

Расчёты и вопросы по индексу сходства Жаккара

I. Чем не устраивает пакет jaccard

Идея и практический успех использования индекса Жаккара J при выявлении ассоциаций организмов в экологии связан с отказом от использования в расчётах ячейки D , когда оба вида не обнаруживаются в местообитании или пробе. В отличие от многих мер сходства это исключает ситуацию, что виды окажутся сходными не столько за счёт совместной встречаемости, сколько за счёт просто отсутствия в пробах.

		Вид b	
		+	-
Вид a	+	A	B
	-	C	D

Т.е. J – это отношение числа случаев совместного обнаружения видов к числу случаев, когда обнаруживался хотя бы один из них:

$$J = \frac{A}{A + B + C}$$
. Изменяется от 0 (нет совстречаемости) до 1 (виды встречаются только вместе).

Рассмотрим такой пример:

		Вид b	
		+	-
Вид a	+	5	1
	-	1	0

$$J = \frac{5}{5 + 1 + 1} = 0,7142857$$
. Ассоциация достаточно сильная, но незначимая:

```
library(jaccard)
a<-c(1,1,1,1,1,0,1)
b<-c(1,1,1,1,1,1,0)
jaccard(a,b)
[1] 0.7142857
jaccard.test.exact(a,b)
$pvalue
[1] 0.4210541
```

Теперь будем наращивать число проб в ячейке D .

```
a<-c(1,1,1,1,1,0,1,0)
b<-c(1,1,1,1,1,1,0,0)
jaccard(a,b)
[1] 0.7142857
```

```

jaccard.test.exact(a,b)
$pvalue
[1] 0.2799831
a<-c(1,1,1,1,1,0,1,0,0)
b<-c(1,1,1,1,1,1,0,0,0)
jaccard(a,b)
[1] 0.7142857
> jaccard.test.exact(a,b)
$pvalue
[1] 0.08125382

```

```

a<-c(1,1,1,1,1,0,1,0,0,0)
b<-c(1,1,1,1,1,1,0,0,0,0)
jaccard(a,b)
[1] 0.7142857
jaccard.test.exact(a,b)
$pvalue
[1] 0.03601733

```

Таким образом, по мере увеличения числа случаев в ячейке D , величина индекса Жаккара закономерно не изменяется. Однако по мере роста объёма выборки за счёт этой ненужной ячейки величина P -значения снижается. И если в первом случае J незначим ($P=0,421$), то в последнем – значим ($P=0,036$). Получается, что на значимость влияет именно та ячейка, которую экологи не хотят брать в расчёт, что концептуально полностью противоречит самой идее данного индекса. Следовательно то, как оценивает значимость пакет jaccard (причём всеми 4-мя способами, а не только продемонстрированным exact) является оценкой вероятности случайности ассоциации, **но не является оценкой значимости собственно индекса Жаккара.**

Логично было бы если P уменьшалось по мере роста объёмов выборки без ячейки D , как для разных коэффициентов корреляции и ассоциации. Удвоим выборку:

		Вид b	
		+	–
Вид a	+	10	2
	–	2	0

$$J = \frac{10}{10 + 2 + 2} = 0,7142857.$$

```

a<-c(1,1,1,1,1,0,1,1,1,1,1,0,1)
b<-c(1,1,1,1,1,1,0,1,1,1,1,1,0)
jaccard(a,b)
[1] 0.7142857
jaccard.test.exact(a,b)
$pvalue
[1] 0.4221532

```

Учетверим выборку:

```
a<-c(1,1,1,1,1,0,1,1,1,1,1,1,0,1,1,1,1,1,1,0,1,1,1,1,1,0,1)
```

```
b<-c(1,1,1,1,1,1,0,1,1,1,1,1,1,0,1,1,1,1,1,1,0,1,1,1,1,1,1,0)
```

```
jaccard(a,b)
```

```
[1] 0.7142857
```

```
jaccard.test.exact(a,b)
```

```
$pvalue
```

```
[1] 0.3611397
```

Коэффициент закономерно остаётся тем же, но и P почти не меняется. Эта ситуация противоречит логике статистического оценивания: увеличение объёмов выборок должно увеличивать надёжность оценки.

Вывод. В целом, то что считает пакет не имеет отношения к индексу сходства Жаккара. Те вероятностные оценки которые он выдаёт и концептуально и численно близки к значимости индекса Раупа – Крика (Raup-Crick metric).

II. А мне нужна оценка значимости именно индекса Жаккара.

Нашёл такую статью (прикрепил в форум),

Syst. Biol. 45(3):380-385, 1996

The Probabilistic Basis of Jaccard's Index of Similarity

RAIMUNDO REAL AND JUAN M. VARGAS

*Department of Animal Biology, Faculty of Science, University of Málaga, Málaga 29071, Spain;
E-mail: rrgimenez@ccuma.uma.es (R.R.)*

Но не получается с ней разобраться. Нужно понять формулу (17), чтобы затем перенести её в R.

$$P = 1 - \frac{\sum_{x=0}^{C-1} \binom{N}{x} VR, 2N - x}{VR3, N}$$

Вопросы:

1) $VR, 2N - x = 2^{N-x}$?

2) $VR3, N = 3^N$?

3) Тогда формулу перепишем по-человечески как:

$$P = 1 - \frac{\sum_{x=0}^{C-1} \binom{N}{x} 2^{N-x}}{3^N} ?$$

4) Чему в формуле равно N ? Из статьи выходит, что оно равно для примера $(A+B+C+D)=N$, хотя по идее D не должно фигурировать в расчётах, раз эта

ячейка не задействуется в расчёте J . Если же N считается так, то она неявным образом присутствует и чем больше будет D , тем больше и N .

По терминологии статьи:

		Вид 2	
		+	-
Вид 1	+	C	A
	-	B	D

Т.е. C в формуле число проб с обоими видами.