effects in the presence of an interaction**--because it's the right advice in almost every case. There are two exceptions worth mentioning. I hesitate only because it might make the general rule seem less important than it is.

The first exception has to do with the distinction between statistical significance and practical importance. It is quite possible for an interaction to be statistically significant yet not large enough to blur the message of the main effects. For example, consider two cholesterol lowering drugs. Suppose both are effective and while drug X has the same effect on men and women, drug Y on average lowers cholesterol an additional 10 mg/dl in men and 5 mg/dl in women. There is a drug-by-sex interaction because the difference between the drugs is not the same for men and women. Yet, the message of the main effects--take drug Y--is unaffected by the interaction.

The second exception comes from a hand-written note to myself on a scrap of paper I found in one of my files. (Perhaps someone can provide me with the original source if it wasn't something I concocted on the spur of the moment. It must be from a few years ago, because the page makes reference to SPSS-X.) The note reads, "Recall story of dairy farmer who could use only one type of feed for all breeds in herd." The story must go something like this...

A dairy farmer wished to determine which type of feed will produce the greatest yield of milk. From the research literature she is able to determine the mean milk output for each of the breeds she owns for each type of feed she is considering. As a practical matter, she can use only one type of feed for her herd.

Since she can use only one type of feed, she wants the one that will produce the greatest yield from her herd. She wants the feed type that produces the greatest yield when averaged over all breeds, even if it means using a feed that is not optimal for a particular breed. (In fact, it is easy to construct examples where the feed-type that is best on average is not the best for *any* breed!) The dairy farmer is interested in what the main effects have to say even in the presence of the interaction. She wants to compare

$$\bar{\mu}_{feed_1 \cdot}, \ldots, \bar{\mu}_{feed_n \cdot}$$

where the means are obtained by averaging over breed.

For the sake of rigor, it is worth remarking that this assumes the herd is composed of equal numbers of each breed. Otherwise, the feed-types would be compared through weighted averages with weights determined by the composition of the herd. For example, suppose feed A is splendid for Jerseys but mundane for Holsteins while feed B is splendid for Holsteins but mundane for Jerseys. Finally, let feed C be pretty good for both. In a mixed herd, feed C would be the feed of choice. If the composition of the herd were to become predominantly Jerseys, A might be the feed of choice with the gains in the Jerseys more than offsetting the losses in the Holsteins. A similar argument applies to feed B and a herd that is predominantly Holsteins.

[back to LHSP]

Last modified: undefined.

Pooling Effects

Gerard E. Dallal, Ph.D.

Analysis of *Variance* is a set of techniques for studying *means*. It works by looking at the variability in a response variable, breaking the variability apart, and assigning pieces to different effects. Consider the analysis of a balanced two-factor study where the common cell count is *n*.

	Sum of Squares	Degrees of Freedom
A	$\sum_{i} nb(\bar{y}_{i..} - \bar{y}_{...})^2$	a-1
B	$\sum_{j} na(\bar{y}_{.j.} - \bar{y}_{...})^2$	b-1
AB	$\sum_{i,j} n(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$	(a-1)(b-1)
Residual	$\sum_{i,j,k} (\bar{y}_{ijk} - \bar{y}_{ij.})^2$	(n-1)ab
Total	$\sum_{i,j,k} (\bar{y}_{ijk} - \bar{y}_{...})^2$	nab-1

For both sums of squares and degrees of freedom, Total=A+B+AB, that is the total variability in the data set is partitioned into three pieces. One piece describes how the row means differ from each other. Another describes how the column means differ from each other. The third describes the extent to which the row and column effects are not additive.

Each piece of the variability is associated with a particular piece of the ANOVA model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\ \beta)_{ij} + \varepsilon_{ij}$$

The following dicussion of pooling is an immediate consequence of a few facts.

- The Total Sum of Squares is unaffected by the model fitted to the data, that is, it is the same regardless of the model being used.
- Any variability the model fails to account for ends up in the Residual Sum of Squares.
- For this balanced experiment, the sum of squares for each of the treatment effects is the same regardless of whatever other effects are in the model. (This assumes the "usual constraints" are being used to constrain the parameters.)

Pooling

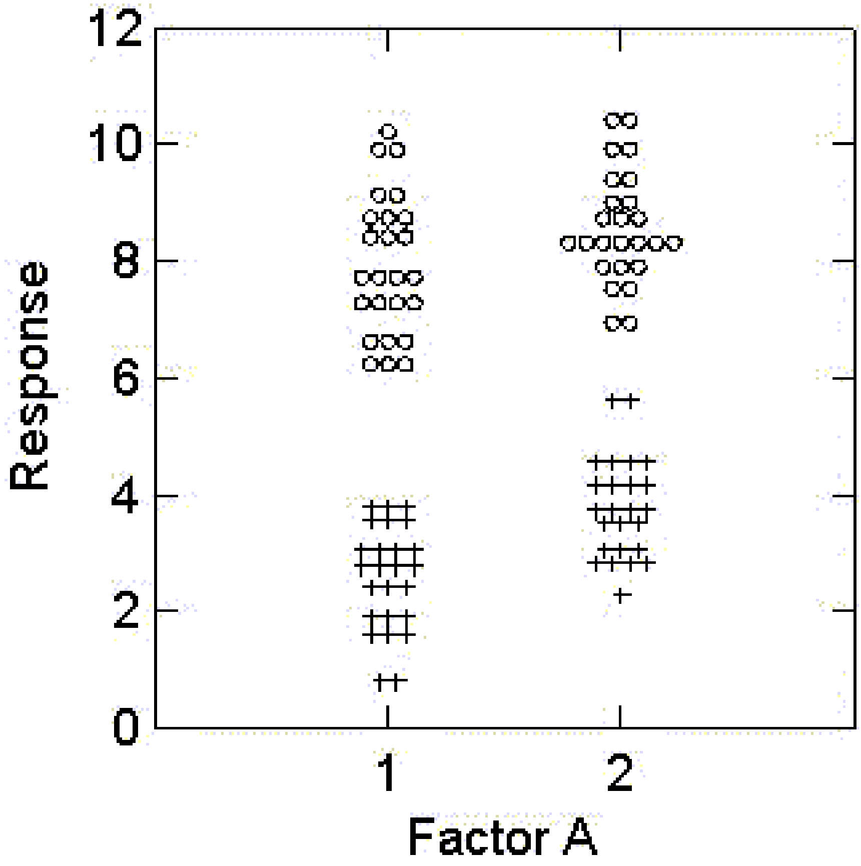
The idea behind pooling is that any effect that is not statistically significant can be eliminated from the model and the model can be refitted. In that case, the sums of squares and degrees of freedom corresponding to the eliminated terms are added into the residual sum of squares and degrees of freedom.

The first question should be, "Why bother?! What does it gain?" Primarly, residual degrees of freedom. This can help if the residual degrees of freedom for the full model is small--less than 10 or 20, say. In most studies, however, this is not an issue.

Pooling is a bad idea because the decision whether to pool is based on looking at the data. Any time the decision whether to do something is based on looking at the data, P values end up being different from what was originally though. Simulation studies have shown what might be expected. If the absence of an effect were known beforehand, pooling would be automatic regardless of the F ratio for the effect. In practice, pooling takes only after the mean squares for effects being pooled are seen not to be large compared to the original Residual Mean Square. When their sums of squares and degrees of freedom are combined with those of the original Residual, the new Residual mean squares is typically smaller than it would be if there were no peeking allowed. This has the effect of making ratios with this new Residual Mean Square in the denominator larger than they should be and other effects are more likely to appear statistically significant.

Other issues

More important than pooling is the notion that effects that do not appear in the model get folded into the residual sum of squares. Consider a two factor experiment once again. To keep things simple, let both factors have two levels. The issues surrounding pooling illustrate why it is inappropriate to use a simple t test to test the main effects even in the absence of an interaction.



In the diagram to the left, the levels of Factor A ar indicated by the tick marks on the horizontal axis. The levels of factor B are indicated by a 'o' or '+'. The reponse is higher for level 'o' of Factor B. The difference between 'o' and '+' is not significantly different for A_1 and A_2 (interaction P = 0.152). If Student's t test for independent samples is used to compare the levels of A--that is, if the presence of factor B is ignored--the P value is 0.109. However, in a two-factor analysis of variance, the P value for the main effect of Factor A is <0.001.

Both tests look at the same mean difference in levels of Factor A. The reason the P values are so different is the variabilty against which the mean difference is compared. In the t test, it is compared to the pooled estimate of variability within a strip of observations defined by the tick marks (2.75). In the two factor ANOVA, it is compared to the pooled estimate of within cell variability (0.98). The estimate of variability used for the t test is so much larger because it overlooks the Factor B effect. Variabilty that could be assigned to Factor B is left in the Residual Sum of Squares, inflating it. Both analyses follow, with the t test presented as a single factor ANOVA to make the visual comparison easier.

```
Source      Sum of    df     Mean     F-ratio  P
            Squares          Square

A            19.828    1    19.828     2.617  0.109
Error       742.488   98     7.576
Total       762.316   99
-----------------------------------------------------
A            19.828    1    19.828    20.553  0.000
B           647.858    1   647.858   671.542  0.000
A*B           2.016    1     2.016     2.090  0.152
Error        92.614   96     0.965
Total       762.316   99
```

Both analyses have the same lines labeled A and Total. The line labeled Residual in the t test has been broken apart into three pieces in the two-factor ANOVA--B, AB, and Residual. The bulk of the variability goes to the main effect for B. It is no longer considered a part of the unexplained variability,

The same principle applies to every regression analysis. Whenever a potential eplanatory variable is overlooked, its explanatory capability remains in the residual sum of squares. In this balanced ANOVA example, the sums of squares were additive because balance makes the effects uncorrelated. In the general regression problem predictors will be correlated. The various sums of squares--each variable adjusted for the presence of the others--will not be exactly additive, but the residual sum of squares will be inflated to the extent to which important predictor variables not appearing in the model are not perfectly correlated with the predictors in the model.

[back to LHSP]

Last modified: undefined.

Fixed and Random Factors

Gerard E. Dallal, Ph.D.

The source of these data is lost to memory. It may have started out as a textbook exercise. They are not from an actual experiment. The next time around I'll change the labels to make the exercise more realistic, but for now it's easier to go with the data as they are currently constituted,

In this experiment four nurses use each of three methods for measuring blood pressure. Nurses and methods are crossed factors with a total of 12 combinations. Thirty-six subjects participate. Each subject is measured by only one combination of nurse and method, that is, three different subjects are measured for each combination of nurse and method. The research question is *whether systolic blood pressure depends on the method use to measure it*.

The analysis of variance table for the experiment is

	Sum of Squares	Degrees of Freedom	Mean Square
Method	679.2	2	339.6
Nurse	815.8	3	271.9
Method*Nurse	612.4	6	102.1
Residual	163.9	24	68.3

Factors can either be **fixed** or **random**. A factor is **fixed** when the levels under study are the only levels of interest. A factor is **random** when the levels under study are a random sample from a larger population and the goal of the study is to make a statement regarding the larger population.

In this example, METHOD is a fixed factor. The purpose of this study is to examine these three methods of measuring blood pressure. There may be other methods, but they do not concern us here. When we are done, the hope is to make a statement comparing these three methods.

From the description of the study, the status of NURSE is less clear. If the investigator cares only about these four nurses, NURSE is a fixed factor. This might be the case where the study concerns the staff of a particular research unit and there is no goal of generalizing beyond the unit. Since only these four nurses matter, NURSE is a fixed factor. However, it might be that the point of the study is to generalize the results to all nurses. In that case, these four nurses might be viewed as a random sample of the population of all nurses, making NURSE a random factor.

One way to decide whether a factor is fixed or random is to ask what would happen if the study were repeated. If the same set of nurses would be used (as in the case of studying a particular research unit) the factor is fixed. If any set of nurses would do equally well, the factor is random.

There is an extra source of variability when factors random, so it should come as no surprise to learn that the analysis of METHOD changes according to whether NURSE is fixed or random. If NURSE is fixed, the analysis proceeds as usual. An F-ratio is constructed with the METHOD Mean Square in the numerator and the Residual Mean Square in the Denominator. The F-ratio is 4.97 [=339.6/68.3]. When it is compared to the percentiles of the F distribution with 2 numerator degrees of freedom and 24 denominator degrees of freedom, the resulting P value is 0.0156. Most statistical program packages produce this analysis by default. If NURSE is random, the F-ratio is still constructed with the METHOD Mean Square in the numerator, but the denominator is now the mean square for the METHOD*NURSE interaction. This F-ratio is 3.33 [=339.6/102.1]. When it is compared to the percentiles of the F distribution with 2 numerator degrees of freedom and 6 denominator degrees of freedom, the resulting P value is 0.1066.

When factors are fixed, the measure of underlying variability is the within cell standard deviation. Differences between methods are compared to the within cell standard deviation. When NURSES is random, methods are evaluated by seeing how much they differ on average relative to the way they differ from nurse to nurse. If two methods differ exactly the same way for all nurses, then that's the way they differ. However, if the differences between methods vary from nurse to nurse, many nurses must be examined to determine how the methods differ on average.

One critical consequence of NURSE being random is that the test for a METHOD effect depends on the number of nurses rather than the number of subjects measured by each nurse. This makes sense at the conceptual level because the determination of a METHOD effect is accomplished by seeing how methods differ from nurse to nurse. Therefore, the more nurses the better. Without going into too much detail, the measure of variability to which methods are compared when nurses are random behaves something like

$$\frac{\sigma_r + \frac{\sigma_\varepsilon}{n}}{r}$$

where σ_ε is an expression depending on the variability in individual subjects measured under the same conditions, *n* is the number of subjects per cell, σ_r is an expression depending on the variabilty between nurses, and *r* is the number of nurses. There is some advantage to be had by increasing *n*, but clearly the big gains are to be had by increasing *r*.

A faulty analysis?

If it is known that there is no interaction between method and nurse, a simpler model can be fitted.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

In that case, the error term for testing the METHOD effect will be the Residual term from the simple model. Because this was a balanced experiment, the METHOD mean square will the same for both models while the Residual mean square for the simpler model will be obtained by adding the Residual and interaction sums of squares and degrees of freedom from the full model. That is, the Residual mean square will be 75.0 [=(612.4+163.9)/(6+24)]. The F-ratio is 4.53 with a corresponding P value of 0.0192.

While this is similar to pooling, I view it as different. With pooling, the error term remains the same while nonsignificant effects are added to it primarily to increase the error degrees of freedom. With the kind of model simplification described here, the error term changes. My own take on this kind of model simplification is that it usually represents an attempt to salvage a study that was not designed properly in the first place.

The issue of fixed and random factors is currently making itself felt in an area called *group randomized trials*. An example of a group randomized study is a comparison of teaching methods in which randomization is achieved by randomizing classes to methods. When CLASS is treated as a random factor, the unit of observation is effectively the class, not the student. The precision of estimates is governed by the expression above with CLASS in place of NURSE. It is not uncommon to see group randomized trials improperly analyzed by treating the grouping variable as fixed rather than random.

Multi-Center Trials

Sometimes it is known from the outset that sufficient numbers of subjects cannot be recruited from a single location. It is common for such studies to be carried out as multi-center trials where subjects are enrolled from many centers. Each center has its own randomization list to insure that each center has subjects on each treatment.

An important question is whether the factor CENTER should be treated as fixed or random. If it is treated as fixed (or, equivalently except for a few degrees of freedom, there is assumed to be no center-by- treatment interation), the sample size is effectively the number of subjects. If CENTER is treated as random, the sample size is effectively the number of centers, and the study is much less powerful.

Standard practice is to treat CENTER as fixed. The rationale is that the same protocol under the control of a single set of investigators is used at all centers. However, if a statistically significant center-by-method interaction is encountered, it must be explained fully.

[back to LHSP]

Last modified: undefined.

Randomized (Complete) Block Designs

Gerard E. Dallal, Ph.D.

The Randomized Complete Block Designs is a bit of an odd duck. The design itself is straightforward. It's the analysis that might seem somewhat peculiar.

The design is best described in terms of the agricultural field trials that gave birth to it. When conducting field trials to compare fertilizers, plant varieties, or whatever, there is concern that some parts of a field may be more fertile than others. So, if one were comparing three fertilizers, say, it would not be a good idea to use one fertilizer here, another fertilizer over there, and the third fertilizer way out back because the effects of the fertilizers would be confounded with the natural fertility of the land.

The randomized block design goes this way.

- The field is divided into blocks and
- each block is divided into a number of units equal to the number of treatments.
- Within each block, the treatments are assigned at random so that a different treatment is applied to each unit. That is, all treatments are observed within each block. **The defining feature of the Randomized (Complete) Block Design is that each block sees each treatment *exactly* once.**[*]

The analysis assumes that there is no interaction between block and treatment, that is, it fits the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where the αs are the treatment effects and the βs are the block effects.

The ANOVA table contains three lines.

Source	Degrees of Freedom
Treatment	a-1
Blocks	b-1
Residual	(a-1)(b-1)

Since there are a*b units, the total number of degrees of freedom is *ab-1* and the residual degrees of freedom is

$$(ab-1)-(a-1)-(b-1) = (a-1)(b-1)$$

.

There is a better way to view randomized blocks that brings them into the mixed model framework. A Randomized (Complete) Block Design is a two-factor study in which the fixed factor TREATMENT is crossed with the random factor BLOCKS. If we fit the standard factorial model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\ \beta)_{ij} + \varepsilon_{ijk}$$

the ANOVA table becomes

Source	Degrees of Freedom
Treatment	a-1
Blocks	b-1
Treatment*Blocks	(a-1)(b-1)
Residual	(n-1)ab = **0**

where **n** is the number of observations per unit. Since n=1 (and **k** in the model takes on only the value 1), there are no degrees for pure error. However, in the mixed model, the proper way to test for a treatment effect is by comparing the Treatment mean square to the *Interaction* mean square, and we can estimate that. One might argue that the way the classic analysis works is by pooling the interaction and residual terms together and labeling them "Residual". However, since there is no pure residual, the term labeled Residual is really the Interaction, so everything works out properly!

While randomized block designs started out in agricultural field trials, they can apply to almost any field of investigation.

- In the lab, it is common for scientists to work with plates of cells divided into wells. Here, the plates are the blocks and the wells are the units. The same thing applies to gels divided into lines.
- When an experiment is replicated, each replicate can be considered a block.
- Often, the blocks are people, as in the case of a paired t test.

----

[*] There are also ***In*****complete Block Designs**, in which the number of units is less than the number of treatments, so that each block sees only a subset of treatments. Incomplete Block Designs are currently beyond the scope of these notes.

[back to LHSP]

Last modified: undefined.

Repeated Measures Analysis Of Variance
Part I: Before SAS's Mixed Procedure

Gerard E. Dallal, Ph.D.

Introduction

Repeated measures analysis of variance generalizes Student's t test for paired samples. It is used when two or more measurements of the same type are made on the same subject. At one time, some statisticians made a sharp distinction between measurement made using different but related techniques and serial measurements made the same way over time. The term *repeated measures* was reserved for nonserial measures. Lately the distinction has blurred. Any time multiple measurements are made on the same subject, they tend to be called repeated measures. It's unfortunate that the distinction has blurred because serial measures should be approached differently from other types of repeated measures. This will be discussed later. However, to keep the discussion simple, all we'll ask of repeated measures here is that multiple measurements of some kind be made on the same subject.

Analysis of variance is characterized by the use of *factors*, which are composed of *levels*. Repeated measures analysis of variance involves two types of factors--*between subjects factors* and *within subjects factors*.

The repeated measures make up the levels of the **within subjects factor**. For example, suppose each subject has his/her reaction time measured under three different conditions. The conditions make up the levels of the within subjects factor, which might be called CONDITION. Depending on the study, subjects may divided into groups according to levels of other factors called **between subjects factors**. Each subject is observed at only a single level of a between-subjects factor. For example, if subjects were randomized to aeorbic or stretching exercise, *form of exercise* would be a between-subjects factor. The levels of a within-subject factor change as we move within a subject, while levels of a between-subject factor change only as we move between subjects.

Technical Issues

Most statistical program packages report two separate analyses for repeated measures data. One is labeled *Univariate Repeated Measures Analysis of Variance*; the other is labeled *Multivariate Repeated Measures Analysis of Variance (MANOVA)*.

The univariate approach is more widely known and used because it was developed long before the ready availability of computers. The calculations can be performed by hand if necessary. It is essentially a multi-factor analysis of variance in which one of the factors is the random factor "Subject". The advantage to using a program's repeated measures routines is that the special handling required for the random "Subjects" factor is taken care of automatically. However, the univariate analysis demands that every pair of measures have the same correlation coefficient across subjects. While this may be reasonable when repeated measures come from different procedures, it is not realistic with serial measurements, where consecutive measurements are usually more highly correlated than measurements made far apart. Two adjustments--Greenhouse-Geisser and Huynh-Feldt--have been proposed to correct observed significance levels for unequal correlation coefficients.

The multivariate approach is computationally complex, but this is no longer an issue now that computers can do the work. The multivariate analysis does not require the correlations to be equal but it is less powerful (able to detect real differences) in small samples when the underlying conditions for the univariate approach are met. While both analyses require the data to follow a multivariate normal distribution, they differ in the way they are sensitive to violations of the assumption.

The multivariate approach includes many statistics--Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root. They are different ways of summarizing the data. In the vast majority of cases, the observed significance levels for these statistics will be the same, so multiple testing concerns will not apply.

In summary, there are two accepted ways to analyze repeated measures designs--the univariate approach and the multivariate approach. While they often agree, they need not. Looney and Stanley (The American Statistician, 43(1989), 220-225) suggest a Bonferroni approach: declare an effect significant at the 0.05 level if either test is significant at the 0.025 level. In my experience, this recommendation is too simplistic. When the tests disagree, it can be due to an outlier, or that the requirements of one or both tests are not met by the data, or because one test has much less power than the other. Further study is needed to determine the cause of the disagreement. Wilkinson (Systat Statistics manual, 1990, page 301) states, "If they [univariate and multivariate analyses] lead to different conclusions, you are usually [in 1988, it read "almost always"] safer trusting the multivariate statistic because it does not require the compound symmetry assumption." I tend to agree with Wilkinson, but not for small samples. There, the reduced number of degrees of freedom for error in the multivariate approach may cause it to fail to identify effects that are significant in the univariate analysis.

When there are significant differences between levels of the within subject factor or when there are interactions involving the within subjects factor, it is common to want to describe them in some detail. The major drawback to most of today's repeated measures analysis routines is that they do not provide the standard set of multiple comparison procedures. The only method most programs provide is paired t tests. Even that isn't easy because most programs, in a single run, will only compare a specified level to all other. When there are four levels, three separate analyses are required to generate all of the comparisons. The first might generate (1,2), (1,3), (1,4). The second might generate (2,1), (2,3), (2,4), which obtains (2,3) and (2,4) but duplicates (1,2). A third analysis is required to obtain (3,4). It is up to the user to apply a Bonferroni adjustment manually.

Another approach, which makes the standard set of multiple comparison procedures available, is to perform a multi-factor analysis of variance in which SUBJECT appears explicitly as a factor. Because SUBJECT is a random factor and the standard analysis of variance routines assume all factors are fixed, it is up to the user to see that test statistics are constructed properly by using whatever features the software provides. Also, the G-G and H-F corrections to observed significance levels are not provided, so it is up to the user to determine whether it is appropriate to assume the data possess compound symmetry. The data will have to be rearranged to perform the analysis.

In the standard data matrix, each row corresponds to a different subject and each column contains a different measurement. If three measurements are made on subjects randomized to one of two treatments (A/B), a subject's data might look like

```
  ID  TREAT   M1   M2   M3
1001    A     99  102  115
```

This is the subject-by-variables format required by most repeated measures analysis of variance programs.

In the rearranged data file, There will be as many records for each subject as there are repeated measures. In this type of file, a subject's data might look like

```
  ID  TREAT METHOD   X
1001    A       1   99
1001    A       2  102
1001    A       3  115
```

Most statistical program packages have their own routines for rearranging data. Some are easier to use than others. [I use the program SYSTAT for most of my work. However, I became so dissatisfied with its routines for rearranging data that I wrote my own program to rearrange my SYSTAT files for me. It will accept either version 7 (.SYS) or version 8 (.SYD) files as input. It produces version 7 files of rearranged data as output, so long variable names (longer than 8 characters) are not permitted. It can be downloaded by clicking on the link.]

The rearranged data can be analyzed by using a multi-factor, mixed model analysis of variance. It is a mixed model because the factor **subject**--here, ID--is a random factor, while TREAT and METHOD are fixed. It gets somewhat more complicated because ID is **nested** within TREAT. Nested factors is a special topic that deserves its own discussion. Because it's a mixed model, test statistics must be constructed carefully. Default tests that assume all factors are fixed will be inappropriate and often too liberal, that is, lead to statistical significance more often than is proper.

One way of looking at a program's repeated measures ANOVA module is that in exchange for the lack of multiple comparison procedures, it performs a univariate analysis properly, that is, it saves users from having to know how to define test statistics for themselves, and adds the G-G and H-F correction to the observed significance levels.

I oversimplify somewhat.

Practical Considerations

The types of analyses that can be obtained from a computer program depend on the way the data are arranged as well as the particular software package. Multiple arrangements of the data may be needed to analyze the data properly. Data should not have to be entered twice. Almost every full-featured package makes it possible to construct one type of file from the other but, at present, this is not a task for the novice.

The analysis of repeated measures data, like any other analysis, begins with a series of graphical displays to explore the data. After studying scatterplots, box plots, and dot plots, a line plot showing profiles of each treatment group is constructed by plotting mean response against time for each treatment group. Each treatment's data are connected by a distinctive line style or color so that the treatments can be distinguished. If the number of subjects is suitably small, parallel plots can be constructed similar to the line plot in which each data for individual subjects are plotted. There is typically one plot for each treatment group containing one line for each subject. In some program packages, these plots are more easily constructed when the data are arranged with one record per measurement by using a program's ability to construct separate plots for each subgroup, defined by subject or treatment.

SAS, SPSS, and SYSTAT all allow the use of nested factors, but only SYSTAT can specify them through menus. SPSS lets factors be specified as fixed or random and will generate the proper F ratios for repeated measures analyses. SAS has a similar feature but requires that interactions be declared fixed or random, too. SYSTAT has no such feature; each F ratio must be specified explicitly. Part of the reason for this inconsistency is that there is no general agreement about the proper analysis of mixed models, which makes vendors reluctant to implement a particular approach.

Comments

When there are only two repeated measures, the univariate and multivariate analyses are equivalent. In addition, the test for treatment-by-measure interaction will be equivalent to a single factor ANOVA of the difference between the two measurements. When there are only two treatment groups, this reduces further to Student's t test for independent samples. It is constructive to take a small data set and verify this by using a statistical program package.

Missing data: The standard statistical program packages provide only two options for dealing with missing data--ignore the subjects with missing data (keeping all the measures) or ignore the measures for which there are missing values (keeping all the subjects). Left to their own devices, the packages will eliminate subjects rather than measures, which is usually the sensible thing to do because typically more data are lost eliminating measures. If the data are rearranged and a multi-factor analysis of variance approach is used, all of the available data can be analyzed.

What to do? If these are the only programs available, I would begin with the standard analysis. If the univariate and multivariate approaches gave the same result and/or if the Greenhouse-Geiser and Huynh-Feldt adjusted P values did not differ from the unadjusted univariate P values, I would rearrange the data so that multiple comparison procedures could be applied to the within subjects factors.

[back to LHSP]

Last modified: undefined.

Repeated Measures Analysis Of Variance
Part II: After SAS's Mixed Procedure

Gerard E. Dallal, Ph.D.

[On occasion, I am asked when this note will be completed. That's a hard question to answer, but the delay is not for lack of interest or enthusiasm. This is arguably the most important topic in linear models today. The techniques described in Part I were developed to be carried out by hand, before computers were invented. They place many constraints on the data, not all of which are met in practice. The class of models that PROC MIXED makes available are computationally intensive, but much better reflect the structure of repeated measures data. However, this is not a simple topic that can be summarized suscintly in a few paragraphs, at least not by me at this time.

Until the time comes when I can do this note justice, and perhaps even afterward, there is no better discussion of repeated measures and longitudinal data than in the book Applied Longitudinal Analysis by Garrett Fitzmaurice, Nan Laird, and James Ware, published by John Wiley & Sons, Inc., ISBN 0-471-21487-6. (The link points to Amazon.com for the convenience of the reader. I am not an Amazon affiliate. I receive no remuneration of any kind if someone buys the book by clicking through. Amazon. I've stripped from the URL everything that looked like it could identify this site as having provided the link.)]

Prologue

SAS's MIXED procedure revolutionized the way repeated measures analyses are performed. It requires the data to be in the one-record-per- measurment (or many-records-per-subject) format. As with other programs that analyze data in that format, PROC MIXED handles missing data and applies multiple comparison procedures to both between and within subjects factors. Unlike other programs, PROC MIXED handles all of the technical details itself. In particular, it knows the proper way to construct its test statistics that account for the fixed and random nature of the study factors. In addition, it provides many important, unique features. For example, it provides for many covariance structures for the repeated measures. However, PROC MIXED has a rich command language which often provides many ways of accomplishing a particular task. Care and attention to detail is necessary so that a model is specified correctly.

In his 1998 book *Design and Analysis of Group-Randomized Trials*, David Murray wrote (p 228),

> From its inception, MIXED has employed approximate methods to compare *ddf* [denominator degrees of freedom] for fixed effects]. Unfortunately, MIXED does not always compute the *ddf* correctly, even in version 6.11. As a result, the analyst should compute the *ddf* based on the talbe of expected mean squares for the design and the partitioning of the total *ddf* among the sources in that table. The analyst can then specify the correct *ddf* using the `ddf=` option in the `model` statement.

One wonders whether the *ddf* are computed incorrectly or whether some of the options are behaving properly but in an unexpected manner!

Consider a simple repeated measures study in which 8 subjects (ID) are randomized to one of 2 treatments (TREAT) and then measured under 3 periods (PERIOD). Although it might be unrealistic, let's fit the model assuming compound symmetry. The command language can be written

```
proc mixed;
  class id treat period;
  model y=treat period treat*period;
  repeat period/sub=id(treat) type=cs;
```

The key elements are that all factors, fixed and random, go into the class statement. Only fixed factors go into the `model` statement, however. The `repeat` statement specifies the repeated measures, while the `sub=` option is used to specify the variable that identifies subjects.

The same results can be obtained by the command language

```
proc mixed;
class id treat period;
  model y=treat period treat*period;
  random id(treat) period*id(treat);
```

or

```
proc mixed;
class id treat period;
  model y=treat period treat*period;
  random id(treat);
```

or

```
proc mixed;
class id treat period;
  model y=treat period treat*period;
  random int/sub=id(treat);
```

In all three examples, the `repeated` statement is replaced by a `random` statement. In the first example, there is no `sub=`option and all random factors are declared explicitly in the `random` statement. In the second example, period*id(treat) is left off the `random` statement. This is possible because its inclusion exhausts the data. When it is eliminated, it is not being pooled with other sources of variation. The third example uses the `sub=` option to specify a subject identifier. In this formulation, the `random` statement specifies a random intercept (`int`) for each subject.

The three most commonly used covariance structures are *compund symmetry* (CS), *unstructured* (UN), and *auto regressive (1)* (AR(1)).

[back to LHSP]

Last modified: undefined.

Why SAS's PROC MIXED Can Seem So Confusing

Gerard E. Dallal, Ph.D.

[Early draft subject to change.]

[The technical details are largely a restatement of the Technical Appendix of Littell RC, Henry PR, and Ammerman CB (1998), "Statistical Analysis of Repeated Measures Data Using SAS Procedures", Journal of Animal Science, 76, 1216-1231.]

Abstract

The **random** and **repeated** statements of SAS's PROC MIXED have different roles. The **random** statement identifies random effects. The **repeated** statement specifies the structure of the within subject errors. They are not interchangeable. However, there are overspecified models that can be specified by using a **random** or **repeated** statement alone. Unfortunately, one such model is the commonly encounterd repeated measures with compound symmetry. This has the potential of leading to confusion over the proper use of the two types of statements.

The simple answer to why SAS's PROC MIXED can seem so confusing is that it's so powerful, but there's more to it than that. Early on, many guides to PROC MIXED present an example of fitting a compound symmetry model to a repeated measures study in which subjects (ID) are randomized to one of many treatments (TREAT) and then measured at multiple time points (PERIOD). The command language to analyze these data can be written

```
proc mixed;
  class id treat period;
  model y=treat period treat*period;
  repeat period/sub=id(treat) type=cs;
```

or

```
proc mixed;
class id treat period;
  model y=treat period treat*period;
  random id(treat);
```

Because both sets of command language produce the correct analysis, this immediately raises confusion over the roles of the repeated and random statements, In order to sort this out, the underlying mathematics must be reviewed. Once the reason for the equivalence is understood, the purposes of the repeated and random statements will be clear.

PROC MIXED is used to fit models of the form

$$y = X\beta + ZU + e$$

where

- y is a vector of responses
- X is a known design matrix for the fixed effects
- β is vector of unknown fixed-effect parameters
- Z is a known design matrix for the random effects
- U is vector of unknown random-effect parameters
- e is a vector of (normally distributed) random errors.

The random statement identifies the random effects. The repeated statement specifies the structure of the within subject errors.

For the repeated measures example,

$$y_{ijk} = \mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik} + u_{ij} + e_{ijk}$$

where

- y_{ijk} is response at time *k* for the *j*-th subject in the *i*-th group
- μ , α_i, γ_k, and $(\alpha\gamma)_{ik}$ are fixed effects
- u_{ij} is the random effect corresponding to the j-th subject in the i-th group
- e_{ijk} is random error

The variance of y_{ijk} is

$$\text{var}(y_{ijk}) = \text{var}(u_{ij} + e_{ijk})$$

The variance of the u-s is typically constant (denoted σ_u^2). The errors e_{ijk} are typically idependent of the random effects u_{ij}. Therefore,

$$\mathrm{var}(y_{ijk}) = \sigma_u^2 + \mathrm{var}(e_{ijk})$$

The covariance between any two observations is

$$\mathrm{cov}(y_{ijk},y_{lmn}) = \mathrm{cov}(u_{ij},u_{lm}) + \mathrm{cov}(u_{ij},e_{lmn}) + \mathrm{cov}(u_{lm},e_{ijk}) + \mathrm{cov}(e_{ijk},e_{lmn})$$

Observations from different animals are typically considered to be independent of each other. Therefore, the covariance between two observations will be 0 unless i=l and j=m, in which case

$$\mathrm{cov}(y_{ijk},y_{ijn}) = \mathrm{cov}(u_{ij},u_{ij}) + \mathrm{cov}(e_{ijk},e_{ijn})$$
$$= \sigma_u^2 + \mathrm{cov}(e_{ijk},e_{ijn})$$

Under the assumption of compound symmetry, $\mathrm{cov}(e_{ijk},e_{ijn})$ is $\sigma_e^2+\sigma$, for k=n, and σ_e^2, otherwise. It therefore follows that

$$\mathrm{var}(y_{ijk}) = \sigma_u^2 + \sigma_e^2 + \sigma$$

and

$$\mathrm{cov}(y_{ijk},y_{ijn}) = \sigma_u^2 + \sigma_e^2.$$

The model is redundant because σ_u^2 and σ_e^2 occur only in the sum $\sigma_u^2 + \sigma_e^2$, so the sum $\sigma_u^2 + \sigma_e^2$ can be estimated, but σ_u^2 and σ_e^2 cannot be estimated individually. The command language file with the **random** statement resolves the redundancy by introducing the **u**-s into the model and treating the repeated measures as independent. The command language file with the **repeated** statement resolves the redundancy by removing the **u**-s from the model.

Littel et al. point out that a similar redundancy exists for the unstructured covariance matrix (TYPE=UN), but there is no reduncancy for an auto-regressive covariance structure (TYPE=AR1). In the latter case, both random and repeated statements should be used. See their article for additional details.

[back to LHSP]

Gerard E. Dallal

Last modified: undefined.

The Analysis of Pre-test/Post-test Experiments

Gerard E. Dallal, Ph.D.

[This is an early draft. **[figure]** is a placeholder
for a figure to be generated when I get the chance.]

Consider a randomized, controlled experiment in which measurements are made before and after treatment.

One way to analyze the data is by comparing the treatments with respect to their **post-test measurements**. [figure]

Even though subjects are assigned to treatment at random, there may be some concern that any difference in the post-test measurements might be due a failure in the randomization. Perhaps the groups differed in their pre-test measurements.* [figure]

One way around the problem is to compare the groups on differences between post-test and pretest, sometimes called **change scores** or **gain scores**. [figure] The test can be carried out in a number of equivalent ways:

- t-test of the differences;
- 2-group ANOVA of the differences,
- repeated measures analysis of variance.

However, there is another approach that could be used--**analysis of covariance**, in which

- the post-test measurement is the response,
- treatment is the design factor, and
- the pre-test is a covariate.

[figure] It is possible for the analysis of covariance to produce a significant treatment effect while the t-test based on differences does not, and vice-versa. The question, then, is which analysis to use.

The problem was first stated by Lord (1967: Psych. Bull., 68, 304-305) in terms of a dietician who measures students' weight at the start and end of the school year to determine sex differences in the effects of the diet provided in the university's dining halls. The data are brought to two statisticians. The first, analyzing the differences (weight changes), claims there is no difference in weight gain between men and women. The second, using analysis of covariance, finds a difference in weight gain. Lord's conclusion was far from optimistic:

> [W]ith the data usually available for such studies, there is simply no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled pre-existing differences between groups. The researcher wants to know how the groups would have compared if there had been no pre-existing uncontrolled differences. The usual research study of this type is attempting to answer a question that simply cannot be answered in any rigorous way on the basis of available data.

Lord was wrong. His confusion is evident in the phrase, "controlling for pre-existing conditions." The two procedures, t-test and ANCOVA, **test different hypotheses**! For Lord's problem,

- the t test answers the question, "Is there a difference in the mean weight change for boys and girls?"
- ANCOVA answers the question, "Are boys and girls of the same initial weight expected to have the same final weight?" or, in Lord's words, "If one selects on the basis of initial weight a subgroup of boys and a subgroup of girls having identical frequency distribution of initial weight, the relative position of the regression lines shows that the subgroup of boys is going to gain substantially more during the year than the subgroup of girls."

Despite how proper and reasonable the ANCOVA question seems, it is **NOT** what the dietician really wanted to know. The reason it's wrong is that when looking at boys and girls of the same weight, one is looking at a relatively light boy and a relatively heavy girl. Even if the school cafeteria had no effect on weight, regression to the mean would have those heavy girls end up weighing less on average and those light boys end up weighing more, even though mean weight in each group would be unchanged.

Campbell and Erlebacher have described a problem that arises in attempts to evaluate gains due to compensatory education in lower-class populations.

> Because randomization is considered impractical, the investigators seek a control group among children who are not enrolled in the compensatory program. Unfortunately, such children tend to be from somewhat higher social-class populations and tend to have relatively greater educational resources. If a technique such as analysis of covariance, blocking, or matching (on initial ability) is used to create treatment and control groups, the posttest scores will regress toward their population means and spuriously cause the compensatory program to appear ineffective or even harmful. Such results may be dangerously misleading if they are permitted to influence education policy. [Bock, p. 496]

Now, consider a case where two teaching methods are being compared in a randomized trial.

Since subjects are randomized to method, we **should** be asking the question, "Are subjects with the same initial value expected to have the same final value irrespective of method?" Even if there is an imbalance in the initial values, the final values should nevertheless follow the regression line of POST on PRE. A test for a treatment effect, then, would involve fitting separate regression lines with common slope and testing for different intercepts. But this is just the analysis of covariance.

Summary

- Use t tests when experimental groups are defined by a variable that is relevant to the change in measurement.
- Use analysis of covariance for experiments in which subjects are assigned randomly to treatment groups, regardless of whether there is any bias with respect to the initial measurement.

NOTES

1. When subjects are randomly assigned to treatment, ANCOVA and t-tests based on differences will usually give the same result because significant imbalances in the pretest measurements are unlikely.

   If the measurements are highly correlated so that the common regression slope is near 1, ANCOVA and t-tests will be nearly identical.

2. ANCOVA using difference (post - pre) as the response and pre-test as the covariate is equivalent to ANCOVA using post-test as the response. Minimizing

   $$\Sigma\ [(\text{POST-PRE}) - (a\ \text{TREAT} + b * \text{PRE})]^2$$

   is equivalent to minimizing

   $$\Sigma\ [\text{POST} - (c\ \text{TREAT} + d * \text{PRE})]^2$$

   with a = c and d = 1 + b.

3. The analysis could be taken one step further to see whether the ANCOVA lines are parallel. If not, then the treatment effect is not constant. It varies with the initial value. This should be reported. There may be a range of covariate values within which the two groups have not been shown to be significantly different. The Johnson-Neyman technique can be used to identify them.

----

* This is actually a thorny problem. It is generally a bad idea to adjust for baseline values *solely* on the basis of a significance test.

- it messes up the level of the test of the outcome variable
- if the randomization were to have failed, differences in the baseline that do not reach statistical significance might still be sufficient to affect the results.

However, there is a good reason, other than imbalance in the initial values, for taking the initial values into account. In most studies involving people, analyses that involve the initial values are typically **more powerful** because they eliminate much of the between-subject variability from the treatment comparison.

Last modified: undefined.

Serial Measurements

Gerard E. Dallal, Ph.D.

When the same quantity is measured repeatedly over time on the same individuals, the resulting values are called **serial measurements**.

Standard repeated measures analyses are almost always inappropriate for serial measurements. When a repeated measures analysis is applied to serial data, the result is invariably one of two types--either the mean response is not the same at all time points or the mean value changes over time varies with to the level of some between subjects factor, that is, there is an interaction between time and the between subjects factor.

These analyses typically raise more questions than they answer. For example, a treatment-by-time interaction will be observed unless the mean response over time is the same for all treatments. However, it rarely is, and many of these interaction will be of questionable biological importance and difficult to interpret. It is common to see reports with a significant treatment-by-time interaction, in which investigators use Student's t test to compare two treatments at every time point and declare the two treatments to be the same at the 1st, 3rd, 4th, 5th, 6th, 8th, 9th, and 10th measurements but different at the 2nd and 7th measurements, without any sense of what this might mean biologically. For this reason, it is usually better to construct a simple summary of the repeated measurements for each subject based on biological considerations and analyze the summary by using familiar univariate statistical techniques, that is, techniques that are appropriate when there is a single measurement per subject. Typical summaries include mean response, difference between first and last measurement, area under the curve as determined by trapezoidal rule, maximum response, linear regression coefficient, and time of maximum response. See Matthews JNS, Altman DG, Campbell MJ, and Royston PG (1990), "Analysis of Serial Measurements In Medical Research," British Medical Journal, 300, 230-5.

[back to LHSP]

Last modified: undefined.

The Computer-Aided Analysis of Crossover Studies

Gerard E. Dallal, Ph.D.

Abstract

This note describes the computer-aided analysis of two treatment, two-period crossover studies. All participants are given both treatments. Half of the subjects receive the treatments in one order, the others receive the treatments in the reverse order. SAS and SYSTAT command language is given for the analysis of such trials.

Introduction

Most studies of two treatments--A and B, say--are parallel groups studies, so-called because the treatments are studied in parallel. One group of subjects receives only treatment A, the other group receives only treatment B. At the end of the study, the two groups are compared on some quantitative outcome measure (a final value of some marker, a change from baseline, or the like), most often by using a t test for independent samples.

It takes little experience with parallel group studies to recognize the potential for great gains in efficiency if each subject could receive both treatments. The comparison of treatments would no longer be contaminated by the variability between subjects since the comparision is carried out within each individual.

If all subjects received the two treatments in the same order, observed differences between treatments would be confounded with any other changes that occur over time. In a study of the effect of treatments on cholesterol levels, for example, subjects might change their diet and exercise behavior for the better as a result of heightened awareness of health issues. This would likely manifest itself as a decrease in cholesterol levels over the later portion of the study and might end up being attributed to the second treatment.

The two treatment, two-period crossover study seeks to overcome this difficulty by having half of the subjects receive treatment A followed by treatment B while the other half receive B followed by A. The order of administration is incorporated into the formal analysis. In essence, any temporal change that might favor B over A in one group will favor A over B in the other group and cancel out of the treatment comparison.

Even though crossover studies are conceptually quite simple, the literature is difficult to read for many reasons.

1. Terminology and notation varies from author to author, making it difficult to compare discussions.

Reference	Terminology		
[2]	TREATMENT	PERIOD	TREATMENT*PERIOD
[3]	TREATMENT	PERIOD	SEQUENCE
[6]	TREATMENT	SEQUENCE*TREATMENT	SEQUENCE

Complicated notational devices are introduced to describe simple comparisons among four cell means.

- In Grizzle [1], y_{ijk} represents the response of subject j to treatment k applied at time period i.
- In Hills and Armitage [2], y_i represents the response of a subject in period i. The treatment is implied by context. The difference between treatments X and Y for group A is $d_A=y_1-y_2$, while the difference between treatments X and Y for group B is $d_B=y_2-y_1$.
- In Fleiss [3], X_j represent the response in period j of a subject who receives the treatments in the first order and Y_j represent the response in period j of a subject who receives the treatments in the second order.

2. Factors such 'time period' and 'sequence in which the treatments are given' are easily confused when reduced to the one word labels required by printed tables or computer programs.

3. The mathematical theory for cross-over studies was developed before the ready availabilty of computers, so practical discussions concentrated on methods of analysis that could be carried out by hand. Because the basic crossover involves only two treatments and two periods, most authors give the analysis in terms of t tests. Virtually all general purpose computer programs analyze crossover studies as special cases of repeated measure analysis of variance and give the results in terms of F tests. The two approaches are algebraically equivalent, but the difference in appearance makes it difficult to reconcile computer output with textbooks and published papers.

4. Many published discussions and examples are incorrect. Grizzle [1] gave an incorrect analysis of studies with unequal numbers of subjects in each sequence group. Nine years later, Grizzle [4] corrected the formula for sums of squares for treatments. After an additional eight years elapsed, Grieve [5] noted, "Although . . . it should be clear that the period sum of squares . . . is also incorrect, the analysis put forward in Grizzle [1,4] still appears to be misleading people. I know of three examples of computer programs, written following Grizzle's analysis, with incorrect period and error sums of squares."

   Grizzle's flawed analysis continues to muddy the waters. In their otherwise excellent book, "intended for everyone who analyzes data," Milliken and Johnson [6] present the flawed analysis in Grizzle [1]; Grizzle [4] is not referenced.

The Crossover Study

In this discussion, the design factors will be denoted

- treatment (TREATMENT), with levels 'A' and 'B',
- time period (PERIOD), with levels '1' and '2',
- sequence group (GROUP), with levels 'A then B' and 'B then A'.

Some authors prefer SEQUENCE to GROUP. There are two powerful reasons for using GROUP. First, the word GROUP is unambiguous. It implies differences between subjects, and there is only one way subjects differ--in the order in which they receive the two treatments. Confusion over SEQUENCE/ PERIOD/ORDER is eliminated. Second, the error term for testing the significance of the GROUP factor is different from the error term for testing PERIOD and TREATMENT, as is true of any repeated measures study with between- and within-subjects factors. The label GROUP helps keep this in mind.

The four observed cell means

Group 1:	A in period 1	B in period 2
Group 2:	B in period 1	A in period 2

will be denoted

x_1	x_2
x_3	x_4

While this prescription introduces yet another set of notation, here the notation is neutral--no attempt has been made to describe the experiment through the notation. When discussing a set of four numbers, this neutrality proves a virtue rather than a vice.

Statistical Details

TREATMENTS are compared by combining the difference between A and B from within each group, specifically

$$((x_1 - x_2) + (x_4 - x_3)) / 2 .$$

PERIODS are compared by looking at the difference between the measurements in period 1 and those made in period 2

$$((x_1 + x_3) - (x_2 + x_4)) / 2 .$$

If period effects are present, they do not influence the comparison of treatments. A period 1 effect appears in the treatment comparisons as part of x_1 and x_3 and cancels out of the treatment difference, while a period 2 effect appears as part of x_2 and x_4 and cancels out, as well. Similarly, treatment effects do not influence the comparison of time periods. Treatment A appears in x_1 and x_4 and cancels out of the PERIODS effect, while treatment B appears as part of x_2 and x_3 and cancels out, too.

Aliasing

If crossover studies were full-factorial designs (with factors GROUP, TREATMENT, and PERIOD), it would be possible to evaluate not only the main effects, but also the GROUP*TREATMENT, PERIOD*TREATMENT, GROUP*PERIOD, and PERIOD*TREATMENT*GROUP interactions. However, crossover studies are not full factorial designs. Not all combinations of factors appear in the study (there is no GROUP='A then B', PERIOD='1', TREATMENT='B' combination, for example). Because only four combinations of the three factors are actually observed, main effects are confounded with two-factor interactions, that is, **each estimate of a main effect also estimates a two-factor interaction**.

As an illustration, notice that the difference between GROUPS is estimated by comparing the two means for group 1 to the two means for group 2, that is,

Group 1		Group 2
$(x_1 + x_2)$	-	$(x_3 + x_4)$

Now consider the PERIOD*TREATMENT interaction, which measures how the difference between treatments change over time. The interaction is estimated by

Period 1		Period 2
A - B		A - B
$(x_1 - x_3)$	-	$(x_4 - x_2)$

But this is the estimate of the GROUP effect. Thus, GROUP and PERIOD*TREATMENT are confounded. They are *aliases*, two names for the same thing. In the two treatment, two period crossover study, each main effect is confounded with the two-factor interaction involving the other factors.

Effect	Alias
TREATMENT	GROUP*PERIOD
PERIOD	GROUP*TREATMENT

TREATMENT*PERIOD   GROUP

If one of the main effects is significant, it is impossible to tell whether the effect, its alias, or both are generating the significant result. One could argue that there is no reason to expect a significant effect involving GROUP because subjects are assigned to GROUPS at random. Therefore, a significant GROUP effect should be interpreted as resulting from a PERIOD*TIME interaction and not from a difference between GROUPS. For similar reasons, a significant PERIOD effect is not considered to be the result of a GROUP*TIME interaction, nor is a significant TREATMENT effect to be the result of a GROUP*PERIOD interaction.

Carryover Effects

Carryover (or residual) effects occur when the effect of a treatment given in the first time period persists into the second period and distorts the effect of the second treatment. Carryover effects will cause the difference between the two treatments to be different in the two time periods, resulting in a significant TREATMENT*PERIOD interaction. Thus TREATMENT*PERIOD is not only an alias for GROUP, it is also another way of labelling CARRYOVER effects.

When the TREATMENT*PERIOD interaction is significant, indicating the presence of carryover, a usual practice is to set aside the results of the second time period and analyze the first period only.

The Computer-Aided Analysis of Crossover Studies

Crossover designs are easily analyzed by any statistical program package, such as SAS (SAS Institute, Cary, NC) and SYSTAT (SPSS Inc., Chicago, IL), that can perform repeated measures analysis of variance.

Within each record, a subject's data can be ordered by either treatment or time period. Both arrangements for the data set given in Grizzle [1] are appended to the end of this note along with the appropriate SAS PROC GLM control language. For data ordered by TREATMENT, the test of treatments will be labelled TREATMENT, the test of treatment by period interaction (which is also the carryover effect) will be labelled GROUP, and the test of time periods will be labelled GROUP*TREATMENT, in keeping with the list of aliases developed earlier. For data ordered by PERIOD the test of treatments will be labelled GROUP*PERIOD, the test of treatment by period interaction will be labelled GROUP, and the test of time periods will be labelled PERIOD.

It might seem more natural to arrange the data by TREATMENT. This has the advantage of having the treatment comparison labelled TREATMENT. If the data are arranged by PERIOD, however, it is easier to analyze only the data from the first period data if a significant PERIOD*TREATMENT interaction is found.

SYSTAT command language is similar. The instructions for data ordered by TREATMENT are

```
CATEGORY GROUP
MODEL A B = CONSTANT + GROUP/REPEATED, NAME = "Treat"
ESTIMATE
```

The instructions for data ordered by PERIOD are

```
CATEGORY GROUP
MODEL P1 P2 = CONSTANT + GROUP/REPEATED, NAME = "Period"
ESTIMATE
```

Practical Issues

> "The intuitive appeal of having each subject serve as his or her own control has made the crossover study one of the most popular experimental strategies since the infancy of formal experimental design. Frequent misapplications of the design in clinical experiments, and frequent misanalyses of the data, motivated the Biometric and Epidemiological Methodology Advisory Committee to the U.S. Food and Drug Administration to recommend in June of 1977 that, in effect, the crossover design be avoided in comparative clinical studies except in the rarest instances." Fleiss [3, p. 263]

Despite the appeal of having each subject serve as his own control, crossover studies have substantial weaknesses, as well, even beyond the possibility of carryover effects mentioned earlier. Because subjects receive both treatments, crossover studies requires subjects to be available for twice as long as would be necessary for a parallel groups study and perhaps even longer, if a washout period is required between treatments. Acute problems might be gone before the second treatment is applied. A washout period between the two treatments might minimize the effects of the carryover, but this will not be feasible for treatments like fat soluble vitamin supplements that can persist in the body for months.

On the other hand, some features of the crossover may make the design preferable to a parallel groups study. In certain cases, volunteers might be willing to participate only if they receive a particular treatment. The crossover insures that each subject will receive both treatments.

References

1. Grizzle JE. The two-period change-over design and its use in clinical trials. Biometrics 21, 467-480 (1965).
2. Hills M, Armitage P. The two-period cross-over clinical trial. British Journal of Clinical Pharmacology 8, 7-20 (1979).
3. Fleiss JL. The Design and Analysis of Clinical Experiments. John Wiley & Sons, Inc., New York (1986).
4. Grizzle JE. Correction. Biometrics 30, 727 (1974).
5. Grieve AP. Correspondence: The two-period changeover design in clinical trials. Biometrics 38, 517 (1982).

6. Milliken GA, Johnson DE. Analysis of Messy Data. Van Nostrand Reinhold Co., New York (1984).

```
        Data from Grizzle [1] arranged for
         analysis by using SAS's PROC GLM

<data arranged                      <data arranged
    by treatment>                      by period>

 DATA;                               DATA;
   INPUT GROUP A B;                    INPUT GROUP P1 P2;
   CARDS;                              CARDS;
 1  0.2  1.0                         1  0.2  1.0
 1  0.0 -0.7                         1  0.0 -0.7
 1 -0.8  0.2                         1 -0.8  0.2
 1  0.6  1.1                         1  0.6  1.1
 1  0.3  0.4                         1  0.3  0.4
 1  1.5  1.2                         1  1.5  1.2
 2  0.9  1.3                         2  1.3  0.9
 2  1.0 -2.3                         2 -2.3  1.0
 2  0.6  0.0                         2  0.0  0.6
 2 -0.3 -0.8                         2 -0.8 -0.3
 2 -1.0 -0.4                         2 -0.4 -1.0
 2  1.7 -2.9                         2 -2.9  1.7
 2 -0.3 -1.9                         2 -1.9 -0.3
 2  0.9 -2.9                         2 -2.9  0.9
 ;                                   ;

 PROC GLM;                           PROC GLM;
   CLASS GROUP;                        CLASS GROUP;
   MODEL A B = GROUP/NOUNI;            MODEL P1 P2 = GROUP;
   REPEATED TREAT 2/SHORT;             REPEATED PERIOD 2/SHORT;
```

Last modified: undefined.

Logistic Regression

Gerard E. Dallal, Ph.D.

Prologue (feel free to skip it, but I can't suppress the urge to write it!)

From the statistican's technical standpoint, logistic regression is very different from linear least-squares regression. The underlying mathematics is different and the computational details are different. Unlike a linear least-squares regression equation which can be solved explicitly--that is, there is a formula for it--logistic regression equations are solved iteratively. A trial equation is fitted and tweaked over and over in order to improve the fit. Iterations stop when the improvement from one step to the next is suitably small.

Also, there are statistical arguments that lead to linear least squares regression. Among other situations, linear least squares regression is the thing to do when one asks for the best way to estimate the response from the predictor variables when they all have a joint multivariate normal distribution. There is no similar argument for logistic regression. In practice it often works, but there's nothing that says it has to.

Logistic Regression

From a practical standpoint, logistic regression and least squares regression are almost identical. Both methods produce prediction equations. In both cases the regression coefficients measure the predictive capability of the independent variables.

The response variable that characterizes logistic regression is what makes it special. With linear least squares regression, the response variable is a quantitative variable. With logistic regression, the response variable is an indicator of some characteristic, that is, a 0/1 variable. Logistic regression is used to determine whether other measurements are related to the presence of some characteristic--for example, whether certain blood measures are predictive of having a disease. If analysis of covariance can be said to be a t test adjusted for other variables, then logistic regression can be thought of as a chi-square test for homogeneity of proportions adjusted for other variables.

While the response variable in a logistic regression is a 0/1 variable, the logistic regression equation, which is a linear equation, does not predict the 0/1 variable itself. In fact, before the development of logistic regression in the 1970s, this is what was done under the name of *discriminant analysis*. A multiple linear least squares regression was fitted with a 0/1 variable as a response. The method fell out of favor because the discriminant function was not easy to interpret. The significance of the regression coefficients could be used to claim specific independent variables had predictive capability, but the coefficients themselves did not have a simple interpretation. In practice, a cutoff prediction value was determined. A case was classified as a 1 or a 0 depending on whether it's predicted value exceeded the cutoff. The predicted value could not be interpreted as a probability because it could be less than 0 or greater than 1.

Instead of classifying an observation into one group or the other, logistic regression predicts the probability that an indicator variable is equal to 1. To be precise, logistic regression equation does not directly predict the probability that the indicator is equal to 1. It predicts the log odds that an observation will have an indicator equal to 1. The odds of an event is defined as the ratio of the probability that an event occurs to the probability that it fails to occur. Thus,

Odds(indicator=1) = Pr(indicator=1) / [1 - Pr(indicator=1)]

or

Odds(indicator=1) = Pr(indicator=1) / Pr(indicator=0)

The log odds is just the (natural) logarithm of the odds.

Probabilities are constrained to lie between 0 and 1, with 1/2 as a neutral value for which both outcomes are equally likely. The constraints at 0 and 1 make it impossible to construct a linear equation for predicting probabilities.

Odds lie between 0 and $+\infty$, with 1 as a neutral value for which both outcomes are equally likely. Odds are asymmetric. When the roles of the two outcomes are switched, each value in the range 0 to 1 is transformed by taking its inverse (1/value) to a value in the range 1 to $+\infty$. For example, if the odds of having a low birthweight baby is 1/4, the odds of not having a low birthweight baby is 4/1.

Log odds are symmetric. They lie in the range $-\infty$ to $+\infty$. The value for which both outcomes are equally likely is 0. When the roles of the two outcomes are switched, the log odds are multiplied by -1, since log(a/b) = -log(b/a). For example, if the log odds of having a low birthweight baby are -1.39, the odds of not having a low birthweight baby are 1.39.

Those new to log odds can take comfort in knowing that as the probability of something increases, the odds and log odds increase, too. Talking about the behavior of the log odds an event is qualitatively the same thing as talking about the behavior of the probability of the event.

Because log odds take on any value between $-\infty$ and $+\infty$, the coefficients from a logistic regression equation can be interpreted in the usual way, namely, they represent the change in log odds of the response per unit change in the predictor.

Some detail...

Suppose we've fitted the logistic regression equation to a group of postmenopausal women, where Y=1 if a subject is osteoporotic and 0 otherwise, with the result

$$\log \text{ odds } (Y=1) = -4.353 + 0.038 \text{ age}$$

or

$$\log [\Pr(\text{osteo})/\Pr(\text{no osteo})] = -4.353 + 0.038 \text{ age}$$

Since the coefficient for AGE is positive, the log odds (and, therefore, the probability) of osteoporosis increases with age. Taking anti-logarithms of both sides gives

$$\Pr(\text{osteo})/\Pr(\text{no osteo}) = \exp(-4.353+ 0.038 \text{ age})$$

With a little manipulation, it becomes

$$\Pr(\text{osteo}) = \exp(-4.353 + 0.038 \text{ age}) / [1 + \exp(-4.353 + 0.038 \text{ age})]$$

or

$$\Pr(\text{osteo}) = 1 / \{1 + \exp[-(-4.353 + 0.038 \text{ age})]\}$$

This is an example of the general result that if

$$\log[P(Y=1)/P(Y=0)] = b_0 + b_1 X_1 + .. + b_p X_p$$

then

$$P(Y=1) = \exp(b_0 + b_1 X_1 + .. + b_p X_p)/[1 + \exp(b_0 + b_1 X_1 + .. + b_p X_p)]$$

or

$$P(Y=1) = 1/\{1 + \exp[-(b_0 + b_1 X_1 + .. + b_p X_p)]\}$$

Interpreting The Coefficients of a Logistic Regression Equation

If *b* is the logistic regression coefficient for AGE, then *exp(b)* is the odds ratio corresponding to a one unit change in age. For example for AGE=a,

$$\text{odds}(\text{osteo}|\text{AGE}=\text{a}) = \exp(-4.353 + 0.038 \text{ a})$$

while for AGE=a+1

$$\text{odds}(\text{osteo}|\text{age}=\text{a}+1) = \exp(-4.353 + 0.038 \text{ (a+1)})$$

Dividing one equation by the other gives

$$\frac{\text{odds}(\text{osteo}|\text{age}=\text{a}+1)}{\text{odds}(\text{osteo}|\text{age}=\text{a})} = \frac{\exp(-4.353 + 0.038 \text{ (a+1)})}{\exp(-4.353 + 0.038 \text{ a})}$$

or

$$\frac{\text{odds(osteo|age=a+1)}}{\text{odds(osteo|age=a)}} = \exp(0.038) \quad [= 1.0387]$$

which equals 1.0387. Thus, the odds that an older individual has osteoporosis increases 3.87% over that of a younger individual with each year of age. For a 10 year age difference, say, the increase is $\exp(b)^{10}$ [= 1.0387^{10}] = 1.46, or a 46% increase.

Virtually any sin that can be committed with least squares regression can be committed with logistic regression. These include stepwise procedures and arriving at a final model by looking at the data. All of the warnings and recommendations made for least squares regression apply to logistic regression as well.

[back to LHSP]

Last modified: undefined.

Degrees of Freedom

Gerard E. Dallal, Ph.D.

[Early draft subject to change.]

One of the questions an instrutor dreads most from a mathematically unsophisticated audience is, "What exactly is degrees of freedom?" It's not that there's no answer. The mathematical answer is a single phrase, "The rank of a quadratic form." The problem is translating that to an audience whose knowledge of mathematics does not extend beyond high school mathematics. It is one thing to say that degrees of freedom is an index and to describe how to calculate it for certain situations, but none of these pieces of information tells what degrees of freedom **means**.

As an alternative to "the rank of a quadratic form", I've always enjoyed Jack Good's 1973 article in the American Statistician "What are Degrees of Freedom?" 27, 227-228, in which he equates degrees of freedom to the difference in dimensionalities of parameter spaces. However, this is a partial answer. It explains what degrees of freedom is for many chi-square tests and the numerator degrees of freedom for F tests, but it doesn't do as well with t tests or the denominator degrees of freedom for F tests.

At the moment, I'm inclined to define **degrees of freedom** as **a way of keeping score.** A data set contains a number of observations, say, *n*. They constitute *n* individual pieces of information. These pieces of information can be used either to estimate parameters or variability. In general, each item being estimated costs one degree of freedom. The remaining degrees of freedom are used to estimate variability. All we have to do is count properly.

A single sample: There are *n* observations. There's one parameter (the mean) that needs to be estimated. That leaves *n-1* degrees of freedom for estimating variability.

Two samples: There are n_1+n_2 observations. There are two means to be estimated. That leaves n_1+n_2-2 degrees of freedom for estimating variability.

One-way ANOVA with *g* groups: There are $n_1+..+n_g$ observations. There are *g* means to be estimated. That leaves $n_1+..+n_g-g$ degrees of freedom for estimating variability. This accounts for the denominator degrees of freedom for the F statistic.

The primary null hypothesis being tested by one-way ANOVA is that the g population means are equal. The null hypothesis is that there is a single mean. The alternative hypothesis is that there are g individual means. Therefore, there are g-*1*--that is g *(H_1)* minus *1* (H_0)--degrees of freedom for testing the null hypothesis. This accounts for the numerator degrees of freedom for the F ratio.

There is another way of viewing the numerator degrees of freedom for the F ratio. The null hypothesis says there is no variability in the g population means. There are g sample means. Therefore, there are g-*1* degrees of freedom for assessing variability among the g means.

Multiple regression with *p* predictors: There are *n* observations with *p+1* parameters to be estimated--one regression coeffient for each of the predictors plus the intercept. This leaves *n-p-1* degrees of freedom for error, which accounts for the error degrees of freedom in the ANOVA table.

The null hypothesis tested in the ANOVA table is that all of coefficients of the predictors are 0. The null hypothesis is that there are no coefficients to be estimated. The alternative hypothesis is that there are *p* coefficients to be estimated. herefore, there are *p-0* or *p* degrees of freedom for testing the null hypothesis. This accounts for the Regression degrees of freedom in the ANOVA table.

There is another way of viewing the Regression degrees of freedom. The null hypothesis says the expected response is the same for all values of the predictors. Therefore there is one parameter to estimate--the common response. The alternative hypothesis specifies a model with *p+1* parameters--*p* regression coefficients plus an intercept. Therefore, there are *p*--that is *p+1* (H_1) minus *1* (H_0)--regression degrees of freedom for testing the null hypothesis.

Okay, so where's the quadratic form? Let's look at the variance of a single sample. If *y* is an n by 1 vector of observations, then

$$\sum (y_i - \overline{y})^2 = \underset{\sim}{y}'M\underset{\sim}{y}, \text{ where } M = \begin{pmatrix} 1-\frac{1}{n} & -1/n & . & -1/n \\ -1/n & 1-\frac{1}{n} & . & -1/n \\ . & . & . & . \\ -1/n & -1/n & . & 1-\frac{1}{n} \end{pmatrix}$$

The number of degrees of freedom is equal to the rank of the n by n matrix *M*, which is n-1.

[back to The Little Handbook of Statistical Practice]

Last modified: undefined.

British Medical Journal: Statistics Notes

Perhaps the finest series of short articles on the use of statistics is the occasional series of Statistics Notes started in 1994 by the British Medical Journal. It should be required reading in any introductory statistics course. The full text of all but the first ten articles is available is available on the World Wide Web. The articles are listed here chronologically.

Absence of evidence is not evidence of absence is something every investigator should know, but too few do. Along with *Interaction 2: compare effect sizes not P values*, these articles describe two of the most common fatal mistakes in manuscripts submitted to research journals. The faulty reasoning leading to these errors is so seductive that papers containing these errors sometimes slip through the reviewing process and misinterpretations of data are published as fact.

Correlation, regression, and repeated data, *Calculating correlation coefficients with repeated observations: Part 1--correlation within subjects*, and *Calculating correlation coefficients with repeated observations: Part 2--correlation between subjects* provide an excellent introduction to the subtleties of analyzing repeated measurements on the same subject.

- Correlation, regression, and repeated data J Martin Bland & Douglas G Altman BMJ 1994;308:896 (2 April)
- Regression towards the mean J Martin Bland & Douglas G Altman BMJ 1994;308:1499 (4 June)
- Diagnostic tests 1: sensitivity and specificity Douglas G Altman & J Martin Bland BMJ 1994;308:1552 (11 June)
- Diagnostic tests 2: predictive values Douglas G Altman & J Martin Bland BMJ 1994;309:102 (9 July)
- Diagnostic tests 3: receiver operating characteristic plots Douglas G Altman & J Martin Bland BMJ 1994;309:188 (16 July)
- One and two sided tests of significance J Martin Bland & Douglas G Altman BMJ 1994;309:248 (23 July)
- Some examples of regression towards the mean J Martin Bland & Douglas G Altman BMJ 1994;309:780 (24 September)
- Quartiles, quintiles, centiles, and other quantiles Douglas G Altman & J Martin Bland BMJ 1994;309:996 (15 October)
- Matching J Martin Bland & Douglas G Altman BMJ 1994;309:1128 (29 October)
- Multiple significance tests: the Bonferroni method J Martin Bland & Douglas G Altman BMJ 1995;310:170 (21 January)
- The normal distribution Douglas G Altman & J Martin Bland BMJ 1995;310:298 (4 February)
- Calculating correlation coefficients with repeated observations: Part 1--correlation within subjects J Martin Bland & Douglas G Altman BMJ 1995;310:446 (18 February)
- Calculating correlation coefficients with repeated observations: Part 2--correlation between

subjects J Martin Bland & Douglas G Altman BMJ 1995;310:633 (11 March)

- Absence of evidence is not evidence of absence Douglas G Altman & J Martin Bland BMJ 1995;311:485 (19 August)
- Presentation of numerical data J Martin Bland & Douglas G Altman BMJ 1996;312:572 (2 March)
- Logarithms J Martin Bland & Douglas G Altman BMJ 1996;312:700 (16 March)
- Transforming data J Martin Bland & Douglas G Altman BMJ 1996;312:770 (23 March)
- Transformations, means, and confidence intervals J Martin Bland & Douglas G Altman BMJ 1996;312:1079 (27 April)
- The use of transformation when comparing two means J Martin Bland & Douglas G Altman BMJ 1996;312:1153 (4 May)
- Comparing several groups using analysis of variance Douglas G Altman & J Martin Bland BMJ 1996;312:1472-1473 (8 June)
- Measurement error and correlation coefficients J Martin Bland & Douglas G Altman BMJ 1996;313:41-42 (6 July)
- Measurement error proportional to the mean J Martin Bland & Douglas G Altman BMJ 1996;313:106 (13 July)
- Interaction 1: heterogeneity of effects Douglas G Altman & John NS Matthews BMJ 1996;313:486 (24 August)
- Measurement error J Martin Bland & Douglas G Altman BMJ 1996;313:744 (21 September)
- Interaction 2: compare effect sizes not P values John NS Matthews & Douglas G Altman BMJ 1996;313:808 (28 September)
- Interaction 3: How to examine heterogeneity John NS Matthews & Douglas G Altman BMJ 1996;313:862 (5 October)
- Detecting skewness from summary information Douglas G Altman & J Martin Bland BMJ 1996;313:1200 (9 November)
- Cronbach's alpha J Martin Bland & Douglas G Altman BMJ 1997;314:572 (22 February)
- Units of analysis Douglas G Altman & J Martin Bland BMJ 1997;314:1874 (28 June)
- Trials randomised in clusters J Martin Bland & Sally M Kerry BMJ 1997;315:600 (6 September)
- Analysis of a trial randomised in clusters Sally M Kerry & Martin Bland BMJ 1998;316:54
- Weighted comparison of means Martin Bland & Sally M Kerry BMJ 1998;316:129 (10 January)
- Sample size in cluster randomisation Sally M Kerry & J Martin Bland BMJ 1998;316:549 (14 February)
- The intracluster correlation coefficient in cluster randomisation Sally M Kerry & J Martin Bland BMJ 1998;316:1455-1460 (9 May)
- Generalisation and extrapolation Douglas G Altman & J Martin Bland BMJ 1998;317:409-410 (8 August)
- Time to event (survival) data Douglas G Altman & J Martin Bland BMJ 1998;317:468-469 (15 August)
- Bayesians and frequentists J Martin Bland & Douglas G Altman BMJ 1998;317:1151-1160 (24

October)

- Survival probabilities (the Kaplan-Meier method) J Martin Bland, & Douglas G Altman BMJ 1998;317:1572-1580 (5 December)
- Treatment allocation in controlled trials: why randomise? Douglas G Altman & J Martin Bland BMJ 1999;318:1209-1209 (1 May)
- Variables and parameters Douglas G Altman & J Martin Bland BMJ 1999;318:1667-1667 (19 June)
- How to randomise Douglas G Altman & J Martin Bland BMJ 1999;319:703-704 (11 September)
- The odds ratio Douglas G Altman & J Martin Bland BMJ 2000;320:1468 (27 May)
- Blinding in clinical trials and other studies Simon J Day & Douglas G Altman BMJ 2000;321:504 (19 August)
- Concealing treatment allocation in randomised trials Douglas G Altman & Kenneth F Schulz BMJ 2001;323:446-447 (25 August)
- Analysing controlled trials with baseline and follow up measurements Andrew J Vickers & Douglas G Altman BMJ 2001;323:1123-1124 (10 November)
- Validating scales and indexes J Martin Bland & Douglas G Altman BMJ 2002;324:606-607 (9 March)
- Interaction revisited: the difference between two estimates Douglas G Altman & J Martin Bland BMJ 2003;326:219 (25 January)
- The logrank test J Martin Bland & Douglas G Altman BMJ 2004 328(7447):1073 (1 May)
- Diagnostic tests 4: likelihood ratios. JJ Deeks & Douglas G Altman BMJ 2004 329:168-169

[back to LHSP]

Last modified: undefined.

HyperStat Online Textbook

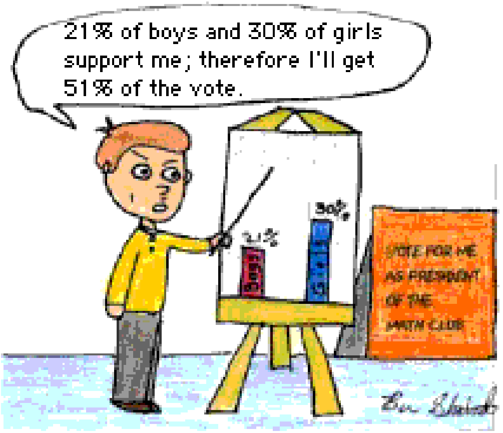
RVLS Home | Glossary | Search | **Free Statistical Analysis Tools** | Instructional Demos | Exercises and Problems | Statistics Help

Email me at dmlane@davidmlane.com
David Lane is an Associate Professor of Psychology, Statistics, and Management at Rice University

statistics solutions
WE PROVIDE RESEARCH AND STATISTICAL ASSISTANCE TO COMPLETE YOUR DEGREE ON SCHEDULE
the country's leader in statistical counsulting
statistics solutions

College Textbooks

Related Material


21% of boys and 30% of girls support me; therefore I'll get 51% of the vote.
21%
30%
Boys
Girls
VOTE FOR ME AS PRESIDENT OF THE MATH CLUB
Ben Shabad

Click here for more cartoons by Ben Shabad.

Follow this link to more statistics humor sites.

Other Sources

The little handbook of statistical practice by Gerard E. Dallal

Visual statistics with multimedia by David J. Krus of Arizona State University

Statistics at square 1 by T. D. V. Swinscow; revised by M. J. Campbell, University of Southampton

Statnotes: An online textbook by G. David Garson, North Carolina State University

Concepts and applications of inferential statistics by Richard Lowry of Vassar College

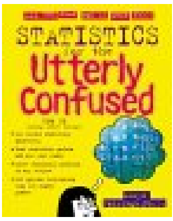
StatPrimer by B. Gerstman

Books/ Music

Visit the HyperStat Bookstore for a listing of **top-rated statistics books**.

Statistics for the Utterly Confused by Lloyd Jaisingh

A good introductory statistics book, somewhat more elementary than HyperStat.



Computer Software Price Comparison

Printed Version
Order from Amazon

Please help some students of mine by filling out an online questionnaire about computer use. Click here to learn more.

Statistics Help
I've compiled a list of Statistical Consultants and Statistics Tutors for help with statistical projects or simply for help with homework.

Online Courses
Find Online e-learning Statistics College Courses for distance learning!
Distance Learning Resources in Statistics
Online Short Courses at Statistics.com

Free stuff
to help students learn statistics

Play Chess? See 49 of the greatest combinations of the world's best players .

Permission to link
Feel free to link to any portion of HyperStat Online from your web site.

Visit the Books for Managers Site!

Interested in entrepreneurship? Follow this link to a site for entrepreneurs, entrepreneurship professors, and students taking entrepreneurship classes.

Personal Training in Southern California

SticiGui by P. B. Stark of UC Berkeley

Investigating Statistics by Robert Hale of Pennsylvania State University

SurfStat by Keith Dear of the University of Newcastle.

Statistics for journalists by Robert Niles of the LA Times.

Introductory statistics: Concepts, models, and applications by David W. Stockburger of Southwest Missouri State University

Multivariate statistics: Concepts, models, and applications by David W. Stockburger of Southwest Missouri State University

Electronic textbook by StatSoft

A new view of statistics by Will Hopkins of the University of Otago

Introduction to data collection and analysis by Albert Goodman of Deakin University

The knowledge base: An online research methods textbook by William M. Trochim of Cornell University

Statistics 30X class notes by H. J. Newton, J. H. Carroll, N. Wang, and D. Whiting of Texas A&M.

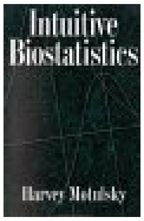
DAU stat refresher by A. Rao, P. McAndrews, A Sunkara, V. Patil, R. Bellary, G. Quisumbing, H. Le, Z. Zhou from Defense Acquisition University

The Visual Display of Quantitative Information by Edward R. Tufte

This book is "must" reading for anyone creating or interpreting graphs. Beautifully done.

Intuitive Biostatistics by Harvey Motulsky

An excellent and easy to understand introduction to biostatistics.



Feedback
Please let me know of any errors you discover and/or any sections you find confusing. Please e-mail me to send feedback.

Statistics and data sources

Web Awards

Click here for help with non-statistical problems.

Looking for a house in San Diego? Click here.

Made with Macintosh

Stat notes: An Online Textbook, by G. David Garson of North Carolina State University

BB&N AP Statistics
by teacher and students of the past and present BB&N AP Statistics classes.

math+blues.com

Humor

For the lighter side of statistics visit Gary Ramseyer's excellent Gallery of Statistics Jokes

Statistics ?! Don't make me laugh!
Compiled by Jill Binker

Statistical Quotes and Song Lyrics
by Tom Short

Using Humor in the Introductory Statistics Course
Hershey H. Friedman, Linda W. Friedman, and Taiwo Amoo

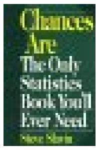
Using Humor to Teach Statistics : Must They Be Orthogonal?
by Richard G. Lomax and Seyed A. Moosavi

Profession Jokes: Statistics
by David Shay

The Canonical List of Math Jokes: Statistics and Statisticians

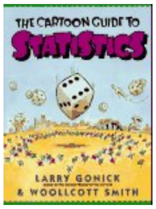
Chances Are : The Only Statistics Book You'll Ever Need
by Stephen Slavin

A good non-mathematical introduction to statistics.



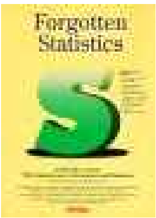
The Cartoon Guide to Statistics
by Larry Gonick and Woollcott Smith

A humorous and easy-to-understand supplement to a textbook on statistics.



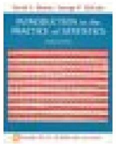
Forgotten Statistics : A Self-Teaching Refresher Course
by Jeffrey Clark

An excellent book for the professional who needs to brush up on statistics and as a supplement to the textbook in a college course.



Introduction to the Practice of Statistics (3rd Ed)
by David S. Moore and George P. McCabe

This textbook sets the standard for introductory statistics books. Extremely well written with lots of examples and exercises. Used frequently in college courses and AP statistics courses.



Last updated: NaN/NaN/ NaN

You have reached
Jerry Dallal's World Home Page
a small cul-de-sac on the Information Superhighway.

Small though it is, the time is fast approaching for some major repaving! There's stuff scattered all over the World Wide Web. It isn't well organized or indexed. It's just sort of... there. In the broadest sense, it divides itself into two groups

- My professional activities
  - Notes for teaching statistical methods
  - Statistical software
  - Randomization plan generator
  - My Erdös number
- My personal activities
  - Ukes, anyone?
  - Generate ukulele chord fingerings and diagrams
  - Generate guitar or mandolin tablature
  - The songs of Charlie Poole
  - Other stuff

Want to write? 'Course you do! gdallal@world.std.com
Last modified: undefined.