Téma 6.: Základní pojmy matematické statistiky

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.

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

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

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

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

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

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

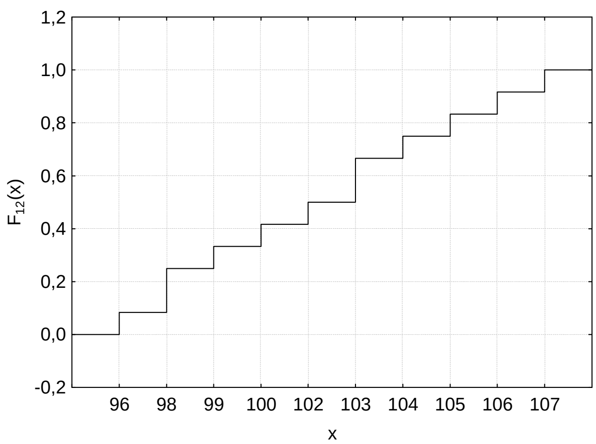
$$103 \le x < 104 : F_{12}(x) = \frac{8}{12} = 0{,}\overline{6}$$

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

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

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

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



Výpočet pomocí systému STATISTICA:
Otevřeme datový soubor 12_cen.sta.

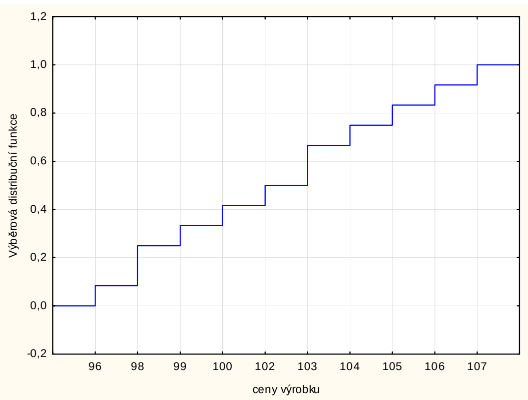
Výpočet realizace výběrového průměru a výběrového rozptylu:
Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – vybereme Průměr a Rozptyl – Výpočet. Dostaneme tabulku:

Proměnná	Popisné statistiky (12_cen.sta) Průměr	Rozptyl
ceny	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 108, do sloupce Kumulativní rel. četnost napíšeme 100). Proměnnou Kumulativní rel. četnost podělíme 100: do jejího Dlouhého jména napíšeme = v2/100.

Kategorie	Kumulativní rel. četnost
95	0,0000
96	0,0833
98	0,2500
99	0,3333
100	0,4167
102	0,5000
103	0,6667
104	0,7500
105	0,8333
106	0,9167
107	1,0000
108	1,0000

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 2.: Přírůstky cen akcií v % na burze v New Yorku u 10 náhodně vybraných společností dosáhly těchto hodnot: 10, 16, 5, 10, 12, 8, 4, 6, 5, 4.
Odhadněte střední hodnotu a směrodatnou odchylku růstu cen akcií a dále odhadněte pravděpodobnost růstu cen akcií aspoň o 8,5 %. Data jsou uložena v souboru akcie.sta.

Výsledky: Průměrný růst cen akcií odhadujeme na 8 % se směrodatnou odchylkou 3,97 %. Dále, u 40 % akcií vzrostla cena aspoň o 8,5 %.

Příklad 3.: 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:
Otevřeme datový soubor fosfor.sta.

Výpočet výběrové kovariance: Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza – Proměnné – Proměnné pro analýzu: X, YX – OK – OK – zvolíme Kovarianční matice. 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: Místo Kovarianční matice vybereme Korelační matice.

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.

Příklad 4.: Pět mužů zjistilo a zapsalo svou hmotnost (v kg) a výšku (v cm):

Číslo muže	1	2	3	4	5
Hmotnost	76	86	73	84	79
Výška	170	177	169	174	175

Najděte nestranný bodový odhad rozptylu hmotnosti, rozptylu výšky a kovariance hmotnosti a výšky. Vypočtěte rovněž realizaci výběrového koeficientu korelace hmotnosti a výšky.
Výsledky: Výběrový rozptyl hmotnosti se realizuje hodnotou 29,3, výběrový rozptyl výšky 11,5 a výběrová kovariance hmotnosti a výšky se realizuje hodnotou 16,5.

Výběrový koeficient korelace hmotnosti a výšky nabývá hodnoty 0,8989.

Příklad 5.: 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.

Příklad 6.: Víme, že výška hochů ve věku 9,5 až 10 let má normální rozložení s neznámou střední hodnotou μ a známým rozptylem $\sigma^2 = 39{,}112\ cm^2$. 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 $U = \dfrac{M-\mu}{\frac{\sigma}{\sqrt{n}}}$ ~ N(0, 1). Testová statistika tedy bude $T_0 = \dfrac{M-c}{\frac{\sigma}{\sqrt{n}}}$ a bude mít rozložení N(0, 1), pokud je nulová hypotéza pravdivá. Vypočítáme realizaci testového kritéria:

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

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.