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

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

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

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

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

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 $x \in R$ 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.

$$x < 96 : F_{12}(x) = 0$$

$$96 \le x < 98 : F_{12}(x) = \frac{1}{12} = 0,08\overline{3}$$

$$98 \le x < 99 : F_{12}(x) = \frac{3}{12} = 0,25$$

$$99 \le x < 100 : F_{12}(x) = \frac{4}{12} = 0,\overline{3}$$

$$100 \le x < 102 : F_{12}(x) = \frac{5}{12} = 0,41\overline{6}$$

$$102 \le x < 103 : F_{12}(x) = \frac{6}{12} = 0,5$$

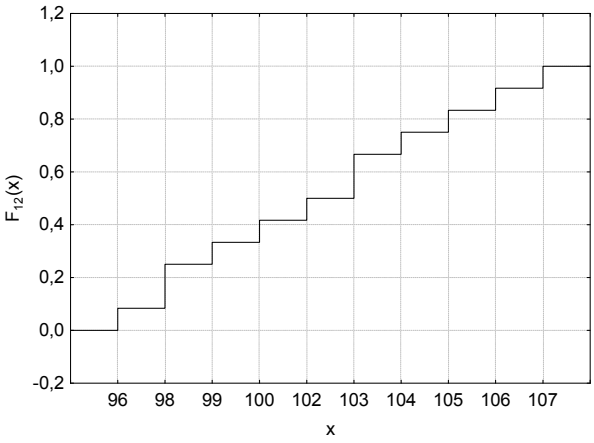
$$103 \le x < 104 : F_{12}(x) = \frac{8}{12} = 0,\overline{6}$$

$$104 \le x < 105 : F_{12}(x) = \frac{9}{12} = 0,75$$

$$105 \le x < 106 : F_{12}(x) = \frac{10}{12} = 0,8\overline{3}$$

$$106 \le x < 107 : F_{12}(x) = \frac{11}{12} = 0,91\overline{6}$$

$$x \ge 107 : F_{12}(x) = 1$$



Výpočet pomocí systému STATISTICA:
Načteme datový soubor ceny_vyrobku.sta.

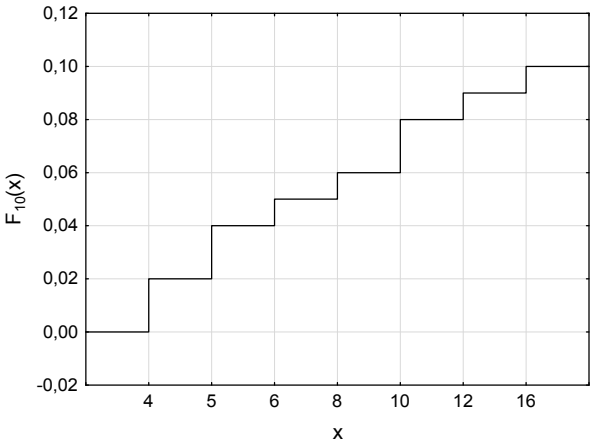
Výpočet realizace výběrového průměru a výběrového rozptylu:
Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – vybereme Průměr a Rozptyl – Výpočet. Dostaneme tabulku:

	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, do sloupce Kumulativní rel. četnost napíšeme 0) a jeden případ za poslední případ (do sloupce Kategorie napíšeme 107, do sloupce Kumulativní rel. četnost napíšeme 100). 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.

Příklad k samostatnému řešení: Přírůstky cen akcií (v procentech) na burze v New Yorku u 10 náhodně vybraných společností dosáhly těchto hodnot: 10, 16, 5, 10, 12, 8, 4, 6, 5, 4. Uvedená čísla považujeme za realizace náhodného výběru s neznámou střední hodnotou μ a neznámým rozptylem σ^2. Data jsou uložena v souboru akcie_na_burze.sta.
a) Najděte bodové odhady střední hodnoty (8), rozptylu (15,78) a směrodatné odchylky (3,97).
b) Najděte odhad pravděpodobnosti, že zvýšení cen akcií překročilo 8,5 % (0,4).
c) Nakreslete graf výběrové distribuční funkce.



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

$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$ 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), ..., (X_9,Y_9) z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ. Najděte bodové odhady kovariance σ_{12} a koeficientu korelace ρ.
Výpočet pomocí systému STATISTICA:
Načteme datový soubor obsah_foforu.sta.
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:

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

Vidíme, že výběrová kovariance veličin 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.

Proměnná	Korelace (Tabulka18	
	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)			
	Průměry	Sm.odch.	X	Y
X	13,00000	9,57862	1,000000	0,804989
Y	80,00000	16,85972	0,804989	1,000000

Příklad k samostatnému řešení: U 10 výrobců byly zjišťovány náklady (veličina X – v Kč) a ceny (veličina Y – v Kč) pro stejný výrobek. Výsledky (X,Y): (30,18; 50,26), (30,19; 50,23), (30,21; 50,27), (30,22; 50,25), (30,25; 50,22), (30,26; 50,32), (30,26; 50,33), (30,28; 50,29), (30,30; 50,37), (30,33; 50,42). Data jsou uložena v souboru ceny_vyrobku.sta.
Těchto 10 dvojic hodnot považujeme za realizace náhodného výběru (X_1,Y_1), ..., (X_{10},Y_{10}) z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ. Najděte bodové odhady kovariance σ_{12} (0,002547) a koeficientu korelace ρ (0,8248).
Upozornění: Povšimněte si, že tyto bodové odhady se nezmění, když od nákladů odečteme 30 a od cen 50.

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)

$$h = m + \frac{\sigma}{\sqrt{n}} u_{0,95} = 3000 + \frac{20}{\sqrt{16}} 1{,}65 = 3008{,}2$$

3008 h a 12 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.

Příklad k samostatnému řešení: Letecká společnost potřebuje odhadnout průměrný počet cestujících na její nově otevřené lince. Podle dosavadních zkušeností jsou údaje za 1. měsíc letů nadhodnocené, ale po tomto období se počet cestujících ustálí. Z tohoto důvodu společnost sledovala počty cestujících v prvních 20 dnech druhého měsíce po otevření linky. Údaje jsou uloženy v souboru cestující.sta a považujeme je za náhodný výběr rozsahu 20 z normálního rozložení s neznámou střední hodnotou μ a známou směrodatnou odchylkou $\sigma = 7$.

a) Najděte bodový odhad neznámé střední hodnoty (107).

b) Najděte 95% empirický interval spolehlivosti pro neznámou střední hodnotu
($103{,}93 < \mu < 110{,}07$ s pravděpodobností 0,95)

c) Najděte 95% empirický levostranný interval spolehlivosti pro neznámou střední hodnotu
($104{,}43 < \mu$ s pravděpodobností 0,95)

Najděte 95% empirický pravostranný interval spolehlivosti pro neznámou střední hodnotu

d) ($\mu < 109{,}57$ s pravděpodobností 0,95)