Téma 6.: Základní pojmy matematické statistiky

Vlastnosti důležitých statistik odvozených z jednorozměrného 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

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

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

pro libovolné, ale pevně dané $x \in \mathbb{R}$ označme

$F_n(x) = \frac{1}{n}$ počet těch veličin $X_1, \ldots, X_n$, které jsou $\leq x$

hodnotu výběrové distribuční funkce.
Pak pro libovolné hodnoty parametrů μ, σ^2 a libovolné, ale pevně dané reálné číslo x platí:
$E(M) = \mu$,
$E(S_n^2) = \sigma^2$,
$E(F_n(x)) = \Phi(x)$,
Znamená to, že
- výběrový průměr M je nestranným odhadem střední hodnoty μ,
- výběrový rozptyl S^2 je nestranným odhadem rozptylu σ^2,
- pro libovolné, ale pevně dané $x \in \mathbb{R}$ je výběrová distribuční funkce $F_n(x)$ nestranným odhadem distribuční funkce $\Phi(x)$.

Příklad 1.: Ve 12 náhodně vybraných prodejnách ve městě byly zjištěny následující ceny určitého výrobku (v Kč): 102, 99, 106, 103, 96, 98, 100, 105, 103, 98, 104, 107. Těchto 12 hodnot považujeme za realizace náhodného výběru $X_1, ..., X_{12}$ z rozložení, které má střední hodnotu μ a rozptyl σ^2.
a) Určete nestranné bodové odhady neznámé střední hodnoty μ a neznámého rozptylu σ^2.
b) Najděte výběrovou distribuční funkci $F_{12}(x)$ a nakreslete její graf.
Řešení:
Vypočteme realizaci výběrového průměru

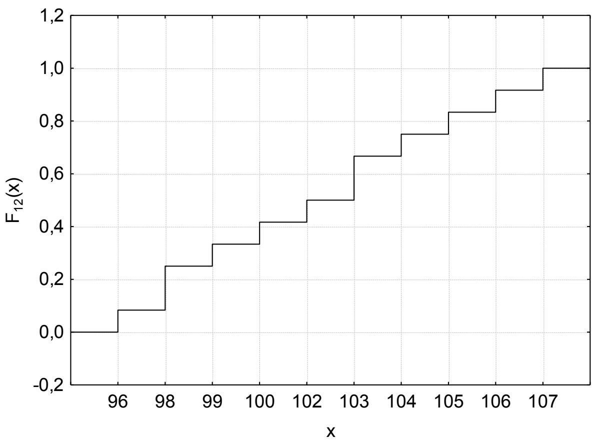
$$m = \frac{1}{12}(102 + 99 + \ldots + 107) = 101{,}75\ \text{Kč}$$

Vypočteme realizaci výběrového rozptylu:

$$s^2 = \frac{1}{11}\left[(102 - 101{,}75)^2 + (99 - 101{,}75)^2 + \ldots + (107 - 101{,}75)^2\right] = 12{,}39\ \text{Kč}^2$$

Pro usnadnění výpočtu hodnot výběrové distribuční funkce $F_{12}(x)$ uspořádáme ceny podle velikosti: 96, 98, 98, 99, 100, 102, 103, 103, 104, 105, 106, 107.
Číselnou osu rozdělíme na 11 intervalů a v každém intervalu stanovíme hodnotu výběrové distribuční funkce.

$$x < 96: F_{12}(x) = 0$$
$$96 \le x < 98: F_{12}(x) = \frac{1}{12} = 0{,}08\overline{3}$$
$$98 \le x < 99: F_{12}(x) = \frac{3}{12} = 0{,}25$$
$$99 \le x < 100: F_{12}(x) = \frac{4}{12} = 0{,}\overline{3}$$
$$100 \le x < 102: F_{12}(x) = \frac{5}{12} = 0{,}41\overline{6}$$
$$102 \le x < 103: F_{12}(x) = \frac{6}{12} = 0{,}5$$
$$103 \le x < 104: F_{12}(x) = \frac{8}{12} = 0{,}\overline{6}$$
$$104 \le x < 105: F_{12}(x) = \frac{9}{12} = 0{,}75$$
$$105 \le x < 106: F_{12}(x) = \frac{10}{12} = 0{,}8\overline{3}$$
$$106 \le x < 107: F_{12}(x) = \frac{11}{12} = 0{,}91\overline{6}$$
$$x \ge 107: F_{12}(x) = 1$$



Výpočet pomocí systému STATISTICA:
Otevřeme nový datový soubor o jedné proměnné (nazveme ji X) a 12 případech. Do proměnné X napíšeme zjištěné ceny.

Výpočet realizace výběrového průměru a výběrového rozptylu:
Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – vybereme Průměr a Rozptyl – Výpočet. Dostaneme tabulku:

	Popisné statistiky (Tabulka15)	
Proměnná	Průměr	Rozptyl
X	101,7500	12,38636

Výpočet hodnot výběrové distribuční funkce:
Statistiky – Základní statistiky/tabulky – Tabulky četností – OK – Proměnné X – OK – Možnosti – ponecháme zaškrtnuté pouze Kumulativní relativní četnosti – Výpočet.
Ke vzniklé tabulce přidáme jeden případ před první případ (do sloupce Kategorie napíšeme 95) a jeden případ za poslední případ (do sloupce Kategorie napíšeme 107). Proměnnou Kumulativní rel. četnost podělíme 100: do jejího Dlouhého jména napíšeme = v2/100.

Kreslení grafu výběrové distribuční funkce:
Nastavíme se kurzorem na proměnnou Kumulativní rel. četnost, klikneme pravým tlačítkem – Grafy bloku dat – Spojnicový graf: celé sloupce. Ve vytvořeném grafu odstraníme značky, spojnici změníme na schodovitou a upravíme měřítko na vodorovné ose od 1 do 12.

Vlastnosti důležitých statistik odvozených z dvourozměrného náhodného výběru:
Nechť $(X_1,Y_1), \ldots, (X_n,Y_n)$ je náhodný výběr z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ. Označme

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

$R_{12} = \frac{S_{12}}{S_1 S_2}$ výběrový koeficient korelace.

Pak pro libovolné hodnoty parametrů σ_{12} a ρ platí:
$E(S_{12}) = \sigma_{12}$,
$E(R_{12}) \approx \rho$ (shoda je vyhovující pro $n \geq 30$).
Znamená to, že výběrová kovariance S_{12} je nestranným odhadem kovariance σ_{12}, avšak výběrový koeficient korelace R_{12} je vychýleným odhadem koeficientu korelace ρ.

Příklad 2.: Bylo zkoumáno 9 vzorků půdy s různým obsahem fosforu (veličina X). Hodnoty veličiny Y označují obsah fosforu v obilných klíčcích (po 38 dnech), jež vyrostly na těchto vzorcích půdy.

číslo vzorku	1	2	3	4	5	6	7	8	9
X	1	4	5	9	11	13	23	23	28
Y	64	71	54	81	76	93	77	95	109

Těchto 9 dvojic hodnot považujeme za realizace náhodného výběru $(X_1,Y_1), \ldots, (X_9,Y_9)$ z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ. Najděte bodové odhady výběrové kovariance σ_{12} a výběrového koeficientu korelace ρ.

Výpočet pomocí systému STATISTICA:
Otevřeme nový datový soubor o dvou proměnných X a Y 9 případech. Do proměnných X a Y zapíšeme zjištěné hodnoty obsafu fosforu v půdě a v obilných klíčcích.
Výpočet výběrové kovariance: Statistiky – Vícerozměrná regrese – Proměnné – Závisle proměnná Y, nezávisle proměnná X – OK – OK – Residua/předpoklady/předpovědi – Popisné statistiky – Další statistiky – Kovariance. Dostaneme tabulku:

	Kovariance (Tabulka18	
Proměnná	X	Y
X	91,7500	130,0000
Y	130,0000	284,2500

Vidíme, že výběrová kovariance veličn X, Y se realizuje hodnotou 130. (Výběrový rozptyl proměnné X resp. Y nabyl hodnoty 91,75 resp. 284,25.)
Výpočet výběrového koeficientu korelace: V menu Další statistiky vybereme Korelace.

	Korelace (Tabulka18)	
Proměnná	X	Y
X	1,000000	0,804989
Y	0,804989	1,000000

Výběrový koeficient korelace veličin X, Y nabyl hodnoty 0,805, tedy mezi veličinami x, Y existuje silná přímá lineární závislost.
Upozornění: Výběrový koeficient korelace lze pomocí systému STATISTICA vypočítat i jiným způsobem: Statistika – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměnných – X, Y – OK – Výpočet. Ve výsledné tabulce máme též realizace výběrových průměrů a směrodatných odchylek.

Proměnná	Korelace (Tabulka18) Označ. korelace jsou významné na hlad. p < ,05000 N=9 (Celé případy vynechány u ChD)			
Proměnná	Průměry	Sm.odch.	X	Y
X	13,00000	9,57862	1,000000	0,804989
Y	80,00000	16,85972	0,804989	1,000000

Vzorce pro meze 100(1-α)% empirického intervalu spolehlivosti pro střední hodnotu μ normálního rozložení při známém rozptylu σ^2:

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

Levostranný: $d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}$.

Pravostranný: $h = m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}$.

Příklad 3.: Při kontrolních zkouškách životnosti 16 žárovek byl stanoven odhad m = 3000 h střední hodnoty jejich životnosti. Z dřívějších zkoušek je známo, že životnost žárovky se řídí normálním rozložením se směrodatnou odchylkou σ = 20 h. Vypočtěte

a) 99% empirický interval spolehlivosti pro střední hodnotu životnosti
b) 90% levostranný empirický interval spolehlivosti pro střední hodnotu životnosti
c) 95% pravostranný empirický interval spolehlivosti pro střední hodnotu životnosti.

Upozornění: Výsledek zaokrouhlete na jedno desetinné místo a vyjádřete v hodinách a minutách.

Řešení:

ad a)

$$d = m - \frac{\sigma}{\sqrt{n}} u_{0,995} = 3000 - \frac{20}{\sqrt{16}} 2,57583 = 2987,1,$$

$$h = m + \frac{\sigma}{\sqrt{n}} u_{0,995} = 3000 + \frac{20}{\sqrt{16}} 2,57583 = 3012,9$$

2987 h a 6 min < μ < 3012 h a 54 min s pravděpodobností 0,99

Výpočet pomocí systému STATISTICA

Otevřeme nový datový soubor o dvou proměnných d, h a jednom případu.
Do Dlouhého jména proměnné d napíšeme vzorec =3000-20/sqrt(16)*VNormal(0,995;0;1)
Do Dlouhého jména proměnné h napíšeme vzorec =3000+20/sqrt(16)*VNormal(0,995;0;1)

ad b)

$$d = m - \frac{\sigma}{\sqrt{n}} u_{0,9} = 3000 - \frac{20}{\sqrt{16}} 1,28155 = 2993,6$$

2993 h a 36 min < μ s pravděpodobností 0,9

Výpočet pomocí systému STATISTICA

Otevřeme nový datový soubor o jedné proměnné d a jednom případu.
Do Dlouhého jména proměnné d napíšeme vzorec =3000-20/sqrt(16)*VNormal(0,9;0;1)

ad c)

$$h = m + \frac{\sigma}{\sqrt{n}} u_{0,975} = 3000 + \frac{20}{\sqrt{16}} 1{,}95996 = 3009{,}8$$

3009 h a 48 min > μ s pravděpodobností 0,95

Výpočet pomocí systému STATISTICA

Otevřeme nový datový soubor o jedné proměnné h a jednom případu.
Do Dlouhého jména proměnné h napíšeme vzorec =3000+20/sqrt(16)*VNormal(0,975;0;1)

Užitečný odkaz: na adrese http://www.prevody-jednotek.cz je program, s jehož pomocí lze převádět různé fyzikální jednotky, v našem případě hodiny na minuty.

Základní poznatky o testování hypotéz

Předpokládáme, že testujeme nulovou hypotézu H_0: $h(\vartheta) = c$, kde $c \in \mathbb{R}$ buď proti oboustranné alternativě H_1: $h(\vartheta) \neq c$ nebo proti levostranné alternativě H_1: $h(\vartheta) < c$ nebo proti pravostranné alternativě H_1: $h(\vartheta) > c$.

Testování pomocí kritického oboru

Najdeme testovou statistiku $T_0 = T_0(X_1, ..., X_n)$. Množina všech hodnot, jichž může testová statistika 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). W aV jsou odděleny kritickými hodnotami (pro danou hladinu významnosti α je lze najít ve statistických tabulkách).
Jestliže číselná realizace t_0 testové statistiky T_0 padne do kritického oboru W, pak nulovou hypotézu zamítáme na hladině významnosti α a znamená to skutečné vyvrácení testované hypotézy. Jestliže t_0 padne do oboru nezamítnutí V, pak jde o pouhé mlčení, které platnost nulové hypotézy jenom připouští.
Stanovení kritického oboru 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
$W = \langle t_{min}, K_{\alpha/2}(T)\rangle \cup \langle K_{1-\alpha/2}(T), t_{max}\rangle$, 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:
$W = \langle t_{min}, K_{\alpha}(T)\rangle$.
Kritický obor v případě pravostranné alternativy má tvar:
$W = \langle K_{1-\alpha}(T), t_{max}\rangle$.

Testování pomocí intervalu spolehlivosti

Sestrojíme $100(1-\alpha)\%$ empirický interval spolehlivosti pro parametrickou funkci $h(\vartheta)$. 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-li $p \leq \alpha$, pak H_0 zamítáme na hladině významnosti α, je-li $p > \alpha$, 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 = $P(T_0 \geq t_0)$.

Příklad 4.: Víme, že výška hochů ve věku 9,5 až 10 let má normální rozložení s neznámou střední hodnotou μ a známým rozptylem $\sigma^2 = 39{,}112\ \text{cm}^2$. Dětský lékař náhodně vybral 15 hochů uvedeného věku, změřil je a vypočítal realizaci výběrového průměru m = 139,13 cm. Podle jeho názoru by výška hochů v tomto věku neměla přesáhnout 142 cm s pravděpodobností 0,95. Lze tvrzení lékaře akceptovat?
Řešení: Testujeme H_0: $\mu = 142$ proti H_1: $\mu < 142$ na hladině významnosti 0,05.
a) Test provedeme pomocí kritického oboru.
Pro úlohy o střední hodnotě normálního rozložení při známém rozptylu používáme pivotovou statistiku $U = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$. Testová statistika tedy bude $T_0 = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}}$ a bude mít rozložení N(0, 1), pokud je nulová hypotéza pravdivá. Vypočítáme realizaci testového kritéria:

$$t_0 = \frac{139{,}13 - 142}{\frac{\sqrt{39{,}112}}{\sqrt{15}}} = -1{,}7773.$$

Stanovíme kritický obor: $W = \left(-\infty, u_\alpha\right\rangle = \left(-\infty, u_{0,05}\right\rangle = \left(-\infty, -u_{0,95}\right\rangle = \left(-\infty, -1{,}6449\right\rangle$.
Protože $-1{,}7773 \in W$, H_0 zamítáme na hladině významnosti 0,05. Tvrzení lékaře lze tedy akceptovat s rizikem omylu 5 %.

b) Test provedeme pomocí intervalu spolehlivosti.
Meze $100(1-\alpha)\%$ 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: $h = 139{,}13 + \frac{\sqrt{39{,}112}}{\sqrt{15}} u_{0,95} = 139{,}13 + \frac{\sqrt{39{,}112}}{\sqrt{15}} 1{,}645 = 141{,}79$.
Protože $142 \notin (-\infty; 141{,}79)$, H_0 zamítáme na hladině významnosti 0,05.

c) Test provedeme pomocí p-hodnoty
$p = P(T_0 \leq t_0) = \Phi(-1{,}7773) = 0{,}0378$
Jelikož $0{,}0378 \leq 0{,}05$, nulovou hypotézu zamítáme na hladině významnosti 0,05.

Při řešení tohoto příkladu použijeme systém STATISTICA pouze jako inteligentní kalkulátor.