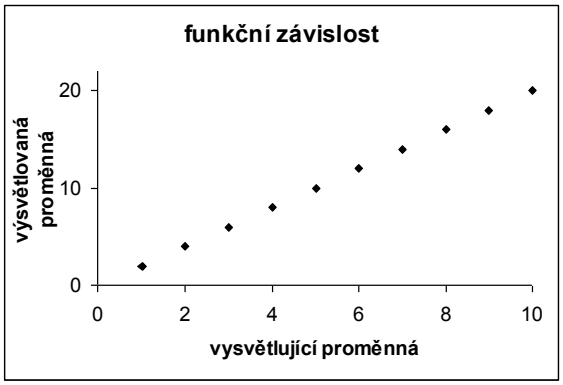
10. Korelační analýza

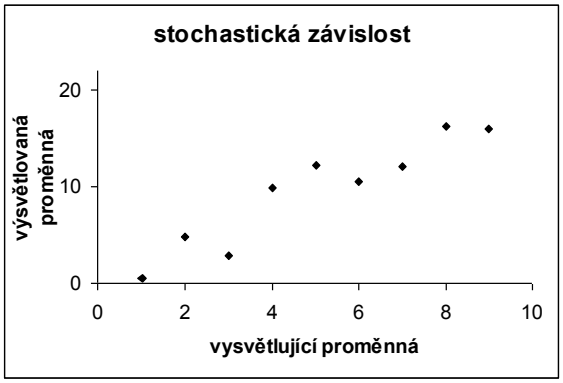
Motivace

Uvažme náhodné veličiny X, Y, které jsou aspoň ordinálního typu. Tyto náhodné veličiny mohou mít různý vztah:

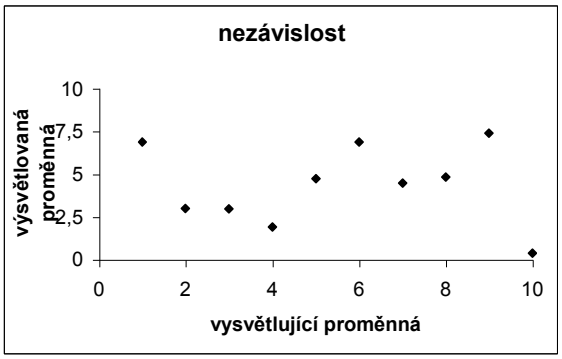
- Deterministická (funkční) závislost: jedna náhodná veličina je spjata s druhou náhodnou veličinou funkční závislostí vyjádřenou předpisem $Y = g(X)$, např. X – poloměr náhodně vybrané sériově vyráběné kuličky do kuličkových ložisek, $Y = \frac{4}{3}\pi\zeta^3$ - objem této kuličky. Každé realizaci náhodné veličiny X (vysvětlující proměnná) je přiřazena právě jedna realizace náhodné veličiny Y (vysvětlovaná proměnná).



- Stochastická závislost: jedna náhodná veličina ovlivňuje v různé míře druhou náhodnou veličinu, např. X – věk pracovníka v letech, Y – počet dnů absence za rok. Každé realizaci náhodné veličiny X může být přiřazeno více realizací náhodné veličiny Y. Závislost může být jednostranná i oboustranná.



- Stochastická nezávislost: náhodné veličiny se navzájem neovlivňují, např. házíme-li naráz dvěma kostkami a označíme X – počet ok padlých na jedné kostce, Y – počet ok padlých na druhé kostce, pak náhodné veličiny X, Y jsou stochasticky nezávislé.



X a Y jsou stochasticky nezávislé, když platí: $\forall x, y \in \mathbb{R}^2 : F(x, y) = F_X(x) \cdot F_Y(y)$

X a Y jsou nekorelované, když platí C(X, Y) = 0 (tj. mezi X a Y není žádný lineární vztah).

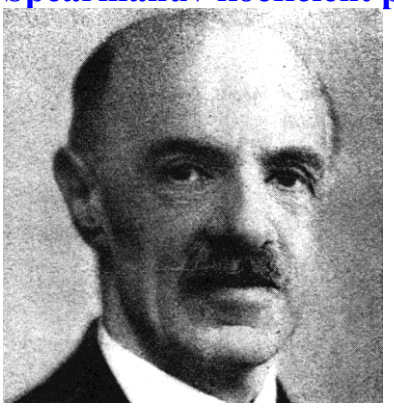
Ze stochastické nezávislosti vyplývá nekorelovanost, avšak z nekorelovanosti nevyplývá stochastická nezávislost.

Korelační analýza:

- zkoumá, zda existuje závislost mezi dvěma náhodnými veličinami X, Y, které jsou buď ordinálního nebo intervalového či poměrového typu. Důležité – nelze se spokojit s formálním matematickým popisem závislosti, závislost musí být logicky zdůvodnitelná!
- pomocí Pearsonova či Spearmanova koeficientu korelace měří těsnost této závislosti
- pro náhodné veličiny intervalového a poměrového typu je založena na předpokladu, že dvourozměrný náhodný vektor $\begin{pmatrix} X \\ Y \end{pmatrix}$ se řídí dvourozměrným normálním rozložením $N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$, kde

  $\mu_1 = E(X)$, $\mu_2 = E(Y)$, $\sigma_1^2 = D(X)$, $\sigma_2^2 = D(Y)$, $\rho = R(X,Y)$
- při výraznějším porušení předpokladu dvourozměrné normality doporučuje použití metod, které jsou určeny pro náhodné veličiny ordinálního typu

Spearmanův koeficient pořadové korelace



Charles Edward Spearman (1863 – 1945): Britský psycholog a statistik, zakladatel faktorové analýzy

Nechť X,Y jsou náhodné veličiny ordinálního typu (tj. obsahová interpretace je možná jenom u relace rovnosti a relace uspořádání).
Pořídíme dvourozměrný náhodný výběr $(X_1, Y_1), ..., (X_n, Y_n)$ z rozložení, jímž se řídí náhodný vektor (X, Y). Označíme R_i pořadí náhodné veličiny X_i a Q_i pořadí náhodné veličiny Y_i, i = 1, ..., n.

Spearmanův koeficient pořadové korelace: $r_S = 1 - \frac{6}{n(n^2-1)} \sum_{j=1}^{n} (R_i - Q_i)^2$.

Tento koeficient nabývá hodnot mezi –1 a 1. Čím je bližší 1, tím je silnější přímá pořadová závislost mezi veličinami X a Y, čím je bližší –1, tím je silnější nepřímá pořadová závislost mezi veličinami X a Y. Teoretická hodnota Spearmanova koeficientu se značí ρ_S.

Testování nezávislosti ordinálních veličin

Na hladině významnosti α testujeme hypotézu H_0: X, Y jsou pořadově nezávislé náhodné veličiny proti
- oboustranné alternativě H_1: X, Y jsou pořadově závislé náhodné veličiny
- levostranné alternativě H_1: mezi X a Y existuje nepřímá pořadová závislost
- pravostranné alternativě H_1: mezi X a Y existuje přímá pořadová závislost).

Jako testová statistika slouží Spearmanův koeficient pořadové korelace r_S.
Nulovou hypotézu zamítáme na hladině významnosti α ve prospěch
- oboustranné alternativy, když $|r_S| \geq r_{S,1-\alpha}(n)$
- levostranné alternativy, když $r_S \leq - r_{S,1-2\alpha}(n)$
- pravostranné alternativy, když $r_S \geq r_{S,1-2\alpha}(n)$,

kde $r_{S,1-\alpha}(n)$ je kritická hodnota, kterou pro α = 0,05 nebo 0,01 a n ≤ 30 najdeme v tabulkách. Pozor – kritické hodnoty pro jednostranné alternativy se v běžně dostupných tabulkách nenajdou.

Asymptotické varianty testu

Pro n > 20 lze použít testovou statistiku $T_0 = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}}$, která se v případě platnosti nulové hypotézy asymptoticky řídí rozložením t(n-2).
Kritický obor pro oboustrannou alternativu:
$W = (-\infty, -t_{1-\alpha/2}(n-2)) \cup (t_{1-\alpha/2}(n-2), \infty)$
Kritický obor pro levostrannou alternativu:
$W = (-\infty, -t_{1-\alpha}(n-2))$
Kritický obor pro pravostrannou alternativu:
$W = (t_{1-\alpha}(n-2), \infty)$.

Hypotézu o pořadové nezávislosti náhodných veličin X, Y zamítáme na asymptotické hladině významnosti α, když $t_0 \in W$.
Upozornění: Systém STATISTICA používá tuto variantu testu pořadové nezávislosti bez ohledu na rozsah náhodného výběru.

Pro n > 30 lze použít testovou statistiku $r_s\sqrt{n-1}$. Platí-li H_0, pak $r_s\sqrt{n-1} \approx N(0, 1)$. Nulovou hypotézu tedy zamítáme na asymptotické hladině významnosti α ve prospěch
oboustranné alternativy, když $r_S\sqrt{n-1} \in (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$,
levostranné alternativy, když $r_S\sqrt{n-1} \in (-\infty, -u_{1-\alpha}\rangle$,
pravostranné alternativy, když $r_S\sqrt{n-1} \in \langle u_{1-\alpha}, \infty)$

Příklad na testování pořadové nezávislosti (jsou známa pořadí):
Dva lékaři hodnotili stav sedmi pacientů po témž chirurgickém zákroku. Postupovali tak, že nejvyšší pořadí dostal nejtěžší případ.

Číslo pacienta	1	2	3	4	5	6	7
Hodnocení 1. lékaře	4	1	6	5	3	2	7
Hodnocení 2. lékaře	4	2	5	6	1	3	7

Vypočtěte Spearmanův koeficient a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou lékařů jsou pořadově nezávislá.
Řešení:
Na hladině významnosti 0,05 testujeme H_0: X, Y jsou pořadově nezávislé náhodné veličiny proti oboustranné alternativě H_1: X, Y jsou pořadově závislé náhodné veličiny. V tomto příkladě přímo známe pořadí R_i (tj. hodnocení 1. lékaře) a pořadí Q_i (tj. hodnocení 2. lékaře). Vypočteme

$$r_S = 1 - \frac{6}{7(7^2-1)}\left[(4-4)^2 + (1-2)^2 + (6-5)^2 + (5-6)^2 + (3-1)^2 + (2-3)^2 + (7-7)^2\right] = 0,857 .$$

Kritická hodnota: $r_{S,0,95}(7) = 0,745$. Protože $0,857 \geq 0,745$, nulovou hypotézu zamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA
Vytvoříme datový soubor o dvou proměnných X (hodnocení 1. lékaře), Y (hodnocení 2. lékaře) a sedmi případech. Do proměnných X a Y zapíšeme zjištěná hodnocení.

	1 X	2 Y
1	4	4
2	1	2
3	6	5
4	5	6
5	3	1
6	2	3
7	7	7

Statistiky – Neparametrické statistiky – Korelace – OK – vybereme Vytvořit detailní report - Proměnné X, Y – OK – Spearmanův koef. R. Dostaneme tabulku

	Spearmanovy korelace (dva lekari.sta) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000			
Dvojice proměnných	Počet plat.	Spearman R	t(N-2)	Úroveň p
X & Y	7	0,857143	3,721042	0,013697

Spearmanův koeficient pořadové korelace nabývá hodnoty 0,857, testová statistika se realizuje hodnotou 3,721, odpovídající p-hodnota je 0,0137, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o pořadové nezávislosti hodnocení dvou lékařů ve prospěch oboustranné alternativy.

Příklad na testování pořadové nezávislosti (pořadí musíme stanovit):

Jsou dány realizace náhodného výběru z dvourozměrného rozložení, kterým se řídí náhodný vektor (X,Y): (2,5 13,4), (3,4 15,2), (1,3 11,8), (5,8 13,1), (3,6 14,5). Na hladině významnosti 0,05 testujte hypotézu, že náhodné veličiny jsou pořadově nezávislé proti oboustranné alternativě.

Řešení:

x_i	2,5	3,4	1,3	5,8	3,6
y_i	13,4	15,2	11,8	13,1	14,5
R_i	2	3	1	5	4
Q_i	3	5	1	2	4
$(R_i-Q_i)^2$	1	4	0	9	0

Testová statistika: $r_S = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2 = 1 - \frac{6}{5 \cdot 24}14 = 0,3$

Kritická hodnota: pro n = 5 a α = 0,05 je kritická hodnota 0,9. Protože testová statistika se realizuje hodnotou 0,3, hypotézu o pořadové nezávislosti veličin X a Y nezamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA

Postupujeme úplně stejně jako v předešlém případě. Výstupní tabulka má tvar:

	Spearmanovy korelace (poradova korelace.sta) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000			
Dvojice proměnných	Počet plat.	Spearman R	t(N-2)	Úroveň p
X & Y	5	0,300000	0,544705	0,623838

Spearmanův koeficient pořadové korelace nabývá hodnoty 0,3, testová statistika se realizuje hodnotou 0,5447, odpovídající p-hodnota je 0,6238, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o pořadové nezávislosti veličin X, Y.

Pearsonův koeficient korelace

Karl Pearson (1857 – 1936): Britský statistik

Číslo

$$R(X,Y) = \begin{cases} E\left(\frac{X - E(X)}{\sqrt{D(X)}} \cdot \frac{Y - E(Y)}{\sqrt{D(Y)}}\right) = \frac{C(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} & \text{pro } \sqrt{D(X)}\sqrt{D(Y)} > 0 \\ \text{jinak} \end{cases}$$

se nazývá Pearsonův koeficient korelace.
(Pro výpočet Pearsonova koeficentu korelace musíme znát simultánní distribuční funkci Φ(x,y) v obecném případě resp. simultánní hustotu pravděpodobnosti φ(x,y) ve spojitém případě resp. simultánní pravděpodobnostní funkci π(x,y) v diskrétním případě.)

Vlastnosti Pearsonova koeficientu korelace

a) $R(a_1, Y) = R(X, a_2) = R(a_1, a_2) = 0$

b) $R(a_1 + b_1X, a_2 + b_2Y) = \operatorname{sgn}(b_1b_2)\, R(X, Y) = \begin{cases} R(X,Y) & \text{pro } b_1b_2 > 0 \\ -R(X,Y) & \text{pro } b_1b_2 < 0 \end{cases}$

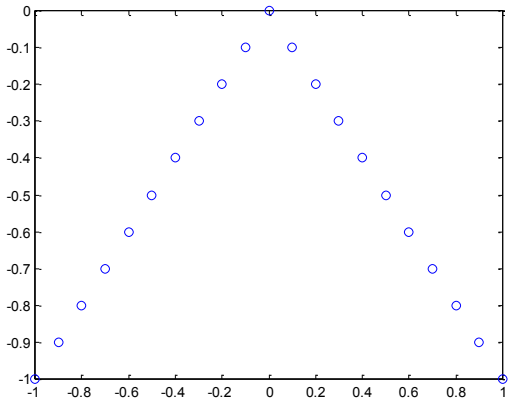
c) $R(X, X) = 1$ pro $D(X) \neq 0$, $R(X, X) = 0$ jinak

d) $R(X, Y) = R(Y, X)$

e) $|R(X,Y)| \leq$ a rovnost nastane tehdy a jen tehdy, když mezi veličinami X, Y existuje s pravděpodobností 1 úplná lineární závislost, tj. existují konstanty a, b tak, že pravděpodobnost P(Y = a + bX) = 1. Přitom R(X, Y) = 1, když b > 0 a R(X, Y) = -1, když b < 0. (Uvedená nerovnost se nazývá Cauchyova – Schwarzova – Buňakovského nerovnost.)

Z vlastností Pearsonova koeficientu korelace vyplývá, že se hodí pouze k měření těsnosti lineárního vztahu veličin X a Y. Při složitějších závislostech může dojít k paradoxní situaci, že Pearsonův koeficient korelace je nulový.

Ilustrace:



Definice nekorelovanosti

Je-li R(X, Y) = 0, pak řekneme, že náhodné veličiny jsou nekorelované. (Znamená to, že mezi X a Y neexistuje žádná lineární závislost.)
Je-li R(X, Y) > 0, pak řekneme, že náhodné veličiny jsou kladně korelované. (Znamená to, že s růstem hodnot veličiny X rostou hodnoty veličiny Y a s poklesem hodnot veličiny X klesají hodnoty veličiny Y.)
Je-li R(X, Y) < 0, pak řekneme, že náhodné veličiny jsou záporně korelované. (Znamená to, že s růstem hodnot veličiny X klesají hodnoty veličiny Y a s poklesem hodnot veličiny X rostou hodnoty veličiny Y.)

Výběrový koeficient korelace

Nechť $(X_1, Y_1), ..., (X_n, Y_n)$ náhodný výběr rozsahu n z dvourozměrného rozložení daného distribuční funkcí $\Phi(x,y)$. Z tohoto dvourozměrného náhodného výběru můžeme stanovit:

výběrové průměry $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$, $M_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i$,

výběrové rozptyly $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - M_1)^2$, $S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - M_2)^2$,

výběrovou kovarianci $S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$ a s jejich pomocí zavedeme

výběrový koeficient korelace $R_{12} = \begin{cases} \frac{1}{n-1}\sum_{i=1}^{n}\frac{X-M_1}{S_1}\cdot\frac{Y-M_2}{S_2} = \frac{S_{12}}{S_1 S_2} & \text{pro } S_1 S_2 > 0 \\ 0 & \text{jinak} \end{cases}$.

Vlastnosti Pearsonova koeficientu korelace se přenášejí i na výběrový koeficient korelace.
(Spearmanův koeficient pořadové korelace odpovídá Pearsonovu koeficientu korelace aplikovanému na pořadí.)

Příklad: Výpočet realizace výběrového koeficientu korelace

U 65 zaměstnanců jisté firmy byla zjišťována délka praxe v letech (veličina X) a výška prémií v Kč (veličina Y). Dvourozměrné rozložení četností je dáno kontingenční tabulkou:

x	y						
	1250	1750	2250	2750	3250	3750	4250
12,5	5	3	0	0	0	0	0
17,5	2	4	4	0	0	0	0
22,5	0	1	6	7	4	0	0
27,5	0	0	1	3	7	1	0
32,5	0	0	0	1	10	5	1

Vypočtěte realizaci r_{12} výběrového koeficientu korelace R_{12} a interpretujte jeho hodnotu. Pro úsporu času máte uvedeny následující součty:

$\sum_{j=} x_{[j]} = 562,5, \sum_{k=} y_{[k]} = 72750, \sum_{j=} x_{[j]}^2 = 10456, \sum_{k=} y_{[k]}^2 = 198562500$,

$\sum_{j=1}^{5}\sum_{k=1}^{7} n_{jk} x_{[j]} y_{[k]} = 4446875$

Řešení:

Vypočteme

průměrnou délku praxe: $m_1 = \frac{562,5}{65} = 8,654$,

průměrnou výšku prémií: $m_2 = \frac{72750}{65} = 1119,231$

rozptyl délky praxe: $s_1^2 = \frac{1}{64}\left(10456 - 65\cdot\left(\frac{562,5}{65}\right)^2\right) = 15,25$

rozptyl výše prémií: $s_2^2 = \frac{1}{64}\left(198562500 - 65\cdot\left(\frac{72750}{65}\right)^2\right) = 516346$

kovariance délky praxe a výše prémií:

$$s_{12} = \frac{1}{64}\left(1446875 - 65 \cdot \frac{1562,5}{65} \cdot \frac{172750}{65} \right) = 1597,4$$

koeficient korelace délky praxe a výše prémií: $r_{12} = \frac{4597,4}{\sqrt{45,25}\sqrt{616346}} = 0,8705$

Hodnota koeficientu korelace svědčí o tom, že mezi délkou praxe a výškou prémií existuje dosti silná přímá lineární závislost – čím delší praxe, tím vyšší prémie.

Pearsonův koeficient korelace dvourozměrného normálního rozložení

Jak bylo uvedeno v motivaci, korelační analýza předpokládá, že daný náhodný výběr pochází z dvourozměrného normálního rozložení. Proč je tento předpoklad tak důležitý? Odpověď poskytne následující věta.

Nechť náhodný vektor (X, Y) má dvourozměrné normální rozložení s hustotou

$$\varphi(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]},$$

přičemž $\mu_1 = E(X)$, $\mu_2 = E(Y)$, $\sigma_1^2 = D(X)$, $\sigma_2^2 = D(Y)$, $\rho = R(X,Y)$.
Marginální hustoty jsou:

$$\varphi_1(x) = \int_{-\infty}^{\infty} \varphi(x,y)\,dy = \ldots = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

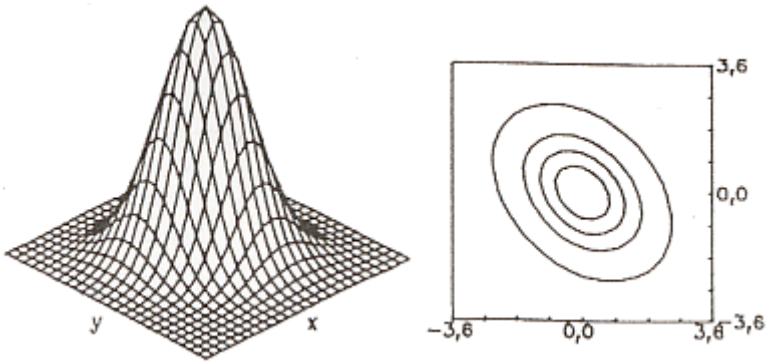
$$\varphi_2(y) = \int_{-\infty}^{\infty} \varphi(x,y)\,dx = \ldots = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

Je-li $\rho = 0$, pak pro $\forall (x,y) \in \mathbb{R}^2 : \varphi(x,y) = \varphi_1(x)\varphi_2(y)$, tedy náhodné veličiny X, Y jsou stochasticky nezávislé. Jinými slovy: stochastická nezávislost složek X, Y normálně rozloženého náhodného vektoru je ekvivalentní jejich nekorelovanosti. Pro jiná dvourozměrná rozložení to neplatí!

Upozornění: nadále budeme předpokládat, že $(X_1, Y_1), \ldots, (X_n, Y_n)$ je náhodný výběr rozsahu n z dvourozměrného normálního rozložení

$$N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right).$$

Předpoklad dvourozměrné normality lze orientačně ověřit pomocí dvourozměrného tečkového diagramu: tečky by měly zhruba rovnoměrně vyplnit vnitřek elipsovitého obrazce. Vrstevnice hustoty dvourozměrného normálního rozložení jsou totiž elipsy:
Graf hustoty a vrstevnice dvourozměrného normálního rozložení s parametry $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = -0,75$:



Do dvourozměrného tečkového diagramu můžeme ještě zakreslit 100(1-α)% elipsu konstantní hustoty pravděpodobnosti. Bude-li více než 100α% teček ležet vně této elipsy, svědčí to o porušení dvourozměrné normality. Bude-li mít hlavní osa elipsy kladnou resp. zápornou směrnici, znamená to, že mezi veličinami X a Y existuje určitý stupeň přímé resp. nepřímé lineární závislosti.

Testování hypotézy o nezávislosti

Na hladině významnosti α testujeme H_0: X, Y jsou stochasticky nezávislé náhodné veličiny (tj. $\rho = 0$) proti
- oboustranné alternativě H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny (tj. $\rho \neq 0$)
- levostranné alternativě H_1: X, Y jsou záporně korelované náhodné veličiny (tj. $\rho < 0$)
- pravostranné alternativě H_1: X, Y jsou kladně korelované náhodné veličiny (tj. $\rho > 0$).

Testová statistika má tvar: $T_0 = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}$.

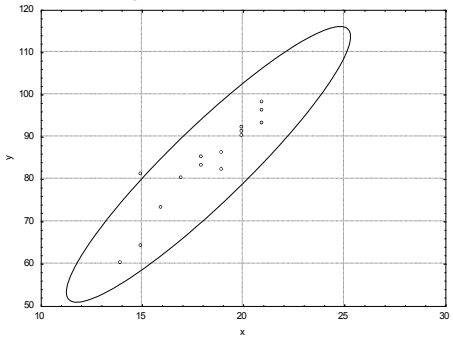
Platí-li nulová hypotéza, pak $T_0 \sim t(n-2)$.
Kritický obor pro test H_0 proti
- oboustranné alternativě: $W = (-\infty, -t_{1-\alpha/2}(n-2)) \cup (t_{1-\alpha/2}(n-2), \infty)$,
- levostranné alternativě: $W = (-\infty, -t_{1-\alpha}(n-2))$,
- pravostranné alternativě: $W = (t_{1-\alpha}(n-2), \infty)$.

H_0 zamítáme na hladině významnosti α, když $t_0 \in W$.

Příklad: Testování hypotézy o nezávislosti proti oboustranné alternativě

V dílně pracuje 15 dělníků. Byl u nich zjištěn počet směn odpracovaných za měsíc (náhodná veličina X) a počet zhotovených výrobků (náhodná veličina Y):
X 20 21 18 17 20 18 19 21 20 14 16 19 21 15 15
Y 92 93 83 80 91 85 82 98 90 60 73 86 96 64 81.
Orientačně ověřte dvourozměrnou normalitu dat, vypočtěte výběrový koeficient korelace mezi X a Y a na hladině 0,01 testujte hypotézu o nezávislosti X a Y proti oboustranné alternativě.

Řešení: Dvourozměrnou normalitu dat ověříme pomocí dvourozměrného tečkového diagramu.



Vidíme, že předpoklad dvourozměrné normality je oprávněný.
Vypočteme realizace

výběrových průměrů: $m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = 18{,}267$, $m_2 = \frac{1}{n}\sum_{i=1}^{n} y_i = 83{,}6$,

výběrových rozptylů: $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - m_1)^2 = 5{,}6381$, $s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - m_2)^2 =$ 121,4,

výběrové kovariance: $s_{12} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - m_1)(y_i - m_2) = 24{,}2571$,

výběrového koeficientu korelace: $r_{12} = \frac{s_{12}}{s_1 s_2} = 0{,}927$.

Realizace testové statistiky: $t_0 = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} = 8{,}912$,

kritický obor $W = (-\infty, -t_{0{,}995}(3)) \cup (t_{0{,}995}(3), \infty) = (-\infty, -3{,}012) \cup (3{,}012, \infty)$.

Protože $t_0 \in W$, hypotézu o nezávislosti veličin X a Y zamítáme na hladině významnosti 0,01. S rizikem omylu nejvýše 1% jsme tedy prokázali, že mezi počtem směn odpracovaných za měsíc a počtem zhotovených výrobků existuje závislost.

Výpočet pomocí systému STATISTICA

Vytvoříme datový soubor o dvou proměnných X, Y a 15 případech. Dvourozměrnou normalitu dat ověříme pomocí dvourozměrného tečkového diagramu – viz výše.
Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměn. – X, Y – OK – na záložce Možnosti vybereme Zobrazit detailní tabulku výsledků – Výpočet.

Prom. X & prom. Y	Korelace (smeny a vyrobky.sta) Označ. korelace jsou významné na hlad. p <,05000 (Celé případy vynechány u ChD) Průměr	Sm.Odch.	r(X,Y)	r2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
X	18,26667	2,37447									
X	18,26667	2,37447	1,000000	1,000000			15	0,000000	1,000000	0,000000	1,000000
X	18,26667	2,37447									
Y	83,60000	11,01817	0,927180	0,859663	8,923795	0,000001	15	5,010135	4,302365	1,562407	0,199812
Y	83,60000	11,01817									
X	18,26667	2,37447	0,927180	0,859663	8,923795	0,000001	15	1,562407	0,199812	5,010135	4,302365
Y	83,60000	11,01817									
Y	83,60000	11,01817	1,000000	1,000000			15	0,000000	1,000000	0,000000	1,000000

Výběrový koeficient korelace se realizoval hodnotou 0,92718, testová statistika nabyla hodnoty 8,924, odpovídající p-hodnota je 0,000001, tedy na hladině významnosti 0,01 zamítáme hypotézu o nezávislosti veličin X, Y.

Příklad: **Testování hypotézy o nezávislosti proti levostranné alternativě**

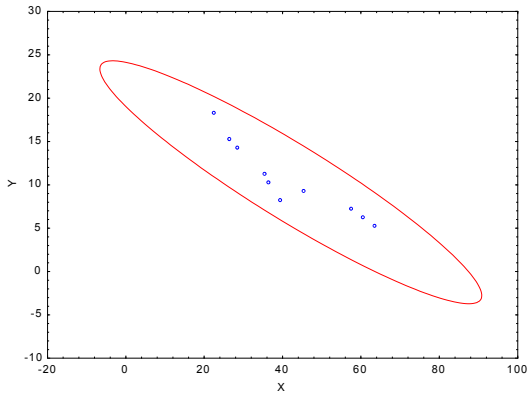
Pracovník personálního oddělení určité firmy zkoumá, zda existuje vztah mezi věkem zaměstnance (náhodná veličina X) a počtem dní absence za rok (náhodná veličina Y). Proto náhodně vybral údaje o 10 zaměstnancích:

X	27	61	37	23	46	58	29	36	64	40
Y	15	6	10	18	9	7	14	11	5	8

Na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny proti alternativě, že X, Y jsou záporně korelované náhodné veličiny.

Řešení:

Předpoklad o dvourozměrné normalitě dat ověříme orientačně pomocí dvourozměrného tečkového diagramu.



Vzhled diagramu svědčí o tom, že předpoklad je oprávněný.

Na hladině významnosti 0,05 testujeme H_0: $\rho = 0$ proti H_1: $\rho < 0$.

Vypočítáme $r_{12} = -0{,}9325$, tedy mezi věkem pracovníka a počtem dnů pracovní neschopnosti existuje silná nepřímá lineární závislost.

Realizace testové statistiky: $t_0 = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} = -,3053$,

kritický obor $W = (-\infty; u_{0,95}) = (-\infty; 1,8595)$.

Jelikož $t_0 \in W$, zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti veličin X a Y ve prospěch levostranné alternativy. S rizikem omylu nejvýše 5% jsme prokázali, že mezi věkem pracovníka a počtem dnů absence za rok existuje nepřímá lineární závislost.

Výpočet pomocí systému STATISTICA

Můžeme využít toho, že již známe r_{12}. Statistiky – Pravděpodobnostní kalkulátor – Korelace – vyplníme n = 10, r = -0,9325, odškrtneme Dvojité, zaškrtneme Výpočet p z r – Výpočet. V okénku p se objeví hodnota 0,000041, tedy na hladině významnosti 0,05 zamítáme hypotézu o nezávislosti veličin X a Y ve prospěch levostranné alternativy.

Příklad: Testování hypotézy o nezávislosti proti pravostranné alternativě

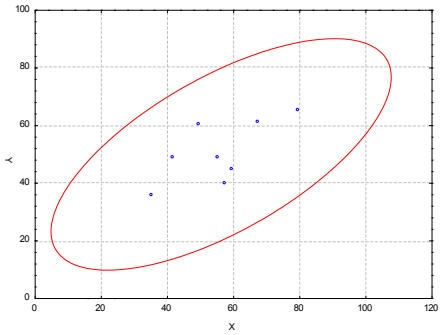
Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

Číslo studenta	1	2	3	4	5	6	7	8
Počet bodů v 1. testu	80	50	36	58	42	60	56	68
Počet bodů ve 2. testu	65	60	35	39	48	44	48	61

Na hladině významnosti 0,05 testujte hypotézu, že výsledky obou testů nejsou kladně korelované.

Řešení:

Nejprve se musíme přesvědčit, že uvedené výsledky lze považovat za realizace náhodného výběru z dvourozměrného normálního rozložení. Lze tak učinit orientačně pomocí dvourozměrného tečkového diagramu. Tečky by měly vytvořit elipsovitý obrazec.



Obrázek svědčí o tom, že předpoklad dvourozměrné normality je oprávněný a že mezi počty bodů z 1. a 2. testu bude existovat určitý stupeň přímé lineární závislosti.

Na hladině významnosti 0,05 testujeme H_0: $\rho = 0$ proti pravostranné alternativě H_1: $\rho > 0$.

Výpočtem zjistíme: $r_{12} = 0{,}6668$, $t_0 = 2{,}1917$. Stanovíme kritický obor: $W = \langle t_{0{,}95}, \infty) = \langle 1{,}9432; \infty)$. Jelikož $t_0 \in W$, zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti veličin X a Y ve prospěch pravostranné alternativy. S rizikem omylu nejvýše 5% jsme prokázali, že mezi výsledky 1. a 2. testu existuje přímá lineární závislost.

Výpočet pomocí systému STATISTICA

Můžeme využít toho, že již známe r_{12}. Statistiky – Pravděpodobnostní kalkulátor – Korelace – vyplníme n = 8, r = 0,6668, odškrtneme Dvojité, zaškrtneme Výpočet p z r – Výpočet. V okénku p se objeví hodnota 0,035455, tedy na hladině významnosti 0,05 zamítáme hypotézu o nezávislosti veličin X a Y ve prospěch pravostranné alternativy.

Postup při nesplnění předpokladu dvourozměrné normality

Máme k dispozici realizace náhodného výběru rozsahu 12 z dvourozměrného rozložení:

X	1	3	4	5	6	8	10	11	13	14	16	17
Y	13	15	18	16	23	31	39	56	45	43	37	0

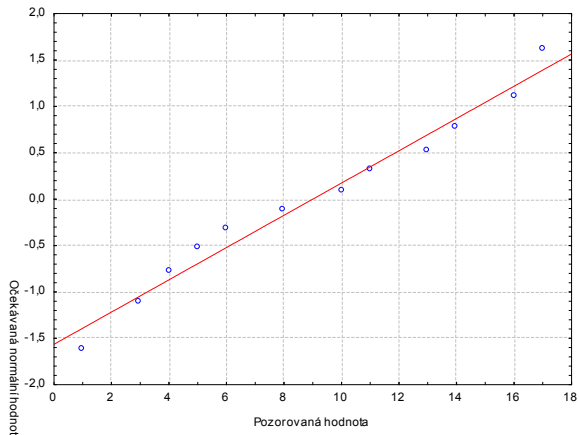
Na hladině významnosti 0,05 testujte hypotézu, že náhodné veličiny X, Y jsou nezávislé proti oboustranné alternativě.

Řešení:

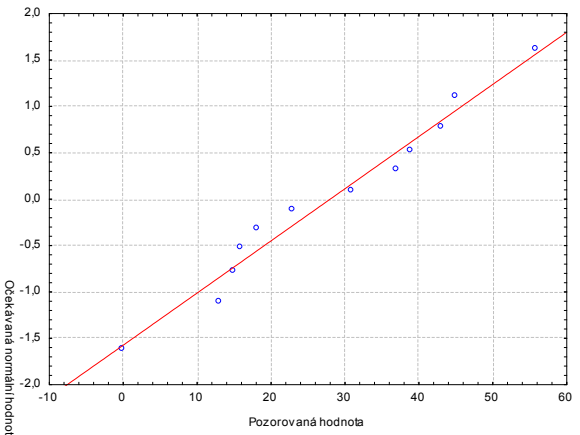
Na hladině významnosti 0,05 testujeme H_0: $\rho = 0$ proti oboustranné alternativě H_1: $\rho \neq 0$. Pokud neověříme předpoklad dvourozměrné normality, obvyklým způsobem vypočteme realizaci výběrového koeficientu korelace $r_{12} = 0{,}3729$ a realizaci testové statistiky $t_0 = 1{,}271$. Stanovíme kritický obor: $W = (-\infty, -t_{0{,}975}) \cup \langle t_{0{,}975}, \infty) = (-\infty, -2{,}2281\rangle \cup \langle t2{,}2281, \infty)$. Protože $t_0 \notin W$, nezamítáme na hladině významnosti 0,05 hypotézu o nezávislosti náhodných veličin X a Y.

Nyní budeme testovat hypotézu o normalitě náhodné veličiny X a náhodné veličiny Y. Grafické ověření pomocí N-P grafů:

N-P graf pro veličinu X



N-P graf pro veličinu Y

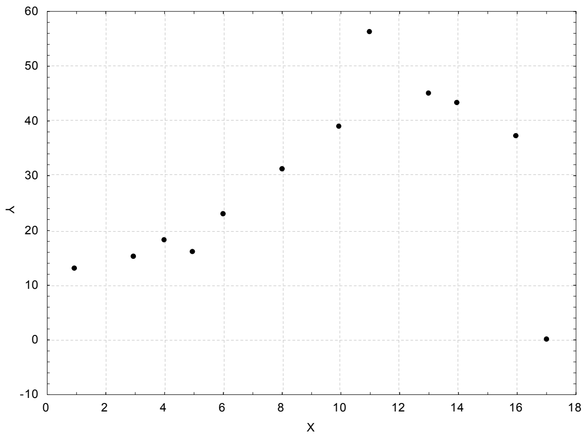


Vzhled grafů svědčí ve prospěch normality.

Testování pomocí Lilieforsovy varianty K - S testu a S – W testu:

Proměnná	Testy normality N	max D	Lilliefors p	W	p
X	12	0,130669	p > .20	0,956714	0,736098
Y	12	0,145742	p > .20	0,968954	0,899540

V obou případech hypotézu o normalitě nezamítáme na hladině významnosti 0,05.
Ověření dvourozměrné normality pomocí dvourozměrného tečkového diagramu:



Dvourozměrná normalita je silně porušena, tečky nevyplňují vnitřek elipsovitého obrazce. Přejdeme tedy k testování hypotézy o pořadové nezávislosti: Testujeme hypotézu
H_0: X, Y jsou pořadově nezávislé náhodné veličiny proti oboustranné alternativě
H_1: X, Y jsou pořadově závislé náhodné veličiny.

Vypočítáme Spearmanův koeficient pořadové korelace.

X	1	3	4	5	6	8	10	11	13	14	16	17
Y	13	15	18	16	23	31	39	56	45	43	37	0
R_i	1	2	3	4	5	6	7	8	9	10	11	12
Q_i	2	3	5	4	6	7	9	12	11	10	8	1

$$r_S = 1 - \frac{6}{12(12^2-1)}\left[\begin{array}{l}(1-2)^2+(2-3)^2+(3-5)^2+(4-4)^2+(5-6)^2+(6-7)^2+(7-9)^2+\\+(8-12)^2+(9-11)^2+(10-10)^2+(11-8)^2+(12-1)^2\end{array}\right]=$$

$$= 1 - \frac{6}{12\cdot 143}(1+1+4+0+1+1+4+16+4+0+9+121) = 1 - \frac{1}{286}\cdot 162 = 0,4336$$

Stanovíme kritický obor:

$$W = \langle -1, -r_{S,1-\alpha}(n)\rangle \cup \langle r_{S,1-\alpha}(n), 1\rangle = \langle -1, -r_{S,0,95}(12)\rangle \cup \langle r_{S,0,95}(12), 1\rangle = \langle -1, -0,5804\rangle \cup \langle 0,5804, 1\rangle$$

Testová statistika se nerealizuje v kritickém oboru, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Porovnání koeficientu korelace s danou konstantou

Nechť c je reálná konstanta. Testujeme H_0: $\rho = c$ proti H_1: $\rho \neq c$. (Tento test se provádí např. tehdy, když experimentátor porovnává vlastnosti svých dat s vlastnostmi uváděnými v literatuře.) Test je založen na statistice

$U = \left(z - \frac{1}{2} \ln \frac{1+}{1-} - \frac{c}{2(n-} \right) \sqrt{n-}$, která má za platnosti H_0 pro $n \geq 10$ asymptoticky rozložení N(0,1), přičemž $Z = \frac{1}{2} \ln \frac{1+ r_{12}}{1- r_{12}}$ je tzv. Fisherova Z-transformace.

Kritický obor pro test H_0 proti oboustranné alternativě tedy je $W = (-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$. H_0 zamítáme na asymptotické hladině významnosti α, když $U \in W$.

Příklad: U 600 vzorků rudy byl stanoven obsah železa dvěma analytickými metodami s výběrovým koeficientem korelace 0,85. V literatuře se uvádí, že koeficient korelace těchto dvou metod má být 0,9. Na asymptotické hladině významnosti 0,05 testujte hypotézu
H_0: $\rho = 0{,}9$ proti H_1: $\rho \neq 0{,}9$.

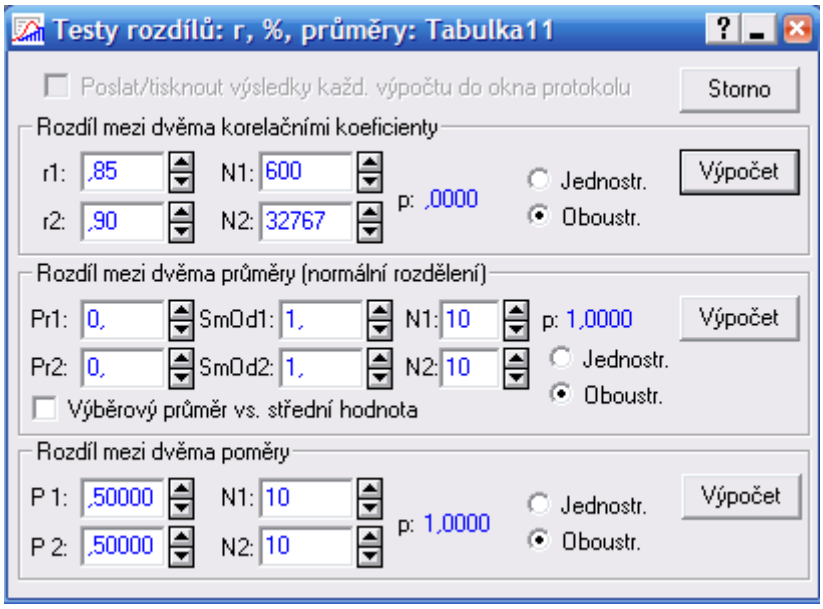
Řešení: $Z = \frac{1}{2} \ln \frac{1+ 0{,}85}{1- 0{,}85} = 1{,}2562$,

$U = \left(1{,}2562 - \frac{1}{2} \ln \frac{1+ 0{,}9}{1- 0{,}9} - \frac{0{,}9}{2(600-} \right) \sqrt{600-} = -2{,}2976$, $u_{0{,}975} = 1{,}96$,

$W = (-\infty, -1{,}96) \cup (1{,}96, \infty)$. Protože $U \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA (pouze přibližný):

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma korelačními koeficienty. Do políčka r1 napíšeme 0,85, do políčka N1 napíšeme 600, do políčka r2 napíšeme 0,9, do políčka N2 napíšeme 32767 (větší hodnotu systém neumožní) - Výpočet. Dostaneme p-hodnotu 0,0000, tedy zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.



Upozornění: Pokud bychom chtěli pomocí systému STATISTICA provést přesnější test s využitím statistiky U, můžeme vypočítat Fisherovu Z- transformaci pomocí Pravděpodobnostního kalkulátoru – Korelace, kde zadáme realizaci výběrového koeficientu korelace, rozsah výběru. Zajímá nás Fisher z.

Porovnání dvou korelačních koeficientů

Nechť jsou dány dva nezávislé náhodné výběry o rozsazích n a n^* z dvourozměrných normálních rozložení s korelačními koeficienty ρ a ρ^*. Testujeme H_0: $\rho = \rho^*$ proti H_1: $\rho \neq \rho^*$. Označme R_{12} výběrový korelační koeficient 1. výběru a R_{12}^* výběrový korelační koeficient 2. výběru. Položme $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ a $Z^* = \frac{1}{2}\ln\frac{1+R_{12}^*}{1-R_{12}^*}$. Platí-li H_0, pak testová statistika $U = \frac{Z - Z^*}{\sqrt{\frac{1}{n-3} + \frac{1}{n^*-3}}}$ má asymptoticky rozložení N(0,1). Kritický obor pro test H_0 proti oboustranné alternativě tedy je $W = (-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$. H_0 zamítáme na asymptotické hladině významnosti α, když $U \in W$.

Příklad: Lékařský výzkum se zabýval sledováním koncentrací látek A a B v moči pacientů trpících určitou ledvinovou chorobou. U 100 zdravých jedinců činil výběrový korelační koeficient mezi koncentracemi obou látek 0,65 a u 142 osob trpících zmíněnou chorobou byl 0,37. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že korelační koeficienty v obou skupinách se neliší.

Řešení: $Z = \frac{1}{2}\ln\frac{1+0,65}{1-0,65} = 0,7753$, $Z^* = \frac{1}{2}\ln\frac{1+0,37}{1-0,37} = 0,3884$,

$U = \frac{0,7753 - 0,3884}{\sqrt{\frac{1}{100-3} + \frac{1}{142-3}}} = 3,9242$, $u_{0,975} = 1,96$, $W = (-\infty, -1,96) \cup (1,96, \infty)$. Protože $U \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma korelačními koeficienty. Do políčka r1 napíšeme 0,65, do políčka N1 napíšeme 100, do políčka r2 napíšeme 0,37, do políčka N2 napíšeme 142 - Výpočet. Dostaneme p-hodnotu 0,0038, tedy zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.

Interval spolehlivosti pro korelační koeficient

Jestliže dvourozměrný náhodný výběr rozsahu n pochází z dvourozměrného normálního rozložení, jehož korelační koeficient se příliš neliší od nuly (je splněna podmínka $|\rho| < 0{,}5$) a rozsah výběru je dostatečně velký ($n \geq 100$), lze odvodit, že 100(1-α)% interval spolehlivosti pro ρ má meze $R_{12} \pm u_{1-\alpha/2}\frac{1-R_{12}^2}{\sqrt{n}}$.

Nejsou-li uvedené podmínky splněny, pak nelze tento vzorec použít, protože rozložení výběrového korelačního koeficientu je příliš zešikmené. V takovém případě využijeme toho, že náhodná veličina $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ má i při malém rozsahu výběru přibližně normální rozložení se střední hodnotou $E(Z) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$ (2. sčítanec lze při větším n zanedbat) a rozptylem $D(Z) = \frac{1}{n-3}$. Standardizací veličiny Z dostaneme veličinu $U = \frac{Z-E(Z)}{\sqrt{D(Z)}}$, která má asymptoticky rozložení N(0,1). Tudíž 100(1-α)% asymptotický interval spolehlivosti pro $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ bude mít meze $Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}}$. Interval spolehlivosti pro ρ pak dostaneme zpětnou transformací.

Poznámka: Jelikož Z = arctgh R_{12}, dostáváme R_{12} = tgh Z a meze intervalu spolehlivosti pro ρ můžeme psát ve tvaru $\mathrm{tgh}\left(Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right)$, přičemž $\mathrm{tgh}\, x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

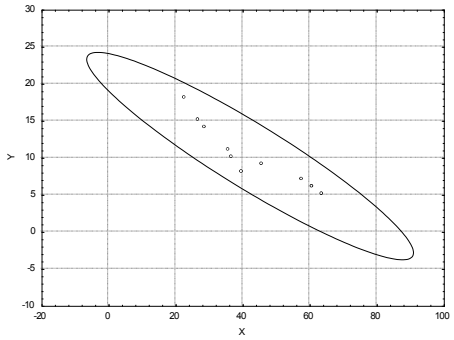
Příklad: Pracovník personálního oddělení určité firmy zkoumá, zda existuje vztah mezi počtem dní absence za rok (veličina Y) a věkem pracovníka (veličina X). Proto náhodně vybral údaje o 10 pracovnících.

Č.prac.	1	2	3	4	5	6	7	8	9	10
X	27	61	37	23	46	58	29	36	64	40
Y	15	6	10	18	9	7	14	11	5	8

Za předpokladu, že uvedené údaje tvoří číselné realizace náhodného výběru rozsahu 10 z dvourozměrného normálního rozložení, vypočtěte výběrový korelační

koeficient a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny. Sestrojte 95% asymptotický interval spolehlivosti pro skutečný korelační koeficient ρ.

Řešení: Předpoklad o dvourozměrné normalitě dat ověříme orientačně pomocí dvourozměrného tečkového diagramu.



Vzhled diagramu svědčí o tom, že předpoklad je oprávněný.
Testujeme H_0: ρ = 0 proti H_1: ρ ≠ 0. Vypočítáme R_{12} = -0,9325, tedy mezi věkem pracovníka a počtem dnů pracovní neschopnosti existuje silná nepřímá lineární závislost. Testová statistika: T = -7,3053, kvantil $t_{0,975}(8)$ = 2,306, kritický obor $W = (-\infty, -2,306) \cup (2,306, \infty)$. Jelikož $T \in W$, zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti veličin X a Y.

Vypočítáme $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}} = \frac{1}{2}\ln\frac{1-0,9325}{1+0,9325} = -1,6772$. Meze 95% asymptotického intervalu spolehlivosti pro ρ jsou $\tanh\left(-1,6772 \pm \frac{1,96}{\sqrt{7}}\right)$, tedy -0,9842 < ρ < -0,7336 s pravděpodobností přibližně 0,95.

Výpočet pomocí systému STATISTICA:

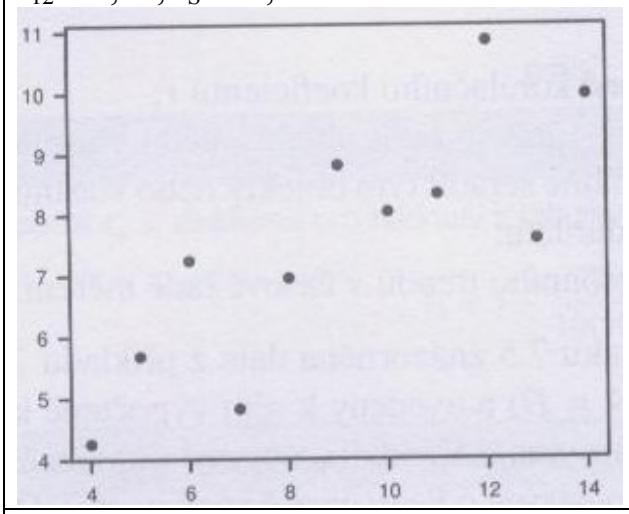
Ve STATISTICE vypočteme meze 100(1-α)% asymptotického intervalu spolehlivosti pro koeficient korelace ρ tak, že otevřeme nový datový soubor se dvěma proměnnými (pojmenujeme je DM a HM) a jedním případem.
Do Dlouhého jména proměnné DM zapíšeme příkaz
= TanH(0,5*log((1-0,9325)/(1+0,9325))-VNormal(0,975;0;1)/sqrt(7))
a do Dlouhého jména proměnné HM zapíšeme příkaz
= TanH(0,5*log((1-0,9325)/(1+0,9325))+VNormal(0,975;0;1)/sqrt(7))

	1 DM	2 HM
1	-0,98425	-0,73358

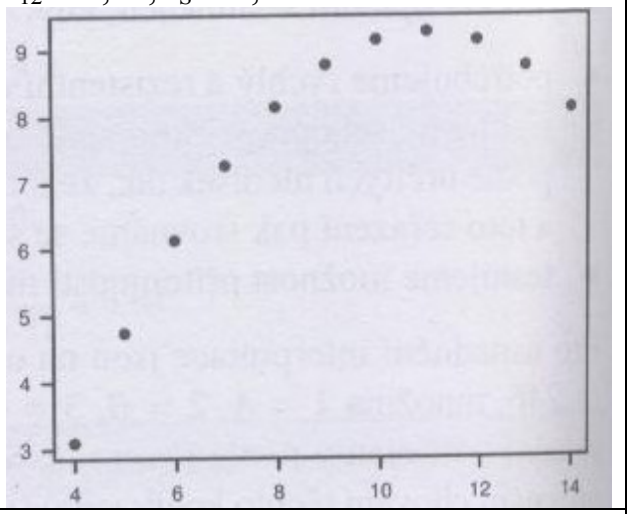
95% asymptotický interval spolehlivosti pro koeficient korelace ρ má tedy meze –0,98425 a -0,73358. (Protože nepokrývá hodnotu 0, zamítáme hypotézu o nezávislosti veličin X, Y na asymptotické hladině významnosti 0,05.)

Vztah mezi koeficienty korelace r_{12} a r_S

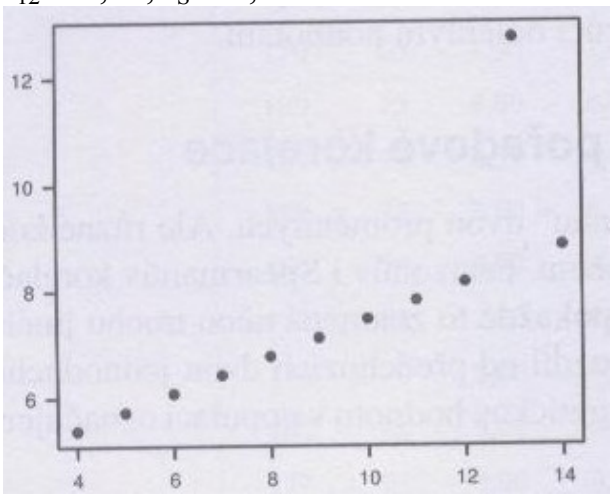
$r_{12} = 0{,}82$, $r_S = 0{,}82$



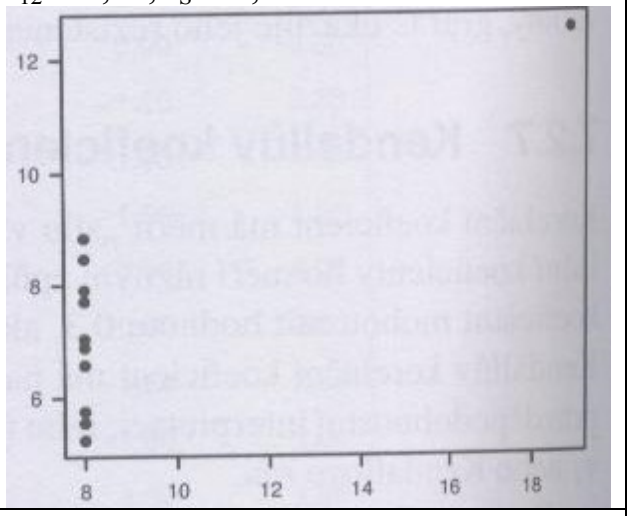
$r_{12} = 0{,}82$, $r_S = 0{,}69$



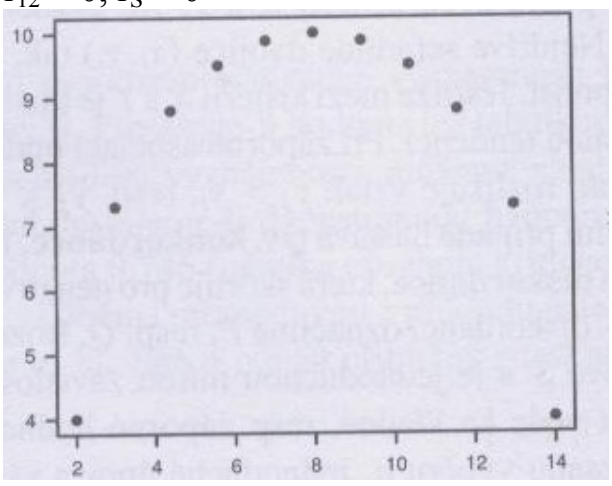
$r_{12} = 0{,}82$, $r_S = 0{,}99$



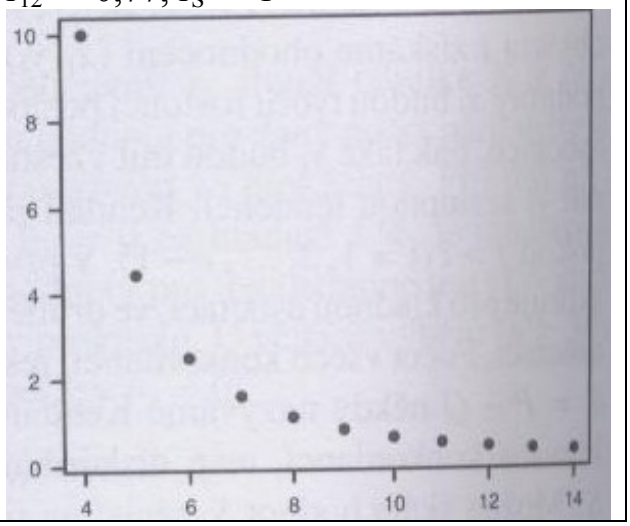
$r_{12} = 0{,}82$, $r_S = 0{,}5$



$r_{12} = 0$, $r_S = 0$



$r_{12} = -0{,}77$, $r_S = -1$



3. obrázek ukazuje odolnost Spearmanova koeficientu vůči odlehlým hodnotám.

6. obrázek dokumentuje schopnost Spearmanova koeficientu měřit monotónní vztahy.
Jestliže náhodný vektor (X, Y) má dvourozměrné normální rozložení s Pearsonovým koeficientem korelace ρ a Spearmanovým koeficientem korelace ρ_S, pak $\rho \approx 2 \sin(0{,}523\rho_S)$.