УДК 519.234

В.Е. Хиценко

Ранговые критерии в задачах защиты информации

Показаны возможности и перспективы ранговых критериев для выявления различий применительно к конкретным задачам технической, организационной и компьютерной безопасности. Методы классифицированы по ситуациям зависимых и независимых выборок, по выявляемым различиям в параметрах положения и масштаба. Изложение иллюстрируется примерами профессионального содержания.

Ключевые слова: непараметрическая статистика, ранговые критерии, защита информации.

Перспективы предлагаемого математического инструментария

Математическая статистика дает нам объективные методы обнаружения различий в данных, настораживающих закономерностей, связей, появление которых маловероятно настолько, что не может быть списано на случай. Это открывает возможность построения систем мониторинга обстановки на основе сравнения текущих значений непрерывно вычисляемой статистики с критическими, пороговыми значениями, вероятность превышения которых при отсутствии оснований для тревоги (нуль-гипотеза) ничтожна. При ином подходе можно непрерывно вычислять эту вероятность, называемую эмпирической значимостью, и информировать ответственных лиц о степени опасности текущей ситуации.

Наиболее приемлемыми для этих задач оказываются непараметрические методы, инвариантные к законам распределения исследуемых признаков. Этими признаками могут быть: интенсивность сетевых атак, специфические черты клавиатурного почерка, эффективность акустического зашумления речи, число нарушений режима секретности и другие случайные величины различной природы и с различными и, как правило, не известными нам законами распределения. Другим преимуществом непараметрических процедур является простота вычислений.

В подавляющем большинстве эти процедуры предполагают переход к ранговой шкале измерения признака и называются ранговыми критериями [1–3]. В задачах организационной безопасности чаще располагаем данными в номинативной шкале, анализируем частоты попадания в разные категории этой шкалы, таблицы сопряженности частот. Такие задачи здесь не рассматриваются. Подробнее об этих задачах и об измерении степени связи признаков, заданных в разных шкалах, можно прочитать в [4].

Выбор критерия

Существует непростая проблема выбора подходящего из достаточно большого арсенала ранговых критериев сравнения двух и более наборов измерений. Здесь, прежде всего, следует различать ситуации независимых и связных выборок. Рассмотрим подробнее этот этап выбора.

Допустим, нам необходимо сравнить одну группу объектов (помещения, методы, отделы, сотрудники) по проявлению некоторого признака в разных условиях. Мы измеряем признак на каждом из объектов и получаем выборку объемом n. Затем на этих же объектах, но в других условиях получаем другую выборку того же объема. Например, мы изучаем изменение внимательности или уровня компетентности конкретной группы людей с течением времени или после обучения, различие фона в одних и тех же помещениях при разных способах экранирования, в разных условиях. В этой ситуации результаты измерений представляют в виде таблицы, где строки соответствуют объектам, а столбцы – условиям. Нельзя перемешивать элементы в столбце, они построчно связаны принадлежностью к одному объекту. Это ситуация связных выборок.

В другой ситуации мы имеем разные наборы объектов, группы сотрудников, возможно, разные по численности, и производим измерения признака в группах для выявления их различия по этому признаку. Здесь нет построчной связанности выборочных значений признака. Нет смысла представлять эти данные в виде таблицы «объект–признак». Это ситуация независимых выборок.

Следующий этап выбора подходящего критерия – это определение интересующего нас параметра распределения. Нас может волновать сдвиг параметра положения как некоего центра распределения признака в разных условиях, либо изменение параметра масштаба как рассеяния значений признака относительно центра распределения, их нестабильности, либо обоих этих параметров. Также существуют критерии сравнения законов распределения двух независимых выборок, так называемые критерии согласия.

Окончательный выбор следует делать на основе сравнения критериев по мощности. Мощностью критерия называют вероятность правильного объявления тревоги. Однако такой анализ непрост и не всегда доступен, и практически при получении сомнительного результата рекомендуется дублировать проверку с помощью других адекватных задаче критериев.

Зависимые выборки

Наиболее известным критерием проверки сдвига параметра положения является знаково-ранговый критерий Уилкоксона. Для реализации критерия мы находим разности, сдвиги между значениями признака на каждой из n пар измерений, выбрасываем из рассмотрения нулевые разности и находим ранги модулей разностей. Затем в качестве статистик Уилкоксона используем сумму рангов положительных и сумму рангов отрицательных разностей. Закон распределения этих статистик в предположении об отсутствии сдвига (нуль-гипотеза) известен, и можно найти критические значения наибольшей или наименьшей из сумм, а также эмпирические значимости.

Более мощным критерием проверки сдвига двух зависимых выборок является критерий Фрезера. Здесь вместо сумм рангов мы суммируем математические ожидания модулей порядковых статистик в выборке объемом n из стандартного нормального распределения. Для определения этих математических ожиданий используются специальные таблицы, а их суммы при справедливости нуль-гипотезы распределены по нормальному закону с известными параметрами.

Рассмотрим пример 1. В течение первой и второй недели фиксировались средние ежедневные интенсивности сетевых атак (табл. 1). Возникло подозрение о росте интенсивности.

Таблица 1

Сравнение интенсивностей сетевых атак

1-я неделя	2-я неделя	Сдвиг	Модуль сдвига	Ранг модуля	Знак сдвига	Математическое ожидание модулей
1,194	0,4	–0,794	0,794	6	–	1,23485
2,8	6,292	3,492	3,492	7	+	1,72385
0,944	1,681	0,737	0,737	5	+	0,93444
0,497	0,767	0,27	0,27	3	+	0,5042
0,65	0,367	–0,283	0,283	4	–	0,70212
0,342	0,564	0,222	0,222	2	+	0,32605
0,317	0,444	0,127	0,127	1	+	0,15967

Статистика Уилкоксона (сумма рангов модулей положительных сдвигов) $T^+ = 18$. Достигнутая значимость 0,289 [2, с. 280]. Сравним с результатом более мощного критерия. Статистика Фрезера [3, с. 350] (сумма математических ожиданий модулей порядковых статистик нормальной выборки объемом 7 для положительных сдвигов) $S^+ = 3{,}64821$. Асимптотическая значимость 0,247 тоже не так уж мала. Не можем отклонить нуль-гипотезу. Роста интенсивности не выявлено.

При сравнении выборок значений признака, измеренных в более чем двух условиях, мы имеем непараметрический аналог двухфакторного дисперсионного анализа. Ранговые критерии могут констатировать влияние условий на признак и даже указать значимую тенденцию в сдвиге признака.

Применяя критерий Фридмана, находим ранги r_{ij} значений признака в каждой из n строк таблицы и используем в дальнейшем средние ранги каждого из k столбцов-условий

$$\overline{r}_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij}, \quad j = 1,...,k.$$

Статистика Фридмана при отсутствии одинаковых рангов в строке вычисляется по формуле

$$S = [\frac{12}{nk(k+1)}\sum_{j=1}^{k}\overline{r}_j^2] - 3n(k+1),$$

и при справедливости нуль-гипотезы имеет известный закон распределения. Имеются таблицы верхних процентных точек и критических значений. Уточненные статистики приведены в [2, с. 12].

Рассмотрим пример 2. Получены результаты экспериментальной проверки четырех различных схем эвакуации людей в виде времени движения каждого из 18 сотрудников до выхода из здания по рекомендуемому схемой маршруту. Средние ранги столбцов $\overline{r}_1 = 2{,}86$, $\overline{r}_2 = 3{,}25$, $\overline{r}_3 = 1{,}94$, $\overline{r}_4 = 1{,}95$. Статистика Фридмана получилась равной $S = 14{,}309$, что превышает критическое значение для 5%-го уровня значимости. Вывод – схемы эвакуации различаются по эффективности. Можно уточнить – третья и четвертая схемы близки по эффективности и лучше первой и второй.

Существуют более мощные критерии для этих задач. Покажем критерий Доксама на примере 3. Исследуются шесть методов акустической защиты речевой информации. К работе привлечена группа из одиннадцати аудиторов. В табл. 2 показаны относительные частоты распознанных слов.

Анализируя модули разностей результатов аудиторов по всем парам методов, находим статистику Доксама $A = 29{,}132$ [2, с. 175] с асимптотической значимостью 2,185E-05 и делаем вывод о существенной разнице в эффективности методов. Затем желательно убедиться в различии лучших методов 2, 4 и 6 между собой и, если различие не обнаружится, сделать окончательный выбор в пользу наиболее комфортного или простого метода.

Таблица 2

К выбору метода акустической защиты

Аудиторы	Мет. 1	Мет. 2	Мет. 3	Мет. 4	Мет. 5	Мет. 6
1	0,17	0,27	0,16	0,37	0,11	0,28
2	0,25	0,29	0,13	0,26	0,1	0,24
3	0,21	0,34	0,19	0,29	0,2	0,37
4	0,3	0,25	0,08	0,19	0,18	0,36
5	0,19	0,3	0,12	0,25	0,06	0,19
6	0,07	0,19	0,11	0,27	0,12	0,32
7	0,24	0,18	0,2	0,25	0,21	0,33
8	0,27	0,31	0,22	0,29	0,1	0,25
9	0,21	0,32	0,24	0,26	0,14	0,31
10	0,23	0,25	0,15	0,31	0,09	0,3
11	0,18	0,22	0,18	0,33	0,2	0,29
Ср. знач.	0,211	0,265	0,162	0,279	0,137	0,295

Представляют интерес задачи проверки различия в рассеянии относительно среднего в парах измерений в зависимых выборках, т.е. построчное различие масштаба. Одним из подходящих является критерий Сэндвика–Олссона [3, с. 507]. Мы рассматриваем пары измерений, сделанные в двух условиях (x_i, y_i), $i = 1,...,n$, и оценки медиан этих измерений med_x и med_y. Вычисляем последовательность $z_i = |y_i - med_y| - |x_i - med_x|, i = 1,...,n.$ Значимая положительная асимметрия в этой последовательности говорит об увеличении параметра масштаба в условии y на каждом объекте и наоборот. Для проверки симметрии используем критерий знаковых рангов Уилкоксона.

Таблица 3

Данные аутентификации

x_i, мс	y_i, мс	z_i	$\|z_i\|$	Ранг $\|z_i\|$
212	211	13	13	9
232	201	3	3	4
200	194	18	18	11
224	230	–6	6	5
217	239	10	10	6
197	224	–15	15	10
211	237	12	12	8
222	195	19	19	12
182	252	–2	2	2,5
200	238	2	2	2,5
207	262	33	33	14
223	224	–11	11	7
213	201	22	22	13
226	209	1	1	1
212	188	36	36	15

Пример 4. В задаче аутентификации пользователя компьютера по клавиатурному почерку длительности y_i нажатий клавиш и пауз между нажатиями при наборе ключевого слова сравниваются с эталонными x_i. В табл. 3 приведены последовательности интервалов при наборе 8 символов. Значения медиан следующие: $med_x = 212$ мс, $med_y = 224$ мс. Сумма рангов модулей положительных значений z_i равна $T^+ = 95{,}5$. Значимость достигает 0,0225. Вывод: различие в рассеянии интервалов существенно, в доступе отказать. Характерно, что в этом же примере при проверке сдвига в длительностях интервалов по критерию Фрезера получили $S^+ = 7{,}6934$ с асимптотической значимостью 0,1789. То есть сдвиг не был выявлен, аутентификация прошла бы успешно. Таким образом, необходима проверка по обоим параметрам положения и масштаба.

Независимые выборки

Рассмотрим несколько наиболее распространенных критериев для этой ситуации.

Пример 5. В двух отделах было проведено тестирование профессионального уровня сотрудников. Упорядоченные по возрастанию результаты в баллах сведены в табл. 4. Различаются ли отделы по исследуемому признаку? Применим критерий Ван-дер-Вардена. Для этого объединяем две выборки $x_1,...,x_m$ и $y_1,...,y_n$ в одну, ранжируем ее и выделяем ранги элементов одной выборки, допустим, y $R_j^y, j = 1,...,n$. Вычисляем статистику Ван-дер-Вардена

$$S_{VdW} = \sum_{j=1}^{n} \Phi^{-1}\left(\frac{R_j^y}{m+n+1}\right),$$

где Φ^{-1} – обратная функция распределения стандартного нормального закона, дающая так называемые инверсные нормальные метки (квантили). При выполнении нуль-гипотезы эта статистика распределена нормально с нулевым математическим ожиданием и с известной дисперсией [1, с. 107]. В данном примере $S_{VdW} = -5{,}081$ и $VarS_{VdW} = 4{,}658$. Асимптотическая значимость достигает лишь 0,0097. Уровень компетентности в отделе 2 явно ниже.

Таблица 4

Результаты тестирования компетентности сотрудников

Отдел 1, x_i	96	100	104	104	120	120	120	121	126	130	134	
Отдел 2, y_j	76	82	82	84	88	100	102	104	110	118	120	126

Если сравниваем более чем 2 независимые выборки, то задача похожа на однофакторный дисперсионный анализ. Обычно используют критерий Краскела–Уоллиса. Объединяем все выборки в одну объемом N и находим ранги совместной выборки. Обозначим ранг i-го измерения j-й выборки R_{ij} и находим суммы рангов каждой выборки R_j и статистику Краскела–Уоллиса H:

$$R_j = \sum_{i=1}^{n_j} R_{ij} \quad \text{и} \quad H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j}) - 3(N+1).$$

Затем из таблиц находим верхние процентные точки или достигнутую значимость.

Таблица 5

Данные по нарушениям режима в филиалах

Филиал 1	2,9	3	2,5	2,6	3,2	
Ранги 1	9	10	4	5	11	R_1= 39
Филиал 2	3,8	2,7	4	3,9		
Ранги 2	12	6	14	13		R_2= 45
Филиал 3	2,8	2,75	1,71	1,9	2	
Ранги 3	8	7	1	2	3	R_3=21

Пример 6. В табл. 5 показаны среднее число нарушений режима секретности в разных отделах трех филиалов организации и ранги этих чисел. Необходимо проверить гипотезу о различии филиалов по данному признаку.

Статистика Краскела–Уоллиса равна $H = 6{,}3514$. Из таблиц [2, с. 305] находим критические ее значения для 5%-го и 1%-го уровней значимости: $H_{\text{кр}(0,05)} = 5{,}6429$ и $H_{\text{кр}(0,01)} = 7{,}7914$. Видно, что различие филиалов проявилось на 5%-м уровне значимости. Из тех же таблиц можно увидеть и достигнутую значимость 0,032. Можно утверждать теперь, что филиал 3 лучше остальных с точки зрения соблюдения режима. Не может быть случайным такое различие в суммах рангов.

В некоторых ситуациях желательно не просто установить различие в выборках, но указать существование тенденции в изменении параметра положения в сравниваемых выборках. Для этого используют критерий Джонкхиера–Терпстры [2, с. 136]. Располагаем все выборки в столбцах таблицы. Причем ожидаемое направление тенденции должно совпадать с расположением, нумерацией столбцов. Это расположение может диктоваться ростом какого-то количественного параметра в условиях, определяющих столбцы, или средними значениями признака в столбцах, если условия носят качественный характер, и нет никаких других доводов в пользу определенной тенденции.

Затем рядом с каждым из значений записываем в скобках сумму превышений его во всех столбцах, расположенных правее. Совпадающие значения добавляют в сумму 0,5. Для последнего столбца эти суммы не вычисляются. Сумма всех превышений дает нам статистику Джонкхиера–Терпстры.

Пример 7. В табл. 6 приведены данные о количестве нарушений режима в течение месяца сотрудниками с разным стажем работы.

Мы ожидаем, что с уменьшением опыта число нарушений должно возрастать, и, судя по средним значениям, это похоже на правду. Разумеется, такое совпадение не всегда получается.

Проверим гипотезу о соответствующей тенденции. Статистика Джонкхиера–Терпстры $J = 68$.

Из [2, с. 322] находим значимость, равную 0,0346, и критические значения $J_{\text{кр}(0,05)} = 66$ и $J_{\text{кр}(0,01)} = 73$. Нуль-гипотезу об отсутствии ожидаемой тенденции отклоняем на уровне 0,0346.

Считаем доказанной рабочую гипотезу – стаж повышает дисциплину.

Нередко нас волнует различие независимых выборок в параметре масштаба.

Если будет доказано значимое увеличение рассеяния при каком-то условии, то это должно привлечь внимание, в частности, необходимо увеличивать количество повторных измерений признака

для получения той же точности результатов по сравнению с условием, более благоприятным с точки зрения стабильности результатов.

Покажем критерий Буша-Винда, позволяющий одновременную проверку различия по положению и масштабу двух выборок: $x_1,...,x_m$ и $y_1,...,y_n$. Обозначим R_i^x – ранг x_i – в совместном вариационном ряду. Найдем инверсные нормальные метки, их квадраты и оценки дисперсий

Таблица 6

Стаж – число нарушений режима

От 3 до 5 лет	от 1 до 3 лет	Менее года
40 (9)	38 (7)	48
35 (11)	40 (6,5)	40
38 (10,5)	47 (1)	45
43 (6,5)	44 (3,5)	43
44 (5)		46
41 (8)		44
		42
Среднее 40,17	Среднее 42,25	Среднее 44,0

$$a_N(R_i^x)=\Phi^{-1}\left(\frac{R_i^x}{N+1}\right),\quad b_N(R_i^x)=a_N^2(R_i^x),\ D_a=\frac{1}{N}\sum_{i=1}^{N}a_N^2(R_i)\ \text{ и }\quad D_b=\frac{1}{N}\sum_{i=1}^{N}b_N^2(R_i),\ \text{где } N=m+n.$$

Вычислим статистики

$$S_{mn}=\sqrt{\frac{N-1}{mn}}\frac{\sum_{i=1}^{m}a_N(R_i^x)}{\sqrt{D_a}}\qquad\text{и}\qquad T_{mn}=\sqrt{\frac{N-1}{mn}}\frac{\sum_{i=1}^{m}b_N(R_i^x)-mD_a}{\sqrt{D_b}}$$

и комбинированную статистику Буша–Винда: $W_{mn} = -2\ln[2(1 - \Phi(|S_{mn}|))] - 2\ln[2(1 - \Phi(|T_{mn}|))]$.

При больших объемах выборок $(\min\{m,n\}>30)$ возможна аппроксимация χ^2-распределением с числом степеней свободы 4 [3, с. 497].

Пример 8. Часть сотрудников $n=28$ прошла тренинг по повышению внимательности. Остальные сотрудники $m=30$ не посещали этих занятий (контрольная группа). Затем в ходе специального тестирования фиксировали число замеченных ими ошибок в тексте. Не приводя таблицы, укажем, что статистика Буша–Винда оказалась равной $W_{mn}=5{,}5052$. Верхняя 5%-я точка χ^2-распределения с числом степеней свободы 4 равна 9,4877. Таким образом, мы не обнаружили значимого сдвига ни в положении, ни в масштабе. Тренинг оказался бесполезным. Характерно, что в таком анализе можно было обнаружить значимый, но негативный результат тренинга.

Заключение. На нескольких продемонстрированных здесь примерах ранговых критериев показана возможность и перспективность использования этих инструментов непараметрической статистики для оперативного контроля и поддержки требуемого уровня информационной безопасности в системах и организациях различного назначения, для обоснования структурных изменений и принятия решений по повышению степени организационной, технической и компьютерной защиты.

Литература

1. Гаек Я. Теория ранговых критериев / Я. Гаек, З. Шидак. – М.: Наука, 1971. – 376 с.

2. Холлендер М. Непараметрические методы статистики / М. Холлендер, Д.А. Вулф. – М.: Финансы и статистика, 1983. – 518 с.

3. Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит, 2006. – 816 с.

4. Хиценко В.Е. Непараметрическая статистика в задачах защиты информации: конспект лекций. – Новосибирск: Изд-во НГТУ (в печати), 2012. – 200 с.

Хиценко Владимир Евгеньевич
Канд. техн. наук, доцент каф. защиты информации Новосибирского государственного университета
Тел.: 8 (383) 346-08-53
Эл. почта: khits@is.cs.nstu.ru

Khitsenko V.E.

Rank tests in the problems of information security

The possibilities and perspectives of rank tests to detect differences in relation to specific problems of technical, organizational and computer security are shown. The methods are classified by situations of dependent and independent samples, the detectable differences in the parameters of position and scale. The presentation is illustrated by examples of professional content.

Keywords: nonparametric statistics, rank tests, protection of information.