1. Параметрические критерии однородности для двух выборок

1.1. Распределение Стьюдента

Пусть $\xi, \xi_1, \ldots, \xi_n$ независимые нормально распределенные величины $\mathcal{N}(0, \sigma)$. Случайная величина вида

$$\zeta = \frac{\xi}{\sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i^2}} = \frac{\xi}{\eta}, \tag{1}$$

имеет распределение *Стьюдента с n степенями свободы.*

Не умаляя общности, будем считать $\sigma = 1$. Тогда $\sum\limits_{i=1}^{n}\xi_i^2$ имеет распределение $\chi^2(n)$ с плотностью

$$\kappa_n(x) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}\,. \tag{2}$$

$$F_{a\xi}(x) = P\{a\xi < x\} = P\left\{\xi < \frac{x}{a}\right\} = F_\xi\left(\frac{x}{a}\right),$$
$$F_{\sqrt{\xi}}(x) = P\left\{\sqrt{\xi} < x\right\} = P\{\xi < x^2\} = F_\xi(x^2), \implies$$
$$f_{a\xi}(x) = \frac{1}{a} f_\xi\left(\frac{x}{a}\right), \qquad f_{\sqrt{\xi}}(x) = 2x f_\xi(x^2).$$

Для величины $\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i^2$ получаем плотность $n\kappa_n(nx)$, а для величины $\eta = \sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i^2}$ плотность $2nx\kappa_n(nx^2)$.
Совместное распределение ξ и η имеет вид:

$$\frac{2n\eta}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\left(n\eta^2\right)^{\frac{n}{2}-1} e^{-\frac{n\eta^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} = c_n \eta^{n-1} e^{-\frac{\xi^2+n\eta^2}{2}},$$
$$\text{где} \quad c_n = \frac{\sqrt{2}\left(\frac{n}{2}\right)^{\frac{n}{2}}}{\sqrt{\pi}\Gamma(\frac{n}{2})}\,.$$

$$P\{\zeta < x\} = P\left\{\frac{\xi}{\eta} < x\right\} = P\{\xi < x\eta\} =$$
$$= c_n \iint\limits_{\substack{\eta>0\\ \xi<x\eta}} \eta^{n-1} e^{-\frac{\xi^2+n\eta^2}{2}} d\xi d\eta = c_n \int\limits_{-\infty}^{x} du \int\limits_{0}^{\infty} v^{n-1} e^{-\frac{u^2v^2+nv^2}{2}} v dv,$$

используем замену $\xi = uv$, $\eta = v$ с якобианом преобразования

$$J = \begin{vmatrix} \frac{\partial\xi}{\partial u} & \frac{\partial\xi}{\partial v} \\ \frac{\partial\eta}{\partial u} & \frac{\partial\eta}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v\,.$$

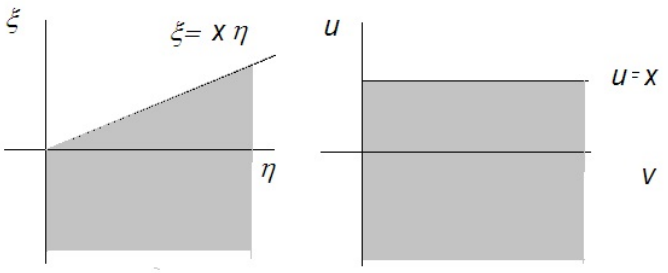

Рис. 1. Замена переменных $\xi = uv$, $\eta = v$ ($\eta > 0$,$\xi < x\eta$).

Далее заменим $\frac{(u^2+n)v^2}{2} = t$, $v = \frac{\sqrt{2t}}{\sqrt{u^2+n}}$, $vdv = \frac{dt}{u^2+n}$.

$$P\{\zeta < x\} = c_n \int\limits_{-\infty}^{x} du \int\limits_{0}^{\infty} \left(\frac{\sqrt{2t}}{\sqrt{u^2+n}}\right)^{n-1} e^{-t} \frac{dt}{u^2+n} =$$

$$= c_n 2^{\frac{n-1}{2}} \int\limits_{-\infty}^{x} \frac{du}{(u^2+n)^{\frac{n+1}{2}}} \Gamma\left(\frac{n+1}{2}\right) =$$

$$= \frac{\sqrt{2}\left(\frac{n}{2}\right)^{\frac{n}{2}}}{\sqrt{\pi}\Gamma(\frac{n}{2})} 2^{\frac{n-1}{2}} \int\limits_{-\infty}^{x} \frac{du}{(u^2+n)^{\frac{n+1}{2}}} \Gamma\left(\frac{n+1}{2}\right)$$

Отсюда плотность распределения имеет вид $f_\zeta(x,n) =$

$$= \frac{n^{\frac{n}{2}+\frac{1}{2}}}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(\frac{n}{2})} \frac{1}{(x^2+n)^{\frac{n+1}{2}}} = \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(\frac{n}{2})} \left(\frac{x^2}{n}+1\right)^{-\frac{n+1}{2}}.$$

1.2. Асимптотическое свойство распределения Стьюдента

Плотность распределения Стьюдента

$$f_\zeta(x,n) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(\frac{n}{2})} \left(\frac{x^2}{n}+1\right)^{-\frac{n+1}{2}}. \tag{3}$$

можно записать в виде

$$f_\zeta(x,n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\frac{n}{2}}\Gamma(\frac{n}{2})} \frac{1}{\sqrt{2\pi}} \left(\frac{x^2}{n}+1\right)^{-\frac{n+1}{2}}.$$

Воспользуемся формулой Стирлинга

$$\ln\Gamma(\lambda) = \left(\lambda - \frac{1}{2}\right)\ln\lambda - \lambda + \frac{1}{2}\ln(2\pi) + \frac{1}{12\lambda} + O\left(\frac{1}{\lambda^3}\right).$$

При $p = \frac{n}{2} \longrightarrow \infty$ имеют место выражения

$$\ln\Gamma\left(p+\frac{1}{2}\right) - \frac{1}{2}\ln p - \ln\Gamma(p) \approx$$
$$\approx \left(p+\frac{1}{2}-\frac{1}{2}\right)\ln\left(p+\frac{1}{2}\right) - p - \frac{1}{2} -$$
$$-\frac{1}{2}\ln p - \left(p-\frac{1}{2}\right)\ln p + p =$$
$$= p\ln\left(p+\frac{1}{2}\right) - \frac{1}{2} - p\ln p = p\ln\left(1+\frac{1}{2p}\right) - \frac{1}{2} \approx 0\,.$$

При любом фиксированном x имеем

$$-\frac{n+1}{2}\ln\left(1+\frac{x^2}{n}\right) \underset{n\longrightarrow\infty}{\longrightarrow} -\frac{x^2}{2}\,,$$

откуда

$$f_\zeta(x,n) \underset{n\longrightarrow\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,.$$

1.3. Распределение Фишера

Распределенная по Фишеру случайная величина имеет вид

$$\zeta = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i^2}{\frac{1}{m}\sum\limits_{j=1}^{m}\eta_j^2} = \frac{m}{n}\zeta_0\,, \tag{4}$$

где ξ_i, η_j независимы, одинаково распределены $\mathcal{N}(0,\sigma)$. Не умаляя общности, будем считать $\sigma^2 = 1$. Тогда $\xi = \sum\limits_{i=1}^{n}\xi_i^2$ и $\eta = \sum\limits_{j=1}^{m}\eta_j^2$ имеют распределение (2) хи-квадрат.

$$P\{\zeta_0 < x\} = P\left\{\frac{\xi}{\eta} < x\right\} = P\{\xi < x\eta\}\,.$$

Совместное распределение величин ξ и η имеет вид

$$\kappa_n(\xi)\kappa_m(\eta) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})}\xi^{\frac{n}{2}-1}e^{-\frac{\xi}{2}}\frac{2^{-\frac{m}{2}}}{\Gamma(\frac{m}{2})}\eta^{\frac{m}{2}-1}e^{-\frac{\eta}{2}} =$$
$$= a_{nm}\xi^{\frac{n}{2}-1}\eta^{\frac{m}{2}-1}e^{-\frac{\xi+\eta}{2}},$$
$$\text{где}\quad a_{nm} = \frac{2^{-\frac{m+n}{2}}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\,.$$

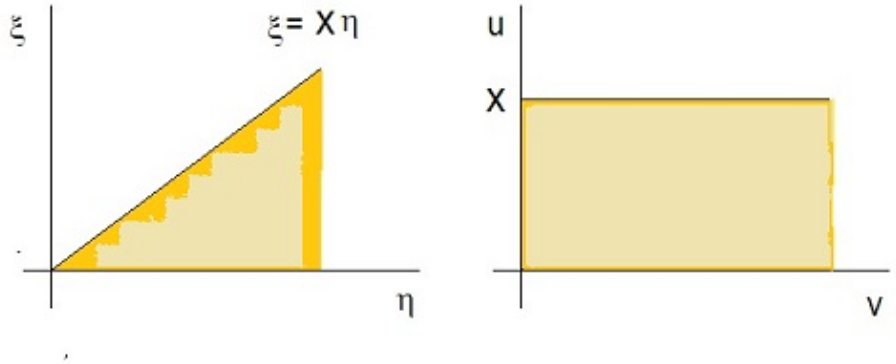

Рис. 2. Замена переменных $\xi = uv$, $\eta = v$.

$$P\{\zeta_0 < x\} = a_{nm} \int\limits_{\substack{\eta>0 \\ 0<\xi<x\eta}} \int \xi^{\frac{n}{2}-1} \eta^{\frac{m}{2}-1} e^{-\frac{\xi+\eta}{2}} d\xi d\eta \,. \tag{5}$$

Используем замену $\xi = uv$, $\eta = v$ с якобианом преобразования

$$J = \begin{vmatrix} \frac{\partial \xi}{\partial u} & \frac{\partial \xi}{\partial v} \\ \frac{\partial \eta}{\partial u} & \frac{\partial \eta}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v \,.$$

Область интегрирования из $\eta > 0$, $0 < \xi < x\eta$ перейдет в область $v > 0, 0 < u < x$ (рис.2). Следовательно,

$$P\{\zeta_0 < x\} = a_{nm} \int\limits_0^x u^{\frac{n}{2}-1} du \int\limits_0^\infty v^{\frac{n+m}{2}-2} e^{-\frac{uv+v}{2}} v dv =$$

$$= a_{nm} \int\limits_0^x u^{\frac{n}{2}-1} du \int\limits_0^\infty v^{\frac{n+m}{2}-1} e^{-\frac{(u+1)v}{2}} dv \,.$$

Заменим переменные $t = \frac{u+1}{2} v,\ v = \frac{2}{u+1}$, $dv = \frac{2}{u+1} dt$.

$$a_{nm} \int\limits_0^x u^{\frac{n}{2}-1} du \int\limits_0^\infty t^{\frac{n+m}{2}-1} \left(\frac{2}{u+1}\right)^{\frac{n+m}{2}-1} \frac{2}{u+1} e^{-t} dt =$$

$$= \frac{2^{-\frac{m+n}{2}}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \int\limits_0^x u^{\frac{n}{2}-1} \left(\frac{2}{u+1}\right)^{\frac{n+m}{2}} du \Gamma\left(\frac{n+m}{2}\right) =$$

$$= \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \int\limits_0^x \frac{u^{\frac{n}{2}-1}}{(u+1)^{\frac{n+m}{2}}} du \,.$$

Таким образом, используя $C_{nm} = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}$, получаем

$$f_0(x) = C_{nm} \frac{x^{\frac{n}{2}-1}}{(x+1)^{\frac{n+m}{2}}},$$

$$f(x) = \frac{n}{m} f_0\left(\frac{n}{m} x\right) = \frac{n}{m} C_{nm} \frac{\left(\frac{n}{m}x\right)^{\frac{n}{2}-1}}{\left(\frac{n}{m}x+1\right)^{\frac{n+m}{2}}} \,.$$

1.4. Совместное распределение $\bar{x}$ и m_2

Лемма 1. *(Фишера) Пусть $x_1, \ldots, x_n$ — независимые и нормально распределенные $\mathcal{N}(0, \sigma)$ компоненты вектора X. В результате ортогонального преобразования $Y = CX$ имеем также независимые и нормально распределенные $\mathcal{N}(0, \sigma)$ компоненты $y_1, \ldots, y_n$. Тогда квадратичная форма*

$$Q(X) = \sum_{i=1}^{n} x_i^2 - y_1^2 - \ldots - y_p^2$$

независима от $y_1, \ldots, y_p$ и имеет плотность $\frac{1}{\sigma^2}\kappa_{n-p}\left(\frac{x}{\sigma^2}\right)$, где $\kappa_n(x)$ — плотность распределения хи-квадрат с n степенями свободы (2).

Теорема 1. *Пусть $x_1, \ldots, x_n \sim \mathbf{N}(\mu, \sigma)$, $\bar{x}$ и m_2 – выборочные среднее и второй центральный момент. Тогда*

1. *$\bar{x}$ и m_2 независимы;*
2. *$\bar{x} \sim \mathbf{N}(\mu; \frac{\sigma}{\sqrt{n}})$;*
3. *статистика $\frac{nm_2}{\sigma^2} \sim \boldsymbol{\chi^2(n-1)}$ имеет распределение хи-квадрат с $(n-1)$ степенью свободы.*

Без ограничения общности будем считать $\mu = 0$.

$$nm_2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Выражение $n\bar{x}^2 = \left(\frac{x_1}{\sqrt{n}} + \ldots + \frac{x_n}{\sqrt{n}}\right)^2$ есть квадрат линейной формы $c_1x_1 + \ldots + c_nx_n$, для которой $c_1^2 + \ldots + c_n^2 = 1$. Поэтому применим лемму 1, положив $p = 1$ и $y_1 = \sqrt{n}\bar{x}$.

1.5. Статистики Стьюдента и Фишера

Пусть $x_1, \ldots, x_n$ — выборка из нормальной совокупности со средним μ и дисперсией σ^2. Тогда отношение

$$\tau = \frac{\bar{x} - \mu}{\sqrt{m_2}} \sqrt{n-1} \tag{6}$$

имеет распределение Стьюдента с $n-1$ степенью свободы.

Действительно, в (1) полагая $\xi_1 = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ и $\sum\limits_{i=1}^{n} \xi_i^2 = \frac{nm_2}{\sigma^2}$, получим

$$\tau = \frac{\xi_1}{\sqrt{\frac{1}{n-1}\sum\limits_{k=2}^{n} \xi_k^2}} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{1}{n-1}\frac{nm_2}{\sigma^2}}} = \frac{\bar{x}-\mu}{\sqrt{m_2}}\sqrt{n-1}.$$

Так как $m_2 = \frac{n-1}{n}S^2$, $\sqrt{m_2} = \frac{\sqrt{n-1}}{\sqrt{n}}S$,

$$\tau = \frac{\bar{x}-\mu}{S}\sqrt{n} \ \sim \ \mathbf{T(n-1)}. \tag{7}$$

Статистика Фишера для проверки гипотезы о равенстве дисперсий строится с учетом структуры распределения Фишера (4) как функция от двух независимых нормально распределенных выборок: $x_1, \ldots, x_n \sim \mathcal{N}(\mu_1; \sigma)$ и $y_1, \ldots, y_m \sim \mathcal{N}(\mu_2; \sigma)$

$$F = \frac{S_x^2}{S_y^2} \ \sim \ \mathbf{F(n;m)}, \tag{8}$$

где S_x^2 и S_y^2 – несмещенные оценки их дисперсий соответствующих выборок. Действительно, так как $(n-1)S^2/\sigma^2 \ \sim \ \boldsymbol{\chi^2(n-1)}$, то в соответствии с (4)

$$F = \frac{\frac{1}{n-1}(n-1)S_x^2/\sigma^2}{\frac{1}{m-1}(m-1)S_y^2/\sigma^2} = \frac{S_x^2}{S_y^2} \ \sim \ \mathbf{F(n;m)}.$$

1.6. Критерий Стьюдента для зависимых выборок

Часто встречаются данные, в которых за одними и теми же индивидами наблюдают в разные моменты времени, например, у $n = 67$ больных с воспалением легких логарифм показателя скорости оседания эритроцитов (СОЭ) в начале заболевания в среднем равен 3.82 со стандартным отклонением 0.13, а при выписке из стационара 3.28 со стандартным отклонением 0.19. Требуется выяснить, можно ли считать уменьшение СОЭ значимым, или это уменьшение можно объснить случайностью.

Итак, имеется парная выборка (x_i, y_i), $i = 1, \ldots, n$, зависимых наблюдений объема n над случайными величинами ξ_1 и ξ_2 с нормальным распределением $\mathcal{N}(\mu_1, \sigma_1)$ и $\mathcal{N}(\mu_2, \sigma_2)$, относительно которых нет предположения о независимости — скорее всего, чем меньше был уровень СОЭ в начале заболевания, тем меньше он при выздоровлении. Рассмотрим разность $\eta = \xi_1 - \xi_2$ и соответствующую выборку $z_i = x_i - y_i$ с выборочным средним $\bar{z}$ и несмещенной оценкой дисперсии S_z^2. Если справедлива нулевая гипотеза $H_0 : \mu_1 = \mu_2$, то для случайной выборки $z_i = x_i - y_i$ должна быт справедлива гипотеза о равенстве нулю ее математического ожидания.

$$t = \frac{\bar{z}}{S_z}\sqrt{n} \ \sim \ \mathbf{T}(n-1) \text{ при справедливости } H_0 : \mu_1 = \mu_2 \,.$$

. Например, $\bar{z} = 0.54$, $S = 0.19$, $t = \frac{\bar{z}-0}{S}\sqrt{n} = \frac{0.54}{0.19}\sqrt{66} = 23.45 = t_*$, двусторонний доверительный уровень вероятности равен

$$p = P\{|t| > |t_*|\} = \text{СТЮДРАСП}(23.45; 66; 2) = 10^{-33} < 0.05,$$

поэтому гипотеза $H_0 : \mu_1 = \mu_2$ отвергается,[1] и уменьшение уровня СОЭ в процессе выздоровления нельзя объяснить случайностью.

[1] Значимость критерия в R считается как $p = 2 * (1 - pt(t_*, n-1))$ или используется общая функция $t.test$ с опцией $paired = TRUE$.

1.7. Распределение разности выборочных средних

Пусть имеется две независимые выборки $x_1, \ldots, x_{n_1}$ и $y_1, \ldots, y_{n_2}$, о которых известно, что они имеют нормальное распределение со средними μ_1 и μ_2 и одинаковой дисперсией σ^2. Выборочные средние обозначим через $\bar{x}$ и $\bar{y}$, а несмещенные оценки дисперсии через S_1^2 и S_2^2. Тогда статистика

$$T = \frac{(\bar{x} - \bar{y} - (\mu_1 - \mu_2))\sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}} \sim \mathbf{T}(n_1 + n_2 - 2) \tag{9}$$

имеет распределение Стьюдента с числом степеней свободы $n_1 + n_2 - 2$.

Во-первых, из независимости выборок выборочные средние $\bar{x}$ и $\bar{y}$ также независимы, а по теореме Фишера (т.1) нормально распределены соответственно с параметрами $\mathcal{N}(\mu_1, \sigma/\sqrt{n_1})$ и $\mathcal{N}(\mu_2, \sigma/\sqrt{n_2})$. Следовательно, их разность имеет нормальное распределение с параметрами

$$\bar{x} - \bar{y} \;\; \sim \;\; \mathcal{N}\left(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right),$$

откуда получаем

$$\xi = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0; 1). \tag{10}$$

С другой стороны, также по теореме 1

$$(n_1 - 1)S_1^2/\sigma^2 \sim \chi^{\mathbf{2}}(\mathbf{n_1 - 1}),$$
$$(n_2 - 1)S_2^2/\sigma^2 \sim \chi^{\mathbf{2}}(\mathbf{n_2 - 1}).$$

Из аддитивности распределения $\chi^{\mathbf{2}}$ их сумма вида

$$\eta = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \;\; \sim \;\; \chi^{\mathbf{2}}(\mathbf{n_1 + n_2 - 2}) \tag{11}$$

имеет распределение хи-квадрат с числом степеней свободы, равным сумме степеней свободы слагаемых. Рассмотрим случайную величину вида

$$T = \frac{\xi}{\sqrt{\frac{1}{n_1 + n_2 - 2}\eta}},$$

которая, согласно структуре (1), имеет распределение Стьюдента с числом степеней свободы, равным $n_1 + n_2 - 2$. Подставив соответствующие ξ из (10) и η из (11), получим вид статистики из (9).

При известных дисперсий σ_1^2 и σ_2^2 (возможно неодинаковых), учитывая нормальность средних

$$\bar{x} \sim \mathcal{N}(\mu_1, \sigma_1/\sqrt{n_1}), \quad \bar{y} \sim \mathcal{N}(\mu_2, \sigma_2/\sqrt{n_2}),$$

вычисляем дисперсию разности

$$\mathbf{D}(\bar{x} - \bar{y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

откуда получаем

$$Z = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \ \sim \ \mathcal{N}(0; 1). \tag{12}$$

В случае неизвестных неодинаковых дисперсий $\sigma_1^2 \neq \sigma_2^2$ статистика T имеет приближенно распределение Стьюдента с нецелым числом степеней свободы.

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \ \sim \ \mathbf{T}(k), \quad \text{где } k = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}. \tag{13}$$