УДК 681.5(075.8)

ГЕРМЕНЕВТИКА НЕПАРАМЕТРИЧЕСКОГО КРИТЕРИЯ АНСАРИ-БРЭДЛИ СРЕДСТВАМИ MATHCAD

© А. Н. Порунов

Самарский государственный технический университет, филиал в г. Сызрани
ул. Советская, 45, г. Сызрань, Россия
e-mail: *rameno@rambler.ru*

Abstract. Procedure is shown hermeneutics nonparametric Ansari-Bradley test by facilities of Mathcad. Hermeneutics procedure is built on the analysis of typical methodical errors of calculation of criterion. The similar technique of penetration in essence of statistical criterion is in an equal degree useful as for beginners so for more experience researchers.

Введение

Сегодня, перефразируя Кевина Аэрна [1], в эпоху «карнавала» статистических пакетов, от пользователя требуется наличие профессиональных навыков и высокой квалификации, широкого первоначального статистического образования. Эти неотъемлемые условия получения корректных, надежных статистических оценок в любом исследовании предполагают глубокое проникновение в суть статистических методов. В прежние времена, «глубокое проникновение» достигалось с помощью обычного учебника, ручки и листа бумаги. Появление мощных математических пакетов, позволяющих не только решать статистические задачи, но и детально изучать, «препарировать» методы ее решения привело к резкому повышению качества и эффективности образовательного процесса. В области статистики сегодня они определяют «дидактический мейнстрим». Речь идет, прежде всего, о таких универсальных математических программных продуктах, как Maple, Mathcad, Mathematica, MATLAB и специализированных типа STATISTICA, SPSS, STATGRAPHICS. Каждый их них по своему хорош, но, практика показывает, что для образовательных и самообразовательных целей в области статистики наибольшее распространение получил Mathcad, поскольку обладает двумя весьма ценными с точки зрения освоения статистических методов характеристиками, — универсальностью и б*о*льшей наглядностью. Как отмечают многие специалисты главное преимущество Mathcad в отличие от аналогичных программных средств, состоит в том, что здесь математические выражения представляются в общепринятой математической нотации. Запись на языке, очень близком к стандартному языку математических расчетов, упрощает постановку и решение многих задач и делает этот продукт наиболее востребованным в самых различных сферах деятельности [2].

Представленная статья иллюстрирует «герменевтику» непараметрического критерия Ансари-Брэдли (AB) средствами Mathcad. Герменевтическая процедура построена на разборе типичных методических ошибок исчисления критерия. Представляется, что подобная техника проникновения в суть статистического критерия в равной степени полезна как для начинающих исследователей так и для более опытных, но еще не знающих, в отличие от А. Лимера, «обратную сторону» таких оценок.

1. Постановка задачи в общем виде

Пусть имеются (здесь и далее автор придерживается синтаксиса принятого в среде Mathcad) две независимые случайные выборки:

$$X_1, X_2, ...X_m \in F(t) \quad \text{и} \quad Y_1, Y_2, ...Y_n \in F(t/\Delta).$$

Решаемые задачи: проверка гипотезы о том, что:

- $H_0 : \Delta = 1$ (дисперсии выборок равны);
- $H_1 : \Delta < 1$ (дисперсия второй выборки больше чем первой);
- $H_2 : \Delta > 1$ (дисперсия первой выборки больше чем второй);
- $H_3 : \Delta \neq 1$ (дисперсии не равны).

Предположения:

1) выборки X и Y должны быть независимы;
2) распределение $F(t)$ — непрерывное, параметры распределения неизвестны;
3) медианы выборок должны быть равны, т.е.:
   $\mu_1 := median(X) \cong \mu_2 := median(Y)$, если $\mu_1 \neq \mu_2$ следует рассмотреть новые выборки $\tilde{X}$ и $\tilde{Y}$ такие что:

$$\tilde{X}, \tilde{Y} := \begin{vmatrix} X - \mu_1 \\ Y - \mu_2 \end{vmatrix}.$$

2. Методика и методические ошибки

Оставаясь в среде Mathcad, рассмотрим на конкретном примере методику и типичные методические ошибки расчета AB-критерия.

В качестве входных данных примера используются результаты исследований компании, оценивающей релевантность показателей отечественных информационно поисковых систем (ИПС). В частности, сравнивались значения двадцати пар измерений F-меры (меры Ван Ризбергена) двух ИПС Rambler и Yandex в случайно выбранные дни марта 2012 года. Эти показатели были получены на основе анализа запроса на слово “cat” и по тому, как много релевантных ссылок выдаёт поисковик

на 100 первых ответов [3]. Основная и альтернативная гипотезы формулировались следующим образом:

H_0: дисперсии F-меры одинаковы ($\Delta = 1$),

H_1: дисперсии F-меры не равны ($\Delta \neq 1$).

Таким образом, имеем две выборки значений F-меры, одну выборку назовем “Rambler”, другую “Yandex”. Данные 40 испытаний по оценке релевантности поисковых систем приведены в табл. 1.

Таблица 1. Значения показателя F-мера выборок “Rambler” и “Yandex”

i	**Rambler**	**Yandex**	**i**	**Rambler**	**Yandex**
1	0.854	0.808	11	0.777	0.877
2	0.823	0.792	12	0.738	0.877
3	0.769	0.846	13	0.746	0.869
4	0.762	0.808	14	0.785	0.831
5	0.785	0.800	15	0.823	0.815
6	0.815	0.769	16	0.869	0.877
7	0.838	0.738	17	0.892	0.869
8	0.831	0.831	18	0.869	0.831
9	0.800	0.792	19	0.846	0.815
10	0.762	0.800	20	0.754	0.762

В векторном виде:

$$Rambler^T := (0.845\ 0.823\ 0.769\ 0.762\ 0.785\ 0.815\ 0.838\ 0.831\ 0.800\ 0.762$$
$$0.777\ 0.738\ 0.746\ 0.785\ 0.823\ 0.869\ 0.892\ 0.869\ 0.846\ 0.754)$$

Вектор $Rambler$ состоит из $n := rows(Rambler) = 20$ наблюдений, с номерами $i := 0..last(Rambler)$.

$$Yandex^T :\ = (0.808\ 0.792\ 0.846\ 0.808\ 0.800\ 0.769\ 0.738\ 0.831\ 0.792\ 0.800\ 0.877$$
$$0.877\ 0.869\ 0.831\ 0.815\ 0.877\ 0.869\ 0.831\ 0.815\ 0.762).$$

Вектор $Yandex$ состоит из $m := rows(Yandex) = 20$ наблюдений с номерами $j := 0..last(Yandex)$.

Медианы выборок Rambler и Yandex соответственно равны:

$$\mu_1 := median(Rambler) = 0.807 \quad \text{и} \quad \mu_2 := median(Yandex) = 0.815.$$

Первая ошибка чаще всего связана с игнорированием требования к медианам выборок. Равенство медиан является необходимым условием АВ-критерия это известно еще со времен Л. Э. Мозеса [4], показавшего насколько требовательны ранговые критерии к равенству медиан или знанию их величин. Тем не менее, даже в учебной литературе, иногда встречаются попытки принизить значимость этого требования [5]. Поскольку в нашем случае медианы не равны, воспользуемся приемом, который был предложен А. Р. Ансари и Р. А. Брэдли несколько позже того как они представили миру одноименный критерий.

Переопределим выборки:

$$Rambler := Rambler - median(Rambler) \tag{1}$$

и

$$Yandex := Yandex - median(Yandex). \tag{2}$$

Графики распределения преобразованных наблюдений в выборках показаны на рис. 1.

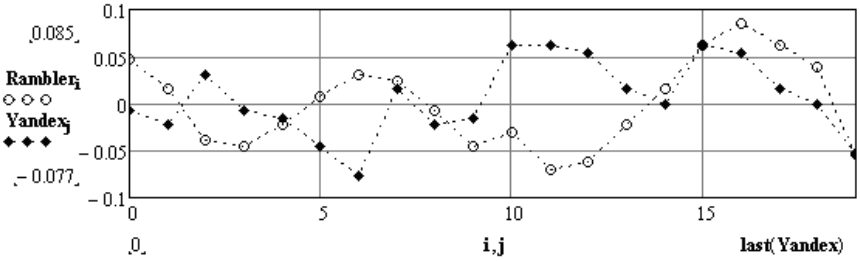

Рис. 1. Распределения F-меры в исходных выборках Rambler и Yandex.

Для проверки основной (нулевой) гипотезы H_0 необходимо:

Этап 1. Объединение выборок и ранжирование наблюдений в порядке возрастания. С этой целью определим вектор RY суть которого — объединенные и отсортированные элементы векторов Rambler и Yandex (рис. 2):

$$RY := sort(stack(submatrix(Rambler, 0, last(Rambler), 0, 0), Yandex)) \tag{3}$$

Этап 2. Присвоение рангов наблюдениям ранжированной объединенной выборки RY производится с концов совокупности по направлению к её медиане так, что

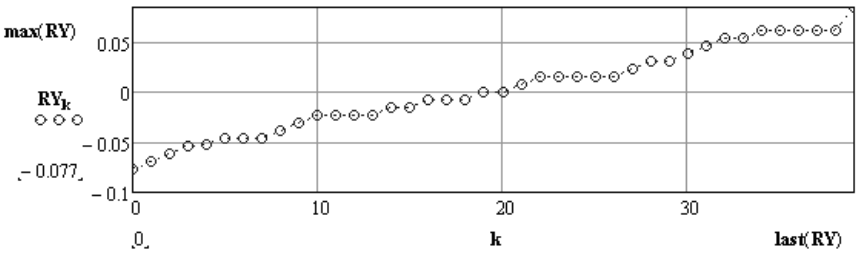

Рис. 2. Распределение F-меры в объединенной выборке RY

образуется симметричный ряд вида:

$$0, 1, 2, \ldots, \frac{N-1}{2}, \frac{N+1}{2}, \frac{N-1}{2}, \ldots, 2, 1, 0 \quad \text{если } N \text{ нечетно}, \tag{4}$$

или

$$0, 1, 2, \ldots, \frac{N}{2}, \frac{N}{2}, \ldots, 2, 1, 0 \quad \text{если } N \text{ четно}. \tag{5}$$

Казалось бы очевидный алгоритм, тем не менее, и на этом этапе нередки ошибки. Связаны они, по-видимому, с «трудностями перевода» статистической литературы. Например, в одной из самых авторитетных и наиболее часто цитируемых работ, находим:

> «Наименьшему и наибольшему из наблюдений в объединенной выборке присвоить ранг 1, следующим среди наименьших и наибольших присвоить ранг 2 и продолжать ранжирование тем же способом.» [6, с. 102].

Такая, не самая удачная версия перевода создает у начинающего исследователя иллюзию симметричности распределения рангов. Возникает соблазн использовать (зеркально отразить) ранги наблюдений левой, от медианы, половины объединенной выборки, для построения рангов её правой половины или наоборот. Как следствие, — неверное распределение рангов. Между тем, как и в первом [7] так и во втором [8] оригинальных изданиях этой книги речь идет о независимом присвоении рангов наблюдениям правой и левой половинам вариационного ряда.

Требования второго этапа в Mathcad можно реализовать следующим образом:

а) разделить вектор RY на две составляющие RY1 и RY2 (равные, если N чётно):

$$RY1 := submatrix\left(RY, floor\left(\frac{last(RY)}{2}, 0, 0\right)\right) \tag{6}$$

$$RY2 := submatrix\left(RY, ceil\left(\frac{last(RY)}{2}, last(RY), 0, 0\right)\right) \tag{7}$$

б) введем новую ранжированную переменную $h := RY_0, \ldots, RY_{last(RY)}$ и, в программном модуле (8), используя mathcad-функцию $Rank(v)$, определим симметричную (относительно медианы вектора RY) функцию средних рангов $rr(h)$. Значения этой функции, аргументами которой являются элементы переопределенного вектора Rambler, приведены в табл. 2.

$$rr(h) := \left| \begin{array}{l} for\ e \in 0..last(RY2) \\ \quad d \leftarrow Rank(-RY2)_e \quad if\ h = RY2_e \\ for\ e \in 0..last(RY1) \\ \quad d \leftarrow Rank(RY1)_e \quad if\ h = RY1_e \end{array} \right. \tag{8}$$

Таблица 2. Переопределенные значения и средние ранги элементов вектора Rambler.

i	$\mathbf{Rambler_i}$	$\mathbf{rr(Rambler_i)}$	**i**	$\mathbf{Rambler_i}$	$\mathbf{rr(Rambler_i)}$
0	0.046	9	10	-0.03	10
1	0.015	17.5	11	-0.07	2
2	-0.038	9	12	-0.062	3
3	-0.045	7.5	13	-0.022	13.5
4	-0.022	13.5	14	0.015	17.5
5	0.007	19	15	0.062	5.5
6	0.03	12	16	0.085	1
7	0.023	13	17	0.062	5.5
8	-0.007	17	18	0.038	10
9	-0.045	7.5	19	-0.054	4

Статистика AB-критерия, — сумма средних рангов элементов вектора Rambler, равна:

$$W_{ab} := \sum_{i=0}^{last(Rambler)} rr(Rambler_i) = 197. \tag{9}$$

Этап 3. Большая выборка. Центрирование и нормирование AB-статистики.

Поскольку таблицы квантилей распределения W_{ab} - статистики доступны только для случаев, когда $m, n < 10$, то при больших выборках (как в нашем случае), с целью приближения распределения W_{ab} - статистики к стандартному, нормальному распределению обычно производят операции её трансформации (центрирования и

нормирования) по известной схеме:

$$M(W) := \left| \begin{array}{l} M(W) \leftarrow \dfrac{m \cdot (m+n+1)^2}{4 \cdot (m+n)} \\ D(W) \leftarrow \dfrac{m \cdot n \cdot (m+n+1) \cdot [3+(m+n)]}{4 \cdot (m+n)} \\ \dfrac{W_{ab} - M(W)}{\sqrt{D(W)}} \quad if\ mod(N,2) \neq 0 \\ otherwise \\ \quad \left| \begin{array}{l} M(W) \leftarrow \dfrac{m \cdot (m+n+2)^2}{4} \\ D(W) \leftarrow \dfrac{m \cdot n \cdot (m+n+2) \cdot (m+n+2)}{48 \cdot (m+n-1)^2} \\ \dfrac{W_{ab} - M(W)}{\sqrt{D(W)}} \end{array} \right. \end{array} \right. \tag{10}$$

где: $M(W)$ и $D(W)$ — соответственно, математическое ожидание и дисперсия W_{ab}-статистики. Одна из часто встречающихся ошибок на данном этапе расчета АВ-критерия, — не учёт поправки на связанные ранги к дисперсии $D(W)$. Подобная практика отчасти объясняется тем, что выходящие сегодня в «свет» отечественные учебники и справочники по прикладной математической статистике, например [9], часто не дают на сей счёт никаких разъяснений. Между тем в данном случае, функция $rr(Rambler_i)$ (табл. 1) имеет дробные значения рангов, что говорит о необходимости расчета дисперсии W_{ab}-статистики с учетом поправки на связи.

Этап 4. Определение связей и числа ранговых групп. Типичная ошибка для этой части расчета АВ -критерия, - неправильное определение числа ранговых групп наблюдений в ранжированной объединенной выборке. Ошибка возникает по причине того, что в состав ранговых групп включают только группы связанных рангов, группы же несвязанных рангов остаются без внимания.

Используя определенную нами ранее (8) функцию $rr(h)$, возвращающую усредненное положение каждого элемента в объединенной выборке-векторе RY, определим число ранговых групп g среди $N = 40$ наблюдений, объём t_g g-й ранговой группы и средний ранг наблюдений r_g в g-й группе с помощью программы:

$$\left(\begin{array}{c} g \\ t \\ r \end{array}\right) := \left|\begin{array}{l} a \leftarrow 0 \\ g \leftarrow 0 \\ t_g \leftarrow 0 \\ for\ i \in 0..last(RY) \\ \quad \left|\begin{array}{l} a \leftarrow a+1 \quad if\ \ i > 0 \wedge rr(RY_i = rr(RY_{i-1}) \\ if\ \ i > 0 \wedge rr(RY_i \neq rr(RY_{i-1}) \\ \quad \left|\begin{array}{l} t_g \leftarrow a+1 \\ g \leftarrow g+1 \\ a \leftarrow 0 \end{array}\right. \\ r_g \leftarrow rr(RY_i) \\ t_g \leftarrow a+1 \end{array}\right. \\ \left(\begin{array}{c} g \\ t \\ r \end{array}\right) \end{array}\right. \tag{11}$$

В данном случае число ранговых групп составило $g + 1 = 27$ (с учетом того, что отсчёт ведется от 0, а не от 1). Значения векторов объема t и среднего ранга наблюдений r в каждой g-й ранговой группе показаны в табл. 3 Среди значений вектора t можно видеть присутствие 1 (единиц), — признаков наличия несвязанных ранговых групп.

Таблица 3: Значения основных характеристик ранговых групп.

Ранговая группа, g	**Средний ранг наблюдений, r_g**	**r_g^2**	**Объем группы, t_g**
0	1	1	1
1	2	4	1
2	3	9	1
3	4	16	1
4	5	25	1
5	6	36	1
6	7.5	56.25	2
7	9	81	1
8	10	100	1

Продолжение таблицы 2

Ранговая группа, g	**Средний ранг наблюдений, r_g**	**r_g^2**	**Объем группы, t_g**
9	11.5	132.25	2
10	13.5	182.25	2
11	15.5	240.25	2
12	17	289	1
13	18.5	342.25	2
14	20	400	2
15	19	361	1
16	17.5	306.25	2
17	15	225	3
18	13	169	1
19	12	144	1
20	11	121	1
21	10	100	1
22	9	81	1
23	7.5	56.25	2
24	5.5	30.25	2
25	3	9	3
26	1	1	1

Этап 5. Преобразование W_{ab}-статистики с учетом связанных рангов. При наличии связанных рангов дисперсия AB-статистики рассчитывается по формуле:

$$D(W) := \left| \begin{array}{ll} I \leftarrow \sum_{j=0}^{g} \left[t_j \cdot (r_j)^2 \right] & \\ \dfrac{m \cdot n \cdot [16 \cdot (m+n) \cdot I - (m+n+1)^4]}{16 \cdot (m+n)^2 \cdot (m+n-1)} & if\ mod(N,2) \neq 0 \\ \dfrac{m \cdot n \cdot [16 \cdot I - (m+n) \cdot (m+n+2)^2]}{16 \cdot (m+n) \cdot (m+n-1)} & otherwise \end{array} \right. \tag{12}$$

С учетом (12) преобразованная к стандартному, нормальному распределению AB-статистика W_{abn} равна:

$$W_{abn} := \left| \begin{array}{l} M(W) \leftarrow \dfrac{m \cdot (m+n+1)^2}{4 \cdot (m+n)} \\ \dfrac{W_{ab} - M(W)}{\sqrt{D(W)}} \quad if\ mod(N,2) \neq 0 \\ otherwise \\ \quad \left| \begin{array}{l} M(W) \leftarrow \dfrac{m \cdot (m+n+2)}{4} \\ \dfrac{W_{ab} - M(W)}{\sqrt{D(W)}} \end{array} \right. \end{array} \right. \tag{13}$$

В итоге имеем $W_{abn} = -0.706$.

В рассматриваемом случае, для двустороннего критерия H_0 против альтернативы вида $\Delta \neq 1$ на уровне значимости $\alpha = 0.05$ надо принять H_0, если:

$$qnorm\left(\frac{\alpha}{2}, 0, 1\right) < W_{abn} < qnorm\left(1 - \frac{\alpha}{2}, 0, 1\right) \tag{14}$$

где $qnorm\left(\frac{\alpha}{2}, 0, 1\right)$ и $qnorm\left(1 - \frac{\alpha}{2}, 0, 1\right)$ mathcad-функции граничных квантилей стандартного нормального распределения уровней значимости соответственно α и $1 - \alpha$.

Условие (14) выполняется, так как $-1.96 < -0.706 < 1.96$, из чего следует, что данные из объединенной выборки не противоречат гипотезе H_0 на уровне доверия 0.95, иными словами поисковые системы Rambler и Yandex на момент исследования имели равную релевантность.

Формулировка итога статистического исследования в таком виде сегодня встречается не так часто. Для абсолютного большинства работ, характерно более безапелляционное, вроде: «...следует принять гипотезу H_0». Между тем, как известно, «поставка» ещё одного нового статистического наблюдения может в корне перевернуть сложившееся представление (в данном случае о паритете поисковых систем).

Заключение

В литературе, посвященной статистическим вычислениям, можно обнаружить такие рекомендации:

> «Вообще, подход к статистическим критериям в анализе данных должен быть прагматическим и не отягощен лишними теоретическими рассуждениями. Имея в своем распоряжении компьютер с системой

«STATISTICA», вы легко примените к своим данным несколько критериев. Зная о некоторых подводных камнях методов, вы путем экспериментирования выберете верное решение.» [10].

Нам же представляется, что только внедрение, новых, в том числе герменевтических технологий в образовательный процесс позволит увеличить интенсивность и качество усвоения статистических знаний, даст возможность специалистам грамотно применять статистический инструментарий и выбирать действительно верное решение, не прибегая к слепому экспериментированию.

Список литературы

1. Kevin Athern's Professional Page [Электронный ресурс]. — Режим доступа к ресурсу: http://www.davinihress.com/professional.html
2. Бородина А. И. Специализированные пакеты для математической обработки данных / А. И. Бородина, Л. И. Крошинская, О. Л. Сапун — Минск: НО ООО «БИП-С», 2003. — Режим доступа к ресурсу: http://bip-ip.com/spetsializirovannyie-paketyi-dlya-matem/
3. Показатели качества информационно-поисковых систем. Полнота и точность выдачи [Электронный ресурс] — Режим доступа к ресурсу: http://koriolan404.narod.ru/tipis/27.htm
4. Moses L. E. Rank tests of dispersion / L. E. Moses // Ann. Main. Statist. — 1963 — №34. — P. 973–983.
5. Сорокин О. Д. Прикладная статистика на компьютере. 2-е изд / О. Д. Сорокин. — Краснообск: ГУП РПО СО РАСХН, 2009. — 222 с.
6. Холлендер М. Непараметрические методы статистики / Холлендер М., Вульф Д. А. — М.: Финансы и статистика, 1983. — 520 с.
7. Hollander M. Non-parametric Statistical Methods / Hollander M., Wolfe D. A. — New York: John Wiley & Sons, 1973. — P. 503.
8. Hollander M. Non-parametric Statistical Methods, 2nd edition / Hollander M., Wolfe D. A. — NewYork: Wiley, 1999. — P. 515.
9. Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников / А. И. Кобзарь. — М.: ФИЗМАТЛИТ, 2006. — 816 с.
10. Боровиков В. STATISTICA — искусство анализа данных на компьютере / В. Боровиков [Электронный ресурс]. — Режим доступа к ресурсу: http://www.statosphere.ru/books-arch/bor-kat/50–13–.html

Статья поступила в редакцию 30.06.2012