Základní pojmy matematické statistiky

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

b) Nechť $(X_1,Y_1), \ldots, (X_n,Y_n)$ jsou stochasticky nezávislé dvourozměrné náhodné vektory, které mají všechny stejné dvourozměrné rozložení $L_2(\vartheta)$. Řekneme, že $(X_1,Y_1), \ldots, (X_n,Y_n)$ je dvourozměrný náhodný výběr rozsahu n z dvourozměrného rozložení $L_2(\vartheta)$. (Číselné realizace $(x_1,y_1), \ldots, (x_n,y_n)$ náhodného výběru $(X_1,Y_1), \ldots, (X_n,Y_n)$ uspořádané do matice typu $2 \times n$ 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 $T = T(X_1, \ldots, X_n)$ náhodného výběru $X_1, \ldots, X_n$ (resp. $T = T(X_1,Y_1, \ldots, X_n,Y_n)$ náhodného výběru $(X_1,Y_1), \ldots, (X_n,Y_n)$) se nazývá (výběrová) statistika.

Definice důležitých statistik:

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

Onačme $M = \frac{1}{n}\sum_{i=1}^{n} X_i$ … 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 $F_n(x) = \frac{1}{n}\operatorname{card}\{i; X_i \leq x\}$

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.

c) Nechť $(X_1,Y_1), ..., (X_n,Y_n)$ je náhodný výběr z dvourozměrného rozložení o rozsahu n.

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, $R_{12} = \begin{cases} \frac{S_{12}}{S_1 S_2} \text{ pro } S_1 S_2 \neq 0 \\ 0 \text{ jinak} \end{cases}$ … výběrový koeficient korelace.

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.

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 (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 $X_1, \dots, X_{10}$. 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:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(2 + 1{,}8 + \dots + 2{,}2) = 2{,}06, s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - nm^2\right) = \frac{1}{9}\left(2^2 + 1{,}8^2 + \dots + 2{,}2^2 - 10 \cdot 2{,}06^2\right) = 0{,}0404$$

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

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.

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

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

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

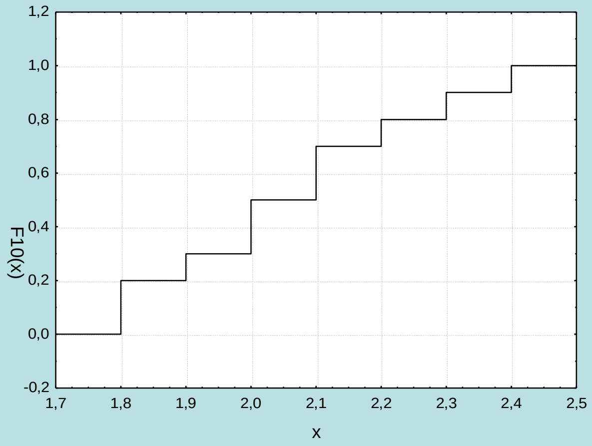
$$2 \le x < 2{,}1 : F_{10}(x) = \frac{5}{10} = 0{,}5$$

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

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

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

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



Bodové a intervalové odhady parametrů a parametrických funkcí

Vycházíme z náhodného výběru $X_1, ..., X_n$ z rozložení L(ϑ), které závisí na parametru ϑ. Množinu všech přípustných hodnot tohoto parametru označíme Ξ. Tato množina se nazývá parametrický prostor.
Např. je-li $X_1, ..., X_n$ náhodný výběr z rozložení N(μ,σ^2), pak $\vartheta = (\mu, \sigma^2)$ a v tomto případě parametrický prostor Ξ = $(-\infty, \infty) \times \langle 0, \infty)$.

Parametr ϑ neznáme a chceme ho odhadnout pomocí daného náhodného výběru (případně chceme odhadnout nějakou parametrickou funkci h(ϑ)).

Bodovým odhadem parametrické funkce h(ϑ) je statistika $T_n = T(X_1, ..., X_n)$, která nabývá hodnot blízkých h(ϑ), ať je hodnota parametru ϑ jakákoliv. Existují různé metody, jak konstruovat bodové odhady (např. metoda momentů či metoda maximální věrohodnosti, ale těmi se zde zabývat nebudeme) a také různé typy bodových odhadů. Omezíme se na odhady nestranné, asymptoticky nestranné a konzistentní.

Intervalovým odhadem parametrické funkce h(ϑ) rozumíme interval (D, H), jehož meze jsou statistiky
$D = D(X_1, ..., X_n)$, $H = H(X_1, ..., X_n)$ a který s dostatečně velkou pravděpodobností pokrývá h(ϑ), ať je hodnota parametru ϑ jakákoliv.

Typy bodových odhadů

Nechť $X_1, \ldots, X_n$ 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
$$\forall\vartheta \in \Xi\,\forall\varepsilon > 0: \lim_{n\to\infty} P\left(\left|T_n - h(\vartheta)\right| > \varepsilon\right) = 0.$$
(Význam konzistence spočívá v tom, že s rostoucím rozsahem výběru klesá pravděpodobnost, že odhad se bude realizovat „daleko“ od parametrické funkce $h(\vartheta)$.)

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

Vlastnosti důležitých statistik

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,

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

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

Znamená to, že M_n je nestranným odhadem μ, S_n^2 je nestranným odhadem σ^2, pro libovolné, ale pevně dané $x \in R$ je výběrová distribuční funkce $F_n(x)$ nestranným odhadem $\Phi(x)$.

Posloupnost $\{M_n\}_{n=1}^{\infty}$ je posloupnost konzistentních odhadů μ,

$\{S_n^2\}_{n=1}^{\infty}$ je posloupnost konzistentních odhadů σ^2,

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

b) Případ $r \geq 2$ 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 μ_1, …, μ_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 $S_*^2 = \dfrac{\sum_{j=1}^{r} (n_j - 1) S_j^2}{n - r}$ je nestranným odhadem rozptylu σ^2.

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 $n \geq 30$).

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.

Pojem intervalu spolehlivosti

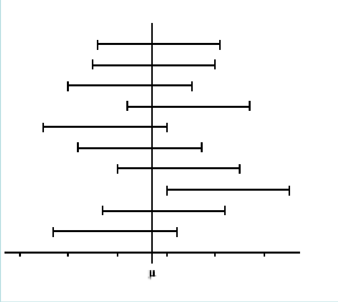
Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $L(\vartheta)$,
$h(\vartheta)$ je parametrická funkce,
$\alpha \in (0,1)$,
$D = D(X_1, \ldots, X_n)$, $H = H(X_1, \ldots, X_n)$ jsou statistiky.

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 $(-\infty, H)$ 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.

Postup při konstrukci intervalu spolehlivosti

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 $D < h(\vartheta) < H$.
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 – α. (Tvrzení, že (d,h) pokrývá $h(\vartheta)$ s pravděpodobností aspoň 1 – α je třeba chápat takto: jestliže mnohonásobně nezávisle získáme realizace $x_1, \dots, x_n$ náhodného výběru $X_1, \dots, X_n$ 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 – α.)

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-ti případech bude ležet parametr μ v intervalech spolehlivosti a asi v 5-ti případech interval spolehlivosti μ nepokryje.



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.

Příklad: Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $N(\mu,\sigma^2)$, 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 $w_{\alpha/2} = u_{\alpha/2} = -u_{1-\alpha/2}$, $w_{1-\alpha/2} = u_{1-\alpha/2}$.

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

Meze $100(1-\alpha)\%$ 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 $100(1-\alpha)\%$ levostranný interval spolehlivosti pro μ je $\left(M - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty\right)$ a pravostranný je $\left(-\infty, M + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}\right)$.

Dosadíme-li do vzorců pro dolní a horní mez číselnou realizaci m výběrového průměru M, dostaneme $100(1-\alpha)\%$ empirický interval spolehlivosti. Postup si ukážeme na následujícím numerickém příkladu.

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 $N(\mu, \sigma^2)$, kde μ neznáme a $\sigma^2 = 0{,}04$. Najděte 95% empirický interval spolehlivosti pro μ, a to a) oboustranný, b) levostranný, c) pravostranný.

Řešení:

Vypočteme realizaci výběrového průměru: $m = 2{,}06$. Riziko α je 0,05. V tabulkách najdeme kvantil $u_{0,975} = 1{,}96$ pro oboustranný interval spolehlivosti a kvantil $u_{0,95} = 1{,}64$ pro jednostranné intervaly spolehlivosti.

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.

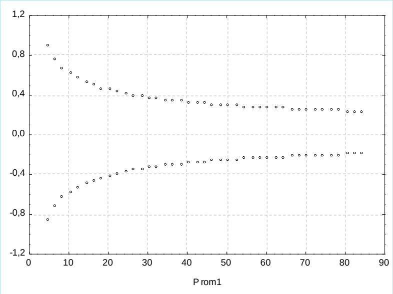
Šířka intervalu spolehlivosti

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

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.

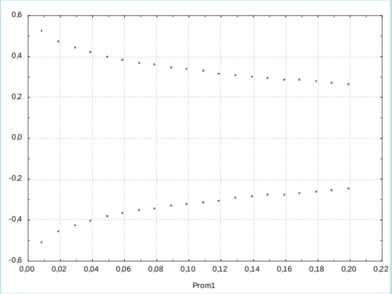
Ilustrace

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:



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

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 $X_1, ..., X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde σ^2 známe. Jaký musí být minimální rozsah výběru n, aby šířka 100(1-α)% 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.

Příklad: Hloubka moře se měří přístrojem, jehož systematická chyba je nulová a náhodné chyby měření mají normální rozložení se směrodatnou odchylkou σ = 1 m. Kolik měření je nutno provést, aby se hloubka stanovila s chybou nejvýše ± 0,25 m při spolehlivosti 0,95?

Řešení: Hledáme rozsah výběru tak, aby šířka 95% intervalu spolehlivosti pro střední hodnotu μ nepřesáhla 0,5 m. Přitom σ známe. Z předešlého příkladu vyplývá, že $n \geq \frac{4\sigma^2 u_{1-\alpha/2}{}^2}{\Delta^2} = \frac{4 \cdot 1{,}96^2}{0{,}5^2} = 61{,}4656$. Nejmenší počet měření je tedy 62.

Úvod do 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.

Nulová a alternativní hypotéza

Nechť $X_1, \ldots, X_n$ 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 alternativní hypotézu H_1: $h(\vartheta) > c$.
Testováním H_0 proti H_1 rozumíme rozhodovací postup založený na náhodném výběru $X_1, \ldots, X_n$, s jehož pomocí zamítneme či nezamítneme platnost nulové hypotézy.

Chyba 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 $\alpha = 0{,}05$, méně často 0,1 či 0,01). Pravděpodobnost chyby 2. druhu se značí β. Číslo $1-\beta$ 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 $1-\beta$, závisí na velikosti efektu, který se snažíme detekovat. Čím drobnější efekt, tím musí být větší rozsah náhodného výběru.

skutečnost	rozhodnutí	
	zdravý	nemocný
jsem zdravý	zdravý a neléčený	zdravý a léčený
jsem nemocný	nemocný a neléčený	nemocný a léčený

Testování pomocí kritického oboru

Najdeme statistiku $T_0 = T_0(X_1, ..., X_n)$, kterou nazveme 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).

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.

Pravděpodobnosti chyb 1. a 2. druhu nyní zapíšeme takto:

$P(T_0 \in W/H_0 \text{ platí}) = \alpha$, $P(T_0 \in V/H_1 \text{ platí}) = \beta$.

Stanovení kritického oboru pro danou hladinu významnosti α:

Označme t_{min} (resp. t_{max}) nejmenší (resp. největší) hodnotu testového kritéria.

Kritický obor v případě oboustranné alternativy má tvar

$W = \left(t_{min}, K_{\alpha/2}(T)\right\rangle \cup \left\langle K_{1-\alpha/2}(T), t_{max}\right)$, 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:

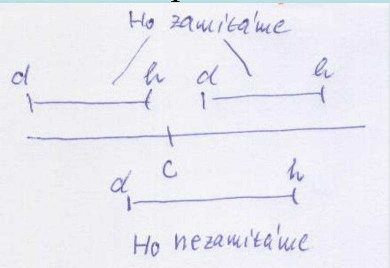
$W = \left(t_{min}, K_{\alpha}(T)\right\rangle$.

Kritický obor v případě pravostranné alternativy má tvar:

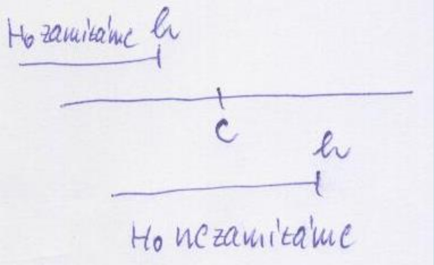
$W = \left\langle K_{1-\alpha}(T), t_{max}\right)$.

Testování pomocí intervalu spolehlivosti

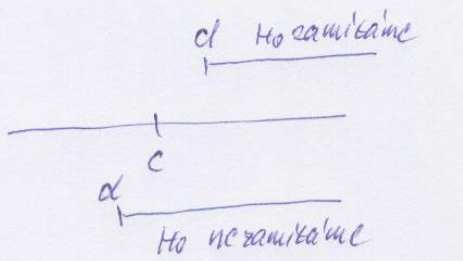
Sestrojíme 100(1-α)% empirický interval spolehlivosti pro parametrickou funkci h(ϑ). Pokryje-li tento interval hodnotu c, pak H_0 nezamítáme na hladině významnosti α, v opačném případě H_0 zamítáme na hladině významnosti α.
Pro test H_0 proti oboustranné alternativě sestrojíme oboustranný interval spolehlivosti.



Pro test H_0 proti levostranné alternativě sestrojíme pravostranný interval spolehlivosti.



Pro test H_0 proti pravostranné alternativě sestrojíme levostranný interval spolehlivosti.



Testování 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.

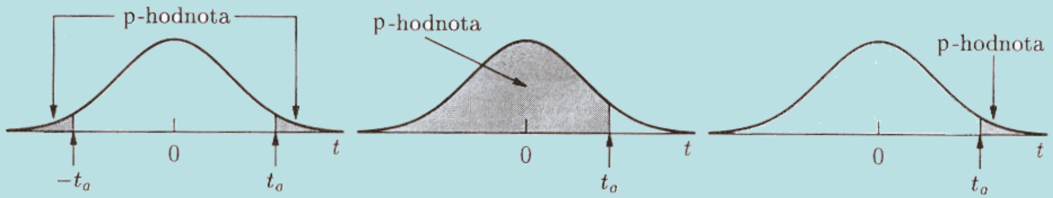
Způsob výpočtu p-hodnoty:

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 $p = P(T_0 \geq t_0)$.

Ilustrace významu p-hodnoty pro test nulové hypotézy proti oboustranné, levostranné a pravostranné alternativě:



(Zvonovitá křivka reprezentuje hustotu rozložení, kterým se řídí testové kritérium, je-li nulová hypotéza pravdivá.)

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.

Doporučený postup při testování hypotéz

1. 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.

2. Zvolíme hladinu významnosti α. Zpravidla volíme $\alpha = 0{,}05$, méně často 0,1 nebo 0,01.

3. Najdeme vhodné testové kritérium a na základě zjištěných dat vypočítáme jeho realizaci.

4.
a) Testujeme-li pomocí kritického oboru, pak ho stanovíme. 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.
b) Testujeme-li pomocí intervalu spolehlivosti, vypočteme empirický $100(1-\alpha)\%$ interval spolehlivosti pro parametrickou funkci $h(\vartheta)$. Pokud číslo c padne do tohoto intervalu, nulovou hypotézu nezamítáme na hladině významnosti α. V opačném případě nulovou hypotézu zamítáme na hladině významnosti α a přijímáme alternativní hypotézu.
c) Testujeme-li pomocí p-hodnoty, vypočteme ji a porovnáme ji s hladinou významnosti α. Jestliže $p \leq \alpha$, pak nulovou hypotézu zamítáme na hladině významnosti α a přijímáme alternativní hypotézu. Je-li $p > \alpha$, pak nulovou hypotézu nezamítáme na hladině významnosti α.

5. Na základě rozhodnutí, které jsme učinili o nulové hypotéze, provedeme 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.)

Příklad: 10 x 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 $N(\mu, 0{,}04)$. Nějaká teorie tvrdí, že $\mu = 1{,}95$.

1. Oboustranná alternativa

Proti nulové hypotéze H_0: $\mu = 1{,}95$ postavíme oboustrannou alternativu H_1: $\mu \neq 1{,}95$. Na hladině významnosti 0,05 testujte H_0 proti H_1 všemi třemi popsanými způsoby.

Řešení:

$m = \frac{1}{10}(2 + \ldots + 2{,}2) = 2{,}06$, $\sigma^2 = 0{,}04$, $n = 10$, $\alpha = 0{,}05$, $c = 1{,}95$

a) Test provedeme pomocí kritického oboru.

Pro úlohy o střední hodnotě normálního rozložení při známém rozptylu používáme pivotovou statistiku $U = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$.

Testové kritérium tedy bude

$T_0 = \frac{M-c}{\frac{\sigma}{\sqrt{n}}}$ a bude mít rozložení N(0, 1), pokud je nulová hypotéza pravdivá. Vypočítáme realizaci testového kritéria: $t_0 = \frac{2{,}06-1{,}95}{\frac{0{,}2}{\sqrt{10}}} = 1{,}74$. Stanovíme kritický obor:

$W = (t_{min}, K_{\alpha/2}(T)\rangle \cup \langle K_{1-\alpha/2}(T), t_{max}) = (-\infty, u_{\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty) = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty) = (-\infty, -u_{0.975}\rangle \cup \langle u_{0,975}, \infty) = (-\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty)$.

Protože $1{,}74 \notin W$, H_0 nezamítáme na hladině významnosti 0,05.

b) Test provedeme pomocí intervalu spolehlivosti.

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

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

V našem případě dostáváme:

$$d = 2{,}06 - \frac{0{,}2}{\sqrt{10}} u_{0{,}975} = 2{,}06 - \frac{0{,}2}{\sqrt{10}} .1{,}96 = 1{,}936,$$

$$h = 2{,}06 + \frac{0{,}2}{\sqrt{10}} u_{0{,}975} = 2{,}06 + \frac{0{,}2}{\sqrt{10}} .1{,}96 = 2{,}184.$$

Protože $1{,}95 \in (1{,}936;\ 2{,}184)$, H_0 nezamítáme na hladině významnosti 0,05.

c) Test provedeme pomocí p-hodnoty.

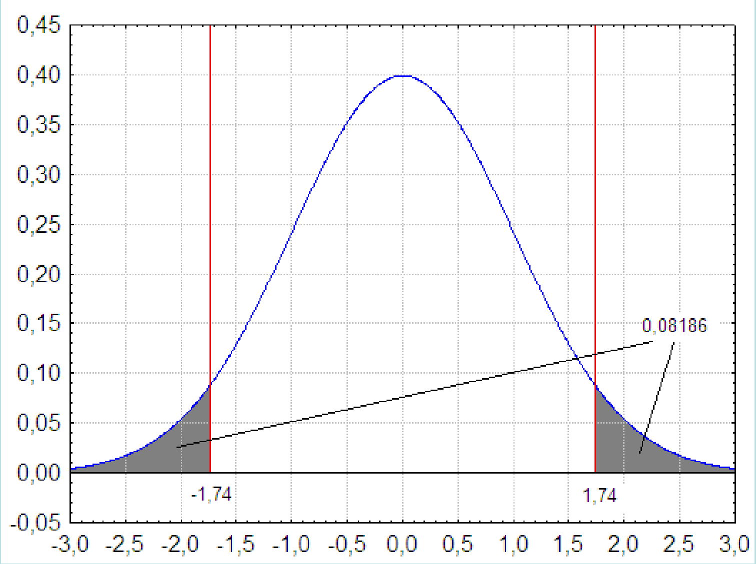
Protože proti nulové hypotéze stavíme oboustrannou alternativu, použijeme vzorec

$p = 2 \min\{P(T_0 \leq t_0), P(T_0 \geq t_0)\} = 2 \min\{P(T_0 \leq 1{,}74), P(T_0 \geq 1{,}74)\} =$

$= 2 \min\{ \Phi(1{,}74), 1 - \Phi(1{,}74) \} = 2 \min\{ 0{,}95907, 1 - 0{,}95907 \} = 0{,}08186.$

Jelikož $0{,}08186 > 0{,}05$, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Ilustrace významu p-hodnoty pro oboustranný test



2. Levostranná alternativa

Proti nulové hypotéze H_0: $\mu = 1{,}95$ postavíme levostrannou alternativu H_1: $\mu < 1{,}95$. Na hladině významnosti 0,05 testujte H_0 proti H_1 všemi třemi popsanými způsoby.

Řešení:

a) Test provedeme pomocí kritického oboru.

Na rozdíl od oboustranné alternativy bude mít kritický obor tvar

$W = \langle -\infty, u_\alpha) = \langle -\infty, u_{0,05}) = \langle -\infty, -1{,}645)$.

Protože $1{,}74 \notin W$, H_0 nezamítáme na hladině významnosti 0,05.

b) Test provedeme pomocí intervalu spolehlivosti.

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

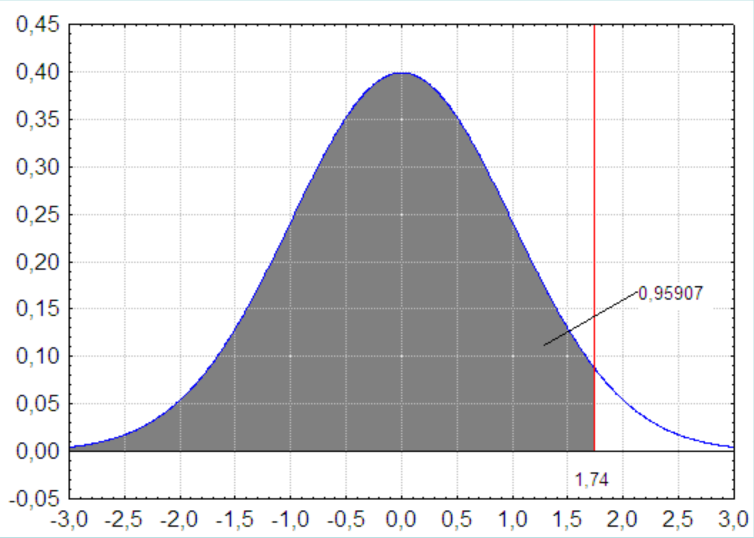
$(-\infty, h) = (-\infty, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha})$.

V našem případě dostáváme: $h = 2{,}06 + \frac{0{,}2}{\sqrt{10}} u_{0{,}95} = 2{,}06 + \frac{0{,}2}{\sqrt{10}} .1{,}645 = 2{,}164$.

Protože $1{,}95 \in (-\infty; 2{,}164)$, H_0 nezamítáme na hladině významnosti 0,05.

c) Test provedeme pomocí p-hodnoty.
Protože proti nulové hypotéze stavíme levostrannou alternativu, použijeme vzorec
$p = P(T_0 \leq t_0) = \Phi(1{,}74) = 0{,}95907$.
Jelikož 0,95907 > 0,05, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Ilustrace významu p-hodnoty pro levostranný test



3. Pravostranná alternativa

Proti nulové hypotéze H_0: $\mu = 1{,}95$ postavíme pravostrannou alternativu H_1: $\mu > 1{,}95$. Na hladině významnosti 0,05 testujte H_0 proti H_1 všemi třemi popsanými způsoby.

Řešení:

a) Test provedeme pomocí kritického oboru.

Na rozdíl od oboustranné alternativy bude mít kritický obor tvar

$W = \langle u_{1-\alpha}, \infty) = \langle u_{0,95}, \infty) = \langle 1{,}645, \infty)$.

Protože $1{,}74 \in W$, H_0 zamítáme na hladině významnosti 0,05 ve prospěch pravostranné alternativy.

b) Test provedeme pomocí intervalu spolehlivosti.

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

$(d, \infty) = (m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty)$.

V našem případě dostáváme: $d = 2{,}06 - \frac{0{,}2}{\sqrt{10}} u_{0{,}95} = 2{,}06 - \frac{0{,}2}{\sqrt{10}} . 1{,}645 = 1{,}956$.

Protože $1{,}95 \notin (1{,}956, \infty)$, H_0 zamítáme na hladině významnosti 0,05 ve prospěch pravostranné alternativy.

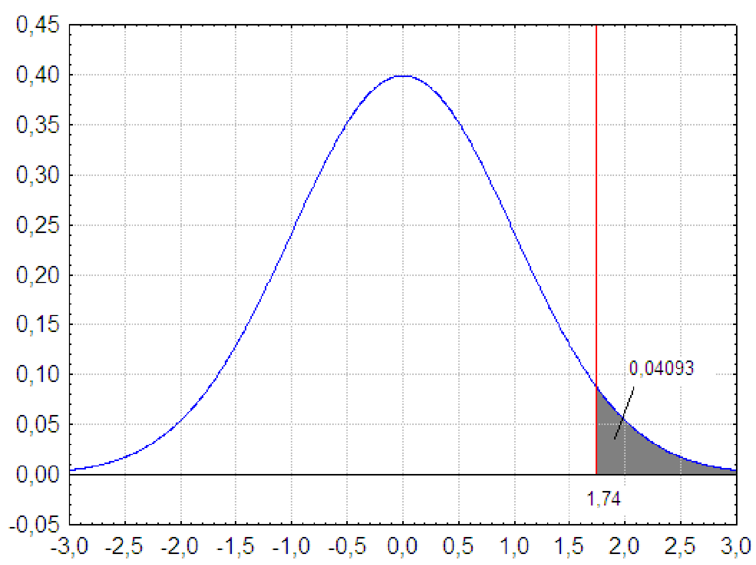
c) Test provedeme pomocí p-hodnoty.

Protože proti nulové hypotéze stavíme pravostrannou alternativu, použijeme vzorec

$p = P(T_0 \geq t_0) = 1 - \Phi(1{,}74) = 1 - 0{,}95907 = 0{,}04093$.

Jelikož $0{,}04093 \leq 0{,}05$, nulovou hypotézu zamítáme na hladině významnosti 0,05 ve prospěch pravostranné alternativy.

Ilustrace významu p-hodnoty pro pravostranný test



Základní typy uspořádání pokusů

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.

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.

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 $X_{11},\ldots,X_{1n_1}$ a $X_{21},\ldots,X_{2n_2}$.

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 $(X_{11}, X_{12}),\ldots,(X_{n1}, X_{n2})$ z dvourozměrného rozložení. Přejdeme k rozdílovému náhodnému výběru $Z_i = X_{i1} - X_{i2}$, $i = 1, \ldots, n$ 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.)

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

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

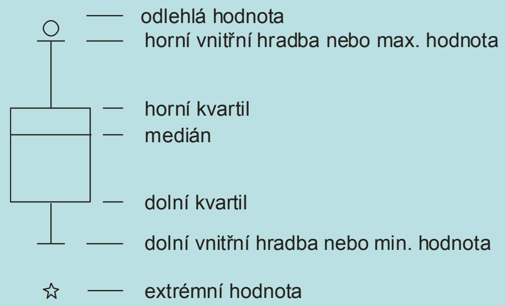
Blokové porovnávání: situace je charakterizována jedním náhodným výběrem $(X_{11}, \ldots, X_{1r}), \ldots, (X_{n1}, \ldots, X_{nr})$ z r-rozměrného rozložení.
Náhodně vylosujeme n vrhů stejně starých selat téhož plemene, z každého odebereme r sourozenců a náhodně jim přiřadíme první až r-tou výkrmnou 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

Diagnostické grafy slouží 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 N-P plotu, Q-Q plotu, histogramu) a nepřítomnost vybočujících hodnot (odhalí je krabicový diagram).
U dvou či více nezávislých náhodných výběrů sledujeme kromě normality též shodu středních hodnot nebo shodu rozptylů - homoskedasticitu (porovnáváme vzhled krabicových diagramů).
V případě jednoho dvourozměrného náhodného výběru často posuzujeme dvourozměrnou normalitu dat (použijeme dvourozměrný tečkový diagram s proloženou 100(1-α)% elipsou konstantní hustoty pravděpodobnosti).

Krabicový diagram

Umožňuje posoudit symetrii a variabilitu datového souboru a existenci odlehlých či extrémních hodnot. Způsob konstrukce je zřejmý z obrázku:



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

U 30 domácností byl zjišťován počet členů.

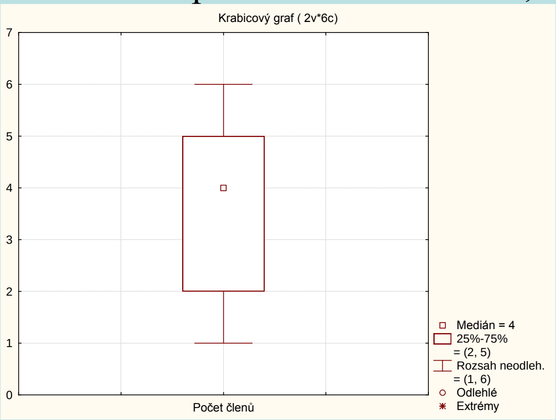
Počet členů	1	2	3	4	5	6
Počet domácností	2	6	4	10	5	3

Pro tyto údaje sestrojte krabicový diagram.

Řešení:

Vytvoříme nový datový soubor se dvěma proměnnými Počet členů, Počet domácností a o 6 případech. Vytvoření krabicového diagramu: Grafy – 2D Grafy – Krabicové grafy. Abychom systému STATISTICA sdělili, že pracujeme s údaji, pro které známe absolutní četnosti, klikneme myší na tlačítko s obrázkem závaží.

V okénku Váhy případů pro analýzu/graf zaškrtneme Status Zapnuto a zadáme Proměnná vah Počet domácností, OK. Na panelu 2D Krabicové grafy zadáme Proměnné – Závisle proměnné Počet členů, OK. Dostaneme krabicový diagram



Z obrázku lze vyčíst, že medián je 4 (aspoň polovina domácností má aspoň 4 členy), dolní kvartil 2 (aspoň čtvrtina domácností má aspoň 2 členy), horní kvartil 5 (aspoň tři čtvrtiny domácností mají aspoň 5 členů), minimum 1, maximum 6. Kvartilová odchylka je 5 – 2 = 3. Datový soubor vykazuje určitou nesymetrii – medián je posunut směrem k hornímu kvartilu, soubor je tedy záporně zešikmen. Odlehlé ani extrémní hodnoty se nevyskytují.

Normální pravděpodobnostní graf (N-P plot)

Před popisem tohoto grafu se musíme seznámit s pojmem pořadí čísla v posloupnosti čísel: Nechť $x_1, \ldots, x_n$ je posloupnost reálných čísel.

a) Jsou-li čísla navzájem různá, pak pořadím R_i čísla x_i rozumíme počet těch čísel $x_1, \ldots, x_n$, která jsou menší nebo rovna číslu x_i.

b) Vyskytují-li se mezi danými čísly skupinky stejných čísel, pak každé takové skupince přiřadíme průměrné pořadí.

Příklad

a) Jsou dána čísla 9, 4, 5, 7, 3, 1.

b) Jsou dána čísla 6, 7, 7, 9, 6, 10, 8, 6, 6, 9.

Stanovte pořadí těchto čísel.

Řešení

ad a)

usp. čísla	1	3	4	5	7	9
pořadí	1	2	3	4	5	6

ad b)

usp. čísla	6	6	6	6	7	7	8	9	9	10
pořadí	1	2	3	4	5	6	7	8	9	10
prům. pořadí	2,5	2,5	2,5	2,5	5,5	5,5	7	8,5	8,5	10

N-P plot umožňuje graficky posoudit, zda data pocházejí z normálního rozložení.

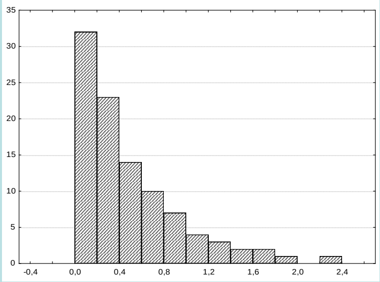
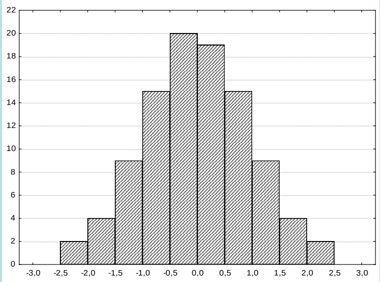
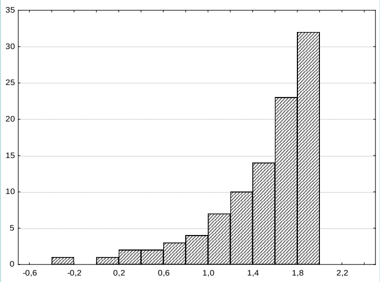
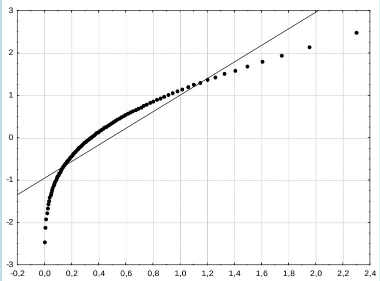
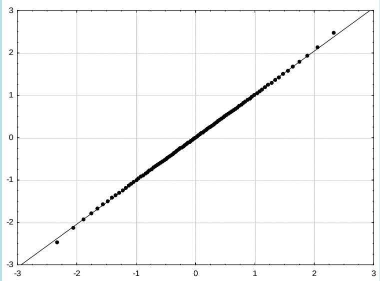
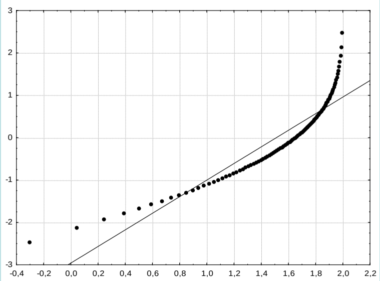
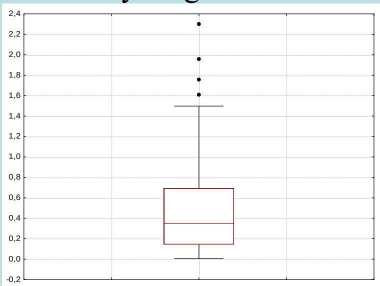
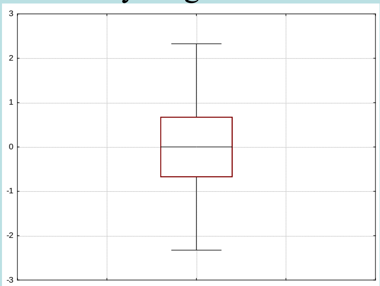
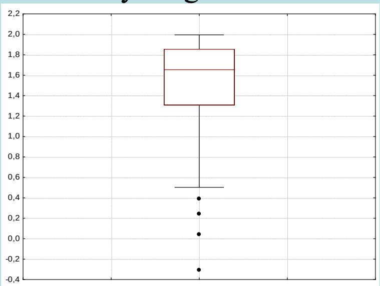
Způsob konstrukce:

na vodorovnou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \ldots \leq x_{(n)}$,

na svislou osu kvantily u_{α_j}, kde $\alpha_j = \frac{3j-1}{3n+1}$, přičemž j je pořadí j-té uspořádané hodnoty (jsou-li některé hodnoty stejné, pak za j bereme průměrné pořadí odpovídající takové skupince).

Pocházejí-li data z normálního rozložení, pak všechny dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou ležet na přímce.

Pro data z rozložení s kladnou šikmostí se dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou řadit do konvexní křivky, zatímco pro data z rozložení se zápornou šikmostí se dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou řadit do konkávní křivky.

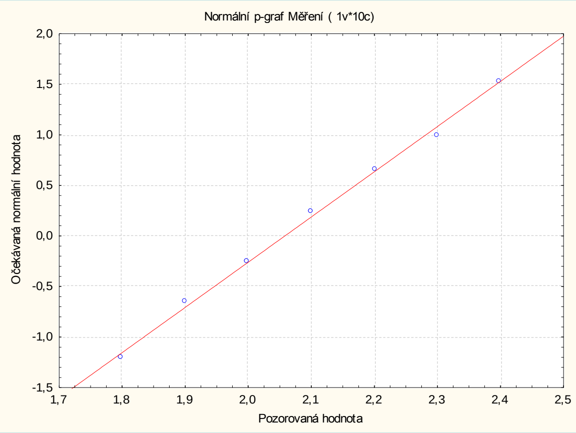
<table>
<tr><td>Rozložení
s kladnou šikmostí</td><td>Normální rozložení</td><td>Rozložení
se zápornou šikmostí</td></tr>
<tr><td>Histogram

</td><td>Histogram

</td><td>Histogram

</td></tr>
<tr><td>N-P plot

</td><td>N-P plot

</td><td>N-P plot

</td></tr>
<tr><td>Krabicový diagram

</td><td>Krabicový diagram

</td><td>Krabicový diagram

</td></tr>
</table>

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

Po zapsání dat do proměnné nazvané Měření zvolíme Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnné Měření, OK.



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

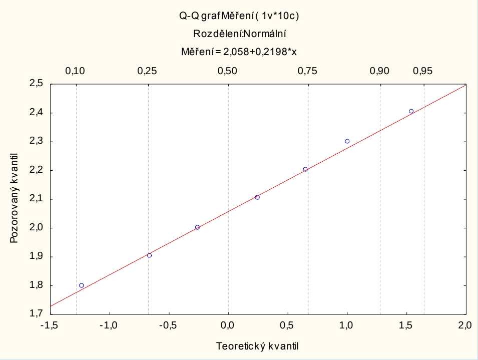
Kvantil-kvantilový graf (Q-Q plot)

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

Způsob konstrukce: na svislou osu vynášíme uspořádané hodnoty $x_{(1)} \le \ldots \le x_{(n)}$ a na vodorovnou osu kvantily $K_{\alpha_j}(X)$ vybraného rozložení, kde $\alpha_j = \frac{j - r_{adj}}{n + n_{adj}}$, přičemž r_{adj} a n_{adj} jsou korigující faktory $\le 0{,}5$, implicitně $r_{adj} = 0{,}375$ a $n_{adj} = 0{,}25$. (Jsou-li některé hodnoty $x_{(1)} \le \ldots \le x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.) Pokud vybrané rozložení závisí na nějakých parametrech, pak se tyto parametry odhadnou z dat nebo je může zadat uživatel. Body $\left(K_{\alpha_j}(X), x_{(j)}\right)$ se metodou nejmenších čtverců proloží přímka. Čím méně se body odchylují od této přímky, tím je lepší soulad mezi empirickým a teoretickým rozložením.

Příklad: Pro data z příkladu o měření konstanty posuďte pomocí kvantil – kvantilového grafu, zda pocházejí z normálního rozložení.

Řešení: Zvolíme Grafy – 2D Grafy – Grafy typu Q-Q – ponecháme implicitní nastavení na normální rozložení (pokud bychom chtěli změnit nastavení na jiný typ rozložení, zvolili bychom ho na záložce Detaily) – Proměnné Měření, OK.



Vzhled grafu nasvědčuje tomu, že data pocházejí z normálního rozložení.

Histogram

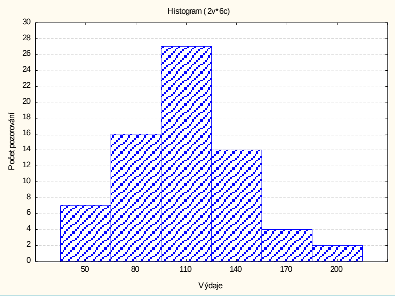
Umožňuje porovnat tvar hustoty četnosti s tvarem hustoty pravděpodobnosti vybraného teoretického rozložení. (Ve STATISTICE je pojem histogramu širší, skrývá se za ním i sloupkový diagram.)
Způsob konstrukce ve STATISTICE: na vodorovnou osu se vynášejí třídicí intervaly (implicitně 10, jejich počet lze 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 může zakreslit tvar hustoty (či pravděpodobnostní funkce) vybraného teoretického rozložení. Kromě osmi typů rozložení uvedených u Q-Q plotu umožňuje STATISTICA použít ještě další čtyři rozložení: Laplaceovo, logistické, geometrické, Poissonovo.

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

Výdaje	$\langle 35,65 \rangle$	$\langle 65,95 \rangle$	$\langle 95,125 \rangle$	$\langle 125,155 \rangle$	$\langle 155,185 \rangle$	$\langle 185,215 \rangle$
Počet dom.	7	16	27	14	4	2

Nakreslete histogram

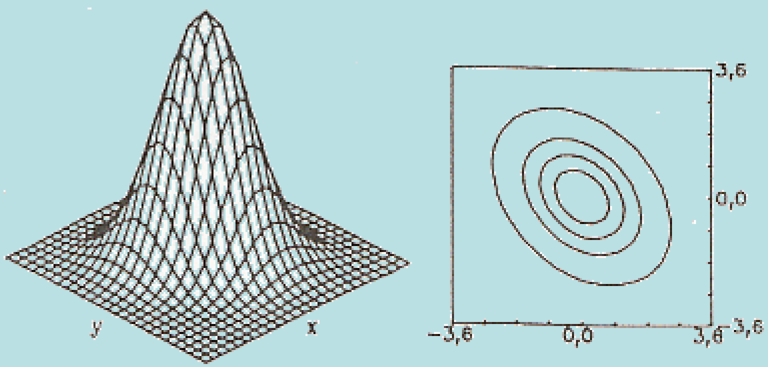
Řešení: Vytvoříme nový datový soubor s dvěma proměnnými Výdaje a Počet domácností. Do proměnné Výdaje zapíšeme středy třídicích intervalů, do proměnné Počet domácností odpovídající absolutní četnosti třídicích intervalů. V menu zvolíme Grafy – Histogramy – pomocí tlačítka s obrázkem závaží zadáme proměnnou vah Počet domácností – OK, Proměnná Výdaje – zapneme volbu Všechny hodnoty – OK. Dostaneme histogram:



Vidíme, že tvar histogramu není symetrický. Malé hodnoty jsou četnější než velké – datový soubor je kladně zešikmen.

Dvourozměrný tečkový diagram
Máme dvourozměrný datový soubor $(x_1, y_1), \dots , (x_n, y_n)$, který je realizací dvourozměrného náhodného výběru $(X_1, Y_1), \dots , (X_n, Y_n)$ z dvourozměrného rozložení. Na vodorovnou osu vyneseme hodnoty x_j , na svislou hodnoty y_k a do příslušných průsečíků nakreslíme tolik teček, jaká je absolutní četnost dvojice (x_j, y_k). Jedná-li se o náhodný výběr z dvourozměrného normálního rozložení, měly by tečky zhruba rovnoměrně vyplnit vnitřek elipsovitého obrazce. Vrstevnice hustoty dvourozměrného normálního rozložení jsou totiž elipsy – viz následující obrázek.

Graf hustoty a vrstevnice dvourozměrného normálního rozložení s parametry $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = -0{,}75$:



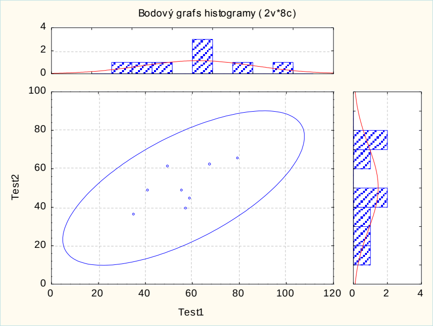
Do dvourozměrného tečkového diagramu můžeme ještě zakreslit $100(1-\alpha)\%$ elipsu konstantní hustoty pravděpodobnosti. Bude-li více než 100α % teček ležet vně této elipsy, svědčí to o porušení dvourozměrné normality. Bude-li mít hlavní osa elipsy kladnou resp. zápornou směrnici, znamená to, že mezi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

Příklad: 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

Pomocí dvourozměrného tečkového diagramu se zakreslenou 95% elipsou konstantní hustoty pravděpodobnosti a histogramy pro počty bodů v 1. a 2. testu posuďte, zda tato data lze považovat za realizace náhodného výběru z dvourozměrného normálního rozložení.

Řešení: Vytvoříme nový datový soubor se dvěma proměnnými Test1 a Test2 a osmi případy. Nyní nakreslíme dvourozměrný tečkový diagram: Grafy – 2D Grafy - Bodové grafy s histogramy. V typu proložení pro bodový graf vypneme lineární proložení. Proměnné – X – Test1, Y – Test2 – OK. Dostaneme dvourozměrný tečkový diagram pro vektorovou proměnnou (Test1, Test2) a histogramy pro Test1 a Test2. Nyní do diagramu zakreslíme 95% elipsu konstantní hustoty pravděpodobnosti: 2x klikneme na pozadí grafu a otevře se okno s názvem Vš. možnosti. Vybereme Graf: Elipsa, zvolíme Přidat novou elipsu. Po vykreslení elipsy změníme měřítko: na vodorovné ose bude minimum 0, maximum 120, na svislé ose bude minimum 0, maximum 100. (Stačí 2x kliknout na číselný popis osy a na záložce Měřítka vybrat manuální mód.)



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, tzn., že u studentů, kteří měli vysoký resp. nízký počet bodů v 1. testu, lze očekávat vysoký resp. nízký počet bodů ve 2. testu.

Testy normality dat

K ověřování normality dat slouží celá řada testů, které jsou podrobně popsány ve statistické literatuře. Zde se omezíme na dva testy, které jsou implementovány v systému STATISTICA, a to Kolmogorovův – Smirnovův test a jeho Lilieforsovu variantu a Shapirův – Wilksův test. K závěrům těchto testů však přistupujeme s určitou opatrností. Máme-li k dispozici rozsáhlejší datový soubor (orientačně n > 30) a test zamítne na obvyklé hladině významnosti 0,01 nebo 0,05 hypotézu o normalitě, i když vzhled diagnostických grafů svědčí jenom o lehkém porušení normality, nedopustíme se závažné chyby, pokud použijeme statistickou metodu založenou na normalitě dat.

Kolmogorovův – Smirnovův test a jeho Lilieforsova varianta

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z normálního rozložení s parametry μ a σ^2.
Distribuční funkci tohoto rozložení označme $\Phi_T(x)$.
Nechť $F_n(x)$ je výběrová distribuční funkce.
Testovou statistikou je statistika $D_n = \sup\limits_{-\infty<x<\infty} \left|F_n(x) - \Phi_T(x)\right|$.

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 $\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}$.

V případě, že neznáme parametry μ a σ^2 normálního rozložení, musíme je odhadnout z dat (střední hodnotu odhadneme pomocí m a rozptyl pomocí s^2). Tím se změní rozložení testové statistiky D_n. Příslušné modifikované kvantily byly určeny pomocí simulačních studií. V této situaci používáme Lilieforsovu variantu Kolmogorovova – Smirnovova testu.

Shapirův – Wilksův test normality dat

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

Testová statistika má tvar:

$$W = \frac{\sum_{i=1}^{m} a_i^{(n)} \left[X_{(n-i+1)} - X_{(i)}\right]^2}{\sum_{i=1}^{m} \left(X_i - M\right)^2},$$

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.)

Shapirův – Wilksův test normality dat

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

Testová statistika má tvar:

$$W = \frac{\sum_{i=1}^{m} a_i^{(n)} \left[X_{(n-i+1)} - X_{(i)}\right]^2}{\sum_{i=1}^{m} \left(X_i - M\right)^2},$$

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.)

Příklad:
Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí K- S testu a S – W testu zjistěte na hladině významnosti 0,05, zda tato data pocházejí z normálního rozložení.

Řešení:
Vytvoříme nový datový soubor o jedné proměnné nazvané X a pěti případech. Do proměnné X zapíšeme uvedené hodnoty. V menu vybereme Statistika – Základní statistiky/tabulky – Tabulky četností – OK, Proměnné X – OK. Na záložce zvolíme Normalita a zaškrtneme Lilieforsův test a Shapiro – Wilksův W test – Testy normality.

	Testy normality (Tabulka1)				
Proměnná	N	max D	Lilliefors p	W	p
X	5	0,22408	p > .20	0,91240	0,48215

Vidíme, že testová statistika K-S testu je d = 0,22409, odpovídající Lilieforsova p-hodnota je větší než 0,2, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.
Testová statistika S-W testu je W = 0,9124, odpovídající p-hodnota je 0,48215, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

Parametrické úlohy o jednom náhodném výběru z normálního rozložení

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 rozptylu

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

a) $M \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, 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.)

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

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

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

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

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

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

Vysvětlení

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

ad d) $U \sim N(0, 1)$, $K \sim \chi^2(n-1)$ 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 ~ N(1002, 64), M ~ $N\left(1002, \frac{64}{9}\right)$

$$P(M \le 999) = P\left(\frac{M-1002}{\sqrt{\frac{64}{9}}} \le \frac{999-1002}{\sqrt{\frac{64}{9}}}\right) = P\left(U \le -\frac{9}{8}\right) = \Phi\left(\frac{-9}{8}\right) = 1-\Phi\left(\frac{9}{8}\right) = 1-\Phi(1{,}125) = 1-0{,}87076 = 0{,}12924$$

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

Řešení 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 $P(M \le 999) = \Phi(999)$, 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $X_1, ..., X_{10}$ 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.

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.

Řešení pomocí systému STATISTICA:
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	1,91613	2,20386	0,13832	0,36714	0,01913	0,13479

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.

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

a) Provedení jednovýběrového z-testu

Vypočteme realizaci testového kritéria $t_0 = \dfrac{m-c}{\dfrac{\sigma}{\sqrt{n}}}$. 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: $\mu = c$ proti H_1: $\mu \neq c$. Kritický obor má tvar: $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$.

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

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

b) Provedení jednovýběrového t-testu

Vypočteme realizaci testového kritéria $t_0 = \dfrac{m-c}{\dfrac{s}{\sqrt{n}}}$. 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: $W = \left(-\infty, -t_{1-\alpha/2}(n-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-1), \infty\right)$.

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

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

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:.

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

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

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

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: $X_1, \ldots, X_{50}$ 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 $\dfrac{m-c}{\frac{s}{\sqrt{n}}} = \dfrac{122-125}{\frac{8,6}{\sqrt{50}}} = -2,4667$.

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

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

Náhodný výběr z dvourozměrného rozložení

Nechť $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ je náhodný výběr z dvourozměrného rozložení, přičemž $n \geq 2$. Označíme $\mu = \mu_1 - \mu_2$ a zavedeme rozdílový náhodný výběr $Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n$, o němž předpokládáme, že se řídí normálním rozložením.

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-α)% empirického intervalu spolehlivosti pro střední hodnotu rozdílového náhodného výběru

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

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

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

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(μ, σ^2). Vypočteme meze 90% oboustranného intervalu spolehlivosti pro μ při neznámém σ:

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

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

-0,0844 < μ < 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 < μ < 0,1244 s pravděpodobností aspoň 0,9.

Párový t-test

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 $t_0 = \dfrac{m - c}{\frac{s}{\sqrt{n}}}$. 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: $W = \left(-\infty, -t_{1-\alpha/2}(n-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-1), \infty\right)$.

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

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

Příklad: 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 rozložení a jejich rozdíl se řídí normálním rozložením, 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 známe 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, …, 12.)

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

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

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.

Úlohy o parametru ϑ alternativního rozložení

S náhodným výběrem rozsahu n z alternativního rozložení se setkáváme v situaci, kdy provádíme n opakovaných nezávislých pokusů a v každém z těchto pokusů sledujeme nastoupení úspěchu. Pravděpodobnost úspěchu je pro všechny pokusy stejná. Náhodná veličina X_i nabude hodnoty 1, pokud v i-tém pokusu nastal úspěch a hodnoty 0, pokud v i-tém pokusu úspěch nenastal, i = 1, 2, …, n. Realizací náhodného výběru $X_1, \ldots, X_n$ je tedy posloupnost 0 a 1.

Opakování:

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

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

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

Centrální limitní věta:

Jsou-li náhodné veličiny $X_1, \dots, X_n$ 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 $U_n \approx N(0,1)$

Věta: Asymptotické rozložení statistiky odvozené z výběrového průměru.

Nechť $X_1, \dots, X_n$ 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 píšeme $U \approx N(0,1)$.)

Důkaz:

Protože $X_1, \dots, X_n$ 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:

$$U = \frac{\frac{Y_n}{n} - \vartheta}{\sqrt{\frac{n\vartheta(1-\vartheta)}{n^2}}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}} = \frac{M - \vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}} \approx N(0,1)$$

Věta: 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:

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

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

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

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

Příklad:

Náhodně bylo vybráno 100 osob a zjištěno, že 34 z nich nakupuje v internetových obchodech. Najděte 95% asymptotický interval spolehlivosti pro pravděpodobnost, že náhodně vybraná osoba nakupuje v internetových obchodech.

Řešení:

Zavedeme náhodné veličiny $X_1, ..., X_{100}$, přičemž $X_i = 1$, když i-tá osoba nakupuje v internetových obchodech a $X_i = 0$ jinak, $i = 1, ..., 100$. 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 $0{,}2472 < \vartheta < 0{,}4328$. Znamená to, že s pravděpodobností přibližně 95% je v uvažované populaci nejméně 24,7% a nejvíce 43,3% osob, které nakupují v internetových obchodech.

Výpočet pomocí systému STATISTICA:

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 nákupu v inetrnetových obchodech bude pohybovat v mezích 0,2471 až 0,4328.

b) Přibližný způsob, použitelný pro dostatečně velký rozsah výběru

Do nového datového souboru o jedné proměnné X a 100 případech uložíme 34 jedniček (nakupování v internetových obchodech) a 66 nul.

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:

Proměnná	Popisné statistiky (Tabulka3) N platných	Průměr	Int. spolehl. -95,000%	Int. spolehl. 95,000
X	100	0,34000	0,24553	0,43446

Dospěli jsme k výsledku, že s pravděpodobností aspoň 0,95 se pravděpodobnost nákupu v inetrnetových obchodech bude pohybovat v mezích 0,2455 až 0,4345.

Testování hypotézy o parametru ϑ

Nechť X_1, ..., X_n je náhodný výběr z rozložení A(ϑ) a nechť je splněna podmínka $n\vartheta(1-\vartheta) > 9$.

Na asymptotické hladině významnosti α testujeme hypotézu

H_0: ϑ = c proti alternativě H_1: $\vartheta \neq$ c (resp. H_1: ϑ < c resp. H_1: ϑ > c).

Testovým kritériem je statistika $T_0 = \dfrac{M - c}{\sqrt{\dfrac{c(1-c)}{n}}}$, která v případě platnosti nulové hypotézy má asymptoticky rozložení N(0,1).

Kritický obor má tvar $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$ (resp. $W = \left(-\infty, -u_{1-\alpha}\right\rangle$ resp. $W = \left\langle u_{1-\alpha}, \infty\right)$).

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

Příklad: 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 $X_1, \dots, X_{1000}$, přičemž $X_i = 1$, když i-tý výrobek byl zmetek a $X_i = 0$ jinak, i = 1, ..., 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: $t_0 = \dfrac{m-c}{\sqrt{\dfrac{c\cdot(1-c)}{n}}} = \dfrac{0{,}016-0{,}01}{\sqrt{\dfrac{0{,}01\cdot 0{,}99}{1000}}} = 1{,}907$.

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

b) Testování pomocí intervalu spolehlivosti

$$d = m - \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2} = 0{,}016 - \sqrt{\frac{0{,}016\cdot 0{,}984}{1000}}1{,}96 = 0{,}0082$$

$$h = m + \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2} = 0{,}016 + \sqrt{\frac{0{,}016\cdot 0{,}984}{1000}}1{,}96 = 0{,}0238$$

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

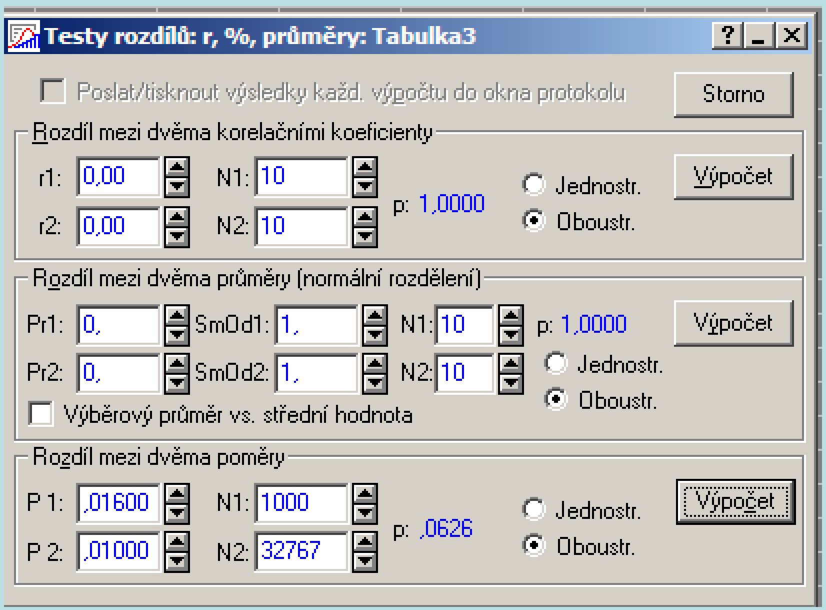
Protože testujeme nulovou hypotézu proti oboustranné alternativě, vypočteme p-hodnotu podle vzorce:

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

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

Výpočet pomocí systému STATISTICA (pouze přibližný):

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.



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

Máme dva nezávislé náhodné výběry, první pochází z rozložení $N(\mu_1, \sigma_1^2)$ a má rozsah $n_1 \geq 2$, druhý pochází z rozložení $N(\mu_2, \sigma_2^2)$ a má rozsah $n_2 \geq 2$. 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) $U = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ ~ $N(0, 1)$. (Pivotová statistika U slouží k řešení úloh o $\mu_1 - \mu_2$, když σ_1^2 a σ_2^2 známe.)

c) Nechť $\sigma_1^2 = \sigma_2^2 =: \sigma^2$, pak $K = \frac{(n_1+n_2-2)S_*^2}{\sigma^2}$ ~ $\chi^2(n_1 + n_2 - 2)$. (Pivotová statistika K slouží k řešení úloh o neznámém rozptylu σ^2.)

d) Jestliže $\sigma_1^2 = \sigma_2^2 =: \sigma^2$, pak $T = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{S_*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ ~ $t(n_1 + n_2 - 2)$. (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) $F = \frac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2}$ ~ $F(n_1 - 1, n_2 - 1)$. (Pivotová statistika F slouží k řešení úloh o σ_1^2/σ_2^2.)

Vysvětlení:

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:

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

$$= 1 - P\left(U \le \frac{-0,28 + 0,25}{\sqrt{\frac{0,09}{16} + \frac{0,04}{25}}} \right) = 1 - P(U \le -0,35294) = 1 - \Phi(-0,35) = \Phi(0,35) = 0,63683$$

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($\mu_1 - \mu_2$, $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$),

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

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

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

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

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:

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

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

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

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

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

Upozornění: Není-li v bodě (b) splněn předpoklad o shodě rozptylů, lze sestrojit aspoň přibližný 100(1-α)% interval spolehlivosti pro μ_1-μ_2.

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

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 z bodu (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:

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

$$= 34{,}48 - 35{,}59 - \sqrt{1{,}7384}\cdot\sqrt{\frac{1}{25}+\frac{1}{10}}\cdot 2{,}035 = -2{,}114$$

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

$$= 34{,}48 - 35{,}59 + \sqrt{1{,}7384}\cdot\sqrt{\frac{1}{25}+\frac{1}{10}}\cdot 2{,}035 = -0{,}106$$

$-2{,}114$ 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	-2,1136	-0,1063

S pravděpodobností aspoň 0,95 tedy -2,114 g/l < μ_1 - μ_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 z bodu (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(μ_1, σ_1^2) a $X_{21},\ldots,X_{2n_2}$ je na něm nezávislý náhodný výběr z rozložení N(μ_2, σ_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(μ_1, σ^2) a $X_{21},\ldots,X_{2n_2}$ je na něm nezávislý náhodný výběr rozložení N(μ_2, σ^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(μ_1, σ_1^2) a $X_{21},\ldots,X_{2n_2}$ je na něm nezávislý náhodný výběr rozložení N(μ_2, σ_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 $\mu_1-\mu_2$, σ_1^2/σ_2^2 pomocí kritického oboru

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: $W = \langle u_{1-\alpha}, \infty)$.

b) Provedení dvouvýběrového t-testu

Vypočteme realizaci testového kritéria $t_0 = \dfrac{(m_1 - m_2) - c}{s_* \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$.

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:
$W = \left(-\infty, -t_{1-\alpha/2}(n_1 + n_2 - 2)\right\rangle \cup \left\langle t_{1-\alpha/2}(n_1 + n_2 - 2), \infty\right)$.

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

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

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 přijím á m e 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:

$$W = \left(0, F_{\alpha/2}(n_1 - 1, n_2 - 1)\right\rangle \cup \left\langle F_{1-\alpha/2}(n_1 - 1, n_2 - 1), \infty\right).$$

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: $W = \left(0, F_{\alpha}(n_1 - 1, n_2 - 1)\right\rangle$.

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: $W = \left\langle F_{1-\alpha}(n_1 - 1, n_2 - 1), \infty\right)$.

Příklad: 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 vzorce z bodu (c) vypočteme realizaci testové statistiky:

$t_0 = \frac{s_1^2}{s_2^2} = \frac{6{,}307}{9{,}41} = 0{,}6702$. Stanovíme kritický obor:

$$W = \langle 0, F_{\alpha/2}(n_1-1, n_2-1)\rangle \cup \langle F_{1-\alpha/2}(n_1-1, n_2-1), \infty) = \langle 0, F_{0{,}025}(19{,}14)\rangle \cup \langle F_{0{,}975}(19{,}14), \infty) =$$

$$= \langle 0{,}1/F_{0{,}975}(14{,}19)\rangle \cup \langle F_{0{,}975}(19{,}14), \infty) = \langle 0{,}1/2{,}6469\rangle \cup \langle 2{,}8607, \infty) = \langle 0; 0{,}3778\rangle \cup \langle 2{,}8607, \infty)$$

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. Podle vzorce z bodu (b) 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:

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

$$= (-\infty, -2{,}035\rangle \cup \langle 2{,}035, \infty)$$

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

Výpočet pomocí systému STATISTICA:

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.

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.

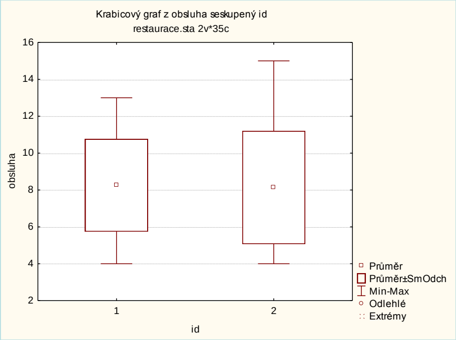
Po kliknutí na tlačítko Souhrn dostaneme tabulku

	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

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“.

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.