Методы современной прикладной статистики
6. Критерии однородности.

Родионов Игорь Владимирович
vecsell@gmail.com

Весна, 2018

Виды задач:

1. Одновыборочные: среднее (медиана) выборки равно заданному числу;
2. Двухвыборочные:
    - Средние (медианы) выборок равны;
        - Выборки связанные;
        - Выборки независимые;
    - Дисперсии выборок равны;
    - Распределения выборок совпадают.

Предположим, что $(X_1, \ldots, X_n)$ – выборка из $N(\mu, \sigma^2)$. Проверим гипотезу $H_0 : \mu = \mu_0$ против альтернативы $H_1 : \mu \neq \mu_0$. Если

$$\left| \sqrt{n}\frac{\overline{X} - \mu_0}{s} \right| > t_{1-\alpha/2},$$

то отвергаем H_0 на уровне значимости α. Здесь $t_{1-\alpha/2}$ – квантиль распределения Стьюдента с $n-1$ степенью свободы, s – корень из выборочной дисперсии.

Если неожиданно оказалось, что мы знаем дисперсию выборки, то в критерии s заменяется на σ, квантиль – на квантиль $N(0, 1)$ того же уровня, а сам тест будет называться Z-тестом.

Предположим, что $(X_1, \ldots, X_n) \sim N(\mu_1, \sigma^2)$, $(Y_1, \ldots, Y_m) \sim N(\mu_2, \sigma^2)$, т.е. дисперсии распределений одинаковы, причем σ неизвестна, и что выборки независимы. Проверим гипотезу $H_0 : \mu_1 = \mu_2$ против альтернативы $H_1 : \mu_1 \neq \mu_2$. Если

$$\left| \sqrt{\frac{nm}{n+m}} \cdot \frac{\overline{X} - \overline{Y}}{S} \right| > t_{1-\alpha/2},$$

то отвергаем H_0 на уровне значимости α. Здесь $t_{1-\alpha/2}$ – квантиль распределения Стьюдента с $n+m-2$ степенью свободы, а

$$S = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}}.$$

Но как проверить, что дисперсии двух нормальных выборок равны? Для этого воспользуемся критерием Фишера.

Пусть $(X_1, \ldots, X_n) \sim N(\mu_1, \sigma_1^2)$, $(Y_1, \ldots, Y_m) \sim N(\mu_2, \sigma_2^2)$, выборки независимы. Проверим гипотезу $H_0 : \sigma_1 = \sigma_2$ против $H_1 : \sigma_1 \neq \sigma_2$. Если

$$\frac{s_X^2}{s_Y^2} \notin (u_{\alpha/2}, u_{1-\alpha/2}),$$

то отвергаем H_0. Здесь u_γ – квантиль распределения Фишера с $n-1$ и $m-1$ степенями свободы.

Критерий Фишера, как и t-критерии, крайне чувствителен к отклонению выборок от нормального распределения.

Критерий Аспина-Уэлча

Что делать, если выяснилось, что дисперсии выборок различны? Пусть, как и ранее, $(X_1, \ldots, X_n) \sim N(\mu_1, \sigma_1^2)$, $(Y_1, \ldots, Y_m) \sim N(\mu_2, \sigma_2^2)$, выборки независимы. Тогда при верной гипотезе $H_0 : \mu_1 = \mu_2$ статистика

$$\frac{\overline{X} - \overline{Y}}{\sqrt{s_X^2/n + s_Y^2/m}}$$

приблизительно распределена по закону Стьюдента с K степенями свободы, где

$$K = \left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2 \cdot \left(\frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)}\right)^{-1} - 2.$$

Парный t-тест

А что делать, если выборки оказались зависимыми? В этом случае проверку гипотезы $H_0 : \mu_1 = \mu_2$ можно свести к одномерной.

Пусть $(X_1, \ldots, X_n) \sim N(\mu_1, \sigma^2)$, $(Y_1, \ldots, Y_n) \sim N(\mu_2, \sigma^2)$, выборки зависимы и одинакового размера. При верной гипотезе H_0 выполнено

$$\sqrt{n} \cdot \frac{\overline{X} - \overline{Y}}{S} \sim St(n-1),$$

где $D_i = X_i - Y_i$ и

$$S = \frac{1}{\sqrt{n-1}} \sqrt{\sum_{i=1}^{n} (D_i - \overline{D})^2}.$$

Одновыборочный критерий знаков

Обсудим теперь ситуацию, когда выборки не являются нормальными. Проще всего в таком случае перейти к непараметрическим тестам.

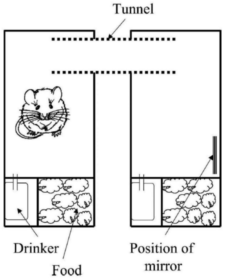
Имеем выборку $X = (X_1, \ldots, X_n)$, где $X_i \neq m_0$. Хотим проверить гипотезу $H_0 : med(X) = m_0$ против $H_1 : med(X) \neq m_0$, где $med(X)$ – медиана распределения выборки X.

Используем для этого статистику $T(X) = \sum_i I(X_i > m_0)$. Ясно, что при верной гипотезе H_0

$$T \sim Bin(n, 1/2).$$

Одновыборочный критерий знаков

Пример: (Shervin, 2004) 16 лабораторных мышей были помещены в двухкомнатные клетки, в одной из комнат висело зеркало. Измерялось доля времени, которое каждая мышь проводила в каждой из своих двух клеток.

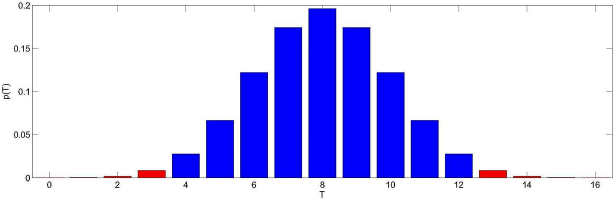


Постановка задачи:
H_0 : мышам всё равно, висит в клетке зеркало или нет.
H_1 : у мышей есть какие-то предпочтения насчёт зеркала.

Одновыборочный критерий знаков

Будем действовать в рамках такой постановки: H_0 : медиана доли времени, проводимого в клетке с зеркалом, равна $1/2$, H_1 : медиана доли времени, проводимого в клетке с зеркалом, не равна $1/2$.

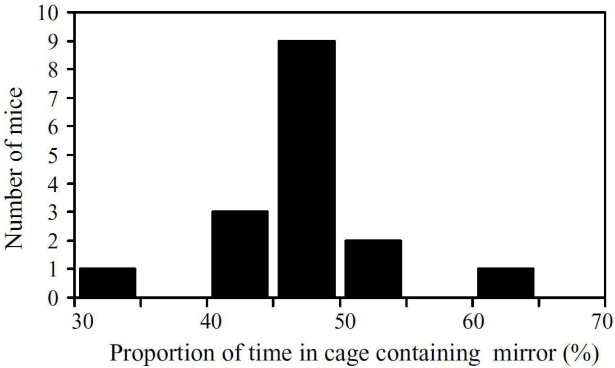
Редуцированные данные: 0 – мышь провела больше времени в комнате с зеркалом, 1 – в комнате без зеркала.
Статистика: T – число единиц в выборке.



13 из 16 мышей провели больше времени в комнате без зеркала.

Критерий знаков: $p = 0.0213$, 95% доверительный интервал для вероятности (что мышь проведёт больше времени в комнате без зеркала) – $[0.54, 0.96]$.

Одновыборочный критерий знаков



По этой гистограмме видим, что отклонение гипотезы об индиффирентности мышей по отношению зеркалам, пожалуй, обосновано.

Двухвыборочный критерий знаков

Пусть зависимость между выборками $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_n)$ неизвестна, причем (желательно) $X_i \neq Y_i \; \forall i$. Потребуем, чтобы распределение выборки величин $\{X_i - Y_i\}_{i=1}^n$ было симметричным.

Хотим проверить гипотезу об отсутствии эффекта
$H_0 : P(X > Y) = 1/2$ против альтернативы
$H_1 : P(X > Y) \neq 1/2$.

Для проверки гипотезы используем статистику $T(X, Y) = \sum_i I(X_i > Y_i)$. Ясно, что при верной H_0

$$T(X, Y) \sim Bin(n, 1/2).$$

Критерий резонно использовать, когда 1) точные разности $X_i - Y_i$ неизвестны, известны только их знаки; 2) разности небольшие по модулю, но систематические по знаку; 3) разности большие по модулю, но случайные по знаку.

Одновыборочный критерий знаковых рангов Уилкоксона

Более точным по сравнению с критерием знаков является критерий Уилкоксона, однако он требует, чтобы распределение выборки было симметричным.

Имеем выборку $X = (X_1, \ldots, X_n)$ из симметричного распределения, где $X_i \neq m_0$. Хотим проверить гипотезу $H_0 : med(X) = m_0$ против $H_1 : med(X) \neq m_0$.

Рассмотрим статистику

$$W = \sum_i rank(|X_i - m_0|) sign(X_i - m_0).$$

Тогда при верной гипотезе H_0 статистика W имеет табличное распределение. При $n > 20$ (и верной H_0) можно использовать аппроксимацию

$$W \sim N(0, n(n+1)(2n+1)/6).$$

Двухвыборочный критерий знаковых рангов Уилкоксона

Пусть зависимость между выборками $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_n)$ неизвестна, причем (желательно) $X_i \neq Y_i \; \forall i$. Потребуем, чтобы распределение выборки величин $\{V_i\}_{i=1}^n$, где $V_i = X_i - Y_i$ было симметричным.

Для проверки гипотезы $H_0 : med(X - Y) = 0$ рассмотрим статистику

$$T = \sum_i rank(|V_i|) I(V_i > 0).$$

Тогда при верной гипотезе H_0 статистика W имеет табличное распределение. При $n > 20$ (и верной H_0) можно использовать аппроксимацию

$$T \sim N(n(n+1)/4, n(n+1)(2n+1)/24).$$

Замечания о знаковых тестах

1) При $n < 20$ пользоваться нормальной аппроксимацией не рекомендуется, следует использовать табличные значения квантилей.

2) Чтобы проверить симметричность распределения, можно нанести на график точки (ξ_i, η_i), где $\xi_i = -V_{(i)} + \widehat{\mu}$, $\eta_i = V_{(n-i+1)} - \widehat{\mu}$, $i = 1, \ldots, [n/2]$, $\widehat{\mu}$ – выборочная медиана. Точки должны приближаться прямой $y = x$.

3) Если некоторые V_i (или другие разности) равны 0, то отбрасываем их и уменьшаем n.

4) Если среди $|V_i|$ есть совпадения, то следует использовать средние ранги (дисперсия нормального приближения тоже изменится).

5) Вместо статистики T в критерии знаковых рангов Уилкоксона может использоваться статистика $W = \sum_i rank(|V_i|)sign(V_i)$.

Критерий Манна-Уитни-Уилкоксона

Пусть $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$ – 2 независимые выборки с функциями распределения F_X и F_Y соответственно, причем $F_X(x - \theta) = F_Y(x)$ и $n < m$. Проверим гипотезу об отсутствии сдвига $H_0 : \theta = 0$ против $H_1 : \theta \neq 0$.

Составим вариационный ряд объединенной совокупности $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ и обозначим через R_i ранг наблюдения X_i в этом вариационном ряду. Определим статистику Манна-Уитни

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

Тогда при верной H_0 W имеет табличное распределение.

Критерий Манна-Уитни-Уилкоксона

1) При $n, m > 20$ и верности гипотезы H_0 можно пользоваться нормальной аппроксимацией

$$W \sim N\left(\frac{n(n+m+1)}{2}, \frac{nm(n+m+1)}{2}\right).$$

В остальных случаях стоит пользоваться табличными значениями квантилей критерия.

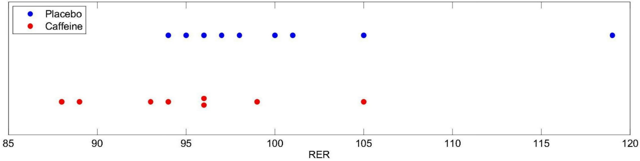
2) Если некоторые наблюдения совпадают (т.е. им присвоены средние ранги), то в аппроксимации следует заменить дисперсию на следующую:

$$\frac{mn(m+n+1)}{2}\left(1 - \frac{\sum_{i=1}^{k} t_i(t_i^2-1)}{(m+n)((m+n)^2-1)}\right),$$

где k – число групп совпавших величин и t_i – число величин в i-той группе.

Критерий Манна-Уитни-Уилкоксона

Пример. RER (респираторный обмен) – соотношение числа молекул CO_2 и O_2 в выдыхаемом воздухе. Является косвенным признаком того, из жиров или углеводов вырабатывается энергия в момент измерения. Изучалось влияние кофеина на мышечный метаболизм. В эксперименте принимало участие 18 испытуемых, респираторный обмен которых измерялся в процессе физических упражнений. За час до этого 9 из них получили таблетку кофеина, оставшиеся 9 – плацебо. Повлиял ли кофеин на значение показателя респираторного обмена?



H_0 : среднее значение RER не отличается в двух группах.
H_1 : среднее значение RER отличается в двух группах.

Критерий Манна-Уитни-Уилкоксона

Ранг	Наблюдение	Номер наблюдения	Наблюдение	Ранг
16.5	105	1	96	9
18	119	2	99	13
14	100	3	94	5.5
11	97	4	89	3
9	96	5	96	9
15	101	6	93	4
5.5	94	7	88	1.5
7	95	8	105	16.5
12	98	9	88	1.5

Статистика W – сумма рангов одной из групп (т.к. $n = m = 9$, то неважно, какой).

$p = 0.0521$, 95% доверительный интервал для медианной разности – $[0.00005, 1.2]$.

С помощью ранговых методов возможно даже проверить гипотезу о равенстве дисперсий двух выборок. Пусть $(X_1, \dots, X_n)$ и $(Y_1, \dots, Y_m)$ – две независимые выборки. Проверим гипотезу $H_0 : \mathrm{D}X_1 = \mathrm{D}X_2$ против альтернативы $H_1 : \mathrm{D}X_1 \neq \mathrm{D}X_2$.

Пусть $N = n + m$, $n < m$. Рассмотрим вариационный ряд объединенной совокупности $\{Z_i\}_{i=1}^N = \{X_j\}_{j=1}^n \cup \{Y_k\}_{k=1}^m$, причем ранги наблюдениям будем присваивать следующим образом:

$$\begin{array}{lccccccccccccc} \widetilde{Z_{(i)}} & Z_{(1)} & \le & Z_{(2)} & \le & Z_{(3)} & \le & \dots & \le & Z_{(N-2)} & \le & Z_{(N-1)} & \le & Z_{(N)} \\ rank(Z_i) & 1 & & 4 & & 5 & & & & 6 & & 3 & & 2 \end{array}$$

При верности H_0 статистика

$$R = \sum_{j=1}^{n} \widetilde{rank}(X_j)$$

имеет табличное распределение. Критерий должен быть двусторонним.

При $m \geq n > 10$ и верности гипотезы H_0 можно пользоваться нормальной аппроксимацией

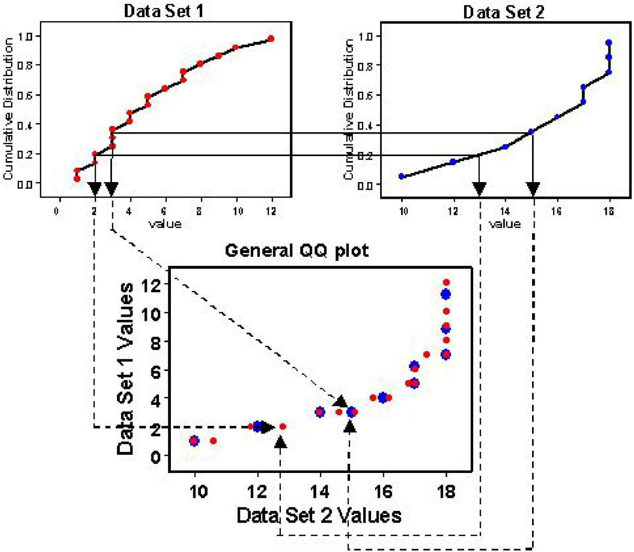
$$R \sim N\left(\frac{n(n+m+1)}{2}, \frac{nm(n+m+1)}{2}\right).$$

В остальных случаях стоит пользоваться табличными значениями квантилей критерия (которые совпадают с квантилями критерия Манна-Уитни).

Пусть имеются две (независимые) выборки $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_n)$ с функциями распределения F и G соответственно. Допустим, мы хотим проверить гипотезу $H_0 : F = G\left(\frac{x-a}{\sigma}\right)$.

General QQ-plot – это график, на который нанесены точки $\left(\widehat{F}_n^{-1}\left(\frac{j}{m}\right), Y_{(j)}\right)$ и $\left(X_{(i)}, \widehat{G}_m^{-1}\left(\frac{i}{n}\right)\right)$. Если точки лежат примерно на одной прямой, то гипотеза H_0 близка к верности.

General QQ-plot



Обсудим теперь критерии проверки двух выборок на однородность. Для решения этой задачи можно адаптировать критерии согласия, например, критерий Колмогорова-Смирнова.

Пусть $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$ – две независимые выборки с непрерывными ф.р. F и G соответственно, а $\widehat{F}_n(x)$ и $\widehat{G}_m(x)$ – эмпирические функции распределения этих выборок. Определим

$$D_{n,m} = \sup_x |\widehat{F}_n(x) - \widehat{G}_m(x)|,$$

тогда при верной гипотезе $H_0 : F = G$ статистика $\sqrt{\frac{nm}{n+m}} D_{n,m}$ имеет табличное распределение. При $n, m \geq 20$ оно приближается распределением Колмогорова с ф.р. $F_K(z) = \sum_{k \in \mathbb{Z}} (-1)^k e^{-2k^2 z^2}$.

ω^2-критерий Розенблатта

Также для проверки гипотезы $H_0 : F = G$ можно воспользоваться модификацией критерий Крамера-фон Мизеса. Определим

$$\widehat{H}_{m,n}(x) = \tfrac{n}{n+m}\widehat{F}_n(x) + \tfrac{m}{n+m}\widehat{G}_m(x)$$

эмпирическую функцию распределения совокупности $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ и статистику критерия

$$\omega^2_{m,n} = \textstyle\int_{\mathbb{R}}(\widehat{F}_n(x) - \widehat{G}_m(x))^2 d\widehat{H}_{m,n}(x).$$

Тогда при верной гипотезе H_0 и достаточно больших n, m статистика $Z := \frac{nm}{n+m}\omega^2_{m,n}$ приближается распределением A_1, приведем таблицу квантилей этого распределения:

α	0.5	0.85	0.9	0.95	0.975	0.99
$y_{1-\alpha}$	0.12	0.28	0.35	0.46	0.58	0.74

1) Статистику критерия Розенблатта можно представить в явной форме

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

где R_i и S_j – ранги наблюдений $X_{(i)}$ и $Y_{(j)}$ в объединенной совокупности.

2) Известно, что

$$\mathrm{E}Z = \frac{1}{6}\left(1 + \frac{1}{n+m}\right), \; \mathrm{D}Z = \frac{1}{45}\left(1 + \frac{1}{n+m}\right)\left(1 + \frac{1}{n+m} - \frac{3}{4}\frac{n+m}{nm}\right).$$

Тогда статистика

$$Z^* = \frac{Z - \mathrm{E}Z}{\sqrt{45\mathrm{D}Z}} - \frac{1}{6}$$

хорошо приближается распределением A_1 уже при $n, m \geq 7$.

Общий критерий Андерсона-Дарлинга

Проблема критерия Розенблатта в том, что он не реализован в Python. Зато реализован критерий Андерсона-Дарлинга для нескольких ($k \geq 2$) выборок.

Пусть $(X_1^{(1)}, \ldots, X_{n_1}^{(1)}), \ldots, (X_1^{(k)}, \ldots, X_{n_k}^{(k)})$ – k независимых выборок с функциями распределений $F_1, \ldots, F_k$ соответственно. Пусть $\widehat{F}_1, \ldots, \widehat{F}_k$ – эмпирические функции распределения этих выборок и $\widehat{H}_N(x), N = \sum_i n_i,$ – эмпирическая функция распределения общей совокупности наблюдений.

Общий критерий Андерсона-Дарлинга

Тогда статистика

$$\Omega^2 = \sum_{i=1}^{k} n_i \int_{\mathbb{R}} \frac{(\widehat{F}_i(x) - \widehat{H}_N(x))^2}{\widehat{H}_N(x)(1 - \widehat{H}_N(x))} d\widehat{H}_N(x)$$

имеет табличное распределение при верной гипотезе $H_0 : F_1 = \ldots = F_k$.

Finita!