УДК 620.1
DOI: 10.15827/0236-235X.118.257-260
Дата подачи статьи: 09.09.16
2017. Т. 30. № 2. С. 257–260

РЕКУРСИВНЫЙ АЛГОРИТМ ТОЧНОГО РАСЧЕТА РАНГОВЫХ КРИТЕРИЕВ ПРОВЕРКИ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Л.В. Агамиров, д.т.н., профессор, mmk@mati.ru
(Национальный исследовательский университет «МЭИ», ул. Красноказарменная, 14, г. Москва, 111250, Россия);
В.А. Вестяк, к.ф.-м.н., доцент, v.a.vestyak@mail.ru
(Московский авиационный институт (национальный исследовательский университет), Волоколамское шоссе, 4, г. Москва, 125993, Россия);
В.Л. Агамиров, к.т.н., аналитик, avl095@mail.ru

В статье рассматривается методика генерации точных распределений ранговых непараметрических критериев средствами компьютерной комбинаторики.

Актуальность работы обусловлена затруднениями в определении точных распределений критических значений ранговых критериев проверки статистических гипотез из-за того, что точные таблицы, рекуррентные формулы для многих критериев не существуют, а аппроксимации часто дают неудовлетворительный результат при ограниченных объемах наблюдений.

Задача расчета распределения ранговых критериев заключается в переборе всех возможных вариантов перестановок выборок и в расчете ранговых статистик, а также накопленных частот их появления.

Для ее решения разработана программа генерации перестановок элементов выборок ранговых непараметрических критериев, основанная на рекурсивном алгоритме прямого перебора вариантов перестановок вектора порядковых статистик со следующим ограничением числа вариантов: во всех вариантах перестановок элементы одной и той же выборки не могут меняться местами, что является универсальным условием для всех точных распределений ранговых критериев.

В работе приводится ссылка на интернет-ресурс, содержащий программный комплекс реализации алгоритма расчета ранговых критериев. В данном комплексе рассмотрены четыре непараметрических критерия: двухвыборочный критерий Уилкоксона, критерий Лемана–Розенблатта, критерий серий и критерий Краскела–Уоллиса, точные распределения статистик которых представляют наибольший интерес для технических задач. Рассматриваемый алгоритм может быть использован и для других ранговых критериев проверки статистических гипотез.

В работе представлена разработанная авторами реализация метода генерации точных распределений ранговых непараметрических критериев средствами компьютерной комбинаторики, основанная на рекурсивном прямом переборе вариантов перестановок вектора порядковых статистик с последующей фильтрацией результатов. Таким образом, решена задача определения критических значений ранговых непараметрических критериев для проверки статистических гипотез.

Ключевые слова: *непараметрические ранговые критерии, проверка гипотез, точное распределение, комбинаторика, алгоритм, программа, JavaScript.*

Задача проверки статистических гипотез во всех случаях сопряжена с необходимостью определения критических значений критериев. В то же время для большинства ранговых критериев определение точных распределений является весьма непростой задачей как с математической, так и с вычислительной точки зрения. Различного рода аппроксимации зачастую дают неудовлетворительный результат при ограниченных объемах наблюдений, свойственных анализу данных в технических задачах, связанных со значительным рассеянием свойств, вследствие структурной неоднородности конструкционных материалов и большой вариативности внешних факторов при проведении испытаний. Точные таблицы, рекуррентные формулы, производящие функции частот и моментов для многих критериев не существуют [1–4]. Кроме того, при современном развитии вычислительной техники более предпочтительным является точный компьютерный расчет. Подробнее эти вопросы обсуждаются в [5], в данной работе предлагается методика генерации точных распределений ранговых непараметрических критериев средствами компьютерной комбинаторики.

С вычислительной точки зрения распределение ранговых критериев представляет собой перебор всех возможных вариантов перестановок элементов выборочных совокупностей при некоторых граничных условиях с последующим расчетом ранговых статистик и накопленных частот их появления. Предлагаемый далее алгоритм применим для большинства критериев, для которых вычисляются выборочные ранговые статистики. Для некоторых критериев существуют более эффективные методы [5, 6], однако с целью обобщения здесь рассматриваются различные критерии независимо от наличия иных методов расчета точных распределений. Авторы сознают, что предлагаемый алгоритм имеет недостаток – перебор большого числа лишних вариантов, что зачастую, особенно при более чем двух выборках, ведет к существенному увеличению машинного времени. Так что, данный алгоритм может рассматриваться как некоторый шаг в направлении применения методов рекурсивной

компьютерной комбинаторики в задачах статистического анализа. К тому же при постоянном увеличении производительности процессоров современных компьютеров проблема машинного времени постепенно отступает на второй план.

В рассматриваемой ниже программе, написанной на языке Javascript, используется метод генерации перестановок элементов выборок ранговых непараметрических критериев. Метод основан на рекурсивном прямом переборе вариантов перестановок вектора порядковых статистик объема *n* с последующей фильтрацией результатов с учетом универсального для всех точных распределений ранговых критериев условия: внутри данной выборки порядковые номера элементов должны сохраняться независимо от места их расположения. Естественно, что за пределами выборки порядковые номера элементов должны принимать все возможные варианты перестановок.

Для примера рассмотрим возможные варианты таких перестановок для трех выборок (k=3) объемами $n_1 = 1$, $n_2 = 1$, $n_3 = 2$.

При суммарном объеме выборок $N = \sum_{i=1}^{k} n_i = 4$ безусловное число вариантов перестановок элементов равно *N!*, то есть 24. С учетом указанного выше условия число таких вариантов сокращается до 12: $kk = \frac{N!}{\prod_{i=1}^{k} n_i} = \frac{4!}{1!\cdot 1!\cdot 2!} = 12$.

В таблице перечислены эти варианты, выборки обозначены символами X, Y, Z.

Варианты перестановок рангов трех выборок объемами n_1=1, n_2=1, n_3=2

Permutations of ranks of 3 samples with a volume of n_1=1, n_2=1, n_3=2

Номер варианта	Перестановка	Вектор *a*	Вектор *aindex*
1	X_1 Y_1 Z_1 Z_2	1 2 3 4	1 2 3 3
2	X_1 Z_1 Y_1 Z_2	1 3 2 4	1 3 2 3
3	X_1 Z_1 Z_2 Y_1	1 3 4 2	1 3 3 2
4	Y_1 X_1 Z_1 Z_2	2 1 3 4	2 1 3 3
5	Y_1 Z_1 X_1 Z_2	2 3 1 4	2 3 1 3
6	Y_1 Z_1 Z_2 X_1	2 3 4 1	2 3 3 1
7	Z_1 X_1 Y_2 Z_2	3 1 2 4	3 1 2 3
8	Z_1 X_1 Z_2 Y_1	3 1 4 2	3 1 3 2
9	Z_1 Y_1 X_1 Z_2	3 2 1 4	3 2 1 3
10	Z_1 Y_1 Z_2 X_1	3 2 4 1	3 2 3 1
11	Z_1 Z_2 X_1 Y_1	3 4 1 2	3 3 1 2
12	Z_1 Z_2 Y_1 X_1	3 4 2 1	3 3 2 1

Отметим еще раз, что номера 3 и 4, обозначающие элементы третьей выборки объемом 2, во всех вариантах сохраняют порядок, что и является условием генерации точного распределения для любого рангового критерия, в то время как порядковые номера первой и второй выборок, имеющих единичные объемы, могут меняться местами.

Далее приведена программа генерации перестановок при наличии граничных условий.

```
1. <html>
2. <script>
3. var kk;
4. kk=0;
5. stat_crit();
//*********************************************
6. function stat_crit() {
7. var kx; // количество выборок
8. var a=[]; // вектор порядковых номеров элементов выборок
9. var aindex=[]; //вспомогательный вектор индексов выборок
10. var m=[]; //вектор объемов выборок
11. var km;
12. var n; //суммарный объем выборок
13. kx=3;n=0;m[1]=1;m[2]=2;m[3]=2;
14. for (i=1;i<=kx;i++) n=n+m[i];
15. for (i=1;i<=n;i++) a[i]=i;
16. km=0;
17. for (i=1;i<=kx;i++) {
18. for (j=1;j<=m[i];j++) aindex[j+km]=i;
19. km=km+m[i];
20. }
21. omega(a,aindex,n,n);
22. }
//*********************************************
23. function omega(a,aindex,n,mx) {
//рекурсивная функция генерации перестановок от 1 до n
24. var i,j,ss;
25. if (mx==1) {
// функция фильтрации и вывода на печать текущей перестановки элементов вектора a[1..n]
26. filter(true,n,a,aindex); }
27. else {
28. for (i=1;i<=mx;i++) {
29. ss=a[1];
30. for (j=1;j<=mx-1;j++) a[j]=a[j+1];
31. a[mx]=ss;
32. omega(a,aindex,n,mx-1);
33. }
34. }
35. }
//*********************************************
// функция фильтрации и печати текущей перестановки из kk=n!/(m[1]!*m[2]!*...*m[kx]!) вариантов
36. function filter(fltr,n,a,aindex) {
37. test=0;
38. if(fltr) {
39. if (test==0) {
40. for (i=1;i<=n;i++) {
41. for (j=i;j<=n;j++) {
//отмена вывода перестановки элементов при нарушении их порядковых номеров внутри выборки
42. if (aindex[a[i]]==aindex[a[j]] && a[i]>a[j])  {test=1;break;}
43. }
44. if (test==1) break;
45. }
46. }
47 }
48. if (test==0) {
49. kk++;
50. document.write(kk+".  ");
// вывод на печать вектора a[1..n]
51. for (i=1;i<=n;i++) document.write(a[i]+" ");
52. document.write("<br>");
53 }
54 }
//*********************************************
55. </script>
56. </html>
```

Программа содержит три основные функции. Первая функция – stat_crit (строки 6–20) формирует исходные данные рассмотренного выше примера и два вспомогательных вектора *a* и *aindex* размерности *N*, первый из которых представляет собой вектор порядковых номеров объединенной выборки, второй – вектор, содержащий номера выборок, как это показано в таблице.

Вторая функция – omega(a, aindex, n, n) (строки 23–35) является рекурсивной функцией генерации перестановок элементов от 1 до *n*, в строке 32 которой представлена процедура рекурсии с уменьшением на единицу объема объединенной выборки от *n* до 1. В этом случае (строка 26) реализуется фильтрация перестановок в соответствии с принятыми граничными условиями, для чего введена третья функция – filter(fltr, n, a, aindex) (строки 36–54). Отметим, что фильтрация производится при значении параметра *fltr=true* (строка 38), в противном случае выводятся все варианты перестановок. На выходе этой функции выводятся элементы вектора *a* с учетом граничных условий, как показано в таблице.

Все остальные действия, необходимые для расчета статистик ранговых критериев, являются достаточно тривиальной вычислительной задачей и в данной работе не обсуждаются.

Полная версия программы с открытым кодом размещена на Javascript по ссылке http://inteh.mpei.ru/programs/stat/Menu/stat_tree.html. В данном комплексе рассмотрены четыре непараметрических критерия: двухвыборочный критерий Уилкоксона [1–4], критерий серий [7], критерий Лемана–Розенблатта [8] и критерий Краскела–Уоллиса [3, 4], точные распределения статистик которых, по мнению авторов, представляют наибольший интерес, по крайней мере, для технических задач. Очевидно, что представленный алгоритм может быть использован и для других ранговых критериев проверки статистических гипотез.

Для полноты картины рассмотрим указанные критерии, за исключением критерия серий, распределение статистики которого не требует вычисления рангов, а может быть подсчитано [7, 9] более простым способом, что и реализовано в полной версии программного комплекса.

Двухвыборочный критерий Уилкоксона предназначен для проверки гипотезы об отсутствии сдвига двух независимых выборок, то есть об отсутствии различия между медианами двух совокупностей при одинаковом, но произвольном распределении [1, 2]. Пусть $x_1, x_2, \ldots, x_m$ – случайная выборка из $F(x-\theta_x)$, $y_1, y_2, \ldots, y_n$ – случайная выборка из $F(y-\theta_y)$ ($m \le n$). Функцию распределения F не предполагают симметричной, но форма распределения должна быть одинаковой для двух совокупностей. Для проверки нулевой гипотезы о том, что обе выборки извлечены из одной и той же совокупности H_0: $\Delta = \theta_y - \theta_x = 0$ против альтернативы H_A: $\Delta \ne 0$ строят вариационный ряд из $k = m + n$ наблюдений и присваивают им ранги, равные порядковому номеру наблюдения в общем вариационном ряду. Далее рассчитывают сумму рангов меньшей выборки в общем вариационном ряду:

$$W = \sum_{i=1}^{m} R_i \ .$$

Для проверки нулевой гипотезы H_0: $\Delta = 0$ при альтернативной гипотезе H_A: $\Delta < 0$ должно выполняться неравенство $W > W_{\alpha l}$. При альтернативной гипотезе H_A: $\Delta > 0$ должно выполняться неравенство $W \le W_{\alpha u}$. При двусторонней альтернативной гипотезе H_A: $\Delta \ne 0$ должно выполняться неравенство $W_{\alpha l} \le W \le W_{\alpha u}$ с уровнем значимости 2α.

Критерий Лемана–Розенблатта проверяет гипотезу об однородности двух выборок. Проверяется нулевая гипотеза о том, что две выборки извлечены из одной и той же генеральной совокупности, то есть H_0: $F(x) = G(x)$ при любом x. Статистика критерия рассчитывается по формуле

$$\omega^2 = \frac{1}{n \cdot m}\left[\frac{1}{6} + \frac{\sum_{i=1}^{n}\left(R_i - i\right)^2}{m} + \frac{\sum_{j=1}^{m}\left(S_j - j\right)^2}{n}\right] + \frac{2}{3},$$

где R_i, S_j – ранги первой выборки объемом n и второй объемом m в общем вариационном ряду. Нулевую гипотезу принимают, если $\omega^2 \le \omega_\alpha^2$ с уровнем значимости α. В противном случае принимают альтернативную гипотезу.

Критерий Краскела–Уоллиса обобщает задачу о двух выборках на случай k выборок: x_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_j$, с функциями распределения $F(x - \theta_j)$, где n_j – число наблюдений в j-й выборке. Нулевая гипотеза утверждает, что k выборок из произвольных совокупностей можно рассматривать как одну (объединенную) выборку из общей совокупности, то есть подтверждается равенство параметров сдвига θ_j, когда не задано значение общего параметра масштаба H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$ против альтернативы H_A: $\theta_1, \ldots, \theta_k$ не все равны. Для проверки нулевой гипотезы строят общий вариационный ряд из $n = \sum_{i=1}^{k} n_i$ наблюдений и рассчитывают статистику:

$$H = \frac{12}{(n+1)} \cdot \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3 \cdot (n+1),$$

где R_i – сумма рангов i-й выборки в общем вариационном ряду.

Нулевую гипотезу принимают, если $H \le H_\alpha$ с уровнем значимости α. В противном случае принимают альтернативную гипотезу.

Другим способом проверки выборочной гипотезы k является попарное сравнение выборок по критерию Уилкоксона с вычислением точных критических значений.

Таким образом, разработанный алгоритм, являющийся реализацией метода рекурсивной компью-

терной комбинаторики, может быть использован для решения задачи проверки статистических гипотез, связанной с определением критических значений критериев. Учитывая, что для большинства ранговых критериев определение точных распределений является сложной задачей с математической и вычислительной точек зрения, представленная в работе методика генерации точных распределений критериев средствами компьютерной комбинаторики является эффективным решением.

Литература

1. Кендалл М.Дж., Стьюарт А. Теория распределений. М.: Наука, 1966.
2. Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. М.: Наука, 1973. 899 с.
3. Хеттманспергер Т. Статистические выводы, основанные на рангах. М.: Финансы и статистика, 1987. 334 с.
4. Холлендер М., Вулф Д. Непараметрические методы статистики. М.: Финансы и статистика, 1983. 518 с.
5. Агамиров Л.В., Агамиров В.Л., Вестяк В.А. Численные методы и алгоритмы расчета точных распределений непараметрических критериев проверки статистических гипотез // Вестн. МАИ. 2013. Т. 20. № 4. С. 212–218.
6. Агамиров Л.В. Методы статистического анализа механических испытаний. М.: Интермет Инжиниринг, 2004. 127 с.
7. Шуленин В.П. Математическая статистика. Ч. 2: Непараметрическая статистика. Томск: Изд-во НТЛ, 2012. 388 с.
8. Леман Э. Проверка статистических гипотез. М.: Наука, 1964. 498 с.
9. Липский В. Комбинаторика для программистов. М.: Мир, 1988. 200 с.

Software & Systems Received 09.09.16
DOI: 10.15827/0236-235X.118.257-260 2017, vol. 30, no. 2, pp. 257–260

RECURSIVE ALGORITHM FOR EXACT CALCULATION OF RANK TESTS FOR TESTING STATISTICAL HYPOTHESES

L.V. Agamirov [1], *Dr.Sc. (Engineering), Professor, mmk@mati.ru*
V.A. Vestyak [2], *Ph.D. (Physics and Mathematics), Associate Professor, v.a.vestyak@mail.ru*
V.L. Agamirov, *Ph.D. (Engineering), analist, avl095@mail.ru*

[1] *National Research University "MPEI", Krasnokazarmennaya St. 14, Moscow, 111250, Russian Federation*
[2] *Moscow Aviation Institute (National Research University), Volokolamskoe Highway 4, Moscow, 125993, Russian Federation*

Abstract. The paper considers the method of generating exact distributions of nonparametric rank tests by means of the computer combinatorial theory.

Relevance of the work consists in the fact that determination of exact distribution of critical values of rank tests for statistical hypotheses testing is complicated by the fact that the exact tables and recurrence formulas for many of the tests do not exist. In addition, approximations often give unsatisfactory results at limited volumes of observations.

The task of calculating the distribution for rank tests is a search of all possible sample permutations and calculations of rank statistics, as well as cumulative frequency of their occurrence.

The program of generating permutations of elements of samples for nonparametric rank criteria based on the recursive brute-force algorithm of direct enumeration of order statistics vector permutation is developed with the following limited number of options: in all permutation options the elements from the same sample cannot be swapped. It is a universal condition for all the rank criteria exact distributions.

The paper refers to the Internet resource that contains the software package implementation of the considered calculation algorithm for a rank test. This complex contains four nonparametric criteria: two-sample Wilcoxon test, Lehmann-Rosenblatt test, series test and Kruskal-Wallis test, whose accurate distribution statistics are of greatest interest for technical problems. The algorithm can be used for other rank tests of statistical hypotheses testing.

The paper presents an implementation of the generation method of nonparametric rank test exact distributions by computer combinatorial means. It is based on the developed by the authors recursive direct enumeration of options of order statistics vector permutation with following filtration of the results. Thus, the authors solve the problem of determining the critical values of nonparametric rank tests for testing statistical hypotheses.

Keywords: nonparametric rank tests, hypothesis testing, exact distribution, combinatorics, algorithm, program, JavaScript.

References

1. Kendall M.G., Stuart A. *The Advanced Theory of Statistics. Vol. 1. Distribution Theory*. Charles Griffin Publ., 4th ed., 1952 (Russ. ed.: Moscow, Nauka Publ., 1966).
2. Kendall M.G., Stuart A. *The Advanced Theory of Statistics. Vol. 1. Inference and Relationship*. Hafner Publ. Company, 3rd ed., 1961, 676 p. (Russ. ed.: Moscow, Nauka Publ., 1973, 899 p.).
3. Hettmansperger T.P. *Statistical Inference Based on Ranks*. NY, J. Wiley and Sons Publ., 1984 (Russ. ed.: Moscow, Finansy i statistika Publ., 1987, 334 p.).
4. Hollander M., Wolfe D. *Nonparametric Statistical Methods*. Wiley Publ., 1999 (Russ. ed.: Moscow, Finansy i statistika Publ., 1983, 518 p.).
5. Agamirov L.V., Agamirov V.L., Vestyak V.A. Numerical methods and algorithms of calculation of exact distributions of nonparametrical criteria statistical hypotheses. *Vestnik MAI* [Vestnik Moskovskogo Aviatsionnogo Instituta]. 2013, vol. 20, no. 4, pp. 212–218 (in Russ.).
6. Agamirov L.V. *Metody statisticheskogo analiza mekhanicheskikh ispytany* [The Methods of Mechanical Test Statistical Analysis]. Moscow, Intermet Inzhiniring Publ., 2004, 127 p.
7. Shulenin V.P. *Matematicheskaya statistika* [Mathematical Statistics]. Tomsk, NTL Publ., 2012, 388 p.
8. Lehman E.L. *Testing Statistical Hypotheses*. J. Wiley & Sons Publ., NY, 1959 (Russ. ed.: Moscow, Nauka Publ., 1964, 498 p.).
9. Lipsky B. *Kombinatorika dlya programmistov* [Combinatorics for Programmers]. Moscow, Mir Publ., 1988, 200 p.