Cvičení 6.: Bodové a intervalové odhady střední hodnoty, rozptylu a koeficientu korelace, test hypotézy o střední hodnotě při známém rozptylu

Příklad 1.: 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 se středními hodnotami μ_1, μ_2, rozptyly σ_1^2, σ_2^2 a koeficientem korelace ρ. Najděte bodové odhady těchto číselných charakteristik, tj. realizace výběrových průměrů, výběrových rozptylů a výběrového koeficientu korelace.

Výpočet pomocí systému STATISTICA:
Otevřeme datový soubor fosfor.sta o dvou proměnných X a Y 9 případech. V proměnné X jsou zjištěné hodnoty obsahu fosforu v půdě a v Y v obilných klíčcích.

Výpočet výběrových průměrů a výběrových rozptylů: Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X, Y – na záložce Detailní výsledky vybereme Průměr, Rozptyl – Výpočet. Dostaneme tabulku:

	Popisné statistiky (fosfor.sta)	
Proměnná	Průměr	Rozptyl
X	13	91,75
Y	80	284,25

Vidíme, že výběrové průměry veličin X, Y se realizují hodnotami 13 a 80, výběrové rozptyly pak nabývají hodnot 91,75 a 284,25.

Výpočet výběrového koeficientu korelace: Aktivujeme Popisné statistiky – Storno – Korelační matice – OK – 2 seznamy – 1. seznam proměnných X, 2. seznam proměnných Y – OK – Výpočet.

	Korelace (fosfor.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=9 (Celé případy vynechány u ChD)
Proměnná	Y
X	0,804989

Výběrový koeficient korelace veličin X, Y nabyl hodnoty 0,805, tedy mezi veličinami x, Y existuje silná přímá lineární závislost.

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 2.: 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 $c \in R$ 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 a V 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
$W = \left(t_{min}, K_{\alpha/2}(T)\right\rangle \cup \left\langle K_{1-\alpha/2}(T), t_{max}\right)$, 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 $h(\vartheta)$.
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 $p \le \alpha$, 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 $p = 2 \min\{P(T_0 \le t_0), P(T_0 \ge t_0)\}$.
Pro levostrannou alternativu $p = P(T_0 \le t_0)$.
Pro pravostrannou alternativu $p = P(T_0 \ge t_0)$.

Příklad 3.: 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: $\mu < 142$ (to je tvrzení lékaře) 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 $T_0 = \frac{M-c}{\frac{\sigma}{\sqrt{n}}}$ a bude mít rozložení N(0, 1), pokud je nulová hypotéza pravdivá. Vypočítáme realizaci testové statistiky:

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

Stanovíme kritický obor: $W = \langle -\infty, u_\alpha \rangle = \langle -\infty, u_{0,05} \rangle = \langle -\infty, -u_{0,95} \rangle = \langle -\infty, -1{,}6449 \rangle$.

Protože $-1{,}7773 \in W$, H_0 zamítáme na hladině významnosti 0,05. Tvrzení lékaře lze tedy akceptovat s rizikem omylu 5 %.

Výpočet pomocí systému STATISTICA

Otevřeme nový datový soubor o dvou proměnných t0 a kvantil a jednom případu.
Do Dlouhého jména proměnné t0 napíšeme =(139,13-142)/sqrt(39,112/15). Do Dlouhého jména proměnné kvantil napíšeme =VNormal (0,05;0;1). Dostaneme tabulku:

	1 t0	2 kvantil
1	-1,7773482	-1,6448536

Protože se testová statistika realizuje v kritickém oboru, nulovou hypotézu zamítáme 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: $h = 139{,}13 + \frac{\sqrt{39{,}112}}{\sqrt{15}} u_{0,95} = 139{,}13 + \frac{\sqrt{39{,}112}}{\sqrt{15}} 1{,}645 = 141{,}79$.

Protože $142 \notin (-\infty; 141{,}79)$, H_0 zamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA

Otevřeme nový datový soubor o jedné proměnné h a jednom případu.
Do Dlouhého jména proměnné h napíšeme =139,13+sqrt(39,112/15)*VNormal(0,95;0;1)

	1 h
1	141,786052

Protože číslo 142 nepatří do intervalu $(-\infty; 141{,}79)$, 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.

Výpočet pomocí systému STATISTICA

Otevřeme nový datový soubor o jedné proměnné p a jednom případu.

Do Dlouhého jména proměnné p napíšeme =INormal(-1,7773;0;1)

	1 p
1	0,03775945

Protože p-hodnota je menší než 0,05, H_0 zamítáme na hladině významnosti 0,05.