1. Непараметрические критерии однородности

1.1. Критерий W-Вилкоксона

Критерий W-Вилкоксона используется для проверки гипотезы однородности двух независимых выборок объема n_1 и n_2, утверждающую, что выборки получены из одной генеральной совокупности.

Для построения статистики критерия выборки $x_1, \ldots, x_{n_1}$ и $y_1, \ldots, y_{n_2}$ объединяют в одну совокупность и упорядочивают ее по возрастанию. Если в этом объединенном вариационном ряду объема $n_1 + n_2 = n$ окажется, что элементы одной выборки в целом больше элементов второй выборки, то понятно, что ни о какой однородности не может быть речи. Измеряют это преобладание при помощи числа инверсий: *если какой-либо элемент из выборки X встречается после какого-то элемента из выборки Y, то говорят, что имеет место одна инверсия.* Для упрощения изложения будем считать, что в общем вариационном ряду нет одинаковых элементов. Число инверсий будем обозначать через $\upsilon = U$, а число контринверсий через $\omega = V$.

Например, в последовательности $x\,y\,y\,x\,y\,x\,y\,y\,x\,x$ число инверсий $U = 15 = 0{+}2{+}3{+}5{+}5$. Максимальное число инверсий равно $U_{max} = 25$ и соответствует $y\,y\,y\,y\,y\,x\,x\,x\,x\,x$, минимальное $x\,x\,x\,x\,x\,y\,y\,y\,y\,y$ и равно $U_{min} = 0$. Очевидно, если число инверсий U близко к нулю, то элементы выборки X в совокупности меньше элементов выборки Y. Вопрос состоит в том, насколько можно объяснить случайностью ситуацию, когда $U > 0$.

Для проверки нулевой гипотезы $H_0 : \upsilon = \omega$, означающей, что число инверсий равно числу контринверсий, при односторонней альтернативной гипотезе $H_1 : \upsilon > \omega$ выбирается критическое число инверсий U_α такое, что если наблюдаемое число инверсий больше критического $U > U_\alpha$, то нулевая гипотеза отвергается в пользу альтернативной с уровнем значимости α. Критическое значение U_α выбирается таким образом, чтобы в случае, когда нулевая гипотеза верна, количество перестановок с числом инверсий $U \geq U_\alpha$, не превышало $\theta_{\text{кр}} = \alpha\theta$, где θ — общее число возможных последовательностей, вычисляемое по формуле:

$$\theta = C_n^{n_1} = \frac{n!}{n_1! n_2!}.$$

Например, $n_1 = 5;\ n_2 = 5,\ \theta = C_{10}^5 = 252$, $\theta_{\text{кр}} = 6.3$ при $\alpha = 0.025$. Подсчет начинают с последовательности с максимальным числом инверсий до тех пор, пока не наберется больше чем $\alpha\theta$ последователь-

ностей. Количество инверсий в последней последовательности принимают за U_α.

$$\begin{array}{lllllllllll} 1\,. & y & y & y & y & y & x & x & x & x & x \quad U_1 = 25 \\ 2\,. & y & y & y & y & x & y & x & x & x & x \quad U_2 = 24 \\ 3\,. & y & y & y & x & y & y & x & x & x & x \quad U_3 = 23 \\ 4\,. & y & y & y & y & x & x & y & x & x & x \quad U_4 = 23 \\ 5\,. & y & y & x & y & y & y & x & x & x & x \quad U_5 = 22 \\ 6\,. & y & y & y & x & y & x & y & x & x & x \quad U_6 = 22 \\ 7\,. & y & y & y & y & x & x & x & y & x & x \quad U_7 = 22 \end{array}$$

В последней последовательности число инверсий равно 22, поэтому $U_{0.025} = 22$. Наблюдаемое число инверсий $U_* = 15 < 22 = U_{0.025}$, поэтому выборки можно считать однородными. При двустороннем критерии нулевая гипотеза отвергается, если

$$U > U_{\alpha/2} \;\text{ или } n_1 n_2 - U > U_{\alpha/2}.$$

Вместо подсчета инверсий можно суммировать ранги (порядок элемента в вариационном ряду).

$$\begin{aligned} U = (r_1 - 1) + (r_2 - 2) + \ldots + (r_{n_1} - n_1) = \\ = \sum r_i - \frac{n_1(n_1+1)}{2}. \end{aligned}$$

1.2. Критерий Манна-Уитни

Представим число инверсий в виде

$$U = \sum_{i=1}^{n_1} \sum_{k=1}^{n_2} Z_{ik}, \quad \text{где} \quad Z_{ik} = \begin{cases} 1 & \text{при } x_i > y_k, \\ 0 & \text{иначе.} \end{cases}$$

Если выборки однородны, то случайная величина

$$Z_{ik} = \begin{cases} 1, & \text{с вероятностью } 0.5, \\ 0, & \text{с вероятностью } 0.5. \end{cases}$$

В этом случае $\mathbf{E}Z_{ik} = \frac{1}{2}$, $\mathbf{E}U = \frac{n_1 n_2}{2}$. Для вычисления дисперсии $\mathbf{D}U$ рассмотрим случайную величину $U - \mathbf{E}U = \sum\limits_{i,k} \left(Z_{ik} - \frac{1}{2}\right)$.

$$\mathbf{D}U = \mathbf{E}(U - \mathbf{E}U)^2 = \sum_{i,k,j,l} \mathbf{E}\left(Z_{ik} - \frac{1}{2}\right)\left(Z_{jl} - \frac{1}{2}\right).$$

Здесь нужно выделить четыре возможных случая:

- Слагаемые с индексами $i \neq j$ и $k \neq l$ равны нулю, так как элементы выборок независимы и

$$\mathbf{E}\left(Z_{ik} - \frac{1}{2}\right)\left(Z_{jl} - \frac{1}{2}\right) = \mathbf{E}\left(Z_{ik} - \frac{1}{2}\right)\mathbf{E}\left(Z_{jl} - \frac{1}{2}\right) = 0.$$

- Пусть $i = j$, $k = l$. Тогда

$$\mathbf{E}\left(Z_{ik} - \frac{1}{2}\right)^2 = \frac{1}{2}\left(1 - \frac{1}{2}\right)^2 + \frac{1}{2}\left(0 - \frac{1}{2}\right)^2 = \frac{1}{4}.$$

- Пусть $i = j$, $k \neq l$. В этом случае рассматриваются три индекса i, k, l, из которых индекс i относится к переменной из выборки X, а индексы k и l – к переменной из выборки Y. В случае однородных выборок возможны шесть равновероятных вариантов расположения этих трех элементов.

	индексы	порядок	Z_{ik}	Z_{il}	$(Z_{ik} - 0.5)(Z_{il} - 0.5)$
1.	(ikl)	$X_iY_kY_l$	0	0	$1/4$
2.	(ilk)	$X_iY_lY_k$	0	0	$1/4$
3.	(kil)	$Y_kX_iY_l$	1	0	$-1/4$
4.	(kli)	$Y_kY_lX_i$	1	1	$1/4$
5.	(lik)	$Y_lX_iY_k$	0	1	$-1/4$
6.	(lki)	$Y_lY_kX_i$	1	1	$1/4$

$$\mathbf{E}\left(Z_{ik} - \frac{1}{2}\right)\left(Z_{il} - \frac{1}{2}\right) =$$
$$= \frac{1}{6}\left(\frac{1}{4} + \frac{1}{4} - \frac{1}{4} + \frac{1}{4} - \frac{1}{4} + \frac{1}{4}\right) = \frac{1}{12}.$$

- В случае $i \neq j,\ k = l$ поступаем аналогично и получаем

$$\mathbf{E}\left(Z_{ik} - \frac{1}{2}\right)\left(Z_{jk} - \frac{1}{2}\right) = \frac{1}{12}.$$

Следовательно,

$$\mathbf{D}U = \left(\frac{1}{4}\right)n_1n_2 + \left(\frac{1}{12}\right)n_1(n_2^2 - n_2) + \left(\frac{1}{12}\right)n_2(n_1^2 - n_1) =$$
$$= \frac{n_1n_2}{12}(n_1 + n_2 + 1).$$

Итак получаем, что число инверисий $U = R_1 - \frac{n_1(n_1+1)}{2}$ имеет математическое ожидание, равное $\mathbf{E}U = \frac{n_1n_2}{2}$ и дисперсию, равную $\mathbf{D}U = n_1n_2(n_1 + n_2 + 1)/12$. При небольших объемах выборок для проверки гипотезы однородности используют по готовым таблицам критические значения числа инверсий или в R функция $wilcox.test$

из пакета $lirbary(stats)$ с опцией $exact = TRUE$. Нормальное приближение (опция $exact = FALSE$) оказывается достаточно точным при $n_1 > 3, n_1 + n_2 > 20$, поэтому для проверки нулевой гипотезы используется статистика

$$Z = \frac{R_1 - \frac{n_1(n_1+1)}{2} - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}} \quad \sim \quad \mathcal{N}(0; 1)\,. \tag{1}$$

1.3. Критерий знаков

Примером зависимых выборок могут служить данные о наблюдениях за массой тела до и после диеты у 10 индивидов.

	1	2	3	4	5	6	7	8	9	10
Масса до диеты	68	80	92	81	70	79	78	66	57	76
Масса после диеты	60	84	87	79	74	71	72	67	57	70
Разность	8	−4	5	2	−4	8	6	−1	0	6

Из 9 индивидов, у которых произошли изменения в массе тела после диеты, шесть индивидов уменьшили массу тела, а трое даже увеличили. Один индивид остался с прежним весом.

Итак, имеется парная выборка (x_{1j}, x_{2j}), $j = 1, \ldots, n$. Рассмотрим разности $z_j = x_{2j} - x_{1j}$, $j = 1, \ldots, n$, в виде выборки наблюдений случайной величины ξ. Если выборки одинаково распределены, то вероятности положительных и отрицательных значений ξ должны быть одинаковы. Вследствие непрерывности случайных величин, $\xi = 0$ с нулевой вероятностью. Если одинаковые наблюдения встречаются из-за погрешности вычислений или ошибок округления, то их следует исключить из выборки. Для определенности будем считать успехом $\xi > 0$. Обозначим через $p = P\{\xi > 0\}$. Нулевая гипотеза имеет вид $H_0 : p = 0.5$.

Пусть из $n = n_+ + n_-$ элементов парной выборки имеется n_+ и n_- разностей со знаком $+$ и $-$, через ν_+ и ν_- обозначим случайное число плюсов и минусов соответственно.

При альтернативной гипотезе $H_1 : p > 0.5$ вычисляется вероятность получить n_- неудач или даже меньше:

$$\alpha_* = P\{\nu_- \le n_-\} = \sum_{k=0}^{n_-} P\{\nu_- = k\} = \frac{1}{2^n} \sum_{k=0}^{n_-} C_n^k.$$

При альтернативной гипотезе $H_1 : p < 0.5$ вычисляется вероятность получить n_+ успехов или даже больше:

$$\alpha_* = P\{\nu_+ \ge n_+\} = \sum_{k=n_+}^{n} P\{\nu_+ = k\} = \frac{1}{2^n} \sum_{k=n_+}^{n} C_n^k.$$

При $H_1 : p \ne 0.5$ вычисляется значимость

$$\alpha_* = \frac{1}{2^{n-1}} \sum_{k=0}^{\min(n_-,n_+)} C_n^k.$$

Если $\alpha_* < \alpha$, то гипотеза о случайности в колебаниях знаков разностей отвергается с уровнем значимости α, иначе нет оснований отвергнуть эту гипотезу.

В R для проверки равентсва нулю медианы разности двух зависимых выборок можно использовать функцию $SIGN.test(x, y)$ из пакета $BSDA$.

Пример. Ниже приводится время (в секундах) решения контрольных задач одиннадцатью учащимися до и после упражнений по устному счету. Можно ли считать, что результативность увеличилась? Принять $\alpha = 0.1$.

до	87	61	98	90	93	74	83	72	81	75	83
после	50	45	79	90	88	65	52	79	84	61	52
разность	37	16	19	0	5	9	31	−7	−3	14	31
знак разности	+	+	+	0	+	+	+	−	−	+	+
ранги абс.разн.	10	6	7		2	4	8.5	3	1	5	8.5

Считаем только ненулевые разности, $n = 10$, $n_=2$, $n_+ = 8$. Двусторонняя значимость имеет вид $\alpha_* = 2(C_{10}^2 + C_{10}^1 + C_{10}^0) \cdot \frac{1}{2^{10}} = \frac{11}{2^9} = 0.109 > 0.1$,

поэтому изменения нельзя считать значимыми. Пограничное значение от 0.05 до 0.2 характеризует неопределенность ситуации. В таком случае рекомендуется применить другой критерий, учитывающий величины положительных и отрицательных разностей.

1.4. Критерий Вилкоксона для зависимых выборок

Также имеется парная выборка (x_{1j}, x_{2j}), $j = 1, \ldots, n$, и разности $z_j = x_{2j} - x_{1j}$. Выборку из модулей разностей $|z_j|$ проранжируем по возрастанию. Вычислим суммы рангов S_+ и S_-, соответствующие положительным и отрицательным разностям. Если все разности одного знака, то

$$\min\{S_+, S_-\} = 0, \quad \max\{S_+, S_-\} = \frac{n(n+1)}{2}.$$

При условии однородности зависимых выборок

$$E(S_-) = \frac{n(n+1)}{4},$$
$$D(S_-) = \frac{n(n+1)(2n+1)}{24}.$$

Таким образом, для проверки гипотезы однородности зависимых выборок используется статистика

$$Z = \frac{S_- - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}},$$

которая при справедливости нулевой гипотезы асимптотически имеет стандартное нормальное распределение.

В R используется функция $wilcox.test$ с опцией $paired = TRUE$.

В примере из предыдущего параграфа $S_- = 4$, $n = 10$,

$$Z = \frac{4 - \frac{10 \cdot 9}{4}}{\sqrt{\frac{10 \cdot 11 \cdot 21}{24}}} = -2.4 < -1.96,$$

различие нельзя объяснить случайностью.

1.5. Критерий Фридмана

Этот критерий предназначен для проверки однородности нескольких зависимых выборок или для сопоставления условий эксперимента.

Имеется n индивидов, у которых измеряется значение одного и того же признака при c условиях. Предполагаем, что между полученными в разных условиях измерениями имеются случайные различия. Наблюдениям вида

$$\begin{matrix} x_{11} & x_{12} & \dots & x_{1c} \\ x_{21} & x_{22} & \dots & x_{2c} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nc} \end{matrix}$$

сопоставляются ранги по строкам

$$\begin{matrix} r_{11} & r_{12} & \dots & r_{1c} \\ r_{21} & r_{22} & \dots & r_{2c} \\ \vdots & \vdots & \dots & \vdots \\ r_{n1} & x_{r2} & \dots & r_{nc} \end{matrix}$$

и вычисляеются суммы рангов $R_i = \sum\limits_{j=1}^{n} r_{ij}$, средний ранг $\bar{R}_i = \frac{1}{n} R_i$, соответствующий условию $i = 1, 2, \dots, c$, и общий средний ранг вида $R = \frac{c+1}{2}$. Статистика критерия имеет вид

$$S = \frac{12n}{c(c+1)} \sum_{i=1}^{c} (\bar{R}_i - R)^2,$$

которая при $n \geq 13$, $c \geq 20$ имеет приближенное распределение хи-квадрат с $c - 1$ степенями свободы. При меньших значениях параметров используются специальные вычисления. Пример кода в R:

library(stats); A <- matrix(c(52, 45, 38, 63, 79, 50, 45, 57, 39, 53, 51, 43, 47, 50, 56, 62, 72, 49, 49, 52, 40), byrow=TRUE, ncol=3)

$friedman.test(A)$

$p - value = 0.049$

1.6. Критерий Краскела-Уоллиса

Критерий Краскела-Уоллиса является непараметрическим аналогом однофакторного дисперсионного анализа и применяется для проверки однородности нескольких независимых выборок. Пусть имеются наблюдения некоторого признака в r группах, $n_1 + \ldots + n_r = n$. Метод заключается в определении рангов каждого из элементов смешанной выборки и вычислении суммы рангов $S_i,\ i = 1, \ldots, r$, по каждой выборке. Однородность по группам проверяется при помощи статистики

$$\chi^2 = \frac{12}{n(n+1)} \sum_{i=1}^{r} \frac{S_i^2}{n_i} - 3(n+1)\,,$$

которая при H_0 имеет распределение $\chi^{\mathbf{2}}(r-1)$.

Если в объединенной выборке имеются m групп одинаковых наблюдений размера t_i, $i = 1, 2, \ldots, m$, то ранги, относящиеся к ним, усредняются, вычисляются поправочные коэффициенты $T_i = t_i^3 - t_i$, а значение статистики H меняется на

$$H' = \frac{H}{T}, \quad \text{где} \quad T = 1 - \sum_{i=1}^{m} \frac{T_i}{n^3 - n}\,.$$

Пример кода в R $library(stats)$; $kruskal.test(yield \sim block, npk)$ и его результат $Kruskal-Wallischi-squared = 8.3684, df = 5, p-value = 0.1371$. Для включения в анализ множественных сравнений используется функция $kruskal$ из $library(agricolae)$.

Медианный критерий

Медианный тест применяется для проверки однородности нескольких независимых выборок. Пусть меются наблюдения в r группах, $n_1+\ldots+n_r = n$. Через $x^{(1)} \leq \ldots \leq x^{(n)}$ обозначим объединенный вариационный ряд, n_{1i}

и n_{2i} — число наблюдений в i-й группе соответственно больше и меньше медианы

$$M = \begin{cases} \frac{x^{(m)}+x^{(m+1)}}{2} & \text{при } n = 2m, \\ x^{(m+1)} & \text{при } n = 2m+1. \end{cases}$$

	Группы			
	1	...	r	
$< M$	n_{11}	...	n_{1r}	n_{1*}
$> M$	n_{21}	...	n_{2r}	n_{2*}
	n_{*1}	...	n_{*r}	n

При однородности выборок статистика

$$\chi^2 = \sum_{j=1}^{r} \sum_{i=1}^{2} \frac{(n_{ij} - \frac{n_{*j}n_{i*}}{n})^2}{\frac{n_{*j}n_{i*}}{n}}$$

имеет распределение хи-квадрат с $r-1$ степенями свободы. При значении $p = \{\chi^2 > \chi^2_*\}$, меньшем α, гипотеза однородности отвергается, и с уровнем значимости α внутригрупповые отклонения от медианы невозможно объяснить случайностью.

Пример кода в $R\, library(nonparam); mediantest(x, y, alpha = NULL, exact = FALSE)$ или $Median.test(y, trt, alpha = 0.05, correct = TRUE, simulate.p.value = FALSE, group = TRUE, main = NULL, console = TRUE)$ из $library(agricolae)$ или $library(RVAideMemoire)$ $median\_test(yield \sim block, npk)$ и его результат $chi - squared = 7.6667, df = 5, p - value = 0.1756$.