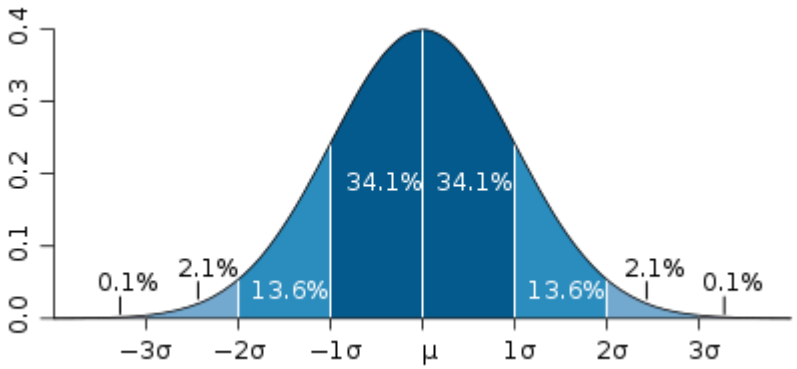
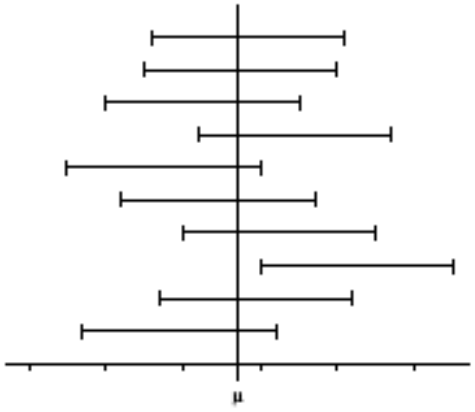
STATISTIKA II





Martin Řezáč, Marie Budíková

2013

Obsah

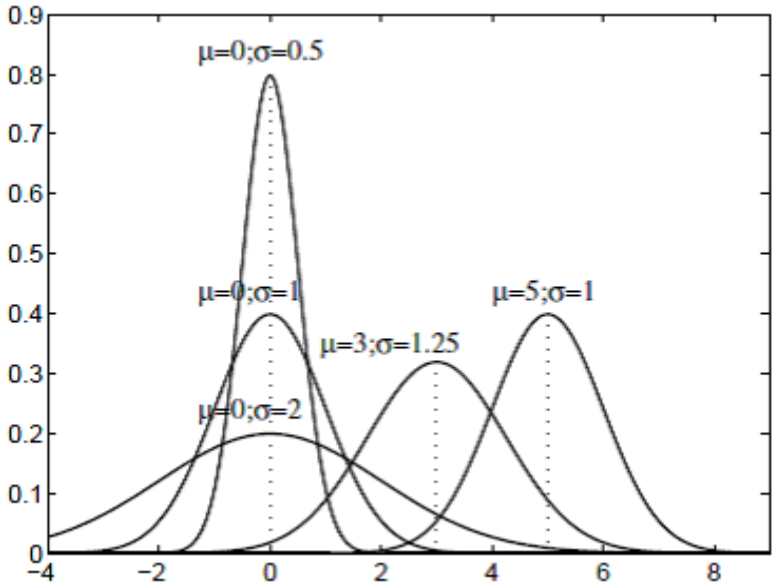
1. Normální rozložení a odvozená rozložení

Náhodná veličina s normálním rozložením $X \sim N(\mu, \sigma^2)$ má dominantní postavení v počtu pravděpodobnosti i v matematické statistice. Vyskytuje se v takových situacích, kdy se ke konstatní střední hodnotě μ přičítá velké množství nezávislých náhodních vlivů, které lehce kolísají kolem nuly. Takto vzniklá variabilita je charakterizována konstantou $\sigma \geq 0$. Normálně rozdělená náhodná veličina je tedy určena dvěma parametry μ a σ^2, kde μ je její střední hodnota a σ^2 je její rozptyl. Speciální případ, kde $\mu = 0$ a $\sigma^2 = 1$ nazýváme standardizované normální rozložení a značíme jej $U \sim N(0,1)$. Příklady: procentové změny v cenách akcií na dobře fungujících trzích (Eugene Chama, 1960), devizové výplatní poměry měn, …

Ze standardizovaného normálního rozložení U lze různými transformacemi odvodit další rozložení, z nichž se seznámíme s Pearsonovým χ^2-rozložením, studentovým t-rozložením a Fisher-Snedecorovým F-rozložením. Tato rozložení nacházejí velké uplatnění především v matematické statistice.

Definice normálního rozložení

O spojité náhodné veličině X říkáme, že má normální rozložení s parametry μ a σ^2, když její hustota je dána vzorcem $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, x \in R$. Zkráceně píšeme $X \sim N(\mu, \sigma^2)$.



Distribuční funkci normální náhodní veličiny X vyjádříme

$$\forall x \in R: F(x) = \int_{-\infty}^{x} f(t)\,dt = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt.$$

Definice standardizované normálné náhodné veličiny

Náhodnou veličinu $U \sim N(0,1)$ nazýváme standardizovaná normální náhodná veličina. Její hustota má tvar $f(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}u^2}, u \in R$

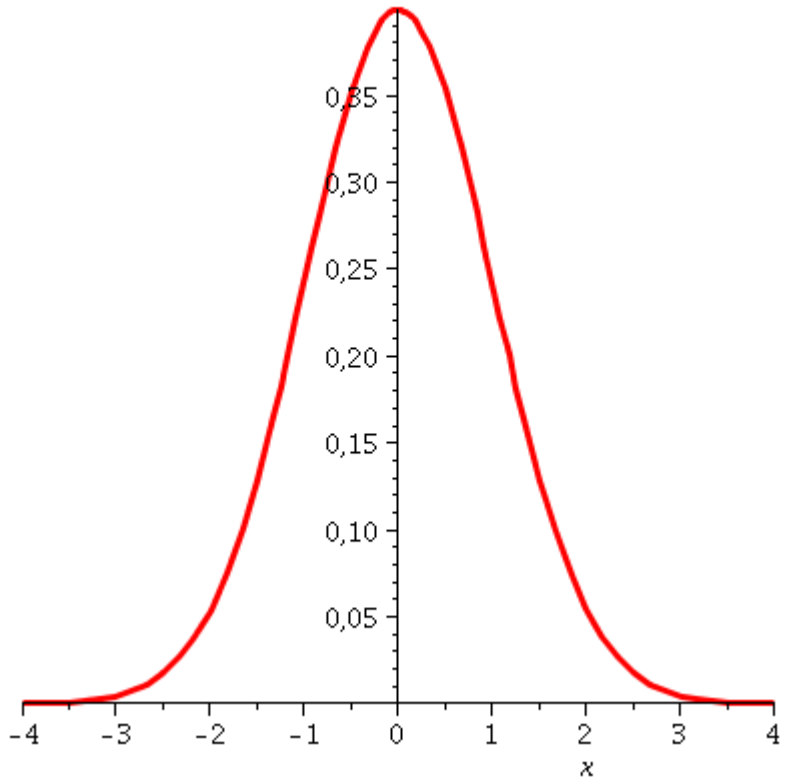
a distribuční funkce má tvar $F(u) = \int_{-\infty}^{u} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt.$

Následující věta uvede vlastnosti normálního rozložení.

Poznámka: Pro standardizovanou normální veličinu je obvyklé značení hustoty: $f(u) = \varphi(u)$ a distribuční funkce $F(u) = \Phi(u)$.

Ilustrace vlastností standardizované normální náhodné veličiny (1)

$$plot\left(\frac{1}{sqrt(2\cdot Pi)}\exp\left(-\frac{x^2}{2}\right),\ x=-4..4\right)$$



$$df := diff\left(\frac{1}{sqrt(2\cdot Pi)}\exp\left(-\frac{x^2}{2}\right), x\right)$$

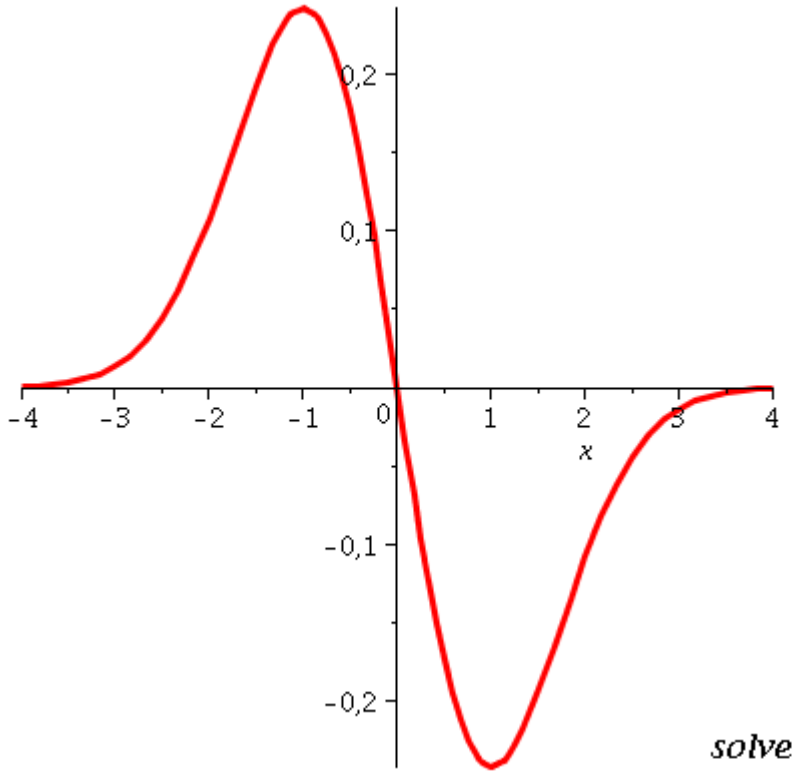
$$-\frac{1}{2}\,\frac{\sqrt{2}\,x\,e^{-\frac{1}{2}x^2}}{\sqrt{\pi}}$$

$$df2 := diff(df, x)$$

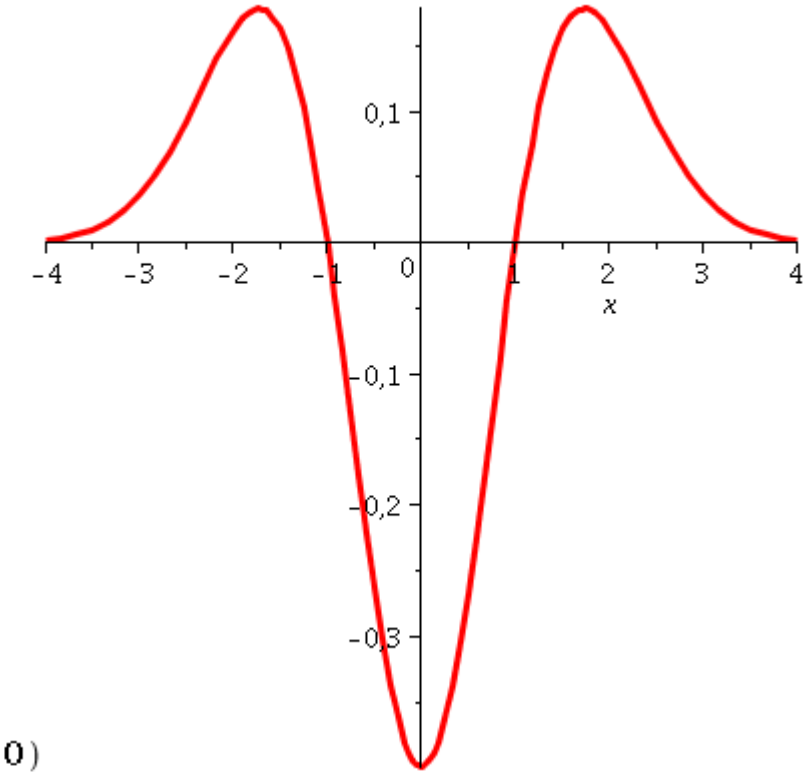
$$-\frac{1}{2}\,\frac{\sqrt{2}\,e^{-\frac{1}{2}x^2}}{\sqrt{\pi}} + \frac{1}{2}\,\frac{\sqrt{2}\,x^2\,e^{-\frac{1}{2}x^2}}{\sqrt{\pi}}$$

Ilustrace vlastností standardizované normální náhodné veličiny (2)

$plot(df,\ x=-4..4)$



$plot(df2,\ x=-4..4)$



$solve(df=0)$

0

$solve(df2=0)$

$1,\ -1$

Ilustrace vlastností standardizované normální náhodné veličiny (3)

Nalezení vrcholu a inflexních bodů v obecném případě:

$$f := \frac{1}{\sigma \cdot sqrt(2 \cdot \mathrm{Pi})} \exp\left(-\frac{\left(\frac{x-\mu}{\sigma}\right)^2}{2}\right)$$

$$\frac{1}{2}\,\frac{e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\sqrt{2}}{\sigma\sqrt{\pi}}$$

$$df := diff(f, x)$$

$$-\frac{1}{2}\,\frac{(x-\mu)\,e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\sqrt{2}}{\sigma^3\sqrt{\pi}}$$

$$solve(df = 0, x)$$

$$\mu$$

$$df2 := diff(df, x)$$

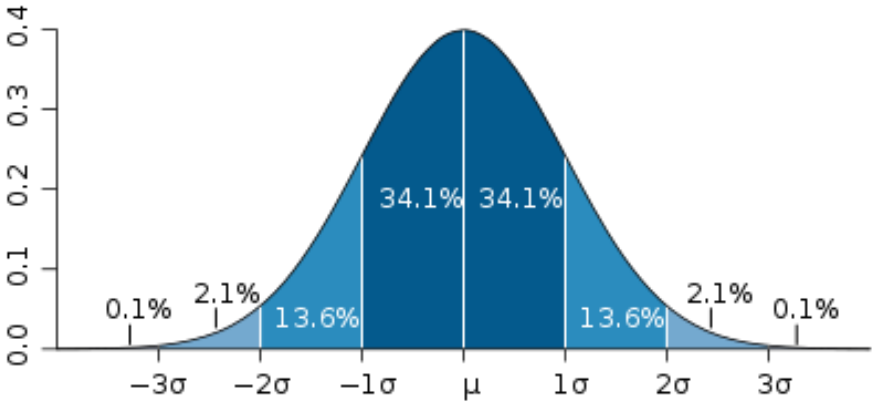
$$-\frac{1}{2}\,\frac{e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\sqrt{2}}{\sigma^3\sqrt{\pi}} + \frac{1}{2}\,\frac{(x-\mu)^2\,e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\sqrt{2}}{\sigma^5\sqrt{\pi}}$$

$$solve(df2 = 0, x)$$

$$\sigma + \mu,\ -\sigma + \mu$$

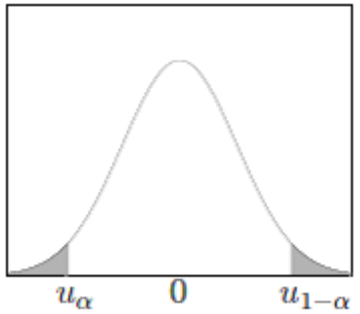
Hlavní charakteristiky křivky normálního rozdělení

- Unimodální
- Symetrická okolo průměru
- Asymptoticky se přibližuje k ose x
- Má zvonovitý tvar
- Plocha pod křivkou = 1
- Inflexní body leží ve vzdálenosti $\pm\sigma$ od průměru
- 99 % plochy pod křivkou se rozprostírá ve vzdálenosti $\pm 3\sigma$ od průměru



Vlastnosti normální náhodné veličiny

Věta 1.3



a) Jestliže $X \sim N(\mu, \sigma^2)$, pak $E(X) = \mu$, $D(X) = \sigma^2$.

b) Nechť $a, b \in \mathbf{R}$, $b \neq 0$.
Jestliže $X \sim N(\mu, \sigma^2)$ a $Y = a + bX$, pak $Y \sim N(a + b\mu, b^2\sigma^2)$.
[Lineární transformace normální náhodné veličiny normalitu neporuší.]

c) Nechť $X_1, \ldots X_n$ jsou stochasticky nezávislé náhodné veličiny,
$X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots n$. Pak $Y = \sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$
[Součet nezávislých normálních náhodných veličin je opět normální náhodná veličina.]

d) Nechť $X \sim N(\mu, \sigma^2)$. Pak $U = \frac{X-\mu}{\sigma} \sim N(0, 1)$
[Normální náhodnou veličinu X standardizujeme tak, že od ní odečteme její střední hodnotu a tento rozdíl pak dělíme její směrodatnou odchylkou.]

Poznámka 1.4
Distribuční funkce náhodné veličiny $U \sim N(0, 1)$ je tabelována ve statistických tabulkách pro $u \geq 0$. Jinak se užívá přepočtový vzorec $F(-u) = 1 - F(u)$. Kvantily náhodné veličiny $U \sim N(0, 1)$ se značí u_α a jsou tabelovány pro $\alpha \geq 0,5$. Jinak se užívá přepočtový vzorec $u_\alpha = -u_{1-\alpha}$.

Příklady

Příklad 1.5
Výsledky u přijímací zkoušky na jistou VŠ jsou normálně rozloženy se střední hodnotou $\mu = 550$ bodů a směrodatnou odchylkou $\sigma = 100$ bodů. Jaká je pravděpodobnost, že náhodně vybraný uchazeč bude mít aspoň 600 bodů?

Řešení
Náhodná veličina X udává bodový výsledek náhodně vybraného uchazeče, $X \sim N(550, 100^2)$

$$P(X \geq 600) = 1 - P(X < 600) = 1 - P(X \leq 600) + \overbrace{P(X = 600)}^{0} =$$
$$= 1 - P(\tfrac{X-550}{100} \leq \tfrac{600-550}{100}) = 1 - P(U \leq 0,5) = 1 - F(0,5) = 1 - 0,69146 \doteq 0,31.$$

$F(0,5)$ je distribuční funkce standardizovaného normálního rozložení v bodě 0,5 - viz. tabulky.

Příklad 1.6
Nechť $X \sim N(-1, 4)$. Najděte kvantil $K_{0,025}(X)$.

Řešení
$U = \frac{X+1}{2} \sim N(0,1), \qquad K_{0,025}(X) = ?$

$$0,025 = P(X \leq K_{0,025}(X)) = P(\tfrac{X+1}{2} \leq \tfrac{K_{0,025}(X)+1}{2}) = P(U \leq \tfrac{K_{0,025}(X)+1}{2}).$$
Tedy $\frac{K_{0,025}(X)+1}{2} = u_{0,025}$
Proto $K_{0,025}(X) = 2u_{0,025} - 1 = 2 \cdot (-u_{1-0,025}) - 1 = -2 \cdot u_{0,975} - 1 = -2 \cdot 1,96 - 1 = -4,92$

Příklad

Příklad: K danému číslu α, $0 < \alpha < 1$, určete interval tak, aby pro náhodnou veličinu U, která má normované normální rozdělení $N(0;1)$ platilo:

a) (♣) $P(|U| < a) = 1 - \alpha$;

b) (♣♣) $P(U < a) = 1 - \alpha$;

c) (♣♣♣) $P(U > a) = 1 - \alpha$.

Řešení: a) Z podmínky vyplývá

$1 - \alpha = P(-a < U < a) = \Phi(a) - \Phi(-a) = \Phi(a) - (1 - \Phi(a)) = 2\Phi(a) - 1 \Rightarrow \Phi(a) = 1 - \frac{\alpha}{2}$.

Odtud plyne, že $a = u_{1-\frac{\alpha}{2}}$ kvantil. Je tedy

$-a < U < a \Leftrightarrow -u_{1-\frac{\alpha}{2}} < U < u_{1-\frac{\alpha}{2}}$. Viz obr. 8.5.

b) Obdobně jako v a) dostaneme

$1 - \alpha = P(U < a) = \Phi(a) \Rightarrow a = u_{1-\alpha}$. Je tedy

$U < a \Leftrightarrow U < u_{1-\alpha}$. Viz obr. 8.6.

c) Z podmínky pro interval plyne

$1 - \alpha = P(U > a) = 1 - \Phi(a) \Rightarrow \Phi(a) = \alpha \Rightarrow a = u_{\alpha}$. Je tedy

$U > a \Leftrightarrow U > u_{\alpha}$.

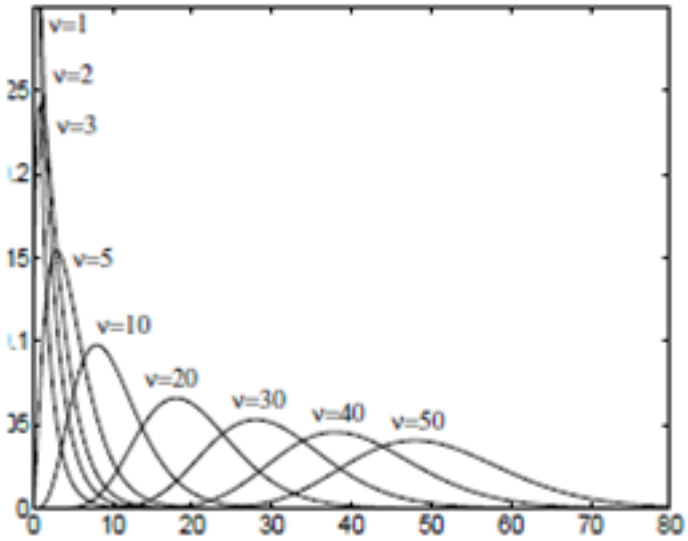
Definice odvozených rozložení (χ^2 - „chí kvadrát“)

Nyní budou následovat definice odvozených rozložení (Pearsonovo rozložení, Studentovo rozložení a Fisher-Snedecorovo rozložení) a související příklady. V definicích nepřehlédněte požadavky na nezávislost!

Definice 1.7

Nechť $U_1, \ldots, U_n$ jsou stochasticky nezávislé náhodné veličiny, $U_i \sim N(0,1)$, $i = 1, \ldots, n$. Pak náhodná veličina $V = \sum_{i=1}^{n} U_i^2 \sim \chi^2(n)$.

Říkáme, že náhodná veličina V má Pearsonovo rozložení "chí kvadrát" a parametr n nazýváme stupně volnosti.



Poznámka 1.8

$\alpha-$kvantil Pearsonova rozložení s n stupni volnosti značíme $\chi^2_\alpha(n)$. Tyto kvantily jsou tabelovány a pro $n > 30$ užíváme přibližný vztah $\chi^2_\alpha(n) \approx \frac{1}{2}(u_\alpha + \sqrt{2n-1})^2$

Příklad

Příklad 1.9

a) Nechť $V \sim \chi^2(10)$. Najděte kvantil $\chi^2_{0,975}(10)$.

b) Nechť $V \sim \chi^2(3)$. Najděte kvantil $\chi^2_{0,05}(3)$.

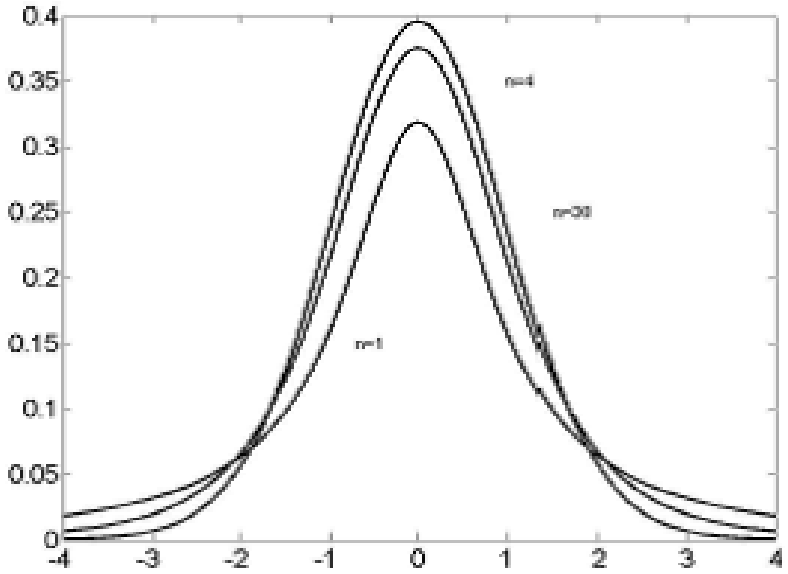
Řešení

a) $\chi^2_{0,975}(10) = 20,483.$

b) $\chi^2_{0,05}(3) = 0,352.$

Definice odvozených rozložení (*t*-rozložení)

Definice 1.10
Nechť U, V jsou stochasticky nezávislé náhodné veličiny, $U \sim N(0,1),\ V \sim \chi^2(n)$.
Pak náhodná veličina $T = \frac{U}{\sqrt{\frac{V}{n}}} \sim t(n)$. Říkáme, že náhodná veličina T má Studentovo rozložení s n stupni volnosti.



Poznámka 1.11
$\alpha-$kvantil Studentova rozložení s n stupni volnosti značíme $t_\alpha(n)$. Tyto kvantily jsou tabelovány. Pro $\alpha < 0,5$ se používá přepočtový vzorec $t_\alpha(n) = -t_{1-\alpha}(n)$ a pro distribuční funkci platí vztah $F(-x) = 1 - F(x)$.

Příklady

Příklad 1.12

a) Nechť $T \sim t(8)$. Najděte kvantil $t_{0,9}(8)$.

b) Nechť $T \sim t(6)$. Najděte kvantil $t_{0,05}(6)$.

Řešení

a) $t_{0,9}(8) = 1,3968.$

b) $t_{0,05}(6) = -t_{0,95}(6) = -1,9432.$

Příklad 1.13

Nechť $X \sim t(14)$. Určete konstantu c tak, aby platilo: $P(-c < X < c) = 0,9.$

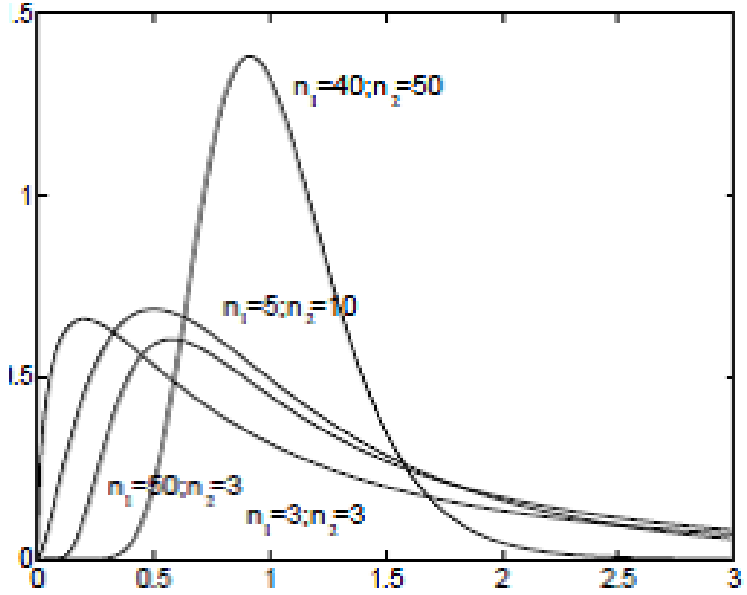
Řešení

$0,9 = P(-c < X < c) = F(c) - F(-c) = F(c) - [1 - F(c)] = 2F(c) - 1$

Tedy $0,9 = 2F(c) - 1 \Rightarrow F(c) = \frac{1,9}{2} = 0,95 \Rightarrow c = t_{0,95}(14) = 1,7613$

Definice odvozených rozložení (F-rozložení)

Definice 1.14
Nechť V_1, V_2 jsou stochasticky nezávislé náhodné veličiny, $V_1 \sim \chi^2(n_1)$, $V_2 \sim \chi^2(n_2)$. Pak náhodná veličina $F = \frac{V_1/n_1}{V_2/n_2} \sim F(n_1, n_2)$. Říkáme, že náhodná veličina F má Fisher-Snedecorovo rozložení, kde n_1 je počet stupňů volnosti čitatele a n_2 je počet stupňů volnosti jmenovatele.



Poznámka 1.15
$\alpha-$kvantil Fisher-Snedecorova rozložení se stupni volnosti n_1, n_2 značíme $F_\alpha(n_1, n_2)$. Tyto kvantily jsou tabelovány. Pro $\alpha < 0,5$ se používá přepočtový vzorec $F_\alpha(n_1, n_2) = \frac{1}{F_{1-\alpha}(n_2, n_1)}$.

Příklady

Příklad 1.16

a) Nechť $F \sim F(5,7)$. Najděte kvantil $F_{0,975}(5,7)$.

b) Nechť $F \sim F(8,6)$. Najděte kvantil $F_{0,025}(8,6)$.

Řešení

a) $F_{0,975}(5,7) = 5,2852.$

b) $F_{0,025}(8,6) = \frac{1}{F_{0,975}(6,8)} = \frac{1}{4,6517} = 0,215.$

Příklad 1.17

Nechť $X \sim F(5,8)$. Určete konstantu c tak, aby platilo: $P(X < c) = 0,05$.

Řešení

$0,05 = P(X < c) = F(c)$

Tedy $c = F_{0,05}(5,8) = \frac{1}{F_{0,95}(8,5)} = \frac{1}{4,8183} = 0,2075.$

N-rozměrné normální rozložení

Nyní se budeme věnovat náhodnému vektoru s $n-$rozměrným normálním rozložením, pro jednoduchost budeme uvažovat $n = 2$. Konvenci při zapisování náhodných vektorů ilustrujme následovně:

sloupcový vektor náhodných veličin značíme velkým tlustým písmenem, např. $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$

sloupcový vektor konstant značíme malým tlustým písmenem, např. $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$

Náhodný vektor s dvourozměrným normálním rozložením

Definice 1.18

O spojitém náhodném vektoru $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ říkáme, že má dvojrozměrné normální rozložení s parametry $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ a $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, když jeho hustota je dána vzorcem

$$f(\mathbf{x}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\cdot\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}, \quad \mathbf{x} \in \mathbf{R}^2.$$

Zkráceně píšeme $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2(\mu, \Sigma)$.

Pro $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ a $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ mluvíme o standardizovaném dvojrozměrném normálním rozložení.

Poznámka 1.19

Význam parametrů je následující:

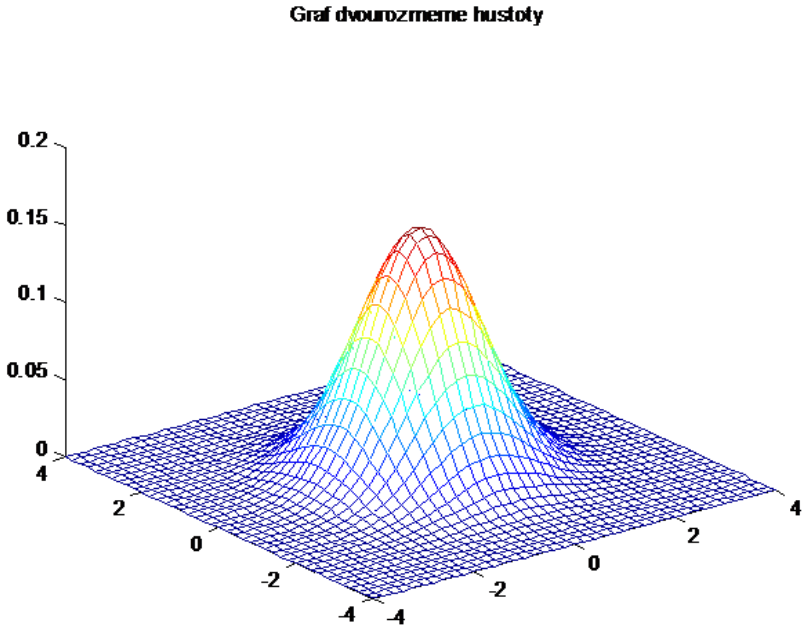
$\mu_1 = E(X_1),\ \mu_2 = E(X_2),\ \sigma_1^2 = D(X_1),\ \sigma_2^2 = D(X_2),\ \rho = R(X_1, X_2)$

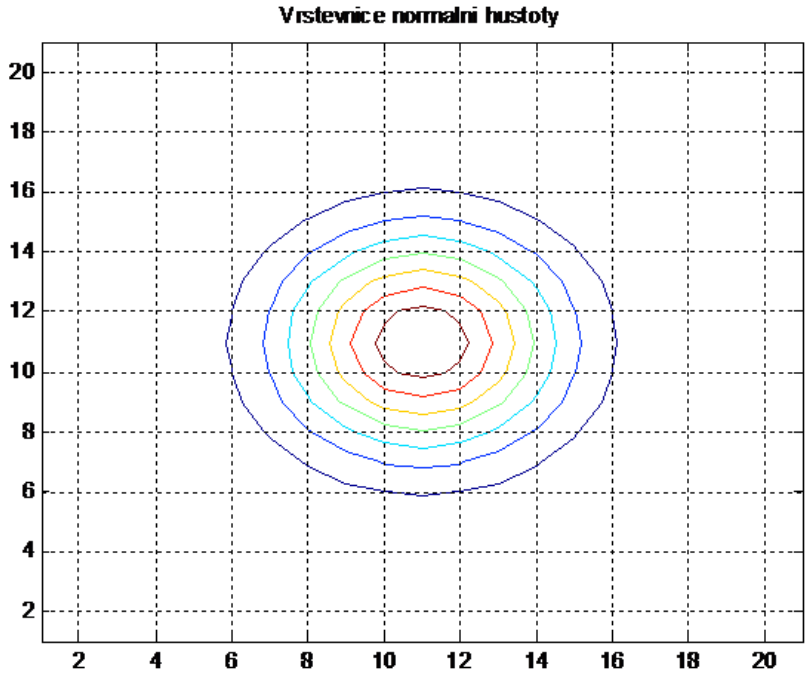
Graf dvourozměrné hustoty (1)

$$\mu_1 = \mu_2 = 0$$

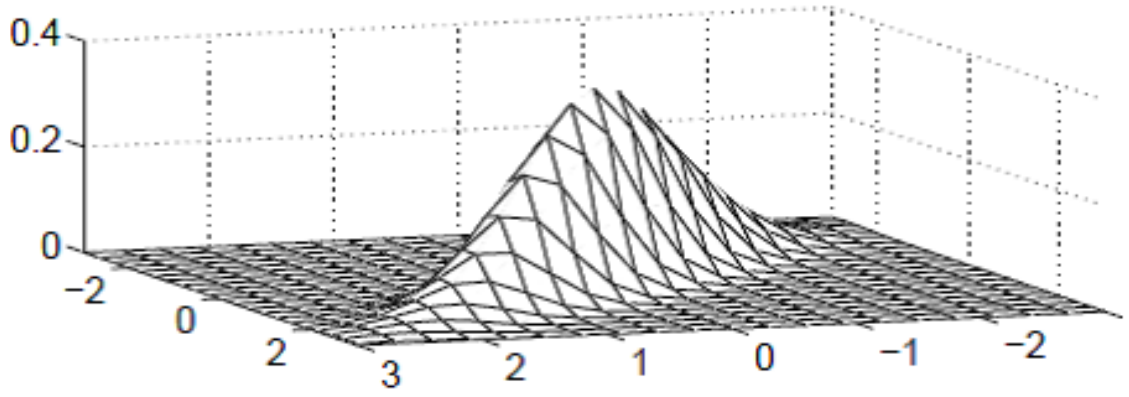
$$\sigma_1 = \sigma_2 = 1$$

$$\rho_{12} = 0$$





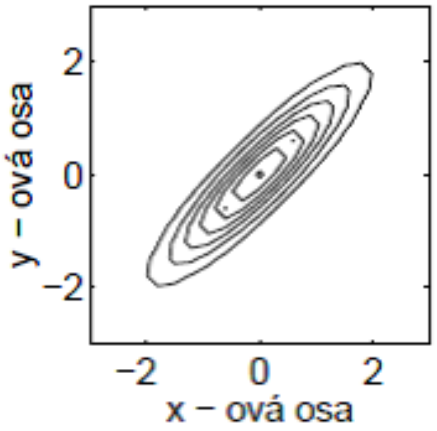
Graf dvourozměrné hustoty (2)



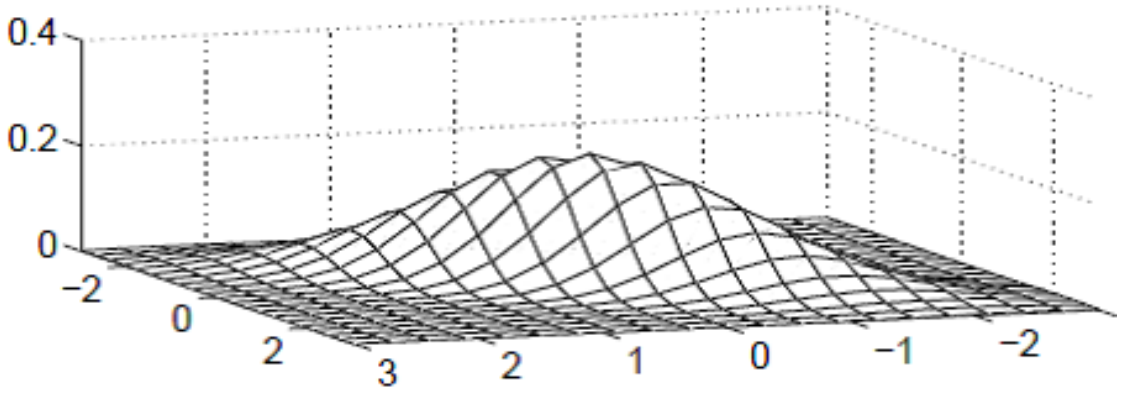
$\mu_1 = \mu_2 = 0$

$\sigma_1 = \sigma_2 = 1$

$\rho_{12} = 0.9$



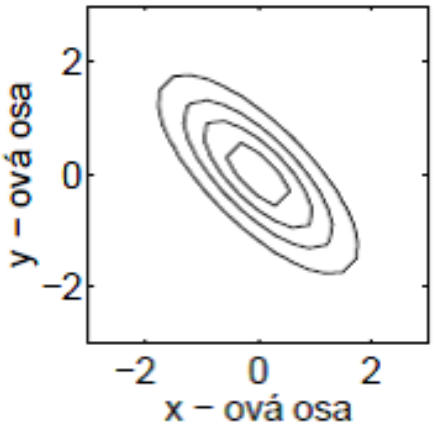
Graf dvourozměrné hustoty (3)



$\mu_1 = \mu_2 = 0$

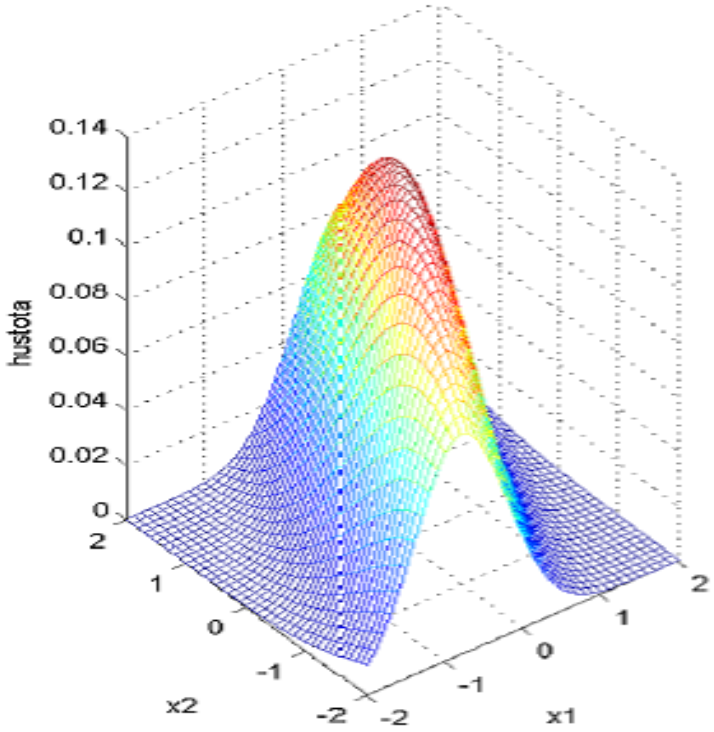
$\sigma_1 = \sigma_2 = 1$

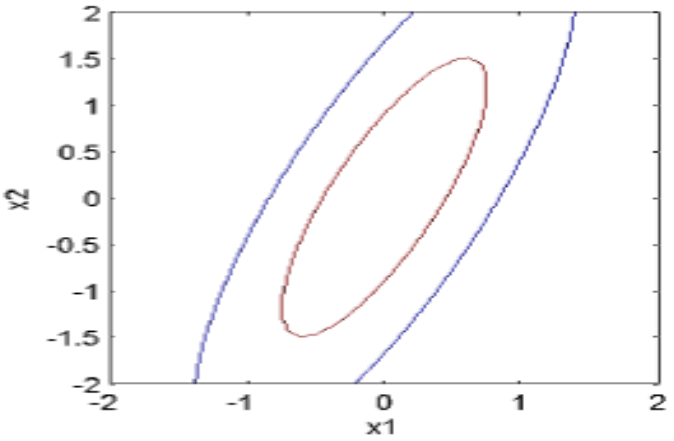
$\rho_{12} = -0.75$



Graf dvourozměrné hustoty (4)

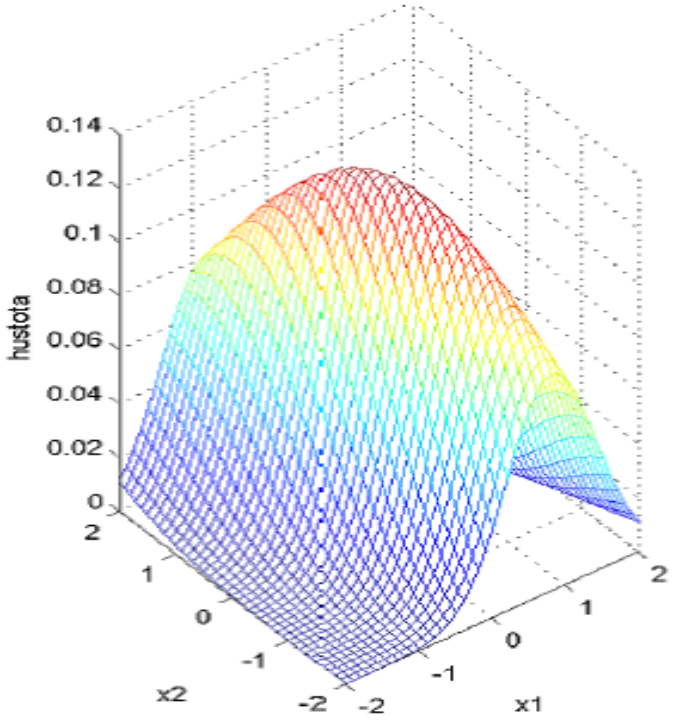
$$\mu_1 = \mu_2 = 0,\ \sigma_1 = 1,\ \sigma_2 = 2,\ \rho = 0.8$$

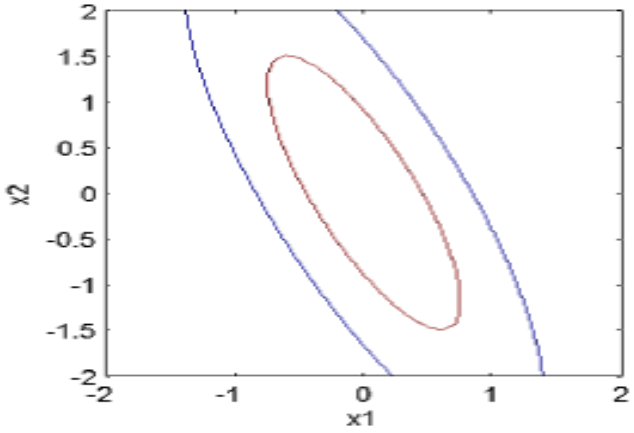




Graf dvourozměrné hustoty (5)

$$\mu_1 = \mu_2 = 0,\ \sigma_1 = 1,\ \sigma_2 = 2,\ \rho = -0.8$$





Marginální rozložení skalární NV a lineární transformace

Věta 1.20

Nechť dvojrozměrný vektor $\mathbf{X}$ má dvojrozměrné normální rozložení

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right).$$

Potom pro marginální rozložení skalární náhodné veličiny X_i platí: $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$.

[Složky normálního náhodného vektoru normalitu "podědí".]

Věta 1.21

Nechť $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$ je normální náhodný vektor, nechť $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ je vektor reálných čísel, nechť $\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$ je matice reálných čísel.

Potom transformovaný náhodný vektor $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X} \sim N_2(\mathbf{a} + \mathbf{B}\mu,\ \mathbf{B}\Sigma\mathbf{B}')$.

[Lineární transformace zachovává normalitu.]

Příklad

Příklad 1.22

Nechť devizový kurs marky je náhodná veličina $X_1 \sim N(19, 0.5^2)$ a devizový kurs dolaru je náhodná veličina $X_2 \sim N(32, 0.6^2)$. Korelace $R(X_1, X_2) = -0.8$. Jaká je pravděpodobnost, že měnový koš $0.65X_1 + 0.35X_2$ bude mít hodnotu větší než 24? Návod:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 19 \\ 32 \end{pmatrix}, \begin{pmatrix} 0.25 & -0.072 \\ -0.072 & 0.36 \end{pmatrix}\right)$$

$$0.65X_1 + 0.35X_2 = (0.65\ 0.35)\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

Příklad

Příklad 1.23
Nechť kursy dvou akcií jsou náhodné veličiny $X_1 \sim N(600, 40^2)$, $X_2 \sim N(800, 30^2)$. Korelace $R(X_1, X_2) = -0.4$. Jaká je pravděpodobnost, že index $X_1 + X_2$ nepoklesne pod 1300 bodů?

Řešení
$P(X_1 + X_2 \geq 1300) = ?$
Jelikož X_1, X_2 jsou korelované, nelze užít věty 1.3.c). Abychom mohli užít vět 1.20 a 1.21, musíme nejdříve určit rozložení náhodného vektoru $\mathbf{X} = \binom{X_1}{X_2}$.

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 600 \\ 800 \end{pmatrix}, \begin{pmatrix} 1600 & -480 \\ -480 & 900 \end{pmatrix} \right).$$

(Pro prvek σ_{12} matice Σ platí: $\sigma_{12} = \sigma_{21} = \rho\sigma_1\sigma_2 = -0,4 \cdot 40 \cdot 30 = -480$)

Užijeme-li ve větě 1.21 $\mathbf{a} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ a $\mathbf{B} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$ potom transformovaný náhodný vektor $\mathbf{Y} = \mathbf{a} + \mathbf{BX} = \begin{pmatrix} X_1 + X_2 \\ 0 \end{pmatrix}$ zůstává normálně rozložený a dle věty 1.20 je normálně rozložená i každá jeho složka. Tedy náhodná veličina $Z = X_1 + X_2$ má normální rozložení a pro její parametry platí:

$E(Z) = E(X_1 + X_2) = E(X_1) + E(X_2) = 600 + 800 = 1400$
$D(Z) = D(X_1 + X_2) = D(X_1) + D(X_2) + 2C(X_1, X_2) = 1600 + 900 - 2 \cdot 480 = 1540$
$Z \sim N(1400, 1540)$

$$P(X_1 + X_2 \geq 1300) = P(Z \geq 1300) = 1 - P(Z \leq 1300) + \overbrace{P(Z = 1300)}^{0} =$$
$$= 1 - P\left(\frac{Z-1400}{\sqrt{1540}} \leq \frac{1300-1400}{\sqrt{1540}}\right) = 1 - P(U \leq -2,55) = P(U \leq 2,55) = 0,9946.$$

Index $X_1 + X_2$ nepoklesne pod 1300 bodů s pravděpodobností 0,9946.

Vlastnosti vícerozměrného normálního rozdělení

Náhodný vektor $\mathbf{X}$ má vícerozměrné normální rozdělení, jestliže jeho hustota je dána vztahem

$$f(\mathbf{x}) = (2\pi)^{p/2} \ |\boldsymbol{\Sigma}|^{-1/2} \ \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right),$$

kde $\boldsymbol{\mu}$ je vektor středních hodnot a $\boldsymbol{\Sigma}$ je kovarianční matice.

Vícerozměrné normální rozdělení má tyto vlastnosti:

- lineární kombinace prvků z $\mathbf{X}$ mají normální rozdělení
- všechny podmnožiny $\mathbf{X}$ mají normální rozdělení
- nekorelovanost veličin z $\mathbf{X}$ (složek vektoru $\mathbf{X}$) znamená i jejich nezávislost
- všechna podmíněná rozdělení jsou normální

Mahalanobisova vzdálenost

I pro vícerozměrné normální rozdělení je možno chápat kvadratickou formu v exponentu jako čtverec vzdálenosti vektoru $\mathbf{x}$ od vektoru $\boldsymbol{\mu}$, ve kterém je obsažena i informace z kovarinační matice

$$C^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}).$$

C je Mahalanobisova vzdálenost, pro zvolenou hodnotu $f(\mathbf{x})$ její čtverec je geometricky plocha elipsoidu se středem μ a osami $c\sqrt{\lambda_j \mathbf{v}_j}$ pro $j = 1, 2, \cdots, p$, kde λ_j jsou vlastní čísla matice $\boldsymbol{\Sigma}$ a $\mathbf{v}_j$ jsou vlastní vektory matice $\boldsymbol{\Sigma}$.

$$C^2 = (\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(p)$$

2. Základní pojmy matematické statistiky. Diagnostické grafy.

Motivace: Matematická statistika je věda, která analyzuje a interpretuje data především za účelem získání předpovědi a zlepšení rozhodování v různých oborech lidské činnosti. Přitom se řídí principem statistické indukce, tj. na základě znalostí o náhodném výběru z určitého rozložení pravděpodobností se snaží učinit závěry o vlastnostech tohoto rozložení.

Ústředním pojmem matematické statistiky je tedy pojem náhodného výběru.

Definice náhodného výběru

a) Nechť $X_1, ..., X_n$ jsou stochasticky nezávislé náhodné veličiny, které mají všechny stejné rozložení $L(\vartheta)$. Řekneme, že $X_1, ..., X_n$ je náhodný výběr rozsahu n z rozložení $L(\vartheta)$. (Číselné realizace $x_1, ..., x_n$ náhodného výběru $X_1, ..., X_n$ uspořádané do sloupcového vektoru odpovídají datovému souboru zavedenému v popisné statistice.)

b) Nechť $(X_1,Y_1), ..., (X_n,Y_n)$ jsou stochasticky nezávislé dvourozměrné náhodné vektory, které mají všechny stejné dvourozměrné rozložení $L_2(\vartheta)$. Řekneme, že $(X_1,Y_1), ..., (X_n,Y_n)$ je dvourozměrný náhodný výběr rozsahu n z dvourozměrného rozložení $L_2(\vartheta)$. (Číselné realizace $(x_1,y_1), ..., (x_n,y_n)$ náhodného výběru $(X_1,Y_1), ..., (X_n,Y_n)$ uspořádané do matice typu n x 2 odpovídají dvourozměrnému datovému souboru zavedenému v popisné statistice.)

c) Analogicky lze definovat p-rozměrný náhodný výběr rozsahu n z p-rozměrného rozložení $L_p(\vartheta)$.

Definice statistiky

Libovolná funkce $T = T(X_1, ..., X_n)$ náhodného výběru $X_1, ..., X_n$ (resp. $T = T(X_1, Y_1, ..., X_n, Y_n)$ náhodného výběru $(X_1, Y_1), ..., (X_n, Y_n)$) se nazývá (výběrová) statistika.

Důsledek:

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení s distribuční funkcí $\Phi(x)$. Pak simultánní distribuční funkce náhodného vektoru $(X_1, ..., X_n)$ je

$\Phi(x_1) \dots \Phi(x_n)$.

Definice důležitých statistik (1)

a) Nechť $X_1, \ldots, X_n$ je náhodný výběr, $n \geq 2$.

$M = \frac{1}{n}\sum_{i=1}^{n} X_i$ … výběrový průměr,

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2$ … výběrový rozptyl,

$S = \sqrt{S^2}$ … výběrová směrodatná odchylka

Pro libovolné, ale pevně dané reálné číslo x je statistikou též hodnota výběrové distribuční funkce $F_n(x) = \frac{1}{n} card\{i; X_i \leq x\}$

Definice důležitých statistik (2)

b) Nechť je dáno $r \geq 2$ stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2, \ldots, n_r \geq 2$. Celkový rozsah je $n = \sum_{j=1}^{r} n_j$. Označme $M_1, \ldots, M_r$ výběrové průměry a $S_1^2, \ldots, S_r^2$ výběrové rozptyly jednotlivých výběrů. Nechť $c_1, \ldots, c_r$ jsou reálné konstanty, aspoň jedna nenulová.

$\sum_{j=1}^{r} c_j M_j$ … lineární kombinace výběrových průměrů,

$S_*^2 = \frac{\sum_{j=1}^{r} (n_j - 1) S_j^2}{n - r}$ … vážený průměr výběrových rozptylů.

Definice důležitých statistik (3)

c) Nechť $(X_1,Y_1), \ldots, (X_n,Y_n)$ je náhodný výběr z dvourozměrného rozložení.

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

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

$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$ … výběrová kovariance,

$$R_{12} = \begin{cases} \frac{1}{n-1}\sum_{i=1}^{n}\frac{X_i - M_1}{S_1}\cdot\frac{Y_i - M_2}{S_2} = \frac{S_{12}}{S_1 S_2} & \text{pro } S_1 S_2 \neq 0 \\ 0 & \text{jinak} \end{cases}$$

… výběrový koeficient korelace.

Definice důležitých statistik (4)

Pro libovolnou, ale pevně zvolenou dvojici reálných čísel x,y je statistikou též hodnota výběrové simultánní distribuční funkce

$$F_n(x,y) = \frac{1}{n} card\{ i; X_i \leq x \wedge Y_i \leq y\}.$$

Upozornění: Číselné realizace statistik M, S^2, S, S_{12}, R_{12} odpovídají číselným charakteristikám m, s^2, s, s_{12}, r_{12} zavedeným v popisné statistice, ale u rozptylu, směrodatné odchylky, kovariance a koeficientu korelace je multiplikativní konstanta $\frac{1}{n-1}$, nikoliv $\frac{1}{n}$, jak tomu bylo v popisné statistice. Jak uvidíme později, uvedené číselné realizace mohou být považovány za odhady číselných realizací náhodných veličin zavedených v počtu pravděpodobnosti.

Definice důležitých statistik (5)

Charakteristika vlastnosti	Počet pravděpodobnosti	Matematická statistika	Popisná statistika
poloha	$E(X) = \mu$	M	m
variabilita	$D(X) = \sigma^2$	S^2	$\frac{n-1}{n} s^2$
variabilita	$\sqrt{D(X)} = \sigma$	S	$\sqrt{\frac{n-1}{n}} s$
společná variabilita	$C(X_1, X_2) = \sigma_{12}$	S_{12}	$\frac{n-1}{n} s_{12}$
těsnost vztahu	$R(X_1, X_2) = \rho$	R_{12}	r_{12}
rozložení	$\Phi(x)$	$F_n(x)$	F(x)

Příklad (1)

(Výpočet realizací výběrového průměru, výběrového rozptylu a hodnot výběrové distribuční funkce)

Desetkrát nezávisle na sobě byla změřena jistá konstanta μ. Výsledky měření byly: 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2. Tyto výsledky považujeme za číselné realizace náhodného výběru $X_1, \ldots, X_{10}$. Vypočtěte realizaci m výběrového průměru M, realizaci s^2 výběrového rozptylu S^2, realizaci s výběrové směrodatné odchylky S a hodnoty výběrové distribuční funkce $F_{10}(x)$.

Řešení:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(2 + 1{,}8 + \ldots + 2{,}2) = 2{,}06$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} {x_i}^2 - nm^2\right) =$$

$$= \frac{1}{9}(2^2 + 1{,}8^2 + \ldots + 2{,}2^2 - 10 \cdot 2{,}06^2) = 0{,}0404$$

$$s = \sqrt{s^2} = \sqrt{0{,}0404} = 0{,}2011$$

Příklad (2)

(Výpočet realizací výběrového průměru, výběrového rozptylu a hodnot výběrové distribuční funkce)

Pro usnadnění výpočtu hodnot výběrové distribuční funkce $F_{10}(x)$ uspořádáme měření podle velikosti: 1,8 1,8 1,9 2 2 2,1 2,1 2,2 2,3 2,4.

$$x < 1{,}8 : F_{10}(x) = 0$$

$$1{,}8 \le x < 1{,}9 : F_{10}(x) = \frac{2}{10} = 0{,}2$$

$$1{,}9 \le x < 2 : F_{10}(x) = \frac{3}{10} = 0{,}3$$

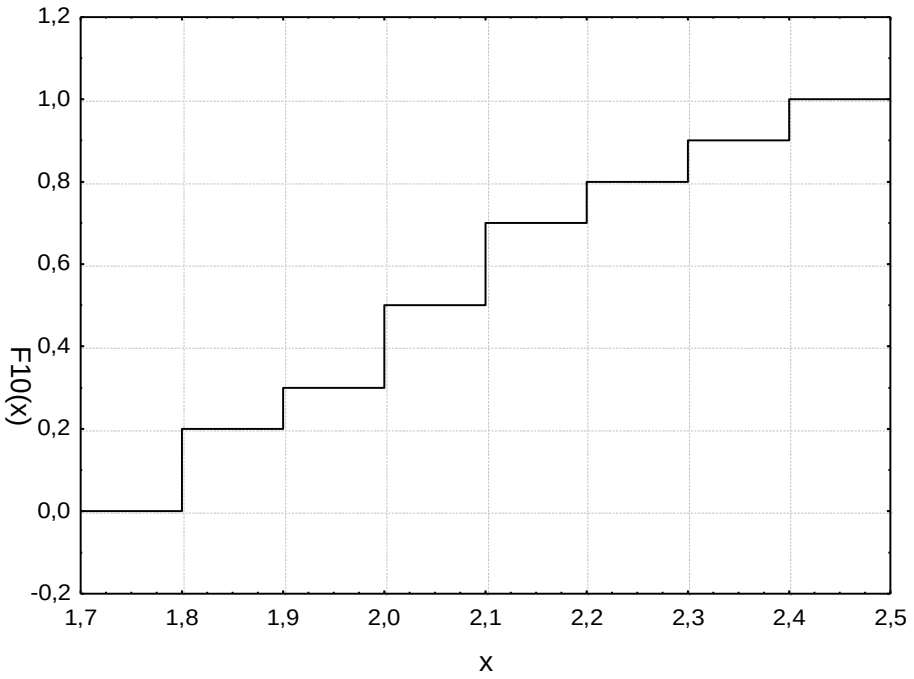
$$2 \le x < 2{,}1 : F_{10}(x) = \frac{5}{10} = 0{,}5$$

$$2{,}1 \le x < 2{,}2 : F_{10}(x) = \frac{7}{10} = 0{,}7$$

$$2{,}2 \le x < 2{,}3 : F_{10}(x) = \frac{8}{10} = 0{,}8$$

$$2{,}3 \le x < 2{,}4 : F_{10}(x) = \frac{9}{10} = 0{,}9$$

$$x \ge 2{,}4 : F_{10}(x) = 1$$



Příklad

(Výpočet realizace výběrového koeficientu korelace)

U 11 náhodně vybraných aut jisté značky bylo zjišťováno jejich stáří (náhodná veličina X – v letech) a cena (náhodná veličina Y – v tisících Kč). Výsledky: (5, 85), (4, 103), (6, 70), (5, 82), (5, 89), (5, 98), (6, 66), (6, 95), (2, 169), (7, 70), (7, 48). Vypočtěte a interpretujte číselnou realizaci r_{12} výběrového koeficientu korelace R_{12}.

Řešení:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{11}(5 + 4 + \ldots + 7) = 5{,}28$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{11}(85 + 103 + \ldots + 48) = 88{,}63$$

$$s_1{}^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i{}^2 - n m_1{}^2\right) = \frac{1}{10}(5^2 + 4^2 + \ldots + 7^2 - 11 \cdot 5{,}28^2) = 2{,}02$$

$$s_2{}^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i{}^2 - n m_2{}^2\right) = \frac{1}{10}(85^2 + 103^2 + \cdots + 48^2 - 11 \cdot 88{,}63^2) = 970{,}85$$

$$s_{12} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n m_1 m_2\right) = \frac{1}{10}(5 \cdot 85 + 4 \cdot 103 + \ldots + 7 \cdot 48 - 11 \cdot 5{,}28 \cdot 88{,}63) = -40{,}89$$

$$r_{12} = \frac{s_{12}}{s_1 \cdot s_2} = \frac{-40{,}82}{\sqrt{2{,}02} \cdot \sqrt{970{,}85}} = -0{,}92$$

Mezi náhodnými veličinami X a Y existuje silná nepřímá lineární závislost. Čím starší auto, tím nižší cena.

Vlastnosti důležitých statistik (1)

a) Případ jednoho náhodného výběru: Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení se střední hodnotou μ, rozptylem σ^2 a distribuční funkcí $\Phi(x)$. Nechť $n \geq 2$. Označme M_n výběrový průměr, S_n^2 výběrový rozptyl a pro libovolné, ale pevně dané $x \in R$ označme $F_n(x)$ hodnotu výběrové distribuční funkce. Pak pro libovolné hodnoty parametrů μ , σ^2 a libovolné, ale pevně dané reálné číslo x platí:

$E(M_n) = \mu,$

$D(M_n) = \frac{\sigma^2}{n},$

$E(S_n^2) = \sigma^2,$

$D(S_n^2) = \frac{\gamma_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$, kde γ_4 je 4. centrální moment,

$E(F_n(x)) = \Phi(x),$

$$D(F_n(x)) = \frac{\Phi(x)[1-\Phi(x)]}{n}$$

Vlastnosti důležitých statistik (2)

b) Případ r ≥ 2 stochasticky nezávislých náhodných výběrů: Nechť $X_{11}, \ldots, X_{1n_1}$, ..., $X_{r1}, \ldots, X_{rn_r}$ je r stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2$, ..., $n_r \geq 2$ z rozložení se středními hodnotami μ_1, ..., μ_r a rozptylem σ^2. Celkový rozsah je $n = \sum_{j=1}^{r} n_j$. Nechť c_1, ..., c_r jsou reálné konstanty, aspoň jedna nenulová. Pak pro libovolné hodnoty parametrů $\mu_1, \ldots, \mu_r$ a σ^2 platí:

$$E\left(\sum_{j=1}^{r} c_j M_j\right) = \sum_{j=1}^{r} c_j \mu_j,$$

$E(S_*^2) = \sigma^2$.

Vlastnosti důležitých statistik (3)

c) Případ jednoho náhodného výběru z dvourozměrného rozložení: Nechť (X_1,Y_1), ..., (X_n,Y_n) je náhodný výběr z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ. Pak pro libovolné hodnoty parametrů σ_{12} a ρ platí:

$E(S_{12}) = \sigma_{12}$,

$E(R_{12}) \approx \rho$ (shoda je vyhovující pro $n \geq 30$).

Poznámka: Metody matematické statistiky často slouží k vyhodnocování výsledků pokusů. Aby mohl být pokus správně vyhodnocen, musí být dobře naplánován. Uvedeme zde nejjednodušší typy uspořádání pokusů.

Předpokládejme například, že sledujeme hmotnostní přírůstky selat téhož plemene při různých výkrmných dietách.

Typy pozorování (1)

a) Jednoduché pozorování: Náhodná veličina X je pozorována za týchž podmínek. Situace je charakterizována jedním náhodným výběrem $X_1, ..., X_n$.

Náhodně vylosujeme n selat téhož plemene, podrobíme je jediné výkrmné dietě a zjistíme u každého selete hmotnostní přírůstek. Tím dostaneme realizaci jednoho náhodného výběru.

Typy pozorování (2)

b) Dvojné pozorování: Náhodná veličina X je pozorována za dvojích různých podmínek. Existují dvě odlišná uspořádání tohoto pokusu.

Dvouvýběrové porovnávání: situace je charakterizována dvěma nezávislými náhodnými výběry $X_{11}, \dots, X_{1n_1}$ a $X_{21}, \dots, X_{2n_2}$.

Náhodně vylosujeme n_1 a n_2 selat téhož plemene, náhodně je rozdělíme na dva soubory o n_1 a n_2 jedincích, první podrobíme výkrmné dietě č. 1 a druhý výkrmné dietě číslo 2. Tak dostaneme realizace dvou nezávislých náhodných výběrů.

Párové porovnávání: situace je charakterizována jedním náhodným výběrem $(X_{11}, X_{12}), \dots, (X_{n1}, X_{n2})$ z dvourozměrného rozložení. Přejdeme k rozdílovému náhodnému výběru $Z_i = X_{i1} - X_{i2}$, i = 1, …, n a tím dostaneme jednoduché pozorování.

Náhodně vylosujeme n vrhů stejně starých selat téhož plemene, z každého odebereme dva sourozence a náhodně jim přiřadíme první a druhou výkrmnou dietu. Tak dostaneme realizaci jednoho dvourozměrného náhodného výběru, kde první složka odpovídá první dietě a druhá složka druhé dietě.

(Párové porovnávání je efektivnější, protože skutečný rozdíl v účinnosti obou diet je překrýván pouze náhodnými vlivy při samotném krmení a trvání, kdežto vliv různých dědičných vloh, který byl losováním znárodněn, je u sourozeneckého páru selat částečně vyloučen.)

Typy pozorování (3)

c) Mnohonásobné pozorování: Náhodná veličina X je pozorována za r ≥ 3 různých podmínek. Existují dvě odlišná uspořádání tohoto pokusu.

Mnohovýběrové porovnávání: situace je charakterizována r nezávislými náhodnými výběry $X_{11}, \dots, X_{1n_1}$ až $X_{r1}, \dots, X_{rn_r}$.

Náhodně vylosujeme n_1 , n_2, …, n_r selat téhož plemene, náhodně je rozdělíme na r souborů o n_1 , n_2, …, n_r jedincích, první podrobíme výkrmné dietě č. 1, druhý výkrmné dietě číslo 2 atd. až r-tý podrobíme výkrmné dietě číslo r. Tak dostaneme realizace r nezávislých náhodných výběrů.

Blokové porovnávání: situace je charakterizována jedním náhodným výběrem $(X_{11}, \dots, X_{1r}), \dots, (X_{n1}, \dots, X_{nr})$ z r-rozměrného rozložení.

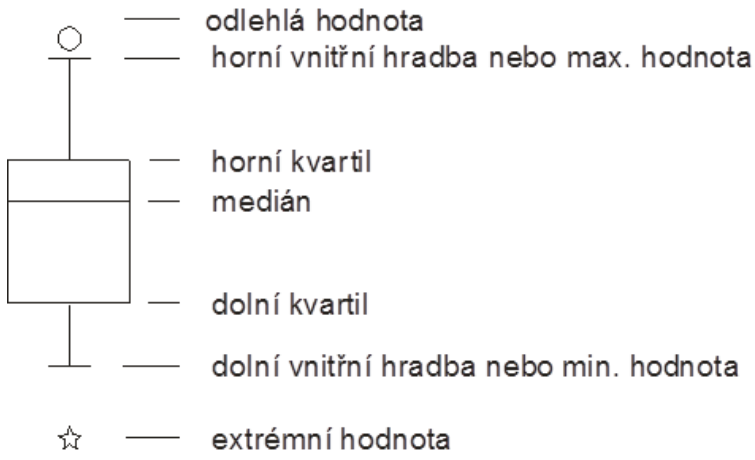
Náhodně vylosujeme n vrhů stejně starých selat téhož plemene, z každého odebereme r sourozenců a náhodně jim přiřadíme první až r-tou výkrmnou dietu. Tak dostaneme realizaci jednoho r-rozměrného náhodného výběru, kde první složka odpovídá první dietě , druhá složka druhé dietě atd. až r-tá složka odpovídá r-té dietě.

Diagnostické grafy

Motivace: Diagnostické grafy slouží především k tomu, aby nám pomohly orientačně posoudit povahu dat a určit směr další statistické analýzy. Při zpracování dat se často předpokládá splnění určitých podmínek. V případě jednoho náhodného výběru je to především normalita (posuzujeme ji pomocí NP plotu či histogramu) a nepřítomnost vybočujících hodnot (odhalí je krabicový diagram).

Krabicový diagram

Umožňuje posoudit symetrii a variabilitu datového souboru a existenci odlehlých či extrémních hodnot.



Způsob konstrukce: Odlehlá hodnota leží mezi vnějšími a vnitřními hradbami, tj. v intervalu $(x_{0,75} + 1{,}5q, x_{0,75} + 3q)$ či v intervalu $(x_{0,25} - 3q, x_{0,25} - 1{,}5q)$. Extrémní hodnota leží za vnějšími hradbami, tj. v intervalu $(x_{0,75} + 3q, \infty)$ či v intervalu $(-\infty, x_{0,25} - 3q)$.

Příklad (1)

U 30 domácností byl zjišťován počet členů. Pro tyto údaje sestrojte krabicový diagram.

Počet členů	1	2	3	4	5	6
Počet domácností	2	6	4	10	5	3

Řešení:

Připomeneme nejprve definici α-kvantilu. Je-li $\alpha \in (0; 1)$, pak α-kvantil x_α je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl α všech dat a na horní úsek obsahující aspoň podíl 1 – α všech dat. Pro výpočet α-kvantilu slouží algoritmus:

$$n\alpha = \begin{cases} \text{celé číslo } c \Rightarrow x_\alpha = \frac{x_{(c)}+x_{(c+1)}}{2} \\ \text{necelé číslo} \Rightarrow \text{zaokrouhlíme nahoru na nejbližší celé číslo } c \Rightarrow x_\alpha = x_{(c)} \end{cases}$$

Pro speciálně zvolená α užíváme názvů:

$x_{0,50}$ – medián, $x_{0,25}$ – dolní kvartil, $x_{0,75}$ – horní kvartil,

$x_{0,1}$, ..., $x_{0,9}$ – decily, $x_{0,01}$, ..., $x_{0,99}$ – percentily.

Jako charakteristika variability slouží kvartilová odchylka: $q = x_{0,75} - x_{0,25}$. V našem případě rozsah souboru $n = 30$.

Výpočty potřebných kvantilů uspořádáme do tabulky.

α	nα	c		x_α
0,25	7,5	8	$x_{(c)}=x_{(8)}$	2
0,50	15	15	$\frac{x_{(15)} + x_{(16)})}{2}$	4
0,75	22,5	23	$x_{(c)}=x_{(23)}$	5

Příklad (2)

Dolní kvartil je 2, tedy aspoň čtvrtina domácností má aspoň dva členy.

Medián je 4, tedy aspoň polovina domácností má aspoň 4 členy.

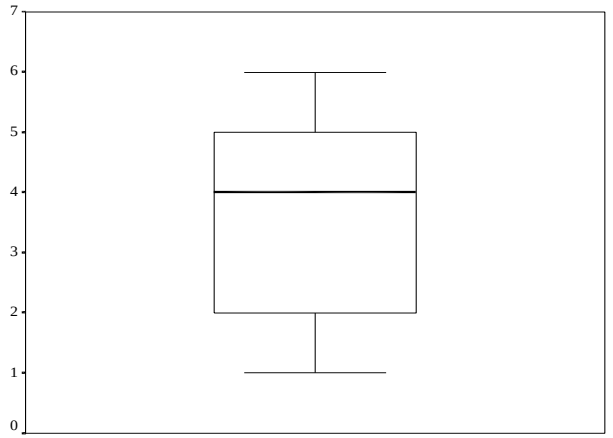
Horní kvartil je 5, tedy aspoň tři čtvrtiny domácností mají aspoň 5 členů.

Vypočteme kvartilovou odchylku: $q = x_{0,75} - x_{0,25} = 5 - 2 = 3$.

Dolní vnitřní hradba: $x_{0,25} - 1,5q = 2 - 1,5.3 = -2,5$

Horní vnitřní hradba: $x_{0,75} + 1,5q = 5 + 1,5.3 = 9,5$

Nakonec sestrojíme krabicový diagram:



Vidíme, že datový soubor vykazuje určitou nesymetrii – medián je posunut směrem k hornímu kvartilu, soubor je tedy záporně sešikmen. V souboru se nevyskytují žádné odlehlé ani extrémní hodnoty.

Pravděpodobnostně – pravděpodobnostní graf (P – P plot)

Umožňuje graficky posoudit, zda data pocházejí z nějakého známého rozložení (např. STATISTICA nabízí 8 typů rozložení: beta, exponenciální, Gumbelovo, gamma, log-normální, normální, Rayleighovo a Weibulovo).

Vypočtou se standardizované hodnoty $z_{(j)} = \frac{x_{(j)} - m}{s}$, j = 1, ..., n. Na vodorovnou osu se vynesou hodnoty teoretické distribuční funkce $\Phi(z_{(j)})$ a na svislou osu hodnoty empirické distribuční funkce $F(z_{(j)}) = j/n$. (Jsou-li některé hodnoty $x_{(1)} \leq \ldots \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.) Pokud se body $(\Phi(z_{(j)}), F(z_{(j)}))$ řadí kolem hlavní diagonály čtverce [0,1] x [0,1], lze usuzovat na dobrou shodu empirického a teoretického rozložení.

Pro posouzení normality dat se používá normální pravděpodobnostní graf (N – P plot): na vodorovnou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \ldots \leq x_{(n)}$ a na svislou osu kvantily u_{α_j}, kde $\alpha_j = \frac{3j-1}{3n+1}$ (jsou-li některé hodnoty stejné, pak za j bereme průměrné pořadí odpovídající takové skupince).

- Pocházejí-li data z normálního rozložení, pak dvojice $(x_{(j)}, u_{\alpha_j})$ budou ležet na přímce.
- Pocházejí-li data z rozložení s kladnou šikmostí, pak dvojice $(x_{(j)}, u_{\alpha_j})$ se budou řadit do konkávní křivky.
- Pocházejí-li data z rozložení se zápornou šikmostí, pak dvojice $(x_{(j)}, u_{\alpha_j})$ se budou řadit do konvexní křivky.

Příklad

Desetkrát nezávisle na sobě byla změřena jistá konstanta. Výsledky měření: 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2. Pomocí normálního pravděpodobnostního grafu posuďte, zda se tato data řídí normálním rozložením.

Řešení:

usp. hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

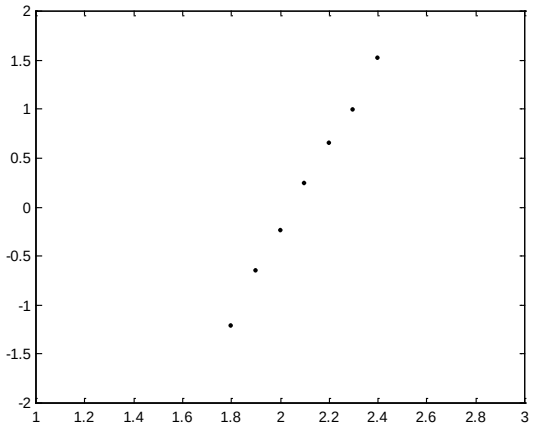
Vektor hodnot průměrného pořadí:

j = (1,5 3 4,5 6,5 8 9 10),

vektor hodnot $\alpha_j = \frac{3j-1}{3n+1} = (0{,}1129; 0{,}2581; 0{,}4032; 0{,}5968; 0{,}7419; 0{,}8387; 0{,}9355)$,

vektor kvantilů $u_{\alpha_j} = (-1{,}2112; -0{,}6493; -0{,}245; 0{,}245; 0{,}6493; 0{,}9892; 1{,}5179)$.

Normální pravděpodobnostní graf:



Protože dvojice $(x_{(j)}, u_{\alpha_j})$ téměř leží na přímce, lze usoudit, že data pocházejí z normálního rozložení.

Histogram

Umožňuje porovnat tvar hustoty četnosti s tvarem hustoty pravděpodobnosti vybraného teoretického rozložení. (Ve STATISTICE je pojem histogramu širší, skrývá se za ním i sloupkový diagram.)

Způsob konstrukce ve STATISTICE: na vodorovnou osu se vynášejí třídicí intervaly (implicitně 10, jejich počet lze změnit, stejně tak i meze třídicích intervalů) či varianty znaku a na svislou osu absolutní nebo relativní četnosti třídicích intervalů či variant. Do histogramu se zakreslí tvar hustoty (či pravděpodobnostní funkce) vybraného teoretického rozložení.

Příklad

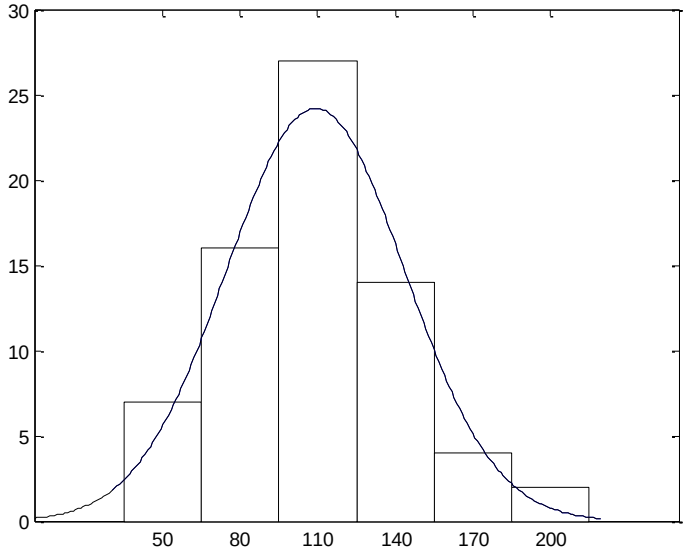
U 70 domácností byly zjišťovány týdenní výdaje na nealkoholické nápoje (v Kč).

Výdaje	(35,65⟩	(65,95⟩	(95,125⟩	(125,155⟩	(125,155⟩	(185,215⟩
Počet dom.	7	16	27	14	4	2

Nakreslete histogram.

Řešení:

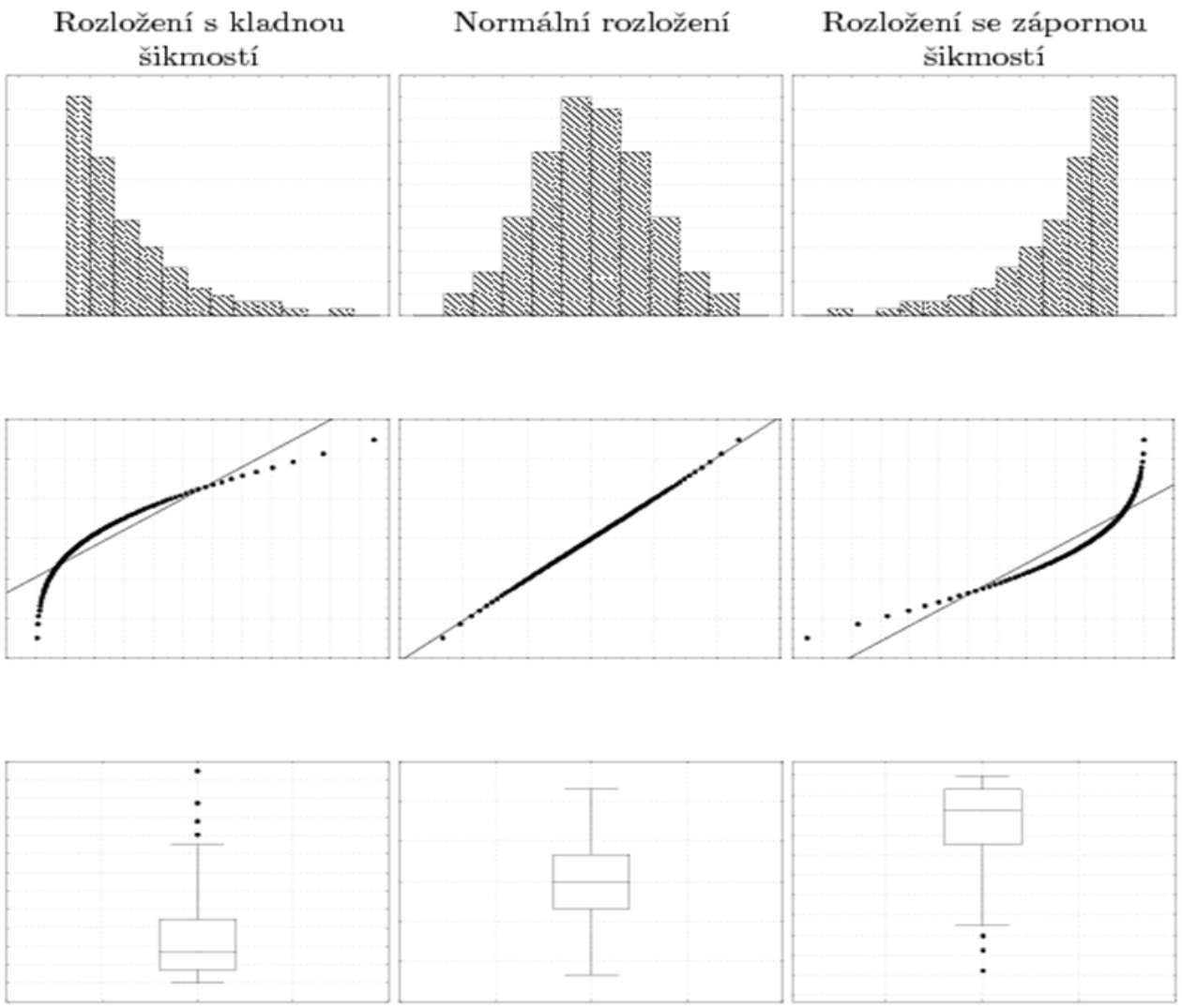
Histogram s proloženou hustotou pravděpodobnosti normálního rozložení:



Vidíme, že tvar histogramu se poněkud odchyluje od tvaru hustoty pravděpodobnosti normálního rozložení. Malé hodnoty jsou četnější než velké – datový soubor je kladně sešikmen.

Vlastnosti rozložení četností datového souboru se projeví ve vzhledu histogramu, N–P plotu a krabicového diagramu, jak vidíme na na následujícím obrázku:

Vlastnosti rozložení četností datového souboru



3. Bodové a intervalové odhady parametrů a parametrických funkcí

Motivace: Vycházíme z náhodného výběru $X_1, \ldots, X_n$ z rozložení $L(\vartheta)$, které závisí na parametru ϑ. Parametr ϑ neznáme a chceme ho odhadnout pomocí daného náhodného výběru (případně chceme odhadnout nějakou parametrickou funkci $h(\vartheta)$). Bodovým odhadem parametrické funkce $h(\vartheta)$ je statistika $T_n = T(X_1, \ldots, X_n)$, která nabývá hodnot blízkých $h(\vartheta)$, ať je hodnota parametru ϑ jakákoliv. Existují různé metody, jak konstruovat bodové odhady (např. metoda momentů či metoda maximální věrohodnosti) a také různé typy bodových odhadů. Omezíme se na odhady nestranné, asymptoticky nestranné a konzistentní.

Intervalovým odhadem parametrické funkce $h(\vartheta)$ rozumíme interval

(D, H), jehož meze jsou statistiky $D = D(X_1, \ldots, X_n)$, $H = H(X_1, \ldots, X_n)$ a který s dostatečně velkou pravděpodobností pokrývá $h(\vartheta)$, ať je hodnota parametru ϑ jakákoliv.

Definice parametrického prostoru a parametrické funkce

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $L(\vartheta)$. Množina všech hodnot, jichž může parametr ϑ nabývat, se nazývá parametrický prostor a značí se Ξ. Libovolná funkce $h(\vartheta)$ se nazývá parametrická funkce.

Definice nestranného odhadu, lepšího nestranného odhadu, posloupnosti asymptoticky nestranných odhadů a konzistentních odhadů

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $L(\vartheta)$, $h(\vartheta)$ je parametrická funkce, $T, T_1, T_2, ...$ jsou statistiky.

a) Řekneme, že statistika T je nestranným odhadem parametrické funkce $h(\vartheta)$, jestliže $\forall \vartheta \in \Xi: E(T) = h(\vartheta)$.
(Význam nestrannosti spočívá v tom, že odhad T nesmí parametrickou funkci $h(\vartheta)$ systematicky nadhodnocovat ani podhodnocovat. Není-li tato podmínka splněna, jde o vychýlený odhad.)

b) Jsou-li T_1, T_2 nestranné odhady téže parametrické funkce $h(\vartheta)$, pak řekneme, že T_1 je lepší odhad než T_2, jestliže $\forall \vartheta \in \Xi: D(T_1) < D(T_2)$.

c) Posloupnost $\{T_n\}_{n=1}^{\infty}$ se nazývá posloupnost asymptoticky nestranných odhadů parametrické funkce $h(\vartheta)$, jestliže $\forall \vartheta \in \Xi: \lim_{n\to\infty} E(T_n) = h(\vartheta)$.
(Význam asymptotické nestrannosti spočívá v tom, že s rostoucím rozsahem výběru klesá vychýlení odhadu.)

d) Posloupnost $\{T_n\}_{n=1}^{\infty}$ se nazývá posloupnost konzistentních odhadů parametrické funkce $h(\vartheta)$, jestliže

$$\forall \vartheta \in \Xi \forall \varepsilon > 0: \lim_{n\to\infty} P(|T_n - h(\vartheta)| > \varepsilon) = 0.$$

(Význam konzistence spočívá v tom, že s rostoucím rozsahem výběru klesá pravděpodobnost, že odhad se bude realizovat „daleko“ od parametrické funkce $h(\vartheta)$.)arametrické funkce $h(\vartheta)$.)

Důsledek: Vztah mezi jednotlivými typy bodových odhadů

Lze dokázat, že z nestrannosti odhadu vyplývá jeho asymptotická nestrannost a z asymptotické nestrannosti vyplývá konzistence, pokud posloupnost rozptylů odhadu konverguje k nule.

Věta o vlastnostech bodových odhadů odvozených z jednoho náhodného výběru (1)

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení se střední hodnotou μ, rozptylem σ^2 a distribuční funkcí $\Phi(x)$. Nechť $n \geq 2$. Označme M_n výběrový průměr, S_n^2 výběrový rozptyl a pro libovolné, ale pevně dané $x \in R$ označme $F_n(x)$ hodnotu výběrové distribuční funkce. Pak pro libovolné hodnoty parametrů μ, σ^2 a libovolnou hodnotu distribuční funkce $\Phi(x)$ platí:

a) M_n je nestranným odhadem μ (tj. $E(M_n) = \mu$) s rozptylem $D(M_n) = \frac{\sigma^2}{n}$,

b) S_n^2 je nestranným odhadem σ^2 (tj. $E(S_n^2) = \sigma^2$) s rozptylem $D(S_n^2) = \frac{\gamma_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$, kde γ_4 je 4. centrální moment

c) pro libovolné, ale pevně dané $x \in R$ je výběrová distribuční funkce $F_n(x)$ nestranným odhadem $\Phi(x)$ (tj. $E(F_n(x)) = \Phi(x)$) s rozptylem $D(F_n(x)) = \frac{\Phi(x)[1-\Phi(x)]}{n}$.

Věta o vlastnostech bodových odhadů odvozených z jednoho náhodného výběru (2)

d) Posloupnost $\{M_n\}_{n=1}^{\infty}$ je posloupnost asymptoticky nestranných a konzistentních odhadů μ,

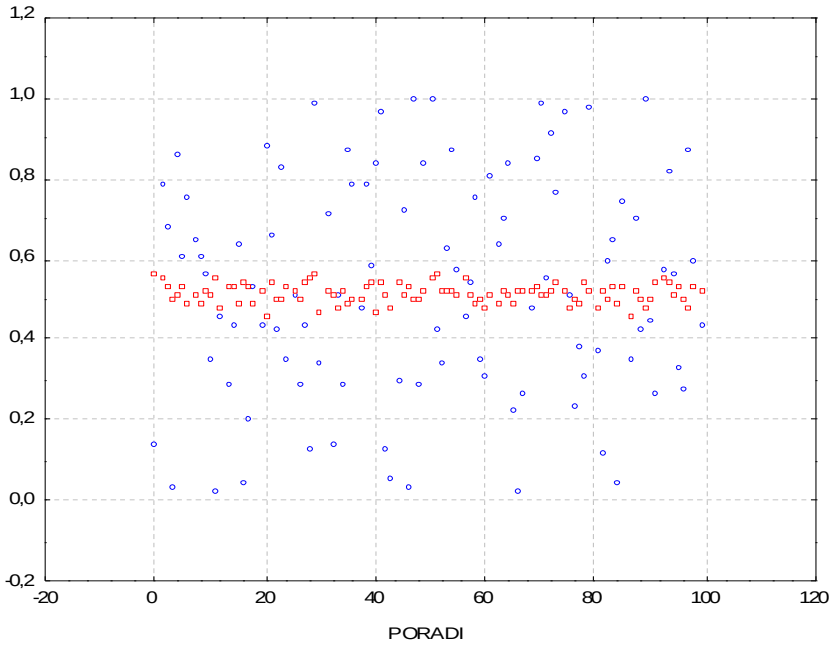
e) $\{S_n^2\}_{n=1}^{\infty}$ je posloupnost asymptoticky nestranných a konzistentních odhadů σ^2,

f) pro libovolné, ale pevně dané $x \in R$ je $\{F_n(x)\}_{n=1}^{\infty}$ posloupnost asymptoticky nestranných a konzistentních odhadů $\Phi(x)$.

Poznámka: Výběrová směrodatná odchylka S není nestranným odhadem směrodatné odchylky σ. To by platilo, pokud S by byla náhodná veličina s degenerovaným rozložením, tj. nabývala by pouze konstantní hodnoty. Pak totiž $D(S) = E(S^2) - [E(S)]^2 = \sigma^2 - [\sigma]^2 = 0$.

Ilustrace (1)

Vlastnosti výběrového průměru a výběrového rozptylu budeme ilustrovat na náhodném výběru rozsahu 100 z rozložení Rs(0,1). V tomto případě $E(X_i) = 1/2$, $D(X_i) = 1/12$, $i = 1, \ldots, 100$. Pomocí systému STATISTICA vygenerujeme pro každou z náhodných veličin $X_1, \ldots, X_{100}$ 100 realizací a uložíme je do proměnných $v_1, \ldots, v_{100}$. Dále vypočítáme průměr a rozptyl těchto realizací, uložíme je do proměnných PRUMER a ROZPTYL. Graficky znázorníme hodnoty některé z proměnných $v_1, \ldots, v_{100}$ (např. v_1) a hodnoty proměnné PRUMER:



Ilustrace (2)

Vidíme, že hodnoty proměnné v_1 kolísají od 0 do 1, zatímco hodnoty proměnné PRUMER se nacházejí v úzkém pásu kolem 1/2.

Dále vypočteme průměr a rozptyl např. proměnné v1 a proměnné PRUMER a dále vypočtěte průměr proměnné ROZPTYL.

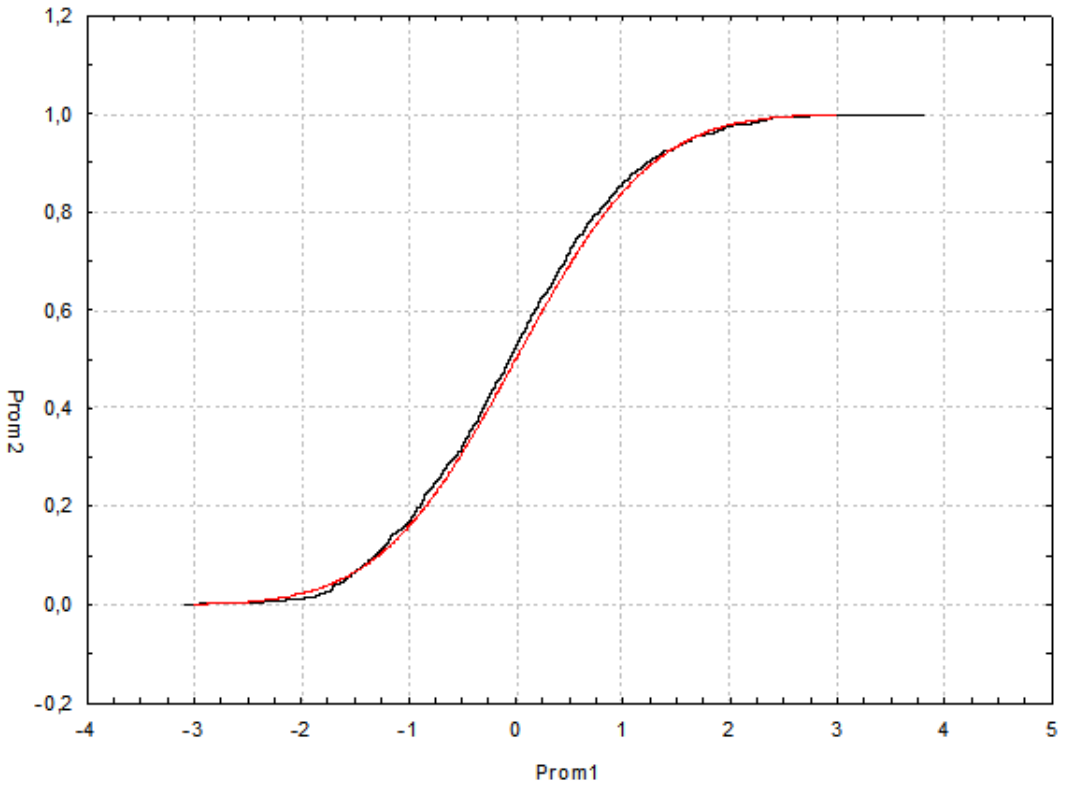
	Popisné statistiky (uniform	
Proměnná	Průměr	Rozptyl
Prom1	0,536605	0,078676
PRUMER	0,503984	0,000783

	Popisné statistiky (uniform
Proměnná	Průměr
ROZPTYL	0,083143

Průměr proměnné v1 by měl být blízký 0,5, rozptyl 1/12 = 0,083. Průměr proměnné PRUMER by se měl blížit 0,5, zatímco rozptyl by měl být n = 100 x menší než 1/12, tj. 0,00083. Dále průměr proměnné ROZPTYL by se měl blížit 1/12 = 0,083.

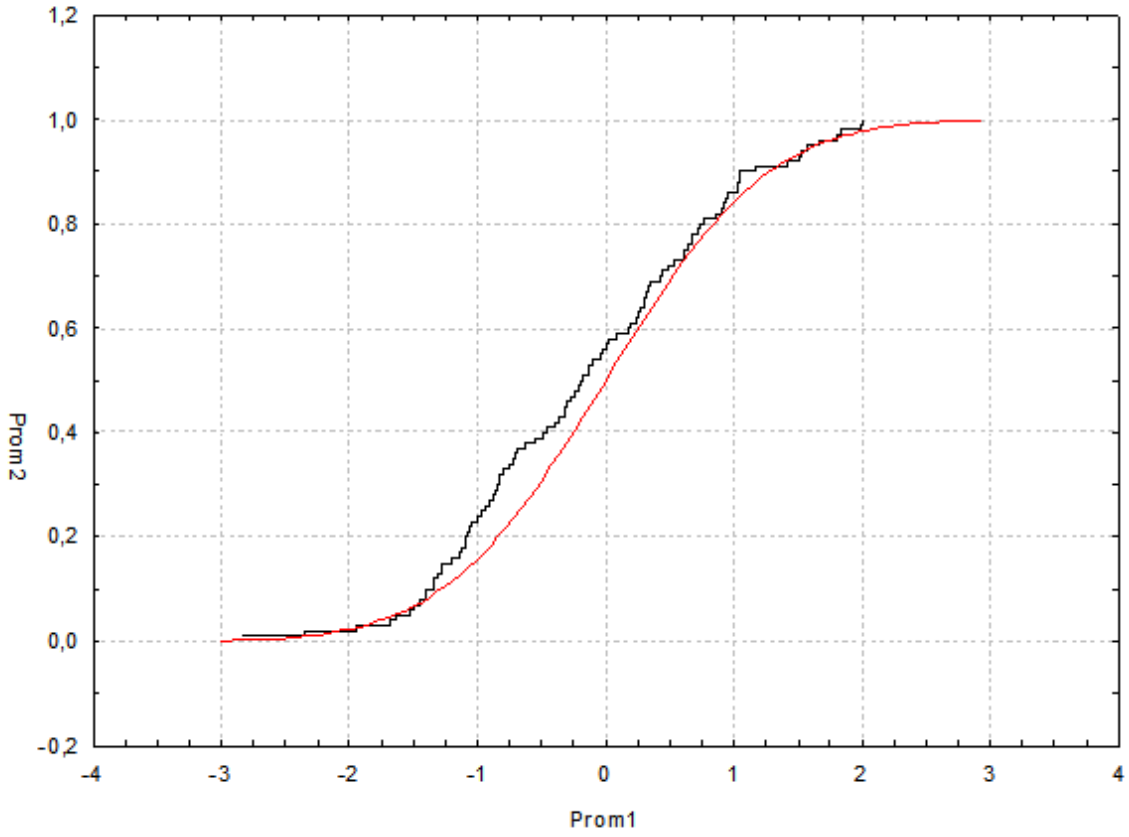
Ilustrace (3)

Nestrannost výběrové distribuční funkce budeme ilustrovat na náhodném výběru rozsahu 1000 z rozložení N(0,1). Získáme výběrovou distribuční funkci tohoto výběru a její graf porovnáme s grafem distribuční funkce náhodné veličiny se standardizovaným normálním rozložením. Graf výběrové distribuční funkce má černou barvu, graf distribuční funkce standardizovaného normálního rozložení má červenou barvu.



Ilustrace (4)

Průběh výběrové distribuční funkce $F_{1000}(x)$ je velmi podobný průběhu distribuční funkce $\Phi(x)$. Pokud bychom postup zopakovali s podstatně menším rozsahem náhodného výběru (např. n = 100), průběh obou funkcí by se lišil výrazněji:



Věta o vlastnostech bodových odhadů odvozených z r ≥ 2 nezávislých náhodných výběrů

Nechť $X_{11}, \dots, X_{1n_1}, \dots, X_{r1}, \dots, X_{rn_r}$ je r stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2, \dots, n_r \geq 2$ z rozložení se středními hodnotami $\mu_1, \dots, \mu_r$ a rozptylem σ^2. Celkový rozsah je $n = \sum_{j=1}^{r} n_j$. Nechť $c_1, \dots, c_r$ jsou reálné konstanty, aspoň jedna nenulová. Označme $\sum_{j=1}^{r} c_j M_j$ lineární kombinaci výběrových průměrů a

$S_*^{\,2} = \frac{\sum_{j=1}^{r} (n_j - 1) S_j^{\,2}}{n - r}$ vážený průměr výběrových rozptylů. Pak pro libovolné hodnoty parametrů $\mu_1, \dots, \mu_r$ a σ^2 platí: $E\left(\sum_{j=1}^{r} c_j M_j\right) = \sum_{j=1}^{r} c_j \mu_j$, $E(S_*^{\,2}) = \sigma^2$.

Znamená to, že lineární kombinace výběrových průměrů $\sum_{j=1}^{r} c_j M_j$ je nestranným odhadem lineární kombinace středních hodnot $\sum_{j=1}^{r} c_j \mu_j$ a vážený průměr výběrových rozptylů $S_*^{\,2} = \frac{\sum_{j=1}^{r} (n_j - 1) S_j^{\,2}}{n - r}$ je nestranným odhadem rozptylu σ^2.

Věta o vlastnostech bodových odhadů odvozených z jednoho dvourozměrného náhodného výběru

Nechť (X_1,Y_1), ..., (X_n,Y_n) je náhodný výběr z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ. Označme S_{12} výběrovou kovarianci a R_{12} výběrový koeficient korelace. Pak pro libovolné hodnoty parametrů σ_{12} a ρ platí:

$E(S_{12}) = \sigma_{12}$,

$E(R_{12}) \approx \rho$ (shoda je vyhovující pro $n \geq 30$).

Znamená to, že výběrová kovariance S_{12} je nestranným odhadem kovariance σ_{12}, avšak výběrový koeficient korelace R_{12} je vychýleným odhadem koeficientu korelace ρ.

Definice intervalu spolehlivosti

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $L(\vartheta)$, $h(\vartheta)$ je parametrická funkce, $\alpha \in (0,1)$, $D = D(X_1, ..., X_n)$, $H = H(X_1, ..., X_n)$ jsou statistiky.

a) Interval (D, H) se nazývá 100(1-α)% (oboustranný) interval spolehlivosti pro parametrickou funkci $h(\vartheta)$, jestliže: $\forall \vartheta \in \Xi$: $P(D < h(\vartheta) < H) \geq 1-\alpha$.

b) Interval (D, ∞) se nazývá 100(1-α)% levostranný interval spolehlivosti pro parametrickou funkci $h(\vartheta)$, jestliže: $\forall \vartheta \in \Xi$: $P(D < h(\vartheta)) \geq 1-\alpha$.

c) Interval (-∞, H) se nazývá 100(1-α)% pravostranný interval spolehlivosti pro parametrickou funkci $h(\vartheta)$, jestliže: $\forall \vartheta \in \Xi$: $P(h(\vartheta) < H) \geq 1-\alpha$.

Číslo α se nazývá riziko (zpravidla $\alpha = 0,05$, méně často 0,1 či 0,01), číslo $1 - \alpha$ se nazývá spolehlivost.

Doporučený postup při konstrukci intervalu spolehlivosti (1)

a) Vyjdeme ze statistiky V, která je nestranným bodovým odhadem parametrické funkce $h(\vartheta)$.

b) Najdeme tzv. pivotovou statistiku W, která vznikne transformací statistiky V, je monotónní funkcí $h(\vartheta)$ a přitom její rozložení je známé a na $h(\vartheta)$ nezávisí. Pomocí známého rozložení pivotové statistiky W najdeme kvantily $w_{\alpha/2}$, $w_{1-\alpha/2}$, takže platí:

$$\forall \vartheta \in \Xi: P(w_{\alpha/2} < W < w_{1-\alpha/2}) \geq 1 - \alpha.$$

c) Nerovnost $w_{\alpha/2} < W < w_{1-\alpha/2}$ převedeme ekvivalentními úpravami na nerovnost $D < h(\vartheta) < H$.

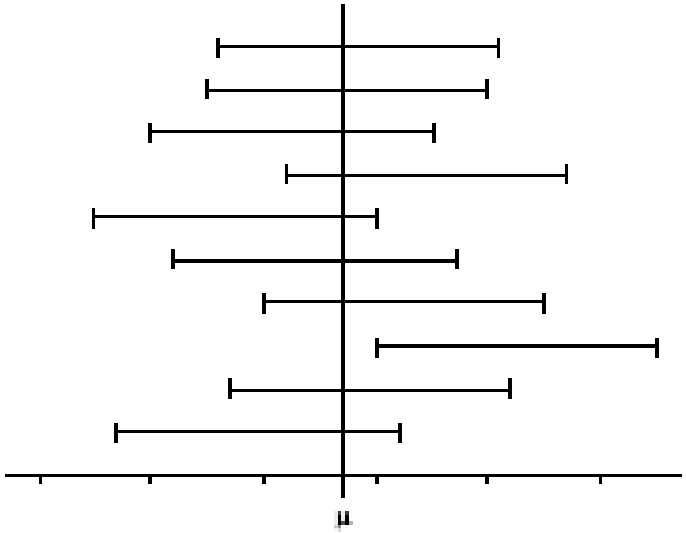
Doporučený postup při konstrukci intervalu spolehlivosti (2)

d) Statistiky D, H nahradíme jejich číselnými realizacemi d, h a získáme tak 100(1-α)% empirický interval spolehlivosti, o němž prohlásíme, že pokrývá $h(\vartheta)$ s pravděpodobností aspoň $1 - \alpha$. (Tvrzení, že (d,h) pokrývá $h(\vartheta)$ s pravděpodobností aspoň $1 - \alpha$ je třeba chápat takto: jestliže mnohonásobně nezávisle získáme realizace $x_1, \ldots, x_n$ náhodného výběru $X_1, \ldots, X_n$ z rozložení $L(\vartheta)$ a pomocí každé této realizace sestrojíme 100(1-α)% empirický interval spolehlivosti pro $h(\vartheta)$, pak podíl počtu těch intervalů, které pokrývají $h(\vartheta)$ k počtu všech sestrojených intervalů bude přibližně $1 - \alpha$.)

(Volba oboustranného, levostranného, nebo pravostranného intervalu závisí na konkrétní situaci. Např. oboustranný interval spolehlivosti použije konstruktér, kterého zajímá dolní i horní hranice pro skutečnou délku μ nějaké součástky. Levostranný interval spolehlivosti použije výkupčí drahých kovů, který potřebuje znát dolní mez pro skutečný obsah zlata μ v kupovaném slitku. Pravostranný interval spolehlivosti použije chemik, který potřebuje znát horní mez pro obsah nečistot μ v analyzovaném vzorku.)

Ilustrace

Jestliže 100x nezávisle na sobě uskutečníme náhodný výběr z rozložení se střední hodnotou μ a pokaždé sestrojíme 95% empirický interval spolehlivosti pro μ, pak přibližně v 95 případech bude ležet parametr μ v intervalech spolehlivosti a asi v 5 případech interval spolehlivosti μ nepokryje.



Příklad (1)

Nechť $X_1, ..., X_n$ je náhodný výběr z $N(\mu,\sigma^2)$, kde $n \geq 2$ a rozptyl σ^2 známe. Sestrojte $100(1-\alpha)\%$ interval spolehlivosti pro neznámou střední hodnotu μ.

Řešení: V tomto případě parametrická funkce $h(\vartheta) = \mu$. Nestranným odhadem střední hodnoty je výběrový průměr $M = \frac{1}{n}\sum_{i=1}^{n} X_i$. Protože M je lineární kombinací normálně rozložených náhodných veličin, bude mít také normální rozložení se střední hodnotou $E(M) = \mu$ a rozptylem $D(M) = \frac{\sigma^2}{n}$. Pivotovou statistikou W bude standardizovaná náhodná veličina $U = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$.

Kvantil $w_{\alpha/2} = u_{\alpha/2} = -u_{1-\alpha/2}$, $w_{1-\alpha/2} = u_{1-\alpha/2}$.

$$\forall \vartheta \in \Xi: 1-\alpha \leq P(-u_{1-\alpha/2} < U < u_{1-\alpha/2}) = P(-u_{1-\alpha/2} < \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} <$$

$$u_{1-\alpha/2}) = P(M - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} < \mu < M + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}).$$

Příklad (2)

Meze 100(1-α)% intervalu spolehlivosti pro střední hodnotu μ při známém rozptylu σ^2 tedy jsou:

D $= M - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}$, H $= M + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}$.

Při konstrukci jednostranných intervalů spolehlivosti se riziko nepůlí, tedy 100(1-α)% levostranný interval spolehlivosti pro μ je $(M - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty)$ a pravostranný je $(-\infty, M + \frac{\sigma}{\sqrt{n}} u_{1-\alpha})$.

Dosadíme-li do vzorců pro dolní a horní mez číselnou realizaci m výběrového průměru M, dostaneme 100(1-α)% empirický interval spolehlivosti.

Příklad

10 krát nezávisle na sobě byla změřena jistá konstanta μ. Výsledky měření byly: 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2. Tyto výsledky považujeme za číselné realizace náhodného výběru $X_1, \ldots, X_{10}$ z rozložení $N(\mu, 0{,}04)$, kde parametr μ neznáme. Najděte 95% empirický interval spolehlivosti pro μ, a to

a) oboustranný,

b) levostranný,

c) pravostranný.

Řešení: $m = 2{,}06$, $\sigma^2 = 0{,}04$, $\sigma = 0{,}2$, $\alpha = 0{,}05$, $u_{0{,}975} = 1{,}96$, $u_{0{,}95} = 1{,}64$.

ad a) $d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} = 2{,}06 - \frac{0{,}2}{\sqrt{10}} 1{,}96 = 1{,}94$

$h = m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} = 2{,}06 + \frac{0{,}2}{\sqrt{10}} 1{,}96 = 2{,}18$

$1{,}94 < \mu < 2{,}18$ s pravděpodobností aspoň 0,95.

ad b) $d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha} = 2{,}06 - \frac{0{,}2}{\sqrt{10}} 1{,}64 = 1{,}96$

$1{,}96 < \mu$ s pravděpodobností aspoň 0,95.

ad c) $h = m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha} = 2{,}06 + \frac{0{,}2}{\sqrt{10}} 1{,}64 = 2{,}16$

$\mu < 2{,}16$ s pravděpodobností aspoň 0,95.

Poznámka o šířce intervalu spolehlivosti

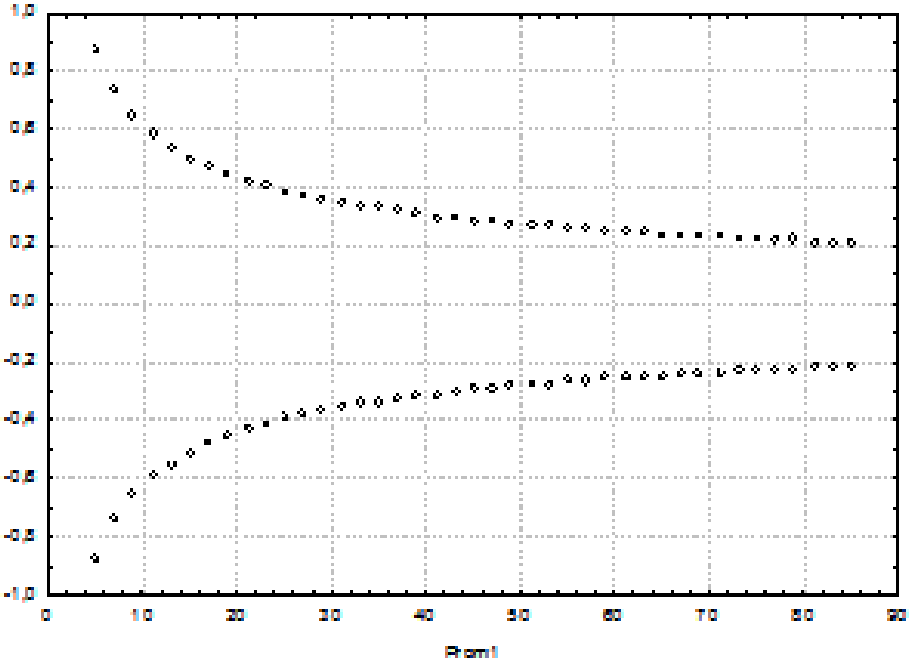
Nechť (d, h) je $100(1-\alpha)\%$ empirický interval spolehlivosti pro $h(\vartheta)$ zkonstruovaný pomocí číselných realizací $x_1, \ldots, x_n$ náhodného výběru $X_1, \ldots, X_n$ z rozložení $L(\vartheta)$.

a) Při konstantním riziku klesá šířka h-d s rostoucím rozsahem náhodného výběru.

b) Při konstantním rozsahu náhodného výběru klesá šířka h-d s rostoucím rizikem.

Ilustrace (1)

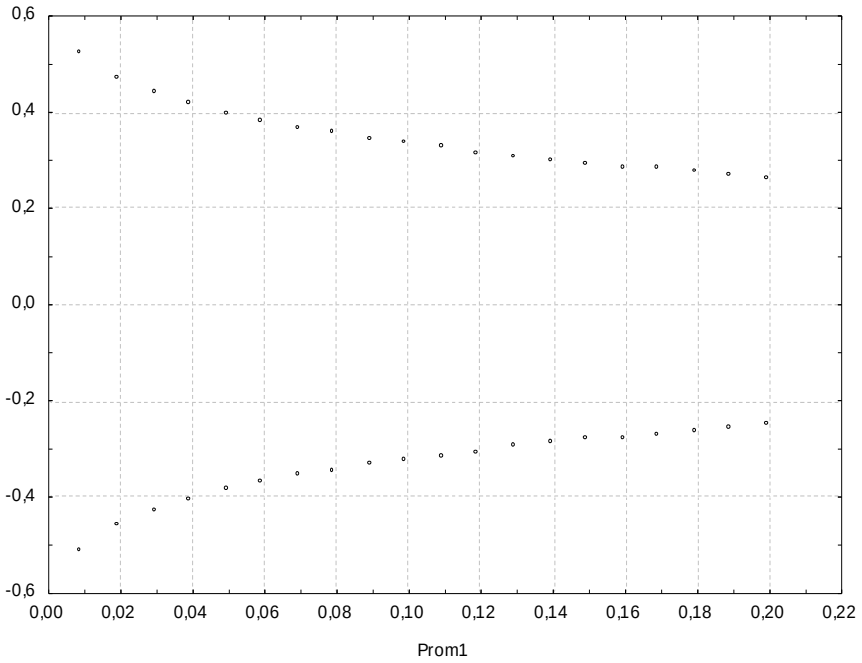
ad a) Grafické znázornění závislosti dolních a horních meze 95% empirických intervalů spolehlivosti pro střední hodnotu normálního rozložení při známém rozptylu na rozsahu náhodného výběru:



Vidíme, že šířka intervalu spolehlivosti klesá se zvětšujícím se rozsahem náhodného výběru, zprvu rychle a pak stále pomaleji.

Ilustrace (2)

ad b) Grafické znázornění závislosti dolních a horních mezí 100(1-α)% empirických intervalů spolehlivosti pro střední hodnotu normálního rozložení při známém rozptylu a konstantním rozsahu výběru na riziku:



Vidíme, že šířka intervalu spolehlivosti s rostoucím rizikem klesá.

Příklad

(Stanovení minimálního rozsahu výběru z normálního rozložení)

Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde σ^2 známe. Jaký musí být minimální rozsah výběru n, aby šířka $100(1-\alpha)\%$ empirického intervalu spolehlivosti pro střední hodnotu μ nepřesáhla číslo Δ?

Řešení: Požadujeme, aby $\Delta \geq h - d = m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} - (m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}) = \frac{2\sigma}{\sqrt{n}} u_{1-\alpha/2}$.

Z této podmínky dostaneme, že $n \geq \frac{4\sigma^2 {u_{1-\alpha/2}}^2}{\Delta^2}$. Za rozsah výběru zvolíme nejmenší přirozené číslo vyhovující této podmínce.

Odvozený vzorec použijeme v této situaci: v příkladu 3.12. (a) se uživateli zdá 95% empirický interval spolehlivosti (1,94; 2,18) pro střední hodnotu μ příliš široký. Přál by si, aby šířka 95% empirického intervalu spolehlivosti nepřesáhla číslo 0,16. Dostáváme tedy $n \geq \frac{4\sigma^2 {u_{1-\frac{\alpha}{2}}}^2}{0{,}16^2} = \frac{4 \cdot 0{,}04 \cdot {u_{0{,}975}}^2}{0{,}16^2} = \frac{4 \cdot 0{,}04 \cdot 1{,}96^2}{0{,}16^2} = 24{,}01$. Podmínku tedy splňuje číslo 25.

Poznámky

Cokoliv z náhodného výběru můžeme pokládat za bodový odhad parametru.

Bodový odhad by měl být nestranný.

Konzistentní odhad – rozptyly se asymptoticky blíží k 0.

4. Metody hledání bodových odhadů parametrů. Úvod do testování hypotéz.

Motivace: Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $L(\vartheta)$, které závisí na parametru ϑ. Úkolem je najít statistiku $T = T(X_1, ..., X_n)$, která nabývá hodnot blízkých parametru ϑ resp. parametrické funkci $h(\vartheta)$, ať je hodnota parametru ϑ jakákoliv. Seznámíme se se dvěma metodami hledání bodových odhadů, a to metodou maximální věrohodnosti a metodou momentů.

Definice maximálně věrohodného odhadu

Nechť $X_1, ..., X_n$ je náhodný výběr z diskrétního rozložení (je popsáno pravděpodobnostní funkcí $\pi(x; \vartheta)$ resp. ze spojitého rozložení (je popsáno hustotou $\varphi(x; \vartheta)$).

Simultánní pravděpodobnostní funkce resp. simultánní hustota náhodného vektoru $(X_1, ..., X_n)$ je $\pi(x_1; \vartheta)\ldots\pi(x_n; \vartheta)$ resp. $\varphi(x_1; \vartheta)\ldots\varphi(x_n; \vartheta)$. Pro pevně dané $\mathbf{x} = (x_1, ..., x_n)$ zavedeme věrohodnostní funkci $L(\vartheta) = \prod_{i=1}^{n} \pi(x_i; \vartheta)$ v diskrétním případě resp. $L(\vartheta) = \prod_{i=1}^{n} \varphi(x_i; \vartheta)$ ve spojitém případě.

Statistika $\hat{\vartheta}(X)$, která má tu vlastnost, že $\forall \vartheta \in \Xi: L(\hat{\vartheta}) \geq L(\vartheta)$, se nazývá maximálně věrohodný odhad parametru ϑ.

(Místo věrohodnostní funkce $L(\vartheta)$ používáme logaritmickou věrohodnostní funkci $\ln L(\vartheta)$.)

Definice věrohodnostních rovnic

Nechť $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$. Logaritmickou věrohodnostní funkci $\ln L(\vartheta)$ parciálně derivujeme podle $\vartheta_1, \ldots, \vartheta_k$ a derivace položíme rovny 0:

$$\frac{\partial \ln L(\vartheta_1, \ldots, \vartheta_k)}{\partial \vartheta_i} = 0,\ \mathrm{i} = 1, 2, \ldots, \mathrm{k}.$$

Dostaneme systém věrohodnostních rovnic. Jeho řešením je maximálně věrohodný odhad parametru ϑ: $\hat{\vartheta}(X) = (\hat{\vartheta}_1(X), \ldots, \hat{\vartheta}_k(X))$.

Příklad (1)

(Maximálně věrohodný odhad v diskrétním skalárním případě)

Nechť $X_1, \ldots, X_n$ je náhodný výběr z alternativního rozložení $A(\vartheta)$. Metodou maximální věrohodnosti najděte odhad parametru ϑ.

Řešení:

$$X \sim A(\vartheta) \Rightarrow \pi(x) = \begin{cases} \vartheta^{x_i}(1-\vartheta)^{1-x_i} \text{ pro } x = 0, 1 \\ 0 \text{ jinak} \end{cases}$$

Věrohodnostní funkce:

$$L(\vartheta) = \begin{cases} \prod_{i=1}^{n} \vartheta^{x_i}(1-\vartheta)^{1-x_i} = \vartheta^{\sum_{i=1}^{n} x_i}(1-\vartheta)^{n-\sum_{i=1}^{n} x_i} & \text{pro } x_i = 0,1 \\ 0 \text{ jinak} \end{cases}$$

Příklad (2)

(Maximálně věrohodný odhad v diskrétním skalárním případě)

Logaritmická funkce věrohodnosti:

$$\ln L(\vartheta) = \ln\vartheta \cdot \sum_{i=1}^{n} x_i + (n - \sum_{i=1}^{n} x_i)\ln(1-\vartheta)$$

Věrohodnostní rovnice:

$$\frac{d\ln L(\vartheta)}{d\vartheta} = \frac{\sum_{i=1}^{n} x_i}{\vartheta} - \frac{n - \sum_{i=1}^{n} x_i}{1-\vartheta} = 0 \Rightarrow$$

$$\Rightarrow (1-\vartheta)\sum_{i=1}^{n} x_i - \vartheta(n - \sum_{i=1}^{n} x_i) = 0 \Rightarrow$$

$$\Rightarrow \sum_{i=1}^{n} x_i - \vartheta\sum_{i=1}^{n} x_i - n\vartheta + \vartheta\sum_{i=1}^{n} x_i = 0 \Rightarrow \hat{\vartheta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Maximálně věrohodným odhadem parametru ϑ alternativního rozložení $A(\vartheta)$ je tedy statistika $\hat{\vartheta}(X) = M$, tj. výběrový průměr.

Příklad (1)

(Maximálně věrohodný odhad ve spojitém vektorovém případě)

Nechť $X_1, \ldots, X_n$ je náhodný výběr z normálního rozložení $N(\mu, \sigma^2)$. Metodou maximální věrohodnosti najděte odhad vektorového parametru $\vartheta = (\mu, \sigma^2)$.

Řešení: $X \sim N(\mu, \sigma^2) \Rightarrow \varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Věrohodnostní funkce:

$$L(\vartheta) = \prod_{i=1}^{n} \phi(x_i; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)}{\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Logaritmická funkce věrohodnosti: $\ln L(\vartheta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$.

Věrohodnostní rovnice:

$$\frac{\partial \ln L(\vartheta)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

$$\frac{\partial \ln L(\vartheta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

Příklad (2)

(Maximálně věrohodný odhad ve spojitém vektorovém případě)

Z první rovnice plyne $\sum_{i=1}^{n} x_i - n\mu = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Maximálně věrohodným odhadem parametru μ je tedy statistika $\hat{\mu}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i = M$, tj. výběrový průměr.

Z druhé rovnice plyne $-n\sigma^2 + \sum_{i=1}^{n}(x_i - \mu)^2 = 0$. Za μ dosadíme odhad $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = m$ a získáme $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$. Maximálně věrohodným odhadem parametru σ^2 je tedy statistika $\hat{\sigma}^2(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2$.

Definice momentového odhadu

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $L(\vartheta)$, $\vartheta \in \Xi$. Předpokládáme, že existuje prvních k počátečních momentů $\mu_r' = E(X^r)$ rozložení $L(\vartheta)$, r = 1, 2, …, k.

Označme $M_r' = \frac{1}{n} \sum_{i=1}^{n} X_i^r$ výběrové počáteční momenty, r = 1, 2, …, k.

Statistika $\hat{\vartheta}(X)$, která je řešením systému momentových rovnic

$\mu_r' = M_r'$, r = 1, 2, …, k, se nazývá momentový odhad parametru ϑ.

Příklad

Nechť $X_1, ..., X_n$ je náhodný výběr z geometrického rozložení $Ge(\vartheta)$. Metodou momentů najděte odhad parametru ϑ.

Řešení: $X \sim Ge(\vartheta) \Rightarrow \pi(x) = \begin{cases} (1-\vartheta)^x \vartheta & \text{pro x} = 0, 1, ... \\ 0 & \text{jinak} \end{cases}$.

Lze odvodit, že $E(X) = \frac{1-\vartheta}{\vartheta}$.

Momentová rovnice: $\mu_1' = M_1'$, tj. $\frac{1-\vartheta}{\vartheta} = M \Rightarrow 1 - \vartheta = \vartheta M \Rightarrow \hat{\vartheta}(X) = \frac{1}{1+M}$.

Testování hypotéz

Motivace: Častým úkolem statistika je na základě dat ověřit předpoklady o parametrech nebo typu rozložení, z něhož pochází náhodný výběr. Takovému předpokladu se říká nulová hypotéza. Nulová hypotéza vyjadřuje nějaký teoretický předpoklad, často skeptického rázu a uživatel ji musí stanovit předem, bez přihlédnutí k datovému souboru. Proti nulové hypotéze stavíme alternativní hypotézu, která říká, co platí, když neplatí nulová hypotéza. Alternativní hypotéza je formulována tak, aby mohla platit jenom jedna z těchto dvou hypotéz. Pravdivost alternativní hypotézy by znamenala objevení nějakých nových skutečností nebo zásadnější změnu v dosavadních představách.

Např. výzkumník by chtěl na základě dat prověřit tezi (nový objev), že pasivní kouření škodí zdraví. Jako nulovou hypotézu tedy položí tvrzení, že pasivní kouření neškodí zdraví a proti nulové hypotéze postaví alternativní, že pasivní kouření škodí zdraví.

Testováním hypotéz se myslí rozhodovací postup, který je založen na daném náhodném výběru a s jehož pomocí rozhodneme o zamítnutí či nezamítnutí nulové hypotézy.

Definice nulové a alternativní hypotézy

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $L(\vartheta)$, kde parametr $\vartheta \in \Xi$ neznáme. Nechť $h(\vartheta)$ je parametrická funkce a c daná reálná konstanta.

a) Oboustranná alternativa: Tvrzení H_0: $h(\vartheta) = c$ se nazývá jednoduchá nulová hypotéza. Proti nulové hypotéze postavíme složenou oboustrannou alternativní hypotézu H_1: $h(\vartheta) \neq c$.

b) Levostranná alternativa: Tvrzení H_0: $h(\vartheta) \geq c$ se nazývá složená pravostranná nulová hypotéza. Proti jednoduché nebo složené pravostranné nulové hypotéze postavíme složenou levostrannou alternativní hypotézu H_1: $h(\vartheta) < c$.

c) Pravostranná alternativa: Tvrzení H_0: $h(\vartheta) \leq c$ se nazývá složená levostranná nulová hypotéza. Proti jednoduché nebo složené levostranné nulové hypotéze postavíme složenou pravostrannou

Testování nulové a alternativní hypotézy

Testováním H_0 proti H_1 rozumíme rozhodovací postup založený na náhodném výběru $X_1, ..., X_n$, s jehož pomocí zamítneme či nezamítneme platnost nulové hypotézy.

(Volba alternativní hypotézy není libovolná, ale vyplývá z konkrétní situace. Např. při současné technologii je pravděpodobnost vyrobení zmetku $\vartheta = 0{,}01$.

a) Po rekonstrukci výrobní linky byla obnovena výroba, přičemž technologie zůstala stejná. Chceme ověřit, zda se změnila kvalita výrobků. Testujeme H_0: $\vartheta = 0{,}01$ proti H_1: $\vartheta \neq 0{,}01$.

b) Byly provedeny změny v technologii výroby s cílem zvýšit kvalitu. V tomto případě tedy testujeme H_0: $\vartheta = 0{,}01$ proti H_1: $\vartheta < 0{,}01$.

c) Byly provedeny změny v technologii výroby s cílem snížit náklady. V této situaci testujeme H_0: $\vartheta = 0{,}01$ proti H_1: $\vartheta > 0{,}01$.)

Definice chyby 1. a 2. druhu

Při testování H_0 proti H_1 se můžeme dopustit jedné ze dvou chyb: chyba 1. druhu spočívá v tom, že H_0 zamítneme, ač ve skutečnosti platí a chyba 2. druhu spočívá v tom, že H_0 nezamítneme, ač ve skutečnosti neplatí. Situaci přehledně znázorňuje tabulka:

skutečnost	rozhodnutí	
	H_0 nezamítáme	H_0 zamítáme
H_0 platí	správné rozhodnutí	**chyba 1. druhu**
H_0 neplatí	**chyba 2. druhu**	správné rozhodnutí

Pravděpodobnost chyby 1. druhu se značí **α** a nazývá se **hladina významnosti testu** (většinou bývá α = 0,05, méně často 0,1 či 0,01). Pravděpodobnost chyby 2. druhu se značí **β**. Číslo **1–β** se nazývá **síla testu** a vyjadřuje pravděpodobnost, že bude H_0 zamítnuta za předpokladu, že neplatí. Obvykle se snažíme, aby síla testu byla aspoň 0,8. Obě hodnoty, α i 1–β, závisí na velikosti efektu, který se snažíme detekovat. Čím drobnější efekt, tím musí být větší rozsah náhodného výběru.

Poznámka: Testování nulové hypotézy proti alternativní hypotéze třemi způsoby.

Testování nulové hypotézy proti alternativní hypotéze lze provést pomocí kritického oboru, pomocí intervalu spolehlivosti nebo pomocí p-hodnoty.

Definice testového kritéria, oboru nezamítnutí, kritického oboru a kritických hodnot

Statistika $T_0 = T_0(X_1, ..., X_n)$ se nazývá testovým kritériem. Množina všech hodnot, jichž může testové kritérium nabýt, se rozpadá na obor nezamítnutí nulové hypotézy (značí se V) a obor zamítnutí nulové hypotézy (značí se W a nazývá se též kritický obor). Tyto dva obory jsou odděleny kritickými hodnotami (pro danou hladinu významnosti α je lze najít ve statistických tabulkách).

Rozhodnutí o nulové hypotéze pomocí realizace testového kritéria v oboru nezamítnutí či v kritickém oboru

Jestliže číselná realizace t_0 testového kritéria T_0 padne do kritického oboru W, pak nulovou hypotézu zamítáme na hladině významnosti α a znamená to skutečné vyvrácení testované hypotézy. Jestliže t_0 padne do oboru nezamítnutí V, pak jde o pouhé mlčení, které platnost nulové hypotézy jenom připouští.

Stanovení kritického oboru v případě oboustranné alternativy, levostranné alternativy, pravostranné alternativy

Kritický obor v případě oboustranné alternativy má tvar

$W = (t_{\min}, K_{\alpha/2}(T)\rangle \cup \langle K_{1-\alpha/2}(T), t_{\max})$, kde $K_{\alpha/2}(T)$ a $K_{1-\alpha/2}(T)$ jsou kvantily rozložení, jímž se řídí testové kritérium T_0, je-li nulová hypotéza pravdivá.

Kritický obor v případě levostranné alternativy má tvar:

$W = (t_{\min}, K_{\alpha}(T)\rangle$.

Kritický obor v případě pravostranné alternativy má tvar:

$W = \langle K_{1-\alpha}(T), t_{\max})$.

Doporučený postup při testování nulové hypotézy proti alternativní hypotéze pomocí kritického oboru

- Stanovíme nulovou hypotézu a alternativní hypotézu. Přitom je vhodné zvolit jako alternativní hypotézu ten předpoklad, jehož přijetí znamená závažné opatření a mělo by k němu dojít jen s malým rizikem omylu.
- Zvolíme hladinu významnosti α. Zpravidla volíme $\alpha = 0{,}05$, méně často 0,1 nebo 0,01.
- Najdeme vhodné testové kritérium a na základě zjištěných dat vypočítáme jeho realizaci.
- Jestliže realizace testového kritéria padla do kritického oboru, nulovou hypotézu zamítáme na hladině významnosti α a přijímáme alternativní hypotézu. V opačném případě nulovou hypotézu nezamítáme na hladině významnosti α.
- Na základě rozhodnutí, které jsme učinili o nulové hypotéze, učiníme nějaké konkrétní opatření, např. seřídíme obráběcí stroj.
- (Při testování hypotéz musíme mít k dispozici odpovídající nástroje, nejlépe vhodný statistický software. Nemáme-li ho k dispozici, musíme znát příslušné vzorce. Dále potřebujeme statistické tabulky a kalkulačku.)

Testování nulové hypotézy proti alternativní hypotéze pomocí 100(1-α)% empirického intervalu spolehlivosti pro parametrickou funkci h(ϑ)

Sestrojíme 100(1-α)% empirický interval spolehlivosti pro parametrickou funkci h(ϑ). Pokryje-li tento interval hodnotu c, pak H_0 nezamítáme na hladině významnosti α, v opačném případě H_0 zamítáme na hladině významnosti α.

Pro test H_0 proti oboustranné alternativě sestrojíme oboustranný interval spolehlivosti.

Pro test H_0 proti levostranné alternativě sestrojíme pravostranný interval spolehlivosti.

Pro test H_0 proti pravostranné alternativě sestrojíme levostranný interval spolehlivosti.

Testování nulové hypotézy proti alternativní hypotéze pomocí p-hodnoty

p-hodnota udává nejnižší možnou hladinu významnosti pro zamítnutí nulové hypotézy. Je to riziko, že bude zamítnuta H_0 za předpokladu, že platí (riziko planého poplachu). Jestliže p-hodnota $\leq \alpha$, pak H_0 zamítáme na hladině významnosti α, je-li p-hodnota $> \alpha$, pak H_0 nezamítáme na hladině významnosti α.

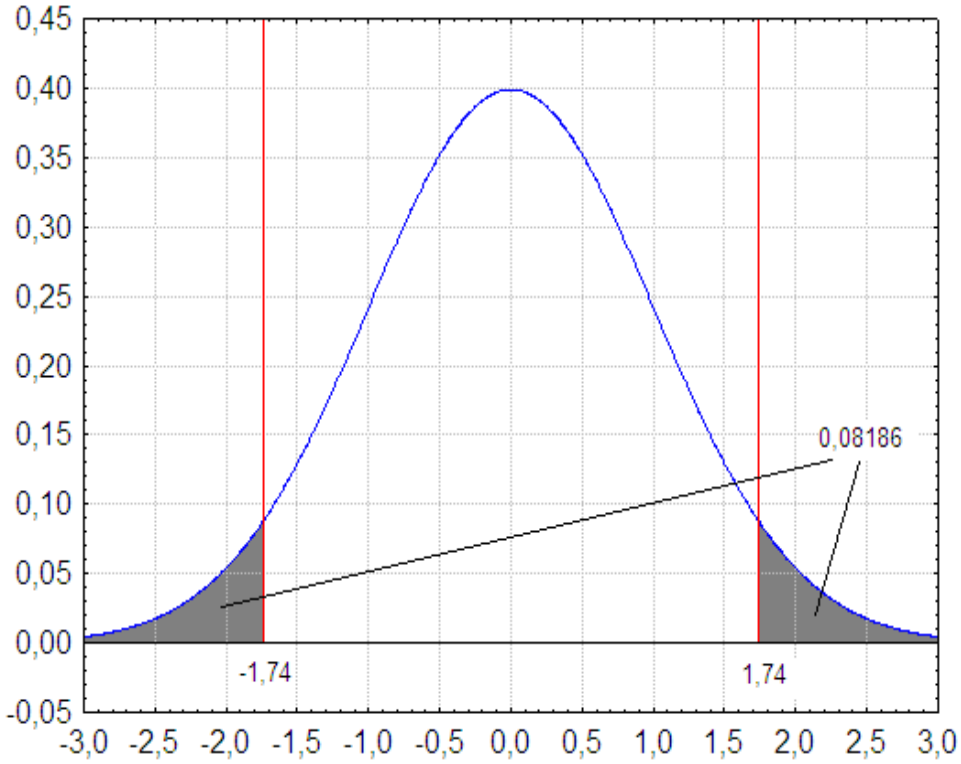
Způsob výpočtu p-hodnoty:

- Pro oboustrannou alternativu $p = 2 \min\{P(T_0 \leq t_0), P(T_0 \geq t_0)\}$.
- Pro levostrannou alternativu $p = P(T_0 \leq t_0)$.
- Pro pravostrannou alternativu $p = P(T_0 \geq t_0)$.

(p-hodnota vyjadřuje pravděpodobnost, s jakou číselné realizace $x_1, ..., x_n$ náhodného výběru $X_1, ..., X_n$ podporují H_0, je-li pravdivá. Statistické programové systémy poskytují ve svých výstupech p-hodnotu. Její výpočet vyžaduje znalost distribuční funkce rozložení, kterým se řídí testové kritérium T_0, je-li H_0 pravdivá. Vzhledem k tomu, že v běžných statistických tabulkách jsou uvedeny pouze hodnoty distribuční funkce standardizovaného normálního rozložení, bez použití speciálního software jsme schopni vypočítat p-hodnotu pouze pro test hypotézy o střední hodnotě normálního rozložení při známém rozptylu.)

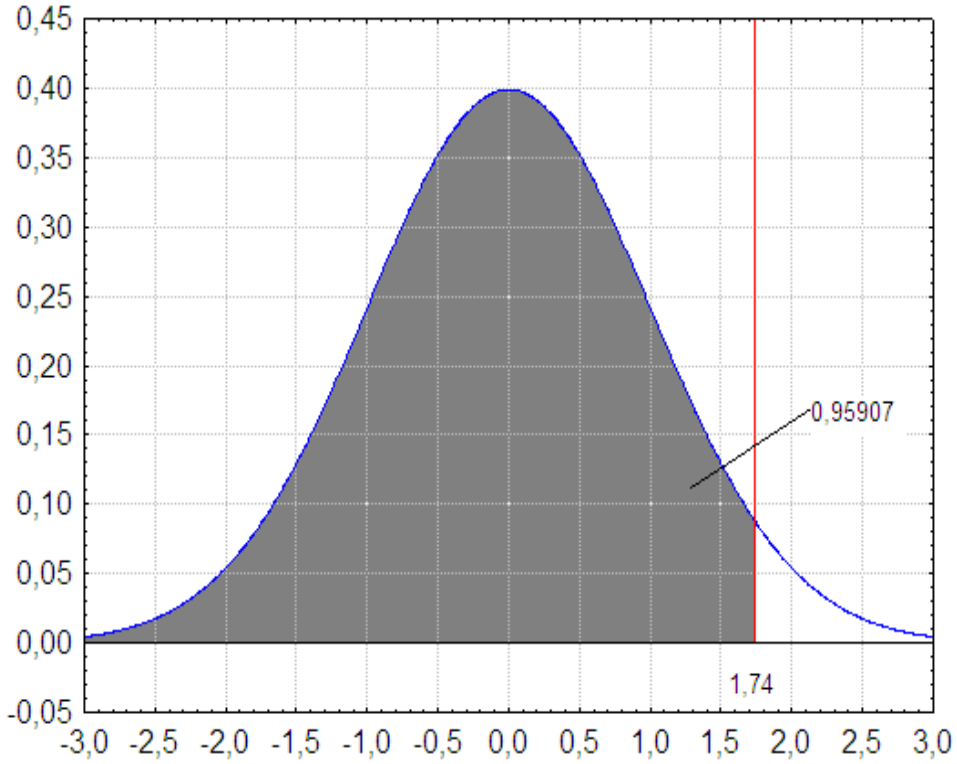
Ilustrace významu p-hodnoty (1)

Oboustranný test:



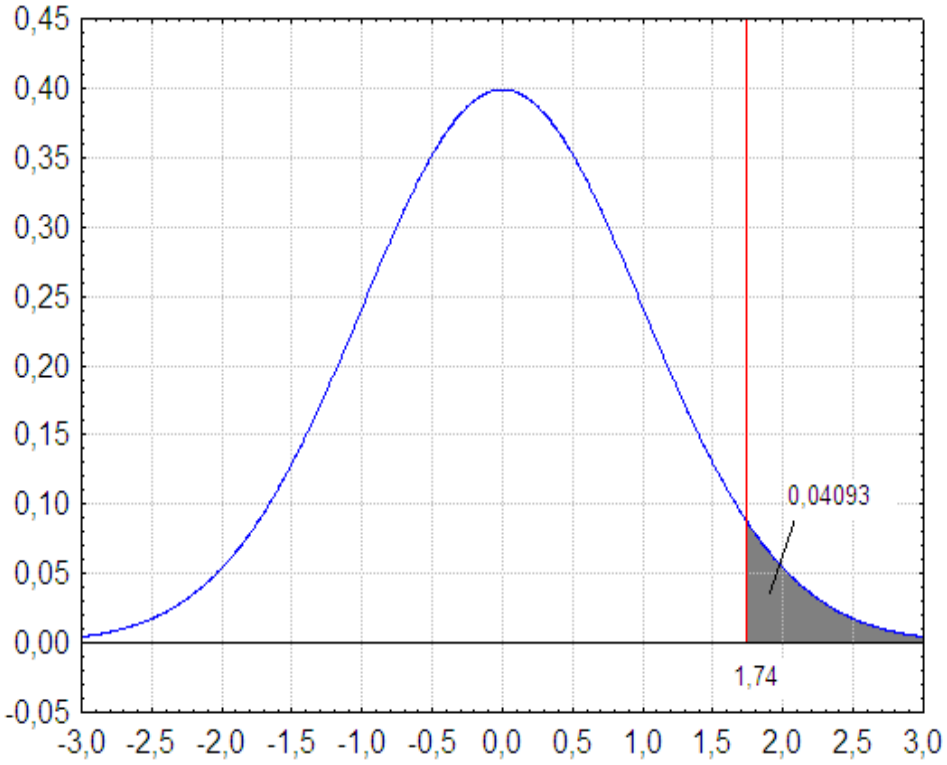
Ilustrace významu p-hodnoty (2)

Levostranný test:



Ilustrace významu p-hodnoty (3)

Pravostranný test:



Příklad (1)

Nechť $X_1, \ldots, X_{400}$ je náhodný výběr z $N(\mu,0{,}01)$. Je známo, že výběrový průměr se realizoval hodnotou 0,01. Na hladině významnosti 0,05 testujte hypotézu H_0: $\mu = 0$ proti pravostranné alternativě H_1: $\mu > 0$

a) pomocí intervalu spolehlivosti

b) pomocí kritického oboru

c) pomocí p-hodnoty.

Řešení:

ad a) Při testování nulové hypotézy proti pravostranné alternativě používáme levostranný interval spolehlivosti.

$$d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha} = 0{,}01 - \frac{0{,}1}{\sqrt{400}} u_{0{,}95} = 0{,}01 - \frac{0{,}1}{20} 1{,}64485 = 0{,}0018$$

Protože číslo c = 0 neleží v intervalu $(0{,}0018; \infty)$, H_0 zamítáme na hladině

Příklad (2)

ad b) Vypočteme realizaci testové statistiky: $t_0 = \frac{m-c}{\frac{\sigma}{\sqrt{n}}} = \frac{0{,}01-0}{\frac{0{,}1}{\sqrt{400}}} = \frac{0{,}01 \cdot 20}{0{,}1} = 2$

Stanovíme kritický obor: $W = \langle u_{1-\alpha}, \infty) = \langle u_{0{,}95}, \infty) = \langle 1{,}64485, \infty)$

Protože testová statistika se realizuje v kritickém oboru, H_0 zamítáme na hladině významnosti 0,05.

ad c) Při testování nulové hypotézy proti pravostranné alternativě se p-hodnota počítá podle vzorce: $p = P(T_0 \geq t_0)$. V našem případě: $p = P(T_0 \geq 2) = 1 - \Phi(2) = 1 - 0{,}97725 = 0{,}02275$. Protože p-hodnota je menší než hladina významnosti 0,05, H_0 zamítáme na hladině významnosti 0,05.

5. Porovnání empirického a teoretického rozložení

Motivace: Možnost použití statistických testů je podmíněna nějakými předpoklady o datech. Velmi často je to předpoklad o typu rozložení, z něhož získaná data pocházejí. Mnoho testů je založeno na předpokladu normality.

Opomíjení předpokladů o typu rozložení může v praxi vést i ke zcela zavádějícím výsledkům, proto je nutné věnovat tomuto problému patřičnou pozornost.

Popis Kolmogorovova – Smirnovova testu a jeho Lilieforsovy varianty

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z rozložení s distribuční funkcí $\Phi(x)$. Nechť $F_n(x) = \frac{1}{n} card\{i; X_i \leq x\}$ je výběrová distribuční funkce. Testovou statistikou je statistika $D_n = \sup_{-\infty<x<\infty} |F_n(x) - \Phi(x)|$. Nulovou hypotézu zamítáme na hladině významnosti α, když $D_n \geq D_n(\alpha)$, kde $D_n(\alpha)$ je tabelovaná kritická hodnota.

(Pro $n \geq 30$ lze $D_n(\alpha)$ aproximovat výrazem $\sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}$.)

Upozornění: Nulová hypotéza musí specifikovat distribuční funkci zcela přesně, včetně všech jejích případných parametrů. Např. K-S test lze použít pro testování hypotézy, že náhodný výběr $X_1, ..., X_n$ pochází z rozložení Rs(0,1), což se využívá při testování generátorů náhodných čísel.

Lilieforsova modifikace Kolmogorovova – Smirnovova testu

Nechť nulová hypotéza tvrdí, že náhodný výběr pochází z normálního rozložení, jehož parametry μ a σ^2 neznáme. Tyto parametry musíme odhadnout z dat. Tím se změní rozložení testové statistiky D_n. V takovém případě jde o Lilieforsovu modifikaci Kolmogorovova – Smirnovova testu. Příslušné modifikované kvantily byly určeny pomocí simulačních studií.

Poznámka ke K-S testu ve STATISTICE: Test normality poskytuje hodnotu testové statistiky (ozn. d) a dvě p-hodnoty. První se vztahuje k případu, kdy μ a σ^2 známe předem, druhá (ozn. Liliefors p) se vztahuje k případu, kdy μ a σ^2 neznáme. Objeví-li se ve výstupu p = n.s. (tj. non significant), pak hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

Příklad (1)

Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí Lilieforsovy varianty K- S testu ověřte na hladině významnosti 0,05, zda tato data pocházejí z normálního rozložení.

Řešení: Odhadem střední hodnoty je výběrový průměr m = 11, odhadem rozptylu je výběrový rozptyl $s^2 = 10$. Uspořádaný náhodný výběr je (8, 9, 10, 12, 16). Vypočteme hodnoty výběrové distribuční funkce:

$$x < 8: F_5(x) = 0,\ 8 \le x < 9: F_5(x) = \frac{1}{5} = 0{,}2\ ,\ 9 \le x < 10: F_5(x) = \frac{2}{5} = 0{,}4,$$

$$10 \le x < 12: F_5(x) = \frac{3}{5} = 0{,}6,\ 12 \le x < 16: F_5(x) = \frac{4}{5} = 0{,}8,\ x \ge 16: F_5(x) = 1$$

Hodnoty teoretické distribuční funkce $\Phi_T(x)$ v bodech 8, 9, 10, 12, 16:

$$\Phi_T(8) = \Phi\left(\frac{8-11}{\sqrt{10}}\right) = \Phi(-0{,}95) = 1 - \Phi(0{,}95) = 1 - 0{,}82894 = 0{,}17106$$

$$\Phi_T(9) = \Phi\left(\frac{9-11}{\sqrt{10}}\right) = \Phi(-0{,}63) = 1 - \Phi(0{,}63) = 1 - 0{,}73565 = 0{,}26435$$

$$\Phi_T(10) = \Phi\left(\frac{10-11}{\sqrt{10}}\right) = \Phi(-0{,}32) = 1 - \Phi(0{,}32) = 1 - 0{,}62552 = 0{,}37448$$

Příklad (2)

$$\Phi_T(12) = \Phi\left(\frac{12-11}{\sqrt{10}}\right) = \Phi(0{,}32) = 0{,}62552$$

$$\Phi_T(16) = \Phi\left(\frac{16-11}{\sqrt{10}}\right) = \Phi(1{,}58) = 0{,}94295$$

(Φ je distribuční funkce rozložení N(0,1).)

Rozdíly mezi výběrovou distribuční funkcí $F_5(x)$ a teoretickou distribuční funkcí $\Phi_T(x)$:

$d_1 = 0{,}2 - 0{,}17106 = 0{,}02894$; $d_2 = 0{,}4 - 0{,}26435 = 0{,}13565$;

$d_3 = 0{,}6 - 0{,}37448 = 0{,}22552$; $d_4 = 0{,}8 - 0{,}62552 = 0{,}17448$;

$d_5 = 1 - 0{,}94295 = 0{,}05705$.

Testová statistika: $D_5 = 0{,}22552$, modifikovaná kritická hodnota pro $n = 5$, $\alpha = 0{,}05$ je 0,343. Protože $0{,}22552 < 0{,}343$, hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

Popis Shapirova – Wilkova testu

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení $N(\mu, \sigma^2)$.

Testová statistika má tvar: $W = \frac{\left[\sum_{i=1}^{m} a_i^{(n)} (X_{(n-i+1)} - X_{(i)})\right]^2}{\sum_{i=1}^{n} (X_i - M)^2}$, kde m = n/2 pro n sudé a m = (n-1)/2 pro n liché. Koeficienty $a_i^{(n)}$ jsou tabelovány.

Na testovou statistiku W lze pohlížet jako na korelační koeficient mezi uspořádanými pozorováními a jim odpovídajícími kvantily standardizovaného normálního rozložení. V případě, že data vykazují perfektní shodu s normálním rozložením, bude mít W hodnotu 1. Hypotézu o normalitě tedy zamítneme na hladině významnosti α, když se na této hladině neprokáže korelace mezi daty a jim odpovídajícími kvantily rozložení N(0,1).

Lze také říci, že S – W test je založen na zjištění, zda body v Q-Q grafu jsou významně odlišné od regresní přímky proložené těmito body.

(S-W test se používá především pro výběry menších rozsahů, n < 50, ale v systému STATISTICA je implementováno jeho rozšíření i na výběry velkých rozsahů, kolem 2000.)

Výpočet pomocí systému STATISTICA (1)

V sedmi náhodně vybraných prodejnách byly zjištěny následující ceny určitého druhu zboží (v Kč): 35, 29, 30, 33, 45, 33, 36. Rozhodněte pomocí Lilieforsovy varianty K-S testu a S-W testu na hladině významnosti 0,05, zda lze tyto ceny považovat za realizace náhodného výběru z normálního rozložení.

Řešení:

Otevřeme nový datový soubor o jedné proměnné a 7 případech. Do proměnné X jsou zapíšeme zjištěné ceny.

Statistiky – Základní statistiky a tabulky – Tabulky četností - OK – Proměnné X, OK – Normalita – zaškrtneme Lilieforsův test a Shaphiro - Wilksův W test –Testy normality

	Testy normality (Tabulka22)				
Proměnná	N	max D	Lilliefors p	W	p
x	7	0,24029	p > .20	0,86866	0,18067

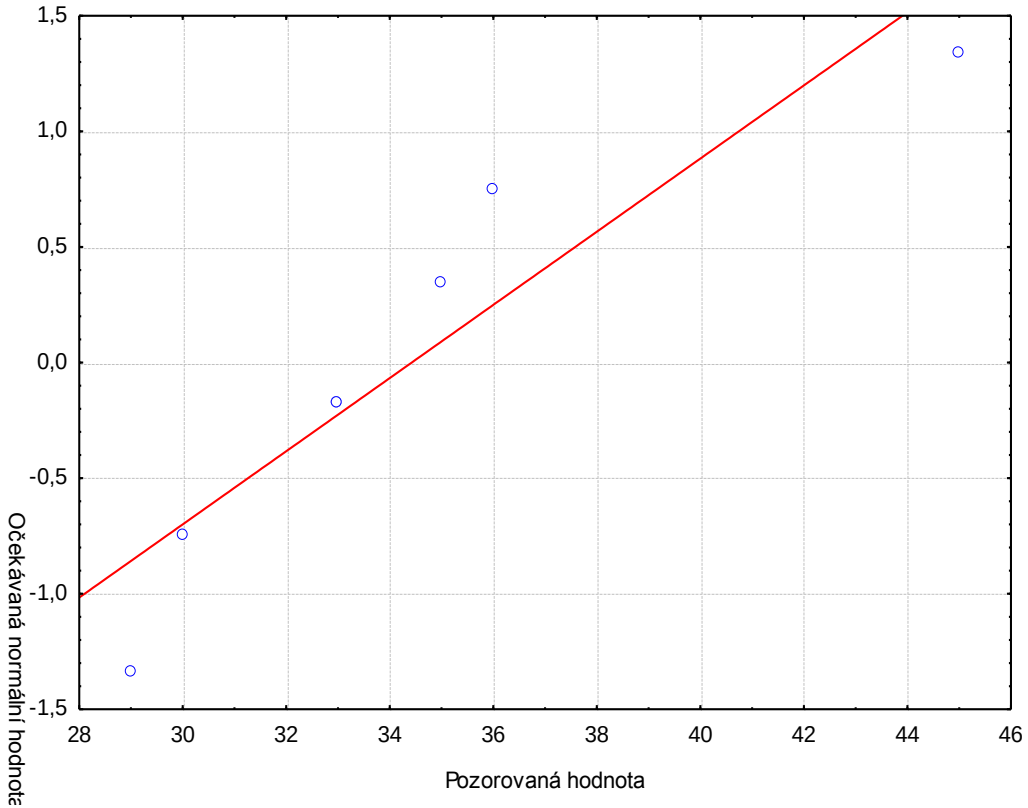
V tabulce je uvedena hodnota testové statistiky pro Lilieforsův test (d = 0,24029) a pro S-W test (W = 0,86866) a odpovídající p-hodnoty. Lilieforsovo p je počítáno na základě parametrů odhadnutých z dat. V našem případě p > 0,2 a pro S-W test p = 0,18068. Ani jeden z testů nezamítá nulovou hypotézu o normalitě.

Výpočet doplníme normálním pravděpodobnostním grafem a kvantil – kvantilovým grafem:

Graphs – 2D Graphs - Normal Probability Plots (resp. Quantile- Quantile plot)- Variables X – OK.

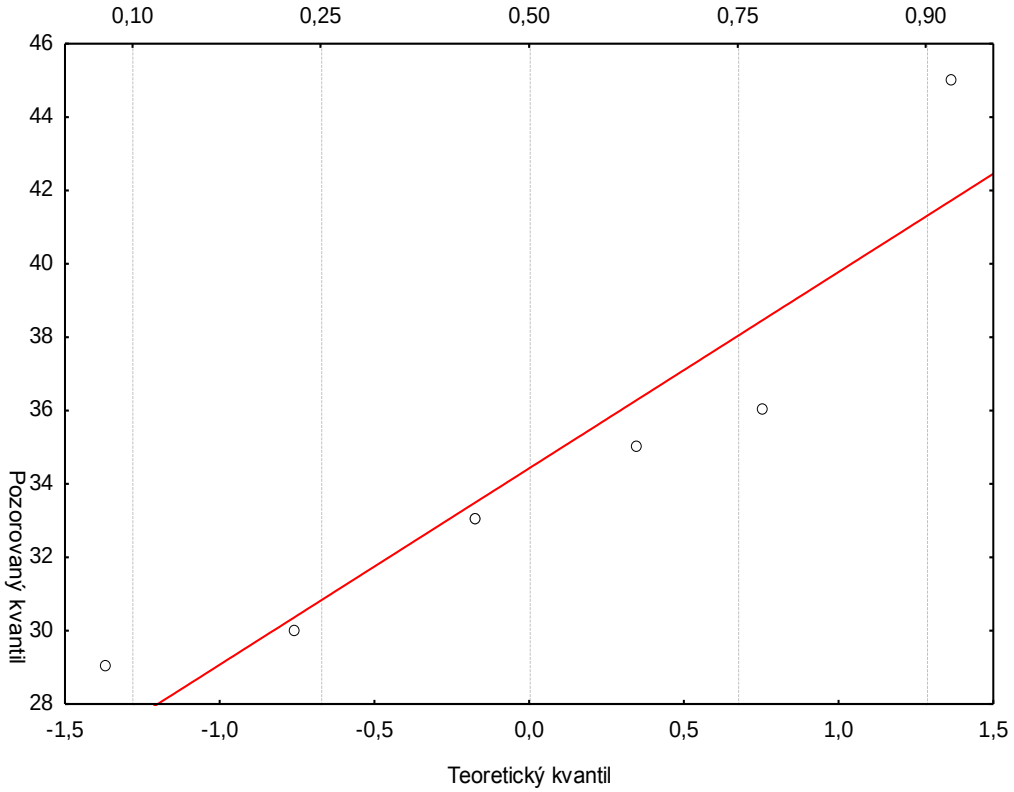
Výpočet pomocí systému STATISTICA (2)

N-P plot:



Výpočet pomocí systému STATISTICA (3)

Q-Q plot:



Další testy normality (1)

Existují testy normality založené na výběrové šikmosti a špičatosti. Pro náhodnou veličinu s normálním rozložením platí, že její šikmost i špičatost jsou nulové. Pro výběr z normálního rozložení by tedy výběrová šikmost a špičatost měly být blízké 0.

Nechť $X_1, \ldots, X_n$ je náhodný výběr.

Výběrová šikmost: $A_3 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^3}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2}\right]^3}$

Výběrová špičatost: $A_4 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^4}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2}\right]^4} - 3$

Lze dokázat, že pro výběr z normálního rozložení platí:

$$E(A_3) = 0,\ D(A_3) = \frac{6(n-2)}{(n+1)(n+3)},\ E(A_4) = -\frac{6}{n+1},\ D(A_4) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

Další testy normality (2)

Test založený na šikmosti zamítne hypotézu o normalitě na asymptotické hladině významnosti α, když $U_3 = \frac{|A_3|}{\sqrt{D(A_3)}} \geq u_{1-\alpha/2}$.

D'Agostinův test: zavedeme pomocné veličiny

$$b = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)},$$

$$W^2 = \sqrt{2(b-1)} - 1,$$

$$d = \frac{1}{\sqrt{\ln W}}, \quad a = \sqrt{\frac{2}{W^2-1}}.$$

Testová statistika má tvar $Z_3 = d \cdot \ln\left[\frac{U_3}{a} + \sqrt{(\frac{U_3}{a})^2+1}\right]$ a platí, že má přibližně rozložení N(0,1). Pro n>8 zamítáme hypotézu o normalitě pokud $|Z_3| \geq u_{1-\alpha/2}$.

Další testy normality (3)

Test založený na špičatosti zamítne hypotézu o normalitě na asymptotické hladině významnosti α, když $U_4 = \frac{|A_4 - E(A_4)|}{\sqrt{D(A_4)}} \geq u_{1-\alpha/2}$.

Také v tomto případě existuje D'Agostinova modifikace testu, nebudeme ji ale uvádět. Z dalších testů normality lze jmenovat např. Andersonův-Darlingův nebo Jarque-Beraův test.

Popis testu dobré shody v diskrétním a spojitém případě (1)

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení s distribuční funkcí $\Phi(x)$.

Je-li distribuční funkce spojitá, pak data rozdělíme do r třídicích intervalů $\langle u_j, u_{j+1} \rangle$, j = 1, ..., r. Zjistíme absolutní četnost n_j j-tého třídicího intervalu a vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat v j-tém třídicím intervalu. Platí-li nulová hypotéza, pak $p_j = \Phi(u_{j+1}) - \Phi(u_j)$.

Má-li distribuční funkce nejvýše spočetně mnoho bodů nespojitosti, pak místo třídicích intervalů použijeme varianty $x_{[j]}$, j = 1, ..., r. Pro variantu $x_{[j]}$ zjistíme absolutní četnost n_j a vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat variantou $x_{[j]}$. Platí-li nulová hypotéza, pak $p_j = \Phi(x_{[j]}) - \lim_{x \to x_{[j]}-} \Phi(x) = P(X = x_{[j]})$.

Popis testu dobré shody v diskrétním a spojitém případě (2)

Testová statistika: $K = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j}$.

Platí-li nulová hypotéza, pak $K \approx \chi^2(r-1-p)$, kde p je počet odhadovaných parametrů daného rozložení. (Např. pro normální rozložení p = 2, protože z dat odhadujeme střední hodnotu a rozptyl.) Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}(r-1-p)$. Aproximace se považuje za vyhovující, když tzv. teoretické četnosti $np_j \geq 5$, $j = 1, \ldots, r$.

Upozornění: Hodnota testové statistiky K je silně závislá na volbě třídicích intervalů. Navíc při nesplnění podmínky $np_j \geq 5$, $j = 1, \ldots, r$ je třeba některé intervaly resp. varianty slučovat, což vede ke ztrátě informace.

Příklad

(Test dobré shody pro diskrétní rozložení)

Byl zjišťován počet poruch určitého zařízení za 100 hodin provozu ve 150 disjunktních 100 h intervalech. Výsledky měření:

Počet poruch za 100 hodin provozu	0	1	2	3	4 a víc
Absolutní četnost	52	48	36	10	4

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že náhodný výběr $X_1, \ldots, X_{150}$ pochází z rozložení Po(1,2).

Řešení: Pravděpodobnost, že náhodná veličina s rozložením Po(λ), kde $\lambda = 1{,}2$ bude nabývat hodnot $p_0, \ldots, p_4$ a víc je $p_j = \frac{\lambda^j}{j!}e^{-\lambda} = \frac{1{,}2^j}{j!}e^{-1{,}2}$, $j = 0,1,2,3$, $p_4 = 1 - (p_0 + p_1 + p_2 + p_3)$.

Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_j	$\frac{(n_j - np_j)^2}{np_j}$
0	52	0,301	150.0,301=45,15	1,039
1	48	0,361	150.0,361=54,15	0,698
2	36	0,217	150.0,217=32,55	0,366
3	10	0,087	150.0,087=13,05	0,713
4	4	0,034	150.0,034=5,1	0,237

$K = 1{,}039 + 0{,}698 + 0{,}713 + 0{,}237 = 3{,}053$, $r = 5$, $\chi^2_{0,95}(4) = 9{,}488$. Protože $3{,}053 < 9{,}488$, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA (1)

Vytvoříme datový soubor o dvou proměnných (POČET a ČETNOST) a pěti případech a zapíšeme do něj hodnoty 0 1 2 3 4 a 52 48 36 10 4.

Statistiky – Prokládání rozdělení – Diskrétní rozdělení – Poissonovo – OK – Proměnná POČET – Proměnná vah ČETNOST – Stav zapnuto – OK – Parametry Lambda 1,2, OK.

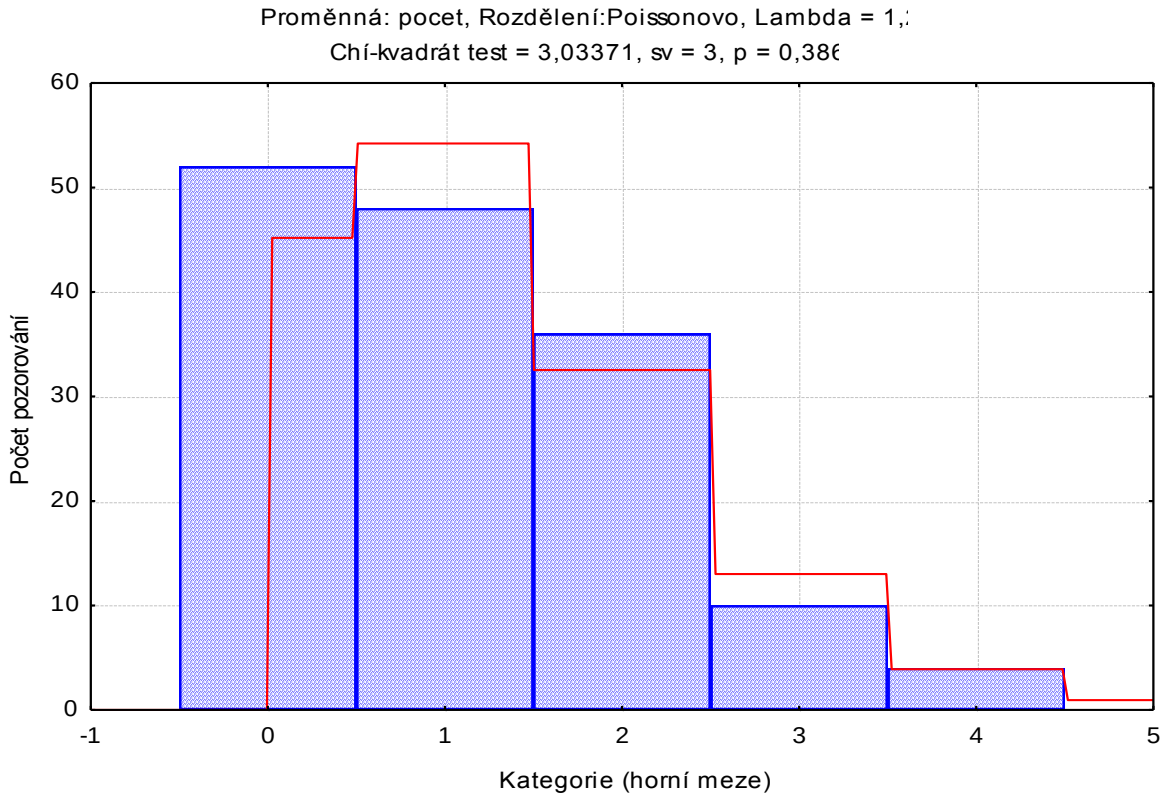
	Proměnná: pocet, Rozdělení:Poissonovo, Lambda = 1,20000 (Tabulka4) Chí-kvadrát = 3,03371, sv = 3, p = 0,38646							
Kategorie	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 0,00000	52	52	34,6666	34,666	45,1791	45,179	30,1194	30,119
1,00000	48	100	32,0000	66,666	54,2149	99,394	36,1433	66,262
2,00000	36	136	24,0000	90,666	32,5289	131,923	21,6859	87,948
3,00000	10	146	6,6666	97,333	13,0115	144,934	8,6743	96,623
< Nekonečno	4	150	2,6666	100,000	5,0653	150,000	3,3769	100,000

Ve výstupní tabulce je uvedena hodnota testového kritéria (3,03371) a odpovídající p-hodnota (0,38646). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

(Podmínky dobré aproximace jsou splněny, všechny teoretické četnosti - uvedené ve sloupci Očekávané četnosti – jsou větší než 5.)

Výpočet pomocí systému STATISTICA (2)

Pro vytvoření grafu se vrátíme do Proložení diskrétních rozložení – Základní výsledky – Graf pozorovaného a očekávaného rozdělení.



Příklad

(Test dobré shody pro spojité rozložení)

Byl pořízen náhodný výběr rozsahu n = 100. Jeho číselné realizace byly roztříděny do 5 ekvidistantních třídicích intervalů o délce 0,04, přičemž dolní mez prvního třídicího intervalu je 3,92. Absolutní četnosti jednotlivých třídicích intervalů jsou: 11, 20, 44, 19, 6.

Výběrový průměr se realizoval hodnotou m = 4,02 a výběrová směrodatná odchylka hodnotou s = 0,04.

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že náhodný výběr pochází z normálního rozložení.

Řešení:

Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

Přitom symbolem Φ značíme distribuční funkci rozložení $N(\mu,\sigma^2)$, kde $\mu = 4{,}02$ a $\sigma = 0{,}04$.

$(u_j, u_{j+1}\rangle$	n_j	$p_j=\Phi(u_{j+1})-\Phi(u_j)$	np_j	$(n_j - np_j)^2$	$\frac{(n_j - np_j)^2}{np_j}$
$(3{,}92;3{,}96\rangle$	11	0,060598	6,0598	24,4060	4,0276
$(3{,}96;4{,}00$	20	0,241730	24,1730	17,4142	0,7204
$(4{,}00;4{,}04\rangle$	44	0,382925	38,2925	32,5756	0,8507
$(4{,}04;4{,}08\rangle$	19	0,241730	24,1730	26,7608	1,1070
$(4{,}08;4{,}12\rangle$	6	0,060598	6,0598	0,0036	0,0006

$K = 4{,}0276 + 0{,}7204 + 0{,}8507 + 1{,}1070 + 0{,}0006 = 6{,}7063$

Kritický obor: $W = \langle \chi^2_{1-\alpha}(r-1-p), \infty) = \langle \chi^2_{0.95}(5-1-2), \infty) = \langle 5{,}9915, \infty)$

Výpočet pomocí systému STATISTICA (1)

Protože nemáme k dispozici původní data, ale jenom třídicí intervaly a jejich četnosti, do nového datového souboru o dvou proměnných xj a nj zadáme středy třídicích intervalů a jejich absolutní četnosti:

	1 xj	2 nj
1	3,94	11
2	3,98	20
3	4,02	44
4	4,06	19
5	4,1	6

Statistiky – Prokládání rozdělení – ponecháme implicitní nastavení pro Normální rozdělení – OK – Proměnná xj – klikneme na ikonu se závažím – Proměnná vah nj – Stav Zapnuto – OK – Parametry – Počet kategorií 5, Průměr 4,02, Rozptyl 0,0016, OK.

Dostaneme výstupní tabulku:

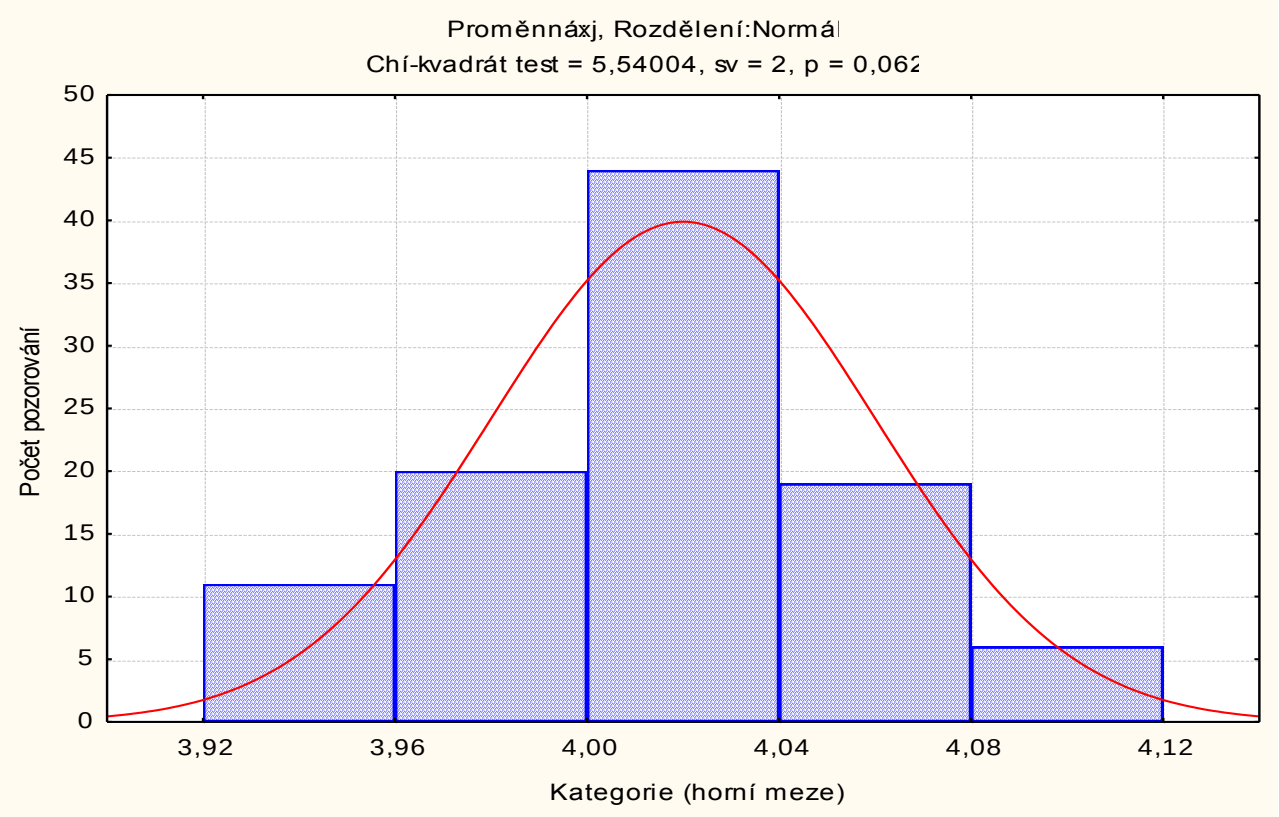
	Proměnná:xj, Rozdělení:Normální (Tabulka10) Chí-kvadrát = 5,54004, sv = 2, p = 0,06266							
Horní hranice	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 3,96000	11	11	11,0000	11,000	6,6807	6,6807	6,6807	6,6807
4,00000	20	31	20,0000	31,000	24,1730	30,853	24,1730	30,853
4,04000	44	75	44,0000	75,000	38,2924	69,146	38,2924	69,146
4,08000	19	94	19,0000	94,000	24,1730	93,319	24,1730	93,319
< Nekonečno	6	100	6,0000	100,000	6,6807	100,000	6,6807	100,000

V záhlaví výstupní tabulky je uvedena hodnota testového kritéria (5,54004), počet stupňů volnosti = 2 a p-hodnota (0,06266). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

Rozdíl oproti ručnímu výpočtu je způsoben tím, že systém STATISTICA uvažuje první interval $(-\infty, 3{,}96\rangle$ a poslední interval $\langle 4{,}08, \infty)$.

Výpočet pomocí systému STATISTICA (2)

Pro vytvoření grafu se vrátíme do Proložení spojitých rozdělení – Základní výsledky – Graf pozorovaného a očekávaného rozdělení.



Poznámka o testu dobré shody

Test dobré shody může být použit i v těch případech, kdy rozložení, z něhož daný náhodný výběr pochází, neodpovídá nějakému známému rozložení (např. exponenciálnímu, normálnímu, Poissonovu, ...), ale je určeno intuitivně nebo na základě zkušenosti.

Příklad

Ve svých pokusech pozoroval J.G. Mendel 10 rostlin hrachu a na každé z nich počet žlutých a zelených semen. Výsledky pokusu:

č.rostliny	1	2	3	4	5	6	7	8	9	10
počet žlutých semen	25	32	14	70	24	20	32	44	50	44
počet zelených semen	11	7	5	27	13	6	13	9	14	18
celkem	36	39	19	97	37	26	45	53	64	62

Z genetických modelů vyplývá, že pravděpodobnost výskytu žlutého semene by měla být 0,75 a zeleného 0,25. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že výsledky Mendelových pokusů se shodují s modelem.

Řešení: Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_j	$\frac{(n_j - np_j)^2}{np_j}$
1	25	0,75	36.0,75=27	0,148148
2	32	0,75	39.0,75=29,25	0,258547
⋮	⋮	⋮	⋮	⋮
10	44	0,75	62.0,75=46,5	0,134409

K = 0,148148 + 0,258547 + ... + 0,134409 = 1,797495, r = 10, $\chi^2_{0,95}(9) = 16,9$.

Protože 1,797495 < 16,9, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA

Vytvoříme datový soubor se třemi proměnnými Celkem, X a Y a 10 případy. Do proměnné Celkem zapíšeme celkový počet žlutých a zelených semen, do X zapíšeme pozorované absolutní četnosti žlutých semen, do proměnné Y vypočítané teoretické četnosti (v našem případě Celkem*0,75).

Statistiky – Neparametrická statistika – Pozorované vs. očekávané χ^2 – Proměnné Pozorované četnosti X, Očekávané četnosti Y, OK – Výpočet.

Pozorované vs. očekávané četnosti (Mendel hr
Chi-Kvadr. = 1,797495 sv = 9 p = ,994280
POZN.: Nestejné součty pozor. a oček. četnosti

Případ	pozorov. X	očekáv. Y	P - O	(P-O)^2 /O
C: 1	25,000(	27,000(	-2,0000(	0,14814
C: 2	32,000(	29,250(	2,7500(	0,25854
C: 3	14,000(	14,250(	-0,2500(	0,00438
C: 4	70,000(	72,750(	-2,7500(	0,10395
C: 5	24,000(	27,750(	-3,7500(	0,50675
C: 6	20,000(	19,500(	0,5000(	0,01282
C: 7	32,000(	33,750(	-1,7500(	0,09074
C: 8	44,000(	39,750(	4,2500(	0,45440
C: 9	50,000(	48,000(	2,0000(	0,08333
C: 10	44,000(	46,500(	-2,5000(	0,13440
Sčt	355,000(	358,500(	-3,5000(	1,79749

Ve výstupní tabulce najdeme hodnotu testové statistiky (Chi-Kvadr. = 1,797495) a odpovídající p-hodnotu, kterou porovnáme se zvolenou hladinou významnosti. V našem případě je p-hodnota 0,99428, takže nulová hypotéza se nezamítá na asymptotické hladině významnosti 0,05.

6. Parametrické úlohy o jednom náhodném výběru z normálního rozložení

Motivace: Mnoho náhodných veličin, s nimiž se setkáváme ve výzkumu i praxi, se řídí normálním rozložením. Za jistých předpokladů obsažených v centrální limitní větě se dá rozložení jiných náhodných veličin aproximovat normálním rozložením. Proto je zapotřebí věnovat velkou pozornost právě náhodným výběrům z normálního rozložení.

Rozložení statistik odvozených z výběrového průměru a výběrového rozptylu

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $N(\mu, \sigma^2)$. Pak platí

a) Výběrový průměr M a výběrový rozptyl S^2 jsou stochasticky nezávislé.

b) $M \sim N(\mu, \frac{\sigma^2}{n})$, tedy $U = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$.

(Pivotová statistika U slouží k řešení úloh o μ, když σ^2 známe.)

c) $K = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

(Pivotová statistika K slouží k řešení úloh o σ^2, když μ neznáme.)

d) $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$.

(Tato pivotová statistika slouží k řešení úloh o σ^2, když μ známe.)

e) $T = \frac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$.

(Pivotová statistika T slouží k řešení úloh o μ, když σ^2 neznáme.)

Důkaz

ad a) Nebudeme provádět, viz Jiří Anděl: Matematická statistika, SNTL/Alfa, Praha 1978, str. 82

ad b) Výběrový průměr M je lineární kombinace náhodných veličin s normálním rozložením, má tedy normální rozložení s parametry $E(M) = \mu$, $D(M) = \sigma^2/n$. Statistika U se získá standardizací M.

ad c) Vhodnou úpravou výběrového rozptylu S^2, kde použijeme obrat $X_i - M = (X_i - \mu) - (M - \mu)$, lze statistiku K vyjádřit jako součet kvadrátů n - 1 stochasticky nezávislých náhodných veličin se standardizovaným normálním rozložením. Tento součet se řídí rozložením $\chi^2(n-1)$.

ad d) Tato statistika je součet kvadrátů n stochasticky nezávislých náhodných veličin se standardizovaným normálním rozložením, řídí se tedy rozložením $\chi^2(n)$.

ad e) $U \sim N(0, 1)$, $K \sim \chi^2(n-1)$ jsou stochasticky nezávislé, protože M a S^2 jsou stochasticky nezávislé, tudíž statistika $T = \frac{U}{\sqrt{\frac{K}{n-1}}} = \frac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$.

Příklad

Hmotnost balíčku krystalového cukru baleného na automatické lince se řídí normálním rozložením se střední hodnotou 1002 g a směrodatnou odchylkou 8 g. Kontrolor náhodně vybírá 9 balíčků z jedné série a zjišťuje, zda jejich průměrná hmotnost je alespoň 999 g. Pokud ne, podnik musí zaplatit pokutu 20 000 Kč. Jaká je pravděpodobnost, že podnik bude muset zaplatit pokutu?

Řešení:

$X \sim N(1002, 64)$, $M \sim N(1002, \frac{64}{9})$

$$P(M \leq 999) = P\left(\frac{M-1002}{\sqrt{\frac{64}{9}}} \leq \frac{999-1002}{\sqrt{\frac{64}{9}}}\right) = P\left(U \leq -\frac{9}{8}\right) = \Phi\left(\frac{-9}{8}\right) = 1 - - \Phi\left(\frac{9}{8}\right) =$$

$$1 - \Phi(1{,}125) = 1 - \Phi(1{,}13) = 1 - 0{,}87076 = 0{,}12924$$

Pravděpodobnost, že podnik bude platit pokutu, je asi 12,9 %.

Výpočet pomocí systému STATISTICA

Využijeme toho, že STATISTICA pomocí funkce INormal(x;mu;sigma) umí vypočítat hodnotu distribuční funkce normálního rozložení se střední hodnotou mu a směrodatnou odchylkou sigma. Tedy $P(M \leq 999) = \Phi(999)$, kde Φ je distribuční funkce rozložení N(1002, 64/9).

Otevřeme nový datový soubor o jedné proměnné a jednom případu. Dvakrát klikneme na název proměnné Prom1. Do Dlouhého jména této proměnné napíšeme = INormal(999;1002;8/3).

V proměnné Prom1 se objeví hodnota 0,130295.

Vzorce pro meze 100(1-α)% empirických intervalů spolehlivosti pro μ a σ^2 (1)

a) Interval spolehlivosti pro μ, když σ^2 známe (využití pivotové statistiky U)

Oboustranný: $(d, h) = \left(m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right)$

Levostranný: $(d, \infty) = \left(m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}\right)$

b) Interval spolehlivosti pro μ, když σ^2 neznáme (využití pivotové statistiky T)

Oboustranný: $(d, h) = \left(m - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1), m + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right)$

Levostranný: $(d, \infty) = \left(m - \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1), \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty, m + \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1)\right)$

Vzorce pro meze 100(1-α)% empirických intervalů spolehlivosti pro μ a σ^2 (2)

c) Interval spolehlivosti pro σ^2, když μ neznáme (využití pivotové statistiky K)

Oboustranný: $(d, h) = \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} \right)$

Levostranný: $(d, \infty) = \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha}(n-1)}, \infty \right)$

Pravostranný: $(-\infty, h) = \left(-\infty, \frac{(n-1)s^2}{\chi^2_{\alpha}(n-1)} \right)$

d) Interval spolehlivosti pro σ^2, když μ známe (využití pivotové statistiky $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}$)

Oboustranný: $(d, h) = \left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha/2}(n)} \right)$

Levostranný: $(d, \infty) = \left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha}(n)}, \infty \right)$

Pravostranný: $(-\infty, h) = \left(-\infty, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha}(n)} \right)$

Příklad (1)

10 krát nezávisle na sobě byla změřena jistá konstanta μ. Výsledky měření byly: 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2. Tyto výsledky považujeme za číselné realizace náhodného výběru $X_1, \ldots, X_{10}$ z rozložení $N(\mu, \sigma^2)$, kde parametry μ, σ^2 neznáme. Najděte 95% empirický interval spolehlivosti jak pro μ, tak pro σ^2 a to

a) oboustranný,

b) levostranný,

c) pravostranný.

Řešení: $m = 2{,}06$, $s^2 = 0{,}0404$, $s = 0{,}2011$, $\alpha = 0{,}05$, $t_{0{,}975}(9) = 2{,}2622$, $t_{0{,}95}(9) = 1{,}8331$, $\chi^2_{0{,}975}(9) = 19{,}023$, $\chi^2_{0{,}025}(9) = 2{,}7$, $\chi^2_{0{,}95}(9) = 16{,}919$, $\chi^2_{0{,}05}(9) = 3{,}325$

ad a) Oboustranný interval spolehlivosti pro střední hodnotu μ

$$d = m - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) = 2{,}06 - \frac{0{,}2011}{\sqrt{10}}\, 2{,}2622 = 1{,}92$$

$$h = m + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) = 2{,}06 + \frac{0{,}2011}{\sqrt{10}}\, 2{,}2622 = 2{,}20$$

$1{,}92 < \mu < 2{,}20$ s pravděpodobností aspoň 0,95.

Oboustranný interval spolehlivosti pro rozptyl σ^2

$$d = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} = \frac{9 \cdot 0{,}0404}{19{,}023} = 0{,}0191$$

$$h = \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} = \frac{9 \cdot 0{,}0404}{2{,}7} = 0{,}1347$$

$0{,}0191 < \sigma^2 < 0{,}1347$ s pravděpodobností aspoň 0,95.

Příklad (2)

ad b) Levostranný interval spolehlivosti pro střední hodnotu μ

$$d = m - \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1) = 2{,}06 - \frac{0{,}2011}{\sqrt{10}} 1{,}8331 = 1{,}94$$

$1{,}94 < \mu$ s pravděpodobností aspoň 0,95.

Levostranný interval spolehlivosti pro rozptyl σ^2

$$d = \frac{(n-1)s^2}{\chi^2_{1-\alpha}(n-1)} = \frac{9 \cdot 0{,}0404}{16{,}919} = 0{,}0215$$

$\sigma^2 > 0{,}0215$ s pravděpodobností aspoň 0,95.

ad c) Pravostranný interval spolehlivosti pro střední hodnotu μ

$$h = m + \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1) = 2{,}06 + \frac{0{,}2011}{\sqrt{10}} 1{,}8331 = 2{,}18$$

$\mu < 2{,}18$ s pravděpodobností aspoň 0,95.

Pravostranný interval spolehlivosti pro rozptyl σ^2

$$h = \frac{(n-1)s^2}{\chi^2_{\alpha}(n-1)} = \frac{9 \cdot 0{,}0404}{3{,}325} = 0{,}1094$$

$\sigma^2 < 0{,}1094$ s pravděpodobností aspoň 0,95.

Výpočet pomocí systému STATISTICA (1)

Vytvoříme nový datový soubor o jedné proměnné X a 10 případech. Do proměnné X napíšeme dané hodnoty.

Statistika – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme Meze spolehl. prům. a Meze sp. směr. odch. (ostatní volby zrušíme) – pro oboustranný 95% interval spolehlivosti ponecháme implicitní hodnotu pro Interval 95,00, pro jednostranné intervaly změníme hodnotu na 90,00.

Výsledky pro oboustranné 95% intervaly spolehlivosti pro střední hodnotu μ, pro směrodatnou odchylku σ a rozptyl σ^2:

Proměnná	Int. spolehl. -95,000%	Int. spolehl. 95,000	Spolehlivost Sm.Odch. -95,000%	Spolehlivost Sm.Odch. +95,000%	NProm1 =v3 ^2	NProm2 =v4 ^2
X	1,916136	2,203864	0,138329	0,367145	0,019135	0,134795

Vidíme, že

$1{,}92 < \mu < 2{,}20$ s pravděpodobností aspoň 0,95,

$0{,}1383 < \sigma < 0{,}3671$ s pravděpodobností aspoň 0,95.

$0{,}0191 < \sigma^2 < 0{,}1348$ s pravděpodobností aspoň 0,95.

Výpočet pomocí systému STATISTICA (2)

Výsledky pro jednostranné 95% intervaly spolehlivosti pro střední hodnotu μ, pro směrodatnou odchylku σ a rozptyl σ^2:

Proměnná	Int. spolehl. -90,000%	Int. spolehl. 90,000	Spolehlivost Sm.Odch. -90,000%	Spolehlivost Sm.Odch. +90,000%	NProm1 =v3^2	NProm2 =v4^2
X	1,94342	2,17657	0,14667	0,33086	0,02151	0,1094

Vidíme, že

$\mu > 1{,}94$ s pravděpodobností aspoň 0,95,

$\mu < 2{,}20$ s pravděpodobností aspoň 0,95,

$\sigma > 0{,}1467$ s pravděpodobností aspoň 0,95,

$\sigma < 0{,}3309$ s pravděpodobností aspoň 0,95,

$\sigma^2 > 0{,}0215$ s pravděpodobností aspoň 0,95,

$\sigma^2 < 0{,}1095$ s pravděpodobností aspoň 0,95.

Jednotlivé typy testů pro parametry normálního rozložení

a) Nechť $X_1, ..., X_n$ je náhodný výběr $N(\mu, \sigma^2)$, kde σ^2 známe. Nechť $n \geq 2$ a c je konstanta. Test $H_0: \mu = c$ proti $H_1: \mu \neq c$ se nazývá jednovýběrový z-test.

b) Nechť $X_1, ..., X_n$ je náhodný výběr $N(\mu, \sigma^2)$, kde σ^2 neznáme. Nechť $n \geq 2$ a c je konstanta. Test $H_0: \mu = c$ proti $H_1: \mu \neq c$ se nazývá jednovýběrový t-test.

c) Nechť $X_1, ..., X_n$ je náhodný výběr $N(\mu, \sigma^2)$, kde μ neznáme. Nechť $n \geq 2$ a c je konstanta. Test $H_0: \sigma^2 = c$ proti $H_1: \sigma^2 \neq c$ se nazývá test o rozptylu.

Provedení testů o parametrech μ, σ^2 pomocí kritického oboru (1)

a) Provedení jednovýběrového z-testu

Vypočteme realizaci testového kritéria $t_0 = \frac{m-c}{\frac{\sigma}{\sqrt{n}}}$. Stanovíme kritický obor W. Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: μ = c proti H_1: μ ≠ c. Kritický obor má tvar: $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$.

Levostranný test: Testujeme H_0: μ = c proti H_1: μ < c. Kritický obor má tvar: $W = (-\infty, -u_{1-\alpha}\rangle$.

Pravostranný test: Testujeme H_0: μ = c proti H_1: μ > c. Kritický obor má tvar: $W = \langle u_{1-\alpha}, \infty)$.

Provedení testů o parametrech μ, σ^2 pomocí kritického oboru (2)

b) Provedení jednovýběrového t-testu

Vypočteme realizaci testového kritéria $t_0 = \frac{m-c}{\frac{s}{\sqrt{n}}}$. Stanovíme kritický obor W. Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: μ = c proti H_1: μ ≠ c. Kritický obor má tvar: $W = (-\infty, -t_{1-\alpha/2}(n-1)\rangle \cup \langle t_{1-\alpha/2}(n-1), \infty)$.

Levostranný test: Testujeme H_0: μ = c proti H_1: μ < c. Kritický obor má tvar: $W = (-\infty, -t_{1-\alpha}(n-1)\rangle$.

Pravostranný test: Testujeme H_0: μ = c proti H_1: μ > c. Kritický obor má tvar: $W = \langle t_{1-\alpha}(n-1), \infty)$.

Provedení testů o parametrech μ, σ^2 pomocí kritického oboru (3)

c) Provedení testu o rozptylu

Vypočteme realizaci testového kritéria $t_0 = \frac{(n-1)s^2}{c}$. Stanovíme kritický obor W. Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: $\sigma^2 = c$ proti H_1: $\sigma^2 \neq c$. Kritický obor má tvar:.

$$W = \langle 0, \chi^2{}_{\alpha/2}(n-1)\rangle \cup \langle \chi^2{}_{1-\alpha/2}(n-1), \infty)$$

Levostranný test: Testujeme H_0: $\sigma^2 = c$ proti H_1: $\sigma^2 < c$. Kritický obor má tvar: $W = \langle 0, \chi^2{}_{\alpha}(n-1)\rangle$.

Pravostranný test: Testujeme H_0: $\sigma^2 = c$ proti H_1: $\sigma^2 > c$. Kritický obor má tvar: $W = \langle \chi^2{}_{1-\alpha}(n-1), \infty)$.

Příklad

Podle údajů na obalu čokolády by její čistá hmotnost měla být 125 g. Výrobce dostal několik stížností od kupujících, ve kterých tvrdili, že hmotnost čokolád je nižší než deklarovaných 125 g. Z tohoto důvodu oddělení kontroly náhodně vybralo 50 čokolád a zjistilo, že jejich průměrná hmotnost je 122 g a směrodatná odchylka 8,6 g. Za předpokladu, že hmotnost čokolád se řídí normálním rozložením, můžeme na hladině významnosti 0,01 považovat stížnosti kupujících za oprávněné?

Řešení: $X_1, ..., X_{50}$ je náhodný výběr z $N(\mu, \sigma^2)$. Testujeme hypotézu H_0: $\mu = 125$ proti levostranné alternativě H_1: $\mu < 125$. Protože neznáme rozptyl σ^2, použijeme jednovýběrový t-test.

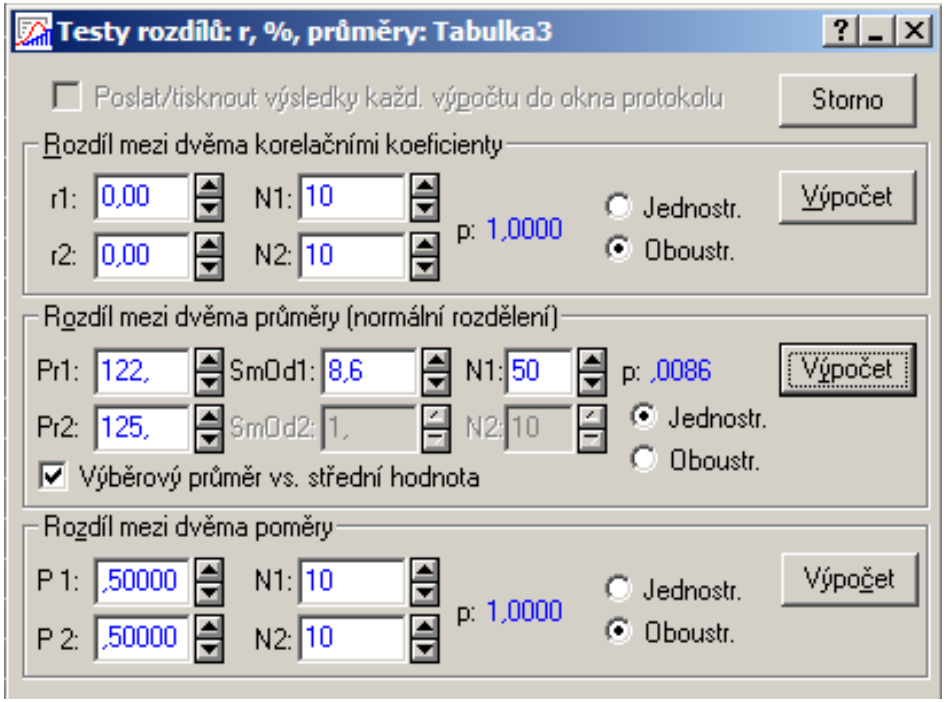
Testové kritérium $\frac{m-c}{\frac{s}{\sqrt{n}}} = \frac{122-125}{\frac{8,6}{\sqrt{50}}} = -2,4667.$

Kritický obor $W = (-\infty, -t_{1-\alpha}(n-1)\rangle = (-\infty, -t_{0,99}(49)\rangle = (-\infty, -2,4049\rangle.$

Jelikož testové kritérium se realizuje v kritickém oboru, zamítáme nulovou hypotézu na hladině významnosti 0,01. Stížnosti kupujících tedy lze považovat za oprávněné.

Výpočet pomocí systému STATISTICA

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma průměry (normální rozdělení) – zaškrtneme Výběrový průměr vs. Střední hodnota a zvolíme jednostr. – do políčka Pr1 napíšeme 122, do políčka SmOd1 napíšeme 8,6, do políčka N1 napíšeme 50, do políčka Pr2 napíšeme 125 - Výpočet. Dostaneme p-hodnotu 0,0086, tedy zamítáme nulovou hypotézu na hladině významnosti 0,01.



Definice rozdílového náhodného výběru a vzorec pro meze na základě náhodného výběru z dvourozměrného normálního rozložení

Nechť $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ je náhodný výběr z rozložení $N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} {\sigma_1}^2 & \sigma_{12} \\ \sigma_{12} & {\sigma_2}^2 \end{pmatrix}\right)$, přičemž $n \geq 2$. Označíme $\mu = \mu_1 - \mu_2$ a zavedeme rozdílový náhodný výběr $Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n$. Vypočteme $M = \frac{1}{n}\sum_{i=1}^{n} Z_i$, $S^2 = \frac{1}{n}\sum_{i=1}^{n}(Z_i - M)^2$.

Vzorec pro meze $100(1-\alpha)\%$ empirického intervalu spolehlivosti pro střední hodnotu rozdílového náhodného výběru

Oboustranný: $(d, h) = \left(m - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1), m + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right)$

Levostranný: $(d, \infty) = \left(m - \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1), \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty, m + \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1)\right)$

Příklad

Dvěma rozdílnými laboratorními metodami se zjišťoval obsah chemické látky v roztoku (v procentech). Bylo vybráno 5 vzorků a proměřeno oběma metodami. Výsledky měření jsou obsaženy v tabulce:

číslo vzorku	1	2	3	4	5
1. metoda	2,3	1,9	2,1	2,4	2,6
2. metoda	2,4	2,0	2,0	2,3	2,5

Za předpokladu, že data mají normální rozložení, sestrojte 90% empirický interval spolehlivosti pro rozdíl středních hodnot výsledků obou metod.

Řešení:

Přejdeme k rozdílovému náhodnému výběru, jehož realizace jsou: -0,1 -0,1 0,1 0,1 0,1. Vypočteme $m = 0{,}02$, $s^2 = 0{,}012$, $s = 0{,}109545$. Předpokládáme, že tato data pocházejí z normálního rozložení $N(\mu, \sigma^2)$. Vypočteme meze 90% oboustranného intervalu spolehlivosti pro μ při neznámém σ:

$$d = m - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1) = 0{,}02 - \frac{0{,}109545}{\sqrt{5}} t_{0{,}95}(4) = 0{,}02 - \frac{0{,}109545}{\sqrt{5}} 2{,}1318 = -0{,}0844$$

$$h = m + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) = 0{,}02 + \frac{0{,}109545}{\sqrt{5}} t_{0{,}95}(4) = 0{,}02 + \frac{0{,}109545}{\sqrt{5}} 2{,}1318 = 0{,}1244$$

$-0{,}0844 < \mu < 0{,}1244$ s pravděpodobností aspoň 0,9.

Výpočet pomocí systému STATISTICA

Vytvoříme nový datový soubor o 3 proměnných a 5 případech. Do 1. proměnné X napíšeme hodnoty pro 1. metodu, do 2. proměnné Y hodnoty pro 2. metodu a do 3. proměnné Z rozdíly mezi X a Y.

Statistiky – Základní statistiky a tabulky – Popisné statistiky, OK - Proměnné Z, Detailní výsledky – zaškrtneme Meze spolehl. Prům. – Interval 90% - Výpočet. Dostaneme tabulku:

	Popisné statistiky (chemicka latka	
Proměnná	Int. spolehl. -90,000%	Int. spolehl. 90,000
Z	-0,08443	0,12443

Vidíme tedy, že $-0{,}0844 < \mu < 0{,}1244$ s pravděpodobností aspoň 0,9.

Definice párového t-testu

Nechť $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ je náhodný výběr z rozložení $N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} {\sigma_1}^2 & \sigma_{12} \\ \sigma_{12} & {\sigma_2}^2 \end{pmatrix}\right)$, přičemž $n \geq 2$. Testujeme H_0: $\mu_1 - \mu_2 = c$ (tj. $\mu = c$) proti H_1: $\mu_1 - \mu_2 \neq c$ (tj. $\mu \neq c$) nebo testujeme nulovou hypotézu proti jedné z jednostranných alternativ. Tento test se nazývá párový t-test.

Provedení párového t-testu

Vypočteme realizaci testového kritéria $t_0 = \frac{m-c}{\frac{s}{\sqrt{n}}}$. Stanovíme kritický obor W. Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: $\mu = c$ proti H_1: $\mu \neq c$. Kritický obor má tvar: $W = (-\infty, -t_{1-\alpha/2}(n-1)\rangle \cup \langle t_{1-\alpha/2}(n-1), \infty)$.

Levostranný test: Testujeme H_0: $\mu = c$ proti H_1: $\mu < c$. Kritický obor má tvar:

Příklad (1)

V následující tabulce jsou údaje o výnosnosti dosažené 12 náhodně vybranými firmami při investování do mezinárodního podnikání (veličina X) a do domácího podnikání (veličina Y):

č.firmy	1	2	3	4	5	6	7	8	9	10	11	12
X	10	12	14	12	12	17	9	15	9	11	7	15
Y	11	14	15	11	13	16	10	13	11	17	9	19

(Výnosnost je vyjádřena v procentech a představuje podíl na zisku vložených investic za rok.)

Za předpokladu, že data pocházejí z dvourozměrného normálního rozložení, na hladině významnosti 0,1 testujte hypotézu, že neexistuje rozdíl mezi střední hodnotou výnosnosti investic do mezinárodního a domácího podnikání proti oboustranné alternativě.

Testování proveďte

a) pomocí intervalu spolehlivosti

b) pomocí kritického oboru.

(Pro úsporu času máte uvedeny realizace výběrového průměru $m = -1,\overline{3}$ a výběrového rozptylu $s^2 = 4,\overline{78}$ rozdílového náhodného výběru $Z_i = X_i - Y_i$, $i = 1, \ldots, 12$.)

Příklad (2)

Řešení:

Testujeme H_0: $\mu = 0$ proti H_1: $\mu \neq 0$

ad a)

90% interval spolehlivosti pro střední hodnotu μ při neznámém rozptylu σ^2 má meze:

$$d = m - \frac{s}{\sqrt{n}} t_{0,95}(n-1) = -1,\overline{3} - \frac{\sqrt{4,\overline{78}}}{\sqrt{12}} 1,7959 = -2,4677$$

$$h = m - \frac{s}{\sqrt{n}} t_{0,95}(n-1) = -1,\overline{3} + \frac{\sqrt{4,\overline{78}}}{\sqrt{12}} 1,7959 = -0,1989$$

Protože číslo c = 0 neleží v intervalu (-2,4677; -0,1989), H_0 zamítáme na hladině významnosti 0,1.

ad b)

Vypočítáme realizaci testové statistiky $t_0 = \frac{m-c}{\frac{s}{\sqrt{n}}} = \frac{-1,\overline{3}}{\frac{\sqrt{4,\overline{78}}}{\sqrt{12}}} = -2,11085$

Stanovíme kritický obor $W = (-\infty, -t_{0,95}(11)\rangle \cup \langle t_{0,95}(11), \infty) = (-\infty, -1,7959\rangle \cup \langle 1,7959, \infty)$

Protože testová statistika se realizuje v kritickém oboru, H_0 zamítáme na hladině významnosti 0,1.

Výpočet pomocí systému STATISTICA

Vytvoříme nový datový soubor o 2 proměnných a 12 případech. Do 1. proměnné X napíšeme hodnoty pro mezinárodní podnikání, do 2. proměnné hodnoty pro domácí podnikání.

Statistiky – Základní statistiky a tabulky – t-test pro závislé vzorky, OK - Proměnné X, Y – OK – Výpočet. Dostaneme tabulku:

	t-test pro závislé vzorky (investovani) Označ. rozdíly jsou významné na hlad. p < ,05000							
Proměnná	Průměr	Sm.odch.	N	Rozdíl	Sm.odch. rozdílu	t	sv	p
X	11,9166	2,93748						
Y	13,2500	3,04884	12	-1,3333	2,18812	-2,1108	11	0,05849

Vypočtenou p-hodnotu 0,05849 porovnáme se zvolenou hladinou významnosti $\alpha = 0{,}1$. Protože $p \leq \alpha$, zamítáme nulovou hypotézu na hladině významnosti 0,1.

7. Parametrické úlohy o dvou nezávislých náhodných výběrech z normálních rozložení

Motivace: V této situaci je naším úkolem porovnat střední hodnoty či rozptyly dvou normálních rozložení na základě znalosti dvou nezávislých náhodných výběrů pořízených z těchto rozložení. Zpravidla konstruujeme intervaly spolehlivosti pro rozdíl středních hodnot respektive hodnotíme shodu středních hodnot pomocí dvouvýběrového t-testu či dvouvýběrového z-testu a shodu rozptylů pomocí F-testu.

Rozložení statistik odvozených z výběrových průměrů a výběrových rozptylů (1)

Nechť $X_{11} \dots, X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$ a $X_{21}, \dots, X_{2n_2}$ je na něm nezávislý náhodný výběr z rozložení $N(\mu_2, \sigma_2^2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$. Označme M_1, M_2 výběrové průměry, S_1^2, S_2^2 výběrové rozptyly a $S_*^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ vážený průměr výběrových rozptylů. Pak platí:

a) Statistiky $M_1 - M_2$ a $S_*^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ jsou stochasticky nezávislé.

b) $U = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$.

(Pivotová statistika U slouží k řešení úloh o μ_1- μ_2, když σ_1^2 a σ_2^2 známe.)

Rozložení statistik odvozených z výběrových průměrů a výběrových rozptylů (2)

c) Jestliže $\sigma_1^2 = \sigma_2^2 =: \sigma^2$, pak $K = \frac{(n_1+n_2-2)S_*^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$.

(Pivotová statistika K slouží k řešení úloh o neznámém společném rozptylu σ^2.)

d) Jestliže $\sigma_1^2 = \sigma_2^2 =: \sigma^2$, pak $T = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{S_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$.

(Pivotová statistika T slouží k řešení úloh o $\mu_1 - \mu_2$, když σ_1^2 a σ_2^2 neznáme, ale víme, že jsou shodné.)

e) $F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$.

(Pivotová statistika F slouží k řešení úloh o σ_1^2/σ_2^2.)

Důkaz

ad a) Nebudeme provádět.

ad b) M_1-M_2 je lineární kombinace náhodných veličin s normálním rozložením, má tedy normální rozložení s parametry $E(M_1-M_2) = \mu_1 - \mu_2$, $D(M_1-M_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2$.

U se získá standardizací M_1-M_2.

ad c) $K_1 = \frac{(n_1-1)S_1^2}{\sigma^2} \sim \chi^2(n_1-1)$ a $K_2 = \frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_2-1)$ jsou stochasticky nezávislé náhodné veličiny, tedy $K = K_1+K_2 \sim \chi^2(n_1+n_2-2)$.

ad d) $U = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}}} \sim N(0,1)$, $K = \frac{(n_1+n_2-2)S_*^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$ jsou stochasticky nezávislé, protože $M_1 - M_2$ a S_*^2 jsou stochasticky nezávislé. $T = \frac{U}{\sqrt{\frac{K}{n_1+n_2-2}}} = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{S_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$.

ad e) $K_1 = \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1)$ a $K_2 = \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$ jsou stochasticky nezávislé náhodné veličiny, tedy $F = \frac{\frac{K_1}{n_1-1}}{\frac{K_2}{n_2-1}} = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$.

Příklad

Nechť jsou dány dva nezávislé náhodné výběry, první pochází z rozložení N(0,28; 0,09) a má rozsah 16, druhý pochází z rozložení N(0,25; 0,04) a má rozsah 25. Jaká je pravděpodobnost, že výběrový průměr 1. výběru bude větší než výběrový průměr 2. výběru?

Řešení:

$$P(M_1 > M_2) = P(M_1 - M_2 > 0) = 1 - P(M_1 - M_2 \leq 0) =$$

$$= 1 - P\left(\frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}} \leq \frac{0 - (\mu_1 - \mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}\right) =$$

$$= 1 - P\left(U \leq \frac{-0{,}28 + 0{,}25}{\sqrt{\frac{0{,}09}{16} + \frac{0{,}04}{25}}}\right) = 1 - P(U \leq -0{,}35294) = 1 - \Phi(-0{,}35) =$$

$$= \Phi(0{,}35) = 0{,}63683$$

S pravděpodobností přibližně 63,7% je výběrový průměr 1. výběru větší než výběrový průměr 2. výběru.

Výpočet pomocí systému STATISTICA

Statistika $M_1 - M_2$ se podle bodu (a) řídí rozložením $N\left(\mu_1 - \mu_2, \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right)$,

kde $\mu_1 - \mu_2 = 0{,}28 - 0{,}25 = 0{,}03$, $\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2} = \frac{0{,}09}{16} + \frac{0{,}04}{25} = 0{,}007225$, tj.

statistika $M_1 - M_2 \sim N(0{,}03;0{,}007225)$.

Otevřeme nový datový soubor o jedné proměnné a jednom případu. Do Dlouhého jména této proměnné napíšeme
= 1-INormal(0;0,03;sqrt(0,007225)).

V proměnné Prom1 se objeví hodnota 0,637934.

Intervaly spolehlivosti pro parametrické funkce $\mu_1-\mu_2$, σ_1^2/σ_2^2 (1)

Uvedeme přehled vzorců pro meze 100(1-α)% empirických intervalů spolehlivosti pro parametrické funkce $\mu_1 - \mu_2$, σ_1^2/σ_2^2.

a) Interval spolehlivosti pro $\mu_1-\mu_2$, když σ_1^2, σ_2^2 známe (využití pivotové statistiky U)

Oboustranný:

$$(\mathrm{d},\mathrm{h}) = \left(m_1-m_2-\sqrt{\frac{{\sigma_1}^2}{n_1}+\frac{{\sigma_2}^2}{n_2}}u_{1-\alpha/2}, m_1-m_2+\sqrt{\frac{{\sigma_1}^2}{n_1}+\frac{{\sigma_2}^2}{n_2}}u_{1-\alpha/2}\right)$$

Levostranný: $$(\mathrm{d},\infty) = \left(m_1-m_2-\sqrt{\frac{{\sigma_1}^2}{n_1}+\frac{{\sigma_2}^2}{n_2}}u_{1-\alpha}, \infty\right)$$

Pravostranný: $$(-\infty,\mathrm{h}) = \left(-\infty, m_1-m_2+\sqrt{\frac{{\sigma_1}^2}{n_1}+\frac{{\sigma_2}^2}{n_2}}u_{1-\alpha}\right)$$

Intervaly spolehlivosti pro parametrické funkce $\mu_1-\mu_2$, σ_1^2/σ_2^2 (2)

b) Interval spolehlivosti pro $\mu_1-\mu_2$, když σ_1^2, σ_2^2 neznáme, ale víme, že jsou shodné (využití pivotové statistiky T)

Oboustranný:

$$(\mathrm{d},\mathrm{h}) = \left(m_1 - m_2 - s_* \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\, t_{1-\alpha/2}(n_1+n_2-2),\ m_1 - m_2 + s_* \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\, t_{1-\alpha/2}(n_1+n_2-2)\right)$$

Levostranný: $$(\mathrm{d},\infty) = \left(m_1 - m_2 - s_* \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\, t_{1-\alpha}(n_1+n_2-2),\infty\right)$$

Pravostranný: $$(-\infty,\mathrm{h}) = \left(-\infty, m_1 - m_2 + s_* \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\, t_{1-\alpha}(n_1+n_2-2)\right)$$

Upozornění: Není-li v bodě (b) splněn předpoklad o shodě rozptylů, lze sestrojit aspoň přibližný 100(1-α)% interval spolehlivosti pro $\mu_1-\mu_2$.

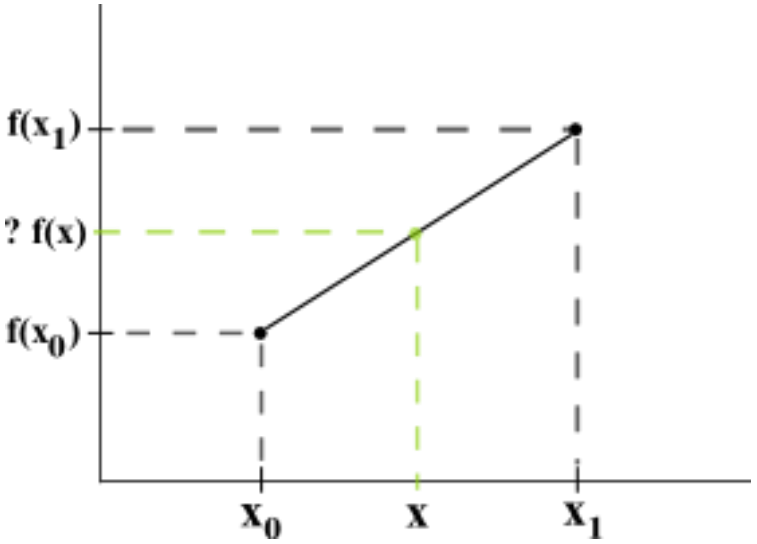
V tomto případě má statistika T přibližně rozložení t(ν), kde počet stupňů volnosti $\nu = \frac{(s_1^2/n_1+s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}}$. Není-li ν celé číslo, použijeme v tabulkách kvantilů Studentova rozložení lineární interpolaci.

Lineární interpolace

$$f(x) - f(x_0) = \frac{\Delta f(x)}{\Delta x}(x - x_0) \quad \text{– rovnice směrnice}$$

$$f(x) - f(x_0) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0)$$

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0)$$



Příklad: máme α=0,05, vyjde ν=5,25, v tabulkách najdeme $t_{0,95}(5)= 2,015$ a $t_{0,95}(6)= 1,943$.
Interpolací dostáváme:

$$t_{0,95}(5.25) = 2,015 + \frac{1,943 - 2,015}{6 - 5}(5,25 - 5) = 1,997$$

Intervaly spolehlivosti pro parametrické funkce $\mu_1-\mu_2$, σ_1^2/σ_2^2 (3)

c) Interval spolehlivosti pro společný neznámý rozptyl σ^2 (využití pivotové statistiky K)

Oboustranný: $(d, h) = \left(\frac{(n_1+n_2-2)s_*^2}{\chi^2_{1-\alpha/2}(n_1+n_2-2)}, \frac{(n_1+n_2-2)s_*^2}{\chi^2_{\alpha/2}(n_1+n_2-2)} \right)$

Levostranný: $(d, \infty) = \left(\frac{(n_1+n_2-2)s_*^2}{\chi^2_{1-\alpha}(n_1+n_2-2)}, \infty \right)$

Pravostranný: $(-\infty, h) = \left(-\infty, \frac{(n_1+n_2-2)s_*^2}{\chi^2_{\alpha}(n_1+n_2-2)} \right)$

Intervaly spolehlivosti pro parametrické funkce μ_1-μ_2 , σ_1^2/σ_2^2 (4)

d) Interval spolehlivosti pro podíl rozptylů $\frac{\sigma_1^2}{\sigma_2^2}$ (využití pivotové statistiky F)

Oboustranný: $(d, h) = \left(\frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1,n_2-1)}, \frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1,n_2-1)} \right)$

Levostranný: $(d, \infty) = \left(\frac{s_1^2/s_2^2}{F_{1-\alpha}(n_1-1,n_2-1)}, \infty \right)$

Pravostranný: $(-\infty, h) = \left(-\infty, \frac{s_1^2/s_2^2}{F_{\alpha}(n_1-1,n_2-1)} \right)$

Příklad

Ve dvou nádržích se zkoumal obsah chlóru (v g/l). Z první nádrže bylo odebráno 25 vzorků, z druhé nádrže 10 vzorků. Byly vypočteny realizace výběrových průměrů a rozptylů: $m_1 = 34{,}48$, $m_2 = 35{,}59$, $s_1^2 = 1{,}7482$, $s_2^2 = 1{,}7121$. Hodnoty zjištěné z odebraných vzorků považujeme za realizace dvou nezávislých náhodných výběrů z rozložení $N(\mu_1, \sigma^2)$ a $N(\mu_2, \sigma^2)$. Sestrojte 95% empirický interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$.

Řešení: Úloha vede na vzorec (b). Vypočteme vážený průměr výběrových rozptylů a najdeme odpovídající kvantily Studentova rozložení:

$$s_*^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{24\cdot 1{,}7482+9\cdot 1{,}7121}{33} = 1{,}7384,\ t_{0{,}975}(33) = 2{,}035$$

Dosadíme do vzorců pro dolní a horní mez intervalu spolehlivosti:

$$d = m_1 - m_2 - s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\,t_{1-\frac{\alpha}{2}}(n_1+n_2-2) = 34{,}48 - 35{,}59 - -\sqrt{1{,}7384}\cdot\sqrt{\frac{1}{25}+\frac{1}{10}}\cdot 2{,}035 = -2{,}114$$

$$h = m_1 - m_2 + s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\,t_{1-\frac{\alpha}{2}}(n_1+n_2-2) = 34{,}48 - 35{,}59 + +\sqrt{1{,}7384}\cdot\sqrt{\frac{1}{25}+\frac{1}{10}}\cdot 2{,}035 = -0{,}106$$

-2,114 g/l < $\mu_1 - \mu_2$ < -0,106 g/l s pravděpodobností aspoň 0,95.

Výpočet pomocí systému STATISTICA

Otevřeme nový datový soubor o dvou proměnných d a h a jednom případu.

Do Dlouhého jména proměnné d napíšeme

=34,48-35,59-
sqrt((24*1,7482+9*1,7121)/33)*sqrt((1/25)+(1/10))*VStudent(0,975;33)

Do Dlouhého jména proměnné h napíšeme

=34,48-35,59+
sqrt((24*1,7482+9*1,7121)/33)*sqrt((1/25)+(1/10))*VStudent(0,975;33)

	1 d	2 h
1	-2,11368	-0,10632

S pravděpodobností aspoň 0,95 tedy -2,114 g/l < $\mu_1 - \mu_2$ < -0,106 g/l.

Příklad

V předešlém příkladě nyní předpokládáme, že dané dva náhodné výběry pocházejí z rozložení $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$. Sestrojte 95% empirický interval spolehlivosti pro podíl rozptylů.

Řešení:

Úloha vede na vzorec 7.3. (d).

$$d = \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1, n_2-1)} = \frac{1{,}7482/1{,}7121}{F_{0{,}975}(24{,}9)} = \frac{1{,}7482/1{,}7121}{3{,}6142} = 0{,}28$$

$$h = \frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1, n_2-1)} = \frac{1{,}7482/1{,}7121}{F_{0{,}025}(24{,}9)} = \frac{1{,}7482/1{,}7121}{1/F_{0{,}975}(9{,}24)} = \frac{1{,}7482/1{,}7121}{1/2{,}7027} = 2{,}76$$

$0{,}28 < \frac{\sigma_1^2}{\sigma_2^2} < 2{,}76$ s pravděpodobností aspoň 0,95.

Výpočet pomocí systému STATISTICA

Otevřeme nový datový soubor o dvou proměnných d a h a jednom případu.

Do Dlouhého jména proměnné d napíšeme

=(1,7482/1,7121)/VF(0,975;24;9)

(Funkce VF(x;ný;omega) počítá x-kvantil Fisherova – Snedecorova rozložení F(ný, omega).)

Do Dlouhého jména proměnné h napíšeme

=(1,7482/1,7121)/VF(0,025;24;9)

	1 d	2 h
1	0,28252	2,75969

S pravděpodobností aspoň 0,95 tedy platí: $0{,}28 < \sigma_1^2/\sigma_2^2 < 2{,}76$.

Jednotlivé typy testů o parametrických funkcích μ_1-μ_2 , σ_1^2/σ_2^2

a) Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$ a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr z rozložení $N(\mu_2, \sigma_2^2)$, přičemž $n_1 \geq 2$, $n_2 \geq 2$ a σ_1^2, σ_2^2 známe. Nechť c je konstanta.

Test H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 \neq c$ se nazývá dvouvýběrový z-test.

b) Nechť $X_{11} \ldots, X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma^2)$ a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr rozložení $N(\mu_2, \sigma^2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$ a σ^2 neznáme. Nechť c je konstanta.

Test H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 \neq c$ se nazývá dvouvýběrový t-test.

c) Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$ a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr rozložení $N(\mu_2, \sigma_2^2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$.

Test H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$ proti H_1: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$ se nazývá F-test.

Provedení testů o parametrických funkcích μ_1-μ_2, σ_1^2/σ_2^2 pomocí kritického oboru (1)

a) Provedení dvouvýběrového z-testu

Vypočteme realizaci testového kritéria $t_0 = \frac{(M_1 - M_2) - c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$. Stanovíme kritický obor W. Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 \neq c$. Kritický obor má tvar: $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$.

Levostranný test: Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 < c$. Kritický obor má tvar: $W = (-\infty, -u_{1-\alpha}\rangle$.

Pravostranný test: Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 > c$. Kritický obor má tvar: $W = \langle u_{1-\alpha}, \infty)$.

Provedení testů o parametrických funkcích μ_1-μ_2, σ_1^2/σ_2^2 pomocí kritického oboru (2)

b) Provedení dvouvýběrového t-testu

Vypočteme realizaci testového kritéria $t_0 = \frac{(M_1 - M_2) - c}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$. Stanovíme kritický obor W. Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 \neq c$. Kritický obor má tvar: $W = (-\infty, -t_{1-\alpha/2}(n_1 + n_2 - 2)\rangle \cup \langle t_{1-\alpha/2}(n_1 + n_2 - 2), \infty)$.

Levostranný test: Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 < c$. Kritický obor má tvar: $W = (-\infty, -t_{1-\alpha}(n_1 + n_2 - 2)\rangle$.

Pravostranný test: Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 > c$. Kritický obor má tvar: $W = \langle t_{1-\alpha}(n_1 + n_2 - 2), \infty)$.

Provedení testů o parametrických funkcích μ_1-μ_2, ${\sigma_1}^2/{\sigma_2}^2$ pomocí kritického oboru (3)

c) Provedení F-testu

Vypočteme realizaci testového kritéria $t_0 = \frac{{s_1}^2}{{s_2}^2}$. Stanovíme kritický obor W. Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: $\frac{{\sigma_1}^2}{{\sigma_2}^2} = 1$ proti H_1: $\frac{{\sigma_1}^2}{{\sigma_2}^2} \neq 1$. Kritický obor má tvar: $W = (0, F_{\alpha/2}(n_1 - 1, n_2 - 1)\rangle \cup \langle F_{1-\alpha/2}(n_1 - 1, n_2 - 1), \infty)$.

Levostranný test: Testujeme H_0: $\frac{{\sigma_1}^2}{{\sigma_2}^2} = 1$ proti H_1: $\frac{{\sigma_1}^2}{{\sigma_2}^2} < 1$. Kritický obor má tvar: $W = (0, F_{\alpha}(n_1 - 1, n_2 - 1)\rangle$.

Pravostranný test: Testujeme H_0: $\frac{{\sigma_1}^2}{{\sigma_2}^2} = 1$ proti H_1: $\frac{{\sigma_1}^2}{{\sigma_2}^2} > 1$. Kritický obor má tvar: $W = \langle F_{1-\alpha}(n_1 - 1, n_2 - 1), \infty)$.

Příklad (1)

V restauraci "U bílého koníčka" měřili ve 20 případech čas obsluhy zákazníka. Výsledky v minutách: 6, 8, 11, 4, 7, 6, 10, 6, 9, 8, 5, 12, 13, 10, 9, 8, 7, 11, 10, 5. V restauraci "Zlatý lev" bylo dané pozorování uskutečněno v 15 případech s těmito výsledky: 9, 11, 10, 7, 6, 4, 8, 13, 5, 15, 8, 5, 6, 8 ,7. Za předpokladu, že uvedené hodnoty pocházejí ze dvou normálních rozložení, na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty doby obsluhy jsou v obou restauracích stejné.

Řešení:

Na hladině významnosti 0,05 testujeme nulovou hypotézu H_0: $\mu_1 - \mu_2 = 0$ proti oboustranné alternativě H_1: $\mu_1 - \mu_2 \neq 0$. Je to úloha na dvouvýběrový t-test. Před provedením tohoto testu je však nutné pomocí F-testu ověřit shodu rozptylů. Na hladině významnosti 0,05 tedy testujeme H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$ proti H_1: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$. Nejprve vypočteme $m_1 = 8{,}25$, $m_2 = 8{,}13$, $s_1^2 = 6{,}307$, $s_2^2 = 9{,}41$, $s_*^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{19 \cdot 6{,}307+14 \cdot 9{,}41}{33} = 7{,}623$. Podle 7.6. (c) vypočteme realizaci testové statistiky: $t_0 = \frac{s_1^2}{s_2^2} = \frac{6{,}307}{9{,}41} = 0{,}6702$.

Příklad (2)

Stanovíme kritický obor:

$$W = \langle 0, F_{\alpha/2}(n_1 - 1, n_2 - 1)\rangle \cup \langle F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1), \infty\rangle = \langle 0, F_{0,025}(19,14)\rangle \cup$$

$$\cup \langle F_{0,925}(19,14), \infty) = \langle 0,1/F_{0,925}(14,19)\rangle \cup \langle F_{0,925}(19,14), \infty\rangle =$$

$$= \langle 0,1/2,649\rangle \cup \langle 2,8607, \infty) = \langle 0; 0,3778\rangle \cup \langle 2,8607, \infty)$$

Protože se testová statistika nerealizuje v kritickém oboru, nulovou hypotézu nezamítáme na hladině významnosti 0,05. Rozptyly tedy můžeme považovat za shodné.

Nyní se vrátíme k dvouvýběrovému t-testu. Vypočteme realizaci testové statistiky:

$t_0 = \frac{m_1 - m_2 - c}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{8,25 - 8,13}{\sqrt{7,623}\sqrt{\frac{1}{20} + \frac{1}{15}}} = 0,124$. Stanovíme kritický obor:

$$W = \left(-\infty, -t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)\right\rangle \cup \left\langle t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2), \infty\right) = \left(-\infty, -t_{0,975}(33)\right\rangle \cup$$

$$\langle t_{0,975}(33), \infty\rangle = (-\infty, -2,035\rangle \cup \langle 2,035, \infty)$$

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

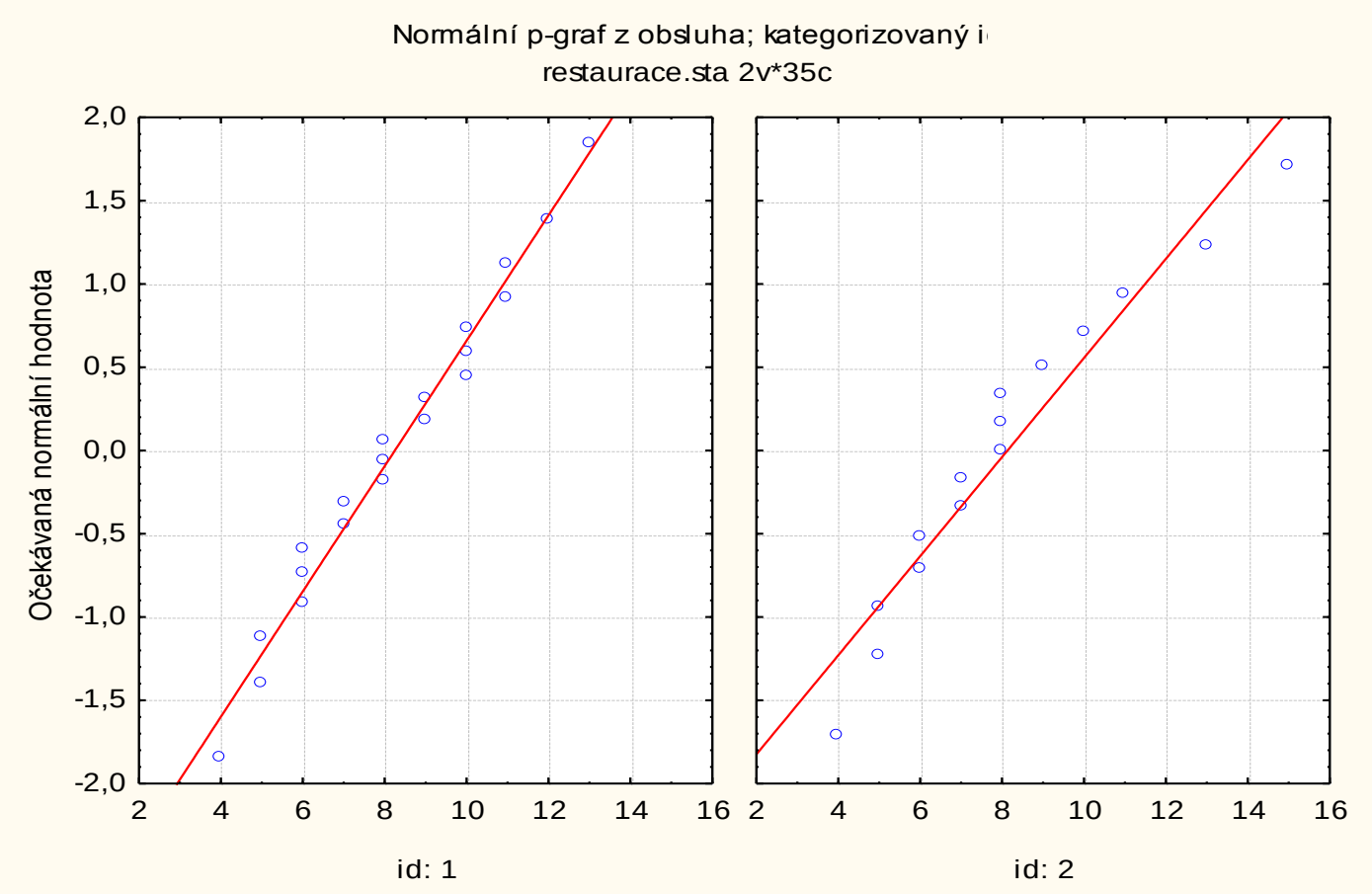
Výpočet pomocí systému STATISTICA (1)

Otevřeme nový datový soubor o dvou proměnných a 35 případech. První proměnnou nazveme OBSLUHA, druhou ID. Do proměnné OBSLUHA napíšeme nejprve doby obsluhy v první restauraci a poté doby obsluhy ve druhé restauraci. Do proměnné ID, která slouží k rozlišení první a druhé restaurace, napíšeme 20 krát jedničku a 15 krát dvojku.

Pomocí NP-grafu ověříme normalitu dat v obou skupinách. Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnné OBSLUHA, OK, Kategorizovaný – Kategorie X, zaškrtneme Zapnuto, Změnit proměnnou – ID, OK.

Výpočet pomocí systému STATISTICA (2)

Dostaneme graf



Výpočet pomocí systému STATISTICA (3)

V obou případech se tečky odchylují od přímky jenom málo. Předpoklad o normálním rozložení dat v obou skupinách je oprávněný.

Nyní provedeme dvouvýběrový t-test současně s testem o shodě rozptylů:

Statistika – Základní statistiky a tabulky – t-test, nezávislé, dle skupin – OK, Proměnné –Závislé proměnné OBSLUHA, Grupovací proměnná ID – OK.

Po kliknutí na tlačítko Souhrn dostaneme tabulku

	t-testy; grupováno: ID (restaurace) Skup. 1: 1 Skup. 2: 2										
Proměnná	Průměr 1	Průměr 2	t	sv	p	Poč.plat 1	Poč.plat. 2	Sm.odch. 1	Sm.odch. 2	F-poměr rozptyly	p rozptyly
OBSLUHA	8,25000	8,13333	0,12373	33	0,90227	20	15	2,51050	3,06749	1,49295	0,41044

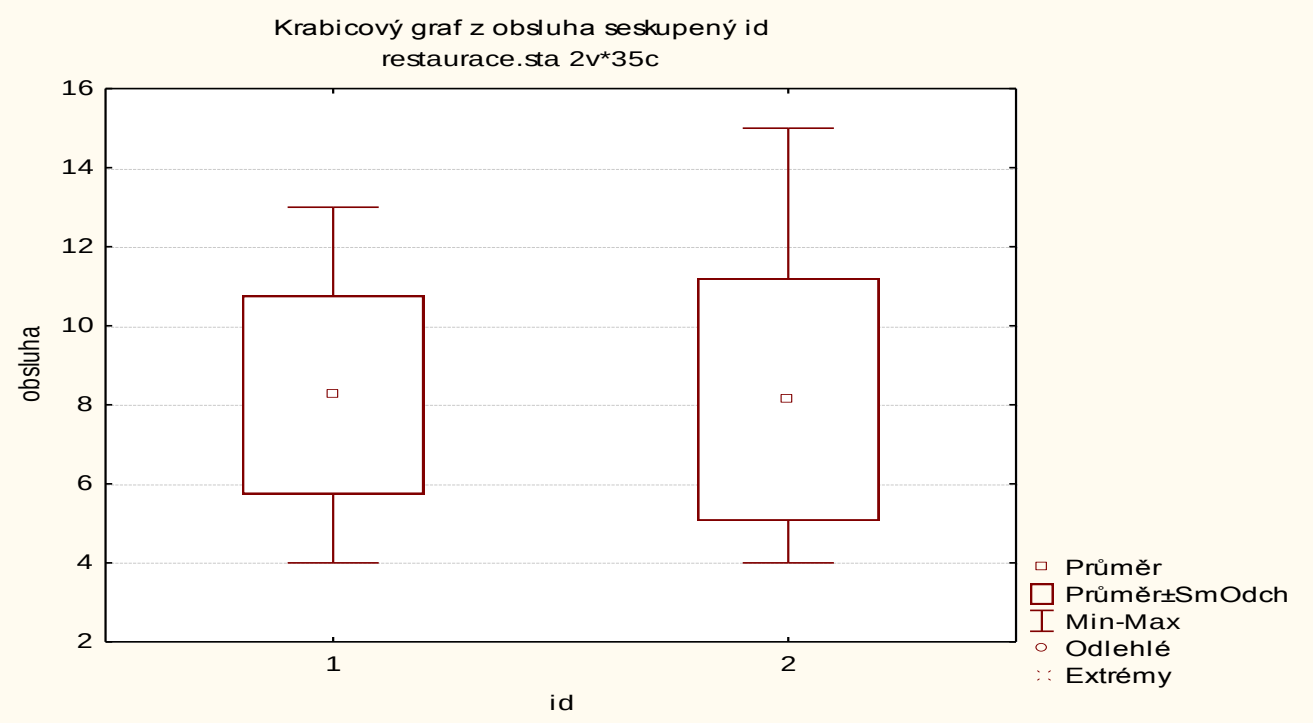
Výpočet pomocí systému STATISTICA (4)

Vidíme, že testová statistika pro test shody rozptylů se realizuje hodnotou 1,492952 (je to převrácená hodnota k číslu 0,6702, které jsme vypočítali při ručním postupu), odpovídající p-hodnota je 0,41044, tedy na hladině významnosti 0,05 nezamítáme hypotézu o shodě rozptylů. (Upozornění: v případě zamítnutí hypotézy o shodě rozptylů je zapotřebí v tabulce t-testu pro nezávislé vzorky dle skupin zaškrtnout volbu Test se samostatnými odhady rozptylu.)

Dále z tabulky plyne, že testová statistika pro test shody středních hodnot se realizuje hodnotou 0,12373, počet stupňů volnosti je 33, odpovídající p-hodnota 0,902279, tedy hypotézu o shodě středních hodnot nezamítáme na hladině významnosti 0,05. Znamená to, že s rizikem omylu nejvýše 5% se neprokázal rozdíl ve středních hodnotách dob obsluhy v restauracích "U bílého koníčka" a „Zlatý lev“.

Výpočet pomocí systému STATISTICA (5)

Tabulku ještě doplníme krabicovými diagramy. Na záložce Detaily zaškrtneme krabicový graf a vybereme volbu Průměr/SmOdch/Min-Max.



Z grafu je vidět, že průměrná doba obsluhy v první restauraci je nepatrně delší a má menší variabilitu než ve druhé restauraci. Extrémní ani odlehlé hodnoty se zde nevyskytují.

Cohenův koeficient věcného účinku – doplnění významu dvouvýběrového t-testu (1)

Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma^2)$ a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr rozložení $N(\mu_2, \sigma^2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$ a σ^2 neznáme. Nechť c je konstanta.

Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 \neq c$. Označme m_1, m_2 realizace výběrových průměrů hodnot dané veličiny v těchto dvou skupinách, s_1^2, s_2^2 realizace výběrových rozptylů a $s_*^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ realizaci váženého průměru výběrových rozptylů.

Cohenův koeficient d vypočteme podle vzorce: $d = \frac{|m_1-m_2|}{s_*}$.

Tento koeficient slouží k posouzení velikosti rozdílu průměrů, který je standardizován pomocí odmocniny z váženého průměru výběrových rozptylů. Jedná se o tzv. věcnou významnost neboli velikost účinku skupiny na variabilitu hodnot sledované náhodné veličiny.

Cohenův koeficient věcného účinku – doplnění významu dvouvýběrového t-testu (2)

Velikost účinku hodnotíme podle následující tabulky:

Hodnota d	účinek
aspoň 0,8	velký
mezi 0,5 až 0,8	střední
mezi 0,2 až 0,5	malý
pod 0,2	zanedbatelný

(Uvedené hodnoty nemají samozřejmě absolutní platnost, posouzení, jaký účinek považujeme za velký či malý, závisí na kontextu.)

Je zapotřebí si uvědomit, že při dostatečně velkých rozsazích náhodných výběrů i malý rozdíl ve výběrových průměrech způsobí zamítnutí nulové hypotézy na hladině významnosti α, i když z věcného hlediska tak malý rozdíl nemá význam. Naopak, máme-li výběry malých rozsahů, pak i značně velký rozdíl ve výběrových průměrech nemusí vést k zamítnutí nulové hypotézy na hladině významnosti α.

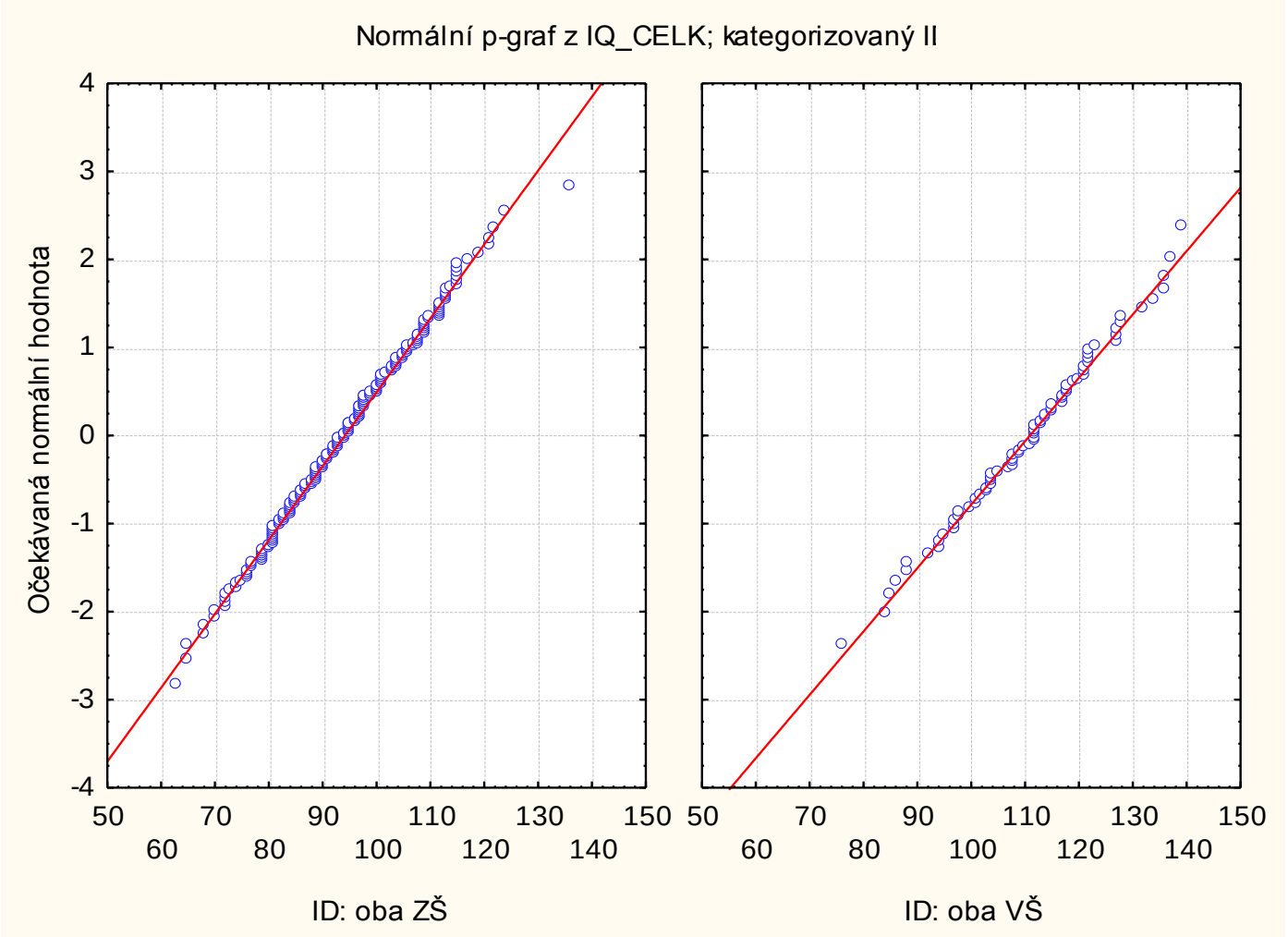
Příklad (1)

Máme k dispozici údaje o celkovém IQ 856 žáků ZŠ. Zajímáme se jednak o skupinu dětí, jejichž oba rodiče mají pouze základní vzdělání (je jich 296) a jednak o skupinu dětí, jejichž oba rodiče mají vysokoškolské vzdělání (těch je 75). Na hladině významnosti 0,05 budeme testovat hypotézu, že střední hodnota celkového IQ je v obou skupinách stejná a také vypočteme Cohenův koeficient věcného účinku.

Řešení:

Normalitu dat v obou skupinách posoudíme pomocí N-P plotu:

Příklad (2)


Normální p-graf z IQ_CELK; kategorizovaný II
Očekávaná normální hodnota
4
3
2
1
0
-1
-2
-3
-4
50
60
70
80
90
100
110
120
130
140
150
ID: oba ZŠ
ID: oba VŠ

Příklad (3)

Vzhled N- P plotů v obou skupinách podporuje domněnku o normalitě dat.

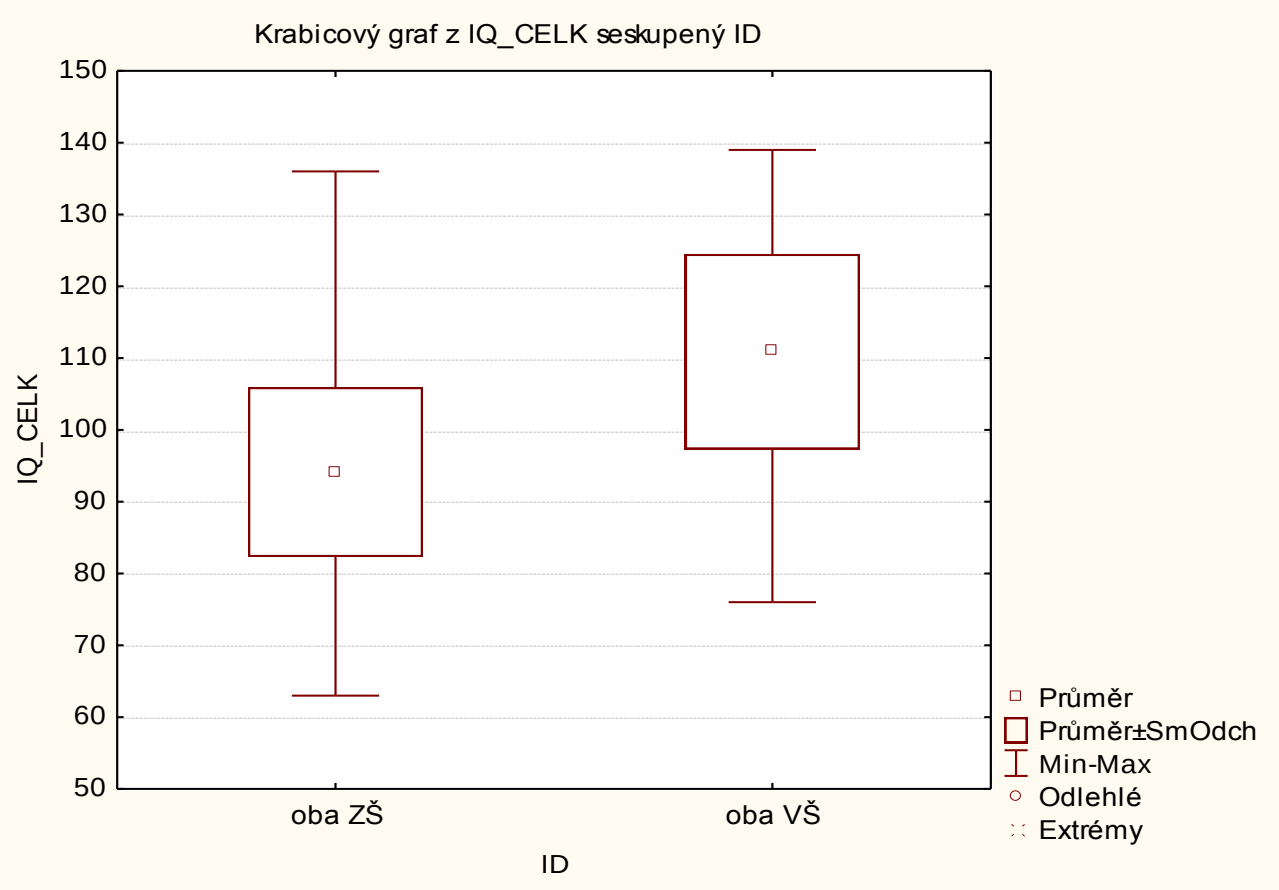
Provedeme dvouvýběrový t-test:

Proměnná	t-testy; grupováno:ZŠ a VŠ (IQ) Skup. 1: oba ZŠ Skup. 2: oba VŠ Průměr oba ZŠ	Průměr oba VŠ	t	sv	p	Poč.plat oba ZŠ	Poč.plat. oba VŠ	Sm.odch. oba ZŠ	Sm.odch. oba VŠ	F-poměr Rozptyly	p Rozptyly
IQ_CELK	94,1385	110,906	-10,629	369	0,00000	296	75	11,8260	13,6016	1,32282	0,11012

Hypotézu o shodě středních hodnot zamítáme na hladině významnosti 0,05, protože odpovídající p-hodnota je velmi blízká 0 (hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05, p-hodnota F-testu je 0,110124, což je větší než 0,05).

Příklad (4)

Krabicový diagram:



Příklad (5)

Vidíme, že průměrné celkové IQ dětí v 1. skupině je 94,1, zatímco ve 2. skupině 110,9. Vliv skupiny na variabilitu hodnot celkového IQ posoudíme pomocí Cohenova koeficientu.

	1 n1	2 n2	3 m1	4 m2	5 s1	6 s2	7 d
1	296	75	94,1385	110,906	11,8260	13,6016	1,37411

Cohenův koeficient nabývá hodnoty 1,37, tudíž vliv skupiny na variabilitu hodnot celkového IQ lze považovat za velký.

8. Parametrické úlohy o jednom náhodném výběru a dvou nezávislých náhodných výběrech z alternativních rozložení

Opakování:

Alternativní rozložení: Náhodná veličina X udává počet úspěchů v jednom pokusu, přičemž pravděpodobnost úspěchu je ϑ. Píšeme $X \sim A(\vartheta)$.

$$\pi(x) = \begin{cases} 1-\vartheta \text{ pro } x = 0 \\ \vartheta \text{ pro } x = 1 \\ 0 \text{ jinak} \end{cases} \quad \text{neboli } \pi(x) = f(x) = \begin{cases} \vartheta^x (1-\vartheta)^{1-x} \text{ pro } x = 0, 1 \\ 0 \text{ jinak} \end{cases}$$

Binomické rozložení: Náhodná veličina X udává počet úspěchů v posloupnosti n nezávislých opakovaných pokusů, přičemž pravděpodobnost úspěchu je v každém pokusu ϑ. Píšeme $X \sim Bi(n,\vartheta)$.

$$\pi(x) = \begin{cases} \binom{n}{x} \vartheta^x (1-\vartheta)^{n-x} \text{ pro } x = 0, \ldots, n \\ 0 \text{ jinak} \end{cases}$$

$E(X) = n\vartheta$, $D(X) = n\vartheta(1-\vartheta)$

(Alternativní rozložení je speciálním případem binomického rozložení pro $n = 1$.

Jsou-li $X_1, \ldots, X_n$ stochasticky nezávislé náhodné veličiny, $X_i \sim A(\vartheta)$, $i = 1, \ldots, n$, pak $X = \sum_{i=1}^{n} X_i \sim Bi(n, \vartheta)$.)

Centrální limitní věta

Jsou-li náhodné veličiny $X_1, \ldots, X_n$ stochasticky nezávislé a všechny mají stejné rozložení se střední hodnotou μ a rozptylem σ^2, pak pro velká n ($n \geq 30$) lze rozložení součtu $\sum_{i=1}^{n} X_i$ aproximovat normálním rozložením $N(n\mu, n\sigma^2)$. Zkráceně píšeme $\sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2)$. Pokud součet $\sum_{i=1}^{n} X_i$ standardizujeme, tj. vytvoříme náhodnou veličinu

$U_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$, pak rozložení této náhodné veličiny lze aproximovat standardizovaným normálním rozložením. Zkráceně píšeme $U_n \approx N(0,1)$

Asymptotické rozložení statistiky odvozené z výběrového průměru

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $A(\vartheta)$ a nechť je splněna podmínka $n\vartheta(1-\vartheta) > 9$. Pak statistika $U = \frac{M-\vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}}$ konverguje v distribuci k náhodné veličině se standardizovaným normálním rozložením. (Říkáme, že U má asymptoticky rozložení N(0,1) a píšemu $U \approx N(0,1)$.)

Důkaz:

Protože $X_1, ..., X_n$ je náhodný výběr z rozložení $A(\vartheta)$, bude mít statistika $Y_n = \sum_{i=1}^{n} X_i$ (výběrový úhrn) rozložení $Bi(n, \vartheta)$. Y_n má střední hodnotu $E(Y_n) = n\vartheta$ a rozptyl $D(Y_n) = n\vartheta(1-\vartheta)$. Podle centrální limitní věty se standardizovaná statistika $U = \frac{Y_n - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}$ asymptoticky řídí standardizovaným normálním rozložením N(0,1). Pokud čitatele i jmenovatele podělíme n, dostaneme vyjádření: $U = \frac{\frac{Y_n}{n}-\vartheta}{\sqrt{\frac{n\vartheta(1-\vartheta)}{n^2}}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}} = \frac{M-\vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}} \approx N(0,1)$

Vzorec pro meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametr ϑ

Meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametr ϑ jsou: $d = m - \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2}$, $\mathrm{h} = m + \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2}$.

Důkaz:

Pokud rozptyl $D(M) = \frac{\vartheta(1-\vartheta)}{n}$ nahradíme odhadem $\frac{M(1-M)}{n}$, konvergence náhodné veličiny U k veličině s rozložením N(0,1) se neporuší. Tedy

$$\forall\vartheta \in \Xi: 1-\alpha \leq P\left(-u_{1-\alpha/2} < \frac{M-\vartheta}{\sqrt{\frac{M(1-M)}{n}}} < u_{1-\alpha/2}\right) =$$

$$= P\left(M - \sqrt{\frac{M(1-M)}{n}}u_{1-\alpha/2} < \vartheta < M + \sqrt{\frac{M(1-M)}{n}}u_{1-\alpha/2}\right)$$

Příklad

Náhodně bylo vybráno 100 osob a zjištěno, že 34 z nich používá zubní kartáček zahraniční výroby. Najděte 95% asymptotický interval spolehlivosti pro pravděpodobnost, že náhodně vybraná osoba používá zubní kartáček zahraniční výroby.

Řešení:

Zavedeme náhodné veličiny $X_1, \ldots, X_{100}$, přičemž $X_i = 1$, když i-tá osoba používá zahraniční zubní kartáček a $X_i = 0$ jinak, $i = 1, \ldots, 100$. Tyto náhodné veličiny tvoří náhodný výběr z rozložení $A(\vartheta)$.

$n = 100$, $m = 34/100$, $\alpha = 0{,}05$, $u_{1-\alpha/2} = u_{0{,}975} = 1{,}96$.

Ověření podmínky $n\vartheta(1 - \vartheta) > 9$: parametr ϑ neznáme, musíme ho nahradit výběrovým průměrem. Pak $100.0{,}34.0{,}66 = 22{,}44 > 9$.

$$d = 0{,}34 - \sqrt{\frac{0{,}34(1-0{,}34)}{100}}\,1{,}96 = 0{,}2472,\ h = 0{,}34 + \sqrt{\frac{0{,}34(1-0{,}34)}{100}}\,1{,}96 = 0{,}4328.$$

S pravděpodobností přibližně 0,95 tedy $0{,}2472 < \vartheta < 0{,}4328$.

Výpočet pomocí systému STATISTICA (1)

a) Přesný způsob

Otevřeme nový datový soubor se dvěma proměnnými a jednom případu.

První proměnnou nazveme d a do jejího Dlouhého jména napíšeme

=0,34-sqrt(0,34*0,66/100)*VNormal(0,975;0;1)

Druhou proměnnou nazveme h a do jejího Dlouhého jména napíšeme

=0,34+sqrt(0,34*0,66/100)*VNormal(0,975;0;1)

Dostaneme výsledek:

	1 d	2 h
1	0,24715	0,43284

Vidíme, že s pravděpodobností aspoň 0,95 se pravděpodobnost používání zubního kartáčku zahraniční výroby bude pohybovat v mezích 0,2471 až 0,4328.

Výpočet pomocí systému STATISTICA (2)

b) Přibližný způsob, použitelný pro dostatečně velký rozsah výběru

Do nového datového souboru o jedné proměnné X a 100 případech uložíme 34 jedniček (indikují používání zubního kartáčku zahraniční výroby) a 66 nul (indikují používání zubního kartáčku domácí výroby).

Statistika – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme Meze spolehl. prům. – ponecháme implicitní hodnotu pro Interval 95,00 – Výpočet.

Dostaneme tabulku:

	Popisné statistiky (Tabulka3)			
Proměnná	N platných	Průměr	Int. spolehl. -95,000%	Int. spolehl. 95,000
X	100	0,340000	0,245533	0,434468

Dospěli jsme k výsledku, že s pravděpodobností aspoň 0,95 se pravděpodobnost používání zubního kartáčku zahraniční výroby bude pohybovat v mezích 0,2455 až 0,4345.

Výpočet pomocí systému STATISTICA (3)

c) Výpočet pomocí modulu Analýza síly testu

Statistiky – Analýza síly testu – Odhad intervalu – Jeden podíl, Z, Chí-kvadrát test – OK – Pozorovaný podíl p: 0,34, Velikost vzorku: 100, Spolehlivost: 0,95 – Vypočítat.

Dostaneme tabulku:

	Hodnota
Podíl vzorku p	0,340(
Velikost vz. ve skup. (N)	100,000(
Interval spolehlivosti	0,950(
Meze spolehlivosti:	
Pí (přesně):	
Dolní mez	0,248
Horní mez	0,441
Pí (přibližně):	
Dolní mez	0,250
Horní mez	0,442
Pí (původ.):	
Dolní mez	0,247
Horní mez	0,432

Zajímá nás výsledek uvedený v dolní části tabulky, tj. Pí (původ.). Zjišťujeme, že s pravděpodobností aspoň 0,95 se hledaná pravděpodobnost bude pohybovat v mezích 0,2472 až 0,4328.

Příklad (1)

Kolik osob musíme vybrat, abychom podíl modrookých osob v populaci odhadli se spolehlivostí 90% a šířka intervalu spolehlivosti byla nanejvýš a) 0,06, b) 0,01?

Řešení:

Šířka 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametr ϑ:

$$h - d = m + \sqrt{\frac{m(1-m)}{n}} u_{1-\frac{\alpha}{2}} - \left(m - \sqrt{\frac{m(1-m)}{n}} u_{1-\frac{\alpha}{2}} \right) = 2\sqrt{\frac{m(1-m)}{n}} u_{1-\alpha/2}$$

Požadujeme, aby $h - d \leq \Delta$, tedy $2\sqrt{\frac{m(1-m)}{n}} u_{1-\alpha/2} \leq \Delta$. Odtud vyjádříme

$$n \geq \frac{4m(1-m){u_{1-\alpha/2}}^2}{\Delta^2}.$$

Příklad (2)

Předpokládejme, že nemáme žádné předběžné informace o podílu modrookých osob v populaci. Musíme tedy zvolit takové m, aby šířka intervalu spolehlivosti byla maximální. Maximalizujeme výraz $m(1-m) = m - m^2$. Derivujeme podle m a položíme rovno 0: $1 - 2m = 0 \Rightarrow m = \frac{1}{2}$.V tomto případě volíme relativní četnost m = 0,5.

ad a) $n \geq \frac{4m(1-m){u_{1-\alpha/2}}^2}{\Delta^2} = \frac{4 \cdot 0{,}5 \cdot 0{,}5 \cdot {u_{0{,}95}}^2}{0{,}06^2} = \frac{4 \cdot 0{,}5 \cdot 0{,}5 \cdot 1{,}645^2}{0{,}06^2} = 751{,}67$

Uvedenou podmínku tedy splníme, když vybereme aspoň 752 osob.

ad b) $n \geq \frac{4m(1-m){u_{1-\alpha/2}}^2}{\Delta^2} = \frac{4 \cdot 0{,}5 \cdot 0{,}5 \cdot {u_{0{,}95}}^2}{0{,}01^2} = \frac{4 \cdot 0{,}5 \cdot 0{,}5 \cdot 1{,}645^2}{0{,}01^2} = 27060{,}25$

Chceme-li dosáhnout podstatně užšího intervalu spolehlivosti, musíme vybrat aspoň 27 061 osob.

Příklad (3)

Modifikace: Předpokládejme, že v populaci je nanejvýš 30% modrookých osob. Pak relativní četnost m = 0,3.

ad a) $n \geq \frac{4m(1-m)u_{1-\alpha/2}{}^2}{\Delta^2} = \frac{4 \cdot 0{,}3 \cdot 0{,}7 \cdot u_{0,95}{}^2}{0{,}06^2} = \frac{4 \cdot 0{,}3 \cdot 0{,}7 \cdot 1{,}645^2}{0{,}06^2} = 631{,}41$

V tomto případě stačí vybrat 632 osob.

Ve srovnání s předešlým případem vidíme, že rozsah výběru skutečně klesl.

ad b)

$$n \geq \frac{4m(1-m)u_{1-\alpha/2}{}^2}{\Delta^2} = \frac{4 \cdot 0{,}3 \cdot 0{,}7 \cdot u_{0,95}{}^2}{0{,}01^2} = \frac{4 \cdot 0{,}3 \cdot 0{,}7 \cdot 1{,}645^2}{0{,}01^2} = 22730{,}61$$

V tomto případě musíme vybrat aspoň 22 731 osob.

Testování hypotézy o parametru ϑ

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $A(\vartheta)$ a nechť je splněna podmínka $n\vartheta(1-\vartheta) > 9$. Na asymptotické hladině významnosti α testujeme hypotézu H_0: $\vartheta = c$ proti alternativě H_1: $\vartheta \neq c$ (resp. H_1: $\vartheta < c$ resp. H_1: $\vartheta > c$). Testovým kritériem je statistika $T_0 = \frac{M-c}{\sqrt{\frac{c(1-c)}{n}}}$, která v případě platnosti nulové hypotézy má asymptoticky rozložení N(0,1). Kritický obor má tvar $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$ (resp. $W = (-\infty, -u_{1-\alpha}\rangle$ resp. $W = \langle u_{1-\alpha}, \infty)$).

(Testování hypotézy o parametru ϑ lze samozřejmě provést i pomocí 100(1-α)% asymptotického intervalu spolehlivosti nebo pomocí p-hodnoty.)

Příklad (1)

Podíl zmetků při výrobě určité součástky činí $\vartheta = 0{,}01$. Bylo náhodně vybráno 1000 výrobků a zjistilo se, že mezi nimi je 16 zmetků. Na asymptotické hladině významnosti 0,05 testujte hypotézu H_0: $\vartheta = 0{,}01$ proti oboustranné alternativě H_1: $\vartheta \neq 0{,}01$.

Řešení:

Zavedeme náhodné veličiny $X_1, \ldots, X_{1000}$, přičemž $X_i = 1$, když i-tý výrobek byl zmetek a $X_i = 0$ jinak, $i = 1, \ldots, 1000$. Tyto náhodné veličiny tvoří náhodný výběr z rozložení $A(\vartheta)$.

Testujeme hypotézu H_0: $\vartheta = 0{,}01$ proti alternativě H_1: $\vartheta \neq 0{,}01$.

Známe: $n = 1000$, $m = \frac{16}{1000} = 0{,}016$, $c = 0{,}01$, $\alpha = 0{,}05$, $u_{1-\alpha/2} = u_{0{,}975} = 1{,}96$

Ověření podmínky $n\vartheta(1-\vartheta) > 9$: $1000.0{,}01.0{,}99 = 9{,}9 > 9$.

a) Testování pomocí kritického oboru:

Realizace testového kritéria: $t_0 = \frac{m-c}{\sqrt{\frac{c\cdot(1-c)}{n}}} = \frac{0{,}016-0{,}01}{\sqrt{\frac{0{,}01\cdot 0{,}99}{1000}}} = 1{,}907.$

Kritický obor: $W = (-\infty, -u_{0{,}975}\rangle \cup \langle u_{0{,}975}, \infty) = (-\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty)$. Protože $1{,}907 \notin W$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Příklad (2)

b) Testování pomocí intervalu spolehlivosti

$$d = m - \sqrt{\frac{m(1-m)}{n}}\, u_{1-\alpha/2} = 0{,}016 - \sqrt{\frac{0{,}016 \cdot 0{,}984}{500}}\, 1{,}96 = 0{,}0082$$

$$h = m + \sqrt{\frac{m(1-m)}{n}}\, u_{1-\alpha/2} = 0{,}016 + \sqrt{\frac{0{,}016 \cdot 0{,}984}{500}}\, 1{,}96 = 0{,}0238$$

Protože číslo c = 0,01 leží v intervalu 0,0082 až 0,0238, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

c) Testování pomocí p-hodnoty

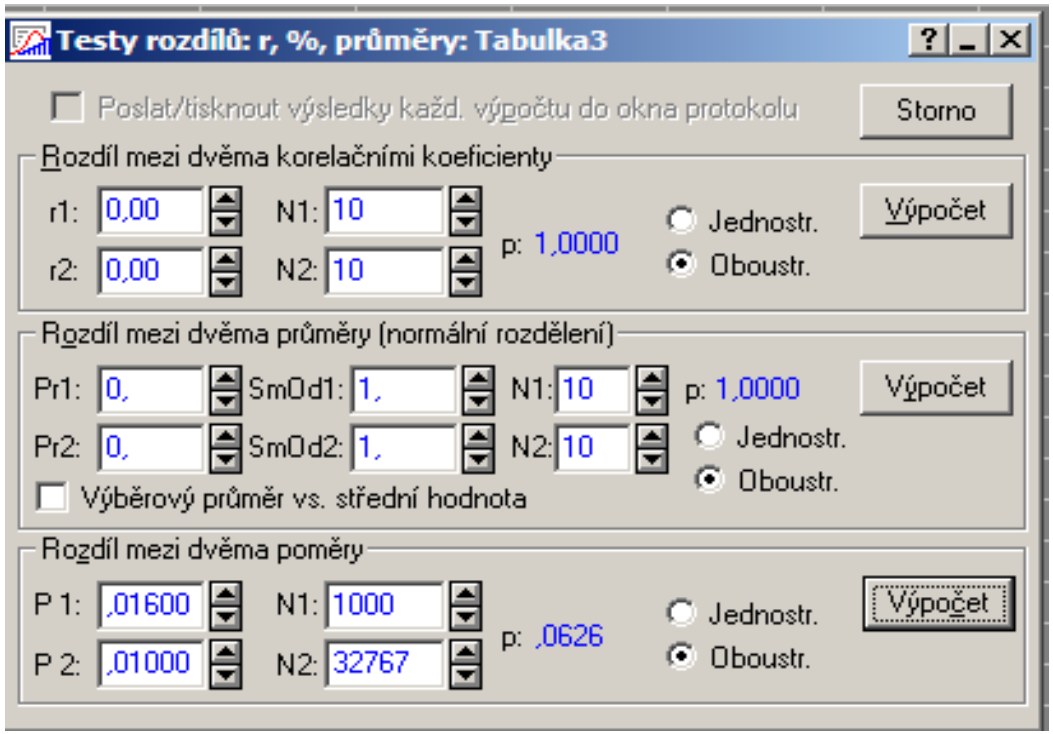
Protože testujeme nulovou hypotézu proti oboustranné alternativě, vypočteme p-hodnotu podle vzorce:

$p = 2 \min\{ \Phi(1{,}907), 1 - \Phi(1{,}907) \} = 2 \min \{ 0{,}97104, 1 - 0{,}97104 \} = 0{,}05792.$

Protože vypočtená p-hodnota je větší než hladina významnosti 0,05, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma poměry – do políčka P 1 napíšeme 0,016, do políčka N1 napíšeme 1000, do políčka P 2 napíšeme 0,01, do políčka N2 napíšeme 32767 (větší hodnotu systém neumožní) - Výpočet. Dostaneme p-hodnotu 0,0626, tedy nezamítáme nulovou hypotézu na hladině významnosti 0,05.



Příklad

Nový léčebný postup považujeme za úspěšný, pokud po jeho ukončení bude dosaženo zlepšení zdravotního stavu u alespoň 50% zúčastněných pacientů. Nová terapie byla vyzkoušena u 40 pacientů a ke zlepšení došlo u 24 osob. Je možné na asymptotické hladině významnosti 0,05 zamítnout hypotézu, že tato terapie nedosahuje úspěšnosti aspoň 50%?

Řešení:

Zavedeme náhodné veličiny $X_1, ..., X_{40}$, přičemž $X_i = 1$, když terapie u i-tého pacienta byl úspěšná a $X_i = 0$ jinak, $i = 1, ..., 40$. Tyto náhodné veličiny tvoří náhodný výběr z rozložení $A(\vartheta)$.

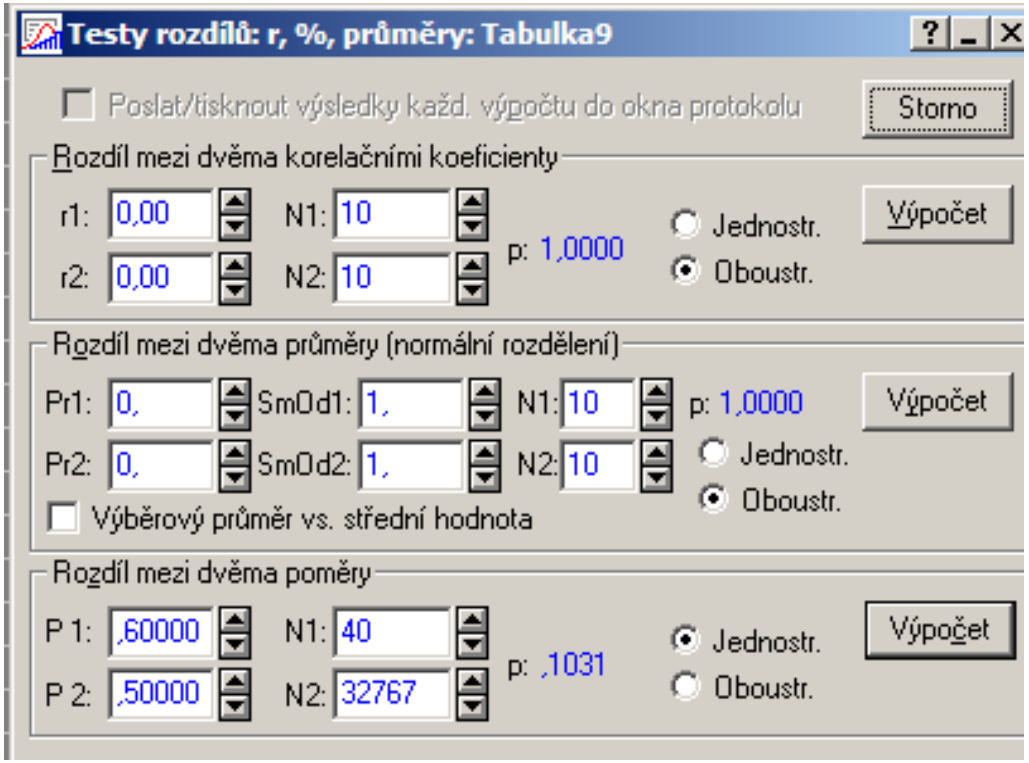
Testujeme hypotézu H_0: $\vartheta \leq 0{,}5$ proti pravostranné alternativě H_1: $\vartheta > 0{,}5$.

Známe: $n = 40$, $m = \frac{24}{40} = 0{,}6$, $c = 0{,}5$, $\alpha = 0{,}05$, $u_{1-\alpha} = u_{0{,}95} = 1{,}645$

Ověření podmínky $n\vartheta(1-\vartheta) > 9$: $40.0{,}6.0{,}4 = 9{,}6 > 9$.

Realizace testového kritéria: $t_0 = \frac{m-c}{\sqrt{\frac{c\cdot(1-c)}{n}}} = \frac{0{,}6-0{,}5}{\sqrt{\frac{0{,}5\cdot 0{,}5}{40}}} = 1{,}2649.$

Výpočet pomocí systému STATISTICA



Vypočtená p-hodnota jednostranného testu je 0,1031, tedy menší než asymptotická hladina významnosti 0,05. H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Asymptotické rozložení statistiky odvozené ze dvou výběrových průměrů

Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z alternativního rozložení $A(\vartheta_1)$ a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr alternativního rozložení $A(\vartheta_2)$ a nechť jsou splněny podmínky $n_1\vartheta_1(1-\vartheta_1) > 9$ a $n_2\vartheta_2(1-\vartheta_2) > 9$. Označme M_1, M_2 výběrové průměry.

Pak statistika $U = \frac{M_1 - M_2 - (\vartheta_1 - \vartheta_2)}{\sqrt{\frac{\vartheta_1(1-\vartheta_1)}{n_1} + \frac{\vartheta_2(1-\vartheta_2)}{n_2}}} \approx N(0,1)$.

Důkaz:

Analogicky jako v případě jednoho náhodného výběru z alternativního rozložení.

Vzorec pro meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$.

Meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro $\vartheta_1 - \vartheta_2$ jsou:

$$d = m_1 - m_2 - \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\, u_{1-\alpha/2},$$

$$\mathrm{h} = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\, u_{1-\alpha/2}$$

Důkaz:

Pokud rozptyl $D(M_i) = \frac{\vartheta_i(1-\vartheta_i)}{n_i}$ nahradíme odhadem $\frac{M_i(1-M_i)}{n_i}$, i = 1, 2, konvergence náhodné veličiny U k veličině s rozložením N(0,1) se neporuší. Tedy

$$\forall \vartheta_1 - \vartheta_2 \in \Xi: 1 - \alpha \leq P\left(-u_{1-\alpha/2} < \frac{M_1 - M_2 - (\vartheta_1 - \vartheta_2)}{\sqrt{\frac{M_1(1-M_1)}{n_1} + \frac{M_2(1-M_2)}{n_2}}} < u_{1-\alpha/2}\right) =$$

$$= P\left(M_1 - M_2 - \sqrt{\frac{M_1(1-M_1)}{n_1} + \frac{M_2(1-M_2)}{n_2}}\, u_{1-\alpha/2} < \vartheta_1 - \vartheta_2 < \right.$$

$$\left. < M_1 - M_2 + \sqrt{\frac{M_1(1-M_1)}{n_1} + \frac{M_2(1-M_2)}{n_2}}\, u_{1-\alpha/2}\right)$$

Příklad (1)

Management supermarketu vyhlásil týden slev a sledoval, zda toto vyhlášení má vliv na podíl větších nákupů (nad 500 Kč). Na základě náhodného výběru 200 zákazníků v týdnu bez slev bylo zjištěno 97 velkých nákupů, zatímco v týdnu se slevou z 300 náhodně vybraných zákazníků učinilo velký nákup 162 zákazníků. Sestrojte 95% asymptotický interval spolehlivosti pro rozdíl pravděpodobností uskutečnění většího nákupu v týdnu bez slevy a v týdnu se slevou.

Řešení:

Zavedeme náhodnou veličinu X_{1i}, která bude nabývat hodnoty 1, když v týdnu bez slevy i-tý náhodně vybraný zákazník uskuteční větší nákup a hodnoty 0 jinak, i = 1, …, 200. Náhodné veličiny $X_{1,1}$, …, $X_{1,200}$ tvoří náhodný výběr z rozložení $A(\vartheta_1)$. Dále zavedeme náhodnou veličinu X_{2i}, která bude nabývat hodnoty 1, když v týdnu se slevou i-tý náhodně vybraný zákazník uskuteční větší nákup a hodnoty 0 jinak, i = 1, …, 300. Náhodné veličiny $X_{2,1}$, …, $X_{2,300}$ tvoří náhodný výběr z rozložení $A(\vartheta_2)$.

$n_1 = 200$, $n_2 = 300$, $m_1 = 97/200 = 0{,}485$, $m_2 = 162/300 = 0{,}54$.

Ověření podmínek $n_1\vartheta_1(1-\vartheta_1) > 9$ a $n_1\vartheta_2(1-\vartheta_2) > 9$: Parametry ϑ_1 a ϑ_2 neznáme, nahradíme je odhady m_1 a m_2. $97.(1\text{-}97/200) = 49{,}955 > 9$, $162.(1\text{-}162/300) = 74{,}52 > 9$.

Příklad (2)

Meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$ jsou:

$$d = m_1 - m_2 - \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\, u_{1-\alpha/2} =$$

$$= \frac{97}{200} - \frac{162}{300} - \sqrt{\frac{\frac{97}{200}(1-\frac{97}{200})}{200} + \frac{\frac{162}{300}(1-\frac{162}{300})}{300}}\, 1{,}96 = -0{,}1443$$

$$\mathrm{h} = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\, u_{1-\alpha/2} =$$

$$= \frac{97}{200} - \frac{162}{300} + \sqrt{\frac{\frac{97}{200}(1-\frac{97}{200})}{200} + \frac{\frac{162}{300}(1-\frac{162}{300})}{300}}\, 1{,}96 = 0{,}0343$$

Zjistili jsme tedy, že s pravděpodobností přibližně 0,95: $-0{,}1443 < \vartheta_1 - \vartheta_2 < 0{,}0343$.

Testování hypotézy o parametrické funkci $\vartheta_1 - \vartheta_2$

Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z alternativního rozložení A(ϑ_1) a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr alternativního rozložení A(ϑ_2) a nechť jsou splněny podmínky $n_1\vartheta_1(1-\vartheta_1) > 9$ a $n_2\vartheta_2(1-\vartheta_2) > 9$. Na asymptotické hladině významnosti α testujeme nulovou hypotézu H_0: $\vartheta_1 - \vartheta_2 = c$ proti alternativě H_1: $\vartheta_1 - \vartheta_2 \neq c$ (resp. H_1: $\vartheta_1 - \vartheta_2 < c$ resp. H_1: $\vartheta_1 - \vartheta_2 > c$). Testovým kritériem je statistika

$T_0 = \frac{M_1 - M_2 - c}{\sqrt{\frac{M_1(1-M_1)}{n_1} + \frac{M_2(1-M_2)}{n_2}}}$, která v případě platnosti nulové hypotézy má asymptoticky rozložení N(0,1). Kritický obor má tvar $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$ (resp. $W = (-\infty, -u_{1-\alpha}\rangle$ resp. $W = \langle u_{1-\alpha}, \infty)$).

(Testování hypotézy o parametrické funkci $\vartheta_1 - \vartheta_2$ lze provést též pomocí 100(1-α)% asymptotického intervalu spolehlivosti nebo pomocí p-hodnoty.)

Postup při testování hypotézy $\vartheta_1 - \vartheta_2 = 0$

Je-li c = 0, pak označme $M_* = \frac{n_1 M_1 + n_2 M_2}{n_1 + n_2}$ vážený průměr výběrových průměrů. Jako testová statistika slouží $T_0 = \frac{M_1 - M_2}{\sqrt{M_*(1-M_*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, která v případě platnosti nulové hypotézy má asymptoticky rozložení N(0,1). Kritický obor má tvar $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$ (resp. $W = (-\infty, -u_{1-\alpha}\rangle$ resp. $W = \langle u_{1-\alpha}, \infty)$). Testová statistika T_0 vznikne standardizací statistiky $M_1 - M_2$, kde neznámé parametry ϑ_1, ϑ_2 nahradíme společným odhadem M_*.

Příklad (1)

Pro údaje z předchozího příkladu testujte na asymptotické hladině významnosti 0,05 hypotézu, že týden se slevami nezvýší pravděpodobnost uskutečnění většího nákupu.

Řešení:

Testujeme hypotézu $\vartheta_1 - \vartheta_2 = 0$ proti levostranné alternativě H_1: $\vartheta_1 - \vartheta_2 < 0$ na asymptotické hladině významnosti 0,05.

$n_1 = 200$, $n_2 = 300$, $m_1 = 97/200$, $m_2 = 162/300$, $m_* = (97 + 162)/500 = 0{,}518$.

Podmínky dobré aproximace byly ověřeny v příkladu 8.10.

Testování pomocí intervalu spolehlivosti:

Pro levostrannou alternativu používáme pravostranný interval spolehlivosti:

$$h = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\,u_{1-\alpha} =$$

$$= \frac{97}{200} - \frac{162}{300} + \sqrt{\frac{\frac{97}{200}\left(1-\frac{97}{200}\right)}{200} + \frac{\frac{162}{300}\left(1-\frac{162}{300}\right)}{300}}\,1{,}645 = 0{,}02$$

Protože číslo c = 0 je obsaženo v intervalu $(-\infty; 0{,}02\rangle$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Příklad (2)

Testování pomocí kritického oboru:

Realizace testového kritéria:

$$t_0 = \frac{m_1 - m_2}{\sqrt{m_*(1-m_*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{\frac{97}{200} - \frac{162}{300}}{\sqrt{0{,}518(1-0{,}518)\left(\frac{1}{200}+\frac{1}{300}\right)}} = -1{,}2058.$$

Kritický obor je $W = (-\infty, -u_{1-\alpha}\rangle = (-\infty, -u_{0,95}\rangle = (-\infty, -1{,}645\rangle$. Protože testové kritérium nepatří do kritického oboru, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Testování pomocí p-hodnoty:

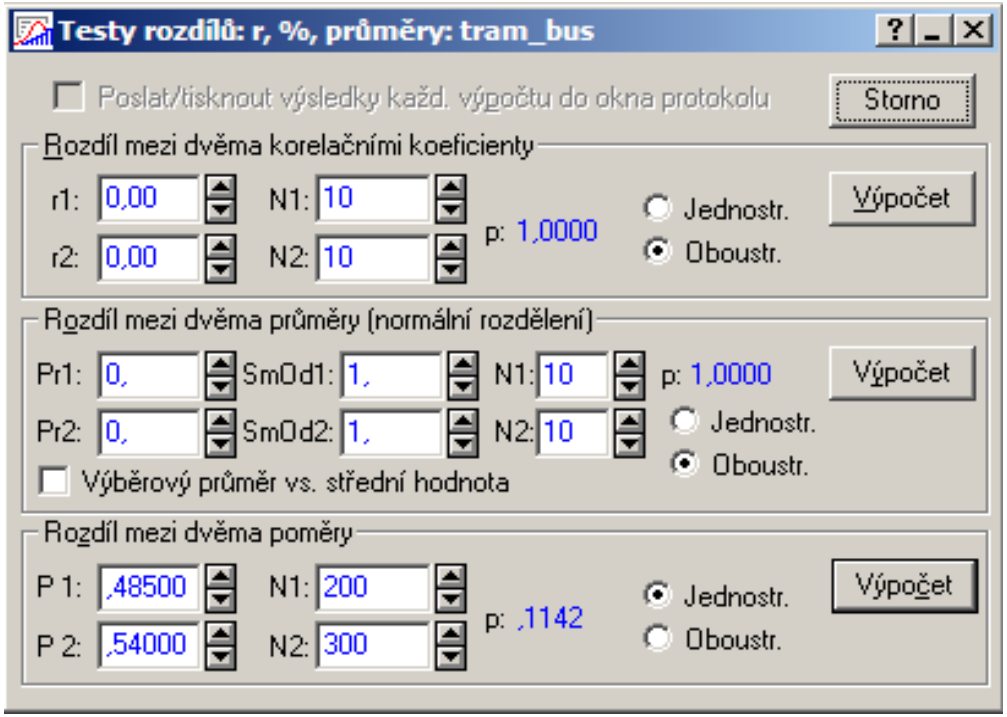
Pro levostrannou alternativu se p-hodnota počítá podle vzorce $p = P(T_0 \leq t_0)$:

$$p = P(T_0 \leq -1{,}2058) = \Phi(-1{,}2058) = 1 - \Phi(1{,}2058) = 1 - 0{,}8861 == 0{,}1139$$

Protože p-hodnota je větší než 0,05, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma poměry – do políčka P 1 napíšeme 0,485, do políčka N1 napíšeme 200, do políčka P 2 napíšeme 0,54, do políčka N2 napíšeme 300 – zaškrtneme Jednostr. - Výpočet. Dostaneme p-hodnotu 0,1142, tedy nezamítáme nulovou hypotézu na hladině významnosti 0,05.



9. Analýza rozptylu jednoduchého třídění

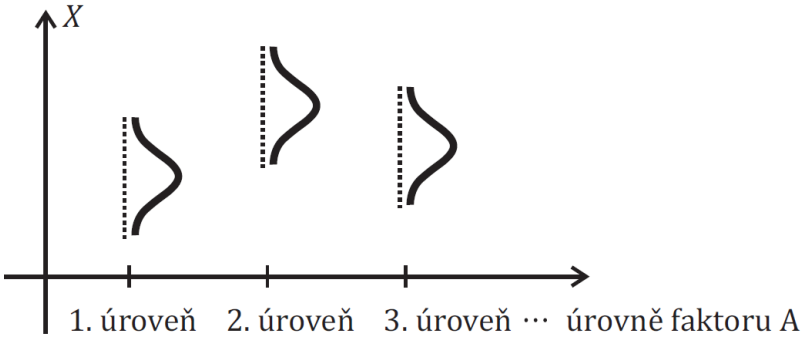
Motivace: Zajímáme se o problém, zda lze určitým faktorem (tj. nominální náhodnou veličinou A) vysvětlit variabilitu pozorovaných hodnot náhodné veličiny X, která je intervalového či poměrového typu. Např. zkoumáme, zda metoda výuky určitého předmětu (faktor A) ovlivňuje počet bodů dosažených studenty v závěrečném testu (náhodná veličina X).

Předpokládáme, že faktor A má $r \geq 3$ úrovní a přitom i-té úrovni odpovídá n_i pozorování $X_{i1}, \ldots, X_{in_i}$, které tvoří náhodný výběr z rozložení $N(\mu_i, \sigma^2)$, $i = 1, \ldots, r$ a jednotlivé náhodné výběry jsou stochasticky nezávislé, tedy $X_{ij} = \mu_i + \varepsilon_{ij}$, kde ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$, $i = 1, \ldots, r$, $j = 1, \ldots, n_i$.

Výsledky lze zapsat do tabulky

faktor A	výsledky
úroveň 1	$X_{11}, \ldots, X_{1n_1}$
úroveň 2	$X_{21}, \ldots, X_{2n_2}$
…	…
úroveň r	$X_{r1}, \ldots, X_{rn_r}$

Ilustrace (1)



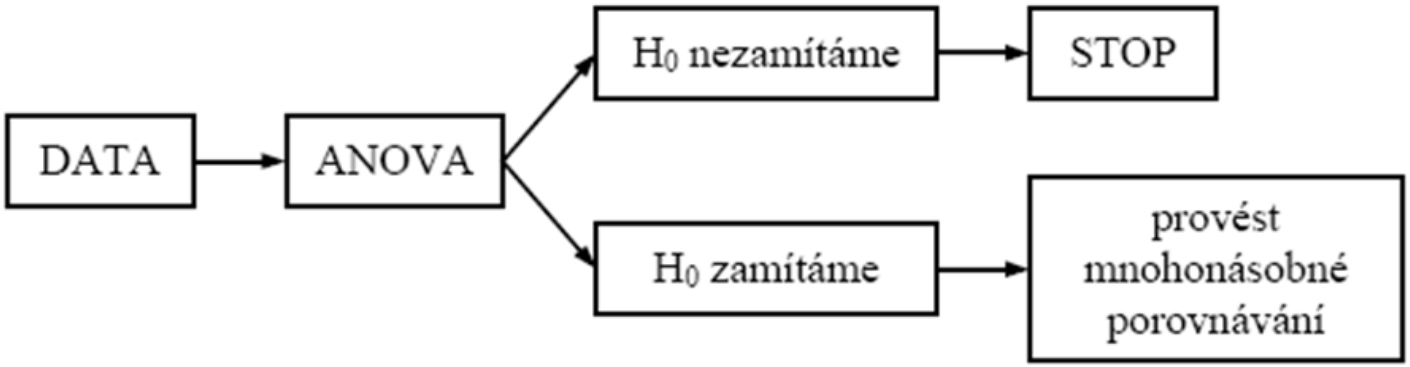
Na hladině významnosti α testujeme nulovou hypotézu, která tvrdí, že všechny střední hodnoty jsou stejné, tj. H_0: $\mu_1 = \ldots = \mu_r$ proti alternativní hypotéze H_1, která tvrdí, že aspoň jedna dvojice středních hodnot se liší.

Jedná se tedy o zobecnění dvouvýběrového t-testu a na první pohled se zdá, že stačí utvořit $\binom{r}{2}$ dvojic náhodných výběrů a na každou dvojici aplikovat dvouvýběrový t-test. Hypotézu o shodě všech středních hodnot bychom pak zamítli, pokud aspoň v jednom případě z $\binom{r}{2}$ porovnávání se prokáže odlišnost středních hodnot. Odtud je vidět, že k neoprávněnému zamítnutí nulové hypotézy (tj. k chybě 1. druhu) může dojít s pravděpodobností větší než α.

Ilustrace (2)

Proto ve 30. letech 20. století vytvořil R. A. Fisher metodu ANOVA (analýza rozptylu, v popsané situaci konkrétně analýza rozptylu jednoduchého třídění), která uvedenou podmínku splňuje.

Pokud na hladině významnosti α zamítneme nulovou hypotézu, zajímá nás, které dvojice středních hodnot se od sebe liší. K řešení tohoto problému slouží metody mnohonásobného porovnávání, např. Scheffého nebo Tukeyova metoda.



Tečková notace

V analýze rozptylu jednoduchého třídění se používá tzv. tečková notace.

$n = \sum_{i=1}^{r} n_i$ … celkový rozsah všech r výběrů

$X_{i.} = \sum_{j=1}^{n_i} X_{ij}$ … součet hodnot v i-tém výběru

$M_{i.} = \frac{1}{n_i} X_{i.}$ … výběrový průměr v i-tém výběru

$X_{..} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} X_{ij}$ … součet hodnot všech výběrů

$M_{..} = \frac{1}{n} X_{..}$ … celkový průměr všech r výběrů

Testování hypotézy o shodě středních hodnot (1)

Náhodné veličiny X_{ij} se řídí modelem

M0: $X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

pro $i = 1, \ldots, r$, $j = 1, \ldots, n_i$, přičemž

ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$,

μ je společná část střední hodnoty závisle proměnné veličiny,

α_i je efekt faktoru A na úrovni i.

Parametry μ, α_i neznáme.

Požadujeme, aby platila tzv. reparametrizační rovnice: $\sum_{i=1}^{r} n_i \alpha_i = 0$. (Pokud je třídění vyvážené, tj. pokud mají všechny výběry stejný rozsah: $n_1 = n_2 = \ldots = n_r$, pak lze použít zjednodušenou podmínku $\sum_{i=1}^{r} \alpha_i = 0$.)

Zavedeme součty čtverců $S_T = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - M_{..})^2$ … celkový součet čtverců (charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru), počet stupňů volnosti $f_T = n - 1$,

Testování hypotézy o shodě středních hodnot (2)

$S_A = \sum_{i=1}^{r} n_i (M_{i.} - M_{..})^2$ … skupinový součet čtverců (charakterizuje variabilitu mezi jednotlivými náhodnými výběry), počet stupňů volnosti $f_A = r - 1$.

Sčítanec $(M_{i.} - M_{..})$ představuje bodový odhad efektu α_i.

$$S_E = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - M_{i.})^2$$ … reziduální součet čtverců (charakterizuje variabilitu uvnitř jednotlivých výběrů), počet stupňů volnosti $f_E = n - r$.

Lze dokázat, že $S_T = S_A + S_E$.

(Důkaz je proveden např. ve skriptech Budíková, Mikoláš, Osecký: Popisná statistika v poznámce 5.20.)

Kdyby nezáleželo na faktoru A, platila by hypotéza $\alpha_1 = \ldots = \alpha_r = 0$ a dostali bychom model

M1: $X_{ij} = \mu + \varepsilon_{ij}$.

Během analýzy rozptylu tedy zkoumáme, zda výběrové průměry $M_1, \ldots, M_r$ se od sebe liší pouze v mezích náhodného kolísání kolem celkového průměru M nebo zda se projevuje vliv faktoru A.

Testování hypotézy o shodě středních hodnot (3)

Rozdíl mezi modely M0 a M1 ověřujeme pomocí testové statistiky

$F_A = \frac{S_A/f_A}{S_E/f_E}$, která se řídí rozložením F(r-1,n-r), je-li model M1 správný. Hypotézu o nevýznamnosti faktoru A tedy zamítneme na hladině významnosti α, když platí: $F_A \geq F_{1-\alpha}(r-1,n-r)$.

Výsledky výpočtů zapisujeme do tabulky analýzy rozptylu jednoduchého třídění.

Zdroj variability	součet čtverců	stupně volnosti	podíl	F_A
skupiny	S_A	$f_A = r - 1$	S_A/f_A	$\frac{S_A/f_A}{S_E/f_E}$
reziduální	S_E	$f_E = n - r$	S_E/f_E	-
celkový	S_T	$f_T = n - 1$	-	-

Sílu závislosti náhodné veličiny X na faktoru A můžeme měřit pomocí poměru determinace: $P^2 = \frac{S_A}{S_T}$. Nabývá hodnot z intervalu $\langle 0,1 \rangle$.

Testování hypotézy o shodě rozptylů (1)

Před provedením analýzy rozptylu je zapotřebí ověřit předpoklad o shodě rozptylů v daných r výběrech.

a) Levenův test: Položme $Z_{ij} = |X_{ij} - M_{i.}|$. Označíme

$$M_{Zi} = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij},$$

$$\mathrm{M_Z} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} Z_{ij},$$

$$\mathrm{S_{ZE}} = \sum_{i=1}^{r}\sum_{j=1}^{n_i} (Z_{ij} - M_{Zi})^2,$$

$$\mathrm{S_{ZA}} = \sum_{i=1}^{r} n_i (M_{Zi} - M_Z)^2 .$$

Platí-li hypotéza o shodě rozptylů, pak statistika $F_{ZA} = \frac{S_{ZA}/(r-1)}{S_{ZE}/(n-r)} \approx \mathrm{F(r\text{-}1,\ n\text{-}r)}$.

Hypotézu o shodě rozptylů tedy zamítáme na asymptotické hladině významnosti α, když $\mathrm{F_{ZA} \geq F_{1\text{-}\alpha}(r\text{-}1,\ n\text{-}r)}$.

Testování hypotézy o shodě rozptylů (2)

(Levenův test je vlastně založen na analýze rozptylu absolutních hodnot centrovaných pozorování. Vzhledem k tomu, že náhodné veličiny $X_{ij} - M_i$ nejsou stochasticky nezávislé a absolutní hodnoty těchto veličin nemají normální rozložení, je Levenův test pouze aproximativní.)

Modifikací Levenova testu je Brownův – Forsytheův test. Modifikace spočívá v tom, že místo výběrového průměru i-tého výběru se při výpočtu veličiny Z_{ij} používá medián i-tého výběru.

b) Bartlettův test: Platí-li hypotéza o shodě rozptylů a rozsahy všech výběrů jsou větší než 6, pak statistika

$B = \frac{1}{C}\left[(n-r)\ln S_*^2 - \sum_{i=1}^{r}(n_i - 1)\ln S_i^2\right] \approx \chi^2(r-1)$, kde

$C = 1 + \frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{n_i - 1} - \frac{1}{n-r}\right)$ a S_*^2 je vážený průměr výběrových rozptylů.

H_0 zamítáme na asymptotické hladině významnosti α, když $B \geq \chi^2_{1-\alpha}(r-1)$.

(Bartlettův test je poměrně slabý a je citlivý na porušení normality. Nedá se použít pro malé rozsahy výběrů.)

Post – hoc metody mnohonásobného porovnávání (1)

Zamítneme-li na hladině významnosti α hypotézu o shodě středních hodnot, chceme zjistit, které dvojice středních hodnot se liší na dané hladině významnosti α, tj. na hladině významnosti α testujeme H_0: $\mu_l = \mu_k$ proti H_1: $\mu_l \neq \mu_k$ pro všechna l, k = 1, .., r, $l \neq k$.

a) Mají-li všechny výběry týž rozsah p (říkáme, že třídění je vyvážené), použijeme Tukeyovu metodu. Testová statistika má tvar $\frac{|M_{k.} - M_{l.}|}{\frac{S_*}{\sqrt{p}}}$. Rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když $\frac{|M_{k.} - M_{l.}|}{\frac{S_*}{\sqrt{p}}} \geq q_{1-\alpha}(r, n-r)$, kde hodnoty $q_{1-\alpha}(r, n-r)$ jsou kvantily studentizovaného rozpětí a najdeme je ve statistických tabulkách. (Studentizované rozpětí je náhodná veličina $Q = \frac{X_{(n)} - X_{(1)}}{S}$.)

Existuje modifikace Tukeyovy metody pro nestejné rozsahy výběrů, nazývá se Tukeyova HSD metoda. V tomto případě má testová statistika tvar $\frac{|M_{k.} - M_{l.}|}{S_* \sqrt{\frac{1}{2}(\frac{1}{n_k} + \frac{1}{n_l})}}$.

Rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když $\frac{|M_{k.} - M_{l.}|}{S_* \sqrt{\frac{1}{2}(\frac{1}{n_k} + \frac{1}{n_l})}} \geq q_{1-\alpha}(r, n-r)$.

Post – hoc metody mnohonásobného porovnávání (2)

b) Nemají-li všechny výběry stejný rozsah, použijeme Scheffého metodu: rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když

$$|M_{k.} - M_{l.}| \geq S_* \sqrt{(r-1)\left(\frac{1}{n_k} + \frac{1}{n_l}\right) F_{1-\alpha}(r-1, n-r)}.$$

Výhodou Scheffého testu je, že k jeho provedení nepotřebujeme speciální statistické tabulky s hodnotami kvantilů studentizovaného rozpětí, ale stačí běžné statistické tabulky s kvantila Fisherova – Snedecorova rozložení.

V případě vyváženého třídění, kdy lze aplikovat Tukeyovu i Scheffého metodu, použijeme tu, která je citlivější. Tukeyova metoda tedy bude výhodnější, když

$q_{1-\alpha}^2(r, n-r) < 2(r-1)F_{1-\alpha}(r-1, n-r)$.

Metody mnohonásobného porovnávání mají obecně menší sílu než ANOVA.

Může nastat situace, kdy při zamítnutí H_0 nenajdeme metodami mnohonásobného porovnávání významný rozdíl u žádné dvojice středních hodnot. K tomu dochází zvláště tehdy, když p-hodnota pro ANOVU je jen o málo nižší než zvolená hladina významnosti. Pak slabší test patřící do skupiny metod mnohonásobného porovnávání nemusí odhalit žádný rozdíl.

Tukeyova metoda je citlivější při stejném rozsahu.

Plánované porovnávání - testování významnosti kontrastů

Plánované porovnávání je navrženo před prováděním ANOVY. Provádí se pomocí kontrastů, tj. pomocí lineárních kombinací středních hodnot. Kontrast $q = \sum_{i=1}^{r} c_i \mu_i$, kde $\sum_{i=1}^{r} c_i = 0$ a $\sum_{i=1}^{r} {c_i}^2 > 0$. Odhadem kontrastu je veličina $\hat{q} = \sum_{i=1}^{r} c_i M_{i.}$. Testování H_0: $q = 0$ proti H_1: $q \neq 0$ je založeno na statistice $F_q = \frac{\hat{q}^2}{{S_*}^2 \sum_{i=1}^{r} \frac{{c_i}^2}{n_i}}$, která se v případě platnosti H_0 řídí rozložením F(1, n-r). Nulovou hypotézu zamítáme na hladině významnosti α, když platí

$F_q \geq F_{1-\alpha}(1, n - r)$.

Porovnávání s kontrolou

V tomto případě neporovnáváme jednotlivé skupiny mezi sebou, ale každou skupinu porovnáme s kontrolní skupinou. Na hladině významnosti α testujeme H_0: $\mu_i = \mu_{kontrola}$ proti H_1: $\mu_i \neq \mu_{kontrola}$. Provedeme tedy celkem r-1 porovnání. Používá se Dunettův test, jehož testové kritérium je $\frac{|M_{i.} - M_{kontrola}|}{S_*}$.

Nulovou hypotézu zamítáme na hladině významnosti α, když $\frac{|M_{i.} - M_{kontrola}|}{S_*} \geq q_{1-\alpha}(r, n-r)$.

Příklad (1)

U čtyř odrůd brambor (označených symboly A, B, C, D) se zjišťovala celková hmotnost brambor vyrostlých vždy z jednoho trsu. Výsledky (v kg):

odrůda	hmotnost
A	0,9 0,8 0,6 0,9
B	1,3 1,0 1,3
C	1,3 1,5 1,6 1,1 1,5
D	1,1 1,2 1,0

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota hmotnosti trsu brambor nezávisí na odrůdě. Zamítnete-li nulovou hypotézu, zjistěte, které dvojice odrůd se liší na hladině významnosti 0,05.

Řešení:

Data považujeme za realizace čtyř nezávislých náhodných výběrů ze čtyř normálních rozložení se stejným rozptylem. Testujeme hypotézu, že všechny čtyři střední hodnoty jsou stejné.

Vypočítáme výběrové průměry v jednotlivých výběrech:

$M_{1.} = 0{,}8$, $M_{2.} = 1{,}2$, $M_{3.} = 1{,}4$, $M_{4.} = 1{,}1$,

celkový průměr $M_{..} = 1{,}14$,

výběrové rozptyly:

$S_1^2 = 0{,}02$, $S_2^2 = 0{,}03$, $S_3^2 = 0{,}04$, $S_4^2 = 0{,}01$,

Příklad (2)

vážený průměr výběrových rozptylů:

$S_*{}^2 = \frac{\sum_{i=1}^{r}(n_i-1)S_i{}^2}{n-r} = \frac{3\cdot 0{,}02+2\cdot 0{,}03+4\cdot 0{,}04+2\cdot 0{,}01}{11} = \frac{3}{110} = 0{,}0\overline{27}$, reziduální součet čtverců: $S_E = (n-r)S_*{}^2 = 11\cdot\frac{3}{110} = 0{,}3$,

skupinový součet čtverců: $S_A = \sum_{i=1}^{r} n_i(M_{i.} - M_{..})^2 = 4\cdot(0{,}8-1{,}14)^2 + 3\cdot(1.2-1{,}14)^2 + 5\cdot(1{,}4-1{,}14)^2 + 3\cdot(1{,}1-1{,}14)^2 = 0{,}816$

celkový součet čtverců: $S_T = S_A + S_E = 0{,}816 + 0{,}3 = 1{,}116$,

testová statistika $F_A = \frac{S_A/f_A}{S_E/f_E} = \frac{0{,}816/3}{0{,}3/11} = 9{,}97$,

Kritický obor $W = \langle F_{0,95}(3,11), \infty) = \langle 3{,}59, \infty)$. Protože testová statistika se realizuje v kritickém oboru, H_0 zamítáme na hladině významnosti 0,05.

Vypočteme poměr determinace: $P^2 = \frac{S_A}{S_T} = \frac{0{,}816}{1{,}116} = 0{,}7312$

Příklad (3)

Výsledky zapíšeme do tabulky ANOVA:

Zdroj variability	Součet čtverců	Stupně volnosti	podíl	F_A
skupiny	$S_A = 0{,}816$	3	$S_A/3 = 0{,}272$	$\frac{S_A/(r-1)}{S_E/(n-r)} = 9{,}97$
reziduální	$S_E = 0{,}3$	11	$S_E/11 = 0{,}02727$	-
celkový	$S_T = 1{,}116$	14	-	-

Nyní pomocí Scheffého metody zjistíme, které dvojice odrůd se liší na hladině významnosti 0,05.

Srovnávané odrůdy	Rozdíly $\lvert M_{k.} - M_{l.} \rvert$	Pravá strana vzorce
A, B	0,4	0,41
A, C	0,6	0,36
A, D	0,3	0,41
B, C	0,2	0,40
B, D	0,1	0,44
C, D	0,3	0,40

Na hladině významnosti 0,05 se liší odrůdy A a C.

Výpočet pomocí systému STATISTICA (1)

Otevřeme nový datový soubor o dvou proměnných X a odrůda a 15 případech. Do proměnné X zapíšeme zjištěné hmotnosti, do proměnné odrůda kódy pro dané odrůdy (1 pro A, 2 pro B, 3 pro C a 4 pro D).

	1 X	2 odruda
1	0,9	A
2	0,8	A
3	0,6	A
4	0,9	A
5	1,3	B
6	1	B
7	1,3	B
8	1,3	C
9	1,5	C
10	1,6	C
11	1,1	C
12	1,5	C
13	1,1	D
14	1,2	D
15	1	D

Výpočet pomocí systému STATISTICA (2)

Vypočteme výběrové průměry a výběrové rozptyly:

Statistiky – Základní statistiky a tabulky – Rozklad & jednofakt. ANOVA – OK – Proměnné – Závislé – X, Grupovací - odrůda – OK – Skupiny tabulek - zaškrtneme Rozptyly - Výpočet.

Rozkladová tabulka popisných statistik (priklad83
N=15 (V seznamu záv. prom. nejsou ChD)

odruda	X průměr	X N	X Sm.odch.	X Rozptyl
A	0,80000	4	0,14142	0,02000
B	1,20000	3	0,17320	0,03000
C	1,40000	5	0,20000	0,04000
D	1,10000	3	0,10000	0,01000
Vš.skup.	1,14000	15	0,28233	0,07971

Nyní ověříme předpoklad shody rozptylů.

Na záložce Skupiny tabulek zaškrtneme Levenův test – Výpočet.

Výpočet pomocí systému STATISTICA (3)

Proměnná	Leveneův test homogenity rozpylů (priklad8301) Označ. efekty jsou význ. na hlad. p < ,05000							
	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
X	0,01866	3	0,00622	0,06533	11	0,00593	1,04761	0,41002

Vidíme, že p-hodnota Levenova testu je 0,41, tedy větší než hladina významnosti 0,05. Hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05.

Přistoupíme k testu hypotézy o shodě středních hodnot.

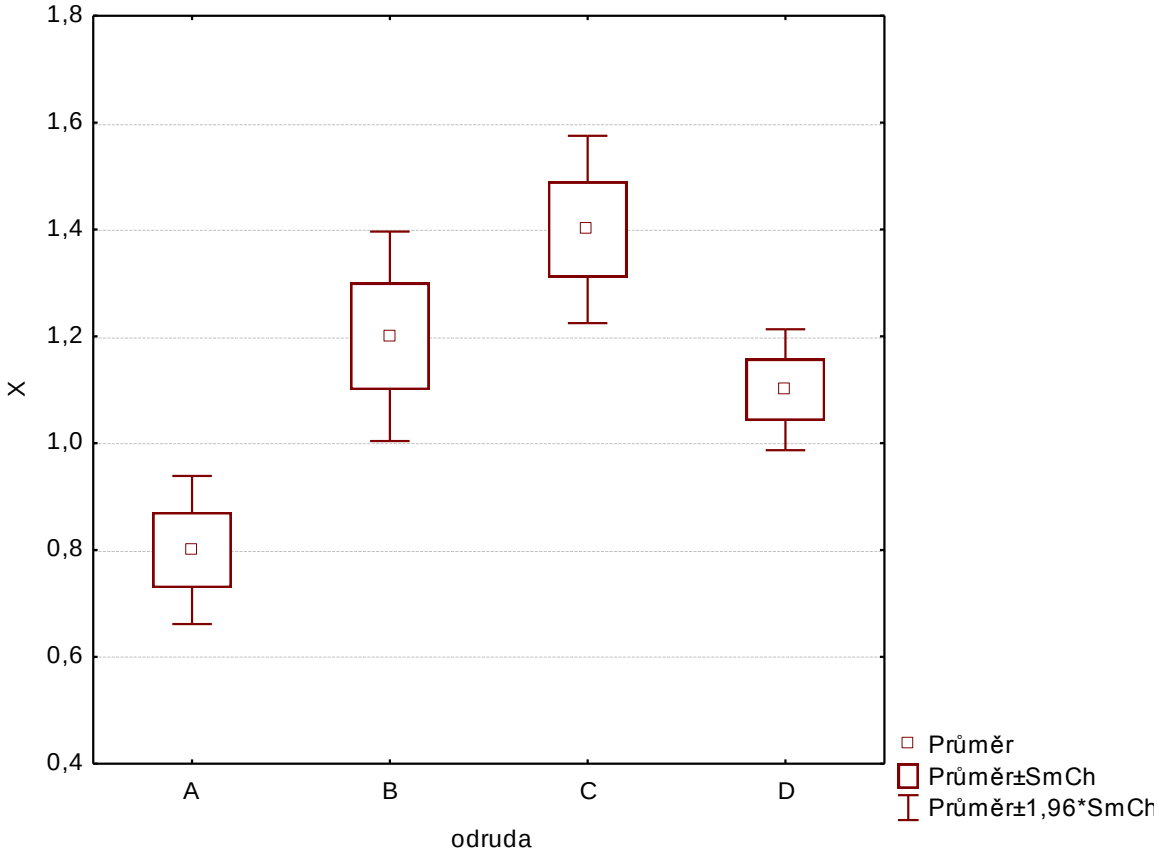
Na záložce Skupiny tabulek zaškrtneme Analýza rozptylu – Výpočet.

Proměnná	Analýza rozptylu (priklad8301) Označ. efekty jsou význ. na hlad. p < ,05000							
	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
X	0,81600	3	0,27200	0,30000	11	0,02727	9,97333	0,00180

Jelikož p-hodnota = 0,001805 je menší než hladina významnosti 0,05, hypotézu o shodě středních hodnot zamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA (4)

Výpočet doplníme krabicovými diagramy:



Výpočet pomocí systému STATISTICA (5)

Nyní aplikujeme Scheffého metodu mnohonásobného porovnávání, abychom zjistili, které dvojice odrůd se liší na hladině významnosti 0,05. Na záložce Post – hoc zvolíme Schefféův test.

		Scheffeho test; proměn.:X (priklad8301) Označ. rozdíly jsou významné na hlad. p < ,05			
odruda		{1} M=,8000(	{2} M=1,200(	{3} M=1,400(	{4} M=1,100(
A	{1}		0,05916	0,00195	0,19046
B	{2}	0,05916		0,46453	0,90550
C	{3}	0,00195	0,46453		0,16349
D	{4}	0,19046	0,90550	0,16349	

Tabulka obsahuje p-hodnoty pro vzájemné porovnání středních hodnot hmotnosti všech čtyř odrůd. Vidíme, že na hladině významnosti 0,05 se liší odrůdy A, C.

Význam předpokladů v analýze rozptylu

a) Nezávislost jednotlivých náhodných výběrů – velmi důležitý předpoklad, musí být splněn, jinak dostaneme nesmyslné výsledky.

b) Normalita – ANOVA není příliš citlivá na porušení normality, zvlášť pokud mají všechny výběry rozsah nad 20 (důsledek centrální limitní věty). Při výraznějším porušení normality se doporučuje Kruskalův – Wallisův test.

c) Shoda rozptylů – mírné porušení nevadí, při větším se doporučuje Kruskalův – Wallisův test. Test shody rozptylů má smysl provádět až po ověření předpokladu normality.

10. Neparametrické testy o mediánech

Motivace: Při aplikaci t-testů či analýzy rozptylu by měly být splněny určité předpoklady:

- normalita dat (pro výběry větších rozsahů ($n \geq 30$) nemá mírné porušení normality závažný dopad na výsledky)
- homogenita rozptylů
- intervalový či poměrový charakter dat

Pokud nejsou tyto předpoklady splněny, použijeme tzv. neparametrické testy, které nevyžadují předpoklad o konkrétním typu rozložení (např. normálním), stačí např. předpokládat, že distribuční funkce rozložení, z něhož náhodný výběr pochází, je spojitá.

Nevýhoda – ve srovnání s klasickými parametrickými testy jsou neparametrické testy slabší, tzn., že nepravdivou hypotézu zamítají s menší pravděpodobností než testy parametrické.

V této kapitole se omezíme na ty neparametrické testy, které jsou založeny na pořadí a týkají se mediánů. Nazývají se pořadové testy.

Pojem pořadí a průměrného pořadí

Nechť $X_1, ..., X_n$ je náhodný výběr.

Vektor $(X_{(1)}, ..., X_{(n)})$, kde $X_{(1)} \leq ... \leq X_{(n)}$ se nazývá uspořádaný náhodný výběr a statistika $X_{(i)}$ se nazývá i-tá pořádková statistika, $i = 1, ..., n$.

Pořadím R_i statistiky X_i rozumíme počet těch náhodných veličin $X_1, ..., X_n$, které nabývají hodnoty menší nebo rovné X_i, tj.

$$R_i = card\{j; X_j \leq X_i\}.$$

V praxi se může stát, že některá pozorování jsou si rovna a vytvářejí skupiny shodných čísel. Pak těmto shodným číslům přiřadíme průměrné pořadí odpovídající takové skupině.

Příklad

Máme čísla 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2. Stanovte jejich pořadí.

Řešení:

usp.hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

Jednovýběrový znaménkový test a jeho asymptotická varianta (1)

Nechť $X_1, ..., X_n$ je náhodný výběr ze spojitého rozložení se spojitou distribuční funkcí $\Phi(x)$. Nechť $x_{0,50}$ je mediánem tohoto rozložení, tj. $\Phi(x_{0,50}) = 0,5$. Nechť c je reálná konstanta. Testujeme hypotézu H_0: $x_{0,50} = c$ proti oboustranné alternativě H_1: $x_{0,50} \neq c$ (resp. proti levostranné alternativě H_1: $x_{0,50} < c$ resp. proti pravostranné alternativě H_1: $x_{0,50} > c$).

Postup provedení testu:

a) Utvoříme rozdíly $Y_i = X_i - c$, $i = 1, ..., n$. (Jsou-li některé rozdíly nulové, pak za n bereme jen počet nenulových hodnot.)

b) Zavedeme statistiku S_Z^+, která udává počet těch rozdílů, které jsou kladné. Platí-li H_0, pak $S_Z^+ \sim Bi(n,1/2)$, tedy $E(S_Z^+) = n/2$, $D(S_Z^+) = n/4$.

c) Stanovíme kritický obor.

Jednovýběrový znaménkový test a jeho asymptotická varianta (2)

Pro oboustrannou alternativu ho budou tvořit ty hodnoty testové statistiky S_Z^+, které jsou blízké 0 nebo n, tedy $W = \langle 0, k_1 \rangle \cup \langle k_2, n \rangle$, kde nezáporná celá čísla k_1, k_2, splňují podmínky $P(S_Z{}^+ \leq k_1) \leq \frac{\alpha}{2}$, $P(S_Z{}^+ \geq k_2) \leq \frac{\alpha}{2}$

Pro levostrannou alternativu: $W = \langle 0, k_1 \rangle$, kde nezáporné celé číslo k_1 splňuje podmínku $P(S_Z{}^+ \leq k_1) \leq \alpha$

Pro pravostrannou alternativu: $W = \langle k_2, n \rangle$, kde nezáporné celé číslo k_2 splňuje podmínku $P(S_Z{}^+ \geq k_2) \leq \alpha$

(Čísla k_1, k_2 pro oboustranný test i pro jednostranné testy lze najít ve statistických tabulkách.)

d) H_0 zamítáme na hladině významnosti α, když $S_Z{}^+ \in W$.

Jednovýběrový znaménkový test a jeho asymptotická varianta (3)

Asymptotická varianta testu:

Pro velká n (prakticky n > 20) lze využít asymptotické normality statistiky S_Z^+.

Testová statistika $U_0 = \frac{S_Z^+ - E(S_Z^+)}{\sqrt{D(S_Z^+)}} = \frac{S_Z^+ - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$ má za platnosti H_0 asymptoticky rozložení N(0,1).

Kritický obor

– pro oboustrannou alternativu: $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$,

– pro levostrannou alternativu: $W = (-\infty, -u_{1-\alpha}\rangle$,

– pro pravostrannou alternativu: $W = \langle u_{1-\alpha}, \infty)$.

H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Aproximace rozložením N(0,1) se zlepší, když použijeme tzv. korekci na nespojitost. Testová statistika pak má tvar $U_0 = \frac{S_Z^+ - \frac{n}{2} \pm \frac{1}{2}}{\sqrt{\frac{n}{4}}}$, přičemž 1/2 přičteme, když $S_Z^+ < n/2$ a odečteme v opačném případě.

Příklad

U 10 náhodně vybraných vzorků benzínu byly zjištěny následující hodnoty oktanového čísla: 98,2 96,8 96,3 99,8 96,9 98,6 95,6 97,1 97,7 98,0. Na hladině významnosti 0,05 testujte hypotézu, že medián oktanového čísla je 98 proti oboustranné alternativě.

Řešení:

Testujeme H_0: $x_{0,50} = 98$ proti oboustranné alternativě H_1: $x_{0,50} \neq 98$, kde $x_{0,50}$ je medián rozložení, z něhož pochází náhodný výběr $X_1, \ldots, X_{10}$.

rozdíly $x_i - 98$: 0,2 -1,2 -1,7 1,8 -1,1 0,6 -2,4 -0,9 -0,3 0,0

$S_Z^+ = 3$, nenulových rozdílů je 9. Ve statistických tabulkách najdeme pro $n = 9$ a $\alpha = 0,05$ kritické hodnoty $k_1 = 1$, $k_2 = 8$. Protože kritický obor $W = \langle 0,1 \rangle \cup \langle 8,9 \rangle$ neobsahuje hodnotu 3, nemůžeme H_0 zamítnout na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA

Vytvoříme nový datový soubor se dvěma proměnnými a 10 případy. Do proměnné X napíšeme hodnoty oktanového čísla a do proměnné konst uložíme číslo 98.

Statistiky –Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných X, Druhý seznam proměnných konst – OK – Znaménkový test.

	Znaménkový test (oktanove cislo) Označené testy jsou významné na hladině p <,05			
Dvojice proměnnýc	Počet různých	procent v < V	Z	Úroveň p
X & konst	9	66,6666	0,66666	0,50498

Vidíme, že nenulových hodnot n = 9. Z nich záporných je 66,7%, tj. 6. Hodnota testové statistiky $S_Z^+ = 9 - 6 = 3$. Asymptotická testová statistika U_0 (zde označená jako Z) se realizuje hodnotou 0,6667. Odpovídající asymptotická p-hodnota je 0,505, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu, že medián oktanového čísla je 98.

Upozornění: V tomto případě není splněna podmínka pro využití asymptotické normality statistiky S_Z^+, tj. $n > 20$. Je tedy vhodnější najít v tabulkách kritické hodnoty pro znaménkový test. Pro $n = 9$ a $\alpha = 0{,}05$ jsou kritické hodnoty $k_1 = 1$, $k_2 = 8$. Protože kritický obor $W = \langle 0,1\rangle \cup \langle 8,9\rangle$ neobsahuje hodnotu 3, nezamítáme H_0 na hladině významnosti 0,05. Dostáváme týž výsledek jako při použití asymptotického testu.

Párový znaménkový test

Nechť $(X_1, Y_1), ..., (X_n, Y_n)$ je náhodný výběr ze spojitého dvourozměrného rozložení. Testujeme H_0: $x_{0,50} - y_{0,50} = c$ proti H_1: $x_{0,50} - y_{0,50} \neq c$ (resp. proti jednostranným alternativám). Utvoříme rozdíly $Z_i = X_i - Y_i$, $i = 1, ..., n$ a testujeme hypotézu o mediánu $z_{0,50}$, tj. H_0: $z_{0,50} = c$ proti H_1: $z_{0,50} \neq c$.

Příklad

U osmi osob byl změřen systolický krevní tlak před pokusem a po něm.

č. osoby	1	2	3	4	5	6	7	8
tlak před	130	185	162	136	147	181	138	139
tlak po	139	190	175	135	155	175	158	149

Na hladině významnosti 0,05 testujte hypotézu, že pokus neovlivní systolický krevní tlak

Řešení:

Testujeme H_0: $z_{0,50} = 0$ proti oboustranné alternativě H_1: $z_{0,50} \neq 0$, kde $z_{0,50}$ je medián rozložení, z něhož pochází rozdílový náhodný výběr $Z_1 = X_1 - Y_1, \ldots Z_8 = X_8 - Y_8$. Vypočteme rozdíly mezi tlakem před pokusem a po pokusu, čímž úlohu převedeme na jednovýběrový test.

rozdíly $x_i - y_i$: -9 -5 -13 1 -8 6 -30 -10

Testová statistika $S_Z^+ = 2$.Ve statistických tabulkách najdeme pro $n = 8$ a $\alpha = 0,05$ kritické hodnoty $k_1 = 0$, $k_2 = 8$. Protože kritický obor $W = 0 \cup 8$ neobsahuje hodnotu 2, nemůžeme H0 zamítnout na hladině významnosti 0,05. Znamená to, že s rizikem omylu nejvýše 0,05 je zvýšení krevního tlaku stejně pravděpodobné jako jeho pokles.

Výpočet pomocí systému STATISTICA

Vytvoříme nový datový soubor se dvěma proměnnými a 8 případy. Do proměnné X napíšeme hodnoty tlaku před pokusem, do proměnné Y hodnoty tlaku po pokusu.

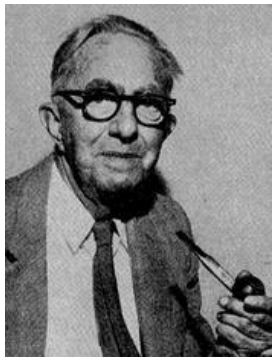
Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných X, 2. seznam proměnných Y – OK – Znaménkový test.

	Znaménkový test (tlak.sta) Označené testy jsou významné na hladině p <,05			
Dvojice proměnnýc	Počet různých	procent v < V	Z	Úroveň p
X & Y	8	75,0000	1,06066	0,28884

Vidíme, že nenulových hodnot n = 8. Z nich záporných je 75%, tj. 6. Hodnota testové statistiky $S_Z^+ = 8 - 6 = 2$. Asymptotická testová statistika U_0 (zde označená jako Z) se realizuje hodnotou 1,06066. Odpovídající asymptotická p-hodnota je 0,2888, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu, že zvýšení krevního tlaku stejně pravděpodobné jako jeho pokles.

Upozornění: Stejně jako u příkladu na str. 237 (jednovýběrový test) není splněna podmínka pro použití asymptotického testu. Správný postup je tedy ten, který je uveden na předchozí str. 240.

Jednovýběrový Wilcoxonův test a jeho asymptotická varianta (1)



Frank Wilcoxon (1892 – 1965): Americký statistik a chemik

Nechť $X_1, ..., X_n$ je náhodný výběr ze spojitého rozložení s hustotou $\varphi(x)$, která je symetrická kolem mediánu $x_{0,50}$, tj. $\varphi(x_{0,50} + x) = \varphi(x_{0,50} - x)$. Nechť c je reálná konstanta.

Testujeme hypotézu H_0: $x_{0,50} = c$

proti oboustranné alternativě H_1: $x_{0,50} \neq c$ nebo

proti levostranné alternativě H_1: $x_{0,50} < c$ nebo

proti pravostranné alternativě H_1: $x_{0,50} > c$.

Jednovýběrový Wilcoxonův test a jeho asymptotická varianta (2)

Postup provedení testu:

a) Utvoříme rozdíly $Y_i = X_i - c$, $i = 1, \ldots, n$. (Jsou-li některé rozdíly nulové, pak za n bereme jen počet nenulových hodnot.)

b) Absolutní hodnoty $|Y_i|$ uspořádáme vzestupně podle velikosti a spočteme pořadí R_i.

c) Zavedeme statistiky

$S_W^+ = \sum_{Y_i>0} R_i^+$, což je součet pořadí přes kladné hodnoty Y_i,

$S_W^- = \sum_{Y_i<0} R_i^-$, což je součet pořadí přes záporné hodnoty Y_i.

Přitom platí, že součet $S_W^+ + S_W^- = n(n+1)/2$.

Je-li H_0 pravdivá, pak $E(S_W^+) = n(n+1)/4$ a $D(S_W^+) = n(n+1)(2n+1)/24$.

d) Testová statistika $= \min(S_W^+, S_W^-)$ pro oboustrannou alternativu,

$= S_W^+$ pro levostrannou alternativu,

$= S_W^-$ pro pravostrannou alternativu.

e) H_0 zamítáme na hladině významnosti α, když testová statistika je menší nebo rovna tabelované kritické hodnotě.

Jednovýběrový Wilcoxonův test a jeho asymptotická varianta (3)

Asymptotická varianta jednovýběrového Wilcoxonova testu:

Pro n $\geq$ 30 lze využít asymptotické normality statistiky S_W^+.

Platí-li H_0, pak $U_0 = \frac{S_W^+ - E(S_W^+)}{\sqrt{D(S_W^+)}} = \frac{S_W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \approx N(0,1)$.

Kritický obor:

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

pro levostrannou alternativu $W = (-\infty, -u_{1-\alpha}\rangle$,

pro pravostrannou alternativu $W = \langle u_{1-\alpha}, \infty)$

H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Předpoklady použití jednovýběrového Wilcoxonova testu:

- rozložení, z něhož daný náhodný výběr pochází, je spojité
- hustota tohoto rozložení je symetrická kolem mediánu
- sledovaná veličina X má aspoň ordinální charakter

(Není-li splněn předpoklad o symetrii hustoty kolem mediánu, lze použít např. znaménkový test.)

Příklad

Pro zadání příkladu o oktanovém číslu benzínu, proveďte jednovýběrový Wilcoxonův test.

Řešení:

Testujeme hypotézu H_0: $x_{0,50} = 98$ proti oboustranné alternativě H_1: $x_{0,50} \neq 98$.

Absolutní hodnoty rozdílů $x_i - 98$ setřídíme vzestupně podle velikosti (přitom vynecháme nulový rozdíl a kladné rozdíly značíme tučně):

abs $(x_i - 98)$	**0,2**	0,3	**0,6**	0,9	1,1	1,2	1,7	**1,8**	2,4
pořadí R_i	**1**	2	**3**	4	5	6	7	**8**	9

Součet pořadí přes kladné hodnoty rozdílů: $S_W^+ = 12$

Součet pořadí přes záporné hodnoty rozdílů: $S_W^- = 33$

Testová statistika = min(12,33) = 12, tabelovaná kritická hodnota pro $\alpha = 0,05$ a n = 9 je 5. Protože $12 > 5$, H_0 nezamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA

Utvoříme nový datový soubor se dvěma proměnnými a 10 případy. Do proměnné oktan napíšeme zjištěné hodnoty a do proměnné konst uložíme číslo 98.

Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných oktan, 2. seznam proměnných konst – OK – Wilcoxonův párový test.

	Wilcoxonův párový test (oktan.sta) Označené testy jsou významné na hladině p <,05			
Dvojice proměnnýc	Počet platných	T	Z	Úroveň p
oktan & konst	10	12,0000	1,24393	0,21352

Výstupní tabulka poskytne hodnotu testové statistiky SW^+ (zde označena T), hodnotu asymptotické testové statistiky U_0 a p-hodnotu pro U_0. V tomto případě je p-hodnota 0,213525, tedy nulová hypotéza se nezamítá na asymptotické hladině významnosti 0,05.

Upozornění: I v tomto případě není splněna podmínka pro použití asymptotického testu. Správný postup je tedy ten, který je uveden na předchozí str. 245.

Párový Wilcoxonův test

Nechť $(X_1, Y_1), ..., (X_n\ Y_n)$ je náhodný výběr ze spojitého dvourozměrného rozložení. Testujeme H_0: $x_{0,50} - y_{0,50} = c$ proti H_1: $x_{0,50} - y_{0,50} \neq c$ (resp. proti jednostranným alternativám). Utvoříme rozdíly $Z_i = X_i - Y_i$, $i = 1, ..., n$ a testujeme hypotézu o mediánu $z_{0,50}$, tj. H_0: $z_{0,50} = c$ proti H_1: $z_{0,50} \neq c$.

Příklad

Pro data z příkladu o krevním tlaku proveďte párový Wilcoxonův test.

Řešení:

Testujeme H_0: $z_{0,50} = 0$ proti oboustranné alternativě H_1: $z_{0,50} \neq 0$, kde $z_{0,50}$ je medián rozložení, z něhož pochází rozdílový náhodný výběr $Z_1 = X_1 - Y_1$, … $Z_8 = X_8 - Y_8$.

Absolutní hodnoty rozdílů $x_i - y_i$ setřídíme vzestupně podle velikosti (kladné rozdíly značíme tučně):

abs $(x_i - y_i)$	**1**	5	**6**	8	9	10	13	20
pořadí R_i	**1**	2	**3**	4	5	6	7	8

Součet pořadí přes kladné hodnoty rozdílů: $S_W^+ = 4$

Součet pořadí přes záporné hodnoty rozdílů: $S_W^- = 32$

Testová statistika = min(4,32) = 4, tabelovaná kritická hodnota pro $\alpha = 0,05$ a n = 8 je 3. Protože $4 > 3$, H_0 nezamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA

Použijeme datový soubor, který jsme již vytvořili pro aplikaci znaménkového testu.

Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných X, 2. seznam proměnných Y – OK – Wilcoxonův párový test.

	Wilcoxonův párový test (tlak.sta) Označené testy jsou významné na hladině p <,05			
Dvojice proměnnýc	Počet platných	T	Z	Úroveň p
X & Y	8	4,00000(	1,96039:	0,04995:

Testová statistika (zde označená jako T) nabývá hodnoty 4, asymptotická testová statistika (označená jako Z) nabývá hodnoty 1,960392, odpovídající asymptotická p-hodnota je 0,049951, tedy na asymptotické hladině významnosti 0,05 nulovou hypotézu zamítáme. To je v rozporu s výsledkem, k němuž jsme dospěli při ručním výpočtu. Je to způsobeno tím, že není dodržena podmínka pro použití asymptotické varianty Wilcoxonova testu – rozsah výběru má být aspoň 30.

Příklad

(Asymptotická varianta Wilcoxonova testu)

30 náhodně vybraných osob mělo nezávisle na sobě bez předchozího nácviku odhadnout, kdy od daného signálu uplyne 1 minuta. Byly získány následující výsledky (v sekundách): 53 48 45 55 63 51 66 56 50 58 61 51 64 63 59 47 46 58 52 56 61 57 48 62 54 49 51 46 53 58.

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že medián rozložení, z něhož daný náhodný výběr pochází, je 60 sekund proti oboustranné alternativě (nulová hypotéza vlastně tvrdí, že polovina osob délku jedné minuty podhodnotí a druhá nadhodnotí).

Řešení:

Testujeme H_0: $x_{0,50} = 60$ proti oboustranné alternativě H_1: $x_{0,50} \neq 60$.

Obvyklým způsobem stanovíme statistiku $S_W^+ = 55$.

Asymptotická testová statistika:

$$U_0 = \frac{S_W^+ - E(S_W^+)}{\sqrt{D(S_W^+)}} = \frac{S_W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{55 - \frac{30(30+1)}{4}}{\sqrt{\frac{30(30+1)(2.30+1)}{24}}} = -3{,}65$$

Kritický obor:

$$W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty) = (-\infty, -u_{0,975}\rangle \cup \langle u_{0,975}, \infty) = (-\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty).$$

Testová statistika se realizuje v kritickém oboru, tedy H_0 zamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA

Utvoříme nový datový soubor se dvěma proměnnými a 30 případy. Do proměnné odhad napíšeme zjištěné hodnoty a do proměnné konst uložíme číslo 60.

Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných odhad, Druhý seznam proměnných konst – OK – Wilcoxonův párový test.

	Wilcoxonův párový test (odhad minuty) Označené testy jsou významné na hladině p <,05			
Dvojice proměnnýc	Počet platných	T	Z	Úroveň p
odhad& konst	30	55,0000	3,65088	0,00026

Testová statistika (zde označená jako T) nabývá hodnoty 55, asymptotická testová statistika (označená jako Z) nabývá hodnoty 3,65088, odpovídající asymptotická p-hodnota je 0,000261, tedy na asymptotické hladině významnosti 0,05 nulovou hypotézu zamítáme.

Dvouvýběrový Wilcoxonův test a jeho asymptotická varianta (1)

Nechť $X_1, ..., X_n$ a $Y_1, ..., Y_m$ jsou dva nezávislé náhodné výběry ze dvou spojitých rozložení, jejichž distribuční funkce se mohou lišit pouze posunutím. Označme $x_{0,50}$ medián prvního rozložení a $y_{0,50}$ medián druhého rozložení. Testujeme hypotézu, že distribuční funkce těchto rozložení jsou shodné neboli mediány jsou shodné proti alternativě, že jsou rozdílné, tj.

$H_0: x_{0,50} - y_{0,50} = 0$ proti $H_1: x_{0,50} - y_{0,50} \neq 0$.

Postup provedení testu:

a) Všech n + m hodnot $X_1, ..., X_n$ a $Y_1, ..., Y_m$ uspořádáme vzestupně podle velikosti.

b) Zjistíme součet pořadí hodnot $X_1, ..., X_n$ a označíme ho T_1. Součet pořadí hodnot $Y_1, ..., Y_m$ označíme T_2.

c) Vypočteme statistiky $U_1 = mn + n(n+1)/2 - T_1$, $U_2 = mn + m(m+1)/2 - T_2$. Přitom platí $U_1 + U_2 = mn$.

d) Pokud $\min(U_1,U_2) \leq$ tabelovaná kritická hodnota (pro dané rozsahy výběrů m, n a dané α), pak nulovou hypotézu o totožnosti obou distribučních funkcí zamítáme na hladině významnosti α. V tabulkách: $n = \min\{m,n\}$ a $m = \max\{m,n\}$.

Dvouvýběrový Wilcoxonův test a jeho asymptotická varianta (2)

Asymptotická varianta dvouvýběrového Wilcoxonova testu:

Pro velká n, m (n, m > 30) lze využít asymptotické normality statistiky U_1.

Platí-li H_0, pak $U_0 = \frac{U_1 - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \approx N(0,1)$, kde $U_1 = \min(U_1, U_2)$.

Kritický obor:

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

pro levostrannou alternativu $W = (-\infty, -u_{1-\alpha}\rangle$,

pro pravostrannou alternativu $W = \langle u_{1-\alpha}, \infty)$

H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Předpoklady použití dvouvýběrového Wilcoxonova testu:

- dané dva náhodné výběry jsou nezávislé
- rozložení, z nichž dané dva náhodné výběry pocházejí, jsou spojitá
- distribuční funkce těchto rozložení se mohou lišit pouze posunutím
- sledovaná veličina má aspoň ordinální charakter

(Není-li splněn předpoklad, že distribuční funkce se mohou lišit pouze posunutím, lze použít např. dvouvýběrový Kolmogorovův – Smirnovův test.)

Příklad

Výrobce určitého výrobku se má rozhodnout mezi dvěma dodavateli polotovarů vyrábějících je různými technologiemi. Rozhodující je procentní obsah určité látky.

1. technologie: 1,52 1,57 1,71 1,34 1,68

2. technologie: 1,75 1,67 1,56 1,66 1,72 1,79 1,64 1,55

Na hladině významnosti 0,05 posuďte pomocí dvouvýběrového Wilcoxonova testu, zda je oprávněný předpoklad, že obě technologie poskytují stejné procento účinné látky.

Řešení:

Na hladině významnosti 0,05 testujeme H_0: $x_{0,50} - y_{0,50} = 0$ proti oboustranné alternativě H_1: $x_{0,50} - y_{0,50} \neq 0$.

usp.h. **1,34** **1,52** 1,55 1,56 **1,57** 1,64 1,66 1,67 **1,68** **1,71** 1,72 1,75 1,79

pořadí **1** **2** 3 4 **5** 6 7 8 **9** **10** 11 12 13

$T_1 = 1 + 2 + 5 + 9 + 10 = 27$, $T_2 = 3 + 4 + 6 + 7 + 8 + 11 + 12 + 13 = 64$

$U_1 = 5.8 + 5.6/2 - 27 = 28$, $U_2 = 5.8 + 8.9/2 - 64 = 12$

Kritická hodnota pro $\alpha = 0,05$, $\min(5,8) = 5$, $\max(5,8) = 8$ je 6. Protože $\min(28,12) = 12 > 6$, nemůžeme na hladině významnosti 0,05 zamítnout hypotézu, že obě technologie poskytují stejné procento účinné látky.

Výpočet pomocí systému STATISTICA (1)

Utvoříme nový datový soubor se dvěma proměnnými a 13 případy. Do proměnné X napíšeme zjištěné hodnoty a do proměnné ID napíšeme 5x číslo 1 pro první technologii a 8x číslo 2 pro starý druhou technologii.

Statistiky – Neparametrická statistika – Porovnání dvou nezávislých vzorků – OK – Proměnné – Seznam závislých proměnných X, Nezáv. (grupov.) proměnná ID – OK – M-W U test.

Upozornění: Ve STATISTICE je dvouvýběrový Wilcoxonův test uveden pod názvem Mannův – Whitneyův test.

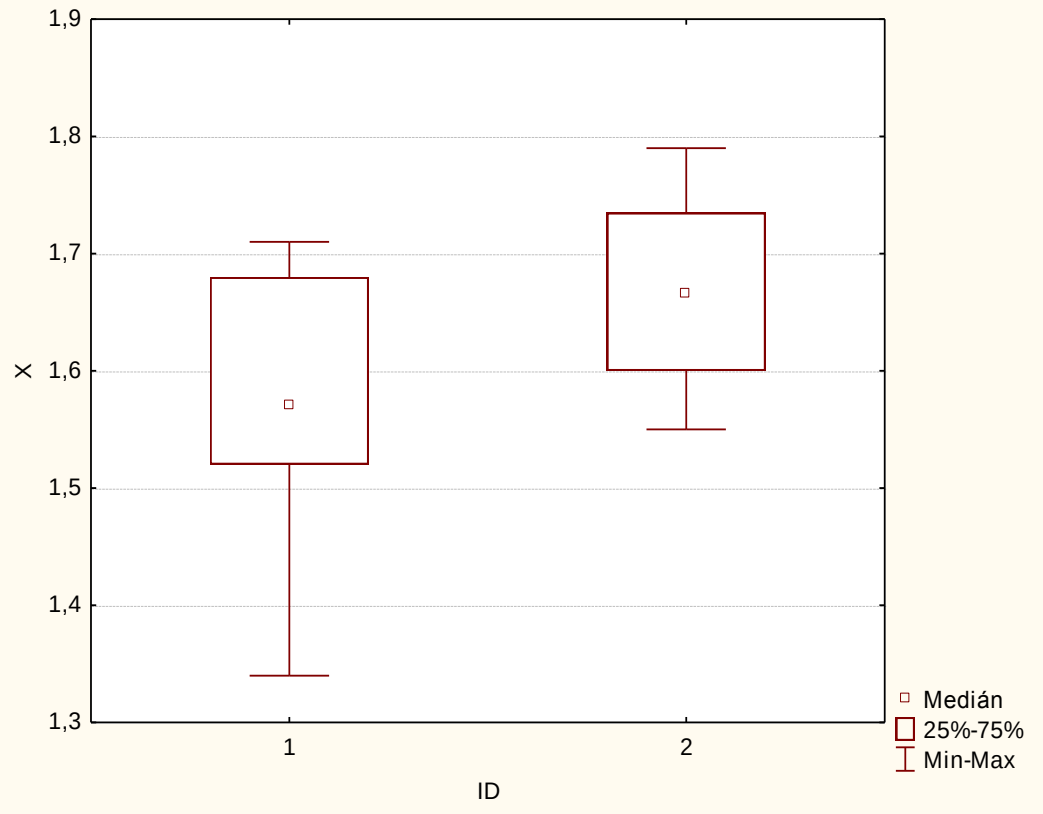
Mann-Whitneyův U test (dve technologie.sta)
Dle proměn. ID
Označené testy jsou významné na hladině p <,05000

Proměnná	Sčt poř. skup. 1	Sčt poř. skup. 2	U	Z	Úroveň p	Z upravené	Úroveň p	N platn. skup. 1	N platn. skup. 2	2*1str. přesné p
X	27,0000	64,0000	12,0000	-1,1710	0,24156	-1,1710	0,24156	5	8	0,28438

Ve výstupní tabulce jsou součty pořadí T_1, T_2, hodnota testové statistiky $\min(U_1, U_2)$ označená U, hodnota asymptotické testové statistiky U_0 (označená Z), asymptotická p-hodnota pro U_0 a přesná p-hodnota (ozn. 2*1str. přesné p – ta se používá pro rozsahy výběrů pod 30). V našem případě přesná p-hodnota = 0,284382, tedy H_0 nezamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA (2)

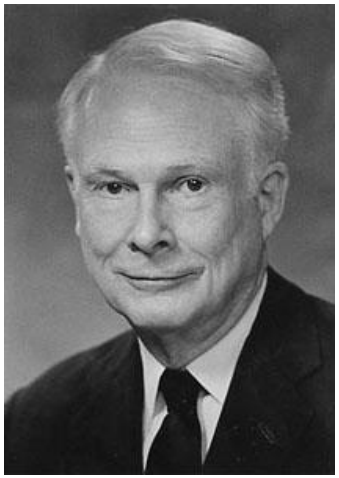
Výpočet je vhodné doplnit krabicovým diagramem.



Je zřejmé, že první technologie poskytuje vesměs nižší procento účinné látky než druhá technologie a také vykazuje poněkud větší variabilitu.

Kruskalův - Wallisův test (1)

William Kruskal (1919 – 2005):
Americký matematik



Wilson Allen Wallis (1912 – 1988):
Americký matematik

Nechť je dáno $r \geq 3$ nezávislých náhodných výběrů o rozsazích $n_1, \ldots, n_r$. Předpokládáme, že tyto výběry pocházejí ze spojitých rozložení. Označme

$n = n_1 + \ldots + n_r$. Na asymptotické hladině významnosti α chceme testovat hypotézu, že všechny tyto výběry pocházejí z téhož rozložení.

Kruskalův - Wallisův test (2)

Postup testu:

a) Všech n hodnot seřadíme do rostoucí posloupnosti.

b) Určíme pořadí každé hodnoty v tomto sdruženém výběru.

c) Označme Tj součet pořadí těch hodnot, které patří do j-tého výběru, j = 1, ..., r (kontrola: musí platit $T_1 + ... + T_r = n(n+1)/2$).

d) Testová statistika má tvar: $Q = \frac{12}{n(n+1)} \sum_{j=1}^{r} \frac{T_j^2}{n_j} - 3(n+1)$. Platí-li H_0, má statistika Q asymptoticky rozložení $\chi^2(r-1)$.

e) Kritický obor: $W = \langle \chi^2_{1-\alpha}(r-1), \infty)$.

f) H_0 zamítneme na asymptotické hladině významnosti α, když $Q \geq \chi^2_{1-\alpha}(r-1)$.

Příklad

V roce 1980 byly získány tři nezávislé výběry obsahující údaje o průměrných ročních příjmech (v tisících dolarů) čtyř sociálních skupin ve třech různých oblastech USA.

jižní oblast: 6 10 15 29

pacifická oblast: 11 13 17 131

severovýchodní oblast: 7 14 28 25

Na hladině významnosti 0,05 testujte hypotézu, že příjmy v těchto oblastech se neliší.

Řešení:

Výpočty uspořádáme do tabulky

Usp. hodnoty	6	7	10	11	13	14	15	17	25	28	29	131
Pořadí 1.výběru	1		3				7				11	
Pořadí 2.výběru				4	5			8				12
Pořadí 3.výběru		2				6			9	10		

$T_1 = 1 + 3 + 7 + 11 = 22, T_2 = 4 + 5 + 8 + 12 = 29, T_3 = 2 + 6 + 9 + 10 = 27$,

$$Q = \frac{12}{n(n+1)} \sum_{j=1}^{r} \frac{{T_j}^2}{n_j} - 3(n+1) = \frac{12}{12 \cdot 13}\left(\frac{22^2}{4} + \frac{29^2}{4} + \frac{27^2}{4}\right) - 3 \cdot 13 = 0{,}5,$$

$$W = \langle {\chi^2}_{1-\alpha}(r-1), \infty) = \langle {\chi^2}_{0{,}95}(2), \infty) = \langle 5{,}991, \infty)$$

Protože $Q < 5{,}991$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Rozdíly mezi průměrnými ročními příjmy v uvedených třech oblastech se neprokázaly.

Mediánový test

Výchozí situace je stejná jako u K-W testu

Postup testu:

a) Všech n hodnot uspořádáme do rostoucí posloupnosti.

b) Najdeme medián $x_{0,50}$ těchto n hodnot.

c) Označme P_j počet hodnot v j-tém výběru, které jsou větší nebo rovny mediánu $x_{0,50}$.

d) Testová statistika má tvar $Q_M = 4 \sum_{j=1}^{r} \frac{{P_j}^2}{n_j} - n$. Platí-li H_0, má statistika Q_M asymptoticky rozložení χ^2(r-1).

e) Kritický obor: $W = \langle {\chi^2}_{1-\alpha}(r-1), \infty)$.

f) H_0 zamítneme na asymptotické hladině významnosti α, když $Q_M \geq {\chi_{1-\alpha}}^2$(r-1).

Příklad

Pro data o průměrných ročních příjmech proveďte mediánový test. Hladinu významnosti volte 0,05.

Řešení:

Usp. hodnoty 6 7 10 11 13 14 15 17 25 28 29 131

Medián je průměr 6. a 7. uspořádané hodnoty: $x_{0,50} = \frac{14+15}{2} = 14{,}5$.

V prvním výběru existují 2 hodnoty, které jsou větší nebo rovny 14,5, stejně tak i ve druhém a třetím výběru, tedy $P_1 = P_2 = P_3 = 2$.

Testová statistika: $Q_M = 4 \sum_{j=1}^{r} \frac{{P_j}^2}{n_j} - n = 4\left[\frac{1}{4}(2^2 + 2^2 + 2^2)\right] - 12 = 0$

Kritický obor: $W = \langle {\chi^2}_{1-\alpha}(r-1), \infty) = \langle {\chi^2}_{0,95}(2), \infty) = \langle 5{,}991, \infty)$

Protože $Q_M < 5{,}991$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Metody mnohonásobného porovnávání

Zamítneme-li hypotézu, že všechny náhodné výběry pocházejí z téhož rozložení, zajímá nás, které dvojice náhodných výběrů se liší na zvolené hladině významnosti. Testujeme H_0: k-tý a l-tý náhodný výběr pocházejí z téhož rozložení, k, l = 1, .., r, k ≠ l proti H_1: aspoň jedna dvojice výběrů pochází z různých rozložení.

a) Neményiho metoda (Peter Neményi 1927 – 2002: Americký matematik maďarského původu)

- Všechny výběry mají týž rozsah p (třídění je vyvážené).
- Vypočteme $|T_l - T_k|$.
- V tabulkách najdeme kritickou hodnotu (pro dané p, r, α).
- Pokud $|T_l - T_k| \geq$ tabelovaná kritická hodnota, pak na hladině významnosti α zamítáme hypotézu, že l-tý a k-tý výběr pocházejí z téhož rozložení.

b) Obecná metoda mnohonásobného porovnávání

- Vypočteme $\left|\frac{T_l}{n_l} - \frac{T_k}{n_k}\right|$.
- Ve speciálních statistických tabulkách najdeme kritickou hodnotu $h_{KW}(\alpha)$. Při větších rozsazích výběrů je možno ji nahradit kvantilem $\chi_{1-\alpha}{}^2(r-1)$.

Jestliže $\left|\frac{T_l}{n_l} - \frac{T_k}{n_k}\right| \geq \sqrt{\frac{1}{12}\left(\frac{1}{n_l} + \frac{1}{n_k}\right) n(n+1) h_{KW}(\alpha)}$, pak na hladině významnosti α zamítáme hypotézu, že l-tý a k-tý výběr pocházejí z téhož rozložení.

Příklad

Čtyři laboranti provedli analytické stanovení procenta niklu v oceli. Každý hodnotil pět vzorků.

Laborant A: 4,15 4,26 4,10 4,30 4,25

Laborant B: 4,38 4,40 4,29 4,39 4,45

Laborant C: 4,23 4,16 4,20 4,24 4,27

Laborant D: 4,41 4,31 4,42 4,37 4,43

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že všechny čtyři náhodné výběry pocházejí ze stejného rozložení. Pokud nulovou hypotézu zamítnete, zjistěte, které dvojice výběrů se liší.

Výpočet pomocí systému STATISTICA (1)

Vytvoříme nový datový soubor o dvou proměnných a 20 případech. Do proměnné nikl napíšeme změřené hodnoty, do proměnné laborant napíšeme 5x1 pro 1. laboranta atd. až 5x4 pro 4. laboranta.

Statistiky – Neparametrická statistika – Porovnání více nezávislých vzorků - OK – Seznam závislých proměnných nikl, Nezáv. (grupovací) proměnná laborant – OK – Summary: Kruskal-Wallis ANOVA & Median test. Ve dvou výstupních tabulkách se objeví výsledky K-W testu a mediánového testu.

	Kruskal-Wallisova ANOVA založ. na poř.; nikl (nikl v ocel Nezávislá (grupovací) proměnná: laborant Kruskal-Wallisův test: H (3, N= 20) =13,77714 p =,003		
Závislá: nikl	Kód	Počet platných	Součet pořadí
1	1	5	29,0000
2	2	5	75,0000
3	3	5	27,0000
4	4	5	79,0000

Výpočet pomocí systému STATISTICA (2)

Závislá: nikl	Mediánový test, celk. medián = 4,29500; nikl (nikl v oceli) Nezávislá (grupovací) proměnná : laborant Chi-Kvadr. = 13,60000 sv = 3 p = ,0035				
	1	2	3	4	Celkem
<= Medián: pozorov.	4,00000	1,00000	5,00000	0,00000	10,00000
očekáv.	2,50000	2,50000	2,50000	2,50000	
poz.-oč.	1,50000	-1,50000	2,50000	-2,50000	
> Medián: pozorov.	1,00000	4,00000	0,00000	5,00000	10,00000
očekáv.	2,50000	2,50000	2,50000	2,50000	
poz.-oč.	-1,50000	1,50000	-2,50000	2,50000	
Celkem: oček.	5,00000	5,00000	5,00000	5,00000	20,00000

Oba testy zamítají hypotézu o shodě mediánů v daných čtyřech skupinách, ale K-W test je poněkud silnější (p-hodnota = 0,0032, zatímco p-hodnota pro mediánový test je 0,0035).

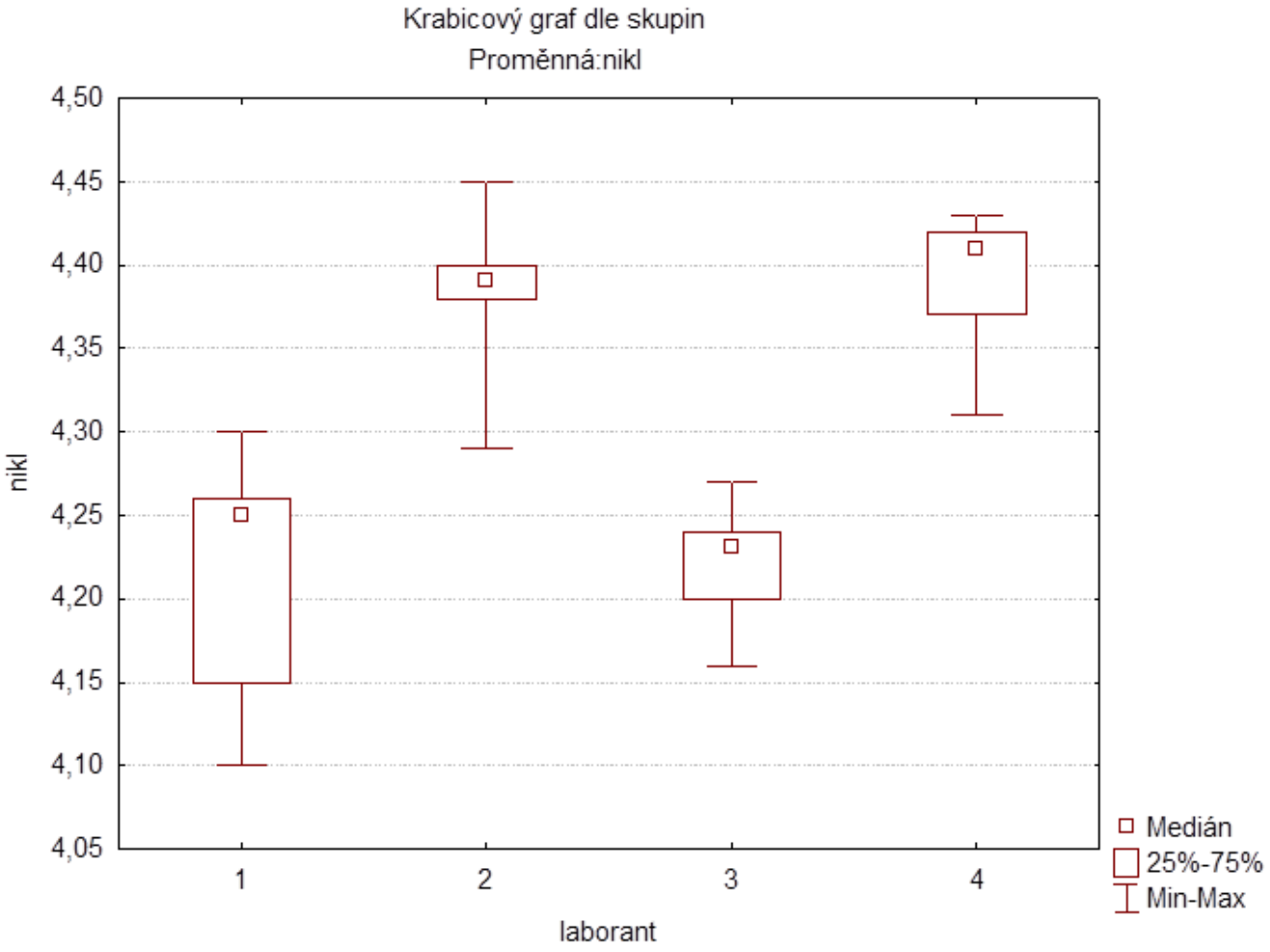
Nyní provedeme mnohonásobné porovnávání, abychom zjistili, které dvojice laborantů se liší. Zvolíme Vícenás. porovnání průměrného pořadí pro vš. skupiny.

Závislá: nikl	Vícenásobné porovnání p hodnot (oboustr.); nikl (nikl v ocel Nezávislá (grupovací) proměnná: laborant Kruskal-Wallisův test: H (3, N= 20) =13,77714 p =,0032			
	1 R:5,800C	2 R:15,00C	3 R:5,400C	4 R:15,80C
1		0,08364	1,00000	0,04515
2	0,08364		0,06177	1,00000
3	1,00000	0,06177		0,03266
4	0,04515	1,00000	0,03266	

Tabulka obsahuje p-hodnoty pro porovnání dvojic skupin. Vidíme, že na hladině významnosti 0,05 se liší laboranti A, D a laboranti C, D.

Výpočet pomocí systému STATISTICA (3)

Grafické znázornění výsledků



11. Testování nezávislosti náhodných veličin

Motivace: Při zpracování dat se velmi často setkáme s úkolem zjistit, zda dvě náhodné veličiny jsou stochasticky nezávislé. Testování hypotézy o nezávislosti se provádí různými způsoby podle toho, jakého typu jsou dané náhodné veličiny – zda jsou nominální, ordinální, intervalové či poměrové. Nominální náhodné veličiny umožňují obsahovou interpretaci pouze u relace rovnosti, ordinální navíc ještě u relace uspořádání, intervalové pak navíc u operace rozdílu a poměrové i u operace podílu.

Např. nás může zajímat, zda ve sledované populaci je barva očí a barva vlasů nezávislá nebo zda počet dnů absence a věk pracovníka jsou nezávislé.

Zpravidla chceme také zjistit intenzitu případné závislosti sledovaných dvou veličin. K tomuto účelu byly zkonstruovány různé koeficienty, které nabývají hodnot od 0 do 1 (resp. od -1 do 1). Čím je takový koeficient bližší 1 (resp. -1), tím je závislost mezi danými dvěma veličinami silnější a čím je bližší 0, tím je slabší.

Definice kontingenční tabulky (1)

Nechť X,Y jsou dvě nominální náhodné veličiny (tj. obsahová interpretace je možná jenom u relace rovnosti). Nechť X nabývá variant $x_{[1]}, \ldots, x_{[r]}$ a Y nabývá variant $y_{[1]}, \ldots, y_{[s]}$.

Označme:

$\pi_{jk} = P(X = x_{[j]} \wedge Y = y_{[k]})$ … simultánní pravděpodobnost dvojice variant $(x_{[j]}, y_{[k]})$

$\pi_{j.} = P(X = x_{[j]})$… marginální pravděpodobnost varianty $x_{[j]}$

$\pi_{.k} = P(Y = y_{[k]})$… marginální pravděpodobnost varianty $y_{[k]}$

Simultánní a marginální pravděpodobnosti zapíšeme do kontingenční tabulky:

	y				
x	π_{jk}	$y_{[1]}$	…	$y_{[s]}$	$\pi_{j.}$
$x_{[1]}$		π_{11}	…	π_{1s}	$\pi_{1.}$
…		…	…	…	…
$x_{[r]}$		π_{r1}	…	π_{rs}	$\pi_{r.}$
$\pi_{.k}$		$\pi_{.1}$	…	$\pi_{.s}$	1

Definice kontingenční tabulky (2)

Nyní pořídíme dvourozměrný náhodný výběr rozsahu n z rozložení, kterým se řídí dvourozměrný diskrétní náhodný vektor (X, Y). Zjištěné absolutní simultánní četnosti n_{jk} dvojice variant $(x_{[j]}, y_{[k]})$ uspořádáme do kontingenční tabulky:

	y	$y_{[1]}$	...	$y_{[s]}$	$n_{j.}$
x	n_{jk}				
$x_{[1]}$		n_{11}	...	n_{1s}	$n_{1.}$
...		...	...	...	...
$x_{[r]}$		n_{r1}	...	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$	...	$n_{.s}$	n

$n_{j.} = n_{j1} + \ldots + n_{js}$ je marginální absolutní četnost varianty $x_{[j]}$

$n_{.k} = n_{1k} + \ldots + n_{rk}$ je marginální absolutní četnost varianty $y_{[k]}$

Simultánní pravděpodobnost π_{jk} odhadneme pomocí simultánní relativní četnosti $p_{jk} = \frac{n_{jk}}{n}$, marginální pravděpodobnosti $\pi_{j.}$ a $\pi_{.k}$ odhadneme pomocí marginálních relativních četností $p_{j.} = \frac{n_{j.}}{n}$ a $p_{.k} = \frac{n_{.k}}{n}$.

Věta o testové statistice K

Testujeme nulovou hypotézu H_0: X, Y jsou stochasticky nezávislé náhodné veličiny proti alternativě H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny.

Kdyby náhodné veličiny X, Y byly stochasticky nezávislé, pak by platil multiplikativní vztah

$\forall j = 1, \dots, r, \forall k = 1, \dots, s$: $\pi_{jk} = \pi_{j.}\ \pi_{.k}$ neboli $\frac{n_{jk}}{n} = \frac{n_{j.}}{n} \cdot \frac{n_{.k}}{n}$, tj. $n_{jk} = \frac{n_{j.} n_{.k}}{n}$. Číslo $m_{jk} = \frac{n_{j.} n_{.k}}{n}$ se nazývá teoretická četnost dvojice variant $(x_{[j]}, y_{[k]})$.

Testová statistika: $K = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j.} n_{.k}}{n}\right)^2}{\frac{n_{j.} n_{.k}}{n}}$.

Platí-li H_0, pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$.

Kritický obor: $W = \langle \chi^2_{1-\alpha}((r-1)(s-1)), \infty)$.

Hypotézu o nezávislosti veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$.

Podmínky dobré aproximace

Rozložení statistiky K lze aproximovat rozložením $\chi^2((r-1)(s-1))$, pokud teoretické četnosti $\frac{n_{j.}n_{.k}}{n}$ aspoň v 80 % případů nabývají hodnoty větší nebo rovné 5 a ve zbylých 20 % neklesnou pod 2. Není-li splněna podmínka dobré aproximace, doporučuje se slučování některých variant.

Definice Cramérova koeficientu, význam jeho hodnot

Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}}$, kde m = min{r,s}. Tento koeficient nabývá hodnot mezi 0 a 1. Čím blíže je 1, tím je těsnější závislost mezi X a Y, čím blíže je 0, tím je tato závislost volnější.

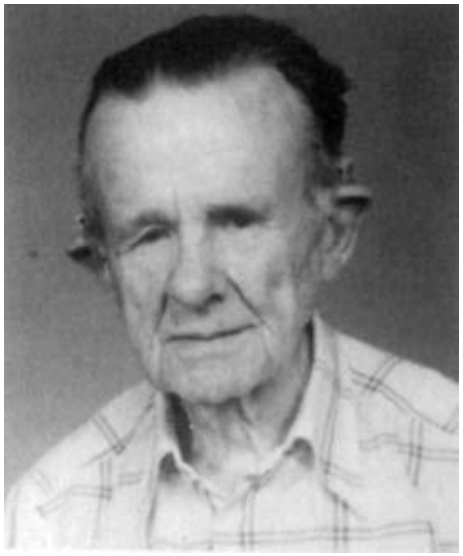
Význam hodnot Cramérova koeficientu:

mezi 0 až 0,1 … zanedbatelná závislost,

mezi 0,1 až 0,3 … slabá závislost,

mezi 0,3 až 0,7 … střední závislost,

mezi 0,7 až 1 … silná závislost.



Příklad (1)

V sociologickém průzkumu byl z uchazečů o studium na vysokých školách pořízen náhodný výběr rozsahu 360. Mimo jiné se zjišťovala sociální skupina, ze které uchazeč pochází a typ školy, na kterou se hlásí. Výsledky jsou zaznamenány v kontingenční tabulce:

Typ školy	Sociální skupina				$n_{j.}$
	I	II	III	IV	
univerzitní	50	30	10	50	140
technický	30	50	20	10	110
ekonomický	10	20	30	50	110
$n_{.k}$	90	100	60	110	360

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti typu školy a sociální skupiny. Vypočtěte Cramérův koeficient.

Příklad (2)

Řešení:

Nejprve vypočteme všech 12 teoretických četností:

$$\frac{n_{1.}n_{.1}}{n} = \frac{140\cdot 90}{360} = 35, \frac{n_{1.}n_{.2}}{n} = \frac{140\cdot 100}{360} = 38{,}9, \frac{n_{1.}n_{.3}}{n} = \frac{140\cdot 60}{360} = 23{,}3, \frac{n_{1.}n_{.4}}{n} = \frac{140\cdot 110}{360} = 42{,}8,$$

$$\frac{n_{2.}n_{.1}}{n} = \frac{110\cdot 90}{360} = 27{,}5, \frac{n_{2.}n_{.2}}{n} = \frac{110\cdot 100}{360} = 30{,}6, \frac{n_{2.}n_{.3}}{n} = \frac{110\cdot 60}{360} = 18{,}3, \frac{n_{2.}n_{.4}}{n} = \frac{110\cdot 110}{360} = 33{,}6,$$

$$\frac{n_{3.}n_{.1}}{n} = \frac{110\cdot 90}{360} = 27{,}5, \frac{n_{3.}n_{.2}}{n} = \frac{110\cdot 100}{360} = 30{,}6, \frac{n_{3.}n_{.3}}{n} = \frac{110\cdot 60}{360} = 18{,}3, \frac{n_{3}n_{.4}}{n} = \frac{110\cdot 110}{360} = 33{,}6.$$

Vidíme, že podmínky dobré aproximace jsou splněny, všechny teoretické četnosti převyšují číslo 5.

Nyní dosadíme do vzorce pro testovou statistiku K:

$$K = \frac{(50-35)^2}{35} + \frac{(30-38{,}9)^2}{38{,}9} + \cdots + \frac{(50-33{,}6)^2}{33{,}6} = 76{,}84,\ \mathrm{r} = 3,\ \mathrm{s} = 4,\ \chi^2_{0{,}95}(6) = 12{,}6.$$

Protože $K \geq 12{,}6$, hypotézu o nezávislosti typu školy a sociální skupiny zamítáme na asymptotické hladině významnosti 0,05. Cramérův koeficient: $V = \sqrt{\frac{76{,}4}{360\cdot 2}} = 0{,}3267.$

Hodnota Cramérova koeficientu svědčí o tom, že mezi veličinami X a Y existuje středně silná závislost.

Výpočet pomocí systému STATISTICA (1)

Vytvoříme nový datový soubor o třech proměnných (X - sociální skupina, Y – typ školy, četnost) a 12 případech:

	1 X	2 Y	3 četnost
1	I	univerzitní	50
2	I	technický	30
3	I	ekonomický	10
4	II	univerzitní	30
5	II	technický	50
6	II	ekonomický	20
7	III	univerzitní	10
8	III	technický	20
9	III	ekonomický	30
10	IV	univerzitní	50
11	IV	technický	10
12	IV	ekonomický	50

Statistiky – Základní statistiky/tabulky – OK – Specif. Tabulky – List 1 X, List 2 Y – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Očekávané četnosti. Dostaneme kontingenční tabulku teoretických četností:

Souhrnná tab.: Očekávané četnosti (typ skoly)
Četnost označených buněk > 10
Pearsonův chí-kv. : 76,8359, sv=6, p=,000000

X	Y univerzitní	Y technický	Y ekonomický	Řádk. součty
I	35,000(	27,500(	27,500(	90,000(
II	38,888(	30,555(	30,555(	100,000(
III	23,333(	18,333(	18,333(	60,000(
IV	42,777(	33,611(	33,611(	110,000(
Vš.skup.	140,000(	110,000(	110,000(	360,000(

Výpočet pomocí systému STATISTICA (2)

Všechny teoretické četnosti jsou větší než 5, podmínky dobré aproximace jsou splněny. V záhlaví tabulky je uvedena hodnota testové statistiky K = 76,8359, počet stupňů volnosti 6 a odpovídající p-hodnota. Je velmi blízká 0, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o nezávislosti typu školy a sociální skupiny.

Hodnotu testové statistiky a Cramérův koeficient dostaneme také tak, že na na záložce Možnosti zaškrtneme Pearsonův & M-V chí kvadrát a Cramérovo V a na záložce Detailní výsledky vybereme Detailní 2 rozm. tabulky.

	Statist. : X(4) x Y(3) (typ skoly.sta		
Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	76,8358	df=6	p=,0000
M-V chí-kvadr.	84,5352	df=6	p=,0000
Fí	,461988		
Kontingenční koeficient	,419394		
Cramér. V	,326674		

Čuprovův koeficient kontingence

Mezi další používané míry závislosti patří následující:

Průměrná čtvercová kontingence: $\Phi^2 = \frac{K}{n}$

ϕ – koeficient: $\phi = \sqrt{\phi^2}$

Pearsonův koeficient kontingence: $P = \sqrt{\frac{K}{K+n}} = \sqrt{\frac{\Phi^2}{\Phi^2+1}}$

$0 \le P < 1$, přičemž hodnoty jedna nemůže nikdy dosáhnout.

Čuprovovův koeficient kontingence: $T = \sqrt{\frac{\Phi^2}{\sqrt{(r-1)(s-1)}}}$

$0 \le T \le 1$ Vhodný zejména pokud se významně liší *r* a *s* (obdélníkové tabulky).

Pro čtvercové tabulky (r=s) platí: $V = T$

Definice čtyřpolní kontingenční tabulky

Nechť r = s = 2. Pak hovoříme o čtyřpolní kontingenční tabulce a používáme označení: $n_{11} = a$, $n_{12} = b$, $n_{21} = c$, $n_{22} = d$.

X	Y		$n_{j.}$
	$y_{[1]}$	$y_{[2]}$	
$x_{[1]}$	a	b	a+b
$x_{[2]}$	c	d	c+d
$n_{.k}$	a+c	b+d	n

Testová statistika K pro čtyřpolní kontingenční tabulku se dá zjednodušit do tvaru:

$$K = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Kritický obor: $W = \langle \chi^2{}_{1-\alpha}(1), \infty)$

Věta o testové statistice K pro čtyřpolní tabulky

Testová statistika K pro čtyřpolní kontingenční tabulku se dá zjednodušit do tvaru:

$$K = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Kritický obor: $W = \langle \chi^2{}_{1-\alpha}(1), \infty)$. Hypotézu o nezávislosti náhodných veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \in W$.

Poznámka: U čtyřpolní KT lze rovněž použít následující podmínky dobré aproximace: $a + b > 5$, $c + d > (a + c)/3$.

Poznámka: Pro čtyřpolní tabulku navrhl R. A. Fisher přesný (exaktní) test nezávislosti známý jako Fisherův faktoriálový test. (Je popsán např. v knize K. Zvára: Biostatistika, Karolinum, Praha 1998.) Jestliže p-hodnota pro tento test $\leq \alpha$, pak hypotézu o nezávislosti zamítáme na hladině významnosti α.

Příklad

U 125 uchazečů o studium na jistou fakultu byl hodnocen dojem, jakým zapůsobili na komisi u ústní přijímací zkoušky. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že přijetí na fakultu nezávisí na dojmu u přijímací zkoušky.

přijetí	dojem		$n_{j.}$
	dobrý	špatný	
ano	17	11	28
ne	39	58	97
$n_{.k}$	56	69	125

Řešení:

Ověříme splnění podmínek dobré aproximace:

$a + b = 28 > 5$, $c + d = 97 > (a + c)/3 = 56/3 = 18{,}66$ – v pořádku

Dosadíme do zjednodušeného vzorce pro testovou statistiku K:

$$K = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{125 \cdot (17 \cdot 58 - 11 \cdot 39)^2}{28 \cdot 97 \cdot 56 \cdot 69} = 3{,}6953$$

Kritický obor: $W = \langle {\chi^2}_{0,95}(1), \infty) = \langle 3{,}841, \infty)$.

Protože testová statistika se nerealizuje k kritickém oboru, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Definice podílu šancí

Ve čtyřpolních tabulkách používáme charakteristiku $OR = \frac{ad}{bc}$, která se nazývá podíl šancí (odds ratio). Můžeme si představit, že pokus se provádí za dvojích různých okolností a může skončit buď úspěchem nebo neúspěchem.

Výsledek pokusu	okolnosti		$n_{j.}$
	I	II	
úspěch	a	b	a+b
neúspěch	c	d	c+d
$n_{.k}$	a+c	b+d	n

Poměr počtu úspěchů k počtu neúspěchů (tzv. šance) za 1. okolností je $\frac{a}{c}$, za druhých okolností je $\frac{b}{d}$. Podíl šancí je $OR = \frac{ad}{bc}$.

Asymptotický interval spolehlivosti pro podíl šancí a jeho využití k testování hypotézy o nezávislosti

Asymptotický 100(1-α)% interval spolehlivosti pro skutečný podíl šancí má meze:

$$d = \exp\left(\ln OR - \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\, u_{1-\frac{\alpha}{2}} \right)$$

$$h = \exp\left(\ln OR + \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\, u_{1-\frac{\alpha}{2}} \right)$$

Jestliže interval spolehlivosti neobsahuje 1, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

Příklad

Pro údaje z minulého příkladu vypočtěte a interpretujte podíl šancí, sestrojte 95% asymptotický interval spolehlivosti pro podíl šancí a s jeho pomocí testujte hypotézu, že přijetí na fakultu nezávisí na dojmu u přijímací zkoušky.

Řešení: $OR = \frac{ad}{bc} = \frac{17 \cdot 58}{11 \cdot 39} = 2{,}298$. Podíl šancí nám říká, že uchazeč, který zapůsobil na komisi dobrým dojmem, má asi 2,3 x větší šanci na přijetí než uchazeč, který zapůsobil špatným dojmem. Provedeme další pomocné výpočty:

$$\ln \mathrm{OR} = 0{,}832, \quad \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{17}+\frac{1}{11}+\frac{1}{39}+\frac{1}{58}} = 0{,}439, \quad u_{0{,}975} = 1{,}96$$

Dosadíme do vzorců pro meze asymptotického intervalu spolehlivosti pro podíl šancí:

$$\ln d = \ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2} = 0{,}832 - 0{,}439 \cdot 1{,}96 = -0{,}028$$

$$\ln h = \ln OR + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2} = 0{,}832 + 0{,}439 \cdot 1{,}96 = 1{,}692$$

Po odlogaritmování dostaneme: $d = e^{-0{,}028} = 0{,}972$, $\mathrm{h} = \mathrm{e}^{1{,}692} = 5{,}433$

Protože interval (0,972; 5,433) obsahuje číslo 1, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

Výpočet pomocí systému STATISTICA

Dolní a horní mez intervalu spolehlivosti pro OR zjistíme pomocí STATISTIKY. Vytvoříme datový soubor o dvou proměnných DM a HM a jednom případu. Do Dlouhého jména proměnné DM napíšeme vzorec pro dolní mez:

=exp(log(2,298)-sqrt(1/17+1/11+1/39+1/58)*VNormal(0,975;0;1))

a analogicky do Do Dlouhého jména proměnné HM napíšeme vzorec pro horní mez:

=exp(log(2,298)+sqrt(1/17+1/11+1/39+1/58)*VNormal(0,975;0;1))

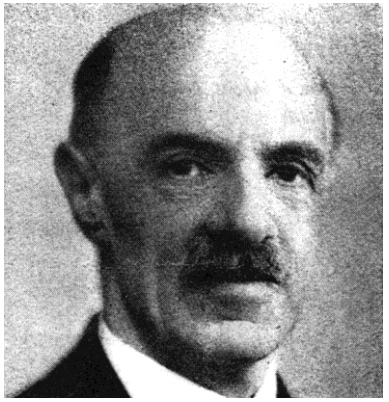
	1 DM	2 HM
1	0,97224	5,43156

Definice Spearmanova koeficientu pořadové korelace, význam jeho hodnot

Nechť X,Y jsou náhodné veličiny aspoň ordinálního typu. Pořídíme dvourozměrný náhodný výběr $(X_1, Y_1), \ldots, (X_n, Y_n)$ 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, $i = 1, \ldots, n$.

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

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.



Věta o testování hypotézy o pořadové nezávislosti veličin X, Y

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 $\alpha = 0{,}05$ nebo $0{,}01$ a $n \leq 30$ najdeme v tabulkách. Pozor – kritické hodnoty pro jednostranné alternativy se v běžně dostupných tabulkách nenajdou.

Asymptotická varianta 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: $W = (-\infty, -t_{1-\alpha/2}(n-2)\rangle \cup \langle t_{1-\alpha/2}(n-2), \infty)$

Kritický obor pro levostrannou alternativu: $W = (-\infty, -t_{1-\alpha}(n-2)\rangle$

Kritický obor pro pravostrannou alternativu: $W = \langle t_{1-\alpha}(n-2), \infty)$.

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

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 r_S a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou lékařů jsou pořadově nezávislá.

Řešení:

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

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 <,05			
Dvojice proměnnýc	Počet plat.	Spearman R	t(N-2)	Úroveň p
X & Y	7	0,85714	3,72104	0,01369

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.

Definice Pearsonova koeficientu korelace

Nechť (X, Y) je náhodný vektor, přičemž náhodné veličiny X, Y jsou aspoň intervalového typu. Číslo

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

se nazývá Pearsonův koeficient korelace.

(Pro výpočet Pearsonova koeficentu korelace musíme znát simultánní distribuční funkci $\Phi(x,y)$ v obecném případě resp. simultánní hustotu pravděpodobnosti $\varphi(x,y)$ ve spojitém případě resp. simultánní pravděpodobnostní funkci $\pi(x,y)$ v diskrétním případě.)

Věta o vlastnostech koeficientu korelace

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

b) $R(a_1 + b_1X, a_2 + b_2Y) = \operatorname{sgn}(b_1b_2)\, R(X, Y) = \begin{cases} R(X,Y) \text{ pro } b_1 b_2 > 0 \\ -R(X,Y) \text{ pro } b_1 b_2 < 0 \end{cases}$

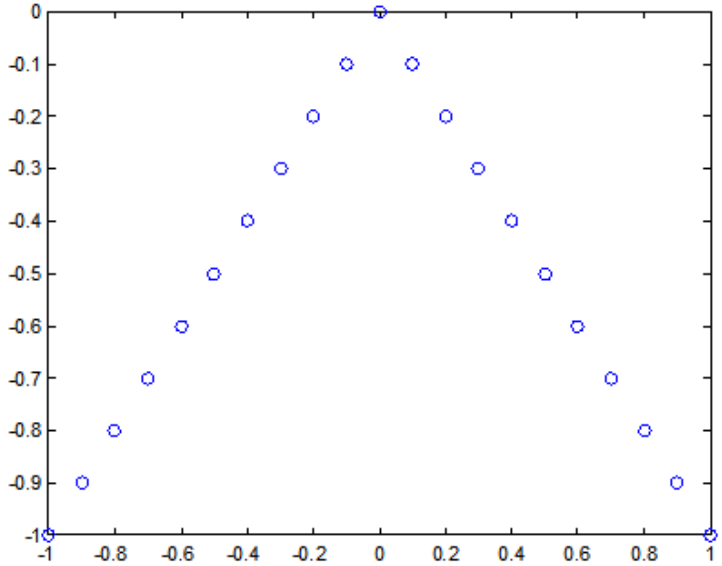
c) $R(X, X) = 1$ pro $D(X) \neq 0$, $R(X, X) = 0$ jinak

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.)

Ilustrace



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.)

Definice výběrového koeficientu korelace

Nechť $(X_1, Y_1), \ldots, (X_n, Y_n)$ 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

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

Poznámka: Vlastnosti Pearsonova koeficientu korelace se přenášejí i na výběrový koeficient korelace.

Věta o koeficientu korelace dvourozměrného normálního rozložení

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

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

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

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

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!

Upozornění: nadále budeme předpokládat, že $(X_1, Y_1), \ldots, (X_n, Y_n)$ je náhodný výběr rozsahu n z dvourozměrného normálního rozložení $N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$

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

– levostranné alternativě: $W = (-\infty, -t_{1-\alpha}(n-2)\rangle$,

– pravostranné alternativě: $W = \langle t_{1-\alpha}(n-2), \infty)$.

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

Příklad (1)

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

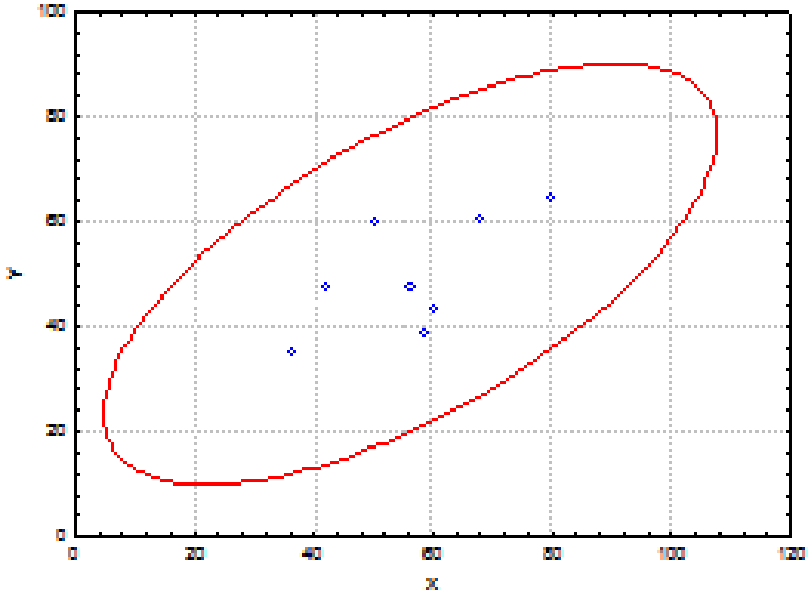
Číslo studenta	1	2	3	4	5	6	7	8
Počet bodů v 1. testu	80	50	36	58	42	60	56	68
Počet bodů ve 2. testu	65	60	35	39	48	44	48	61

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, protože vrstevnice hustoty dvourozměrného normálního rozložení jsou elipsy.

Příklad (2)



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.

Testujeme H_0: $\rho = 0$ proti pravostranné alternativě H_1: $\rho > 0$.

Výpočtem zjistíme: $R_{12} = 0{,}6668$, $T = 2{,}1917$. V tabulkách najdeme $t_{0,95}(6) = 1{,}9432$. Kritický obor: $W = \langle 1{,}9432; \infty)$. Protože $T \in W$, hypotézu o neexistenci kladné korelace výsledků z 1. a 2. testu zamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA

a) Vytvoříme datový soubor o dvou proměnných X, Y a 8 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.

Korelace (dva testy.sta)
Označ. korelace jsou významné na hlad. p < ,05000
(Celé případy vynechány u ChD)

Prom. X & prom. Y	Průměr	Sm.Odch.	r(X,Y)	r2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
X	56,2500	13,9974									
X	56,2500	13,9974	1,00000	1,00000			8	0,0000	1,00000	0,0000	1,00000
X	56,2500	13,9974									
Y	50,0000	10,9283	0,66680	0,44462	2,19169	0,07090	8	20,7163	0,52059	13,5466	0,85406
Y	50,0000	10,9283									
X	56,2500	13,9974	0,66680	0,44462	2,19169	0,07090	8	13,5466	0,85406	20,7163	0,52059
Y	50,0000	10,9283									
Y	50,0000	10,9283	1,00000	1,00000			8	0,0000	1,00000	0,0000	1,00000

Výběrový koeficient korelace se realizoval hodnotou 0,6668, testová statistika nabyla hodnoty 2,1917, odpovídající p-hodnota pro oboustranný test je 0,0709, tedy pro jednostranný test je 0,035045. Na hladině významnosti 0,05 zamítáme hypotézu o nezávislosti veličin X, Y ve prospěch pravostranné alternativy.

b) 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.

Test o 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

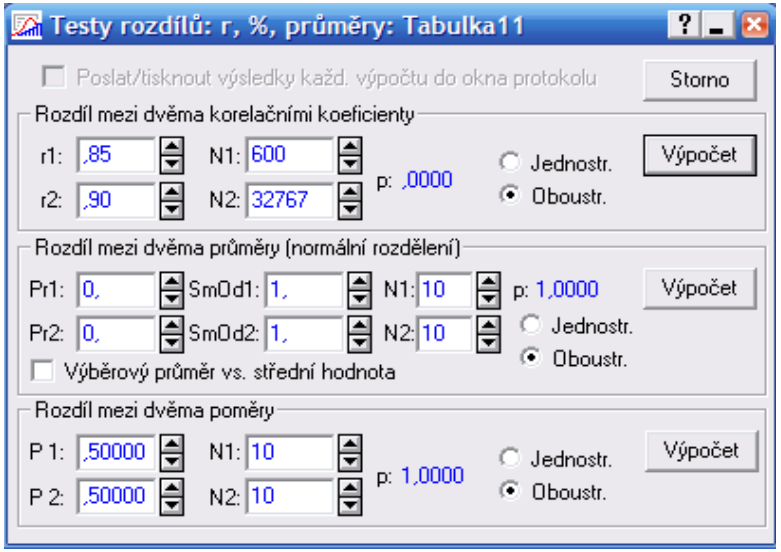
U 600 vzorků rudy byl stanoven obsah železa dvěma analytickými metodami s výběrovým koeficientem korelace 0,85. V literatuře se uvádí, že koeficient korelace těchto dvou metod má být 0,9. Na asymptotické hladině významnosti 0,05 testujte hypotézu H_0: $\rho = 0{,}9$ proti H_1: $\rho \neq 0{,}9$.

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.

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.

Test o 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.

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.

Věta o asymptotickém intervalu spolehlivosti pro koeficient korelace

Nechť dvourozměrný náhodný výběr rozsahu n pochází z dvourozměrného normálního rozložení s koeficientem korelace ρ. Meze 100(1-α)% asymptotického intervalu spolehlivosti pro ρ jsou:

$$d = tgh(Z - \frac{u_{1-\alpha/2}}{\sqrt{n-3}}),\ h = tgh(Z + \frac{u_{1-\alpha/2}}{\sqrt{n-3}})$$

přičemž $tgh\ \mathrm{x} = \frac{\mathrm{e}^{\mathrm{X}} - e^{-x}}{\mathrm{e}^{\mathrm{X}} + e^{-x}}$, $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$.

Příklad (1)

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.

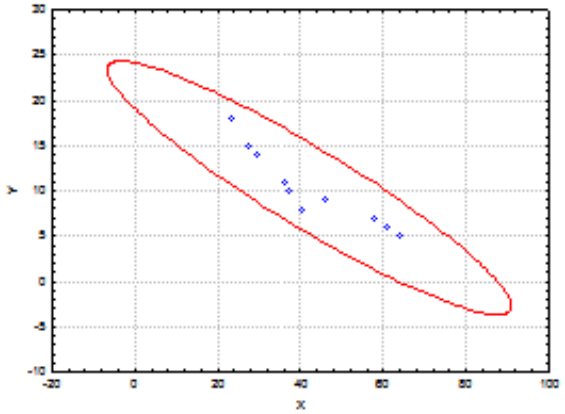
Č.prac.	1	2	3	4	5	6	7	8	9	10
X	27	61	37	23	46	58	29	36	64	40
Y	15	6	10	18	9	7	14	11	5	8

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.

Příklad (2)



Vzhled diagramu svědčí o tom, že předpoklad je oprávněný.

Testujeme H_0: $\rho = 0$ proti H_1: $\rho \neq 0$. 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 $W = (-\infty, -2{,}306\rangle \cup \langle 2{,}306, \infty)$. 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 $tgh\left(-1{,}6772 \pm \frac{1{,}96}{\sqrt{7}}\right)$, tedy $-0{,}9842 < \rho < -0{,}7336$ s pravděpodobností přibližně 0,95.

12. Jednoduchá lineární regrese

Motivace: Cíl regresní analýzy - popsat závislost hodnot veličiny Y na hodnotách veličiny X.

Nutnost vyřešení dvou problémů:

a) jaký typ funkce se použije k popisu dané závislosti;

b) jak se stanoví konkrétní parametry daného typu funkce?

Specifikace klasického modelu lineární regrese

$Y = m(x; \beta_0, \beta_1, \dots, \beta_p) + \varepsilon$, kde

$m(x; \beta_0, \beta_1, \dots, \beta_p)$ – teoretická regresní funkce, která lineárně závisí na neznámých regresních parametrech $\beta_0, \beta_1, \dots, \beta_p$ a známých funkcích $f_1(x), \dots, f_p(x)$, které již neobsahují neznámé parametry, tj. $m(x; \beta_0, \beta_1, \dots, \beta_p) = \sum_{j=0}^{p} \beta_j f_j(x)$, přičemž $f_0(x) \equiv 1$.

Složka ε – náhodná odchylka .

Veličina Y – závisle proměnná (též vysvětlovaná) veličina.

Veličina X – nezávisle proměnná (též vysvětlující) veličina.

Pořídíme n dvojic pozorování $(x_1, y_1), \dots, (x_n, y_n)$, pro i = 1, ..., n platí:
$y_i = m(x_i; \beta_0, \beta_1, \dots, \beta_p) + \varepsilon_i$.

O náhodných odchylkách předpokládáme, že

a) $E(\varepsilon_i) = 0$ (odchylky nejsou systematické)

b) $D(\varepsilon_i) = \sigma^2 > 0$ (všechna pozorování jsou prováděna s touž přesností)

c) $C(\varepsilon_i, \varepsilon_j) = 0$ pro $i \neq j$ (mezi náhodnými odchylkami neexistuje žádný lineární vztah)

d) $\varepsilon_i \sim N(0, \sigma^2)$.

V tomto případě hovoříme o klasickém modelu lineární regrese.

Označení

$b_0, b_1, \ldots, b_p$ – odhady regresních parametrů $\beta_0, \beta_1, \ldots, \beta_p$ (nejčastěji je získáme metodou nejmenších čtverců, tj. z podmínky, že výraz $\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j f_j(x_i) \right)^2$ nabývá svého minima pro $\beta_j = b_j$, $j = 0, 1, \ldots, p$)

$\hat{m}(x; b_0, \ldots, b_p)$ – empirická regresní funkce

$\hat{y}_i = \hat{m}(x_i; b_0, \ldots, b_p) = \sum_{j=0}^{p} b_j f_j(x_i)$ – regresní odhad i-té hodnoty veličiny Y (i-tá predikovaná hodnota veličiny Y)

$e_i = y_i - \hat{y}_i$ – i-té reziduum

$S_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ – reziduální součet čtverců

$s^2 = \frac{S_E}{n-p-1}$ – odhad rozptylu σ^2

$S_R = \sum_{i=1}^{n} (\hat{y}_i - m_2)^2$ – regresní součet čtverců $\left(m_2 = \frac{1}{n} \sum_{i=1}^{n} y_i \right)$

$S_T = \sum_{i=1}^{n} (y_i - m_2)^2$ – celkový součet čtverců $(S_T = S_R + S_E)$

$ID^2 = \frac{S_R}{S_T} = 1 - \frac{S_E}{S_T}$ – index determinace $(0 \leq ID^2 \leq 1)$

$ID_{adj}{}^2 = ID^2 - \frac{(1-ID^2)p}{n-p-1}$ – adjustovaný index determinace

Maticový zápis klasického modelu lineární regrese (1)

$y = X\beta + \varepsilon$, kde $y = (y_1, \dots, y_n)'$ – vektor pozorování závisle proměnné veličiny Y,

$X = \begin{pmatrix} 1 & f_1(x_1) & \dots & f_p(x_1) \\ \dots & \dots & \dots & \dots \\ 1 & f_1(x_n) & \dots & f_p(x_n) \end{pmatrix}$ – regresní matice (předpokládáme, že h(**X**) = p+1 < n)

$\beta = (\beta_0, \beta_1, \dots, \beta_p)'$ – vektor regresních parametrů,

$\varepsilon = (\varepsilon_1, \dots, \varepsilon_p)'$ – vektor náhodných odchylek.

Podmínky (a) až (d) lze zkráceně zapsat ve tvaru $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$.

Maticově zapsaná metoda nejmenších čtverců vede na rovnice

$\mathbf{X'X\beta} = \mathbf{X'y}$ – systém normálních rovnic

$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ – odhad vektoru $\boldsymbol{\beta}$ získaný metodou nejmenších čtverců

$\hat{y} = \mathbf{Xb}$ – vektor regresních odhadů (vektor predikce)

$\mathbf{e} = \mathbf{y} - \hat{y}$ - vektor reziduí

Maticový zápis klasického modelu lineární regrese (2)

Vlastnosti odhadu **b**:

– odhad **b** je lineární, neboť je vytvořen lineární kombinací pozorování $y_1, \ldots, y_n$ s maticí vah $(X'X)^{-1}X'$;

– odhad **b** je nestranný, neboť $E(\mathbf{b}) = \boldsymbol{\beta}$;

– odhad **b** má varianční matici $\text{var } \mathbf{b} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$;

– odhad **b** ~ $N_{p+1}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$ vzhledem k platnosti podmínky (d);

– pro odhad **b** platí Gaussova - Markovova věta: Odhad $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ je nejlepší nestranný lineární odhad vektoru **β**. (Nejlepší v tom smyslu, že rozdíl varianční matice libovolného jiného nestranného odhadu vektoru **β** a varianční matice odhadu **b** je matice pozitivně semidefinitní.)

Intervaly spolehlivosti pro regresní parametry

$s_{b_j} = s\sqrt{v_{jj}}$ – směrodatná chyba odhadu b_j, kde v_{jj} je j-tý diagonální prvek matice $(\mathbf{X}'\mathbf{X})^{-1}$.

Pro j = 0, 1, ..., p statistika $T_j = \frac{b_j - \beta_j}{s_{b_j}} \sim t(n-p-1)$, tedy

100(1 – α)% interval spolehlivosti pro β_j má meze:

$b_j \pm t_{1-\alpha/2}(n-p-1)s_{b_j}$.

Testování významnosti modelu jako celku (celkový F-test)

Na hladině významnosti α testujeme

H_0: $(\beta_1, \dots, \beta_p)' = (0, \dots, 0)'$ proti H_1: $(\beta_1, \dots, \beta_p)' \neq (0, \dots, 0)'$.

(Nulová hypotéza říká, že dostačující je model konstanty.)

Testová statistika: $F = \frac{S_R/p}{S_E/(n-p-1)}$ má rozložení F(p, n-p-1), pokud H_0 platí.

Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty)$.

$F \in W \Rightarrow H_0$ zamítáme na hladině významnosti α.

Výsledky F-testu zapisujeme do tabulky analýzy rozptylu:

zdroj variability	součet čtverců	stupně volnosti	podíl	statistika F
model	S_R	p	S_R/p	$\frac{S_R/p}{S_E/(n-p-1)}$
reziduální	S_E	n-p-1	$S_E/(n-p-1)$	-
celkový	S_T	n-1	-	-

Testování významnosti regresních parametrů (dílčí t-testy)

Na hladině významnosti α pro j = 0,1, ..., p testujeme hypotézu

H_0: $\beta_j = 0$ proti H_1: $\beta j \neq 0$.

Testová statistika: $T_j = \frac{b_j}{s_{b_j}}$ má rozložení $t(n - p - 1)$, pokud H_0 platí.

Kritický obor:

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

$T_j \in W \Rightarrow H_0$ zamítáme na hladině významnosti α.

Příklad (1)

U šesti obchodníků byla zjišťována poptávka po určitém druhu zboží loni (veličina X – v kusech) a letos (veličina Y – v kusech).

číslo. obchodníka	1	2	3	4	5	6
poptávka loni (X)	20	60	70	100	150	260
poptávka letos (Y)	50	60	60	120	230	320

a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení. Vypočtěte výběrový koeficient korelace mezi X a Y, interpretujte jeho hodnotu a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny.

b) Předpokládejte, že závislost letošní poptávky na loňské lze vystihnout regresní přímkou. Sestavte regresní matici, vypočtěte odhady regresních parametrů a napište rovnici regresní přímky. Interpretujte parametry regresní přímky.

c) Najděte odhad rozptylu, vypočtěte index determinace a interpretujte ho.

d) Najděte 95% intervaly spolehlivosti pro regresní parametry.

e) Na hladině významnosti 0,05 proveďte celkový F-test.

f) Na hladině významnosti 0,05 proveďte dílčí t-testy.

g) Vypočtěte regresní odhad letošní poptávky při loňské poptávce 110 kusů.

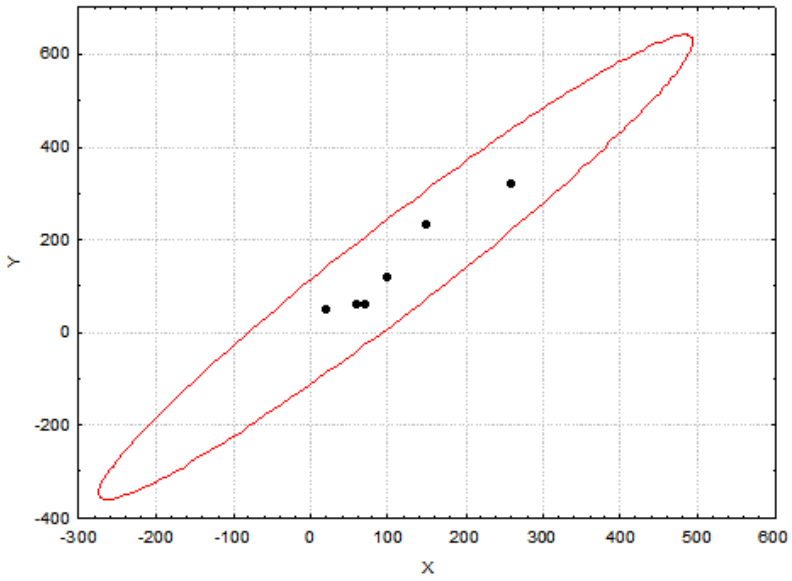
h) Nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou.

Příklad (2)

Řešení:

ad a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení.

Vytvoříme dvourozměrný tečkový diagram s proloženou 95% elipsou konstantní hustoty pravděpodobnosti:



Ze vzhledu diagramu je patrné, že předpoklad dvourozměrné normality je oprávněný a že mezi loňskou a letošní poptávkou existuje vcelku silná přímá lineární závislost.

Příklad (3)

Vypočtěte výběrový koeficient korelace mezi X a Y, interpretujte jeho hodnotu a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny.

Výpočtem zjistíme: $r_{12} = 0{,}972$, tedy mezi poptávkou loni a letos existuje velmi silná přímá lineární závislost.

Realizace testové statistiky: $t = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} = \frac{0{,}972\sqrt{6-2}}{\sqrt{1-0{,}972^2}} = 8{,}2695.$

Kritický obor: $W = (-\infty, -t_{1-\alpha/2}(n-2)\rangle \cup \langle t_{1-\alpha/2}(n-2), \infty) = (-\infty, -t_{0{,}975}(4)\rangle \cup \langle t_{0{,}975}(4), \infty) = (-\infty, -2{,}7764\rangle \cup \langle 2{,}7764, \infty)$

Testová statistika se realizuje v kritickém oboru, hypotézu o nezávislosti veličin X a Y tedy zamítáme na hladině významnosti 0,05.

Příklad (4)

ad b) Sestavíme regresní matici.

$$X = \begin{pmatrix} 1 & f_1(x_1) & \dots & f_p(x_1) \\ \dots & \dots & \dots & \dots \\ 1 & f_1(x_n) & \dots & f_p(x_n) \end{pmatrix}, \text{ tedy } \mathbf{X} = \begin{pmatrix} 1 & 20 \\ 1 & 60 \\ 1 & 70 \\ 1 & 100 \\ 1 & 150 \\ 1 & 260 \end{pmatrix}.$$

Podle vzorce $b = (X'X)^{-1}X'y$ získáme odhady regresních parametrů.

Nejprve vypočítáme matici

$\mathbf{X'X} = \begin{pmatrix} 6 & 660 \\ 660 & 109000 \end{pmatrix}$ a k ní inverzní matici

$(\mathbf{X'X})^{-1} = \begin{pmatrix} 0{,}499084 & -0{,}003022 \\ -0{,}003022 & 0{,}000027 \end{pmatrix}.$

Příklad (5)

Dále získáme součin

$$\mathbf{X'y} = \begin{pmatrix} 840 \\ 138500 \end{pmatrix}$$

a nakonec vektor odhadů regresních parametrů:

$$\mathbf{b} = \begin{pmatrix} 0{,}499084 & -0{,}003022 \\ -0{,}003022 & 0{,}000027 \end{pmatrix} \begin{pmatrix} 840 \\ 138500 \end{pmatrix} = \begin{pmatrix} 0{,}6868 \\ 1{,}2665 \end{pmatrix}$$

Regresní přímka má tedy rovnici

$y = 0{,}6868 + 1{,}2665\ x.$

Znamená to, že při nulové loňské poptávce by letošní poptávka činila 0,6868 kusů a při zvýšení loňské poptávky o 10 kusů by se letošní poptávka zvedla o 12,665 kusů.

Příklad (6)

ad c) Nyní vypočteme vektor regresních odhadů proměnné Y (vektor predikce):

$$\hat{y} = \mathbf{Xb} = \begin{pmatrix} 1 & 20 \\ 1 & 60 \\ 1 & 70 \\ 1 & 100 \\ 1 & 150 \\ 1 & 260 \end{pmatrix} \cdot \begin{pmatrix} 0{,}6868 \\ 1{,}2665 \end{pmatrix} = \begin{pmatrix} 26{,}02 \\ 76{,}68 \\ 89{,}34 \\ 127{,}34 \\ 190{,}66 \\ 329{,}97 \end{pmatrix}.$$

Stanovíme vektor reziduí:

$$e = y - \hat{y} = \begin{pmatrix} 50 \\ 60 \\ 60 \\ 120 \\ 230 \\ 320 \end{pmatrix} - \begin{pmatrix} 26{,}02 \\ 76{,}68 \\ 89{,}34 \\ 127{,}34 \\ 190{,}66 \\ 329{,}97 \end{pmatrix} = \begin{pmatrix} 23{,}98 \\ -16{,}68 \\ -29{,}34 \\ -7{,}34 \\ 39{,}34 \\ -9{,}97 \end{pmatrix}.$$

Příklad (7)

Pomocí vektoru reziduí vypočteme reziduální součet čtverců:

$$S_E = \mathbf{e}'\mathbf{e} = (23{,}98 - 16{,}68 - 29{,}34 - 7{,}34 \;\; 39{,}34 - 9{,}97).\begin{pmatrix} 23{,}98 \\ -16{,}68 \\ -29{,}34 \\ -7{,}34 \\ 39{,}34 \\ -9{,}97 \end{pmatrix} = 3451{,}11.$$

Odhad rozptylu: $s^2 = \frac{S_E}{n-p-1} = \frac{3415{,}11}{6-1-1} = 853{,}78.$

Dále potřebujeme celkový součet čtverců

$S_T = (\mathbf{y} - \mathbf{m}_2)'(\mathbf{y} - \mathbf{m}_2)$,

kde $\mathbf{m}_2$ je sloupcový vektor typu nx1 složený z průměru m_2 závisle proměnné veličiny Y. V našem případě je $m_2 = 140$.

Příklad (8)

Po dosazení do vzorce pro celkový součet čtverců tedy dostaneme

$$S_T = (50-140, 60-140, 60-140, 120-140, 230-140, 320--140) \begin{pmatrix} 50-140 \\ 60-140 \\ 60-140 \\ 120-140 \\ 230-140 \\ 320-140 \end{pmatrix} = 61800.$$

(Celkový součet čtverců lze získat také tak, že výběrový rozptyl veličiny Y vynásobíme n-1: $S_T = 5.12360 = 61800$.) Regresní součet čtverců pak je:

$S_R = S_T - S_E = 61800 - 3451{,}11 = 58348{,}89.$

Index determinace: $ID^2 = \frac{S_R}{S_T} = \frac{58348{,}89}{61800} = 0{,}9442.$

Znamená to, že variabilita hodnot závisle proměnné veličiny je z 94,42% vysvětlena regresní přímkou.

(V případě regresní přímky platí $ID^2 = r_{12}^2$. V našem případě bylo zjištěno, že $r_{12} = 0{,}972$, tedy $ID^2 = 0{,}9447$.)

Příklad (9)

ad d) Vypočteme směrodatné chyby odhadů regresních parametrů b_0 a b_1 podle vzorce $s_{b_j} = s\sqrt{v_{jj}}$, j = 0, 1, kde v_{jj} je j-tý diagonální prvek matice $(\mathbf{X}'\mathbf{X})^{-1}$:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0{,}499084 & -0{,}003022 \\ -0{,}003022 & 0{,}000027 \end{pmatrix}$$

Přitom si uvědomíme, že $v_{00} = 0{,}499084$, $v_{11} = 0{,}000027$

$$s_{b_0} = s\sqrt{v_{00}} = \sqrt{853{,}78} \cdot \sqrt{0{,}499084} = 20{,}6424,$$

$$s_{b_1} = s\sqrt{v_{11}} = \sqrt{853{,}78} \cdot \sqrt{0{,}000027} = 0{,}1532.$$

Stanovíme meze 95% intervalů spolehlivosti pro regresní parametry β_0 a β_1. K tomu slouží vzorec $b_j \pm t_{1-\alpha/2}(n-p-1)s_{b_j}$, j = 0, 1.

Příklad (10)

95% interval spolehlivosti pro β_0:

$$d = b_0 - t_{0,975}(4)s_{b_0} = 0{,}6868 - 2{,}7764 \cdot 20{,}6424 = -56{,}63$$
$$h = b_0 + t_{0,975}(4)s_{b_0} = 0{,}6868 + 2{,}7764 \cdot 20{,}6424 = 58$$

Znamená to, že $-56{,}63 < \beta_0 < 58$ s pravděpodobností aspoň 0,95.

95% interval spolehlivosti pro β_1:

$$d = b_1 - t_{0,975}(4)s_{b_1} = 1{,}2665 - 2{,}7764 \cdot 0{,}1532 = 0{,}841$$
$$h = b_1 + t_{0,975}(4)s_{b_1} = 1{,}2665 + 2{,}7764 \cdot 0{,}1532 = 1{,}692$$

Znamená to, že $0{,}841 < \beta_1 < 1{,}692$ s pravděpodobností aspoň 0,95.

Příklad (11)

ad e) Provedení celkového F-testu: na hladině významnosti $\alpha = 0{,}05$ testujeme H_0: $\beta_1 = 0$ proti H_1: $\beta_1 \neq 0$.

Testová statistika $F = \frac{S_R/p}{S_E/(n-p-1)} = \frac{58348{,}89/1}{3415{,}11/(6-1-1)} = 68{,}384,$

kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty) = \langle F_{0{,}95}(1,4), \infty) = \langle 7{,}7086, \infty).$

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_1 (tj. směrnice regresní přímky) zamítáme na hladině významnosti 0,05. Výsledky testování významnosti modelu jako celku zapíšeme do tabulky ANOVA:

zdroj variab.	součet čtverců	stupně volnosti	podíl	statistika F
model	$S_R = 58348{,}89$	$p = 1$	$S_R/p = 58348{,}89$	68,384
reziduální	$S_E = 3415{,}11$	$n-p-1 = 4$	$S_E/(n-p-1) = 853{,}78$	-
celkový	$S_T = 61800$	$n-1 = 5$	-	-

Příklad (12)

ad f) Provedení dílčích t-testů:

Na hladině významnosti $\alpha = 0{,}05$ testujeme H_0: $\beta_0 = 0$ proti H_1: $\beta_0 \neq 0$.

Testová statistika: $t_0 = \frac{b_0}{s_{b_0}} = \frac{0{,}6868}{20{,}6424} = 0{,}3327,$

kritický obor:

$$W = \left(-\infty, -t_{1-\frac{\alpha}{2}}(n-p-1)\right\rangle \cup \left\langle t_{1-\frac{\alpha}{2}}(n-p-1), \infty\right) = \left(-\infty, -t_{0{,}975}(4)\right\rangle \cup$$
$$\left\langle t_{0{,}975}(4), \infty\right) = (-\infty, -2{,}7764\rangle \cup \langle 2{,}7764, \infty).$$

Protože se testová statistika nerealizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_0 (tj. posunutí regresní přímky) nezamítáme na hladině významnosti 0,05.

Ke stejnému výsledku dospějeme, podíváme-li se na 95% interval spolehlivosti pro β_0. Vypočítali jsme, že $-56{,}63 < \beta_0 < 58$ s pravděpodobností aspoň 0,95. Protože tento interval obsahuje 0, hypotézu H_0: $\beta_0 = 0$ nezamítáme na hladině významnosti 0,05.

Příklad (13)

Na hladině významnosti $\alpha = 0{,}05$ testujeme H_0: $\beta_1 = 0$ proti H_1: $\beta_1 \neq 0$.

Testová statistika: $t_1 = \frac{b_1}{s_{b_1}} = \frac{1{,}2665}{0{,}1532} = 8{,}27$,

kritický obor: $W = (-\infty, -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1), \infty) =$
$(-\infty, -t_{0{,}975}(4)\rangle \cup \langle t_{0{,}975}(4), \infty) = (-\infty, -2{,}7764\rangle \cup \langle 2{,}7764, \infty)$.

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_1 (tj. směrnice regresní přímky) zamítáme na hladině významnosti 0,05.

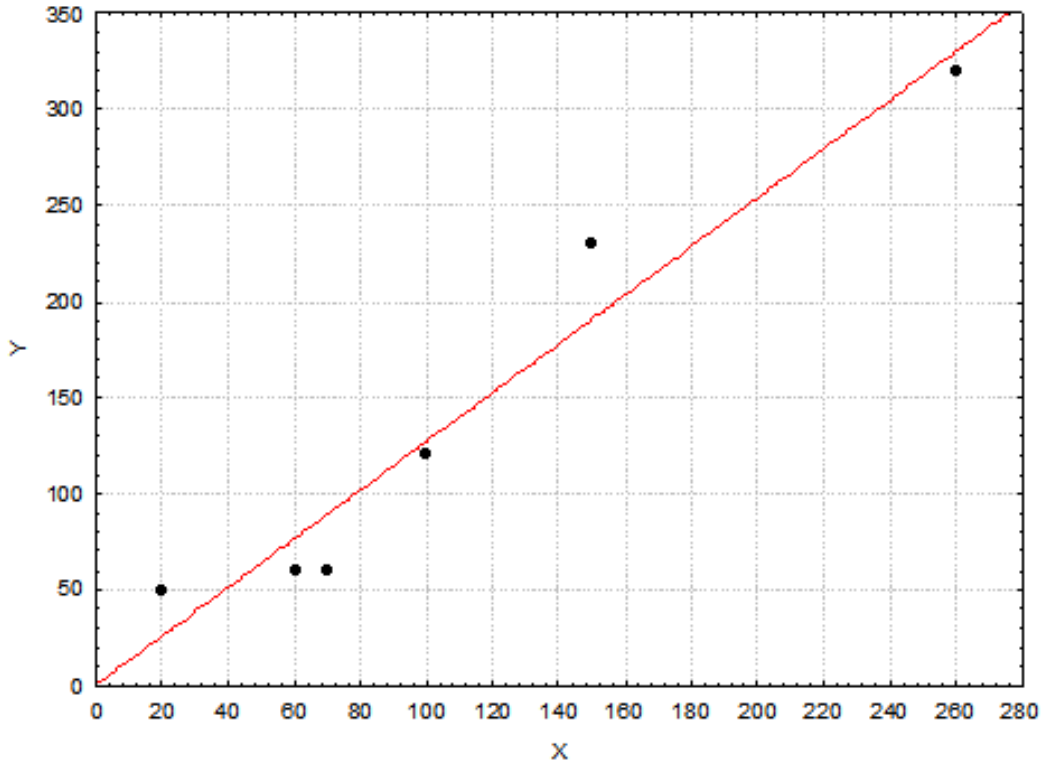
Ke stejnému výsledku dospějeme, podíváme-li se na 95% interval spolehlivosti pro β_1. Vypočítali jsme, že $0{,}841 < \beta_1 < 1{,}692$ s pravděpodobností aspoň 0,95. Protože tento interval neobsahuje 0, hypotézu H_0: $\beta_1 = 0$ zamítáme na hladině významnosti 0,05.

V případě modelu regresní přímky je dílčí t-test pro parametr β_1 ekvivalentní s celkovým F-testem.

Příklad (14)

ad g) Regresní odhad pro x = 110 dostaneme pouhým dosazením do rovnice regresní přímky: $\hat{y} = 0{,}6868 + 1{,}2665 \cdot 110 = 140$.

ad h)



Výpočet pomocí systému STATISTICA (1)

Vytvoříme nový datový soubor se dvěma proměnnými X a Y a 6 případy:

	1 X	2 Y
1	20	50
2	60	60
3	70	60
4	100	120
5	150	230
6	260	320

a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení. Vypočtěte výběrový koeficient korelace mezi X a Y, interpretujte jeho hodnotu a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny.

Zobrazíme dvourozměrný tečkový diagram s proloženou elipsou 95% konstantní hustoty pravděpodobnosti, s jehož pomocí posoudíme dvourozměrnou normalitu dat: Grafy – Bodové grafy – vypneme Typ proložení – Proměnné X, Y - OK.

Výpočet pomocí systému STATISTICA (2)

Na záložce Detaily vybereme Elipsa Normální – OK. Ve vzniklém dvourozměrném tečkovém diagramu změníme rozsah zobrazených hodnot na vodorovné a svislé ose, abychom viděli celou elipsu – viz obrázek výše.

Testování hypotézy o nezávislosti: Statistika – Základní statistiky /Tabulky - Korelační matice – OK – 2 seznamy proměnných X, Y, OK. Na záložce Možnosti zaškrtneme Zobrazit detailní tabulku výsledků – Souhrn.

Prom. X & prom. Y	Korelace (Tabulka1) 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	110,0000	85,3229									
Y	140,0000	111,1755	0,971977	0,944739	8,269474	0,001167	6	0,686813	1,266484	5,566343	0,745955

Ve výstupní tabulce najdeme hodnotu výběrového korelačního koeficientu R_{12} ($r = 0{,}971977$, tzn. že mezi X a Y existuje velmi silná přímá lineární závislost), realizaci testové statistiky $t = 8{,}269474$ a p-hodnotu pro test hypotézy o nezávislosti ($p = 0{,}001167$, H_0 tedy zamítáme na hladině významnosti 0,05).

Výpočet pomocí systému STATISTICA (3)

b) Předpokládejte, že závislost letošní poptávky na loňské lze vystihnout regresní přímkou. Vypočtěte odhady regresních parametrů a napište rovnici regresní přímky. Interpretujte parametry regresní přímky.

Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnná X – OK – OK – Výpočet: Výsledky regrese.

N=6	Výsledky regrese se závislou proměnnou : Y (Tabulka1) R= ,97197702 R2= ,94473932 Upravené R2= ,93092415 F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			0,686813	20,64236	0,033272	0,975052
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167

Ve výstupní tabulce najdeme koeficient b_0 ve sloupci B na řádku označeném Abs. člen, koeficient b_1 ve sloupci B na řádku označeném X. Rovnice regresní přímky:

$y = 0{,}686813 + 1{,}266484\ x.$

Znamená to, že při nulové loňské poptávce by letošní poptávka činila 0,6868 kusů a při zvýšení loňské poptávky o 10 kusů by se letošní poptávka zvedla o 12,665 kusů.

Výpočet pomocí systému STATISTICA (4)

c) Najděte odhad rozptylu, vypočtěte index determinace a interpretujte ho.

Vrátíme se do Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

Efekt	Analýza rozptylu (Tabulka1)				
	Součet čtverců	sv	Průměr čtverců	F	Úroveň p
Regres.	58384,89	1	58384,89	68,38420	0,001167
Rezid.	3415,11	4	853,78		
Celk.	61800,00				

Odhad rozptylu najdeme na řádku Rezid., ve sloupci Průměr čtverců, tedy $s^2 = 853{,}78$.

Index determinace je uveden v záhlaví původní výstupní tabulky pod označením R2. V našem případě $ID^2 = 0{,}9447$, tedy variabilita letošní poptávky je z 94,5 % vysvětlena regresní přímkou.

Výpočet pomocí systému STATISTICA (5)

d) Najděte 95% intervaly spolehlivosti pro regresní parametry.

Ve výstupní tabulce výsledků regrese přidáme za proměnnou Úroveň p dvě nové proměnné dm (pro dolní meze 95% intervalů spolehlivosti pro regresní parametry) a hm (pro horní meze 95% intervalů spolehlivosti pro regresní parametry). Do Dlouhého jména proměnné dm resp. hm napíšeme: =v3-v4*VStudent(0,975;4) resp. =v3+v4*VStudent(0,975;4)

N=6	Výsledky regrese se závislou proměnnou : Y (Tabulka1) R= ,97197702 R2= ,94473932 Upravené R2= ,93092415 F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219							
	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p	dm =v3-v4*V	hm =v3+v4*
Abs.člen			0,686813	20,64236	0,033272	0,975052	-56,6256	57,99918
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167	0,841266	1,691701

Vidíme, že -56,63 < β_0 < 58 s pravděpodobností aspoň 0,95 a 0,841< β_1 < 1,692 s pravděpodobností aspoň 0,95.

Výpočet pomocí systému STATISTICA (6)

e) Na hladině významnosti 0,05 proveďte celkový F-test.

Testovou statistiku F-testu a odpovídající p-hodnotu najdeme v záhlaví výstupní tabulky regrese. Zde F = 68,384, p-hodnota < 0,00117, tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti modelu jako celku. (Výsledky F-testu jsou rovněž uvedeny v tabulce ANOVA.)

f) Na hladině významnosti 0,05 proveďte dílčí t-testy.

Výsledky dílčích t-testů jsou uvedeny ve výstupní tabulce regrese. Testová statistika pro test hypotézy H_0: $\beta_0 = 0$ je 0,033272, p-hodnota je 0,975052. Hypotézu o nevýznamnosti úseku regresní přímky tedy nezamítáme na hladině významnosti 0,05. Testová statistika pro test hypotézy H_0: $\beta_1 = 0$ je 8,269474, p-hodnota je 0,001167. Hypotézu o nevýznamnosti směrnice regresní přímky tedy zamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA (7)

g) Vypočtěte regresní odhad letošní poptávky při loňské poptávce 110 kusů.

Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi Předpovědi závisle proměnné X: 110 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď.

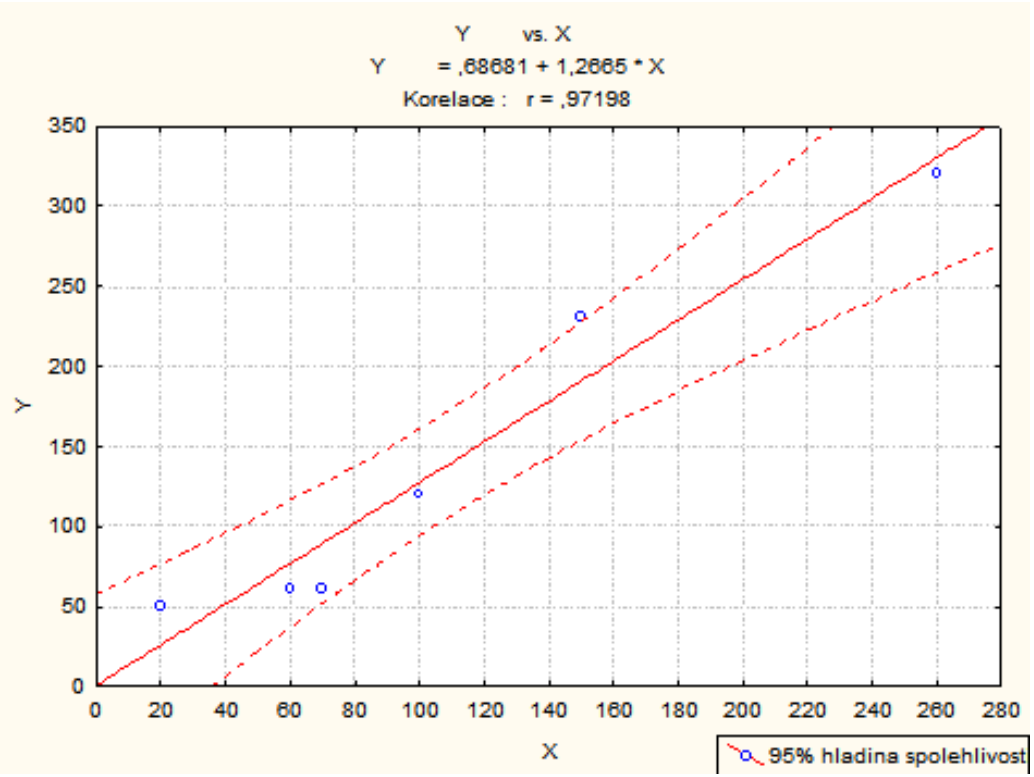
Proměnná	Předpovězené hodnoty (Tabulka1) proměnné: Y		
	B-váž.	Hodnota	B-váž. * Hodnot
X	1,266484	110,0000	139,3132
Abs. člen			0,6868
Předpověď			140,0000
-95,0%LS			106,8803
+95,0%LS			173,1197

Při loňské poptávce 110 kusů je predikovaná hodnota letošní poptávky 140 kusů.

Výpočet pomocí systému STATISTICA (8)

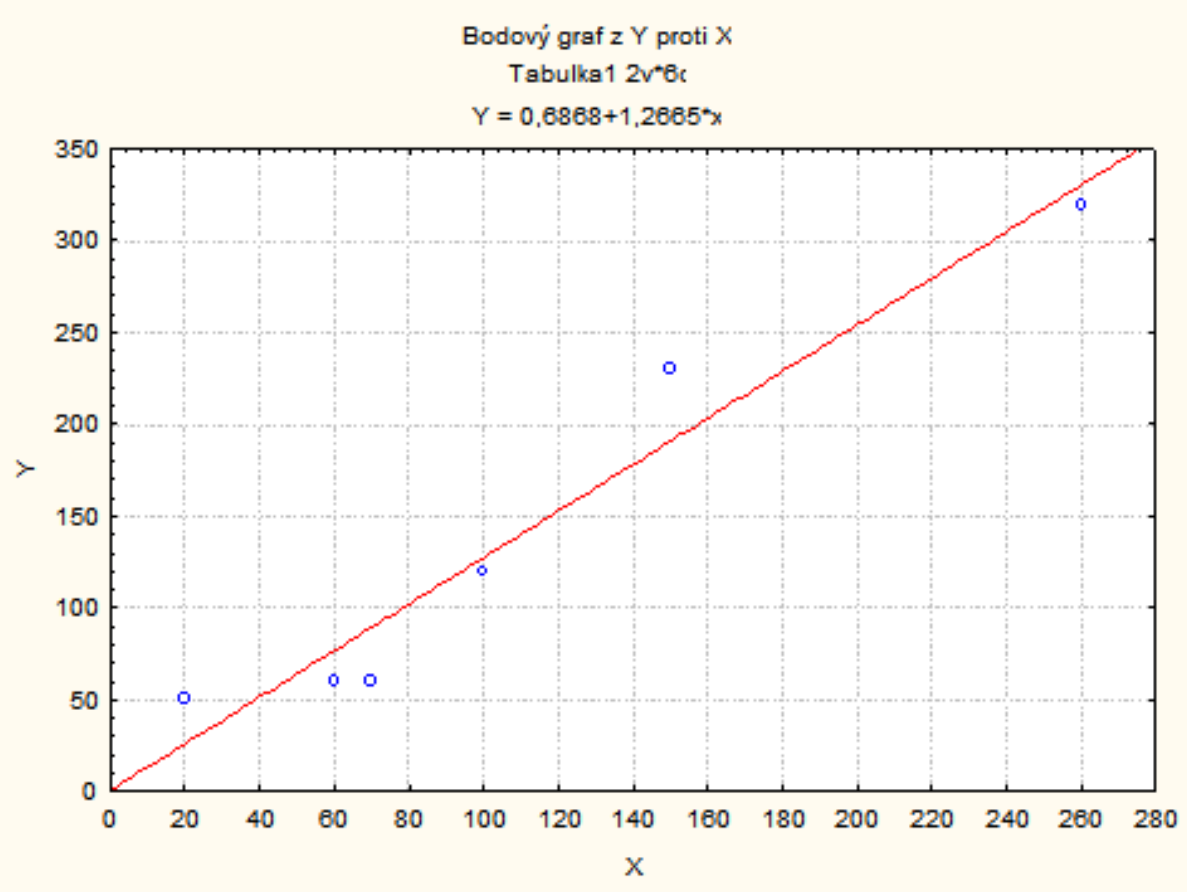
h) Nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou.

Nakreslení regresní přímky: Návrat do Výsledky: Vícenásobná regrese – Rezidua/předpoklady/předpovědi - Reziduální analýza – Bodové grafy – Korelace dvou proměnných – X, Y – OK.



Výpočet pomocí systému STATISTICA (9)

Jiný způsob: Do dvourozměrného tečkového diagramu nakreslíme regresní přímku tak, že v tabulce 2D Bodové grafy zvolíme Typ proložení: Lineární, OK.



13. Statistické tabulky

The Studentized range upper quantiles (0,1)

The Studentized range upper quantiles (0,05)

The Studentized range upper quantiles (0,01)

Kritické hodnoty $D_n(\alpha)$ Kolmogorovova-Smirnovova testu

Modifikované kritické hodnoty $D_n^*(\alpha)$ Kolmogorovova-Smirnovova testu

Koeficienty $a_i^{(n)}$ pro Shapiro – Wilkův test

Kritické hodnoty pro Shapiro – Wilkův test

Kritické hodnoty znaménkového testu

Kritické hodnoty jednovýběrového Wilcoxonova testu

Kritické hodnoty dvouvýběrového Wilcoxonova testu

Kritické hodnoty Neményiho metody

Kritické hodnoty pro Spearmanův koeficient pořadové korelace

The Studentized range upper quantiles q(k, df; 0.10)

df	k-> 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	8.929	13.437	16.358	18.488	20.150	21.504	22.642	23.621	24.477	25.237	25.918	26.536	27.100	27.618	28.097	28.542	28.958	29.347	29.713
2	4.129	5.733	6.772	7.538	8.139	8.633	9.049	9.409	9.725	10.006	10.259	10.488	10.698	10.891	11.070	11.237	11.392	11.538	11.676
3	3.328	4.467	5.199	5.738	6.162	6.511	6.806	7.062	7.287	7.487	7.667	7.831	7.982	8.120	8.248	8.368	8.479	8.584	8.683
4	3.015	3.976	4.586	5.035	5.388	5.679	5.926	6.139	6.327	6.494	6.645	6.783	6.909	7.025	7.132	7.233	7.326	7.414	7.497
5	2.850	3.717	4.264	4.664	4.979	5.238	5.458	5.648	5.816	5.965	6.100	6.223	6.336	6.439	6.536	6.626	6.710	6.788	6.863
6	2.748	3.558	4.065	4.435	4.726	4.966	5.168	5.344	5.499	5.637	5.762	5.875	5.979	6.075	6.164	6.247	6.325	6.398	6.466
7	2.679	3.451	3.931	4.280	4.555	4.780	4.971	5.137	5.283	5.413	5.530	5.637	5.735	5.826	5.910	5.988	6.061	6.130	6.195
8	2.630	3.374	3.834	4.169	4.431	4.646	4.829	4.987	5.126	5.250	5.362	5.464	5.558	5.644	5.724	5.799	5.869	5.935	5.997
9	2.592	3.316	3.761	4.084	4.337	4.545	4.721	4.873	5.007	5.126	5.234	5.333	5.423	5.506	5.583	5.655	5.722	5.786	5.845
10	2.563	3.270	3.704	4.018	4.264	4.465	4.636	4.783	4.913	5.029	5.134	5.229	5.316	5.397	5.472	5.542	5.607	5.668	5.726
11	2.540	3.234	3.658	3.965	4.205	4.401	4.567	4.711	4.838	4.951	5.053	5.145	5.231	5.309	5.382	5.450	5.514	5.573	5.630
12	2.521	3.204	3.621	3.921	4.156	4.349	4.511	4.652	4.776	4.886	4.986	5.076	5.160	5.236	5.308	5.374	5.436	5.495	5.550
13	2.504	3.179	3.589	3.885	4.116	4.304	4.464	4.602	4.724	4.832	4.930	5.019	5.100	5.175	5.245	5.310	5.371	5.429	5.483
14	2.491	3.158	3.563	3.854	4.081	4.267	4.424	4.560	4.679	4.786	4.882	4.969	5.050	5.124	5.192	5.256	5.316	5.372	5.426
15	2.479	3.140	3.540	3.828	4.052	4.235	4.390	4.524	4.641	4.746	4.841	4.927	5.006	5.079	5.146	5.209	5.268	5.324	5.376
16	2.469	3.124	3.520	3.804	4.026	4.207	4.360	4.492	4.608	4.712	4.805	4.890	4.968	5.040	5.106	5.169	5.227	5.282	5.333
17	2.460	3.110	3.503	3.784	4.003	4.182	4.334	4.464	4.579	4.681	4.774	4.857	4.934	5.005	5.071	5.133	5.190	5.244	5.295
18	2.452	3.098	3.487	3.766	3.984	4.161	4.310	4.440	4.553	4.654	4.746	4.829	4.905	4.975	5.040	5.101	5.158	5.211	5.262
19	2.445	3.087	3.474	3.751	3.966	4.142	4.290	4.418	4.530	4.630	4.721	4.803	4.878	4.948	5.012	5.072	5.129	5.182	5.232
20	2.439	3.077	3.462	3.736	3.950	4.124	4.271	4.398	4.510	4.609	4.699	4.780	4.855	4.923	4.987	5.047	5.103	5.155	5.205
df	k-> 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	2.433	3.069	3.451	3.724	3.936	4.109	4.255	4.380	4.491	4.590	4.678	4.759	4.833	4.901	4.965	5.024	5.079	5.131	5.180
22	2.428	3.061	3.441	3.712	3.923	4.095	4.239	4.364	4.474	4.572	4.660	4.740	4.814	4.882	4.944	5.003	5.058	5.109	5.158
23	2.424	3.054	3.432	3.701	3.911	4.082	4.226	4.350	4.459	4.556	4.644	4.723	4.796	4.863	4.926	4.984	5.038	5.089	5.138
24	2.420	3.047	3.423	3.692	3.900	4.070	4.213	4.336	4.445	4.541	4.628	4.707	4.780	4.847	4.909	4.966	5.020	5.071	5.119
25	2.416	3.041	3.416	3.683	3.890	4.059	4.201	4.324	4.432	4.528	4.614	4.693	4.765	4.831	4.893	4.950	5.004	5.055	5.102
26	2.412	3.036	3.409	3.675	3.881	4.049	4.191	4.313	4.420	4.515	4.601	4.680	4.751	4.817	4.878	4.936	4.989	5.039	5.086
27	2.409	3.030	3.402	3.667	3.873	4.040	4.181	4.302	4.409	4.504	4.590	4.667	4.739	4.804	4.865	4.922	4.975	5.025	5.072
28	2.406	3.026	3.396	3.660	3.865	4.032	4.172	4.293	4.399	4.493	4.579	4.656	4.727	4.792	4.853	4.909	4.962	5.012	5.058
29	2.403	3.021	3.391	3.654	3.858	4.024	4.163	4.284	4.389	4.484	4.568	4.645	4.716	4.781	4.841	4.897	4.950	4.999	5.046
30	2.400	3.017	3.386	3.648	3.851	4.016	4.155	4.275	4.381	4.474	4.559	4.635	4.706	4.770	4.830	4.886	4.939	4.988	5.034
31	2.398	3.013	3.381	3.642	3.845	4.009	4.148	4.268	4.372	4.466	4.550	4.626	4.696	4.760	4.820	4.876	4.928	4.977	5.023
32	2.396	3.010	3.376	3.637	3.839	4.003	4.141	4.260	4.365	4.458	4.541	4.617	4.687	4.751	4.811	4.866	4.918	4.967	5.013
33	2.393	3.006	3.372	3.632	3.833	3.997	4.135	4.253	4.357	4.450	4.533	4.609	4.679	4.743	4.802	4.857	4.909	4.957	5.003
34	2.391	3.003	3.368	3.627	3.828	3.991	4.129	4.247	4.351	4.443	4.526	4.602	4.671	4.734	4.794	4.849	4.900	4.949	4.994
35	2.389	3.000	3.364	3.623	3.823	3.986	4.123	4.241	4.344	4.436	4.519	4.594	4.663	4.727	4.786	4.841	4.892	4.940	4.986
36	2.388	2.998	3.361	3.619	3.819	3.981	4.117	4.235	4.338	4.430	4.512	4.588	4.656	4.720	4.778	4.833	4.884	4.932	4.978
37	2.386	2.995	3.357	3.615	3.814	3.976	4.112	4.230	4.332	4.424	4.506	4.581	4.650	4.713	4.771	4.826	4.877	4.925	4.970
38	2.384	2.992	3.354	3.611	3.810	3.972	4.107	4.224	4.327	4.418	4.500	4.575	4.643	4.706	4.765	4.819	4.870	4.918	4.963
39	2.383	2.990	3.351	3.608	3.806	3.967	4.103	4.220	4.322	4.413	4.495	4.569	4.637	4.700	4.758	4.812	4.863	4.911	4.956
40	2.381	2.988	3.348	3.605	3.802	3.963	4.099	4.215	4.317	4.408	4.490	4.564	4.632	4.694	4.752	4.806	4.857	4.904	4.949
48	2.372	2.973	3.330	3.583	3.778	3.937	4.070	4.185	4.285	4.375	4.455	4.528	4.595	4.656	4.713	4.766	4.816	4.863	4.907
60	2.363	2.959	3.312	3.562	3.755	3.911	4.042	4.155	4.254	4.342	4.421	4.493	4.558	4.619	4.675	4.727	4.775	4.821	4.864
80	2.353	2.945	3.294	3.541	3.731	3.885	4.014	4.125	4.223	4.309	4.387	4.457	4.521	4.581	4.636	4.687	4.735	4.780	4.822
120	2.344	2.930	3.276	3.520	3.707	3.859	3.986	4.096	4.191	4.276	4.353	4.422	4.485	4.543	4.597	4.647	4.694	4.738	4.779
240	2.335	2.916	3.258	3.499	3.684	3.834	3.959	4.066	4.160	4.244	4.319	4.386	4.448	4.505	4.558	4.607	4.653	4.696	4.737
Inf	2.326	2.902	3.240	3.478	3.661	3.808	3.931	4.037	4.129	4.211	4.285	4.351	4.412	4.468	4.519	4.568	4.612	4.654	4.694

Pro Tukeyovu metodu (přednáška č.9) odpovídá značení: $q_{1-\alpha}(r, n-r) = q(k, df; \alpha)$

The Studentized range upper quantiles q(k, df; 0.05)

df	k-> 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	17.969	26.976	32.819	37.082	40.408	43.119	45.397	47.357	49.071	50.592	51.957	53.194	54.323	55.361	56.320	57.212	58.044	58.824	59.558
2	6.085	8.331	9.798	10.881	11.734	12.435	13.027	13.539	13.988	14.389	14.749	15.076	15.375	15.650	15.905	16.143	16.365	16.573	16.769
3	4.501	5.910	6.825	7.502	8.037	8.478	8.852	9.177	9.462	9.717	9.946	10.155	10.346	10.522	10.686	10.838	10.980	11.114	11.240
4	3.926	5.040	5.757	6.287	6.706	7.053	7.347	7.602	7.826	8.027	8.208	8.373	8.524	8.664	8.793	8.914	9.027	9.133	9.233
5	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.801	6.995	7.167	7.323	7.466	7.596	7.716	7.828	7.932	8.030	8.122	8.208
6	3.460	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493	6.649	6.789	6.917	7.034	7.143	7.244	7.338	7.426	7.508	7.586
7	3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.997	6.158	6.302	6.431	6.550	6.658	6.759	6.852	6.939	7.020	7.097	7.169
8	3.261	4.041	4.529	4.886	5.167	5.399	5.596	5.767	5.918	6.053	6.175	6.287	6.389	6.483	6.571	6.653	6.729	6.801	6.869
9	3.199	3.948	4.415	4.755	5.024	5.244	5.432	5.595	5.738	5.867	5.983	6.089	6.186	6.276	6.359	6.437	6.510	6.579	6.643
10	3.151	3.877	4.327	4.654	4.912	5.124	5.304	5.460	5.598	5.722	5.833	5.935	6.028	6.114	6.194	6.269	6.339	6.405	6.467
11	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.486	5.605	5.713	5.811	5.901	5.984	6.062	6.134	6.202	6.265	6.325
12	3.081	3.773	4.199	4.508	4.750	4.950	5.119	5.265	5.395	5.510	5.615	5.710	5.797	5.878	5.953	6.023	6.089	6.151	6.209
13	3.055	3.734	4.151	4.453	4.690	4.884	5.049	5.192	5.318	5.431	5.533	5.625	5.711	5.789	5.862	5.931	5.995	6.055	6.112
14	3.033	3.701	4.111	4.407	4.639	4.829	4.990	5.130	5.253	5.364	5.463	5.554	5.637	5.714	5.785	5.852	5.915	5.973	6.029
15	3.014	3.673	4.076	4.367	4.595	4.782	4.940	5.077	5.198	5.306	5.403	5.492	5.574	5.649	5.719	5.785	5.846	5.904	5.958
16	2.998	3.649	4.046	4.333	4.557	4.741	4.896	5.031	5.150	5.256	5.352	5.439	5.519	5.593	5.662	5.726	5.786	5.843	5.896
17	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108	5.212	5.306	5.392	5.471	5.544	5.612	5.675	5.734	5.790	5.842
18	2.971	3.609	3.997	4.276	4.494	4.673	4.824	4.955	5.071	5.173	5.266	5.351	5.429	5.501	5.567	5.629	5.688	5.743	5.794
19	2.960	3.593	3.977	4.253	4.468	4.645	4.794	4.924	5.037	5.139	5.231	5.314	5.391	5.462	5.528	5.589	5.647	5.701	5.752
20	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.895	5.008	5.108	5.199	5.282	5.357	5.427	5.492	5.553	5.610	5.663	5.714
df	k-> 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	2.941	3.565	3.942	4.213	4.424	4.597	4.743	4.870	4.981	5.081	5.170	5.252	5.327	5.396	5.460	5.520	5.576	5.629	5.679
22	2.933	3.553	3.927	4.196	4.405	4.577	4.722	4.847	4.957	5.056	5.144	5.225	5.299	5.368	5.431	5.491	5.546	5.599	5.648
23	2.926	3.542	3.914	4.180	4.388	4.558	4.702	4.826	4.935	5.033	5.121	5.201	5.274	5.342	5.405	5.464	5.519	5.571	5.620
24	2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915	5.012	5.099	5.179	5.251	5.319	5.381	5.439	5.494	5.545	5.594
25	2.913	3.523	3.890	4.153	4.358	4.526	4.667	4.789	4.897	4.993	5.079	5.158	5.230	5.297	5.359	5.417	5.471	5.522	5.570
26	2.907	3.514	3.880	4.141	4.345	4.511	4.652	4.773	4.880	4.975	5.061	5.139	5.211	5.277	5.339	5.396	5.450	5.500	5.548
27	2.902	3.506	3.870	4.130	4.333	4.498	4.638	4.758	4.864	4.959	5.044	5.122	5.193	5.259	5.320	5.377	5.430	5.480	5.528
28	2.897	3.499	3.861	4.120	4.322	4.486	4.625	4.745	4.850	4.944	5.029	5.106	5.177	5.242	5.302	5.359	5.412	5.462	5.509
29	2.892	3.493	3.853	4.111	4.311	4.475	4.613	4.732	4.837	4.930	5.014	5.091	5.161	5.226	5.286	5.342	5.395	5.445	5.491
30	2.888	3.486	3.845	4.102	4.301	4.464	4.601	4.720	4.824	4.917	5.001	5.077	5.147	5.211	5.271	5.327	5.379	5.429	5.475
31	2.884	3.481	3.838	4.094	4.292	4.454	4.591	4.709	4.812	4.905	4.988	5.064	5.134	5.198	5.257	5.313	5.365	5.414	5.460
32	2.881	3.475	3.832	4.086	4.284	4.445	4.581	4.698	4.802	4.894	4.976	5.052	5.121	5.185	5.244	5.299	5.351	5.400	5.445
33	2.877	3.470	3.825	4.079	4.276	4.436	4.572	4.689	4.791	4.883	4.965	5.040	5.109	5.173	5.232	5.287	5.338	5.386	5.432
34	2.874	3.465	3.820	4.072	4.268	4.428	4.563	4.680	4.782	4.873	4.955	5.030	5.098	5.161	5.220	5.275	5.326	5.374	5.420
35	2.871	3.461	3.814	4.066	4.261	4.421	4.555	4.671	4.773	4.863	4.945	5.020	5.088	5.151	5.209	5.264	5.315	5.362	5.408
36	2.868	3.457	3.809	4.060	4.255	4.414	4.547	4.663	4.764	4.855	4.936	5.010	5.078	5.141	5.199	5.253	5.304	5.352	5.397
37	2.865	3.453	3.804	4.054	4.249	4.407	4.540	4.655	4.756	4.846	4.927	5.001	5.069	5.131	5.189	5.243	5.294	5.341	5.386
38	2.863	3.449	3.799	4.049	4.243	4.400	4.533	4.648	4.749	4.838	4.919	4.993	5.060	5.122	5.180	5.234	5.284	5.331	5.376
39	2.861	3.445	3.795	4.044	4.237	4.394	4.527	4.641	4.741	4.831	4.911	4.985	5.052	5.114	5.171	5.225	5.275	5.322	5.367
40	2.858	3.442	3.791	4.039	4.232	4.388	4.521	4.634	4.735	4.824	4.904	4.977	5.044	5.106	5.163	5.216	5.266	5.313	5.358
48	2.843	3.420	3.764	4.008	4.197	4.351	4.481	4.592	4.690	4.777	4.856	4.927	4.993	5.053	5.109	5.161	5.210	5.256	5.299
60	2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646	4.732	4.808	4.878	4.942	5.001	5.056	5.107	5.154	5.199	5.241
80	2.814	3.377	3.711	3.947	4.129	4.277	4.402	4.509	4.603	4.686	4.761	4.829	4.892	4.949	5.003	5.052	5.099	5.142	5.183
120	2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560	4.641	4.714	4.781	4.842	4.898	4.950	4.998	5.043	5.086	5.126
240	2.786	3.335	3.659	3.887	4.063	4.205	4.324	4.427	4.517	4.596	4.668	4.733	4.792	4.847	4.897	4.944	4.988	5.030	5.069
Inf	2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474	4.552	4.622	4.685	4.743	4.796	4.845	4.891	4.934	4.974	5.012

Pro Tukeyovu metodu (přednáška č.9) odpovídá značení: $q_{1-\alpha}(r, n-r) = q(k, df; \alpha)$

The Studentized range upper quantiles q(k, df; 0.01)

df	k-> 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	90.024	135.04	164.25	185.57	202.21	215.77	227.17	236.97	245.54	253.15	259.98	266.16	271.81	277.00	281.80	286.26	290.43	294.33	297.99
2	14.036	19.019	22.294	24.717	26.629	28.201	29.530	30.679	31.689	32.589	33.398	34.134	34.806	35.426	36.000	36.534	37.034	37.502	37.943
3	8.260	10.619	12.170	13.324	14.241	14.998	15.641	16.199	16.691	17.130	17.526	17.887	18.217	18.522	18.805	19.068	19.315	19.546	19.765
4	6.511	8.120	9.173	9.958	10.583	11.101	11.542	11.925	12.264	12.567	12.840	13.090	13.318	13.530	13.726	13.909	14.081	14.242	14.394
5	5.702	6.976	7.804	8.421	8.913	9.321	9.669	9.971	10.239	10.479	10.696	10.894	11.076	11.244	11.400	11.545	11.682	11.811	11.932
6	5.243	6.331	7.033	7.556	7.972	8.318	8.612	8.869	9.097	9.300	9.485	9.653	9.808	9.951	10.084	10.208	10.325	10.434	10.538
7	4.949	5.919	6.542	7.005	7.373	7.678	7.939	8.166	8.367	8.548	8.711	8.860	8.997	9.124	9.242	9.353	9.456	9.553	9.645
8	4.745	5.635	6.204	6.625	6.959	7.237	7.474	7.680	7.863	8.027	8.176	8.311	8.436	8.552	8.659	8.760	8.854	8.943	9.027
9	4.596	5.428	5.957	6.347	6.657	6.915	7.134	7.325	7.494	7.646	7.784	7.910	8.025	8.132	8.232	8.325	8.412	8.495	8.573
10	4.482	5.270	5.769	6.136	6.428	6.669	6.875	7.054	7.213	7.356	7.485	7.603	7.712	7.812	7.906	7.993	8.075	8.153	8.226
11	4.392	5.146	5.621	5.970	6.247	6.476	6.671	6.841	6.992	7.127	7.250	7.362	7.464	7.560	7.648	7.731	7.809	7.883	7.952
12	4.320	5.046	5.502	5.836	6.101	6.320	6.507	6.670	6.814	6.943	7.060	7.166	7.265	7.356	7.441	7.520	7.594	7.664	7.730
13	4.260	4.964	5.404	5.726	5.981	6.192	6.372	6.528	6.666	6.791	6.903	7.006	7.100	7.188	7.269	7.345	7.417	7.484	7.548
14	4.210	4.895	5.322	5.634	5.881	6.085	6.258	6.409	6.543	6.663	6.772	6.871	6.962	7.047	7.125	7.199	7.268	7.333	7.394
15	4.167	4.836	5.252	5.556	5.796	5.994	6.162	6.309	6.438	6.555	6.660	6.756	6.845	6.927	7.003	7.074	7.141	7.204	7.264
16	4.131	4.786	5.192	5.489	5.722	5.915	6.079	6.222	6.348	6.461	6.564	6.658	6.744	6.823	6.897	6.967	7.032	7.093	7.151
17	4.099	4.742	5.140	5.430	5.659	5.847	6.007	6.147	6.270	6.380	6.480	6.572	6.656	6.733	6.806	6.873	6.937	6.997	7.053
18	4.071	4.703	5.094	5.379	5.603	5.787	5.944	6.081	6.201	6.309	6.407	6.496	6.579	6.655	6.725	6.791	6.854	6.912	6.967
19	4.046	4.669	5.054	5.334	5.553	5.735	5.889	6.022	6.141	6.246	6.342	6.430	6.510	6.585	6.654	6.719	6.780	6.837	6.891
20	4.024	4.639	5.018	5.293	5.510	5.688	5.839	5.970	6.086	6.190	6.285	6.370	6.449	6.523	6.591	6.654	6.714	6.770	6.823
df	k-> 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	4.004	4.612	4.986	5.257	5.470	5.646	5.794	5.924	6.038	6.140	6.233	6.317	6.395	6.467	6.534	6.596	6.655	6.710	6.762
22	3.986	4.588	4.957	5.225	5.435	5.608	5.754	5.882	5.994	6.095	6.186	6.269	6.346	6.417	6.482	6.544	6.602	6.656	6.707
23	3.970	4.566	4.931	5.195	5.403	5.573	5.718	5.844	5.955	6.054	6.144	6.226	6.301	6.371	6.436	6.497	6.553	6.607	6.658
24	3.955	4.546	4.907	5.168	5.373	5.542	5.685	5.809	5.919	6.017	6.105	6.186	6.261	6.330	6.394	6.453	6.510	6.562	6.612
25	3.942	4.527	4.885	5.144	5.347	5.513	5.655	5.778	5.886	5.983	6.070	6.150	6.224	6.292	6.355	6.414	6.469	6.522	6.571
26	3.930	4.510	4.865	5.121	5.322	5.487	5.627	5.749	5.856	5.951	6.038	6.117	6.190	6.257	6.319	6.378	6.432	6.484	6.533
27	3.918	4.495	4.847	5.101	5.300	5.463	5.602	5.722	5.828	5.923	6.008	6.087	6.158	6.225	6.287	6.344	6.399	6.450	6.498
28	3.908	4.481	4.830	5.082	5.279	5.441	5.578	5.697	5.802	5.896	5.981	6.058	6.129	6.195	6.256	6.314	6.367	6.418	6.465
29	3.898	4.467	4.814	5.064	5.260	5.420	5.556	5.674	5.778	5.871	5.955	6.032	6.103	6.168	6.228	6.285	6.338	6.388	6.435
30	3.889	4.455	4.799	5.048	5.242	5.401	5.536	5.653	5.756	5.848	5.932	6.008	6.078	6.142	6.202	6.258	6.311	6.361	6.407
31	3.881	4.443	4.786	5.032	5.225	5.383	5.517	5.633	5.736	5.827	5.910	5.985	6.055	6.119	6.178	6.234	6.286	6.335	6.381
32	3.873	4.433	4.773	5.018	5.210	5.367	5.500	5.615	5.716	5.807	5.889	5.964	6.033	6.096	6.155	6.211	6.262	6.311	6.357
33	3.865	4.423	4.761	5.005	5.195	5.351	5.483	5.598	5.698	5.789	5.870	5.944	6.013	6.076	6.134	6.189	6.240	6.289	6.334
34	3.859	4.413	4.750	4.992	5.181	5.336	5.468	5.581	5.682	5.771	5.852	5.926	5.994	6.056	6.114	6.169	6.220	6.268	6.313
35	3.852	4.404	4.739	4.980	5.169	5.323	5.453	5.566	5.666	5.755	5.835	5.908	5.976	6.038	6.096	6.150	6.200	6.248	6.293
36	3.846	4.396	4.729	4.969	5.156	5.310	5.439	5.552	5.651	5.739	5.819	5.892	5.959	6.021	6.078	6.132	6.182	6.229	6.274
37	3.840	4.388	4.720	4.959	5.145	5.298	5.427	5.538	5.637	5.725	5.804	5.876	5.943	6.004	6.061	6.115	6.165	6.212	6.256
38	3.835	4.381	4.711	4.949	5.134	5.286	5.414	5.526	5.623	5.711	5.790	5.862	5.928	5.989	6.046	6.099	6.148	6.195	6.239
39	3.830	4.374	4.703	4.940	5.124	5.275	5.403	5.513	5.611	5.698	5.776	5.848	5.914	5.974	6.031	6.084	6.133	6.179	6.223
40	3.825	4.367	4.695	4.931	5.114	5.265	5.392	5.502	5.599	5.685	5.764	5.835	5.900	5.961	6.017	6.069	6.118	6.165	6.208
48	3.793	4.324	4.644	4.874	5.052	5.198	5.322	5.428	5.522	5.606	5.681	5.750	5.814	5.872	5.926	5.977	6.024	6.069	6.111
60	3.762	4.282	4.594	4.818	4.991	5.133	5.253	5.356	5.447	5.528	5.601	5.667	5.728	5.784	5.837	5.886	5.931	5.974	6.015
80	3.732	4.241	4.545	4.763	4.931	5.069	5.185	5.284	5.372	5.451	5.521	5.585	5.644	5.698	5.749	5.796	5.840	5.881	5.920
120	3.702	4.200	4.497	4.709	4.872	5.005	5.118	5.214	5.299	5.375	5.443	5.505	5.561	5.614	5.662	5.708	5.750	5.790	5.827
240	3.672	4.160	4.450	4.655	4.814	4.943	5.052	5.145	5.227	5.300	5.366	5.426	5.480	5.530	5.577	5.621	5.661	5.699	5.735
Inf	3.643	4.120	4.403	4.603	4.757	4.882	4.987	5.078	5.157	5.227	5.290	5.348	5.400	5.448	5.493	5.535	5.574	5.611	5.645

Pro Tukeyovu metodu (přednáška č.9) odpovídá značení: $q_{1-\alpha}(r, n-r) = q(k, df; \alpha)$

Kritické hodnoty Dn(α) Kolmogorovova-Smirnovova testu n= 4,…,40, α=0,01, α=0,05, α=0,10, α=0,15 a α=0,20

	alfa				
n	0,20	0,15	0,10	0,05	0,01
4	0,4927	0,5221	0,5652	0,6239	0,7342
5	0,4470	0,4754	0,5095	0,5633	0,6685
6	0,4104	0,4334	0,4680	0,5193	0,6166
7	0,3815	0,4043	0,4361	0,4834	0,5758
8	0,3583	0,3801	0,4096	0,4543	0,5418
9	0,3391	0,3591	0,3875	0,4300	0,5133
10	0,3226	0,3416	0,3687	0,4093	0,4889
11	0,3083	0,3266	0,3524	0,3912	0,4677
12	0,2958	0,3134	0,3382	0,3754	0,4491
13	0,2847	0,3016	0,3255	0,3614	0,4325
14	0,2748	0,2911	0,3142	0,3489	0,4176
15	0,2659	0,2816	0,3040	0,3376	0,4042
16	0,2578	0,2731	0,2947	0,3273	0,3920
17	0,2504	0,2652	0,2863	0,3180	0,3809
18	0,2436	0,2580	0,2785	0,3094	0,3706
19	0,2374	0,2514	0,2714	0,3014	0,3612
20	0,2316	0,2452	0,2647	0,2941	0,3524

	alfa				
n	0,20	0,15	0,10	0,05	0,01
21	0,2263	0,2403	0,2587	0,2873	0,3443
22	0,2213	0,2350	0,2529	0,2809	0,3367
23	0,2166	0,2300	0,2475	0,2749	0,3296
24	0,2122	0,2253	0,2425	0,2693	0,3229
25	0,2080	0,2209	0,2377	0,2641	0,3166
26	0,2041	0,2167	0,2333	0,2591	0,3106
27	0,2004	0,2128	0,2290	0,2544	0,3050
28	0,1969	0,2090	0,2250	0,2500	0,2997
29	0,1936	0,2055	0,2212	0,2457	0,2947
30	0,1904	0,2022	0,2176	0,2417	0,2899
31	0,1874	0,1990	0,2142	0,2379	0,2853
32	0,1845	0,1959	0,2109	0,2343	0,2809
33	0,1818	0,1930	0,2078	0,2308	0,2768
34	0,1792	0,1902	0,2048	0,2275	0,2728
35	0,1767	0,1875	0,2019	0,2243	0,2690
36	0,1743	0,1850	0,1991	0,2212	0,2653
37	0,1719	0,1825	0,1965	0,2183	0,2618
38	0,1697	0,1802	0,1940	0,2155	0,2584
39	0,1676	0,1779	0,1915	0,2127	0,2552
40	0,1655	0,1757	0,1892	0,2101	0,2521

Pro n>40 lze Dn(α) aproximovat pomocí $\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}$

Zdroj: kstest v MATLABu.

Modifikované kritické hodnoty Dn*(α) Kolmogorovova-Smirnovova testu

n= 4,…,40, α=0,01, α=0,05, α=0,10, α=0,15 a α=0,20.

	alfa				
n	0,20	0,15	0,10	0,05	0,01
4	0,3028	0,3213	0,3453	0,3754	0,4131
5	0,2893	0,3026	0,3189	0,3431	0,3966
6	0,2688	0,2810	0,2973	0,3236	0,3703
7	0,2523	0,2643	0,2802	0,3041	0,3506
8	0,2387	0,2502	0,2651	0,2880	0,3326
9	0,2271	0,2379	0,2520	0,2740	0,3171
10	0,2171	0,2274	0,2410	0,2620	0,3034
11	0,2082	0,2181	0,2312	0,2515	0,2915
12	0,2002	0,2098	0,2224	0,2418	0,2808
13	0,1932	0,2025	0,2145	0,2333	0,2706
14	0,1868	0,1958	0,2075	0,2257	0,2619
15	0,1811	0,1898	0,2012	0,2189	0,2539
16	0,1759	0,1843	0,1953	0,2126	0,2472
17	0,1711	0,1792	0,1900	0,2068	0,2403
18	0,1666	0,1746	0,1850	0,2013	0,2341
19	0,1625	0,1703	0,1806	0,1965	0,2285
20	0,1587	0,1663	0,1763	0,1920	0,2232

	alfa				
n	0,20	0,15	0,10	0,05	0,01
21	0,1551	0,1626	0,1723	0,1877	0,2183
22	0,1518	0,1591	0,1687	0,1837	0,2137
23	0,1487	0,1558	0,1652	0,1799	0,2093
24	0,1458	0,1528	0,1619	0,1764	0,2052
25	0,1430	0,1499	0,1589	0,1730	0,2014
26	0,1404	0,1471	0,1560	0,1699	0,1977
27	0,1379	0,1445	0,1532	0,1669	0,1943
28	0,1356	0,1421	0,1506	0,1641	0,1910
29	0,1334	0,1398	0,1482	0,1614	0,1879
30	0,1312	0,1375	0,1458	0,1588	0,1849
31	0,1292	0,1354	0,1436	0,1564	0,1821
32	0,1273	0,1334	0,1414	0,1541	0,1794
33	0,1255	0,1315	0,1394	0,1518	0,1768
34	0,1237	0,1296	0,1374	0,1497	0,1743
35	0,1220	0,1279	0,1356	0,1476	0,1720
36	0,1204	0,1262	0,1338	0,1457	0,1697
37	0,1188	0,1245	0,1320	0,1438	0,1675
38	0,1173	0,1230	0,1304	0,1420	0,1654
39	0,1159	0,1214	0,1288	0,1402	0,1634
40	0,1145	0,1200	0,1272	0,1385	0,1614
>40	$\frac{0,741}{f_N}$	$\frac{0,775}{f_N}$	$\frac{0,819}{f_N}$	$\frac{0,895}{f_N}$	$\frac{1,035}{f_N}$

Pro n>40 lze Dn*(α) aproximovat pomocí posledního řádku tabulky, kde $f_n = \frac{0,83+n}{\sqrt{n}} - 0,01$

Zdroj: http://www.utdallas.edu/~herve/Abdi-Lillie2007-pretty.pdf +lillietest v MATLABu.

Koeficienty $a_i^{(n)}$ pro Shapiro – Wilkův test

n→	2	3	4	5	6	7	8	9	10
i↓									
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739
2		0.0000	0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291
3				0.0000	0.0875	0.1401	0.1743	0.1976	0.2141
4						0.0000	0.0561	0.0947	0.1224
5								0.0000	0.0399

n→	11	12	13	14	15	16	17	18	19	20
i↓										
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4963	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565
4	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7			0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8					0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9							0.0000	0.0163	0.0303	0.0422
10									0.0000	0.0140

n→	21	22	23	24	25	26	27	28	29	30
i↓										
1	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407	0.4366	0.4328	0.4291	0.4254
2	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043	0.3018	0.2992	0.2968	0.2944
3	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533	0.2522	0.2510	0.2499	0.2487
4	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151	0.2152	0.2151	0.2150	0.2148
5	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836	0.1848	0.1857	0.1064	0.1870
6	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563	0.1584	0.1601	0.1616	0.1630
7	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316	0.1346	0.1372	0.1395	0.1415
8	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089	0.1128	0.1162	0.1192	0.1219
9	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876	0.0923	0.0965	0.1002	0.1036
10	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672	0.0728	0.0778	0.0822	0.0862
11	0.0000	0.0122	0.0228	0.0321	0.0403	0.0476	0.0540	0.0598	0.0650	0.0697
12			0.0000	0.0107	0.0200	0.0284	0.0358	0.0424	0.0483	0.0537
13					0.0000	0.0094	0.0178	0.0253	0.0320	0.0381
14							0.0000	0.0084	0.0159	0.0227
15									0.0000	0.0076

Zdroj: http://www.kmt.zcu.cz/person/Kohout/info_soubory/letnisem/ruzne/SWkoeficienty.pdf
http://www.santemaghreb.com/algerie/stat/stat_10.htm#28

Kritické hodnoty pro Shapiro – Wilkův test

	α				
n	0,01	0,02	0,05	0,1	0,5
3	0,753	0,756	0,767	0,789	0,959
4	0,687	0,707	0,748	0,792	0,935
5	0,686	0,715	0,762	0,806	0,927
6	0,713	0,743	0,788	0,826	0,927
7	0,73	0,76	0,803	0,838	0,928
8	0,749	0,778	0,818	0,851	0,932
9	0,764	0,791	0,829	0,859	0,935
10	0,781	0,806	0,842	0,869	0,938
11	0,792	0,817	0,85	0,876	0,94
12	0,805	0,828	0,859	0,883	0,943
13	0,814	0,837	0,866	0,889	0,945
14	0,825	0,846	0,874	0,895	0,947
15	0,835	0,855	0,881	0,901	0,95
16	0,884	0,863	0,887	0,906	0,952
17	0,851	0,869	0,892	0,91	0,954
18	0,858	0,874	0,897	0,914	0,956
19	0,863	0,879	0,901	0,917	0,957
20	0,868	0,884	0,905	0,92	0,959
21	0,873	0,888	0,908	0,923	0,96
22	0,878	0,892	0,911	0,926	0,961
23	0,881	0,895	0,914	0,928	0,962
24	0,884	0,898	0,916	0,93	0,963
25	0,888	0,901	0,918	0,931	0,964

	α				
n	0,01	0,02	0,05	0,1	0,5
26	0,891	0,904	0,92	0,933	0,965
27	0,894	0,906	0,923	0,935	0,965
28	0,896	0,908	0,924	0,936	0,966
29	0,898	0,91	0,926	0,937	0,966
30	0,9	0,912	0,927	0,939	0,967
31	0,902	0,914	0,929	0,94	0,967
32	0,904	0,915	0,93	0,941	0,968
33	0,906	0,917	0,931	0,942	0,968
34	0,908	0,919	0,933	0,943	0,969
35	0,91	0,92	0,934	0,944	0,969
36	0,912	0,922	0,935	0,945	0,97
37	0,914	0,924	0,936	0,946	0,97
38	0,916	0,925	0,938	0,947	0,971
39	0,917	0,927	0,939	0,948	0,971
40	0,919	0,928	0,94	0,949	0,972
41	0,92	0,929	0,941	0,95	0,972
42	0,922	0,93	0,942	0,951	0,972
43	0,923	0,932	0,943	0,951	0,973
44	0,924	0,933	0,944	0,952	0,973
45	0,926	0,934	0,945	0,953	0,973
46	0,927	0,935	0,945	0,953	0,974
47	0,928	0,936	0,946	0,954	0,974
48	0,929	0,937	0,947	0,954	0,974
49	0,929	0,937	0,947	0,955	0,974
50	0,93	0,938	0,947	0,955	0,974

Zdroj: http://www.kmt.zcu.cz/person/Kohout/info_soubory/letnisem/ruzne/SWkrithodnoty.pdf

Kritické hodnoty znaménkového testu pro n = 6, 7, .., 20, α = 0,05 a α = 0,01

n	α = 0,05		α = 0,01	
	k_1	k_2	k_1	k_2
6	0	6	-	-
7	0	7	-	-
8	0	8	0	8
9	1	8	0	9
10	1	9	0	10
11	1	10	0	11
12	2	10	1	11
13	2	11	1	12
14	2	12	1	13
15	3	12	2	13
16	3	13	2	14
17	4	13	2	15
18	4	14	3	15
19	4	15	3	16
20	5	15	3	17

Zdroj: Anděl, J.: Matematická statistika. (Tabulka XVIII.8).

Kritické hodnoty jednovýběrového Wilcoxonova testu pro n = 6, 7, .., 30, α = 0,05 a α = 0,01

n	α = 0,05	α = 0,01
	krit. hodnota	krit. hodnota
6	0	-
7	2	-
8	3	0
9	5	1
10	8	3
11	10	5
12	13	7
13	17	9
14	21	12
15	25	15
16	29	19
17	34	23
18	40	27
19	46	32
20	52	37
21	58	42
22	65	48
23	73	54
24	81	61
25	89	68
26	98	75
27	107	83
28	116	91
29	126	100
30	137	109

Zdroj: Anděl, J.: Matematická statistika. (Tabulka XVIII.9).

Kritické hodnoty dvouvýběrového Wilcoxonova testu pro m = 1, 2, .., 30, n = 1, 2, …, 30, α = 0,05

	n																			
m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	-																			
2	-	-																		
3	-	-	-																	
4	-	-	-	0																
5	-	-	0	1	2															
6	-	-	1	2	3	5														
7	-	-	1	3	5	6	8													
8	-	0	2	4	6	8	10	13												
9	-	0	2	4	7	10	12	15	17											
10	-	0	3	5	8	11	14	17	20	23										
11	--	0	3	6	9	13	16	19	23	26	30									
12	-	1	4	7	11	14	18	22	26	29	33	37								
13	-	1	4	8	12	16	20	24	28	33	37	41	45							
14	-	1	5	9	13	17	22	26	31	36	40	45	50	55						
15	-	1	5	10	14	19	24	29	34	39	44	49	54	59	64					
16	-	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75				
17	-	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87			
18	-	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99		
19	-	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	
20	-	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
21	-	2	8	15	22	29	36	43	50	58	65	73	80	88	96	103	111	119	126	134
22	-	3	9	16	23	30	38	45	53	61	69	77	85	93	101	109	117	125	133	141
23	-	3	9	17	24	32	40	48	56	64	73	81	89	98	106	115	123	132	140	149
24	-	3	10	17	25	33	42	50	59	67	76	85	94	102	111	120	129	138	147	156
25	-	3	10	18	27	35	44	53	62	71	80	89	98	107	117	126	135	145	154	161
26	-	4	11	19	28	37	46	55	64	74	83	93	102	112	122	132	141	151	161	171
27	-	4	11	20	29	38	48	57	67	77	87	97	107	117	127	137	147	158	168	178
28	-	4	12	21	30	40	50	60	70	80	90	101	111	122	132	143	154	164	175	186
29	-	4	13	22	32	42	52	62	73	83	94	105	116	127	138	149	160	171	182	193
30	-	5	13	23	33	43	54	65	76	87	98	109	120	131	143	154	166	177	189	200

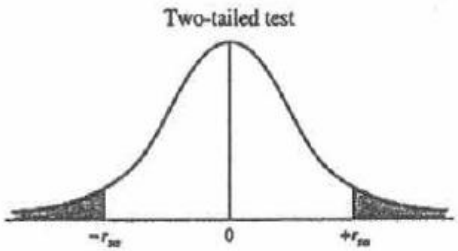
Zdroj: Anděl, J.: Matematická statistika. (Tabulka XVIII.10a).

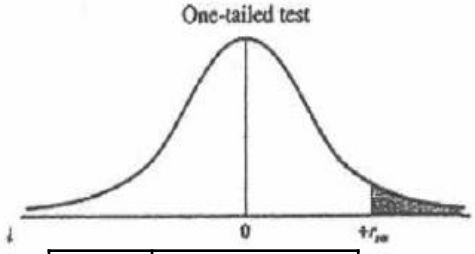
Kritické hodnoty Neményiho metody, r = 3, 4, .., 10, n = 1, 2, …, 25, α = 0,05

	r							
n	3	4	5	6	7	8	9	10
1	3,3	4,7	6,1	7,5	9,0	10,5	12,0	13,5
2	8,8	12,6	16,5	20,5	24,7	28,9	33,1	37,4
3	15,7	22,7	29,9	37,3	44,8	52,5	60,3	68,2
4	23,9	34,6	45,6	57,0	68,6	80,4	92,4	104,6
5	33,1	48,1	63,5	79,3	95,5	112,0	128,8	145,8
6	43,3	62,9	83,2	104,0	125,3	147,0	169,1	191,4
7	54,4	79,1	104,6	130,8	157,6	184,9	212,8	240,9
8	66,3	96,4	127,6	159,6	192,4	225,7	259,7	294,1
9	75,9	114,8	152,0	190,2	229,3	269,1	309,6	350,6
10	92,3	134,3	177,8	222,6	268,4	315,0	362,4	410,5
11	106,3	154,8	205,0	256,6	309,4	363,2	417,9	473,3
12	120,9	176,2	233,4	292,2	352,4	413,6	476,0	539,1
13	136,2	198,5	263,0	329,3	397,1	466,2	536,5	607,7
14	152,1	221,7	293,8	367,8	443,6	520,8	599,4	679,0
15	168,6	245,7	325,7	407,8	491,9	577,4	664,6	752,8
16	185,6	270,6	358,6	449,1	541,7	635,9	732,0	829,2
17	203,1	296,2	392,6	491,7	593,1	696,3	801,5	907,9
18	221,2	322,6	427,6	535,5	646,1	758,5	873,1	989,0
19	239,8	349,7	463,6	580,6	700,5	822,4	946,7	1072,4
20	258,8	377,6	500,5	626,9	756,4	888,1	1022,3	1158,1
21	278,4	406,1	538,4	674,4	813,7	955,4	1099,8	1245,9
22	298,4	435,3	577,2	723,0	872,3	1024,3	1179,1	1335,7
23	318,9	465,2	616,9	772,7	932,4	1094,8	1260,3	1427,7
24	339,8	495,8	657,4	823,5	993,7	1166,8	1343,2	1521,7
25	361,1	527,0	698,8	875,4	1056,3	1240,4	1427,9	1611,6

Zdroj: Blatná, Dagmar: Neparametrické metody. Tabulka T21/1.

Kritické hodnoty pro Spearmanův koeficient pořadové korelace n=5...30, α = 0,05 a α = 0,01





	alfa	
n	**0,05**	**0,01**
5	1,000	*
6	0,886	1,000
7	0,786	0,929
8	0,738	0,881
9	0,700	0,833
10	0,648	0,794
11	0,618	0,755
12	0,587	0,727
13	0,560	0,703
14	0,538	0,675
15	0,521	0,654
16	0,503	0,635
17	0,485	0,615
18	0,472	0,600
19	0,460	0,584
20	0,447	0,570
21	0,435	0,556
22	0,425	0,544
23	0,415	0,532
24	0,406	0,521
25	0,398	0,511
26	0,390	0,501
27	0,382	0,491
28	0,375	0,483
29	0,368	0,475
30	0,362	0,467

	alfa	
n	**0,05**	**0,01**
5	0,900	1,000
6	0,829	0,943
7	0,714	0,893
8	0,643	0,833
9	0,600	0,783
10	0,564	0,745
11	0,536	0,709
12	0,503	0,671
13	0,484	0,648
14	0,464	0,622
15	0,443	0,604
16	0,429	0,582
17	0,414	0,566
18	0,401	0,550
19	0,391	0,535
20	0,380	0,520
21	0,370	0,508
22	0,361	0,496
23	0,353	0,486
24	0,344	0,476
25	0,337	0,466
26	0,331	0,457
27	0,324	0,448
28	0,317	0,440
29	0,312	0,433
30	0,306	0,425

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).

Pro n > 30 lze použít testovou statistiku $r_s\sqrt{n-1}$. Platí-li H_0, pak $r_s\sqrt{n-1} \approx N(0, 1)$.

Zdroj: http://www.ace.upm.edu.my/~bas/5950/Spearman%20Rho%20Table.pdf

Adapted from Zar, J. H. (1972). Significance testing of the Spearman rank correlation. Journal of the American Statistical Association. 67, 578 – 580.

Jednorozměrná analýza a testování normality pomocí SAS, STATA a SPSS

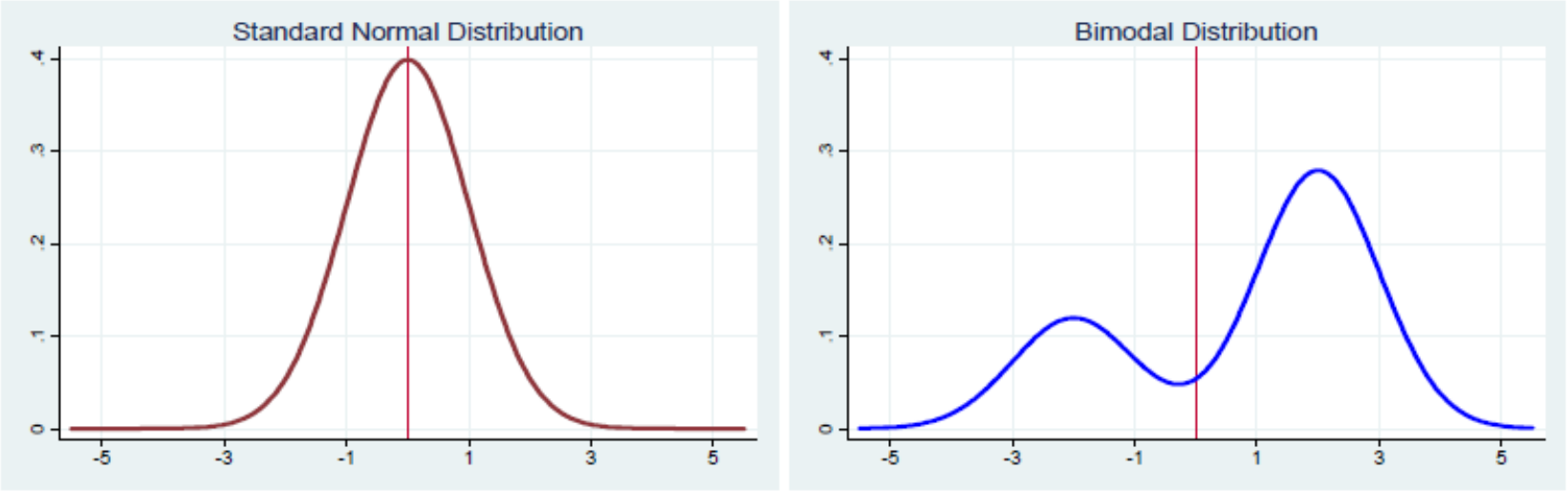
Následující text čerpá z článku Hun Myoung Park: **Univariate Analysis and Normality Test Using SAS, Stata and SPSS** (dostupný z http://www.indiana.edu/~statmath/stat/all/normality/normality.pdf)

1. Introduction

Descriptive statistics provide important information about variables to be analyzed. Mean, median, and mode measure central tendency of a variable. Measures of dispersion include variance, standard deviation, range, and interquantile range (IQR). Researchers may draw a histogram, stem-and-leaf plot, or box plot to see how a variable is distributed.

Statistical methods are based on various underlying assumptions. One common assumption is that a random variable is normally distributed. In many statistical analyses, normality is often conveniently assumed without any empirical evidence or test. But normality is critical in many statistical methods. When this assumption is violated, interpretation and inference may not be reliable or valid.

Figure 1. Comparing the Standard Normal and a Bimodal Probability Distributions



The t-test and ANOVA (Analysis of Variance) compare group means, assuming a variable of interest follows a normal probability distribution. Otherwise, these methods do not make much sense. Figure 1 illustrates the standard normal probability distribution and a bimodal distribution. How can you compare means of these two random variables?

There are two ways of testing normality (Table 1). Graphical methods visualize the distributions of random variables or differences between an empirical distribution and a theoretical distribution (e.g., the standard normal distribution). Numerical methods present

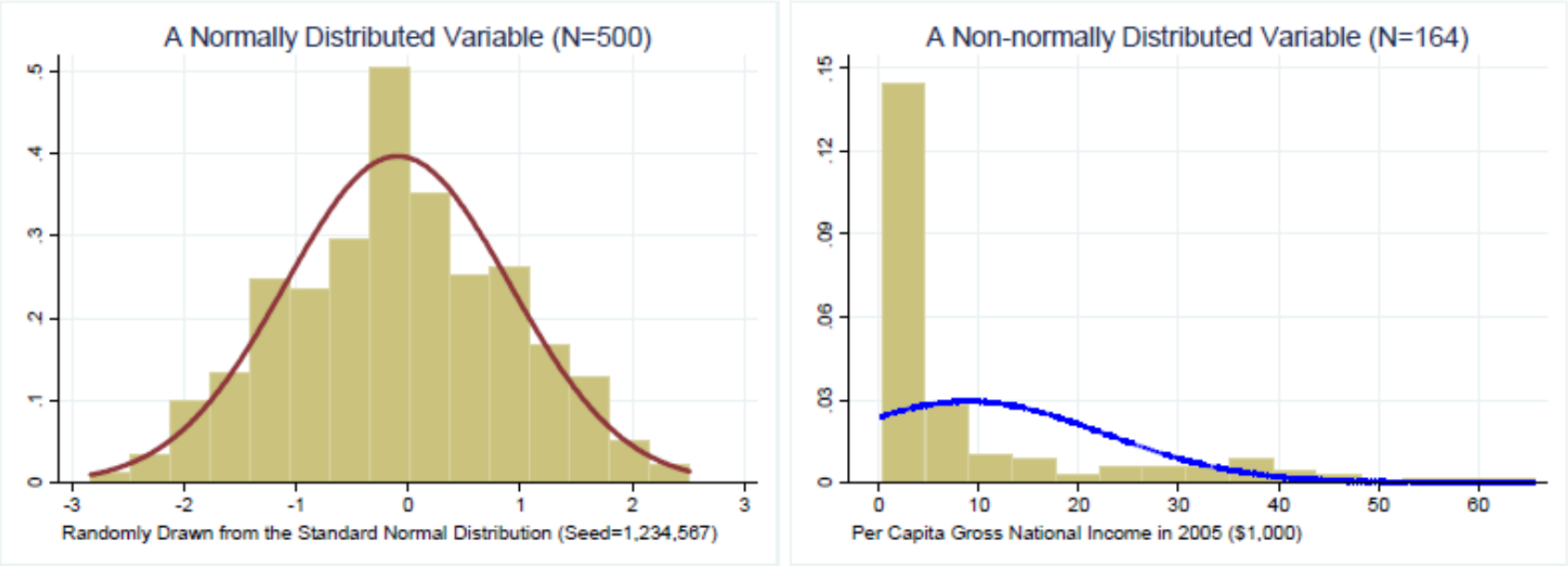
summary statistics such as skewness and kurtosis, or conduct statistical tests of normality. Graphical methods are intuitive and easy to interpret, while numerical methods provide objective ways of examining normality.

Table 1. Graphical Methods versus Numerical Methods

	Graphical Methods	Numerical Methods
Descriptive	Stem-and-leaf plot, (skeletal) box plot, dot plot, histogram	Skewness Kurtosis
Theory-driven	P-P plot Q-Q plot	Shapiro-Wilk, Shapiro- Francia test Kolmogorov-Smirnov test (Lillefors test) Anderson-Darling/Cramer-von Mises tests Jarque-Bera test, Skewness-Kurtosis test

Graphical and numerical methods are either descriptive or theory-driven. A dot plot and histogram, for instance, are descriptive graphical methods, while skewness and kurtosis are descriptive numerical methods. The P-P and Q-Q plots are theory-driven graphical methods for normality test, whereas the Shapiro-Wilk W and Jarque-Bera tests are theory-driven numerical methods.

Figure 2. Histograms of Normally and Non-normally Distributed Variables



Three variables are employed here. The first variable is unemployment rate of Illinois, Indiana, and Ohio in 2005. The second variable includes 500 observations that were randomly drawn from the standard normal distribution. This variable is supposed to be normally distributed with mean 0 and variance 1 (left plot in Figure 2). An example of a non-normal distribution is per capita gross national income (GNI) in 2005 of 164 countries in the world. GNIP is severely skewed to the right and is least likely to be normally distributed (right plot in Figure 2). See the Appendix for details.

2. Graphical Methods

Graphical methods visualize the distribution of a random variable and compare the distribution to a theoretical one using plots. These methods are either descriptive or theory-driven. The former method is based on the empirical data, whereas the latter considers both empirical and theoretical distributions.

2.1 Descriptive Plots

Among frequently used descriptive plots are the stem-and-leaf-plot, dot plot, (skeletal) box plot, and histogram. When N is small, a stem-and-leaf plot and dot plot are useful to summarize continuous or event count data. Figure 3 and 4 respectively present a stem-and-leaf plot and a dot plot of the unemployment rate of three states.

Figure 3. Stem-and-Leaf Plot of Unemployment Rate of Illinois, Indiana, Ohio

```
-> state = IL

Stem-and-leaf plot for rate(Rate)

rate rounded to nearest multiple
of .1
plot in units of .1

  3. | 7889
  4* | 011122344
  4. | 556666666677778888999
  5* | 0011122222333333344444
  5. | 5555667777777888999
  6* | 000011222333444
  6. | 555579
  7* | 0033
  7. |
  8* | 0
  8. | 8
```

```
-> state = IN

Stem-and-leaf plot for rate(Rate)

rate rounded to nearest multiple
of .1
plot in units of .1

  3* | 1
  3. | 89
  4* | 012234
  4. | 566666778889999
  5* | 00000111222222233344
  5. | 555666666777889
  6* | 002222233344
  6. | 5666677889
  7* | 1113344
  7. | 67
  8* | 14
```

```
-> state = OH

Stem-and-leaf plot for rate (Rate)

rate rounded to nearest multiple
of .1
plot in units of .1

  3* | 8
  4* | 014577899
  5* | 012233334455566677778888888999
  6* | 00111112222233444446678899
  7* | 01223335677
  8* | 1223338
  9* | 99
 10* | 1
 11* |
 12* |
 13* | 3
```

A box plot presents the minimum, 25th percentile (1st quartile), 50th percentile (median), 75th percentile (3rd quartile), and maximum in a box and lines.[1] Outliers, if any, appear at the outsides of (adjacent) minimum and maximum lines. As such, a box plot effectively summarizes these major percentiles using a box and lines. If a variable is normally distributed, its 25th and 75th percentile are symmetric, and its median and mean are located at the same point exactly in the center of the box.[2]

In Figure 5, you should see outliers in Illinois and Ohio that affect the shapes of corresponding boxes. By contrast, the Indiana unemployment rate does not have outliers, and its symmetric box implies that the rate appears to be normally distributed.

Figure 5. Box Plots of Unemployment Rates of Illinois, Indiana, and Ohio

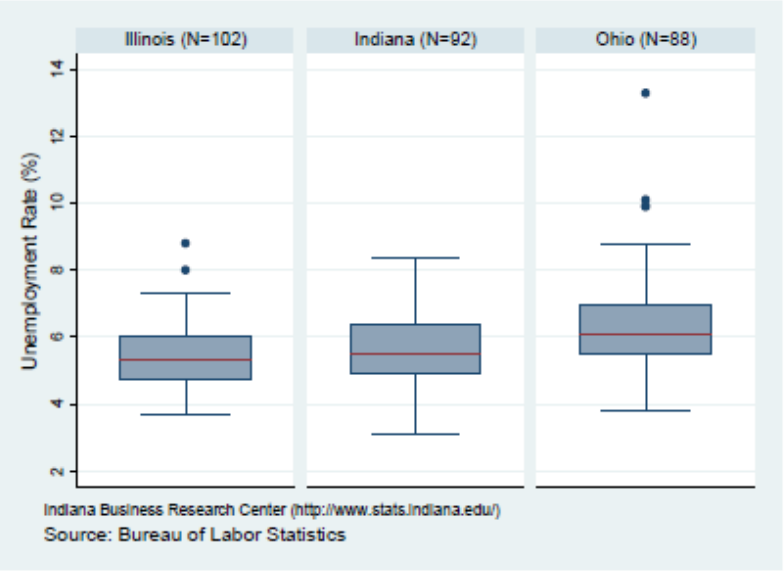
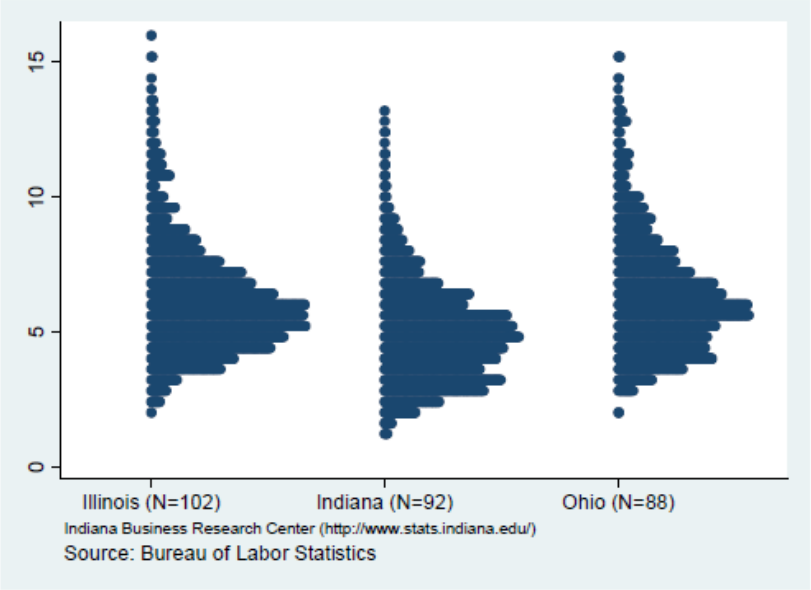
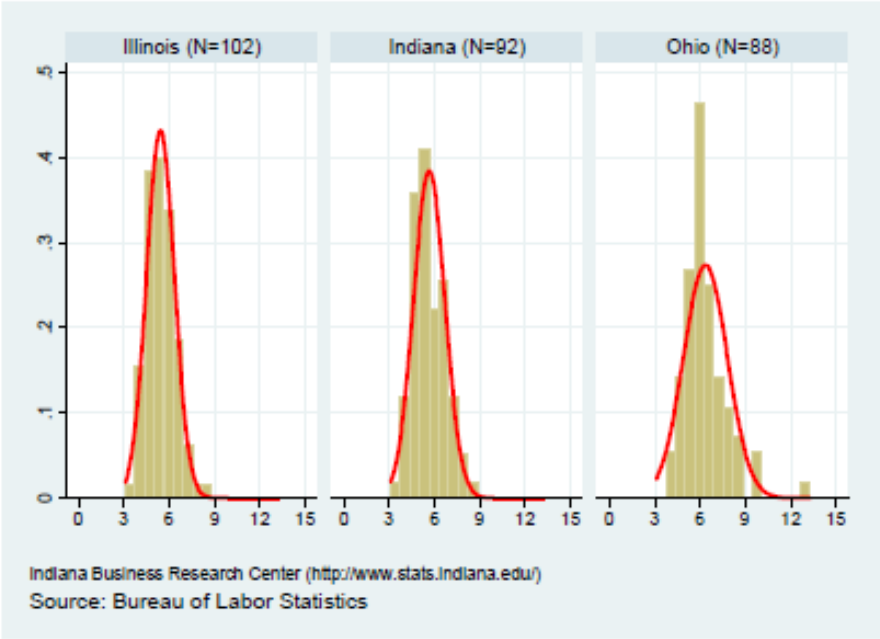


Figure 4. Dot Plot of Unemployment Rate of Illinois, Indiana, Ohio



The histogram graphically shows how each category (interval) accounts for the proportion of total observations and is appropriate when N is large (Figure 6).

Figure 6. Histograms of Unemployment Rates of Illinois, Indiana and Ohio



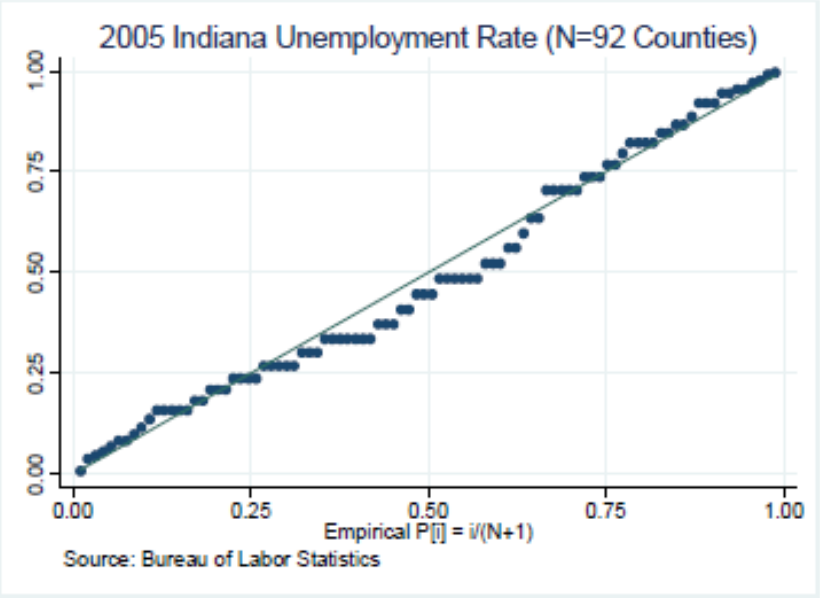
[1] The first quartile cuts off lowest 25 percent of data; the second quartile, median, cuts data set in half; and the third quartile cuts off lowest 75 percent or highest 25 percent of data. See http://en.wikipedia.org/wiki/Quartile

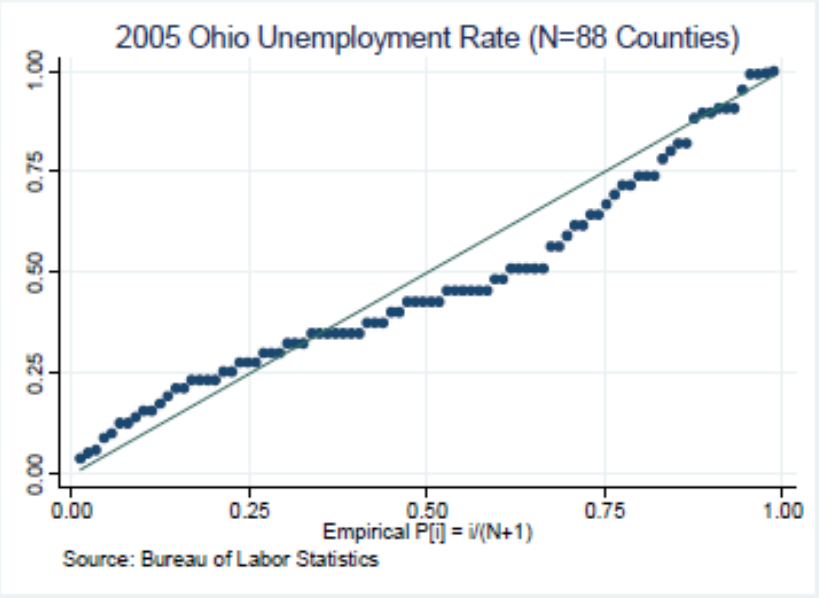
[2] SAS reports a mean as "+" between (adjacent) minimum and maximum lines.

2.2 Theory-driven Plots

P-P and Q-Q plots are considered here. The probability-probability plot (P-P plot or percent plot) compares an empirical cumulative distribution function of a variable with a specific theoretical cumulative distribution function (e.g., the standard normal distribution function). In Figure 7, Ohio appears to deviate more from the fitted line than Indiana.

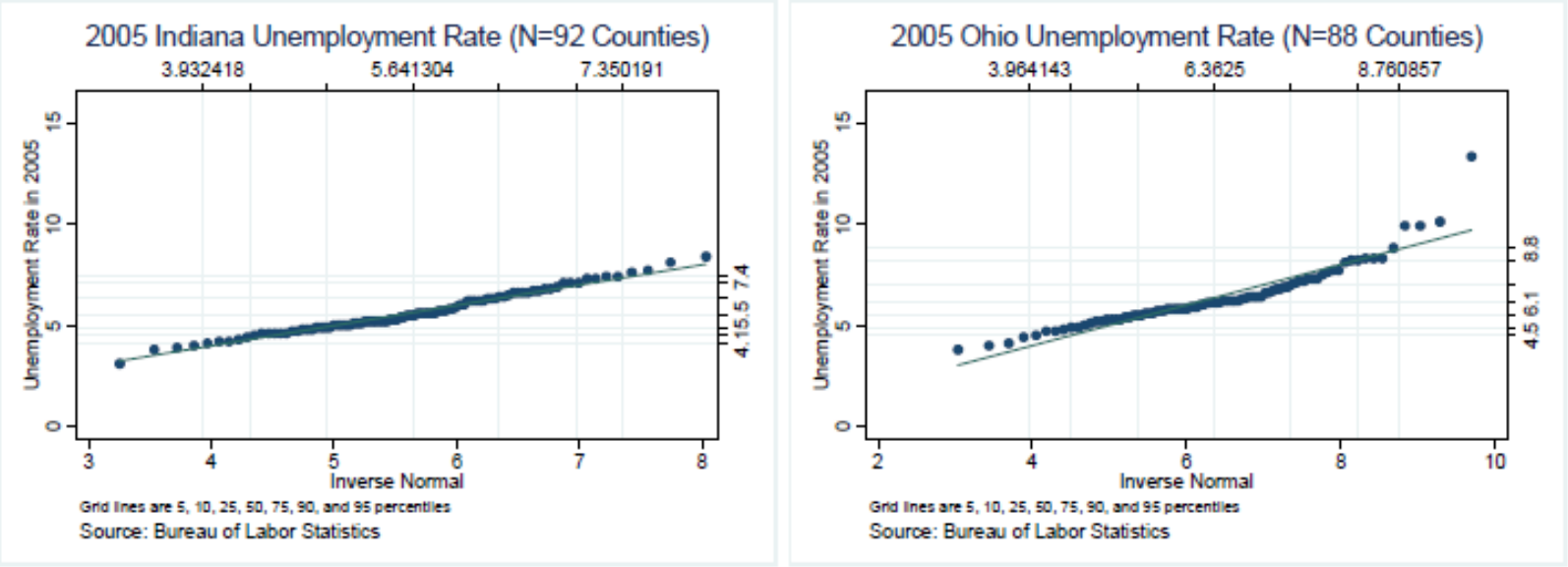
Figure 7. P-P Plots of Unemployment Rates of Indiana and Ohio (Year 2005)





Similarly, the quantile-quantile plot (Q-Q plot) compares ordered values of a variable with quantiles of a specific theoretical distribution (i.e., the normal distribution). If two distributions match, the points on the plot will form a linear pattern passing through the origin with a unit slope. P-P and Q-Q plots are used to see how well a theoretical distribution models the empirical data. In Figure 8, Indiana appears to have a smaller variation in its unemployment rate than Ohio. By contrast, Ohio appears to have a wider range of outliers in the upper extreme.

Figure 8. Q-Q Plots of Unemployment Rates of Indiana and Ohio (Year 2005)



Detrended normal P-P and Q-Q plots depict the actual deviations of data points from the straight horizontal line at zero. No specific pattern in a detrended plot indicates normality of the variable. SPSS can generate detrended P-P and Q-Q plots.

3. Numerical Methods

Graphical methods, although visually appealing, do not provide objective criteria to determine normality of variables. Interpretations are thus a matter of judgments. Numerical methods use descriptive statistics and statistical tests to examine normality.

3.1 Descriptive Statistics

Measures of dispersion such as variance reveal how observations of a random variable deviate from their mean. The second central moment is

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

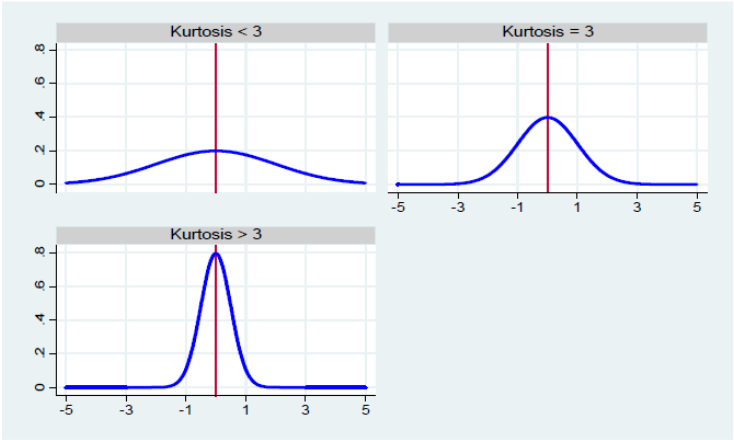
Skewness is a third standardized moment that measures the degree of symmetry of a probability distribution. If skewness is greater than zero, the distribution is skewed to the right, having more observations on the left.

$$\frac{E[(x-\mu)^3]}{\sigma^3} = \frac{\sum (x_i - \bar{x})^3}{s^3(n-1)} = \frac{\sqrt{n-1}\sum (x_i - \bar{x})^3}{[\sum (x_i - \bar{x})^2]^{3/2}}$$

Kurtosis, based on the fourth central moment, measures the thinness of tails or "peakedness" of a probability distribution.

$$\frac{E[(x-\mu)^4]}{\sigma^4} = \frac{\sum (x_i - \bar{x})^4}{s^4(n-1)} = \frac{(n-1)\sum (x_i - \bar{x})^4}{[\sum (x_i - \bar{x})^2]^2}$$

Figure 9. Probability Distributions with Different Kurtosis



If kurtosis of a random variable is less than three (or if kurtosis-3 is less than zero), the distribution has thicker tails and a lower peak compared to a normal distribution (first plot in Figure 9).[3] By contrast, kurtosis larger than 3 indicates a higher peak and thin tails (last plot). A normally distributed random variable should have skewness and kurtosis near zero and three, respectively (second plot in Figure 9).

state	N	mean	median	max	min	variance	skewness	kurtosis
IL	102	5.421569	5.35	8.8	3.7	.8541837	.6570033	3.946029
IN	92	5.641304	5.5	8.4	3.1	1.079374	.3416314	2.785585
OH	88	6.3625	6.1	13.3	3.8	2.126049	1.665322	8.043097
Total	282	5.786879	5.65	13.3	3.1	1.473955	1.44809	8.383285

In short, skewness and kurtosis show how the distribution of a variable deviates from a normal distribution. These statistics are based on the empirical data.

3.2 Theory-driven Statistics

The numerical methods of normality test include the Kolmogorov-Smirnov (K-S) D test (Lilliefors test), Shapiro-Wilk test, Anderson-Darling test, and Cramer-von Mises test (SAS Institute 1995).[4] The K-S D test and Shapiro-Wilk W test are commonly used. The K-S, Anderson-Darling, and Cramer-von Misers tests are based on the empirical distribution function (EDF), which is defined as a set of N independent observations $x_1, x_2, \ldots x_n$ with a common distribution function F(x) (SAS 2004).

Table 2. Numerical Methods of Testing Normality

Test	Statistic	N Range	Dist.	SAS	Stata	SPSS
Jarque-Bera	χ^2		$\chi^2(2)$	-	-	-
Skewness-Kurtosis	χ^2	9≤N	$\chi^2(2)$	-	.sktest	-
Shapiro-Wilk	W	7≤N≤ 2,000	-	YES	.swilk	YES
Shapiro-Francia	W'	5≤N≤ 5,000	-	-	.sfrancia	-
Kolmogorov-Smirnov	D		EDF	YES	*	YES
Cramer-vol Mises	W^2		EDF	YES	-	-
Anderson-Darling	A^2		EDF	YES	-	-

* Stata .ksmirnov command is not used for testing normality.

The Shapiro-Wilk W is the ratio of the best estimator of the variance to the usual corrected sum of squares estimator of the variance (Shapiro and Wilk 1965).[5] The statistic is positive and less than or equal to one. Being close to one indicates normality.

[3] SAS and SPSS produce (kurtosis -3), while Stata returns the kurtosis. SAS uses its weighted kurtosis formula with the degree of freedom adjusted. So, if N is small, SAS, Stata, and SPSS may report different kurtosis.
[4] The UNIVARIATE and CAPABILITY procedures have the NORMAL option to produce four statistics.
[5] The W statistic was constructed by considering the regression of ordered sample values on corresponding expected normal order statistics, which for a sample from a normally distributed population is linear (Royston 1982). Shapiro and Wilk's (1965) original W statistic is valid for the sample sizes between 3 and 50, but Royston extended the test by developing a transformation of the null distribution of W to approximate normality throughout the range between 7 and 2000.

The W statistic requires that the sample size is greater than or equal to 7 and less than or equal to 2,000 (Shapiro and Wilk 1965).[6]

$$W = \frac{\left(\sum a_i x_{(i)}\right)^2}{\sum (x_i - \bar{x})^2}$$

where $a' = (a_1, a_2, \ldots, a_n) = m'V^{-1}[m'V^{-1}V^{-1}m]^{-1/2}$, $m' = (m_1, m_2, \ldots, m_n)$ is the vector of expected values of standard normal order statistics, V is the n by n covariance matrix, $x' = (x_1, x_2, \ldots, x_n)$ is a random sample, and $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$.

The Shapiro-Francia W' test is an approximate test that modifies the Shapro-Wilk W. The S-F statistic uses $b' = (b_1, b_2, \ldots, b_n) = m'(m'm)^{-1/2}$ instead of a'. The statistic was developed by Shapiro and Francia (1972) and Royston (1983). The recommended sample sizes for the Stata `.sfrancia` command range from 5 to 5,000 (Stata 2005). SAS and SPSS do not support this statistic. Table 3 summarizes test statistics for 2005 unemployment rates of Illinois, Indiana, and Ohio. Since N is not large, you need to read Shapiro-Wilk, Shapiro-Francia, Jarque-Bera, and Skewness-Kurtosis statistics.

Table 3. Normality Test for 2005 Unemployment Rates of Illinois, Indiana, and Ohio

State	Illinois		Indiana		Ohio	
	Test	P-value	Test	P-value	Test	P-value
Shapiro-Wilk [sas]	.9714	.0260	.9841	.3266	.8858	.0001
Shapiro-Wilk [stata]	.9728	.0336	.9855	.4005	.8869	.0000
Shapiro-Francia [stata]	.9719	.0292	.9858	.3545	.8787	.0000
Kolmogorov-Smirnov [sas]	.0583	.1500	.0919	.0539	.1602	.0100
Cramer-von Misers [sas]	.0606	.2500	.1217	.0582	.4104	.0050
Anderson-Darling [sas]	.4534	.2500	.6332	.0969	2.2815	.0050
Jarque-Bera	12.2928	.0021	1.9458	.3380	149.5495	.0000
Skewness-Kurtosis [stata]	10.59	.0050	1.99	.3705	43.75	.0000

The SAS UNIVARIATE and CAPABILITY procedures perform the Kolmogorov-Smirnov D, Anderson-Darling A^2, and Cramer-von Misers W^2 tests, which are useful especially when N is larger than 2,000.

3.3 Jarque-Bera (Skewness-Kurtosis) Test

The test statistics mentioned in the previous section tend to reject the null hypothesis when N becomes large. Given a large number of observations, the Jarque-Bera test and Skewness-Kurtosis test will be alternative ways of normality test.

The Jarque-Bera test, a type of Lagrange multiplier test, was developed to test normality, heteroscedasticy, and serial correlation (autocorrelation) of regression residuals (Jarque and Bera 1980). The Jarque-Bera statistic is computed from skewness and kurtosis and asymptotically follows the chi-squared distribution with two degrees of freedom.

[6] Stata `.swilk` command, based on Shapiro and Wilk (1965) and Royston (1992), can be used with from 4 to 2000 observations (Stata 2005).

$$n\left[\frac{skewness^2}{6}+\frac{(kurtosis-3)^2}{24}\right]\sim\chi^2(2)$$, where n is the number of observations.

The above formula gives a penalty for increasing the number of observations and thus implies a good asymptotic property of the Jarque-Bera test. The computation for 2005 unemployment rates is as follows.[7]

For Illinois: `12.292825 = 102*(0.66685022^2/6 + 1.0553068^2/24)`
For Indiana: `1.9458304 = 92*(0.34732004^2/6 + (-0.1583764)^2/24)`
For Ohio: `149.54945 = 88*(1.69434105^2/6 + 5.4132289^2/24)`

The Stata Skewness-Kurtosis test is based on D'Agostino, Belanger, and D'Agostino, Jr. (1990) and Royston (1991) (Stata 2005). Note that in Ohio the Jarque-Bera statistic of 150 is quite different from the S-K statistic of 44 (see Table 3).

Table 4 Comparison of Methods for Testing Normality

N	10	100	500	1,000	5,000	10,000
Mean	.5240	-.0711	-.0951	-.0097	-.0153	-.0192
Standard deviation	.9554	1.0701	1.0033	1.0090	1.0107	1.0065
Minimum	-.8659	-2.8374	-2.8374	-2.8374	-3.5387	-3.9838
1st quantile	-.2372	-.8674	-.8052	-.7099	-.7034	-.7121
Median	.6411	-.0625	-.1196	-.0309	-.0224	-.0219
3rd quantile	1.4673	.7507	.6125	.7027	.6623	.6479
Maximum	1.7739	1.9620	2.5117	3.1631	3.5498	4.3140
Skewness [sas]	-.1620	-.2272	-.0204	.0100	.0388	.0391
Kurtosis-3 [sas]	-1.4559	-.5133	-.3988	-.2633	-.0067	-.0203
Jarque-Bera	.9269 (.6291)	1.9580 (.3757)	3.3483 (.1875)	2.9051 (.2340)	1.2618 (.5321)	2.7171 (.2570)
Skewness [stata]	-.1366	-.2238	-.0203	.0100	.0388	.0391
Kurtosis [stata]	1.6310	2.4526	2.5932	2.7320	2.9921	2.9791
S-K [stata]	1.52 (.4030)	2.52 (.2843)	4.93 (.0850)	3.64 (.1620)	1.26 (.5330)	2.70 (.2589)
Shapiro-Wilk W [sas]	.9359 (.5087)	.9840 (.2666)	.9956 (.1680)	.9980 (.2797)	*.9998 (.8727)*	*.9999 (.8049)*
Shapiro-F W' [stata]	.9591 (.7256)	.9873 (.3877)	.9965 (.2941)	.9983 (.4009)	.9998 (.1000)	*.9998 (.1000)*
Kolmogorov-S D [sas]	.1382 (.1500)	.0708 (.1500)	.0269 (.1500)	.0180 (.1500)	.0076 (.1500)	.0073 (.1500)
Cramer-M W^2 [sas]	.0348 (.2500)	.0793 (.2167)	.0834 (.1945)	.0607 (.2500)	.0304 (.2500)	.0652 (.2500)
Anderson-D A^2 [sas]	.2526 (.2500)	.4695 (.2466)	.5409 (.1712)	.4313 (.2500)	.1920 (.2500)	.4020 (.2500)

* P-value in parentheses

Table 4 presents results of normality tests for random variables with different numbers of observations. The data were randomly generated from the standard normal distribution with a seed of 1,234,567 in SAS. As N grows, the mean, median, skewness, and (kurtosis-3) approach zero, and the standard deviation gets close to 1. The Kolmogorov-Smirnov D, Anderson-

[7] Skewness and Kurtosis are computed using the SAS UNIVARIATE and CAPABILITY procedures that report kurtosis minus 3.

Darling A^2, Cramer-von Mises W^2 are computed in SAS, while the Skewness-Kurtosis and Shapiro-Francia W' are computed in Stata.

All four statistics do not reject the null hypothesis of normality regardless of the number of observations (Table 4). Note that the Shapiro-Wilk W is not reliable when N is larger than 2,000 and S-F W' is valid up to 5,000 observations. The Jarque-Bera and Skewness-Kurtosis tests show consistent results.

3.4 Software Issues

The UNIVARIATE procedure of SAS/BASE and CAPABILITY of SAS/QC compute various statistics and produce P-P and Q-Q plots. These procedures provide many numerical methods including Cramer-vol Mises and Anderson-Darling.[8] The P-P plot is generated only in CAPABILITY.

By contrast, Stata has many individual commands to examine normality. In particular, Stata provides `.sktest` and `.sfrancia` to conduct Skewness-Kurtosis and Shapiro-Francia W' tests, respectively.

SPSS EXAMINE provides numerical and graphical methods for normality test. The detrended P-P and Q-Q plots can be generated in SPSS. Since SPSS has changed graph-related features over time, you need to check menus, syntaxes, and reported bugs.

Table 5 summarizes SAS procedures and Stata/SPSS commands that are used to test normality of random variables.

Table 5. Comparison of Procedures and Commands Available

	SAS	Stata	SPSS
Descriptive statistics	UNIVARIATE	.summarize	Descriptives, Frequencies
(Skewness/Kurtosis)		.tabstat	Examine
Histogram, dot plot	UNIVARIATE	.histogram	Graph, Igraph, Examine,
	CHART, PLOT	.dotplot	Frequencies
Stem-leaf-plot	UNIVARIATE*	.stem	Examine
Box plot	UNIVARIATE*	.graph box	Examine, Igraph
P-P plot	CAPABILITY**	.pnorm	Pplot
Q-Q plot	UNIVARIATE	.qnorm	Pplot, Examine
Detrended Q-Q/P-P plot			Pplot, Examine
Jarque-Bera (S-K) test		.sktest	
Shapiro-Wilk W	UNIVARIATE	.swilk	Examine
Shapiro-Francia W'		.sfrancia	
Kolmogorov-Smirnov	UNIVARIATE		Examine
Cramer-vol Mises	UNIVARIATE		
Anderson-Darling	UNIVARIATE		

* The UNIVARIATE procedure can provide the plot.
** The CAPABILITY procedure can provide the plot.

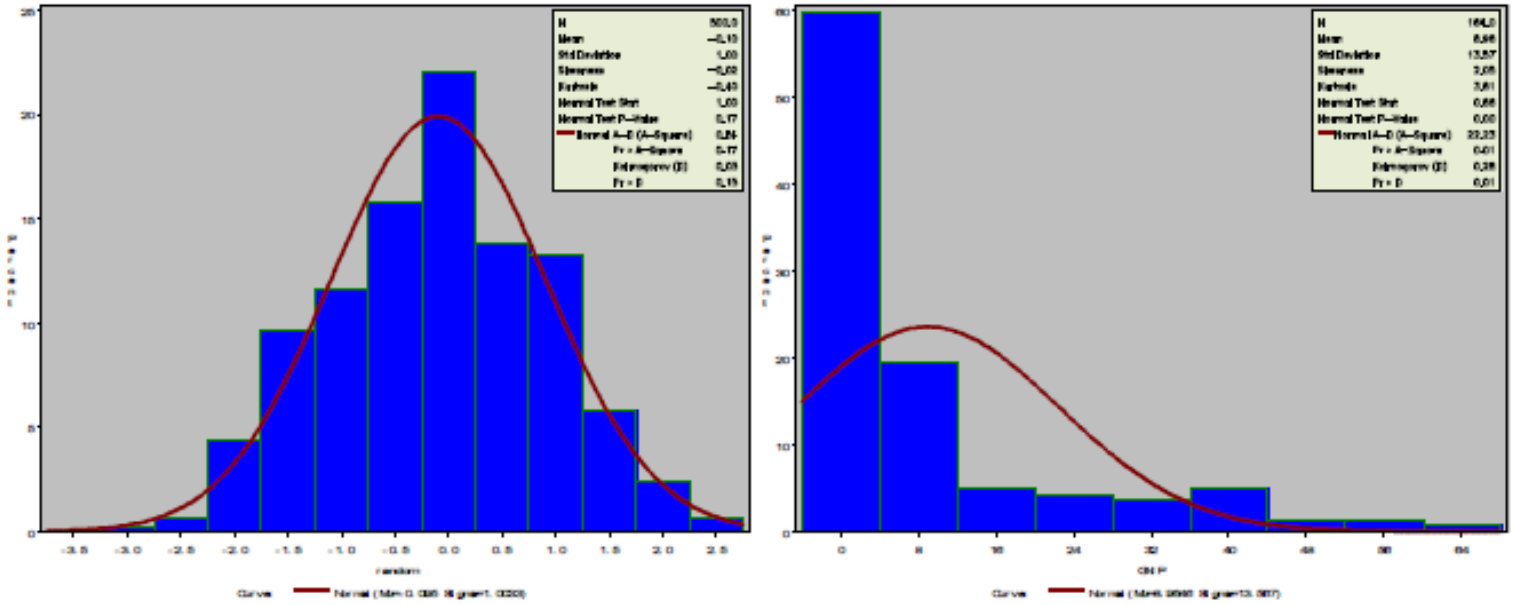
[8] MINITAB also performs the Kolmogorov-Smirnov and Anderson-Darling tests.

4. Testing Normality in SAS

SAS has the UNIVARIATE and CAPABILITY procedures to compute descriptive statistics, draw various graphs, and conduct statistical tests for normality. Two procedures have similar usage and produce similar statistics in the same format. However, UNIVARIATE produces a stem-and-leaf plot, box plot, and normal probability plot, while CAPABILITY provides P-P plot and CDP plot that UNIVARIATE does not.

This section illustrates how to summarize normally and non-normally distributed variables and conduct normality tests of these variables using the two procedures (see Figure 10).

Figure 10. Histogram of Normally and Non-normally Distributed Variables



4.1 A Normally Distributed Variable

The UNIVARIATE procedure provides a variety of descriptive statistics, Q-Q plot, leaf-and-stem-plot, and box plot. This procedure also conducts Kolmogorov-Smirnov test, Shapiro-Wilk' test, Anderson-Darling, and Cramer-von Misers tests.

Let us take a look at an example of the UNIVARIATE procedure. The NORMAL option conducts normality testing; PLOT draws a leaf-and-stem plot and a box plot; finally, the QQPLOT statement draws a Q-Q plot.

```
PROC UNIVARIATE DATA=masil.normality NORMAL PLOT;
       VAR random;
       QQPLOT random /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
RUN;
```

Like UNIVARIATE, the CAPABILITY procedure also produces various descriptive statistics and plots. CAPABILITY can draw a P-P plot using the PPPLOT option but does not support a leaf-and-stem plot, a box plot, and a normal probability plot; this procedure does not have the PLOT option available in UNIVARIATE.

4.1.1 SAS Output of Descriptive Statistics

The following is an example of the CAPABILITY procedure. QQPLOT, PPPLOT, and HISTOGRAM statements respectively draw a Q-Q plot, P-P plot, and histogram. Note that the INSET statement adds summary statistics to graphs such as histogram and Q-Q plot.

```
PROC CAPABILITY DATA=masil.normality NORMAL;
       VAR random;
       QQPLOT random /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
       PPPLOT random /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
       HISTOGRAM /NORMAL(COLOR=MAROON W=4) CFILL  = BLUE CFRAME = LIGR;
       INSET MEAN STD /CFILL=BLANK FORMAT=5.2 ;
RUN;
```

The CAPABILITY Procedure
Variable: random

Moments

N	500	Sum Weights	500
Mean	-0.0950725	Sum Observations	-47.536241
Std Deviation	1.00330171	Variance	1.00661432
Skewness	-0.0203721	Kurtosis	-0.3988198
Uncorrected SS	506.819932	Corrected SS	502.300544
Coeff Variation	-1055.3019	Std Error Mean	0.04486902

Basic Statistical Measures

Location		Variability	
Mean	-0.09507	Std Deviation	1.00330
Median	-0.11959	Variance	1.00661
Mode	.	Range	5.34911
		Interquartile Range	1.41773

Tests for Location: Mu0=0

Test		-Statistic-		-----p Value------
Student's t	t	-2.11889	Pr > |t|	0.0346
Sign	M	-28	Pr >= |M|	0.0138
Signed Rank	S	-6523	Pr >= |S|	0.0435

Tests for Normality

Test		--Statistic---		-----p Value-----
Shapiro-Wilk	W	0.995564	Pr < W	0.168
Kolmogorov-Smirnov	D	0.026891	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.083351	Pr > W-Sq	0.195
Anderson-Darling	A-Sq	0.540894	Pr > A-Sq	0.171

Quantiles (Definition 5)

Quantile	Estimate
100% Max	2.511694336
99%	2.055464409
95%	1.530450397
90%	1.215210586
75% Q3	0.612538495
50% Median	-0.119592165
25% Q1	-0.805191028
10%	-1.413548051
5%	-1.794057126
1%	-2.219479314
0% Min	-2.837417522

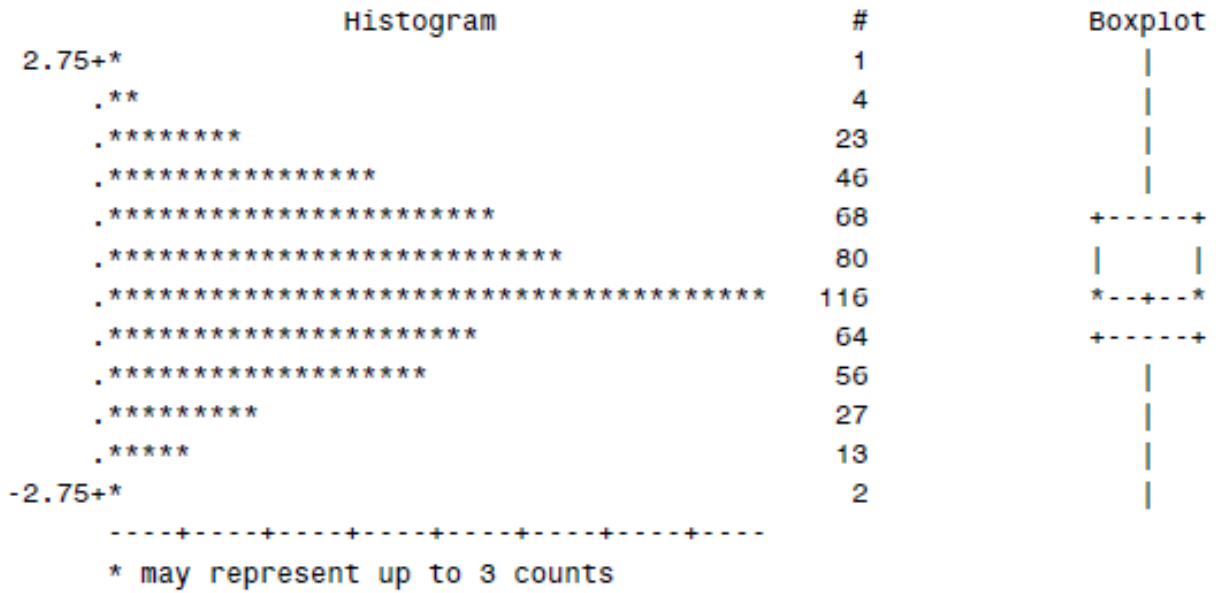
Extreme Observations

-------Lowest-------		-------Highest------	
Value	Obs	Value	Obs
-2.83741752	29	2.14897641	119
-2.59039285	204	2.21109349	340
-2.47829639	73	2.42113892	325
-2.39126554	391	2.42171307	139
-2.24047386	393	2.51169434	332

4.1.2 Graphical Methods

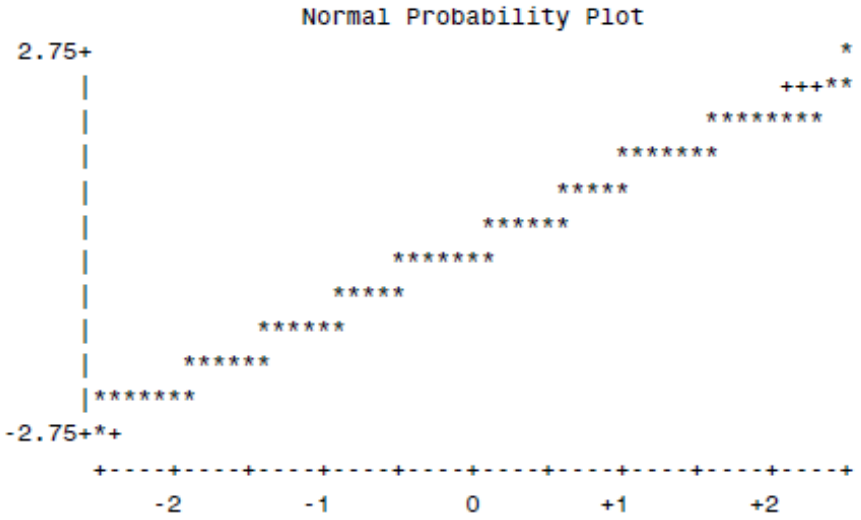
The stem-and-leaf plot and box plot, produced by the UNIVARIATE produre, illustrate that the variable is normally distributed (Figure 11). The locations of first quantile, mean, median, and third quintile indicate a bell-shaped distribution. Note that the mean -.0951 and median -.1196 are very close.

Figure 11. Stem-and-Leaf Plot and Box Plot of a Normally Distributed Variable



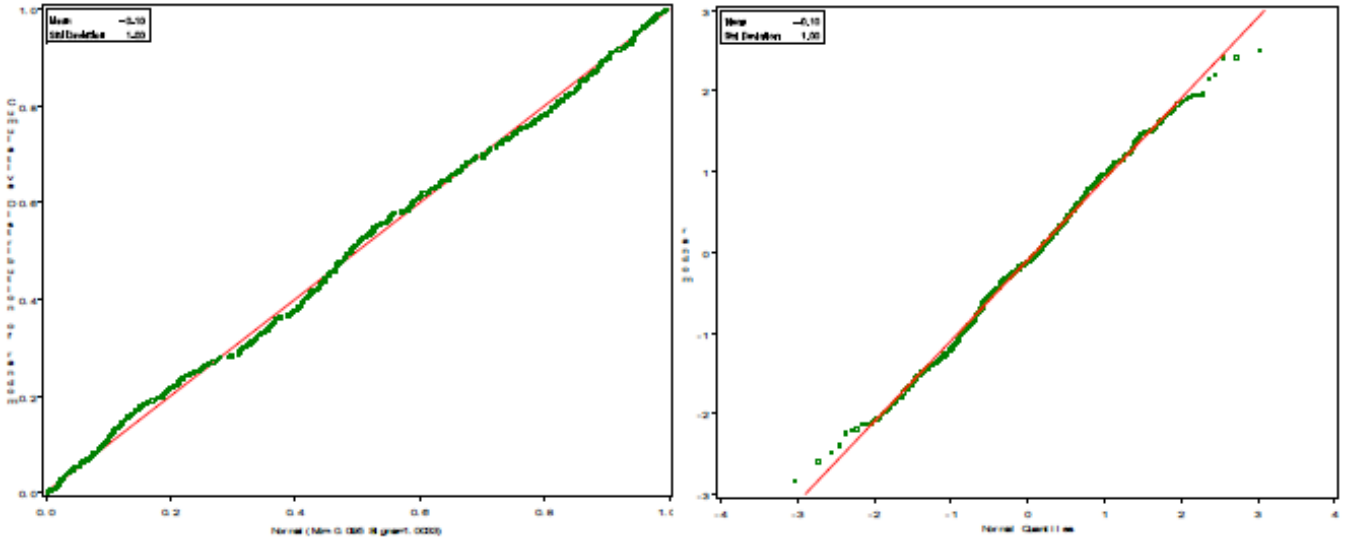
The normal probability plot available in UNIVARIATE shows a straight line, implying the normality of the randomly drawn variable (Figure 12).

Figure 12. Normal Probability Plot of a Normally Distributed Variable



The P-P and Q-Q plots below show that the data points are not seriously deviated from the fitted line. They consistently indicate that the variable is normally distributed.

Figure 13. P-P plot and Q-Q Plot of a Normally Distributed Variable



4.1.3 Numerical Methods

The mean of -.0951 is very close to 0 and variance is almost 1. The skewness and kurtosis-3 are respectively -.0204 and -.3988, indicating an almost normal distribution. However, these descriptive statistics do not provide conclusive information about normality.

SAS provides four different statistics for testing normality. Shapiro-Wilk W of .9956 does not reject the null hypothesis that the variable is normally distributed (p<.168). Similarly, Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests do not reject the null hypothesis. Since the number of observations is less than 2,000, however, Shapiro-Wilk W test will be appropriate for this case.

The Jarque-Bera test also indicates the normality of the randomly drawn variable (p=.1875). Note that -.3988 is kurtosis -3.

$$500\left[\frac{-0.0203721^2}{6}+\frac{-0.3988198^2}{24}\right]\sim 3.3482776(2)$$

Consequently, we can safely conclude that the randomly drawn variable is normally distributed.

4.2 A Non-normally Distributed Variable

Let us examine the per capita gross national income as an example of non-normally distributed variables. See the appendix for details about this variable.

4.2.1 SAS Output of Descriptive Statistics

This section employs the UNIVARIATE procedure to compute descriptive statistics and perform normality tests. The variable has mean 8.9646 and median 2.0495, where are substantially different. Variance 184.0577 is extremely large.

```
PROC UNIVARIATE DATA=masil.gnip NORMAL PLOT;
       VAR gnip;
       QQPLOT gnip /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
       HISTOGRAM / NORMAL(COLOR=MAROON W=4) CFILL = BLUE CFRAME = LIGR;
RUN;
```

The UNIVARIATE Procedure
Variable: GNIP

Moments

N	164	Sum Weights	164
Mean	8.9645732	Sum Observations	1470.19001
Std Deviation	13.5667877	Variance	184.057728
Skewness	2.04947469	Kurtosis	3.60816725
Uncorrected SS	43181.0356	Corrected SS	30001.4096
Coeff Variation	151.337798	Std Error Mean	1.05938813

Basic Statistical Measures

Location		Variability	
Mean	8.964573	Std Deviation	13.56679
Median	2.765000	Variance	184.05773
Mode	1.010000	Range	65.34000
		Interquartile Range	7.72500

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	8.462029	Pr > |t|	<.0001
Sign	M	82	Pr >= |M|	<.0001
Signed Rank	S	6765	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.663114	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.284426	Pr > D	<0.0100
Cramer-von Mises	W-Sq	4.346966	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	22.23115	Pr > A-Sq	<0.0050

Quantiles (Definition 5)

Quantile	Estimate
100% Max	65.630
99%	59.590
95%	38.980
90%	32.600
75% Q3	8.680
50% Median	2.765
25% Q1	0.955
10%	0.450
5%	0.370
1%	0.290
0% Min	0.290

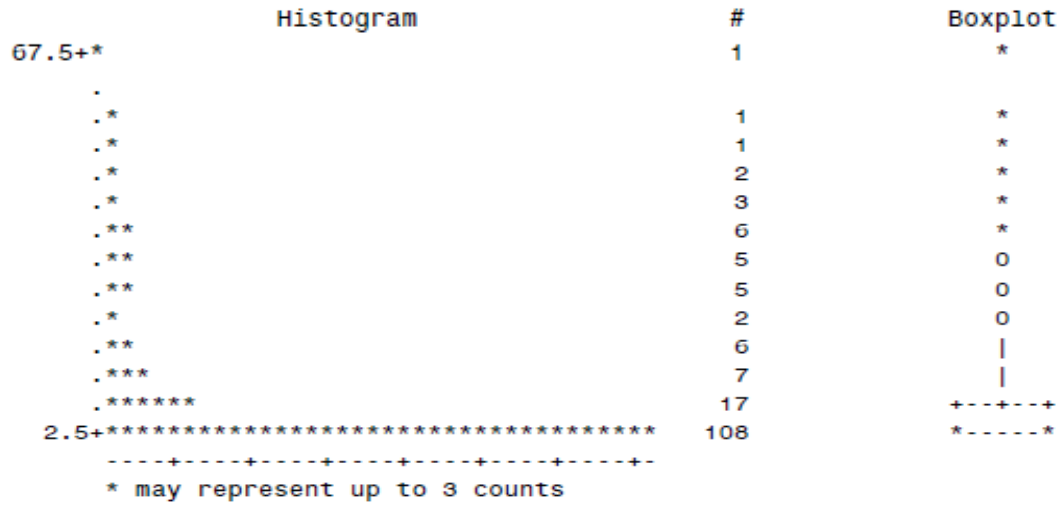
Extreme Observations

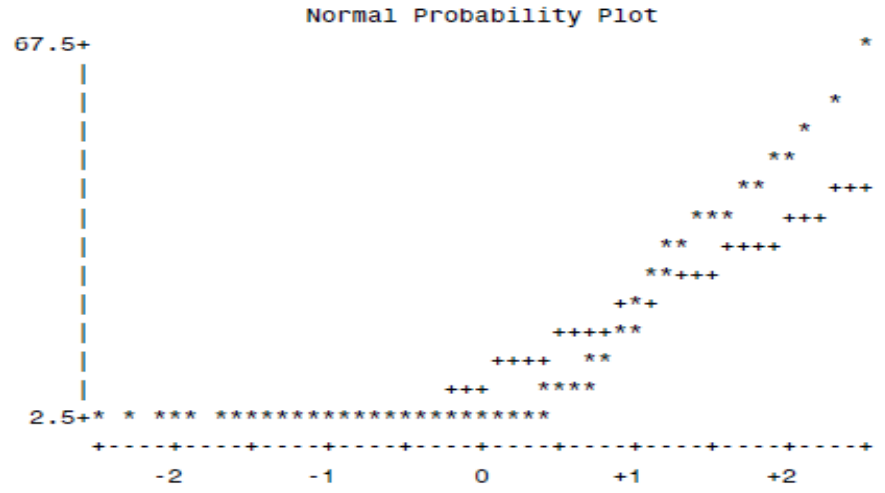
----Lowest----		----Highest----	
Value	Obs	Value	Obs
0.29	164	46.32	5
0.29	163	47.39	4
0.31	162	54.93	3
0.33	161	59.59	2
0.34	160	65.63	1

4.2.2 Graphical Methods

The stem-and-leaf plot, box plot, and normal probability plots all indicate that the variable is not normally distributed (Figure 14). Most observations are highly concentrated on the left side of the distribution. See the stem-and-leaf plot and box plot in Figure 14.

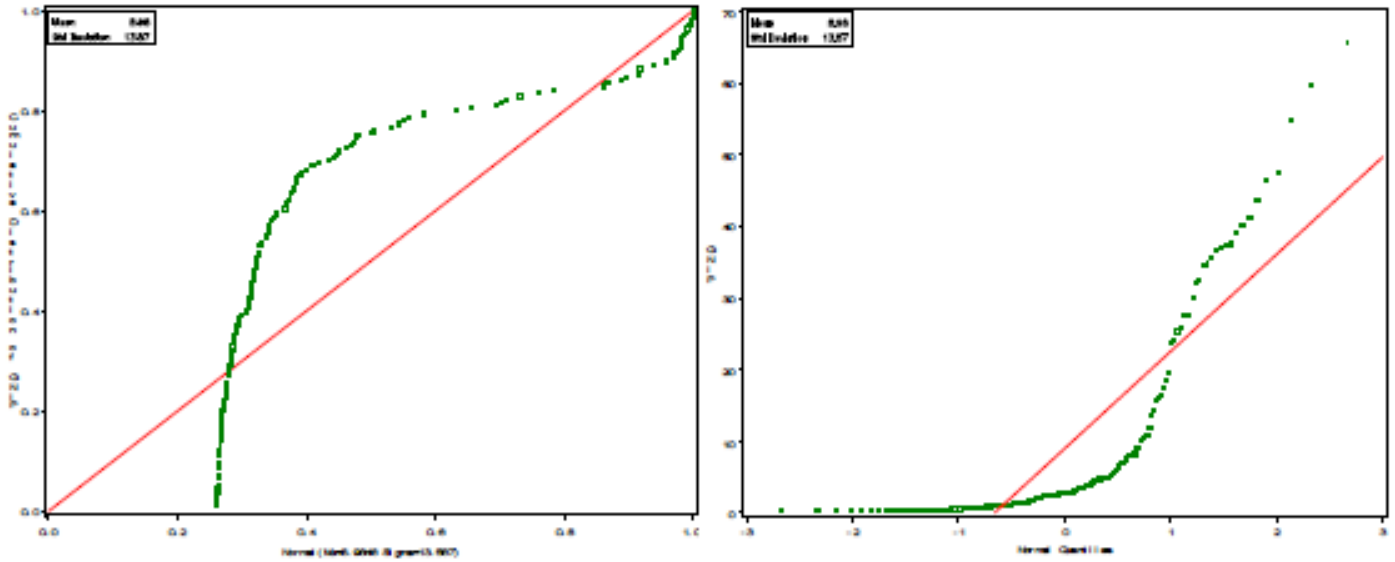
Figure 14. Stem-and-Leaf Plot, Box Plot, and Normally Probability Plot





The following P-P and Q-Q plots show that the data points are seriously deviated from the fitted line (Figure 15).

Figure 15. P-P plot and Q-Q Plot of a Non-normally Distributed Variable



4.2.3 Numerical Methods

Per capita gross national income has a mean of 8.9646 and a large variance of 184.0557. Its skewness and kurtosis-3 are 2.0495 and 3.6082, respectively, indicating that the variable is highly skewed to the right with a high peak and thin tails.

It is not surprising that the Shapiro-Wilk test rejected the null hypothesis; W is .6631 and p-value is less than .0001. Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests also report similar results.

Finally, the Jarque-Bera test returns 203.7717, which rejects the null hypothesis of normality at the .05 level ($p<.0000$).

$$164\left[\frac{2.04947469^2}{6}+\frac{3.60816725^2}{24}\right]\sim 203.77176(2)$$

To sum, we can conclude that the per capita gross national income is not normally distributed.

5. Testing Normality Using Stata

In Stata, you have to use individual commands to get specific statistics or draw various plots. This section contrasts normally distributed and non-normally distributed variables using graphical and numerical methods.

5.1 Graphical Methods

A histogram is the most widely used graphical method. The histograms of normally and non-normally distributed variables are presented in the introduction. The Stata `.histogram` command is followed by a variable name and options. The `normal` option adds a normal density curve to the histogram.

```
. histogram normal, normal
. histogram gnip, normal
```

Let us draw a stem-and-leaf plot using the `.stem` command. The stem-and-leaf plot of the randomly drawn `normal` shows a bell-shaped distribution (Figure 16).

```
. stem normal
```

Figure 16. Stem-and-Leaf Plot of a Normally Distributed Variable

```
Stem-and-leaf plot for normal

normal rounded to nearest multiple of .01
plot in units of .01

 -28* | 4
 -27* |
 -26* |
 -25* | 9
 -24* | 8
 -23* | 9
 -22* | 40
 -21* | 93221
 -20* | 8650
 -19* | 8842
 -18* | 875200
 -17* | 94
 -16* | 9987550
 -15* | 97643320
 -14* | 87755432110
 -13* | 98777655433210
 -12* | 8866666433210
 -11* | 987774332210
 -10* | 875322
  -9* | 88887665542210
  -8* | 99988777533110
  -7* | 77766544100
  -6* | 998332
  -5* | 999888776544332211l0
  -4* | 99987666554444333 21
  -3* | 887666544333222110 0
  -2* | 99998876655554443332211110 0
  -1* | 88887777766555444332222211l0
  -0* | 99887776655433333111
   0* | 01233344445669
   1* | 0111222333445666778
   2* | 0001234444556889999
   3* | 1133444556667899
   4* | 014455667777
   5* | 00112334556888
   6* | 0001123668899
   7* | 00233466799999
   8* | 1122334667889
   9* | 012445666778889
  10* | 1133457799
  11* | 1222334445689
  12* | 122233489
  13* | 26889
  14* | 2777799
  15* | 00112459
  16* | 1347
  17* | 02467
  18* | 358
  19* | 03556
  20* |
  21* | 5
  22* | 1
  23* |
  24* | 22
  25* | 1
```

By contrast, per capita gross national income is highly skewed to the right, having most observations within $10,000 (Figure 17).

```
. stem gnip
```

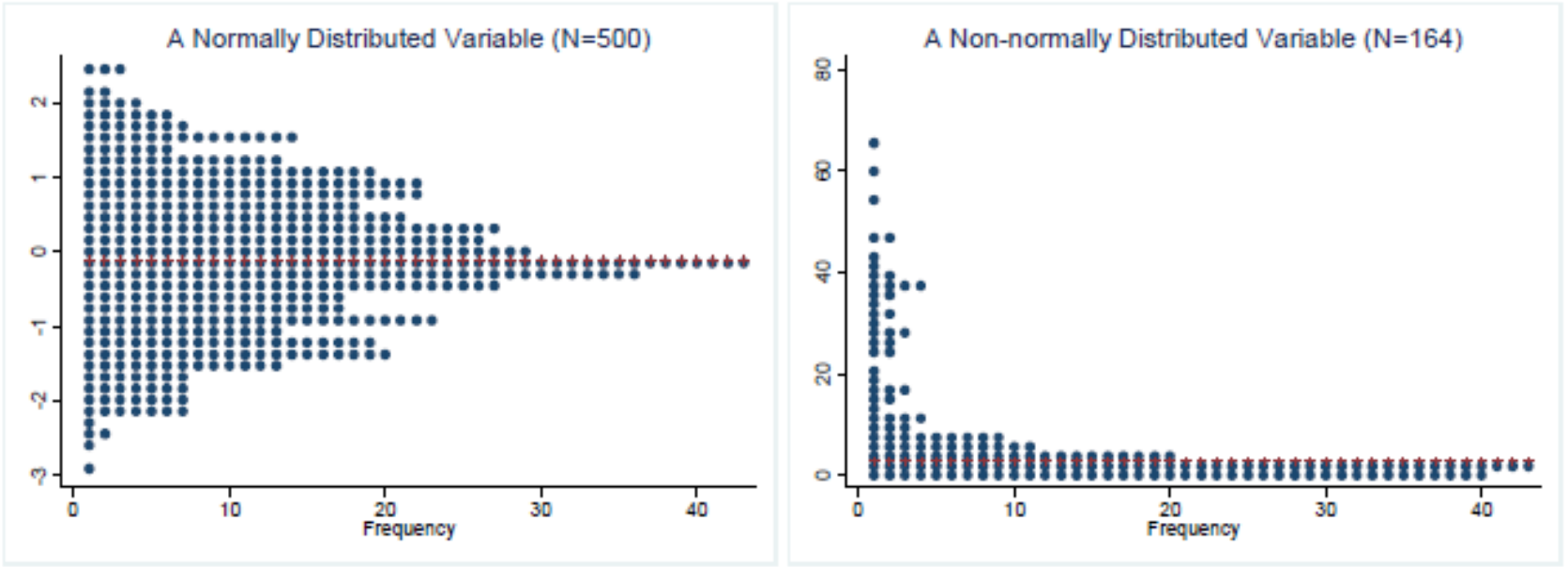
Figure 17. Stem-and-Leaf Plot of a Non-normally Distributed Variable

```
Stem-and-leaf plot for gnip

gnip rounded to nearest multiple of .1
plot in units of .1

  0** | 03,03,03,03,03,03,03,03,04,04,04,04,04,04,04,04,04,04,05,05, ... (64)
  0** | 21,22,23,23,23,24,24,24,24,25,25,25,26,26,26,27,28,28,28,28, ... (34)
  0** | 44,45,45,46,46,47,48,48,50,50,50,52,53,55,59
  0** | 62,68,71,71,73,76,79
  0** | 81,82,83,91,91
  1** | 00,04,07,09,18
  1** | 36
  1** | 44,58
  1** | 62,65,74
  1** | 86,97
  2** |
  2** | 38
  2** | 40,54
  2** | 60,75,77,78
  2** |
  3** | 00
  3** | 22,26
  3** | 46,48,57
  3** | 66,70,75,76
  3** | 90
  4** | 02,11
  4** | 37
  4** |
  4** | 63,74
  4** |
  5** |
  5** |
  5** | 49
  5** |
  5** | 96
  6** |
  6** |
  6** | 56
```

The `.dotplot` command generates a dot plot, very similar to the stem-and leaf plot, in a descending order (Figure 18).

```
. dotplot normal
. dotplot gnip
```

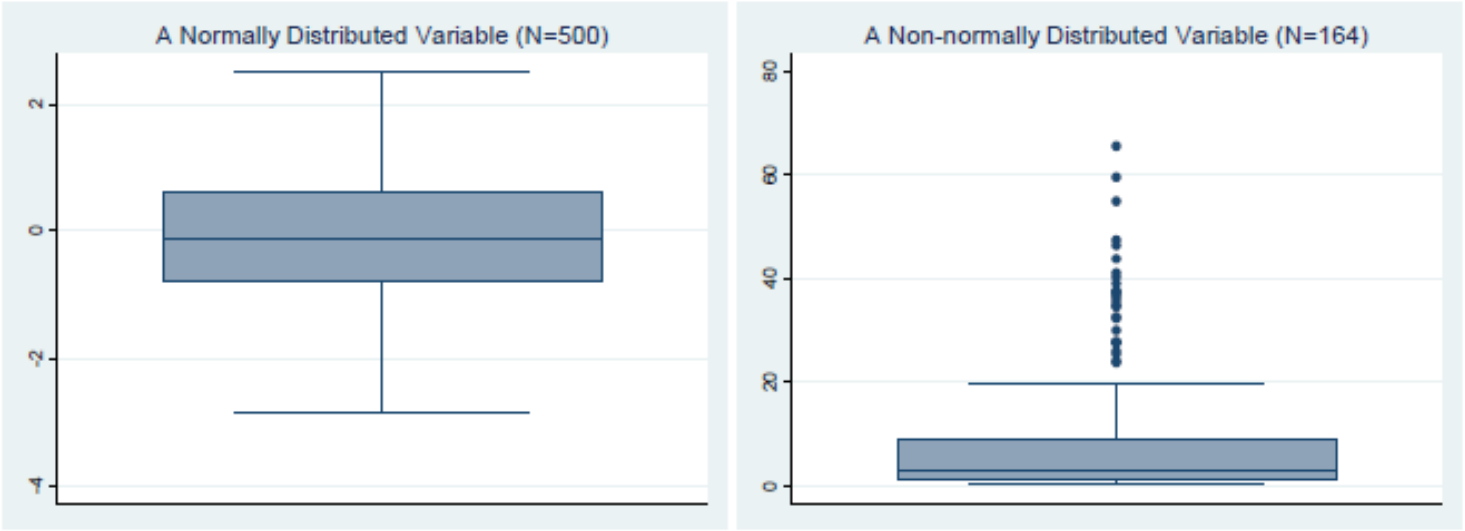
Figure 18. Dotplots of Normally and Non-normally Distributed Variables



The `.graph box` command draws a box plot. In the left plot of Figure 19, the shaded box represents the 25th percentile, median, and 75th percentile, which are symmetrically arranged. The right plot has an asymmetric box with many outliers beyond the adjacent maximum line.

```
. graph box normal
. graph box gnip
```

Figure 19. Box plots of Normally and Non-normally Distributed Variables

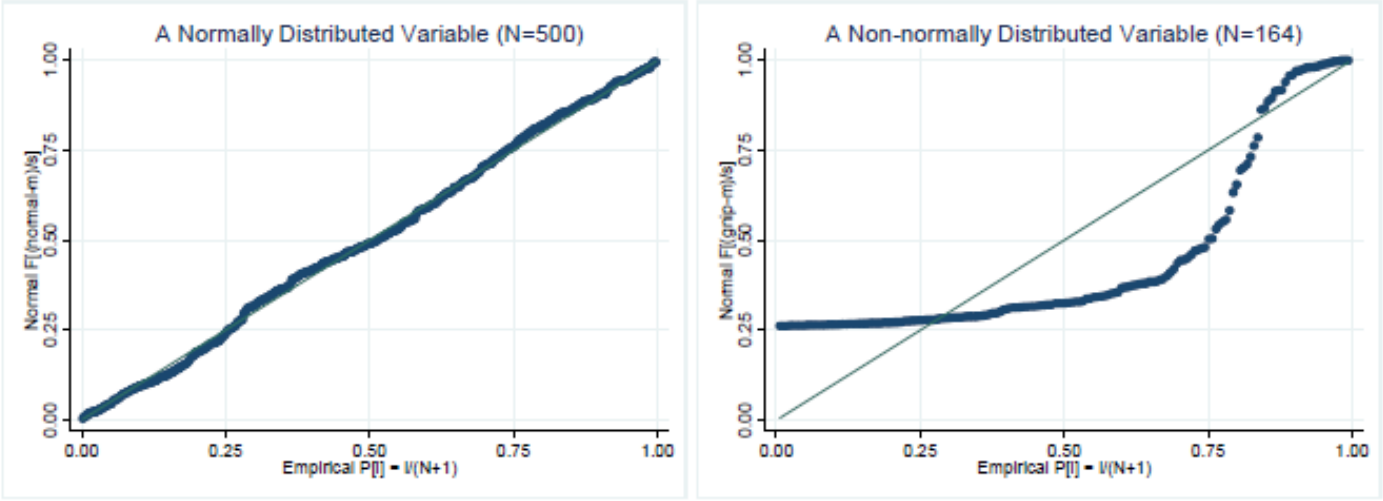


The `.pnorm` command produces standardized normal P-P plot. The left plot shows almost no deviation from the line, while the right depicts an s-shaped curve that is largely deviated from the fitted line. In Stata, a P-P plot has the cumulative distribution of an empirical variable on the x axis and the theoretical normal distribution on the y axis.[9]

```
.pnorm normal
.pnorm gnip
```

[9] In SAS, these distributions are located reversely.

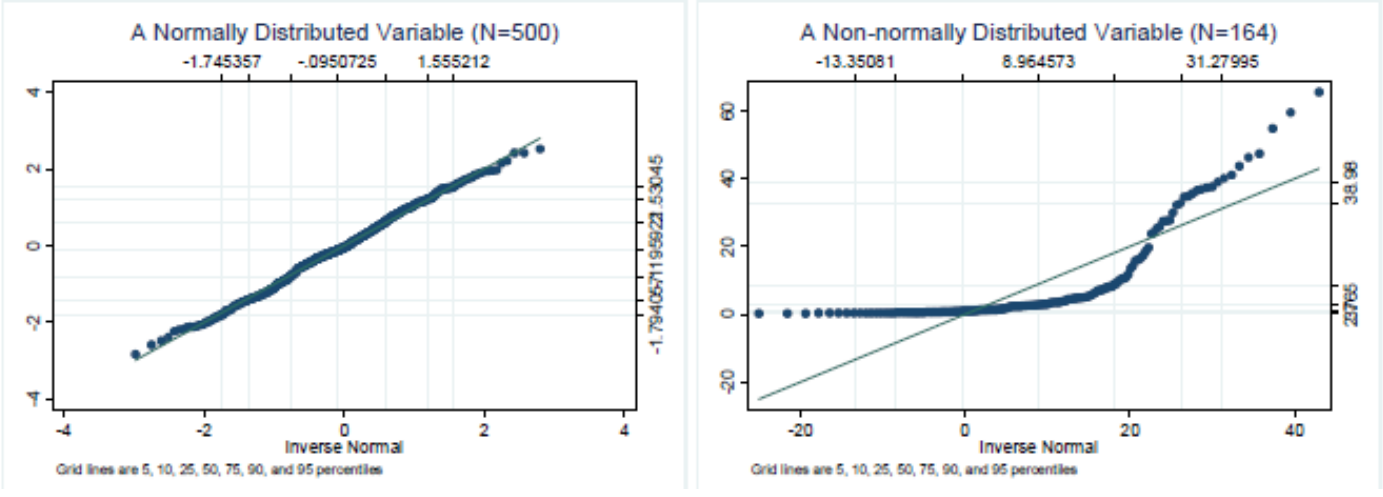
Figure 20. P-P plots of Normally and Non-normally Distributed Variables



The `.qnorm` command produces a standardized normal Q-Q plot. The following Q-Q plots show a similar pattern that P-P plots do (Figure 21). In the right plot, data points are systematically deviated from the straight fitted line.

```
.qnorm normal
.qnorm gnip
```

Figure 21. Q-Q plots of Normally and Non-normally Distributed Variables



5.2 Numerical Methods

Let us first get summary statistics using the `.summarize` command. The `detail` option lists various statistics including mean, standard deviation, minimum, and maximum. Skewness and kurtosis of a randomly drawn variable are respectively close to 0 and 3, implying normality. Per capital gross national income has large skewness of 2.03 and kurtosis of 6.46, being skewed to the right with a high peak and flat tails.

```
. summarize normal, detail
```

```
                            normal
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -2.219479      -2.837418
 5%    -1.794057      -2.590393
10%    -1.413548      -2.478296       Obs                 500
25%     -.805191      -2.391266       Sum of Wgt.         500

50%    -.1195922                      Mean           -.0950725
                      Largest         Std. Dev.       1.003302
75%     .6125385      2.211093
90%     1.215211      2.421139       Variance        1.006614
95%      1.53045      2.421713       Skewness       -.0203109
99%     2.055464      2.511694       Kurtosis        2.593181
```

```
. sum gnip, detail
```

```
                            gnip
-------------------------------------------------------------
      Percentiles      Smallest
 1%          .29            .29
 5%          .37            .29
10%          .45            .31       Obs                 164
25%         .955            .33       Sum of Wgt.         164

50%         2.765                     Mean           8.964573
                      Largest         Std. Dev.      13.56679
75%          8.68          47.39
90%          32.6          54.93      Variance       184.0577
95%         38.98          59.59      Skewness       2.030682
99%         59.59          65.63      Kurtosis       6.462734
```

The `.tabstat` command is vary useful to produce descriptive statistics in a table form. The `column(variable)` option lists statistics vertically (in table rows). The command for the variable `normal` is skipped.

```
. tabstat gnip, stats(n mean sum max min range sd var semean skewness kurtosis ///
  median p1 p5 p10 p25 p50 p75 p90 p95 p99 iqr q) column(variable)
```

stats	normal
N	500
mean	-.0950725
sum	-47.53624
max	2.511694
min	-2.837418
range	5.349112
sd	1.003302
variance	1.006614
se(mean)	.044869
skewness	-.0203109
kurtosis	2.593181
p50	-.1195922
p1	-2.219479
p5	-1.794057
p10	-1.413548
p25	-.805191
p50	-.1195922
p75	.6125385
p90	1.215211
p95	1.53045
p99	2.055464
iqr	1.41773
p25	-.805191
p50	-.1195922
p75	.6125385

stats	gnip
N	164
mean	8.964573
sum	1470.19
max	65.63
min	.29
range	65.34
sd	13.56679
variance	184.0577
se(mean)	1.059388
skewness	2.030682
kurtosis	6.462734
p50	2.765
p1	.29
p5	.37
p10	.45
p25	.955
p50	2.765
p75	8.68
p90	32.6
p95	38.98
p99	59.59
iqr	7.725
p25	.955
p50	2.765
p75	8.68

Now let us conduct statistical tests of normality. Stata provide three testing methods: Shapiro-Wilk test, Shapiro-Francia test, and Skewness-Kurtosis test. The `.swilk` and `.sfrancia` commands respectively conduct the Shapiro-Wilk and Shapiro-Francia tests. Both tests do not

reject normality of the randomly drawn variable and reject normality of per capita gross national income.

```
. swilk normal

                  Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+--------------------------------------------------
      normal |    500    0.99556      1.492      0.962  0.16804

. sfrancia normal

                  Shapiro-Francia W' test for normal data
    Variable |    Obs        W'         V'         z     Prob>z
-------------+--------------------------------------------------
      normal |    500    0.99645      1.273      0.541  0.29412

. swilk gnip

                  Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+--------------------------------------------------
        gnip |    164    0.66322     42.309      8.530  0.00000

. sfrancia gnip

                  Shapiro-Francia W' test for normal data
    Variable |    Obs        W'         V'         z     Prob>z
-------------+--------------------------------------------------
        gnip |    164    0.66365     45.790      7.413  0.00001
```

Stata's `.sktest` command conducts the Skewness-Kurtosis test that is conceptually similar to the Jarque-Bera test. The `noadjust` option suppresses the empirical adjustment made by Royston (1991). The following S-K tests do not reject normality of a randomly drawn variable at the .05 level but surprisingly reject the null hypothesis at the .1 level.

```
. sktest normal

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)     Prob>chi2
-------------+-------------------------------------------------------
      normal |       0.851          0.027            4.93        0.0850

. sktest normal, noadjust

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)      chi2(2)     Prob>chi2
-------------+-------------------------------------------------------
      normal |       0.851          0.027            4.93        0.0850
```

Like the Shapiro-Wilk and Shapiro-Francia tests, both S-K tests below reject the null hypothesis that per capita gross national income is normally distributed at the .01 significance level.

```
. sktest gnip

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)     Prob>chi2
-------------+-------------------------------------------------------
        gnip |       0.000          0.000           55.33        0.0000
  . sktest gnip, noadjust

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)      chi2(2)     Prob>chi2
-------------+-------------------------------------------------------
        gnip |       0.000          0.000           75.39        0.0000
```

The Jarque-Bera statistic of `normal` is 3.4823 = 500*(-.0203109^2/6+(2.593181-3)^2/24), which is not large enough to reject the null hypothesis (p<.1753). The Jarque-Bera statistic of the per capita gross national income is 194.6489 = 164*(2.030682^2/6+(6.462734-3)^2/24). This large chi-squared rejects the null hypothesis (p<.0000). The Jarque-Bera test appears to be more reliable than the Stata S-K test (see Table 4).

In conclusion, graphical methods and numerical methods provide sufficient evidence that per capita gross national income is not normally distributed.

6. Testing Normality Using SPSS

SPSS has the DESCRIPTIVES and FREQUENCIES commands to produce descriptive statistics. DESCRIPTIVES is usually applied to continuous variables, but FREQUENCIES is also able to produce various descriptive statistics in addition to frequency tables. The IGRAPH command draws histogram and box plots. The PPLOT command produces (detrended) P-P and Q-Q plots.

The EXAMINE command can produce both descriptive statistics and various plots, such as a stem-leaf-plot, histogram, box plot, (detrended) P-P plot, and (detrended) Q-Q plot. EXAMINE also performs the Kolmogorov-Smirnov and Shapiro-Wilk tests for normality.

6.1 A Normally Distributed Variable

DESCRIPTIVES summarizes interval or continuous variables and FREQUENCIES reports frequency tables of discrete variables and summary statistics. The /STATISTICS subcommand in both commands specify statistics to be produced.

The following DESCRIPTIVES command reports the number of observations, sum, mean, variance, standard deviation of `normal`.[10] The mean of -.10 and standard deviation 1 implies that the variable is normally distributed.

```
DESCRIPTIVES VARIABLES=normal
        /STATISTICS=MEAN SUM STDDEV VARIANCE.
```

Descriptive Statistics

	N	Sum	Mean	Std. Deviation	Variance
normal	500	-47.54	-.0951	1.00330	1.007
Valid N (listwise)	500				

The following FREQUENCIES produces various statistics of `normal`, a frequency table, and a histogram.[11] Since `normal` is continuous, its frequency table is long and thus skipped here. The /HISTOGRAM subcommand draws a histogram, which is the same as what the GRAPH command in the next page produces.

```
FREQUENCIES VARIABLES=normal /NTILES= 4
        /STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM SEMEAN MEAN MEDIAN MODE
                    SUM SKEWNESS SESKEW KURTOSIS SEKURT
        /HISTOGRAM
        /ORDER= ANALYSIS.
```

Statistics

[10] In order to execute this command, open a syntax window , copy and paste the syntax into the window, and then click Run menu. Alternatively, click Analysis→ Descriptive Statistics→Descriptives and provide a variable of interest.

[11] Click Analysis→ Descriptive Statistics→Frequencies and then specify statistics using the **Statistics** option.

normal

N	Valid	500.000
	Missing	.000
Mean		-.095
Std. Error of Mean		.045
Median		-.120
Mode		-2.837[a]
Std. Deviation		1.003
Variance		1.007
Skewness		-.020
Std. Error of Skewness		.109
Kurtosis		-.399
Std. Error of Kurtosis		.218
Range		5.349
Minimum		-2.837
Maximum		2.512
Sum		-47.536
Percentiles	25	-.807
	50	-.120
	75	.613

a. Multiple modes exist. The smallest value is shown

The variable has a mean -.10 and a unit variance. The median -.120 is very close to the mean. The kurtosis-3 is -.399 and skewness is -.020.

6.1.1 Graphical Methods

Like the /HISTOGRAM subcommand of FREQUENCIES, the GRAPH command draws a histogram of the variable `normal` (left plot in Figure 22).[12]

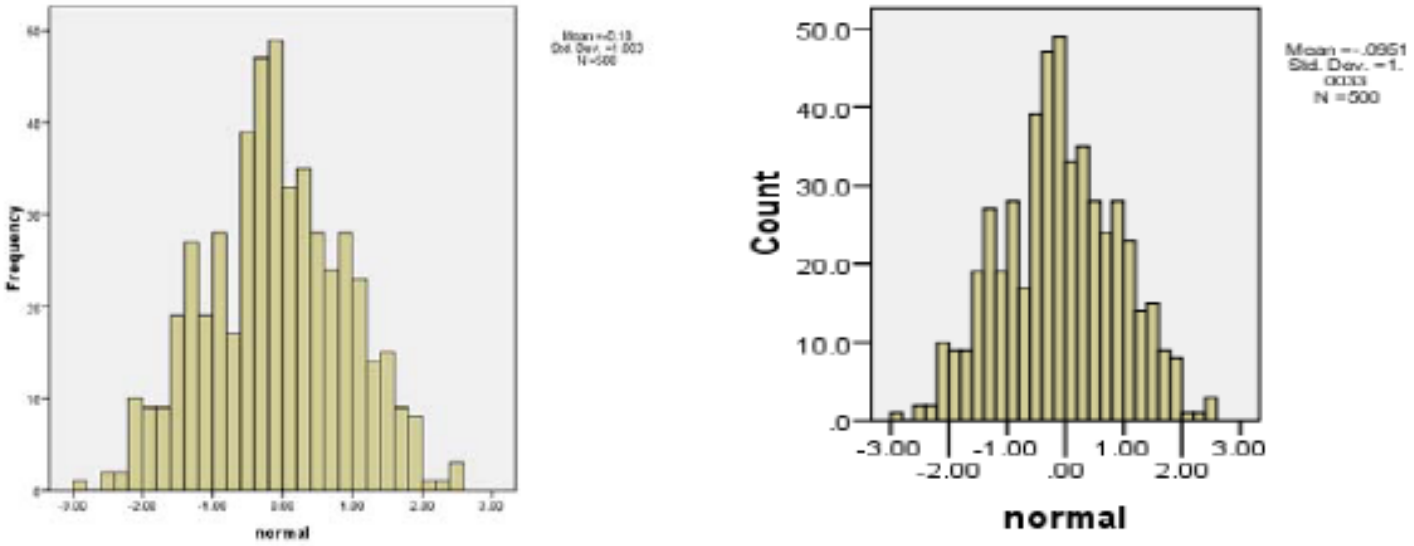
```
GRAPH /HISTOGRAM=normal.
```

The IGRAPH command can produce a similar histogram (right plot in Figure 22) but its syntax appears to be messy.[13] Two histograms report mean -.1 and standard deviation 1 on the right top corner and suggest that the variable is normally distributed.

```
IGRAPH /VIEWNAME='Histogram'
       /X1 = VAR(normal) TYPE = SCALE
       /Y = $count /COORDINATE = VERTICAL
       /X1LENGTH=3.0 /YLENGTH=3.0
       /X2LENGTH=3.0
       /CHARTLOOK='NONE'
       /Histogram SHAPE = HISTOGRAM CURVE = OFF X1INTERVAL AUTO X1START = 0.
```

[12] Click Graphs→Legacy Dialogs→Histogram.
[13] Click Graphs→Legacy Dialogs→Interactive→Histogram.

Figure 22. Histogram of a Normally Distributed Variable



The EXAMINE command can produce descriptive statistics as well as a stem-and-leaf plot and a box plot (Figure 23 and 24).[14] The /PLOT subcommand with STEMLEAF and BOXPLOT draws two plots that is very similar to the histogram in Figure 22.

```
EXAMINE VARIABLES=normal
        /PLOT BOXPLOT STEMLEAF
        /COMPARE GROUP
        /STATISTICS DESCRIPTIVES
        /CINTERVAL 95
        /MISSING LISTWISE
        /NOTOTAL.
```

[14] Click Analyze→Descriptive Statistics→Explore, and then include the variable you want to examine.

Figure 23. Stem-and-Leaf Plot of a Normally Distributed Variable

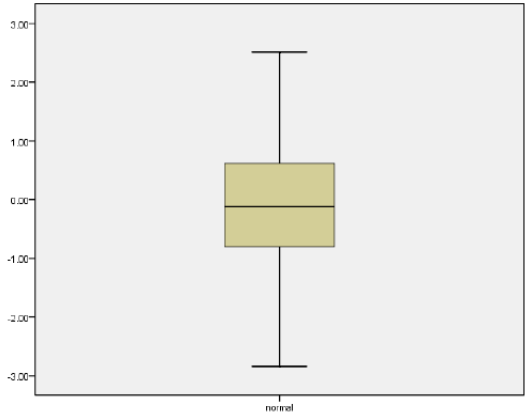
normal Stem-and-Leaf Plot

Frequency	Stem &	Leaf
2.00	-2 .	&
13.00	-2 .	00111&
27.00	-1 .	555566678899
56.00	-1 .	000111111222222333333344444
64.00	-0 .	55555555566677777888888899999
116.00	-0 .	00000000001111111111111222222222222233333333334444444444
80.00	0 .	00000001111111112222222223333333344444
68.00	0 .	55555555666667777777888888999999
46.00	1 .	000001111112222334444
23.00	1 .	55566778899
4.00	2 .	4&
1.00	2 .	&

Stem width: 1.00
Each leaf: 2 case(s)

& denotes fractional leaves.

Figure 24. Box Plot of a Normally Distributed Variable

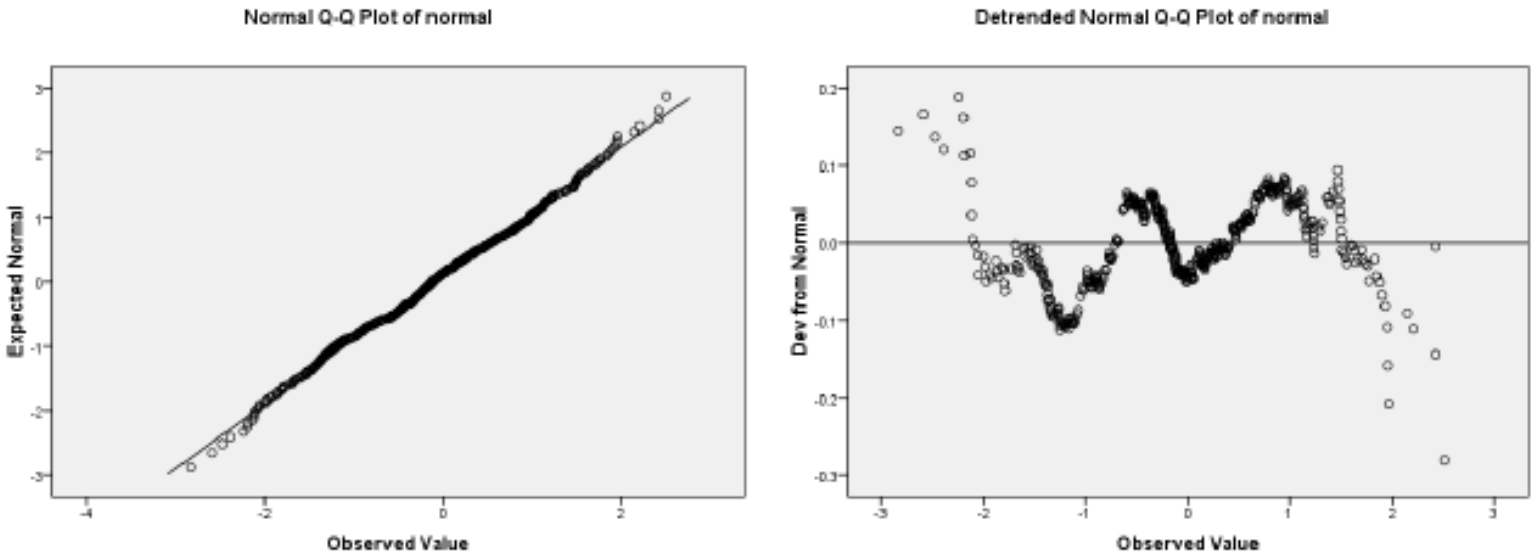


The both extremes (i.e., minimum and maximum), the 25th, 50th, and 75th percentiles are symmetrically arranged in the box plot.

EXAMINE also produces a histogram and normal Q-Q plot and detrended normal Q-Q plot using HISTOGRAM and NPPLOT option (Figure 25).[15] NPPLOT conducts normality test and draw the two Q-Q plots.

```
EXAMINE VARIABLES=normal
        /PLOT HISTOGRAM NPPLOT
        /COMPARE GROUP /STATISTICS DESCRIPTIVES
        /CINTERVAL 95 /MISSING LISTWISE /NOTOTAL.
```

Figure 25. Q-Q and Detrended Q-Q Plots of a Normally Distributed Variable

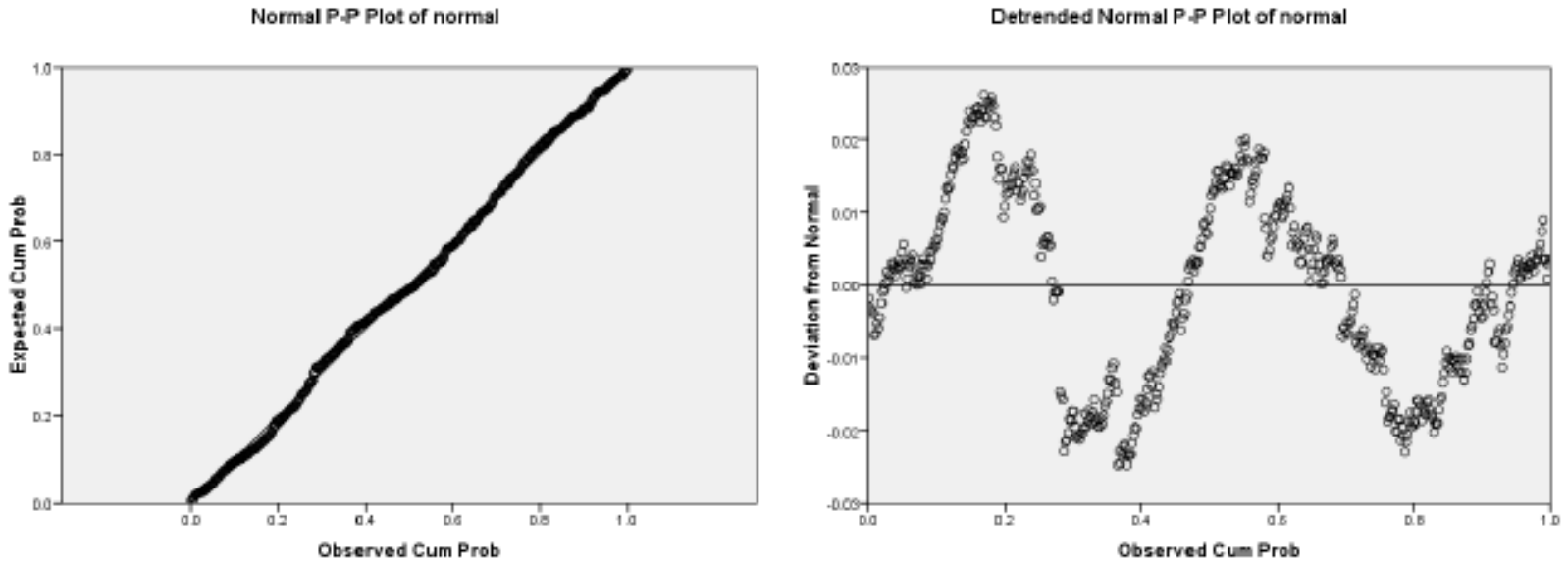


[15] In the **Explore** dialog box, choose **Plots** option and then check **Normality plots with tests** option.

The PPLOT command produces P-P and Q-Q plots as well.[16] The /TYPE subcommand chooses either P-P or Q-Q plot and /DIST specifies a probability distribution (e.g., the standard normal distribution). The following PPLOT command draws normal P-P and detrended normal P-P plots (Figure 26); the output of other descriptive statistics is skipped here.

```
PPLOT /VARIABLES=normal
      /NOLOG /NOSTANDARDIZE
      /TYPE=Q-Q /FRACTION=BLOM /TIES=MEAN /DIST=NORMAL.
```

Figure 26. P-P and Detrended P-P Plots of a Normally Distributed Variable



The following PPLOT command draws normal Q-Q and detrended normal Q-Q plots of the variable (see Figure 25).

```
PPLOT /VARIABLES=normal
      /NOLOG /NOSTANDARDIZE
      /TYPE=Q-Q /FRACTION=BLOM /TIES=MEAN /DIST=NORMAL.
```

Both P-P and Q-Q plots show no significant deviation from the fitted line. As in Stata, the normal Q-Q plot and detrended Q-Q plot has observed quantiles on the X axis and normal quantiles on the Y axis.

6.1.2 Numerical Methods

EXAMINE has the /PLOT NPPLOT subcommand to test normality of a variable. This command produces descriptive statistics (/STATISTICS DESCRIPTIVES), outliers (EXTREME), draws a normal Q-Q plot (/PLOT NPPLOT), and performs the Kolmogorov-Smirnov and Shapiro-Wilk tests.

```
EXAMINE VARIABLES=normal
       /PLOT NPPLOT
       /STATISTICS DESCRIPTIVES EXTREME
       /CINTERVAL 95 /MISSING LISTWISE /NOTOTAL.
```

Case Processing Summary

[16] In SPSS 16.0, you may not see P-P and Q-Q under the Graphs menu, which were available in previous versions.

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
normal	500	100.0%	0	.0%	500	100.0%

Descriptives

			Statistic	Std. Error
normal	Mean		-.0951	.04487
	95% Confidence Interval for Mean	Lower Bound	-.1832	
		Upper Bound	-.0069	
	5% Trimmed Mean		-.0933	
	Median		-.1196	
	Variance		1.007	
	Std. Deviation		1.00330	
	Minimum		-2.84	
	Maximum		2.51	
	Range		5.35	
	Interquartile Range		1.42	
	Skewness		-.020	.109
	Kurtosis		-.399	.218

Extreme Values

			Case Number	Value
Normal	Highest	1	332	2.51
		2	139	2.42
		3	325	2.42
		4	340	2.21
		5	119	2.15
	Lowest	1	29	-2.84
		2	204	-2.59
		3	73	-2.48
		4	391	-2.39
		5	393	-2.24

Since N is less than 2,000, we have to read the Shapiro-Wilk statistic and do not reject the null hypothesis of normality (p<.168). Like SAS, SPSS reports the same Kolmogorov-Smirnov statistic of .027, but it provides an adjusted p-value of .200, a bit larger than the .150 that SAS reports.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Normal	.027	500	.200*	.996	500	.168

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

6.2 A Non-normally Distributed Variable

Let us consider per capita national gross income that is not normally distributed.

6.2.1 Graphical Methods

The following EXAMINE command produce the histogram, stem-and-leaf plot, and box plot of a non-normally distributed variable `gnip`. The stem-and-leaf plot is skipped here.

```
EXAMINE VARIABLES=gnip
      /PLOT BOXPLOT STEMLEAF HISTOGRAM NPPLOT
      /STATISTICS DESCRIPTIVES EXTREME
      /CINTERVAL 95 /MISSING LISTWISE /NOTOTAL.
```

Figure 27 illustrates that the distribution is heavily skewed to the right and there exist many outliers beyond the extreme line in the box plot (right plot). The median and the 25th percentile are close to each other.

Figure 27. Histogram and Box Plot a Non-normally Distributed Variable

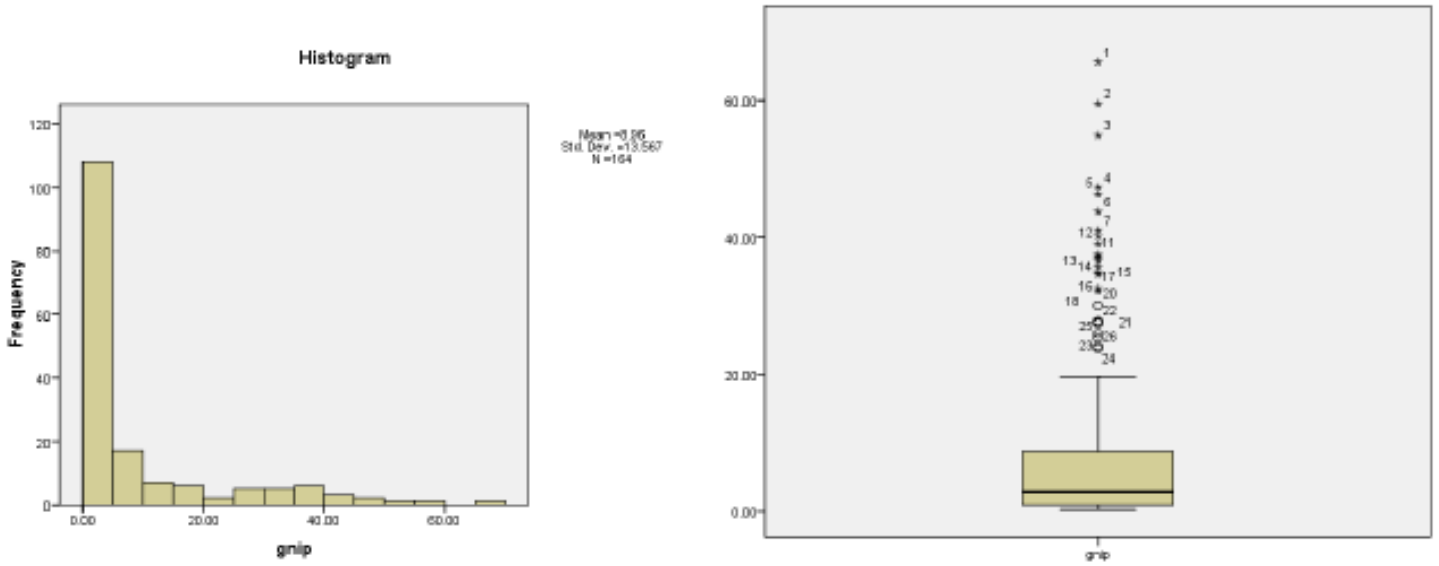
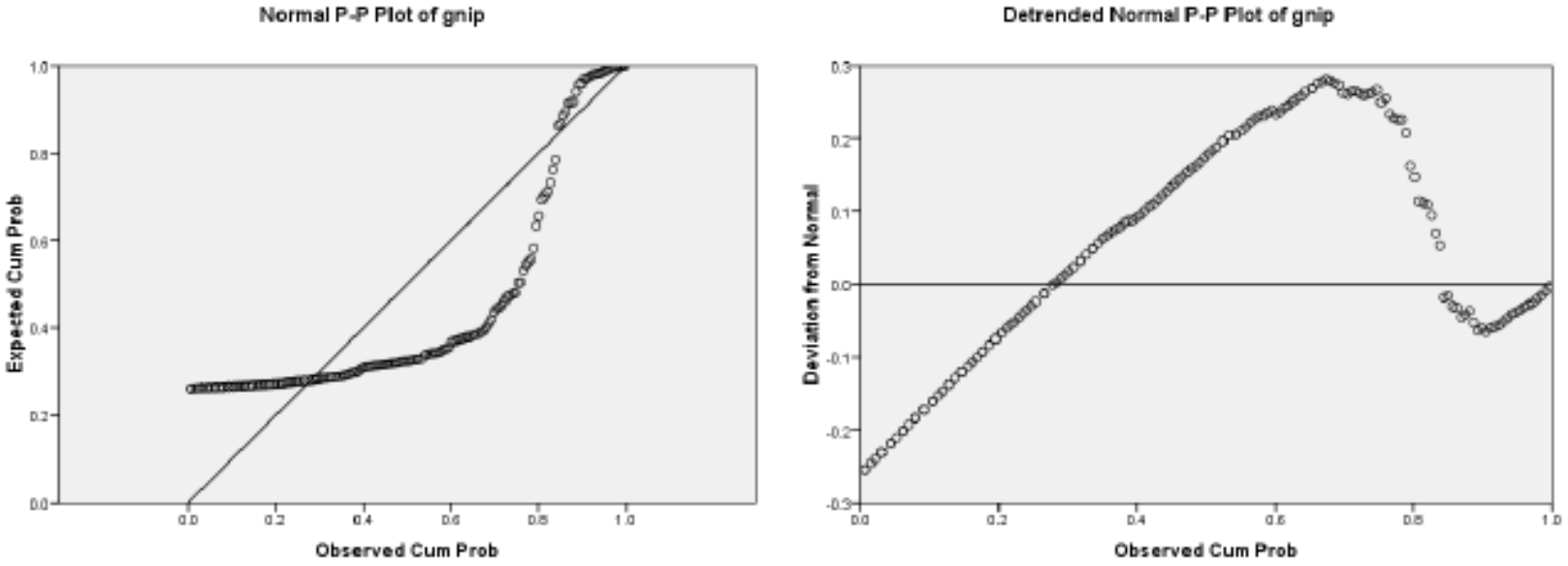


Figure 28 presents the P-P and detrended P-P plots where data points are significantly deviated from the straight fitted line.

```
PPLOT /VARIABLES=gnip
      /NOLOG /NOSTANDARDIZE
      /TYPE=P-P /FRACTION=BLOM
      /TIES=MEAN
     /DIST=NORMAL.
```

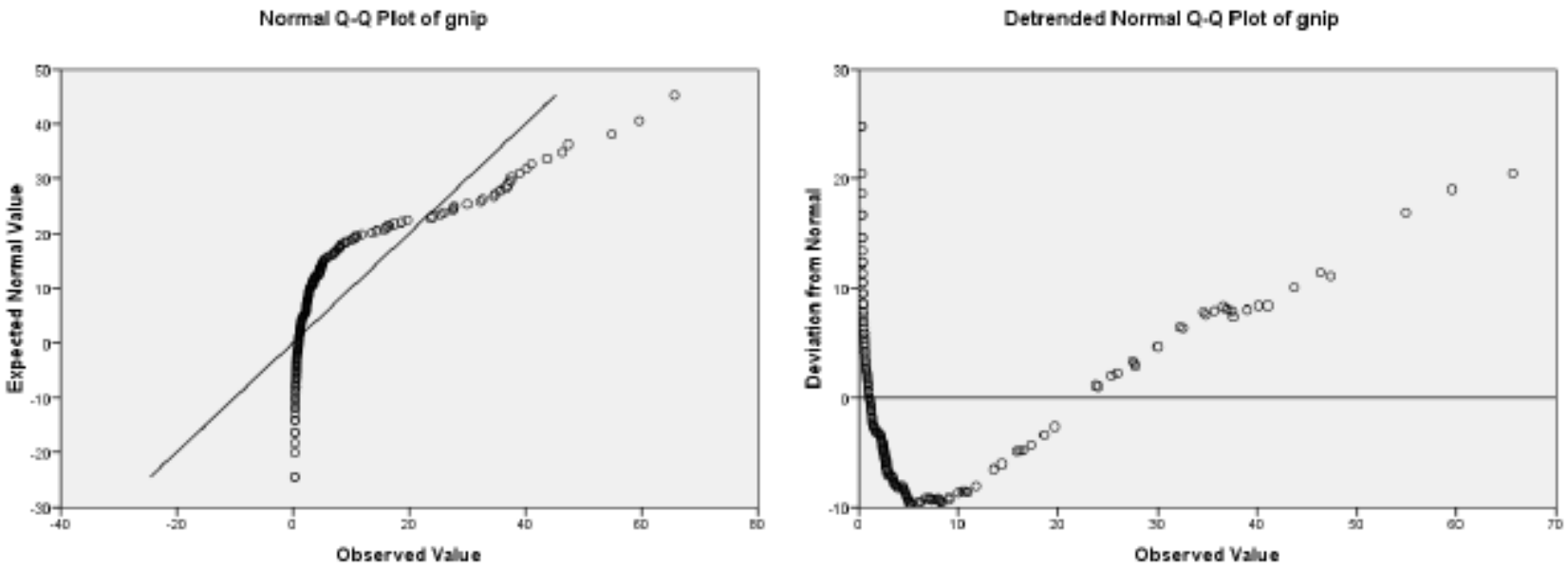
Figure 28. P-P and Detrended P-P Plots of a Non-normally Distributed Variable



The Q-Q and detrended Q-Q plots also show a significant deviation from the fitted line (Figure 26).

```
PPLOT /VARIABLES=gnip
      /NOLOG /NOSTANDARDIZE
      /TYPE=Q-Q /FRACTION=BLOM
      /TIES=MEAN
      /DIST=NORMAL.
```

Figure 29. Q-Q and Detrended Q-Q Plots of a Non-normally Distributed Variable



6.2.2 Numerical Methods

The descriptive statistics of gnip indicates that the variable is not normally distributed. There is a large gap between the mean of 8.9646 and the median of 2.7650. The skewness and kurtosis - 3 are 2.049 and 3.608, respectively. The variable appears severely skewed to the right with a higher peak and flat tails. The following tables are the output of the above EXAMINE command.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
gnip	164	100.0%	0	.0%	164	100.0%

Descriptives

			Statistic	Std. Error
gnip	Mean		8.9646	1.05939
	95% Confidence Interval for Mean	Lower Bound	6.8727	
		Upper Bound	11.0565	
	5% Trimmed Mean		7.1877	
	Median		2.7650	
	Variance		184.058	
	Std. Deviation		13.56679	
	Minimum		.29	
	Maximum		65.63	
	Range		65.34	
	Interquartile Range		7.92	
	Skewness		2.049	.190
	Kurtosis		3.608	.377

Extreme Values

			Case Number	Value
gnip	Highest	1	1	65.63
		2	2	59.59
		3	3	54.93
		4	4	47.39
		5	5	46.32
	Lowest	1	164	.29
		2	163	.29
		3	162	.31
		4	161	.33
		5	160	.34[a]

a. Only a partial list of cases with the value .34 are shown in the table of lower extremes.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
gnip	.284	164	.000	.663	164	.000

a. Lilliefors Significance Correction

The Shapiro-Wilk test rejects the null hypothesis of normality at the .05 level. The Jarque-Bera test also rejects the null hypothesis with a large statistic of 204. Its computation is skipped (see section 4.2.3). Based on a consistent result from both graphical and numerical methods, we can conclude the variable `gnip` is not normally distributed.

7. Conclusion

Univariate analysis is the first step of data analysis once a data set is ready. Various descriptive statistics provide valuable basic information about variables that is used to determine appropriate analysis methods to be employed.

Normality is commonly assumed in many statistical and economic methods, although often conveniently assumed in reality without any empirical test. Violation of this assumption will result in unreliable inferences and misleading interpretations.

There are graphical and numerical methods for conducting univariate analysis and normality tests (Table 1). Graphical methods produce various plots such as a stem-and-leaf plot, histogram, and a P-P plot that are intuitive and easy to interpret. Some are descriptive and others are theory-driven.

Numerical methods compute a variety of measures of central tendency and dispersion such as mean, median, quantile, variance, and standard deviation. Skewness and kurtosis provide clues to the normality of a variable. If skewness and kurtosis-3 are close to zero, the variable may be normally distributed. Keep in mind that SAS and SPSS report kurtosis-3, while Stata returns kurtosis itself.

If the skewness of a varialbe is larger than 0, the variable is skewed to the right with many observations on the left of the distribution; a negative skewness indicates many observations on the right. If kurtosis-3 is greater than 0 (or kurtosis is greater than 3), the distribution has a high peak and flat tails (third plot in Figure 8). If kurtosis is smaller than 3, the variable has a low peak and thick tails (first plot in Figure 9).

In addition to these descriptive statistics, there are formal ways to perform normality tests. The Shapiro-Wilk and Shapiro-Francia tests are proper when N is less than 2,000 and 5,000, respectively. The Kolmogorov-Smirnov, Cramer-vol Mises, and Anderson-Darling tests are recommended when N is large. The Jarque-Bera test, although not supported by most statistical software packages, is a consistent method of normality testing.

The SAS UNIVARIATE and CONTENTS procedures provide a variety of descriptive statistics and normality testing methods including Kolmogorov-Smirnov, Cramer-vol Mises, and Anderson-Darling tests (Table 5). These procedures produce stem-and-leaf, box plot, histogram, P-P plot, and Q-Q plot as well. Stata has various commands for univariate analysis and graphics. In particular, Stata supports the Shapiro-Francia test, a modification of the Shapiro-Wilk test, and the skewness-kurtosis test. But there is no command to conduct the Kolmogorov-Smirnov test for normality in Stata. SPSS can produce detrended P-P and Q-Q plots, and perform the Shapiro-Wilk and Kolmogorov-Smirnov tests with Lilliefors significance correction.

Appendix A: Data Sets

This document uses the following three variables.

1. Unemployment Rate of Illinois, Indiana, and Ohio in 2005

This unemployment rate is provided by Bureau of Labor Statistics. Actual data were downloaded from http://www.stats.indiana.edu/, Indiana Business Research Center of the Kelley School of Business, Indiana University.

```
. tabstat rate, stat(mean sd p25 median p75 skewness kurtosis) by(state)

Summary for variables: rate
     by categories of: state

 state |      mean        sd       p25       p50       p75  skewness  kurtosis
-------+--------------------------------------------------------------------------
    IL |  5.421569  .9242206       4.7      5.35         6  .6570033  3.946029
    IN |  5.641304  1.038929       4.9       5.5      6.35  .3416314  2.785585
    OH |    6.3625  1.458098       5.5       6.1      6.95  1.665322  8.043097
-------+--------------------------------------------------------------------------
 Total |  5.786879  1.214066         5      5.65       6.4   1.44809  8.383285
--------------------------------------------------------------------------------
```

2. A Randomly Drawn Variable

This variable includes 500 observations that were randomly drawn from the standard normal distribution with a seed of 1,234,567. The RANNOR() of SAS was used as a random number generator.

```
%LET n=500; %LET dataset=n500;

DATA masil.&dataset;
seed=1234567;
DO i=1 TO &n;
       normal=RANNOR(seed); OUTPUT;
END;
RUN;
```

```
. tabstat normal, stat(mean sd p25 median p75 skewness kurtosis)

    variable |      mean        sd       p25       p50       p75  skewness  kurtosis
-------------+--------------------------------------------------------------------------
      normal | -.0950725  1.003302  -.805191 -.1195922  .6125385 -.0203109  2.593181
--------------------------------------------------------------------------------------
```

3. Per Capita Gross National Income in 2005.

This data set includes per capita gross national incomes of 164 countries in the world that are provided by World Bank (http://web.worldbank.org/).

```
. tabstat gnip, stat(mean sd p25 median p75 skewness kurtosis)

    variable |      mean        sd       p25       p50       p75  skewness  kurtosis
-------------+--------------------------------------------------------------------------
        gnip |  8.964573  13.56679      .955     2.765      8.68  2.030682  6.462734
--------------------------------------------------------------------------------------
```