Osnova přednášky Korelační analýza

1. Jednoduchá korelace

1.1. Pearsonův koeficient korelace a jeho vlastnosti
1.2. Výběrový koeficient korelace
1.3. Test hypotézy o nezávislosti
1.4. Příklad
1.5. Interval spolehlivosti pro koeficient korelace
1.6. Příklad
1.7. Porovnání koeficientu korelace s danou konstantou
1.8. Porovnání dvou koeficientů korelace

2. Vícenásobná korelace

2.1. Varianční, korelační a kovarianční matice
2.2. Odhady pro jeden náhodný vektor
2.3. Příklad
2.4. Odhady pro dva náhodné vektory
2.5. Příklad
2.6. Koeficient vícenásobné korelace a jeho vlastnosti
2.7. Výběrový koeficient vícenásobné korelace
2.8. Test hypotézy o nevýznamnosti koeficientu vícenásobné korelace
2.9. Příklad

3. Parciální korelace

3.1. Koeficient parciální korelace
3.2. Výběrový koeficient parciální korelace
3.3. Test hypotézy o nevýznamnosti koeficientu parciální korelace
3.4. Příklad

1. Jednoduchá korelace

1.1. Pearsonův koeficient korelace

Definice: Nechť X, Y jsou náhodné veličiny se středními hodnotami E(X), E(Y) a rozptyly D(X), D(Y).
Číslo

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

se nazývá Pearsonův koeficient korelace.

Vlastnosti Pearsonova koeficientu korelace

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

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

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

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

e) $|R(X,Y)| \leq 1$ 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.

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. Jsou-li náhodné veličiny X,Y stochasticky nezávislé, pak jsou samozřejmě i nekorelované.)
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.)

Pearsonův koeficient korelace dvourozměrného normálního rozložení

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

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

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:

$$\varphi_1(x)=\int_{-\infty}^{\infty}\varphi(x,y)dy=\ldots=\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}},$$

$$\varphi_2(y)=\int_{-\infty}^{\infty}\varphi(x,y)dx=\ldots=\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}.$$

Je-li $\rho = 0$, pak pro $\forall(x,y)\in R^2 : \varphi(x,y)=\varphi_1(x)\varphi_2(y)$, 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!

1.2. 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. (Výběrový koeficient korelace není nestranným odhadem skutečného koeficientu korelace, je odhadem vychýleným. Vychýlení je zanedbatelně malé pro rozsahy výběrů nad 30.)

Upozornění: nadále budeme předpokládat, že $(X_1, Y_1), ..., (X_n, Y_n)$ 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.

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.

1.3. 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: $T_0 = \dfrac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^{\ 2}}}$.

Platí-li nulová hypotéza, pak $T_0 \sim t(n-2)$.

Kritický obor pro test H_0 proti

- oboustranné alternativě: $W = \left(-\infty, -t_{1-\alpha/2}(n-2)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-2), \infty\right)$,
- levostranné alternativě: $W = \left(-\infty, -t_{1-\alpha}(n-2)\right\rangle$,
- pravostranné alternativě: $W = \left\langle t_{1-\alpha}(n-2), \infty\right)$.

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

1.4. Příklad

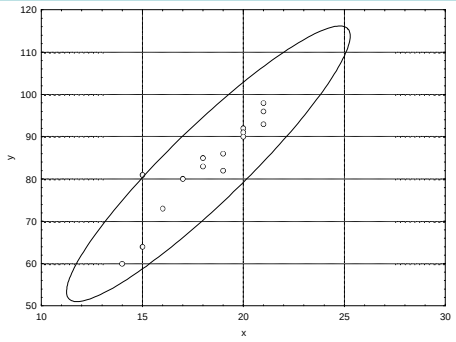
V dílně pracuje 15 dělníků. Byl u nich zjištěn počet směn odpracovaných za měsíc (náhodná veličina X) a počet zhotovených výrobků (náhodná veličina Y):

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.

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 $W = (-\infty, -t_{0,995}(13)\rangle \cup \langle t_{0,995}(13), \infty) = (-\infty, -3{,}012\rangle \cup \langle 3{,}012, \infty).$

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.

Počítačový výstup

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									
Y	83,60000	11,01817	0,927180	0,859663	8,923795	0,000001	15	5,010135	4,302365	1,562407	0,199812

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.

1.5. Interval spolehlivosti pro koeficient korelace

Náhodná veličina $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ (tzv. Fisherova Z-transformace koeficientu korelace) má 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.

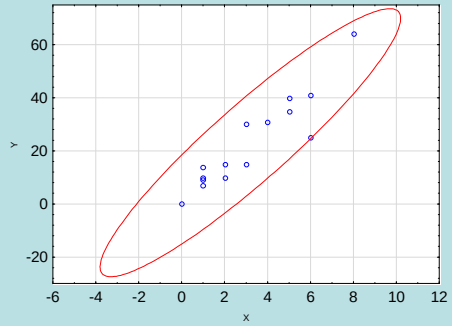
Poznámka: Jelikož Z = arctgh R_{12}, dostáváme R_{12} = tgh Z a meze intervalu spolehlivosti pro ρ můžeme psát ve tvaru $\mathrm{tgh}\left(Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right)$, přičemž $\mathrm{tgh}\, x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

1.6. Příklad: Učitel tělocviku zjišťoval, zda existuje vztah mezi počtem shybů (veličina X) a počtem kliků (veličina Y) u 15 náhodně vybraných chlapců:

Číslo chlapce	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Počet shybů	1	3	2	0	5	6	1	4	3	5	6	2	1	1	8
Počet kliků	10	15	15	0	40	25	7	31	30	35	41	10	14	9	64

Za předpokladu, že uvedené údaje tvoří číselné realizace náhodného výběru rozsahu 15 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} = 0,9276, tedy mezi počtem shybů a počtem kliků existuje silná přímá lineární závislost. Testová statistika: T = 8,9511, kvantil $t_{0,975}(13)$ = 2,1604, kritický obor $W = \langle -\infty, -2{,}1604 \rangle \cup \langle 2{,}1604, \infty \rangle$. Jelikož $T \in W$, zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti veličin X a Y.

Vypočítáme $Z = \frac{1}{2} \ln \frac{1+R_{12}}{1-R_{12}} = \frac{1}{2} \ln \frac{1+0{,}9276}{1-0{,}9276} = 1{,}6409$. Meze 95% asymptotického intervalu spolehlivosti pro ρ jsou

$\operatorname{tgh}\left(1{,}6409 \pm \frac{1{,}96}{\sqrt{12}}\right)$, tedy 0,7914 < ρ < 0,9761 s pravděpodobností přibližně 0,95.

Počítačový výstup

	Odhad intervalu Jedna korelace, t-test
	Hodnota
Pozorovaný korel. koef. R	0,9276
Korelace dle nulové hypotézy (Ró0)	0,0000
Oboustranná p-hodnota	0,0000
Velikost vz. ve skup. (N)	15,0000
Interval spolehlivosti	0,9500
Meze spolehlivosti (Fisher. Z původní):	
Ró:	
Dolní mez	0,7914
Horní mez	0,9761

95% asymptotický interval spolehlivosti pro koeficient korelace ρ má tedy meze 0,7914 a 0,9761. (Protože nepokrývá hodnotu 0, zamítáme hypotézu o nezávislosti veličin X, Y na asymptotické hladině významnosti 0,05.)

1.7. 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 $U = \left(Z - \frac{1}{2} \ln \frac{1+c}{1-c} - \frac{c}{2(n-1)} \right) \sqrt{n-3}$, 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 $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$. H_0 zamítáme na asymptotické hladině významnosti α, když $U \in W$.

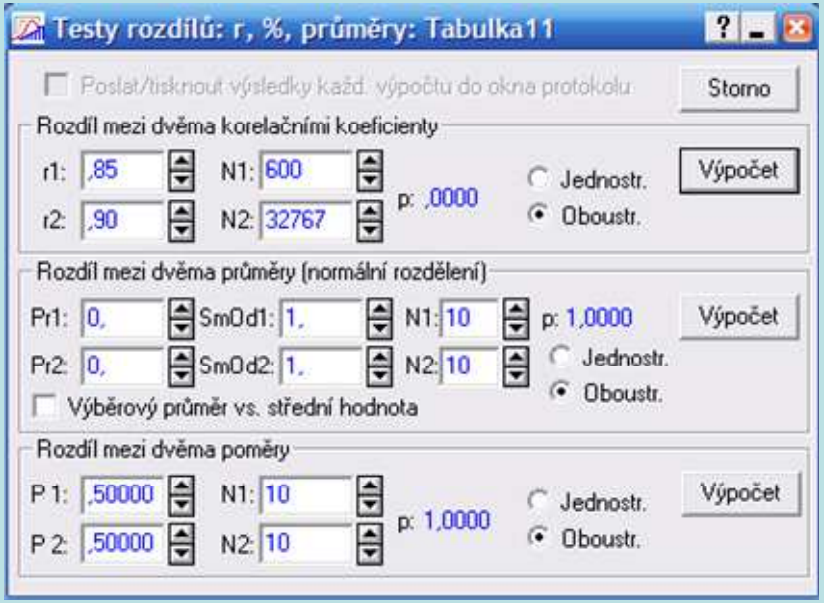
Příklad: U 600 vzorků rudy byl stanoven obsah železa dvěma analytickými metodami s výběrovým koeficientem korelace 0,85. V literatuře se uvádí, že koeficient korelace těchto dvou metod má být 0,9. Na asymptotické hladině významnosti 0,05 testujte hypotézu

H_0: $\rho = 0{,}9$ proti H_1: $\rho \neq 0{,}9$.

Řešení: $Z = \frac{1}{2}\ln\frac{1+0{,}85}{1-0{,}85} = 1{,}2562$, $U = \left(1{,}2562 - \frac{1}{2}\ln\frac{1+0{,}9}{1-0{,}9} - \frac{0{,}9}{2(600-1)}\right)\sqrt{600-3} = -5{,}2976$,

$u_{0{,}975} = 1{,}96$, $W = (-\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty)$. Protože $U \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

Počítačový výstup



1.8. Porovnání dvou koeficientů korelace

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 $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$.

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

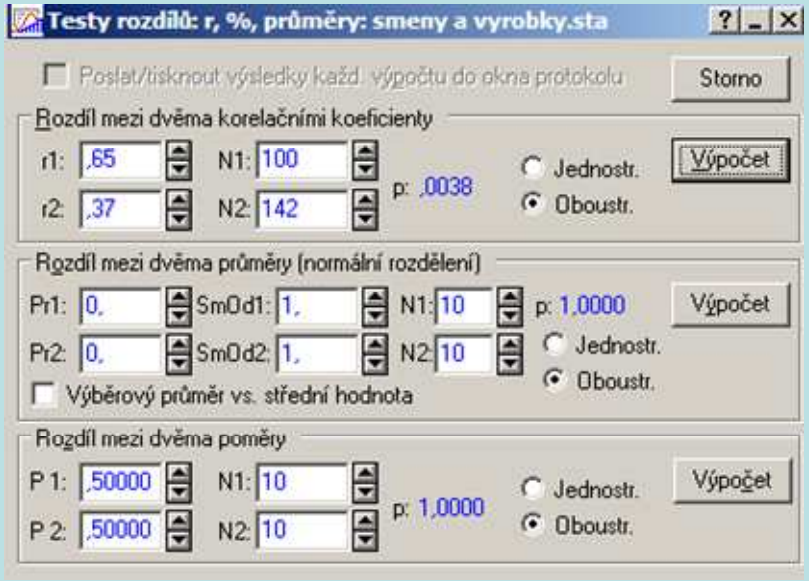
Příklad: Lékařský výzkum se zabýval sledováním koncentrací látek A a B v moči pacientů trpících určitou ledvinovou chorobou. U 100 zdravých jedinců činil výběrový korelační koeficient mezi koncentracemi obou látek 0,65 a u 142 osob trpících zmíněnou chorobou byl 0,37. Na asymptotické hladině významnosti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

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

$U = \frac{0{,}7753 - 0{,}3884}{\sqrt{\frac{1}{100-3} + \frac{1}{142-3}}} = 2{,}9242$, $u_{0{,}975} = 1{,}96$, $W = (-\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty)$.

Protože $U \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

Počítačový výstup



2. Vícenásobná korelace

2.1. Varianční, korelační a kovarianční matice

Nechť $\mathbf{X} = (X_1, \ldots, X_p)'$ je náhodný vektor. Označme

$\mu_i = E(X_i)$ střední hodnotu náhodné veličiny X_i,

$\sigma_i^2 = D(X_i)$ rozptyl náhodné veličiny X_i,

$\sigma_{ij} = C(X_i, X_j)$ kovarianci náhodných veličin X_i, X_j (přitom $\sigma_{ii} = \sigma_i^2$)

$\rho_{ij} = R(X_i, X_j)$ koeficient korelace náhodných veličin X_i, X_j

Vektor $E(\mathbf{X}) = (\mu_1, \ldots, \mu_p)'$ se nazývá vektor středních hodnot náhodného vektoru **X**.

Čtvercová matice řádu p $var(\mathbf{X}) = (\sigma_{ij})_{i,j=1, \ldots, p}$ se nazývá varianční matice náhodného vektoru **X**.

Čtvercová matice řádu p $cor(\mathbf{X}) = (\rho_{ij})_{i,j=1, \ldots, p}$ se nazývá korelační matice náhodného vektoru **X**.

Je zřejmé, že varianční matice a korelační matice jsou symetrické.

Nechť $\mathbf{X} = (X_1, \ldots, X_p)'$ a $\mathbf{Y} = (Y_1, \ldots, Y_q)'$ jsou náhodné vektory.

Matice typu pxq $cov(\mathbf{X},\mathbf{Y}) = (C(X_i, Y_j))$ se nazývá kovarianční matice vektorů **X**, **Y**.

Matice typu pxq $cor(\mathbf{X},\mathbf{Y}) = (\rho(X_i, Y_j))$ se nazývá korelační matice vektorů **X**,**Y**.

2.2. Odhady pro jeden náhodný vektor

Nechť $\mathbf{X}$ je náhodný vektor, který má p-rozměrné rozložení s vektorem středních hodnot $\boldsymbol{\mu}$, varianční maticí var($\mathbf{X}$) a korelační maticí cor($\mathbf{X}$). Nechť je dán náhodný výběr $\mathbf{X}_1 = (X_{11}, \ldots, X_{1p})', \ldots, \mathbf{X}_n = (X_{n1}, \ldots, X_{np})'$ rozsahu n z tohoto rozložení.

Nestranný odhad vektoru $\boldsymbol{\mu}$ je vektor výběrových průměrů $\mathbf{M} = (M_1, \ldots, M_p)'$, kde $M_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ je výběrový průměr j-tého výběru, j = 1, …, p.

Nestranný odhad matice var($\mathbf{X}$) je výběrová varianční matice $\mathbf{S} = (S_{ij}) = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{M})(\mathbf{X}_i - \mathbf{M})'$ řádu p.

Vychýlený odhad matice cor($\mathbf{X}$) je výběrová korelační matice $\mathbf{R} = (R_{ij})$, kde R_{ij} je výběrový korelační koeficient i-té a j-té složky vektoru $\mathbf{X}$, tedy

$R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}}$, i, j = 1, …, p. (Je zřejmé, že diagonální prvky matice $\mathbf{R}$ jsou jedničky a matice $\mathbf{R}$ je symetrická.)

2.3. Příklad: U 28 náhodně vybraných osob byly zjišťovány tyto údaje:

Sex … 1 – muž, 2 – žena (mužů i žen bylo po 14)

výška (v cm), hmotnost (v kg), boty (číslo bot).

Vypočtěte realizaci výběrové varianční matice a výběrové korelační matice. (Soubor udaje_o_lidech_1.sta)

Řešení:

Výběrová varianční matice

Proměnná	vyska	hmotnost	boty
vyska	112,8611	161,0926	41,45370
hmotnost	161,0926	248,4709	61,99206
boty	41,4537	61,9921	16,40608

Výběrová korelační matice

Proměnná	vyska	hmotnost	boty
vyska	1,000000	0,961979	0,963360
hmotnost	0,961979	1,000000	0,970948
boty	0,963360	0,970948	1,000000

Z výběrové varianční matice plyne, že největší variabilitu má hmotnost, pak výška a nakonec číslo bot.

Z výběrové korelační matice plyne, že mezi všemi třemi dvojicemi proměnných existuje velmi silná přímá lineární závislost, nejsilnější je mezi hmotností a velikostí bot.

2.4. Odhady pro dva náhodné vektory

Nechť náhodný vektor **X** má p-rozměrné rozložení a nechť $\mathbf{X}_1, \ldots, \mathbf{X}_n$ je náhodný výběr z tohoto rozložení. Nechť náhodný vektor **Y** má q-rozměrné rozložení a nechť $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ je náhodný výběr z tohoto rozložení. Předpokládejme, že obě rozložení mají konečné druhé momenty. Nechť cov(**X**, **Y**) je kovarianční matice těchto vektorů a cor(**X**, **Y**) je korelační matice těchto vektorů.

Označme $M_{Xj} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, j = 1,\ldots,p, M_{Yj} = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}, j = 1,\ldots,q,$

$\mathbf{M_X} = (M_{X1}, \ldots, M_{Xp})'$, $\mathbf{M_Y} = (M_{Y1}, \ldots, M_{Yq})'$.

Nestranným odhadem kovarianční matice cov(**X**, **Y**) vektorů **X**, **Y** je výběrová kovarianční matice vektorů **X**, **Y** definovaná vzorcem $\mathbf{S_{XY}} = (S_{ij}) = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{M_X})(\mathbf{Y}_i - \mathbf{M_Y})'$, i = 1, …, p, j = 1, …, q.

Vychýleným odhadem korelační matice cor(**X**, **Y**) vektorů **X**, **Y** je výběrová korelační matice vektorů **X**, **Y** definovaná vzorcem $\mathbf{R_{XY}} = (R_{ij})$, kde R_{ij} je výběrový korelační koeficient i-té a j-té složky vektorů **X**, **Y,** i = 1, …, p, j = 1, …, q.

2.5.Příklad: Nechť vektor $\mathbf{X} = (X_1, X_2, X_3)'$ obsahuje údaje o výšce, hmotnosti a číslu bot mužů, vektor $\mathbf{Y} = (Y_1, Y_2)'$ obsahuje údaje výšce a hmotnosti žen. Vypočtěte realizace výběrové kovarianční a výběrové korelační matice vektorů **X**, **Y**. (Soubor udaje_o_lidech_2.sta)

Řešení:

Výběrová kovarianční matice

Efekt	Sloup.4 Vyska_z	Sloup.5 Hmotnost_z
Vyska_m	10,81319	17,39560
Hmotnost_m	15,70879	15,22527
Boty_m	4,43407	5,13736

Výběrová korelační matice

Efekt	Sloup.4 Vyska_z	Sloup.5 Hmotnost_z
Vyska_m	0,467318	0,767160
Hmotnost_m	0,514047	0,508409
Boty_m	0,560289	0,662427

2.6. Koeficient vícenásobné korelace

Intenzitu lineární závislosti mezi náhodnou veličinou Y a náhodným vektorem $\mathbf{X} = (X_1, \ldots, X_p)'$ měříme pomocí koeficientu vícenásobné korelace $\rho_{Y.\mathbf{X}}$. Jeho druhá mocnina je dána vzorcem

$$\rho_{Y.\mathbf{X}}^2 = \mathrm{cor}(Y, \mathbf{X})\, \mathrm{cor}(\mathbf{X})^{-1}\, \mathrm{cor}(\mathbf{X}, Y).$$

Má tyto vlastnosti:

a) $\rho_{Y.\mathbf{X}} \geq 0$

b) $\rho_{Y.\mathbf{X}} \geq \left|\rho(Y, X_i)\right|$ pro $\forall i = 1, \ldots, p$

c) $\rho_{Y.X_1 \ldots X_p} \geq \ldots \geq \rho_{Y.X_1X_2} \geq \rho(Y, X_1)$

d) $\rho_{Y.\mathbf{X}} = 1 \Leftrightarrow$ existují konstanty $\beta_0, \beta_1, \ldots, \beta_p$ tak, že $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

2.7. Výběrový koeficient vícenásobné korelace

Nechť náhodný vektor $(Y, X_1, \ldots, X_p)'$ má (p+1)-rozměrné rozložení s koeficientem mnohonásobné korelace $\rho_{Y.\mathbf{X}}$.

Nechť je dán náhodný výběr $(Y_1, X_{11}, \ldots, X_{1p})', \ldots, (Y_n, X_{n1}, \ldots, X_{np})'$ rozsahu n z tohoto rozložení. Pak jako odhad $\rho_{Y.\mathbf{X}}$ slouží výběrový koeficient vícenásobné korelace $r_{Y.\mathbf{X}}$, jehož druhá mocnina je dána vzorcem

$$r_{Y.\mathbf{X}}^2 = R_{Y\mathbf{X}}\, \mathbf{R}^{-1}\, R_{\mathbf{X}Y},$$

kde $R_{Y\mathbf{X}}$ je výběrová korelační matice veličiny Y a vektoru **X** (v tomto případě se redukuje na vektor $(r_{YX_1}, \ldots, r_{YX_p})$) a **R** je výběrová korelační matice vektoru **X**.

Vlastnosti koeficientu vícenásobné korelace se přenášejí i na výběrový koeficient vícenásobné korelace.

Příklad: Při zkoumání závislosti hodinové výkonnosti dělníka (veličina Y – v kusech) na jeho věku (veličina X_1 – v letech) a době zapracovanosti (veličina X_2 – v letech) byly u 10 náhodně vybraných dělníků zjištěny tyto údaje:

Y	67	65	75	66	77	84	69	60	70	66
X_1	43	40	49	46	41	41	48	34	32	42
X_2	6	8	14	14	8	12	16	1	5	7

Vypočtěte výběrový koeficient vícenásobné korelace $r_{Y,(X_1,X_2)}$ popisující závislost hodinové výkonnosti dělníka na jeho věku a době zapracovanosti.

Řešení:

Koeficient $r_{Y,(X_1,X_2)}$ najdeme v záhlaví výstupní tabulky pod označením R = 0,54.

N=10	Výsledky regrese se závislou proměnnou : Y (vykony delniku.sta) R= ,54005243 R2= ,29165662 Upravené R2= ,08927280 F(2,7)=1,4411 p<,29913 Směrod. chyba odhadu : 6,6491					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(7)	p-hodn.
Abs.člen			86,74217	25,32397	3,425299	0,011056
X1	-0,550937	0,598452	-0,70031	0,76071	-0,920604	0,387883
X2	0,920415	0,598452	1,35062	0,87817	1,537994	0,167937

Jeho druhá mocnina (ozn. R2) nám říká, že variabilita výkonů dělníků je z 29 % vysvětlena jejich věkem a dobou zapracovanosti.

2.8. Test hypotézy o nevýznamnosti koeficientu vícenásobné korelace

Nechť náhodný výběr $(Y_1, X_{11}, \ldots, X_{1p})', \ldots, (Y_n, X_{n1}, \ldots, X_{np})'$ pochází z (p+1)-rozměrného normálního rozložení, které má koeficient vícenásobné korelace $\rho_{Y.\mathbf{X}}$. Musí platit $n > p+1$. Testujeme hypotézu $H_0: \rho_{Y.\mathbf{X}} = 0$ proti $H_1: \rho_{Y.\mathbf{X}} \neq 0$. Vzhledem k tomu, že se jedná o výběr z (p+1)-rozměrného normálního rozložení, testujeme, zda existuje závislost mezi veličinou Y a vektorem **X**. (Je-li $\rho_{Y.\mathbf{X}} = 0$, pak z vlastnosti (b) plyne, že $\rho(Y,X_i) = 0$ pro všechna $i = 1, \ldots, p$, tudíž náhodné veličiny Y a X_i jsou stochasticky nezávislé pro všechna $i = 1, \ldots, p$.)

Testová statistika $F = \frac{n-p-1}{p} \cdot \frac{r_{Y.x}^{\ 2}}{1-r_{Y.x}^{\ 2}}$ se řídí rozložením F(p, n-p-1), pokud H_0 platí. Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty)$. Jestliže $F \in W$, H_0 zamítáme na hladině významnosti α.

2.9. Příklad

Předpokládáme, že údaje o výkonnosti 10 náhodně vybraných dělníků, jejich věku a době zapracovanosti představují číselné realizace náhodného výběru rozsahu 10 ze třírozměrného normálního rozložení. Na hladině významnosti 0,05 testujte hypotézu, že výkon dělníka nezávisí na jeho věku a době zapracovanosti.

Řešení:

N=10	Výsledky regrese se závislou proměnnou : Y (vykony delniku.sta) R= ,54005243 R2= ,29165662 Upravené R2= ,08927280 F(2,7)=1,4411 p<,29913 Směrod. chyba odhadu : 6,6491					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(7)	p-hodn.
Abs.člen			86,74217	25,32397	3,425299	0,011056
X1	-0,550937	0,598452	-0,70031	0,76071	-0,920604	0,387883
X2	0,920415	0,598452	1,35062	0,87817	1,537994	0,167937

Hodnota testové statistiky pro test nevýznamnosti koeficientu vícenásobné korelace $\rho_{Y,(X_1,X_2)}$ je 1,4411, počet stupňů volnosti čitatele je 2, jmenovatele 7, odpovídající p-hodnota je 0,2991, tedy na hladině významnosti 0,05 nezamítáme hypotézu, že výkon dělníka není závislý na jeho věku a době zapracovanosti.

3. Parciální korelace

3.1. Koeficient parciální korelace

Nechť Y, Z jsou náhodné veličiny a $\mathbf{X} = (X_1, \ldots, X_p)'$ je náhodný vektor. Koeficient korelace $\rho(Y,Z)$ udává míru těsnosti lineárního vztahu mezi veličinami Y a Z. Ta však může být ovlivněna i tím, že mezi veličinami $X_1, \ldots, X_p$ existují veličiny, které silně korelují jak s Y, tak se Z. Zajímá nás proto, jaká je „čistá“ korelace mezi Y a Z, když se eliminuje vliv náhodného vektoru **X**. Pokud se omezíme na lineární vztahy, můžeme vliv vektoru **X** na veličinu Y popsat lineární regresní funkcí

$$\widehat{Y} = b_0 + b_1 X_1 + \ldots + b_p X_p .$$

Tu část veličiny Y, kterou vektor **X** nevysvětlí, si můžeme představit jako reziduum $Y - \widehat{Y}$. Analogicky pro veličinu Z dostáváme

$$\widehat{Z} = a_0 + a_1 X_1 + \ldots + a_p X_p ,$$

tudíž reziduum $Z - \widehat{Z}$ chápeme jako tu část veličiny Z, kterou vektor **X** nevysvětlí.

Koeficient korelace mezi rezidui $Y - \widehat{Y}$ a $Z - \widehat{Z}$ se nazývá parciální korelační koeficient mezi náhodnými veličinami Y a Z při pevně daném vektoru $\mathbf{X}$ a značí se $\rho_{Y,Z.\mathbf{X}}$.

Tedy $\rho_{Y,Z.\mathbf{X}} = \rho(Y - \widehat{Y}, Z - \widehat{Z})$. Počítá se podle vzorce

$$\rho_{Y,Z.\mathbf{X}} = \frac{\rho(Y,Z) - \mathrm{cov}(Y,\mathbf{X})\mathrm{cor}(\mathbf{X})^{-1}\mathrm{cov}(\mathbf{X},Z)}{\sqrt{\left[1 - \mathrm{cov}(Y,\mathbf{X})\mathrm{cor}(\mathbf{X})^{-1}\mathrm{cov}(\mathbf{X},Y)\right]\left[1 - \mathrm{cov}(Z,\mathbf{X})\mathrm{cor}(\mathbf{X})^{-1}\mathrm{cov}(\mathbf{X},Z)\right]}}.$$

3.2. Výběrový koeficient parciální korelace

Nechť náhodný vektor $(Y, Z, X_1, \ldots, X_p)'$ pochází z (p+2)-rozměrného rozložení, které má parciální korelační koeficient $\rho_{Y,Z.\mathbf{x}}$.

Nechť je dán náhodný výběr $(Y_1, Z_1, X_{11}, \ldots, X_{1p})', \ldots, (Y_n, Z_n, X_{n1}, \ldots, X_{np})'$ rozsahu n z tohoto rozložení. Musí platit n > p+2. Jako odhad $\rho_{Y,Z.\mathbf{x}}$ slouží výběrový parciální korelační koeficient $r_{Y,Z.\mathbf{x}}$:

$$r_{Y,Z.\mathbf{x}} = \frac{r_{YZ} - \mathbf{S}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{S}_{XZ}}{\sqrt{\left[1 - \mathbf{S}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{S}_{XY}\right]\left[1 - \mathbf{S}_{ZX}\mathbf{R}_{XX}^{-1}\mathbf{S}_{XZ}\right]}}$$

3.3. Test hypotézy o nevýznamnosti koeficientu parciální korelace

Budeme předpokládat, že uvedený náhodný výběr pochází z (p+2)-rozměrného normálního rozložení.

Testujeme hypotézu H_0: $\rho_{Y,Z.\mathbf{X}} = 0$ proti H_1: $\rho_{Y,Z.\mathbf{X}} \neq 0$.

Vzhledem k tomu, že se jedná o výběr z normálního rozložení, testujeme, zda existuje závislost mezi Y a Z při eliminaci vlivu **X**.

Testová statistika $T_0 = \dfrac{r_{Y,Z.\mathbf{X}}\sqrt{n-p-2}}{\sqrt{1-r_{Y,Z.\mathbf{X}}^{\,2}}}$ se řídí rozložením t(n-p-2), pokud H_0 platí.

Kritický obor: $W = (-\infty, t_{1-\alpha/2}(n-p-2)\rangle \cup \langle t_{1-\alpha/2}(n-p-2), \infty)$.

Jestliže $T_0 \in W$, H_0 zamítáme na hladině významnosti α.

3.4. Příklad

Pro data z příkladu o výkonnosti dělníků vypočtěte výběrové parciální korelační koeficienty $r_{Y,X_1.X_2}, r_{Y,X_2.X_1}$, interpretujte je, porovnejte je s obyčejnými výběrovými korelačními koeficienty r_{YX_1}, r_{YX_2} a pro $\alpha = 0{,}05$ otestujte významnost uvedených parciálních korelačních koeficientů.

Řešení:

Nejprve vypočteme párový koeficient korelace mezi výkonem a věkem.

Proměnná	X1
Y	0,2287

Dále vypočteme parciální korelační koeficient mezi výkonem a věkem při vyloučení vlivu doby zapracovanosti a otestujeme jeho významnost.

Proměnná	Y	X1
Y	1,0000	-,3286
	p= ---	p=,388
X1	-,3286	1,0000
	p=,388	p= ---

Korelační koeficient mezi výkonem a věkem vyšel 0,2287, tedy s rostoucím věkem roste výkon. Parciální korelační koeficient mezi výkonem a věkem při vyloučení vlivu doby zapracovanosti vyšel -0,3286, tedy u dělníků se stejnou dobou zapracovanosti klesá s rostoucím věkem výkon.
Odpovídající p-hodnota je 0,388, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nevýznamnosti $\rho_{Y,X_1.X_2}$.

Nyní vypočteme koeficient korelace mezi výkonem a dobou zapracovanosti:

Proměnná	X2
Y	0,4538

Dále vypočteme parciální korelační koeficient mezi výkonem a dobou zapracovanosti při vyloučení vlivu věku pracovníka a otestujeme jeho významnost.

Proměnná	Y	X2
Y	1,0000	,5026
	p= ---	p=,168
X2	,5026	1,0000
	p=,168	p= ---

Korelační koeficient mezi výkonem a dobou zapracovanosti vyšel 0,4538, tedy čím delší doba zapracovanosti, tím lepší výkon dělník podává. Parciální korelační koeficient mezi výkonem a dobou zapracovanosti při vyloučení vlivu věku vyšel 0,5026, tedy u stejně starých dělníků je poněkud silnější přímá lineární vazba mezi výkonem a dobou zapracovanosti.

Odpovídající p-hodnota je 0,168, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nevýznamnosti $\rho_{Y,X_2.X_1}$.