Критерий согласия Колмогорова.

Теорема. Пусть случайная величина Y равномерно распределена на $[0,1]$, пусть F^{-1} обратная функция к функции распределения $F(x)$. Тогда случайная величина $X = F^{-1}(Y)$ распределена по закону $F(x)$.

Доказательство. Для равномерно распределенной на $[0,1]$ случайной величины верно, что $P(Y < x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases}$. Следовательно, поскольку $0 \le F(x) \le 1$, то для всех значений $x \in (-\infty; +\infty)$

$$P(X < x) = P(F^{-1}(Y) < x) = P(Y < F(x)) = F(x).$$

Пусть $X_1, X_2, \ldots . X_n$ выборка из закона распределения, задаваемого ф.р. $F(x)$

Рассмотрим выборочную функцию распределения:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{k=1}^{n} I(x - X_k),\ x \in \mathbb{R},\ I(x) = \begin{cases} 1 & x > 0 \\ 0 & x \le 0 \end{cases},$$

Теорема 1. Если функция распределения непрерывна, то закон распределения статистики D_n не зависит от вида функции $F(x)$.

Доказательство. Рассмотрим сначала случай строго монотонной функции $F(x)$. Наряду с выборкой $X_1, X_2, \ldots . X_n$, представляющей реализацию последовательности независимых случайных величин, распределенных по закону $F(x)$ рассмотрим $Y_1, Y_2, \ldots . Y_n$, где $Y_k = F(X_k)$.

Тогда, положив $y = F(x), x = F^{-1}(y)$, можем записать

$$\sup_{x \in \mathbb{R}} \left| \widehat{F_n}(x) - F(x) \right| = \sup_{y \in [0,1]} \left| \widehat{F_n}(F^{-1}(y)) - y \right| = \sup_{y \in [0,1]} \left| \widehat{\widehat{F}_n}(y) - y \right|$$

$$\widehat{F_n}(F^{-1}(y)) = \frac{1}{n}\sum_{k=1}^{n} I(X_k < F^{-1}(y)) = \frac{1}{n}\sum_{k=1}^{n} I(Y_k < y) = \widehat{\widehat{F}}_n(y).$$

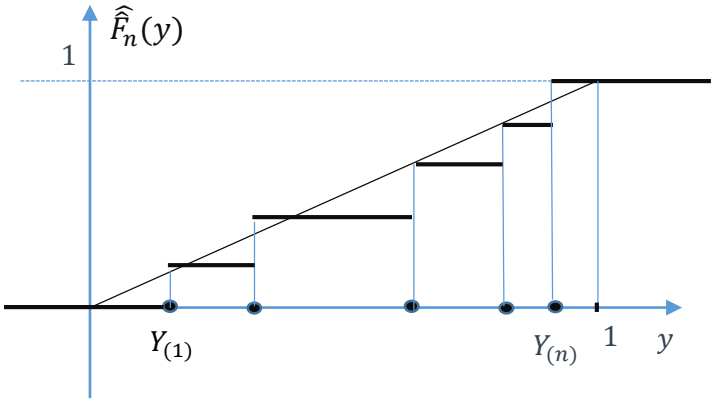
Здесь $\widehat{\widehat{F}}_n(y)$- эмпирическая функция распределения, выборки $Y_1, Y_2, \ldots . Y_n$, распределенной равномерно на $[0,1]$. Таким образом закон распределения статистики D_n не зависит от вида функции $F(x)$.

Если $F(x)$ не является строго монотонной, то в доказательстве теоремы $\sup_{x \in \mathbb{R}} \left| \widehat{F_n}(x) - F(x) \right|$ надо заменить на $\sup_{x \in M} \left| \widehat{F_n}(x) - F(x) \right|$, где M - множество строгой монотонности $F(x)$.

Замечание. Вычисление статистики Колмогорова.

Пусть $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$-вариационный ряд, построенный по пересчитанной выборке. Тогда

$$D_n = \sup_{y\in[0,1]} \left|\hat{\hat{F}}_n(y) - y\right| = \max_{1\le k\le n} \max\left\{\left|Y_{(k)} - \frac{k}{n}\right|, \left|Y_{(k)} - \frac{k-1}{n}\right|\right\} = \max_{1\le k\le n}\left\{\left|Y_{(k)} - \frac{2k-1}{2n}\right| + \frac{1}{2n}\right\}$$



В самом деле, для произвольного отрезка $[a,b]$ и любой точки $x \in [a,b]$ имеет место очевидное соотношение: $\max\{|x-a|, |x-b|\} = \left|x - \frac{a+b}{2}\right| + \frac{b-a}{2}$.

Определение. Статистикой Колмогорова называется $D_n = \sup_{x\in\mathbb{R}} \left|F_n(x) - F(x)\right|$

Критерий Колмогорова для $n \le 20$

Смирнов рассчитал и табулировал критические точки закона распределения D_n для различных уровней значимости , то есть такие $k_{1-\alpha}(n)$, что $P\left(D_n \ge k_{1-\alpha}(n)\right) = \alpha$

Если $D_n \ge k_{1-\alpha}(n)$, то основная гипотеза отклоняется

Если $D_n < k_{1-\alpha}(n)$, то основная гипотеза принимается

Теорема 2 (Колмогоров).

Определим функцию Колмогорова $K(x) = \sum_{j=-\infty}^{\infty} (-1)^n e^{-2k^2x^2}, x > 0$. Тогда для любой непрерывной $F(x)$. статистика $\sqrt{n}D_n$ при $n \to \infty$ по распределению сходится к функции Колмогорова:

$$\lim_{n\to\infty} P\left(\sqrt{n}D_n < x\right) = K(x),\ x > 0.$$

Критерий Колмогорова для $n > 20$

Если $D_n \geq \frac{\lambda_{1-\alpha}}{\sqrt{n}}$, то основная гипотеза отклоняется

Если $D_n < \frac{\lambda_{1-\alpha}}{\sqrt{n}}$, то основная гипотеза принимается

где $\lambda_{1-\alpha}$ -квантиль уровня $1-\alpha$ функции $K(x)$, то есть $K(\lambda_{1-\alpha}) = 1-\alpha$.

При этом ошибка первого рода критерия равна

$$P\left(D_n \geq \frac{\lambda_{1-\alpha}}{\sqrt{n}} | H_0\right) = P\left(\sqrt{n}D_n \geq \lambda_{1-\alpha}\right) = 1 - K(\lambda_{1-\alpha}) = \alpha \ .$$

Квантили функции Колмогорова

$1-\alpha$	0.9	0.95	0.98	0.99
$\lambda_{1-\alpha}$	1.224	1.358	1.515	1.628

$$D_n = \sup_{y\in[0,1]} \left|\hat{F}_n(y) - y\right| = \max_{1\leq k\leq n} \max\left\{\left|Y_{(k)} - \frac{k}{n}\right|, \left|Y_{(k)} - \frac{k-1}{n}\right|\right\} = \max_{1\leq k\leq n}\left\{\left|Y_{(k)} - \frac{2k-1}{2n}\right| + \frac{1}{2n}\right\}$$

Пример 2

Пассажир, приходящий в случайные моменты времени на автобусную остановку, в течение пяти поездок фиксировал своё время ожидания автобуса: 5,1; 3,7; 1,2; 9,2; 4,8 мин. Проверить гипотезу о том, что время ожидания автобуса равномерно распределено на отрезке [0; 10] на уровне значимости 0,05.

Решение. Здесь $F(x) = \frac{x}{10}, 0 \leq x \leq 10,$ - функция распределения проверяемого закона. Упорядочим и пересчитаем выборку: $Y_k = F(X_k) = \frac{X_k}{10}$.

$X_{(k)}$	$Y_{(k)}$	$\frac{2k-1}{2n}$	$\left\|Y_{(k)}-\frac{2k-1}{2n}\right\| + \frac{1}{2n}$
1.2	0.12	0.1	0.12
3.7	0.37	0.3	0.17
4.8	0.48	0.5	0.12
5.1	0.51	0.7	0.29
9.2	0.92	0.9	0.02

Таким образом, $D_5 = 0.29$, в то время, как критическая точка $k_{0.95}(5) = 0.56$

Таким образом, гипотеза о равномерном законе распределения времени ожидания принимается.

Пример 3.

Банк "Стабильный" при построении скоринговой модели предполагает, что месячный доход (в тыс. руб.) его потенциальных клиентов в определённом регионе подчиняется нормальному закону распределения с математическим ожиданием $a = 100$ (тыс. руб.) и стандартным отклонением $\sigma = 15$ (тыс. руб.) Для первичной проверки этого предположения аналитик случайным образом отобрал анкеты 5 клиентов, подавших заявки на ипотеку.

Выборка месячных доходов (в тыс. руб.):

$$X = \{85, 115, 95, 105, 120\}.$$

Нулевая гипотеза H_0: Выборка происходит из нормального распределения $N(\mu = 100, \sigma = 15)$, т.е. $X \sim N(100,15)$.

Альтернативная гипотеза H_1: Распределение не является указанным нормальным.

Упорядочим выборку по возрастанию:

$x_{(1)} = 85, \quad x_{(2)} = 95, \quad x_{(3)} = 105, \quad x_{(4)} = 115, \quad x_{(5)} = 120.$

При гипотезе H_0 доход имеет нормальное распределение $N(100,15)$.

$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-100}{15}\right)$, где $\Phi(z)$ — функция стандартного нормального распределения $N(0,1)$.

Найдём статистику Колмогорова используя $Y_{(k)} = F(X_{(k)})$

$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-100}{15}\right),$

$X_{(k)}$	$Y_{(k)}$	$\left\|Y_{(k)} - \frac{2k-1}{2n}\right\| + \frac{1}{2n}$
85	0.159	0.059+1/10
95	0.369	0.069+1/10
105	0.631	0.131+1/10
115	0.841	0.141+1/10=0.241
120	0.909	0.009+1/10

При проверке гипотезы о нормальном распределении с полностью заданными параметрами (a и σ. известны) критические значения критерия Колмогорова остаются стандартными.

Для $n = 5$:

При $\alpha = 0.10$: $D_{крит} \approx 0.510$

При $\alpha = 0.05$: $D_{крит} \approx 0.563$

Поскольку расчётное значение $D_5 = 0.241$ меньше критического значения для $\alpha = 0.1$, мы не отвергаем нулевую гипотезу H_0. Статистически значимых отклонений выборки от нормального распределения $N(100,15)$ не обнаружено.

Статистический тест не обнаружил противоречий между наблюдаемыми данными (доходами 5 клиентов) и гипотезой о нормальном распределении доходов в популяции со средним 100 тыс. руб. и стандартным отклонением 15 тыс. руб.

1. Исходное предположение банка не опровергнуто на основе данной небольшой выборки. Это означает, что модель, использующая нормальное распределение для описания доходов, может быть адекватной.

2. Банк может продолжать использовать параметры $a = 100$ и $\sigma = 15$ в своих предварительных расчётах рисков и лимитов для данной группы клиентов.

3. Объём выборки ($n = 5$) крайне мал для уверенного вывода о форме распределения генеральной совокупности. Неотвержение гипотезы в данном случае скорее свидетельствует о низкой мощности критерия при малом n, чем о доказанной нормальности. Для серьёзной валидации модели необходим сбор данных с гораздо большим объёмом выборки ($n > 50$).

Критерий Смирнова

Пусть имеется две выборки

$X_1, X_2, \ldots . X_n$ выборка из закона распределения, задаваемого ф.р. $F(x)$

$Y_1, Y_2, \ldots . Y_m$ выборка из закона распределения, задаваемого ф.р. $G(x)$

Проверяется гипотеза об однородности выборок:

H_0: законы распределения выборок совпадают $F(x) = G(x)$

Теорема 3 (Смирнов)

Пусть $F_n(x)$ и $G_m(x)$ – эмпирические функции распределения двух рассматриваемых выборок, пусть $D_{n,m} = \sup\limits_{-\infty<x<\infty} \left|F_n(x) - G_m(x)\right|$ - статистика Смирнова. Тогда в предположении непрерывности общего закона распределения двух выборок

$F(x) = G(x)$ статистика $\sqrt{\dfrac{nm}{n+m}} D_{n,m}$ при $n \to \infty, m \to \infty$ по распределению сходится к функции Колмогорова:

$$\lim_{\substack{n\to\infty \\ m\to\infty}} P\left(\sqrt{\frac{nm}{n+m}} D_{n,m} < x\right) = K(x),\ x > 0.$$

Пример 4

По приведенным в таблице данным проверить по критерию Смирнова что законы распределения выборок совпадают. Принять $\alpha = 0.05$

Первая выборка	Вторая выборка
-1.723	0.985
-1.517	0.862
-0.442	0.916
-1.245	0.673
-0.91	-1.044
0.262	0.069
-1.706	-0.756
-0.998	0.697
0.108	-0.182
-0.107	-0.644

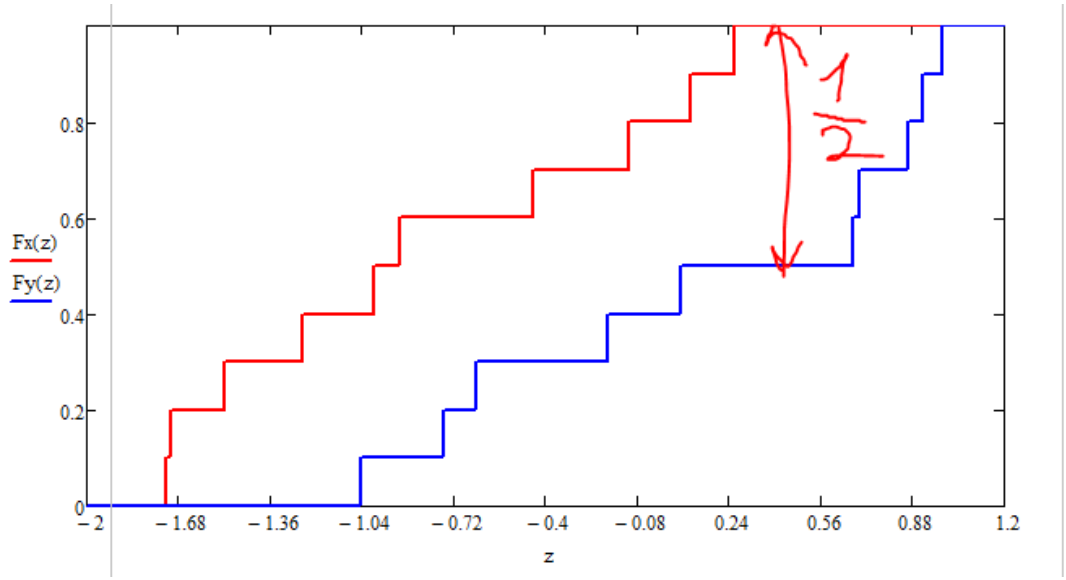
Составим функцию распределения для заданных выборок

$$x := \begin{pmatrix} -1.723 \\ -1.517 \\ -0.442 \\ -1.245 \\ -0.91 \\ 0.262 \\ -1.706 \\ -0.998 \\ 0.108 \\ -0.107 \end{pmatrix} \qquad y := \begin{pmatrix} 0.985 \\ 0.862 \\ 0.916 \\ 0.673 \\ -1.044 \\ 0.069 \\ -0.756 \\ 0.697 \\ -0.182 \\ -0.644 \end{pmatrix}$$

$$\text{ind}(k) := \begin{cases} 0 & \text{if } k < 0 \\ 1 & \text{if } k \geq 0 \end{cases}$$

$$Fx(z) := \frac{1}{10}\left(\sum_{i=0}^{9} \text{ind}\left(z - x_{i,0}\right)\right)$$

$$Fy(z) := \frac{1}{10}\cdot\sum_{i=0}^{9} \text{ind}\left(z - y_{i,0}\right)$$



В данном случае по графику видно в какой точке будет находится максимум разности функций распределения. Найдём статистику Колмогорова следующим образом, каждый скачок функции распределения равен $\frac{1}{n}$, в данном примере $n = 10$, эти скачки происходят в точках равных элементам выборки, отсортируем обе выборки и посмотрим насколько одна «убегает» от другой

Первая выборка	Вторая выборка
-1.723	
-1.706	
-1.517	
-1.245	
	-1.044
-0.998	
-0.91	
	-0.756
	-0.644
-0.442	
	-0.182
-0.107	
	0.069
0.108	
0.262	
	0.673
	0.697
	0.862
	0.916
	0.985

Например в точке -1.246 функция распределения первой выборки равна $\frac{4}{10}$ (4 элемента выборки из 10 меньше, чем -1.246), а функция распределения второй выборки в этой точке равна 0, найдём участок на котором одна из функций дальше всего убежала, это происходит после значения 0.262, в этой точке функция распределения первой выборки равна 1, а функция

распределения второй выборки равна $\frac{5}{10}$, таким образом статистика

Колмогорова равна $\frac{1}{2}$

поскольку $D_{10,10} = 0.5 > 0.41$ гипотеза отклоняется

Пример 5

Исследовательский вопрос: Одинаково ли распределены ежемесячные заработные платы работников в двух смежных отраслях — IT-секторе и цифровом маркетинге?

Для проверки собраны две независимые случайные выборки размером $n_1 = n_2 = 5$ (в тыс. руб.):

Выборка A (IT-специалисты):

$X = \{120, 150, 130, 140, 160\}$.

Выборка B (Специалисты по цифровому маркетингу):

$Y = \{115, 125, 135, 145, 155\}$.

Нулевая гипотеза H_0: Обе выборки происходят из одного и того же распределения ($F_X(t) = F_Y(t)$ для всех t).

Альтернативная гипотеза H_1: Распределения различаются ($F_X(t) \neq F_Y(t)$).

Упорядочим каждую выборку по возрастанию:

$X_{(1)} = 120, \quad X_{(2)} = 130, \quad X_{(3)} = 140, \quad X_{(4)} = 150, \quad X_{(5)} = 160.$
$Y_{(1)} = 115, \quad Y_{(2)} = 125, \quad Y_{(3)} = 135, \quad Y_{(4)} = 145, \quad Y_{(5)} = 155.$

Максимум модуля разности будет достигаться либо в одной из точек выборки X, либо в одной из точек выборки Y. Можно вычислить разности в каждой из 10 точек.

1. $t = 115$: $F_{n_1}(115) = 0$, $F_{n_2}(115) = 0.2$ $\Rightarrow$ разность = $|0 - 0.2| = 0.2$
2. $t = 120$: $F_{n_1}(120) = 0.2$, $F_{n_2}(120) = 0.2$ $\Rightarrow$ разность = $|0.2 - 0.2| = 0$
3. $t = 125$: $F_{n_1}(125) = 0.2$, $F_{n_2}(125) = 0.4$ $\Rightarrow$ разность = $|0.2 - 0.4| = 0.2$
4. $t = 130$: $F_{n_1}(130) = 0.4$, $F_{n_2}(130) = 0.4$ $\Rightarrow$ разность = $|0.4 - 0.4| = 0$
5. $t = 135$: $F_{n_1}(135) = 0.4$, $F_{n_2}(135) = 0.6$ $\Rightarrow$ разность = $|0.4 - 0.6| = 0.2$

6. $t = 140$: $F_{n_1}(140) = 0.6$, $F_{n_2}(140) = 0.6$ $\Rightarrow$ разность = $|0.6 - 0.6| = 0$

7. $t = 145$: $F_{n_1}(145) = 0.6$, $F_{n_2}(145) = 0.8$ $\Rightarrow$ разность = $|0.6 - 0.8| = 0.2$

8. $t = 150$: $F_{n_1}(150) = 0.8$, $F_{n_2}(150) = 0.8$ $\Rightarrow$ разность = $|0.8 - 0.8| = 0$

9. $t = 155$: $F_{n_1}(155) = 0.8$, $F_{n_2}(155) = 1.0$ $\Rightarrow$ разность = $|0.8 - 1.0| = 0.2$

10. $t = 160$: $F_{n_1}(160) = 1.0$, $F_{n_2}(160) = 1.0$ $\Rightarrow$ разность = $|1.0 - 1.0| = 0$

Наибольшая наблюдаемая разность: 0.2 в нескольких точках.

$D_{5,5} = 0.2$

Для двухвыборочного критерия Колмогорова-Смирнова при $n_1 = n_2 = 5$ и уровне значимости $\alpha = 0.10$ критическое значение можно найти по таблице.

Для $n = m = 5$ и $\alpha = 0.10$ Часто используют статистику $K = D_{n,m} \cdot \sqrt{\frac{nm}{n+m}}$, которая имеет известное распределение Колмогорова.

Вычислим:

$$K = 0.2 \cdot \sqrt{\frac{5 \cdot 5}{5+5}} = 0.2 \cdot \sqrt{\frac{25}{10}} = 0.2 \cdot \sqrt{2.5} \approx 0.2 \cdot 1.581 = 0.3162.$$

По таблице распределения Колмогорова для асимптотического случая:

Для $\alpha = 0.10$ критическое значение $K_{\text{крит}} \approx 1.22$,

для $\alpha = 0.05$ — около 1.36,

для $\alpha = 0.01$ — около 1.63.

Наше значение $K = 0.3162$ значительно меньше всех критических значений.

Так как $K = 0.3162 < K_{\text{крит}}$ даже для $\alpha = 0.10$, нет оснований отвергнуть нулевую гипотезу H_0.

На основании имеющихся данных статистически значимых различий в распределении заработных плат между IT-специалистами и специалистами по цифровому маркетингу не обнаружено. Обе выборки согласуются с гипотезой о том, что они взяты из одной генеральной совокупности.

Критические точки для статистики Колмогорова D_n

Объем выборки n	Уровень значимости α			
	0,10	0,05	0,02	0,01
1	0,95	0,98	0,99	0,995
2	0,78	0,84	0,90	0,93
3	0,64	0,71	0,78	0,83
4	0,57	0,62	0,69	0,73
5	0,51	0,56	0,62	0,67
6	0,47	0,52	0,58	0,62
7	0,44	0,48	0,54	0,58
8	0,41	0,45	0,51	0,54
9	0,39	0,43	0,48	0,51
10	0,37	0,41	0,46	0,49
11	0,35	0,39	0,44	0,47
12	0,34	0,38	0,42	0,45
13	0,33	0,36	0,40	0,43
14	0,31	0,35	0,39	0,42
15	0,30	0,34	0,38	0,40
16	0,29	0,33	0,37	0,39
17	0,29	0,32	0,36	0,38
18	0,28	0,31	0,34	0,37
19	0,27	0,30	0,34	0,36
20	0,26	0,29	0,33	0,35

Критические точки распределения Колмогорова

$$Q(\lambda) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2\lambda^2}$$

α	0,10	0,05	0,02	0,01
$\lambda_{кр}$	1,23	1,36	1,52	1,63

Двумерная выборка

(Рассматривается двумерная выборка: $(X_1,Y_1),(X_2,Y_2),\dots(X_n,Y_n)$

Выборочные характеристики:

$$\bar{X},\bar{Y},S_X^2,S_Y^2,S_{XY}^2$$

Свойства $\bar{X},\bar{Y},S_X^2,S_Y^2$ были изучены ранее. Рассмотрим выборочную ковариацию

$$S_{XY}^2=\frac{1}{n-1}\sum_{k=1}^{n}\left(X_k-\bar{X}\right)\left(Y_k-\bar{Y}\right)$$

Выборочная ковариация S_{XY}^2 является состоятельной оценкой ковариации σ_{12} в силу представления

В предположении , что выборка $(X_1,Y_1),(X_2,Y_2),\dots(X_n,Y_n)$ получена из двумерного нормального закона с параметрами $A=\begin{pmatrix}a_x\\a_y\end{pmatrix}$ и $\Sigma=\begin{pmatrix}\sigma_{11}&\sigma_{12}\\\sigma_{12}&\sigma_{22}\end{pmatrix}$ изучим совместный закон распределения пяти статистик

$$\bar{X}=\frac{1}{n}\sum_{k=1}^{n}X_k,\bar{Y}=\frac{1}{n}\sum_{k=1}^{n}Y_k,$$

$$\mu_{20}=\frac{1}{n}\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2,\mu_{02}=\frac{1}{n}\sum_{k=1}^{n}\left(Y_k-\bar{Y}\right)^2,\mu_{11}=\frac{1}{n}\sum_{k=1}^{n}\left(X_k-\bar{X}\right)\left(Y_k-\bar{Y}\right)$$

Определение. Выборочный коэффициент корреляции $r_в=\frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}}=\frac{S_{XY}^2}{\sqrt{S_X^2\cdot S_Y^2}}$

если $r=0$, то

$$t_в=\frac{r_в}{\sqrt{1-r_в^2}}\sqrt{n-2}\sim t(n-2)$$

Если $r\neq 0$, то

Рассматривается статистика $z=\frac{1}{2}\ln\frac{1+r_в}{1-r_в}=Arth\,r_в$ которая при $n\geq 10$

приближенно распределена $N\left(a_z,\frac{1}{\sqrt{n-3}}\right)$, где $a_z=Arth(r)+\frac{r}{2(n-1)}$

Следовательно, $(z-a_z)\sqrt{n-3}\sim N(0,1)$

Следствие. Доверительный интервал для коэффициента корреляции

$$1-\alpha = P\left(-u_{1-\frac{\alpha}{2}} < \left(z - a_z\right)\sqrt{n-3} < u_{1-\frac{\alpha}{2}} \right) =$$

$$= P\left(z - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} < a_z < z + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \right) =$$

$$= P\left(Arth\, r_в - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} < Arth\left(r\right) + \frac{r}{2\left(n-1\right)} < Arth\, r_в + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \right) =$$

$$\approx P\left(\text{th}\left(Arth\, r_в - \frac{r_в}{2\left(n-1\right)} - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \right) < r < \text{th}\left(Arth\, r_в - \frac{r_в}{2\left(n-1\right)} + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \right) \right)$$

Пример 6

Постройте доверительные интервалы для коэффициента корреляции, если $r_в = -0,36,\ n = 28\ ,\ 1-\alpha = 0,95\,;\ 0,99$

Поскольку $u_{1-\alpha/2} = u_{0,975} = 1,96$, получаем:

$$\rho_л = \tanh\left(\frac{1}{2}\ln\frac{0,64}{1,36} - \frac{0,36}{54} - \frac{1,96}{5} \right) = -0.642$$

$$\rho_{пр} = \tanh\left(\frac{1}{2}\ln\frac{0,64}{1,36} - \frac{0,36}{54} + \frac{1,96}{5} \right) = 0,022$$

Пример 7

Из генеральной совокупности, имеющей двумерное нормальное распределение, получена выборка объема $n = 67$. Выборочный коэффициент корреляции оказался равным $r_в = -0,159$. Можно ли считать, что наблюдаемые переменные отрицательно коррелированы, если уровень значимости $\alpha = 0,05$?

Решение.

На основе статистики $z=\frac{1}{2}\ln\frac{1+r_в}{1-r_в}=Arth\,r_в$, которая при $n\geq 10$ приближенно распределена $N\left(a_z,\frac{1}{\sqrt{n-3}}\right)$, где $a_z=Arth(r)+\frac{r}{2(n-1)}$, построим критическое множество для проверки гипотезы

$H_0: r=0$

против левосторонней альтернативы

$H_1: r<0$.

Если основная гипотеза состоит в том, что коэффициент корреляции равен 0, то используем статистику $(Arth\,r_в-a_z)\sqrt{n-3}\sim N(0,1)$, где $a_z=Arth(r)+\frac{r}{2(n-1)}$ при условии $H_0: r=0$ будет иметь вид

$$a_z=Arth(0)+\frac{0}{2(n-1)}=0$$

Найдем критическое множество

$$S=(r_в<C)=(Arth\,r_в<Arth\,C)$$

$$\alpha=P\left(\sqrt{n-3}Arth\,r_в<\sqrt{n-3}Arth\,C|r=0\right)=\Phi\left(\sqrt{n-3}Arth\,C\right)\Rightarrow$$

$$\Rightarrow\sqrt{n-3}Arth\,C=u_\alpha\Rightarrow C=th\left(\frac{u_{0.05}}{\sqrt{n-3}}\right)=-0.203$$

Таким образом критическое множество имеет вид

$$S=(r_в<-0.203)$$

Поскольку $r_в=-0{,}159\notin S=(r_в<-0.203)$, основная гипотеза принимается.

Второй способ

для проверки гипотезы

$H_0: r=0$

против левосторонней альтернативы

$H_1: r<0$.

В этом случае используем статистику

$$t_в = \frac{r_в}{\sqrt{1-r_в^2}}\sqrt{n-2} \sim t(n-2)$$

Найдём значение статистики

$$t_в = \frac{r_в}{\sqrt{1-r_в^2}}\sqrt{n-2} = \frac{-0.159}{\sqrt{1-0.159^2}}\sqrt{65} = -1.298$$

Построим критерий

$$\alpha = P\left(\frac{r_в}{\sqrt{1-r_в^2}}\sqrt{n-2} < t_\alpha(n-2)\middle| H_0\right)$$

$$\boldsymbol{S} = \left(t_в < t_\alpha(n-2)\right) = \left(t_в < t_{0,05}(65)\right) = \left(t_в < -1.669\right)$$

Так как $t_в = -1.298 > -1.669$, то основная гипотеза принимается.

Пример 8

Дано: $n_1 = 28,\ r_1 = 0,71,\ n_2 = 39, r_2 = 0,85\ \ \alpha = 0.01$

а) $H_0 : \rho_1 = \rho_2$

$H_1 : \rho_1 \neq \rho_2$

б) Для каких значений r_2 можно считать, что разность незначима?

Решение. а) Находим $z_B = \frac{Arth\, r_1 - Arth\, r_2}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}} = \frac{Arth\, 0,71 - Arth\, 0,85}{\sqrt{\frac{1}{36}+\frac{1}{25}}} = \frac{30}{\sqrt{61}}(-0.369) = -1.417$.

Находим квантиль: $u_{0.995} = 2.576$

Поскольку $|z_B| < u_{0.995}$ гипотеза H_0 принимается

б) гипотеза H_0 принимается, если $|z_B| < 2.576$, то есть

$$|Arth\, 0.71 - Arth\, r_2| < 2.576\frac{\sqrt{61}}{30} = 0.671$$

$$0.217 = Arth\, 0.71 - 0.671 < Arth\, r_2 < Arth\, 0.71 + 0.671 = 1.558$$

$$\tanh(0.217) < r_2 < \tanh(1.558)$$

$$0.214 < \tanh(0.217) < r_2 < 0.915$$