Лекция 7. Проверка гипотез о равенстве параметров двух нормально распределенных генеральных совокупностей. Однофакторный дисперсионный анализ

Буре В.М., Грауэр Л.В.

ШАД

Санкт-Петербург, 2013

Содержание

Распределение Фишера

Определение 1

Пусть случайные величины $\eta \sim \chi^2_m$, $\xi \sim \chi^2_n$ *независимы. Будем говорить, что случайная величина*

$$\zeta = \frac{\eta/m}{\xi/n} \sim \mathcal{F}_{m,n}$$

подчиняется распределению Фишера со степенями свободы числителя m *и знаменателя* n.

Лемма 1

Плотность распределения случайной величины $\zeta \sim \mathcal{F}_{m,n}$ *имеет вид:*

$$f_\zeta(z) = \begin{cases} \dfrac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \dfrac{m^{\frac{m}{2}} n^{\frac{n}{2}} z^{\frac{m}{2}-1}}{(n+mz)^{\frac{m+n}{2}}}, & \text{если } z > 0; \\ 0, & \text{если } z \leqslant 0. \end{cases}$$

Доказательство

Рассмотрим случайную величину $\tilde{\zeta} = \eta/\xi$ и напишем для нее плотность распределения, считая, что $z > 0$:

$$f_{\tilde{\zeta}}(z) = \int_0^\infty x f_\xi(x) f_\eta(zx) dx = \frac{z^{\frac{m}{2}-1}}{2^{\frac{n+m}{2}} \Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)} \int_0^\infty x^{\frac{m+n}{2}-1} e^{-\frac{x}{2}(z+1)} dx.$$

Далее сделаем замену переменных: $y = x(z+1)/2$, $dx = 2dy/(1+z)$, $x = 2y/(1+z)$. После преобразований получаем формулу плотности распределения для случайной величины $\tilde{\zeta}$:

$$f_{\tilde{\zeta}}(z) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}}, \quad z > 0.$$

Как нетрудно заметить, $\zeta = (n/m)\tilde{\zeta}$, следовательно,

$$f_\zeta(z) = \frac{m}{n} f_{\tilde{\zeta}}\left(\frac{m}{n} z\right).$$

Критерий Фишера

Пусть имеются две независимые генеральные совокупности η с выборкой $X_{[m]}$ и ξ с выборкой $Y_{[n]}$. Будем считать, что η подчиняется нормальному распределению $N(a_1, \sigma_1^2)$ и ξ подчиняется нормальному распределению $N(a_2, \sigma_2^2)$, причем, математические ожидания a_1, a_2 и дисперсии σ_1^2 и σ_2^2 неизвестны.
Сформулируем нулевую и альтернативную гипотезы:

- $H_0 : \sigma_1^2 = \sigma_2^2$.
- $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Альтернативная гипотеза является двусторонней.

Тогда

$$(m-1)\tilde{s}_X^2/\sigma_1^2 \sim \chi^2_{m-1},$$
$$(n-1)\tilde{s}_Y^2/\sigma_2^2 \sim \chi^2_{n-1},$$

где

$$\tilde{s}_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2,$$

$$\tilde{s}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

В соответствии с определением распределения Фишера:

$$F_{m-1,n-1} = \frac{\tilde{s}_X^2}{\tilde{s}_Y^2}\frac{\sigma_2^2}{\sigma_1^2} \sim \mathcal{F}_{m-1,n-1}.$$

Тогда при справедливости нулевой гипотезы:

$$\frac{\tilde{s}_X^2}{\tilde{s}_Y^2} \sim \mathcal{F}_{m-1,n-1}.$$

Пусть $u_{\frac{\alpha}{2}}$, $u_{1-\frac{\alpha}{2}}$ — квантили распределения Фишера $\mathcal{F}_{m-1,n-1}$. Можно сформулировать критерий проверки гипотезы H_0 при альтернативе H_1 с вероятностью ошибки первого рода α (уровнем значимости критерия):

- Если $\frac{\tilde{s}_X^2}{\tilde{s}_Y^2} \in [u_{\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}]$, то принимается гипотеза H_0.
- Если $\frac{\tilde{s}_X^2}{\tilde{s}_Y^2} \notin [u_{\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}]$, то принимается гипотеза H_1.

Если альтернативная гипотеза H_1 односторонняя, т. е. $\sigma_1^2 > \sigma_2^2$, то в качестве критической области для нулевой гипотезы рассматривается $S = (u_{1-\alpha}, +\infty)$, где α — вероятность ошибки первого рода. Случай $\sigma_2^2 > \sigma_1^2$ сводится к предыдущему переменой мест выборок.

Критерий Стьюдента

Пусть заданы независимые случайные величины $\eta \sim N(a_1, \sigma_1^2)$ с выборкой $X_{[m]}$ и $\xi \sim N(a_2, \sigma_2^2)$ с выборкой $Y_{[n]}$. Сформулируем нулевую и альтернативную двустороннюю гипотезы:

- $H_0 : a_1 = a_2$.
- $H_1 : a_1 \neq a_2$.

Рассмотрим три случая:

1. Дисперсии σ_1^2, σ_2^2 известны.
2. Дисперсии неизвестны, но есть основания считать, что $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
3. Дисперсии неизвестны и неравны $\sigma_1^2 \neq \sigma_2^2$.

Дисперсии σ_1^2, σ_2^2 известны

1) Дисперсии σ_1^2, σ_2^2 известны, тогда используем статистику

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1),$$

которая при условии, что верна гипотеза H_0, подчиняется стандартному нормальному распределению.

Получаем критерий с вероятностью ошибки первого рода α:

- Если справедливо неравенство:

$$\frac{\left|\bar{X} - \bar{Y}\right|}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \leqslant u_{1-\frac{\alpha}{2}},$$

  то принимается гипотеза H_0, где $u_{1-\frac{\alpha}{2}}$ — квантиль уровня $1 - \alpha/2$ стандартного нормального распределения.

- Если справедливо неравенство:

$$\frac{\left|\bar{X} - \bar{Y}\right|}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} > u_{1-\frac{\alpha}{2}},$$

  то принимается гипотеза H_1.

Для правосторонней альтернативной гипотезы $H_1 : a_1 > a_2$ критическая область для H_0 будет выглядеть следующим образом:

$$S = (u_{1-\alpha}, +\infty).$$

Для левосторонней альтернативной гипотезы $H_1 : a_1 < a_2$ критическая область для H_0 будет следующей:

$$S = (-\infty, u_\alpha).$$

Дисперсии неизвестны, но $\sigma_1^2 = \sigma_2^2 = \sigma^2$

2) Дисперсии неизвестны, но есть основания считать, что $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
Статистики

$$\frac{(m-1)\tilde{s}_X^2}{\sigma^2}, \quad \frac{(n-1)\tilde{s}_Y^2}{\sigma^2},$$

которые были введены ранее, взаимно независимы и имеют распределения χ^2_{m-1} и χ^2_{n-1} соответственно, тогда статистика

$$\frac{\tilde{s}_X^2(m-1) + \tilde{s}_Y^2(n-1)}{\sigma^2}$$

подчиняется распределению хи-квадрат с $m+n-2$ степенями свободы.

Если верна гипотеза H_0, то статистика

$$t^{(0)}_{n+m-2} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{m+n}{mn}}\sqrt{\frac{\tilde{s}^2_X(m-1)+\tilde{s}^2_Y(n-1)}{m+n-2}}}$$

подчиняется распределению Стьюдента T_{n+m-2} с $n+m-2$ степенями свободы ($t^{(0)}_{n+m-2}$ — дробь Стьюдента).
Если верна гипотеза H_1, то статистика

$$t_{n+m-2} = \frac{\bar{X} - \bar{Y} + (a_2 - a_1)}{\sqrt{\frac{m+n}{mn}}\sqrt{\frac{\tilde{s}^2_X(m-1)+\tilde{s}^2_Y(n-1)}{m+n-2}}}$$

подчиняется распределению Стьюдента T_{n+m-2} с $n+m-2$ степенями свободы.

Сформулируем критерий с вероятностью ошибки первого рода α:

- Если $|t^{(0)}_{n+m-2}| \leqslant t_{1-\frac{\alpha}{2}}$, где $t_{1-\frac{\alpha}{2}}$ — квантиль распределения Стьюдента с $n+m-2$ степенями свободы уровня $1-\alpha/2$, то принимается гипотеза H_0.
- Если $|t^{(0)}_{n+m-2}| > t_{1-\frac{\alpha}{2}}$, то принимается гипотеза H_1.

Для правосторонней альтернативной гипотезы $H_1 : a_1 > a_2$ критическая область для H_0 при использовании статистики t^0_{m+n-2} будет следующей: $S = (t_{1-\alpha}, +\infty)$.
Для левосторонней альтернативной гипотезы $H_1 : a_1 < a_2$ критическая область для H_0 при использовании статистики t^0_{n+m-2} будет следующей: $S = (-\infty, t_\alpha)$.
Границы $t_{1-\alpha}$ и t_α — квантили распределения Стьюдента с $(n+m-2)$ степенями свободы уровней $1-\alpha$ и α соответственно.

Дисперсии неизвестны и неравны $\sigma_1^2 \neq \sigma_2^2$

3) Пусть дисперсии неизвестны и неравны $\sigma_1^2 \neq \sigma_2^2$.
Если верна гипотеза H_0, то статистика

$$t_K = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\tilde{s}_X^2}{m} + \frac{\tilde{s}_Y^2}{n}}}$$

подчиняется распределению Стьюдента T_K с K степенями свободы:

$$K = \frac{\left(\frac{\tilde{s}_X^2}{m} + \frac{\tilde{s}_Y^2}{n}\right)^2}{\frac{(\tilde{s}_X^2/m)^2}{m-1} + \frac{(\tilde{s}_Y^2/n)^2}{n-1}} \tag{1}$$

Сформулируем критерий с вероятностью ошибки первого рода α:

- Если $|t_K| \leqslant t_{1-\frac{\alpha}{2}}$, где $t_{1-\frac{\alpha}{2}}$ — квантиль распределения Стьюдента с K степенями свободы уровня $1-\alpha/2$, то принимается гипотеза H_0.
- Если $|t_K| > t_{1-\frac{\alpha}{2}}$, то принимается гипотеза H_1.

Для правосторонней альтернативной гипотезы $H_1 : a_1 > a_2$ критическая область для H_0 при использовании статистики t_K будет следующей: $S = (t_{1-\alpha}, +\infty)$.
Для левосторонней альтернативной гипотезы $H_1 : a_1 < a_2$ критическая область для H_0 при использовании статистики t_K будет следующей: $S = (-\infty, t_\alpha)$.
Границы $t_{1-\alpha}$ и t_α — квантили распределения Стьюдента с K степенями свободы уровней $1-\alpha$ и α соответственно.

Критерий Стьюдента для парных выборок

Пусть задана двумерная случайная величина η, ξ с парной выборкой $(X, Y)_{[n]}$. Сформулируем нулевую и альтернативную двустороннюю гипотезы:

- $H_0 : a_1 = a_2$.
- $H_1 : a_1 \neq a_2$.

a_1 — математическое ожидание η, a_2 — математическое ожидание ξ.

Рассмотрим случайную величину $\zeta = \eta - \xi$.
Тогда $Z_i = X_i - Y_i$, $i = 1, \ldots, n$, — выборка наблюдений случайной величины ζ.
Проверяемые гипотезы примут вид

- $H_0 : a = 0$.
- $H_1 : a \neq 0$.

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \bar{X} - \bar{Y},$$

$$S_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}((X_i - Y_i) - \bar{Z})^2.$$

Отметим, что в случае положительной корреляции между η и ξ дисперсия $\bar{Z}$ меньше, чем в случае независимых случайных величин

$$D(\bar{Z}) = D(\bar{X}) + D(\bar{Y}) - 2K_{\bar{X},\bar{Y}}.$$

Пусть $\zeta \sim N(a, \sigma)$.
Если верна гипотеза H_0, то статистика

$$t_{n-1} = \frac{\bar{Z}}{S_Z/\sqrt{n}}$$

подчиняется распределению Стьюдента T_{n-1} с $n-1$ степенями свободы.

Сформулируем критерий с вероятностью ошибки первого рода α:

- Если $|t_{n-1}| \leqslant t_{1-\frac{\alpha}{2}}$, где $t_{1-\frac{\alpha}{2}}$ — квантиль распределения Стьюдента с $n-1$ степенями свободы уровня $1-\alpha/2$, то принимается гипотеза H_0.
- Если $|t_{n-1}| > t_{1-\frac{\alpha}{2}}$, то принимается гипотеза H_1.

Для правосторонней альтернативной гипотезы $H_1 : a_1 > a_2$ критическая область для H_0 при использовании статистики t_{n-1} будет следующей: $S = (t_{1-\alpha}, +\infty)$.
Для левосторонней альтернативной гипотезы $H_1 : a_1 < a_2$ критическая область для H_0 при использовании статистики t_{n-1} будет следующей: $S = (-\infty, t_\alpha)$.
Границы $t_{1-\alpha}$ и t_α — квантили распределения Стьюдента с $n-1$ степенями свободы уровней $1-\alpha$ и α соответственно.

Однофакторный дисперсионный анализ

Однофакторный дисперсионный анализ является обобщением T-критерия для двух выборок из генеральной совокупности. Пусть число выборок $k \geqslant 2$:

$$\begin{array}{cccc} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & & \vdots \\ X_{n_1 1} & X_{n_2 2} & \dots & X_{n_k k} \\ N(a_1, \sigma^2) & N(a_2, \sigma^2) & \dots & N(a_k, \sigma^2) \end{array}$$

Пусть все выборки взаимно независимы между собой, при этом выборка $(X_{1i}, \dots, X_{n_i i})$ взята из генеральной совокупности с распределением $N(a_i, \sigma^2)$. Элементы каждой выборки тоже, конечно, взаимно независимы.

Выдвигается гипотеза $H_0 : a_1 = a_2 = \ldots = a_k$ при альтернативной гипотезе H_1, которая заключается в отрицании гипотезы H_0.
Рассмотрим следующие величины:

$$\bar{X}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}, \quad \bar{X}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{n_j}\sum_{j=1}^{k} X_{ij}, \quad N = \sum_{j=1}^{k} n_j,$$

где j — номер выборки.
Независимо от справедливости гипотезы H_0:

$$\sum_{i=1}^{n_j} \frac{(X_{ij} - \bar{X}_{\cdot j})}{\sigma^2} \sim \chi^2_{n_j-1} \sim G\left(\frac{1}{2}, \frac{n_j - 1}{2}\right).$$

Следовательно,

$$S_1 = \sum_{i=1}^{n_j}\sum_{j=1}^{k} \frac{(X_{ij} - \bar{X}_{\cdot j})^2}{\sigma^2} \sim \chi^2_{N-k}.$$

Рассмотрим величину

$$\frac{(\bar{X}_{\cdot j} - a_j)}{\sigma}\sqrt{n_j} \sim N(0,1).$$

Пусть гипотеза H_0 верна. Рассмотри статистику S_2 следующего вида:

$$S_2 = \frac{1}{\sigma^2}\sum_{j=1}^{k} n_j(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2.$$

Прибавим и вычтем величину a, получим:

$$\frac{1}{\sigma^2}\sum_{j=1}^{k} n_j(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2 = \frac{1}{\sigma^2}\sum_{j=1}^{k} n_j\left[(\bar{X}_{\cdot j} - a) - \frac{1}{N}\sum_{j=1}^{k} n_j(\bar{X}_{\cdot j} - a)\right]^2,$$

так как $\bar{X}_{\cdot\cdot} - a = \frac{1}{N}\sum\limits_{i=1}^{n_j}\sum\limits_{j=1}^{k}(X_{ij} - a) = \frac{1}{N}\sum\limits_{j=1}^{k} n_j(\bar{X}_{\cdot j} - a).$

$$S_2 = \frac{1}{\sigma^2}\sum_{j=1}^{k} n_j(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2 =$$

$$= \frac{1}{\sigma^2}\sum_{j=1}^{k} n_j \left[(\bar{X}_{\cdot j} - a) - \frac{1}{N}\sum_{j=1}^{k} n_j(\bar{X}_{\cdot j} - a)\right]^2 =$$

$$= \sum_{j=1}^{k} \frac{(\sqrt{n_j}(\bar{X}_{\cdot j} - a))^2}{\sigma^2} - \frac{\left[\sum_{j=1}^{k} \sqrt{\frac{n_j}{N}}\sqrt{n_j}(\bar{X}_{\cdot j} - a)\right]^2}{\sigma^2} =$$

$$= \sum_{j=1}^{k} \left[\frac{\sqrt{n_j}(\bar{X}_{\cdot j} - a)}{\sigma}\right]^2 - \left[\sum_{j=1}^{k} \sqrt{\frac{n_j}{N}}\frac{\sqrt{n_j}(\bar{X}_{\cdot j} - a)}{\sigma}\right]^2 \sim \chi^2_{k-1},$$

как следует из леммы 4 Л4.

Из леммы 3 Л4 следует, что статистики S_1 и S_2 взаимно независимы. По определению распределения Фишера $\mathcal{F}_{k-1,N-k}$ получаем, что при выполнении гипотезы H_0 статистика

$$F = \frac{S_2/(k-1)}{S_1/(N-k)} \sim \mathcal{F}_{k-1,N-k}$$

подчиняется распределению Фишера с $k-1$ и $N-k$ степенями свободы числителя и знаменателя соответственно.
Заметим, что статистика F не зависит от σ^2. Большие значения статистики F свидетельствуют против нулевой гипотезы, поэтому критическая область S для H_0 с вероятностью ошибки первого рода α имеет вид: $S = (u_{1-\alpha,k-1,N-k}; \infty)$, где $u_{1-\alpha,k-1,N-k}$ — квантиль уровня $1-\alpha$ распределения Фишера $\mathcal{F}_{k-1,N-k}$.

Метод линейных контрастов

Если нулевая гипотеза отклоняется, то требуется определить, какие именно группы имеют значимое различие средних.
Линейный контраст Lk определяется как линейная комбинация

$$Lk = \sum_{j=1}^{k} c_j a_j,$$

где c_j, $j = 1, \ldots, k$, — задаваемые константы, причем $\sum_{j=1}^{k} c_j = 0$.
Оценка линейного контраста имеет следующий вид:

$$\hat{Lk} = \sum_{j=1}^{k} c_j \bar{X}_j.$$

Оценка дисперсии $\hat{Lk}$ вычисляется по формуле:

$$S^2_{\hat{Lk}} = \sum_{j=1}^{k} \frac{c_j^2}{n_j} \hat{\sigma}^2 = \frac{S_1}{n-k} \sum_{j=1}^{k} \frac{c_j^2}{n_j}.$$

Доверительный интервал для Lk имеет вид

$$\left(\hat{Lk} - S_{\hat{Lk}}\sqrt{(k-1)u_{1-\alpha,k-1,n-k}},\right.$$
$$\left.\hat{Lk} + S_{\hat{Lk}}\sqrt{(k-1)u_{1-\alpha,k-1,n-k}}\right), \quad (2)$$

где $u_{1-\alpha,k-1,n-k}$ — квантиль распределния Фишера с $(k-1, n-k)$ степенями свободы уровня $1-\alpha$.

Лемма 2 (Метод Шеффе)

Для любой совокупности векторов $(c_1, \ldots, c_k)$: $\sum_{j=1}^{k} c_j = 0$, *вероятность одновременного выполнения неравенств*

$$\left|\sum_{j=1}^{k} c_j(a_j - \bar{X}_j)\right| < S_{\hat{Lk}}\sqrt{(k-1)u_{1-\alpha,k-1,n-k}}$$

не меньше $1-\alpha$.

Нулевая гипотеза для контраста $H_0 : Lk = 0$ принимается на уровне значимости α, если ноль содержится в доверительном интервале для Lk с доверительной вероятностью 1-α.

Рассмотрим нулевые гипотезы $H_0^{rs} : a_r = a_s,\ s \neq r$ против двусторонних альтернативных гипотез $H_1^{rs} : a_r \neq a_s,\ s \neq r$.
Гипотеза $H_0^{rs} : a_r = a_s$ равносильна гипотезе $H_0^{rs} : Lk_{rs} = 0$ с линейным контрастом вида

$$Lk_{rs} = a_r - a_s, \quad c_r = 1,\ c_s = -1,\ c_j = 0,\ j \neq r,\ j \neq s.$$

Гипотеза H_0^{rs} принимается, если ноль содержится в доверительном интервале (2) для контраста Lk_{rs}, в противном случае H_0^{rs} отклоняется.