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

b) Nechť $(X_1,Y_1), \dots, (X_n,Y_n)$ jsou stochasticky nezávislé dvourozměrné náhodné vektory, které mají všechny stejné dvourozměrné rozložení $L_2(\vartheta)$. Řekneme, že $(X_1,Y_1), \dots, (X_n,Y_n)$ je dvourozměrný náhodný výběr rozsahu n z dvourozměrného rozložení $L_2(\vartheta)$. (Číselné realizace $(x_1,y_1), \dots, (x_n,y_n)$ náhodného výběru $(X_1,Y_1), \dots, (X_n,Y_n)$ 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 $T = T(X_1, \dots, X_n)$ náhodného výběru $X_1, \dots, X_n$ (resp. $T = T(X_1,Y_1, \dots, X_n,Y_n)$ náhodného výběru $(X_1,Y_1), \dots, (X_n,Y_n)$) 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 $n_1 \geq 2, \dots, n_r \geq 2$.

Celkový rozsah je $n = \sum_{j=1}^{r} n_j$.

Označme $M_1, \dots, M_r$ výběrové průměry a $S_1^2, \dots, S_r^2$ výběrové rozptyly jednotlivých výběrů. Nechť $c_1, \dots, c_r$ 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.

Pro libovolnou, ale pevně zvolenou dvojici reálných čísel x,y je statistikou též hodnota výběrové simultánní distribuční funkce $F_n(x,y) = \frac{1}{n}\operatorname{card}\{i; X_i \leq x \wedge Y_i \leq y\}$.

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

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 $X_1, \ldots, X_{10}$. Vypočtěte realizaci m výběrového průměru M, realizaci s^2 výběrového rozptylu S^2, realizaci s výběrové směrodatné odchylky S a hodnoty výběrové distribuční funkce $F_{10}(x)$.

Řešení:

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

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

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

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

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

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

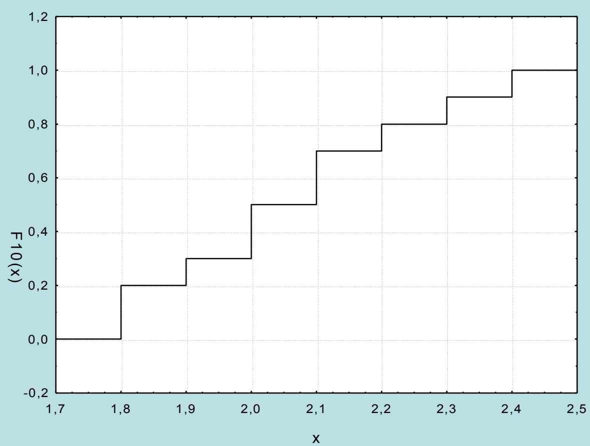
$$2 \le x < 2{,}1 : F_{10}(x) = \frac{5}{10} = 0{,}5$$

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

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

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

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



Příklad (výpočet realizace výběrového koeficientu korelace):

U 11 náhodně vybraných aut jisté značky bylo zjišťováno jejich stáří (náhodná veličina X – v letech) a cena (náhodná veličina Y – v tisících Kč). Výsledky:

(5, 85), (4, 103), (6, 70), (5, 82), (5, 89), (5, 98), (6, 66), (6, 95), (2, 169), (7, 70), (7, 48).

Vypočtěte a interpretujte číselnou realizaci r_{12} výběrového koeficientu korelace R_{12}.

Řešení:

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

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

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

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

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

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

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

Bodové a intervalové odhady parametrů a parametrických funkcí

Vycházíme z náhodného výběru $X_1, \dots, X_n$ z rozložení $L(\vartheta)$, které závisí na parametru ϑ. Množinu všech přípustných hodnot tohoto parametru označíme Ξ. Tato množina se nazývá parametrický prostor.
Např. je-li $X_1, \dots, X_n$ náhodný výběr z rozložení $N(\mu,\sigma^2)$, pak $\vartheta = (\mu, \sigma^2)$ a v tomto případě parametrický prostor $\Xi = (-\infty, \infty) \times (0, \infty)$.
Parametr ϑ neznáme a chceme ho odhadnout pomocí daného náhodného výběru (případně chceme odhadnout nějakou parametrickou funkci $h(\vartheta)$).

Bodovým odhadem parametrické funkce $h(\vartheta)$ je statistika $T_n = T(X_1, \dots, X_n)$, která nabývá hodnot blízkých $h(\vartheta)$, ať je hodnota parametru ϑ jakákoliv. Existují různé metody, jak konstruovat bodové odhady (např. metoda momentů či metoda maximální věrohodnosti, 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 $h(\vartheta)$ rozumíme interval (D, H), jehož meze jsou statistiky
$D = D(X_1, \dots, X_n)$, $H = H(X_1, \dots, X_n)$ a který s dostatečně velkou pravděpodobností pokrývá $h(\vartheta)$, ať je hodnota parametru ϑ jakákoliv.

Typy bodových odhadů

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

a) Řekneme, že statistika T je nestranným odhadem parametrické funkce $h(\vartheta)$, jestliže

$\forall \vartheta \in \Theta$ $E(T) = h(\vartheta)$.

(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

$\forall \vartheta \in \Theta$ $D(T_1) < D(T_2)$.

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

$\forall \vartheta \in \Theta$ $\lim_{n\to\infty} E(T_n) = h(\vartheta)$.

(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

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

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

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

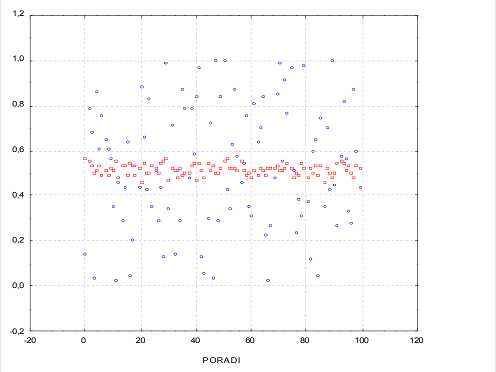
Posloupnost $\{M_n\}_{n=1}^{\infty}$ je posloupnost konzistentních odhadů μ,

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

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

Ilustrace:

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



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

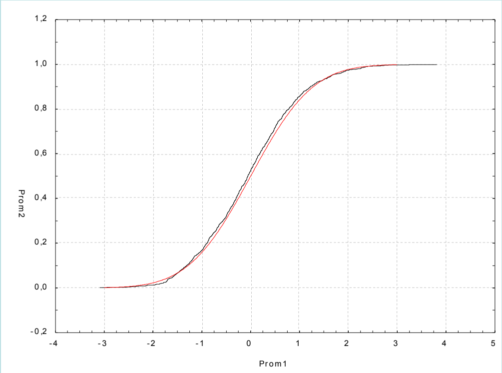
Dále vypočteme průměr a rozptyl např. proměnné v1 a proměnné PRUMER a dále vypočteme průměr proměnné ROZPTYL.

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

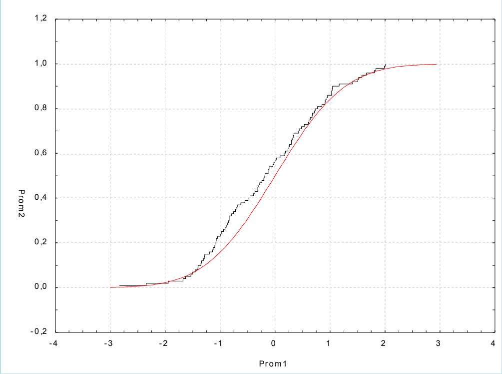
Proměnná	Popisné statistiky (uniform) Průměr
ROZPTYL	0,083143

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

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



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



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 $S_*^2 = \frac{\sum_{j=1}^{r} (n_j - 1) S_j^2}{n - r}$ 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

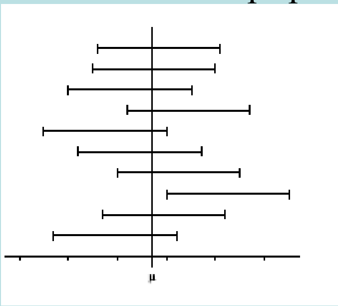
Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $L(\vartheta)$,
$h(\vartheta)$ je parametrická funkce,
$\alpha \in (0,1)$,
$D = D(X_1, ..., X_n)$, $H = H(X_1, ..., X_n)$ jsou statistiky.

a) Interval (D, H) se nazývá 100(1-α)% (oboustranný) interval spolehlivosti pro parametrickou funkci $h(\vartheta)$, jestliže: $\forall \ \in \Xi$ $P(D < h(\vartheta) < H) \geq 1-\alpha$.
b) Interval (D, ∞) se nazývá 100(1-α)% levostranný interval spolehlivosti pro parametrickou funkci $h(\vartheta)$, jestliže: $\forall \ \in \Xi$ $P(D < h(\vartheta)) \geq 1-\alpha$.
c) Interval $(-\infty, H)$ se nazývá 100(1-α)% pravostranný interval spolehlivosti pro parametrickou funkci $h(\vartheta)$, jestliže: $\forall \ \in \Xi$ $P(h(\vartheta) < H) \geq 1-\alpha$.
Číslo α se nazývá riziko (zpravidla α = 0,05, méně často 0,1 či 0,01), číslo 1 – α 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: $\forall \; \in \Xi$: $P(w_{\alpha/2} < W < w_{1-\alpha/2}) \geq 1-\alpha$.
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 $1-\alpha$. (Tvrzení, že (d,h) pokrývá $h(\vartheta)$ s pravděpodobností aspoň $1-\alpha$ 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 $1-\alpha$.)

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-α)% 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 $D(M) = \frac{\sigma}{n}$. 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-α)% intervalu spolehlivosti pro střední hodnotu μ při známém rozptylu σ^2 tedy jsou:

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

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

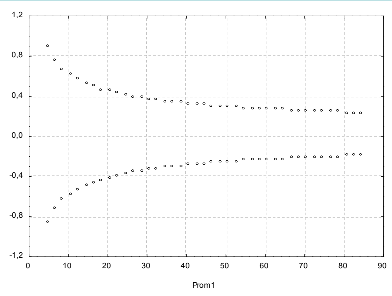
Šířka intervalu spolehlivosti

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.

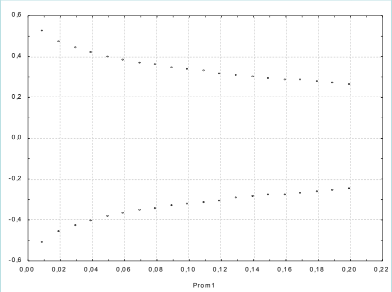
Ilustrace

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 $100(1-\alpha)\%$ empirického intervalu spolehlivosti pro střední hodnotu μ nepřesáhla číslo Δ?

Řešení: Požadujeme, aby $\Delta \geq h - d = \left(m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right) - \left(m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right) = \frac{2\sigma}{\sqrt{n}} u_{1-\alpha/2}$. 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 $\sigma = 1$ m. Kolik měření je nutno provést, aby se hloubka stanovila s chybou nejvýše $\pm$ 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.

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

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

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

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

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

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 $X_{11}, \dots, X_{1n_1}$ až $X_{r1}, \dots, X_{rn_r}$.
Náhodně vylosujeme n_1, n_2, …, n_r selat téhož plemene, náhodně je rozdělíme na r souborů o n_1, n_2, …, n_r jedincích, první podrobíme výkrmné dietě č. 1, druhý výkrmné dietě číslo 2 atd. až r-tý podrobíme výkrmné dietě číslo r. Tak dostaneme realizace r nezávislých náhodných výběrů.

Blokové porovnávání: situace je charakterizována jedním náhodným výběrem $(X_{11}, \dots, X_{1r}), \dots, (X_{n1}, \dots, X_{nr})$ z r-rozměrného rozložení.
Náhodně vylosujeme n vrhů stejně starých selat téhož plemene, z každého odebereme r sourozenců a náhodně jim přiřadíme první až r-tou výkrmnou dietu. Tak dostaneme realizaci jednoho r-rozměrného náhodného výběru, kde první složka odpovídá první dietě , druhá složka druhé dietě atd. až r-tá složka odpovídá r-té dietě.