**Упражнения:**

**1.** От продукцията на две автоматични линии са взети извадки от $n_1$=16 и $n_2$=12 детайли и са изчислени $\bar{x}_1$=180 мм и $\bar{x}_2$=186 мм. От предварителен анализ е установено, че неточността на изработката им се описва с нормално разпределени величини с дисперсии съответно $\sigma_1^2=6$ мм$^2$ и $\sigma_2^2=11$ мм$^2$. С ниво на значимост $\alpha$=0,025 да се провери хипотезата $H_0$: $\mu_1=\mu_2$ срещу хипотезата $H_1$:$\mu_1<\mu_2$.

**2.** Предприятие има два автомата за производство на изделия, диаметърът на които се изследва. Взети са по 100 изделия за проверка. Средният диаметър на изработените от първия автомат изделия е 10,3 със стандартна грешка 1,2, а за изработените от втория – съответно 9,8 и 1,6. С ниво на значимост 0,01 да се проверят хипотезите: а) Точността на изработка на двата автомата е еднаква. б) Двата автомата произвеждат изделия с еднакви диаметри.

**3.** Две съвкупности от обекти, на които се изучава количественият признак X с нормално разпределение, са подложени на контрол. От първото множество е направена извадка с обем $n_1$=21 и е пресметната поправената дисперсия, която е равна на 0,75. От второто множество е получена извадка с обем $n_2$=11 с поправена дисперсия 0,25. Проверете хипотезата $H_0$: $DX_1=DX_2$. при конкурираща хипотеза $H_1$ :$DX_1 \neq DX_2$ и ниво на значимост 0,1.

**4.** При измерване на една и съща величина $X$, са получени следните резултати $x_1=-4,\ x_2=0,\ x_3=-2,\ x_4=2,\ x_5=6$ .а) Да се намерят $\bar{x}$ и $\tilde{s}_X$ . б) Да се определи доверителният интервал за $EX$ на случайната величина $X$ с доверителна вероятност $\gamma=0{,}95$. в) По друг метод от извадка с обем $n=9$ е получена извадъчна дисперсия $\tilde{s}_Y^2=16$ . Да се провери с ниво на значимост $\alpha=0{,}02$ хипотезата за еднаква точност на двата метода, т.е. хипотезата $H_0=\{DX=DY\}$ . Като конкурираща хипотеза да се разгледа $H_1=\{DX\neq DY\}$ (предполага се, че $X$ има нормално разпределение).

**5.** За сравняване на две системи за изучаване на чужд език са избрани случайно по 10 отлични и 10 посредствени студенти. По първата система студентите са усвоили предадения материал както следва (в проценти):

посредствени студенти (извадка $X_1$): 67, 56, 55, 61, 67, 56, 68, 53, 66, 65;

отлични студенти (извадка $X_2$): 87, 78, 86, 90, 77, 78, 81, 91, 82, 75.

След това същите групи студенти са обучавани по втората система като се получени следните резултати:

посредствени студенти (извадка $Y_1$): 32, 41, 51, 34, 55, 36, 39, 45, 36, 40;

отлични студенти (извадка $Y_2$): 90, 88, 83, 85, 94, 91, 95, 87, 90, 83.

а) Кои от двойките извадки са зависими и кои са независими?

Да се намерят: доверителен интервал с доверителна вероятност 0,9 за разликата в степента на усвояване по двете системи а) за посредствените студенти; б) за отличните студенти. Коя от системите е по-подходяща за посредствени студенти и коя за отлично работещи студенти. в) Да се провери с ниво на значимост 0,01 хипотезата, че обучаваните по система 1 по-равномерно усвояват знанията на студентите (отлични и посредствени) отколкото обучаваните по втората система. г) Да се сравнят с доверителен интервал при $\gamma=0{,}95$ средните постижения на всички студенти по система 1 и по система 2.



## КОРЕЛАЦИОНЕН И РЕГРЕСИОНЕН АНАЛИЗ

### §34. Двумерни извадки. Числени характеристики на извадката и точкови оценки. Корелационна таблица.

Много често в практиката се разглеждат два признака $X$ и $Y$ на генералната съвкупност като целта тук е не само изучаване на всеки от признаците, но и намиране на зависимост между тях, ако съществува такава. Например, изучават се теглото и ръстът на населението, броят на пушачите и броят на заболяванията на дихателните пътища и т.н. В такъв случай в резултат на наблюденията се получават двойки числа $(x_i, y_i)$, $i=1,\ldots,n$, затова получената извадка се нарича *двумерна извадка с обем n*. Тъй като приемаме, че обектите на изследване са случайно избрани, то ще считаме, че $(x_i, y_i)$ са реализации на двумерна случайна величина $(X,Y)$.

Резултатите от наблюденията се подреждат във вид на таблица:

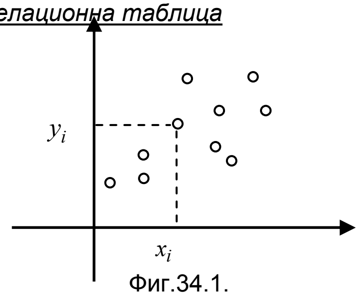
$$\begin{array}{c|cccc} x_i & x_1 & x_2 & \ldots & x_n \\ \hline y_i & y_1 & y_2 & \ldots & y_n \end{array}. \tag{34.1}$$

Графическо представяне на извадката е *диаграмата на разсейването* - съвкупност от точки с координати $(x_i, y_i)$ в равнината $Oxy$ (фиг. 34.1).

**1. Групиране на данните.** Данните, представени чрез таблица (34.1) са негрупирани. Ако има повторения на наблюдаваните стойности по-удобно е:

**1)** ***Групиране по стойности*** - *корелационна таблица*

| $y_j$ \ $x_i$ | $x_1$ … $x_k$ | $m_{y_j}$ |
|---|---|---|
| $y_1$ ⋮ $y_s$ | $m_{11}$ … $m_{k1}$ ⋮ $m_{1s}$ … $m_{ks}$ | $m_{y_1}$ ⋮ $m_{y_s}$ |
| $m_{x_i}$ | $m_{x_1}$ … $m_{x_k}$ | $n$ |



Фиг.34.1.

в която $m_{ij}$ – брой на наблюденията $(x_i, y_j)$, $n=\sum_{i=1}^{k}\sum_{j=1}^{s} m_{ij}$ – обем на извадката. Последната колона се състои от сумите по съответния ред, а последния ред – сумите по съответната колона, при това

$$m_{y_j}=\sum_{i=1}^{k} m_{ij} \text{ е честотата на вариантата } y_j$$

$$m_{x_i}=\sum_{j=1}^{s} m_{ij} \text{ е честотата на вариантата } x_i .$$

2) ***Групиране на данните по интервали***:

| $(y_{j-1},y_j)$ \ $(x_{i-1},x_i)$ | $(x_0,x_1)$ | $\ldots$ | $(x_{k-1},x_k)$ | |
|---|---|---|---|---|
| $(y_0,y_1)$ | $m_{11}$ | $\ldots$ | $m_{k1}$ | $m_{y_1}$ |
| $\vdots$ | $\vdots$ | | | $\vdots$ |
| $(y_{s-1},y_s)$ | $m_{1s}$ | $\ldots$ | $m_{ks}$ | $m_{y_s}$ |
| | $m_{x_1}$ | $\ldots$ | $m_{x_k}$ | $n$ |

,

където смисълът на $m_{ij}$, $m_{x_i}$ и $m_{y_j}$ е аналогичен.

**Пример 34.1.** Дадена е извадка с обем $n$=5: (8,1), (10,3), (5,1), (8,2) и (9,3). Да се получат негрупираното и групираното разпределение на извадката (корелационна таблица).

***Отг.*** Негрупираното и групираното разпределение са съответно

| $x_i$ | 8 | 10 | 5 | 8 | 9 |
|---|---|---|---|---|---|
| $y_i$ | 1 | 3 | 1 | 2 | 3 |

и

| $y_j$ \ $x_i$ | 5 | 8 | 9 | 10 | $m_{y_j}$ |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 2 |
| 2 | 0 | 1 | 0 | 0 | 1 |
| 3 | 0 | 0 | 1 | 1 | 2 |
| $m_{x_i}$ | 1 | 2 | 1 | 1 | 5 |

.

**2. Числени характеристики.**

**1.) *Числени характеристики на съставните компоненти* $X$ *и* $Y$.**

Очевидно, може разгледаме поотделно и да пресметнем всички познати ни числени характеристики на компонентите $X$ и $Y$ :

- *За признака* $X$: $\overline{x}$ , $s_x^2$, $\widetilde{s}_x^2$, начални $\overline{x^p}$ и централни $s_x^{(p)}$ моменти, които са и точкови оценки на съответните моменти на величината $X$. Формулите за пресмятането им се получават от статистическото разпределение на $X$, например,

<u>*извадъчна средна*</u> на $X$: $\overline{x}=\begin{cases}\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i, & \text{за негрупирани данни}\\ \dfrac{1}{n}\sum\limits_{i=1}^{k}m_{x_i}x_i & \text{за групирани данни}\end{cases}$ ;

<u>*начален момент от ред*</u> *p* :

$$\overline{x^p}=\begin{cases}\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i^p, & \text{за негрупирани}\\ \dfrac{1}{n}\sum\limits_{i=1}^{k}m_{x_i}x_i^p=\dfrac{1}{n}\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{s}m_{ij}x_i^p & \text{за групирани}\end{cases};$$

<u>*извадъчна и поправена дисперсия:*</u>

$$s_x^2=\frac{1}{n}\sum_{i=1}^{k}m_{x_i}(x_i-\overline{x})^2=\overline{x^2}-\overline{x}^2 \text{ и } \widetilde{s}_x^2=\frac{n}{n-1}s_x^2 .$$



- *За признака* $Y$: Аналогично получаваме $\overline{y}$, $s_y^2$, $\widetilde{s}_y^2$, начални $\overline{y^p}$ и централни $s_y^{(p)}$ моменти, които са и съответни точкови оценки на съответните моменти на величината $Y$.

**2.) *Числени характеристики, отразяващи връзките между наблюдаваните стойности***:

- <u>*Извадъчна ковариация*</u>:

$$s_{xy}=\begin{cases}\dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) & \text{за негрупирани данни}\\ \dfrac{1}{n}\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{s}m_{ij}(x_i-\overline{x})(y_j-\overline{y}) & \text{за групирани данни}\end{cases}.$$

Извадъчната ковариация се явява отместена оценка на ковариацията $\operatorname{cov}(X,Y)$ на двумерната случайна величина $(X,Y)$.

По-удобна формула за пресмятане на извадъчната ковариация:

$$s_{xy}=\overline{xy}-\overline{x}.\overline{y}\text{ , където }\overline{xy}=\begin{cases}\dfrac{1}{n}\sum\limits_{i=1}^{n}x_iy_i & \text{за негрупирани данни}\\ \dfrac{1}{n}\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{s}m_{ij}x_iy_j & \text{за групирани данни}\end{cases}.$$

- <u>*Поправена извадъчна корелация.*</u> $\widetilde{s}_{xy}=\dfrac{n}{n-1}s_{xy}$ - поправена оценка на $\operatorname{cov}(X,Y)$ :

- <u>*Извадъчен коефициент на корелация (коефициент на Пирсън):*</u>

$$r_{xy}=\frac{s_{xy}}{s_x.s_y} \qquad (34.2)$$

Този коефициент е оценка на коефициента на корелация $\rho_{xy}=\dfrac{\operatorname{cov}(X,Y)}{\sigma X.\sigma Y}$ на случайната величина $(X,Y)$ и има свойствата:

1) *не зависи от мерните единици на съставните*;
2) $-1\le r_{xy}\le 1$ ;
3) $r=\pm 1 \Leftrightarrow$ *между* $X$ *и* $Y$ *има линейна зависимост* $aX+bY+c=0$ .
4) *измерва силата на линейната зависимост между* $X$ и $Y$:

Силата на линейната корелация се определя по скалата:

- *Ако* $|r_{xy}|\le 0{,}3$ *– слаба зависимост (фиг. 35.1а).*
- *Ако* $0{,}3<|r_{xy}|<0{,}8$ *– средна зависимост(фиг.35.1б).;*
- *Ако* $|r_{xy}|\ge 0{,}8$ *– силна зависимост.(фиг. 35.1в).*

*<u>Рангов коефициент на корелация (коефициент на Спирмън).</u>*

Ранговият коефициент на корелация оценява силата на така наречената рангова корелационна зависимост между компонентите $X$ и $Y$. За определянето му по дадена извадка $(x_i, y_i)$, $i=1,\ldots,n$ подреждаме $x_i$ по големина във възходящ ред (т.е. получаваме вариационния ред $\hat{x}_i$).

> *Поредният номер $x_i'$ на $x_i$ във вариационния ред се нарича <u>ранг</u> на вариантата.*

По същия начин получаваме и ранговете $y_i'$ на вариантите $y_i$. Така получаваме извадка, състояща се от ранговете $(x_i', y_i')$ на двойките $(x_i, y_i)$.

> *Коефициентът $r_{x'y'}$ на корелация на извадката $(x_i', y_i')$, $i=1,\ldots,n$ се нарича <u>рангов коефициент</u> $r_s$ <u>на изходната извадка</u> $(x_i, y_i)$. $i=1,\ldots,n$ <u>(коефициент на Спирмън)</u>.*

**Пример 34.2.** Да се намери ранговият коефициент на корелация на извадката

| $x_i$ | 12,7 | 10,2 | 13,6 |
|---|---|---|---|
| $y_j$ | 2,7 | 3,2 | 3,0 |

***Решение:*** Образуваме вариационният ред на $X$: 10,2; 12,7; 13,6. Следователно, поредните номера на вариантите $x_1=12.7$, $x_2=10{,}2$, $x_3=13{,}6$ във вариационния ред са съответно 2, 1 и 3, т.е. $x_1'=2$, $x_2'=1$, $x_3'=3$.

По същия начин определяме, че $y_1'=1$, $y_2'=3$, $y_3'=2$. Така получаваме ранговата извадка

| $x_i'$ | 2 | 1 | 3 |
|---|---|---|---|
| $y_i'$ | 1 | 3 | 2 |

за която пресмятаме коефициента на корелация в следната последователност:

1) $\overline{x}'=\overline{y}'=\frac{1}{3}(1+2+3)=2$,

2) $s_{x'}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1}{3}[(2-2)^2+(1-2)^2+(3-2)^2]=\frac{2}{3}$,

$$s_{y'}^2=s_{x'}^2=\frac{2}{3}\,;\quad s_{x'}=s_{y'}=\sqrt{\frac{2}{3}}\,;$$

3) $s_{x'y'}=\frac{1}{n}\sum(x_i-\overline{x})(y_i-\overline{y})=\frac{1}{3}[(2-2)(1-2)+(1-2)(3-2)+(3-2)(2-2)]=-\frac{1}{3}$.

Следователно, $r_s=r_{x'y'}=\frac{-1/3}{\sqrt{2/3}\sqrt{2/3}}=-0{,}5$. ♦



Едно от предимствата на ранговия коефициент на корелация е лесното му пресмятане. Доказва се, че не е необходимо да се прилага формула (34.2), тъй като

$$r_s=1-\frac{6}{n(n^2-1)}\sum_{i=1}^{n}(x_i'-y_i')^2\ .$$

По тази формула за ранговия коефициент от пример **2** имаме

$$r_s=1-\frac{6}{3(3^2-1)}[(2-1)^2+(1-3)^2+(3-2)^2]=1-\frac{1}{4}.6=-0{,}5\ .$$

Доказва се също, че ако между $X$ и $Y$ има линейна зависимост, т.е. $aX+bY+c=0$, то $|\rho_{xy}|=|r_{xy}|=r_s|=1$.

**Забележка 34.1.** За намиране на числените характеристики при интервално статистическо разпределение на извадката интервалите се заменят със средите им.

**Забележка 34.2.** Пресмятането на числените характеристики се улеснява с помощта на *линейните трансформации*:

$$x_i=au_i+b,\quad y_i=cv_i+d\ .$$

Тогава: $\overline{x}=a\overline{u}+b$, $\overline{y}=c\overline{v}+d$, (34.3)

$s_x^2=a^2s_u^2$, $s_y^2=c^2s_v^2$, $s_{xy}=a.c.s_{uv}$, $r_{xy}=r_{uv}$ (34.4)

**Пример 34.3.** В резултат на 100 измервания е получена корелационната таблица:

| $(y_{j-1},y_j)$ \ $(x_{i-1},x_i)$ | (0,2) | (2,4) | (4,10) |
|---|---|---|---|
| (20,30) | 20 | – | – |
| (30,50) | – | 30 | 10 |
| (50,120) | – | 20 | 20 |

Да се намерят точковите оценки на двумерната случайна величина $(X,Y)$.

***Решение:*** Заменяме интервалите със средите им:

| $y_j$ \ $x_i$ | 1 | 3 | 7 | $m_{y_j}$ |
|---|---|---|---|---|
| 25 | 20 | 0 | 0 | 20 |
| 40 | 0 | 30 | 10 | 40 |
| 85 | 0 | 20 | 20 | 40 |
| $m_{x_i}$ | 20 | 50 | 30 | 100 |

Полагаме $u_i=x_i-3$, $v_j=\frac{1}{5}(y_i-40)=\frac{y_j}{5}-8$ и получаваме таблицата:

| $v_j$ \ $u_i$ | −2 | 0 | 4 | $m_{v_j}$ |
|---|---|---|---|---|
| −3 | 20 | 0 | 0 | 20 |
| 0 | 0 | 30 | 10 | 40 |
| 9 | 0 | 20 | 20 | 40 |
| $m_{u_i}$ | 20 | 50 | 30 | 100 |

, от която пресмятаме:

1) $\overline{u}=\frac{1}{100}(-2.20+0+4.30)=0{,}8$, $\overline{u^2}=\frac{1}{100}(4.20+16.30)=5{,}6$,

$s_u^2 = \overline{u^2} - \bar{u}^2 = 4{,}96$ , $s_u = \sqrt{4{,}96} = 2{,}23$ .

2) $\bar{v} = \frac{1}{100}(-3.20 + 0 + 9.40) = 3$ , $\overline{v^2} = \frac{1}{100}(9.20 + 81.40) = 34{,}20$ ,

$s_v^2 = \overline{v^2} - \bar{v}^2 = 25{,}20$ , $s_v = \sqrt{25{,}20} = 5{,}02$ .

3) $\overline{uv} = \frac{1}{100}[(-2)(-3).20 + 0 + 0 + 0 + 0.0.30 + 0 + 0 + 0 + 9.4.20] = 8{,}4$ ,

$s_{uv} = 8{,}4 - 0{,}8.3 = 6$ ;

4) $r_{uv} = \frac{6}{2{,}23.5{,}02} = 0.536$ .

От формули (34.3),(34.4): $x_i = u_i + 3 \Rightarrow \bar{x} = 0{,}8 + 3 = 3{,}8$ , $s_x^2 = s_u^2 = 4{,}96$ ,

$y_i = 5v_i + 40 \Rightarrow \bar{y} = 5.3 + 40 = 55$ , $s_y^2 = 25.s_v^2 = 630$ ,

$s_{xy} = 1.5.s_{uv} = 30$ $r_{xy} = r_{uv} = 0{,}536$ .

Така получаваме следните точкови оценки:

- За математическите очаквания $EX$ и $EY$ $\bar{x} = 3{,}8$ , $\bar{y} = 55$ .
- За дисперсиите $DX$ и $DY$ – поправените дисперсии
  $\tilde{s}_x^2 = \frac{100}{100-1} s_x^2 = 6{,}06$ и $\tilde{s}_y^2 = \frac{5}{4} 630 = 787{,}5$ .
- За ковариацията $\mathrm{cov}(X,Y)$ – извадковата ковариация
  $\tilde{s}_{xy} = \frac{100}{99} s_{xy} = 30{,}30$ .
- За коефициента на корелация $\rho_{XY}$ – извадъчния коефициент на корелация $r_{xy} = 0{,}536$ .♦

## §35. Елементи на корелационния анализ. Видове зависимости между две величини.

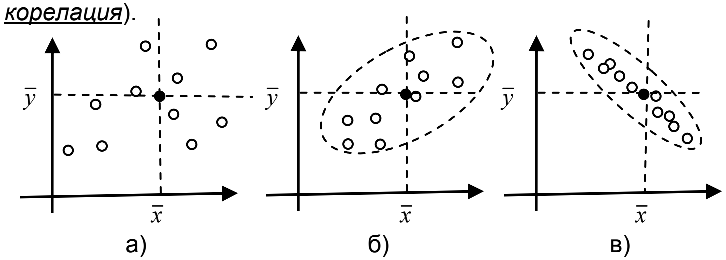
Нека за изучаване на признаците $X$ и $Y$ на генералната съвкупност са направени $n$ наблюдения. Разглеждайки $(X,Y)$ като случайна величина, а получените данни $(x_1, y_1), \ldots, (x_n, y_n)$ като нейни реализации, си поставяме задачата да изследваме зависимостта между компонентите $X$ и $Y$.

**1. Видове зависимости.** Две случайни величини $X$ и $Y$ се наричат *независими*, ако законът на едната величина не зависи от това каква стойност е приела другата. В такъв случай диаграмата на разсейването има вида на фиг. 35.1а.

Две величини могат да бъдат свързани с *функционална зависимост*, например $Y = X^2$ , $Y = aX + b$ - линейна зависимост и т.н.

Всички останали зависимости са *статистически зависимости*. Предмет на изследване тук ще бъде така наречената *корелационна зависимост (зависимост на тенденциите)* (фиг.35.1б,в) – зависимост при която при нарастване на $X$ се забелязва нарастване (*положителна корелация*) на стойностите на $Y$ или намаляване (*отрицателна корелация*).






Фиг.35.1

Както вече показахме (виж.**§19**), за независимите величини $\mathrm{cov}(X,Y) = 0$ , но не и обратното. Величини, за които $\mathrm{cov}(X,Y) = 0$ , се наричат *некорелирани*, а за които $\mathrm{cov}(X,Y) \neq 0$ - *корелирани*.

Основни инструменти за изследване на корелацията на две величини са извадъчната ковариация $s_{xy}$ и особено извадъчният коефициент на корелация $r_{xy}$ , които са оценки на съответно на $\mathrm{cov}(X,Y)$ и $\rho_{XY}$ . Например:

- на фиг.35.1б) повече от произведенията $(x_i - \bar{x})(y_i - \bar{y})$ са положителни, откъдето $s_{xy} = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}) > 0$ , а и от там $r_{xy} > 0$ (*положителна корелация)*
- на фиг.35.1в) е представен случая $s_{xy} < 0$ , $r_{xy} < 0$ (*отрицателна корелация)*
- за фиг.35.1а) $s_{xy}$ и $r_{xy}$ са близки до нула (*некорелирани величини*).

*<u>Корелационният анализ</u> е дял от статистиката, които изследва корелационната зависимост между две величини. Основни задачи:*

- *Намиране на оценки на коефициента на корелация $\rho_{XY}$ .*
- *Доказване на наличието или отсъствието на корелационна зависимост.*
- *Изследване на силата на линейната връзка между величините.*

**2. Доверителни интервали за коефициента на корелация $\rho_{XY}$ на нормално разпределени случайни величини $X$ и $Y$.**

Нека по дадена двумерна извадка на два нормално разпределени признака $X$ и $Y$ на генералната съвкупност и нека е изчислена оценката $r_{xy}$ на коефициента на корелация $\rho_{XY}$ . За определяне на доверителния интервал за $\rho_{XY}$ се използва, че разпределението на $\rho_{XY}$ е близко до нормално разпределение $N(m*,\sigma*)$, с параметри

$$m*=\frac{1}{2}\ln\frac{1+r_{xy}}{1-r_{xy}}, \quad \sigma*=\frac{1}{\sqrt{n-3}}.$$

*Намирането на **доверителния интервал** за $\rho_{XY}$ с доверителна вероятност $\gamma$ се извършва в следната последователност:*

- *изчисляваме* $m*=\frac{1}{2}\ln\frac{1+r_{xy}}{1-r_{xy}}$ *и квантилът* $Z_{\frac{1+\gamma}{2}}$ *на* $Z\sim N(0,1)$;
- *изчисляваме* $z_1=m*-Z_{\frac{1+\gamma}{2}}.\frac{1}{\sqrt{n-3}}$ *и* $z_2=m^*+Z_{\frac{1+\gamma}{2}}.\frac{1}{\sqrt{n-3}}$;
- *Определяме доверителния интервал* $\rho_{xy}\in\left(\frac{e^{2z_1}-1}{e^{2z_1}+1}, \frac{e^{2z_2}-1}{e^{2z_2}+1}\right)$

Посоченото правило за намиране на доверителен интервал може да се приложи и за оценка на ранговия коефициент на корелация.

Освен като интервална оценка за $\rho_{XY}$, с помощта на доверителния интервал могат да се проверяват предположения за вида на зависимостта между $X$ и $Y$:

- *Ако доверителният интервал съдържа числото нула, то с вероятност $\gamma$ може да са твърди, че величините $X$ и $Y$ не са корелирани;*
- *Ако доверителният интервал съдържа числата 1 или -1, то с вероятност $\gamma$ може да се твърди , че между $X$ и $Y$ има силна линейна връзка.*

**3. Проверка на хипотезата за значимостта на корелационната зависимост между нормално разпределени величини $X$ и $Y$.**

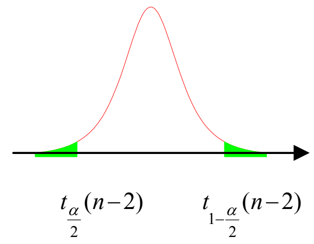
Разглеждаме хипотезата $H_0$ за некорелираност на величините $X$ и $Y$, т.е. $H_0=\{\rho_{XY}=0\}$.

Статистическият критерий, с който се проверява тази хипотеза, е *t*-разпределението на Стюдънт с *n*-2 степени на свобода $t(n-2)$, като наблюдаваната стойност е $t_{\text{набл.}}=\frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$

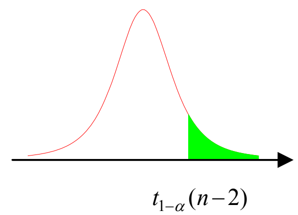
Критичната област се определя в зависимост от конкуриращата хипотеза и избраното ниво на значимост $\alpha$:



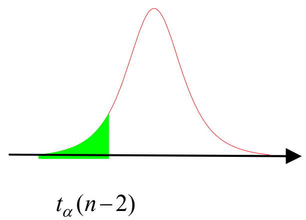
- Ако $H_1=\{\rho_{XY}\neq 0\}$:




- Ако $H_1=\{\rho_{XY}<0\}$:




- Ако $H_1=\{\rho_{XY}>0\}$




Така, като се използва четността на плътността на *t*-разпределението, т.е. $t_\alpha(n-1)=-t_{1-\alpha}(n-1)$, проверката на хипотезата $H_0$ при ниво на значимост $\alpha$ по дадена извадка с обем $n$ се извършва по следния начин:

1) *Изчислява се* $t_{\text{набл.}}=\frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$;

*2) От таблицата за квантилите на t-разпределението с n-1 степени на свобода съгласно конкуриращата хипотеза се определя квантилът:*

- $t_{\text{кр}}=t_{1-\frac{\alpha}{2}}(n-2)$ *при* $H_1=\{\rho_{XY}\neq 0\}$. *Област на приемане на* $H_0$ *е* $|t_{\text{набл.}}|<t_{\text{кр}}$.
- $t_{\text{кр}}=t_{1-\alpha}(n-2)$ *при* $H_1=\{\rho_{XY}>0\}$. *Област на приемане на* $H_0$ *е* $t_{\text{набл.}}<t_{\text{кр}}$
- $t_{\text{кр}}=-t_{1-\alpha}(n-2)$ *при* $H_1=\{\rho_{XY}<0\}$. *Област на приемане на* $H_0$ *е* $t_{\text{набл.}}>t_{\text{кр}}$.

Приемането на хипотезата означава, че с вероятност 1-$\alpha$ между няма корелационна зависимост, а отхвърлянето – че има корелационна зависимост.

Същият тест се прилага и при проверка на хипотезата за значимост на ранговия коефициент на корелация, т.е. за отхвърляне или приемане на рангова корелационна зависимост.

**Пример. 35.1.** В конни състезания състезателните коне, които са номерирани съгласно ръста си, са заели следните места: 6, 5, 1, 4, 2, 7, 8, 10, 3, 9. С ниво на значимост $\alpha$=0,05 да се провери хипотезата, че няма рангова корелационна зависимост между ръста и мястото, което състезателният кон е заел ( $H_0=\{r_s=0\}$ ).

***Решение.*** Означаваме с $x_i$ номерата на конете съгласно ръста им, а с $y_i$ - местата им в класирането. Очевидно, $x_i$ са ранговете на теглата на конете, а $y_i$ - ранговете на времената им за изминаване на разстоянието.

Нанасяме данните в таблицата $\begin{array}{c|cccccccccc} x_i & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \hline y_i & 6 & 5 & 1 & 4 & 2 & 7 & 8 & 10 & 3 & 9 \end{array}$.

Пресмятаме ранговия коефициент:

$$r_s=1-\frac{6}{n(n^2-1)}\sum_{i=1}^{n}(x_i-y_i)^2=1-\frac{6}{990}[5^2+3^2+2^2+3^2+1^2+1^2+2^2+6^2+1^2]=0,4545$$

Изчисляваме наблюдаваната стойност:

$$t_{\text{набл.}}=\frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}=\frac{0,45.\sqrt{8}}{\sqrt{1-0,45^2}}=\frac{1,273}{0,893}=1,42 .$$

От таблицата за $t$-разпределението изчисляваме

$$t_{\text{кр.}}=t_{1-\frac{\alpha}{2}}(10-2)=t_{0,975}(8)=2,31 .$$

Тъй като $|t_{\text{набл.}}|<t_{\text{кр.}}$ , то нямаме основание да отхвърлим хипотезата $H_0$, т.е. няма рангова корелационна зависимост между ръста и мястото на класиране.

**Упражнения.**

1. За изследването на променливите $X$ и $Y$ е получена извадката $\begin{array}{c|ccccc} x_i & -1 & 2 & 2 & 3 & 6 \\ \hline y_i & 0 & -1 & 2 & 4 & 8 \end{array}$. Да се представят графически резултатите, да се намери точкова и интервална оценка с доверителна вероятност 0,99 на коефициента на корелация.

2. Направени са следните наблюдения $\begin{array}{c|ccccc} X & 2 & 4 & 7 & 6 & 8 \\ \hline Y & 2 & 12 & 16 & 18 & 21 \end{array}$ на променливите $X$ и $Y$, за които се предполага, че са нормално разпределени случайни величини. Да се провери хипотезата за корелираност на величините .