

Téma 9: Parametrické úlohy o dvou nezávislých náhodných výběrech z normálních rozložení

Úkol 1.: Vlastnosti rozdílu výběrových průměrů ze dvou normálních rozložení

Jsou dány dva nezávislé náhodné výběry, první pochází z rozložení $N(2; 1,5)$ a má rozsah 10, druhý pochází z rozložení $N(3; 4)$ a má rozsah 5. Jaká je pravděpodobnost, že výběrový průměr 1. výběru bude menší než výběrový průměr 2. výběru?

Návod:

Počítáme $P(M_1 < M_2) = P(M_1 - M_2 < 0) = \Phi(0)$,

kde $\Phi(x)$ je distribuční funkce statistiky $M_1 - M_2$.

Statistika $M_1 - M_2$ se řídí rozložením $N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$, kde $\mu_1 - \mu_2 = 2 - 3 = -1$,

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{1,5}{10} + \frac{4}{5} = 0,95, \text{ tj. statistika } M_1 - M_2 \sim N(-1; 0,95).$$

Otevřeme nový datový soubor o jedné proměnné a jednom případě. Do Dlouhého jména této proměnné napíšeme = INormal(0;-1;sqrt(0,95)). Dostaneme výsledek 0,847549.

Úkol 2.: Intervaly spolehlivosti pro parametrické funkce $\mu_1 - \mu_2, \sigma_1^2 / \sigma_2^2$

Bylo vylosováno 11 stejně starých selat téhož plemene. Šesti z nich byla předepsána výkrmná dieta č. 1 a zbylým pěti výkrmná dieta č. 2. Průměrné denní přírůstky v Dg za dobu půl roku jsou následující:

dieta č. 1: 62, 54, 55, 60, 53, 58

dieta č. 2: 52, 56, 49, 50, 51.

Zjištěné hodnoty považujeme za realizace dvou nezávislých náhodných výběrů pocházejících z rozložení $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$.

a) Sestrojte 95% empirický interval spolehlivosti pro podíl rozptylů.

b) Za předpokladu, že data pocházejí z rozložení $N(\mu_1, \sigma^2)$ a $N(\mu_2, \sigma^2)$, sestrojte 95% empirický interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$.

Návod:

Načteme datový soubor dve_diety.sta o 2 proměnných hmotnost a dieta a 11 případech. Pomocí Popisných statistik zjistíme realizace výběrových průměrů, výběrových rozptylů a výběrových směrodatných odchylek.

Pro první dietu:

Proměnná	Popisné statistiky (Tabulka1)			
	Zhrnout podmínku: v2=1			
	N platných	Průměr	Rozptyl	Sm.odch.
hmotnost	6	57,00000	12,80000	3,577709

Pro druhou dietu:

Proměnná	Popisné statistiky (Tabulka1)			
	Zhrnout podmínku: v2=2			
	N platných	Průměr	Rozptyl	Sm.odch.
hmotnost	5	51,60000	7,300000	2,701851

ad a)

Meze 100(1- α)% empirického intervalu spolehlivosti pro podíl rozptylů jsou:

$$(d, h) = \left(\frac{s_1^2 / s_2^2}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)}, \frac{s_1^2 / s_2^2}{F_{\alpha/2}(n_1 - 1, n_2 - 1)} \right).$$

Otevřeme nový datový soubor o dvou proměnných d a h a jednom případě.

Do Dlouhého jména proměnné d napíšeme

$$=(12,8/7,3)/VF(0,975;5;4)$$

(Funkce VF(x;ný;omega) počítá x-kvantil Fisherova – Snedecorova rozložení F(ný, omega).)

Do Dlouhého jména proměnné h napíšeme

$$=(12,8/7,3)/VF(0,025;5;4)$$

	1	2
	d	h
1	0,187242	12,9541

S pravděpodobností aspoň 0,95 tedy platí: $0,1872 < \sigma_1^2 / \sigma_2^2 < 12,954$.

ad b) Meze 100(1- α)% empirického intervalu spolehlivosti pro rozdíl středních hodnot (v případě, že rozptyly neznáme, ale víme, že jsou shodné) jsou:

$$(d, h) = (m_1 - m_2 - s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1+n_2-2), m_1 - m_2 + s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1+n_2-2)).$$

Otevřeme nový datový soubor o dvou proměnných d a h a jednom případě.

Do Dlouhého jména proměnné d napíšeme

$$=57-51,6-\text{sqrt}((5*12,8+4*7,3)/9)*\text{sqrt}((1/6)+(1/5))*VStudent(0,975;9)$$

Do Dlouhého jména proměnné h napíšeme

$$=57-51,6+\text{sqrt}((5*12,8+4*7,3)/9)*\text{sqrt}((1/6)+(1/5))*VStudent(0,975;9)$$

	1	2
	d	h
1	0,991963	9,808037

S pravděpodobností aspoň 0,95 tedy $0,99 Dg < \mu_1 - \mu_2 < 9,81 Dg$.

Úkol k samostatnému řešení: Jsou dány dva nezávislé náhodné výběry o rozsazích $n_1 = 25$, $n_2 = 10$, první pochází z rozložení $N(\mu_1, \sigma_1^2)$, druhý z rozložení $N(\mu_2, \sigma_2^2)$, kde parametry μ_1 , μ_2 , σ_1^2 , σ_2^2 neznáme. Byly vypočteny realizace výběrových rozptylů: $s_1^2 = 1,7482$, $s_2^2 = 1,7121$. Sestrojte 95% empirický interval spolehlivosti pro podíl rozptylů.

Výsledek:

$$0,28 < \sigma_1^2 / \sigma_2^2 < 2,76 \text{ s pravděpodobností aspoň } 0,95.$$

Úkol 3.: Testování hypotéz o parametrických funkcích $\mu_1 - \mu_2$, σ_1^2 / σ_2^2

Pro datový soubor z úkolu 2 testujte na hladině významnosti 0,05 hypotézu, že

- rozptyly hmotnostních přírůstků selat při obou výkrmných dietách jsou shodné
- obě výkrmné diety mají stejný vliv na hmotnostní přírůstky selat.

Návod:

Provedeme dvouvýběrový t-test současně s testem o shodě rozptylů:

Statistika – Základní statistiky a tabulky – t-test, nezávislé, dle skupin – OK, Proměnné –
Závislé proměnné hmotnost, Grupovací proměnná dieta – OK.

t-testy; grupováno: dieta (Tabulka1)											
Skup. 1: 1											
Skup. 2: 2											
Proměnná	Průměr 1	Průměr 2	t	sv	p	Poč.plat 1	Poč.plat. 2	Sm.odch. 1	Sm.odch. 2	F-poměr Rozptyly	p Rozptyly
hmotnost	57,00000	51,60000	2,771222	9	0,021710	6	5	3,577709	2,701851	1,753425	0,606345

Testová statistika pro test shody rozptylů se realizuje hodnotou 1,7534, odpovídající p-hodnota je 0,6063, tedy na hladině významnosti 0,05 nezamítáme hypotézu o shodě rozptylů. (Upozornění: v případě zamítnutí hypotézy o shodě rozptylů je zapotřebí v tabulce t-testu pro nezávislé vzorky dle skupin zaškrtnout volbu Test se samostatnými odhady rozptylu.)

Dále z tabulky plyne, že testová statistika pro test shody středních hodnot se realizuje hodnotou 2,7712, počet stupňů volnosti je 9, odpovídající p-hodnota 0,0217, tedy hypotézu o shodě středních hodnot zamítáme na hladině významnosti 0,05. Znamená to, že s rizikem omylu nejvýše 5% se prokázalo, že obě výkrmné diety se liší účinností.

Tabulku ještě doplníme krabicovými diagramy. Na záložce Details zaškrtneme krabicový graf a vybereme volbu Průměr/SmOdch/Min-Max.



Upozornění: Dvouvýběrový t-test lze v systému STATISTICA provést ještě jiným způsobem, který je vhodný zvláště tehdy, známe-li realizace výběrových průměrů a výběrových směrodatných odchylek.

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma průměry (normální rozdělení) – do políčka Pr1 napíšeme 57, do políčka SmOd1 napíšeme 3,5777, do políčka N1 napíšeme 6, do políčka Pr2 napíšeme 51,6, do políčka SmOd1 napíšeme 2,7019, do políčka N1 napíšeme 5 - Výpočet.

Testy rozdílů: r, %, průměry: Tabulka1

Poslat/tisknout výsledky každ. výpočtu do okna protokolu

Rozdíl mezi dvěma korelačními koeficienty

r1: 0,00 N1: 10 p: 1,0000 Jednostr. Oboustr.

r2: 0,00 N2: 10

Rozdíl mezi dvěma průměry (normální rozdělení)

Pr1: 57, SmOd1: 3,5777 N1: 6 p: ,0217

Pr2: 51,6 SmOd2: 2,7019 N2: 5 Jednostr. Oboustr.

Výběrový průměr vs. střední hodnota

Rozdíl mezi dvěma poměry

P 1: ,50000 N1: 10 p: 1,0000 Jednostr. Oboustr.

P 2: ,50000 N2: 10

Dostaneme p-hodnotu 0,0217, tedy zamítáme nulovou hypotézu na hladině významnosti 0,05.

Úkol k samostatnému řešení: Do systému STATISTICA načtěte datový soubor studentky.sta, který obsahuje údaje o výšce 48 studentek VŠE v Praze (proměnná vyska) a obor jejich studia (1 – národní hospodářství, 2 – informatika).

a) Na hladině významnosti 0,1 testujte hypotézu o shodě rozptylů výšek studentek v daných dvou oborech studia.

b) Na hladině významnosti 0,1 testujte hypotézu o shodě středních hodnot výšek studentek v daných dvou oborech studia.

(Výpočet doplňte krabicovými diagramy.)

Výsledek:

ad a) Protože p-hodnota F-testu je 0,1249, což je větší než hladina významnosti 0,1, nulovou hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,1.

ad b) Protože p-hodnota dvouvýběrového t-testu je 0,0878, což je menší než hladina významnosti 0,1, nulovou hypotézu o shodě středních hodnot zamítáme na hladině významnosti 0,1.