

## Téma 8: Parametrické úlohy o dvou nezávislých náhodných výběrech z normálních rozložení a jednom náhodném výběru z alternativního 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  $P(M_1 < M_2) = P(M_1 - M_2 < 0) = \Phi\left(\frac{\bar{x}}{s}\right)$ ,

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:

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

Pro druhou dietu:

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

ad a)

Meze 100(1- $\alpha$ )% 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-quantil 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- $\alpha$ )% 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  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1^2$ ,  $\sigma_2^2$  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  $\mu_1 - \mu_2$ ,  $\sigma_1^2 / \sigma_2^2$**

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.

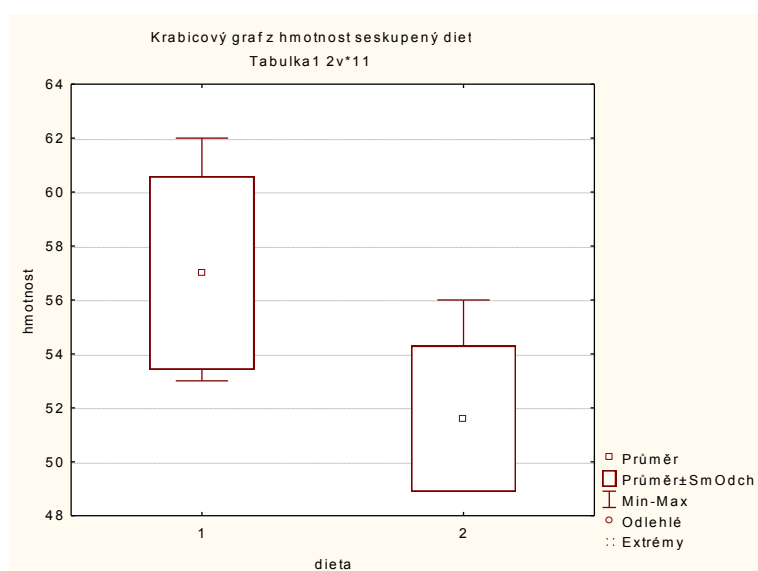
**Návod:**

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.

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.

**Úkol 4.: Asymptotický interval spolehlivosti pro parametr  $\theta$  alternativního rozložení**

Může politická strana, pro niž se v předvolebním průzkumu vyslovilo 60 z 1000 dotázaných osob, očekávat se spolehlivostí 0,95, že by v této době ve volbách překročila 5% hranici pro vstup do parlamentu?

**Návod:** Zavedeme náhodné veličiny  $X_1, \dots, X_{1000}$ , přičemž  $X_i = 1$ , když i-tá osoba se vysloví pro danou politickou stranu a  $X_i = 0$  jinak,  $i = 1, \dots, 1000$ . Tyto náhodné veličiny tvoří náhodný výběr z rozložení  $A(\theta)$ . V tomto případě  $n = 1000$ ,  $m = 60/1000 = 0,06$ ,  $\alpha = 0,05$ ,  $u_{1-\alpha} = u_{0,95} = 1,645$ .

Ověření podmínky  $n\theta(1-\theta) > 9$ : parametr  $\theta$  neznáme, musíme ho nahradit výběrovým průměrem. Pak  $1000 \cdot 0,06 \cdot 0,94 = 56,4 > 9$ .

95% levostranný interval spolehlivosti pro  $\theta$  je

$$\left( m - \sqrt{\frac{m(m-1)}{n}} u_{1-\alpha} ; \infty \right) = \left( 0,06 - \sqrt{\frac{0,06(1-0,06)}{1000}} u_{0,95} ; \infty \right). \text{ V našem případě}$$

$$d = 0,06 - \sqrt{\frac{0,06 \cdot 0,94}{1000}} \cdot 1,645 = 0,0476$$

S pravděpodobností přibližně 0,95 tedy  $\vartheta > 0,048$ . Protože tento interval zahrnuje i hodnoty nižší než 0,05, nelze vyloučit, že strana získá méně než 5% hlasů.

### Postup ve STATISTICE:

Asymptotický způsob: Vytvoříme datový soubor o jedné proměnné (nazveme ji d) a o jednom případě. Do Dlouhého jména proměnné d napíšeme

$$=0,06-\text{sqrt}(0,06*0,94/1000)*\text{VNormal}(0,95;0;1)$$

Vyjde 0,047647.

Přibližný způsob: Do nového datového souboru o jedné proměnné X a 1000 případech uložíme 60 jedniček (indikují volbu dané politické strany) a 940 nul (indikují volbu jiné politické strany).

Statistika – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme Meze spolehl. prům. – Interval 90,00 – Výpočet.

Dostaneme tabulku:

Proměnná	Popisné statistiky (Tabulka1)			
	N platných	Průměr	Int. spolehl. -90,000%	Int. spolehl. 90,000
X	1000	0,060000	0,047630	0,072370

Protože dolní mez oboustranného 90% intervalu spolehlivosti pro střední hodnotu je shodná s dolní mezí 95% jednostranného intervalu spolehlivosti, můžeme konstatovat, že voliči budou volit danou politickou stranu s pravděpodobností aspoň 4,76%. Na základě uvedených dat strana tedy nemá zaručeno, že překročí 5% hranici pro vstup do parlamentu.

**Úkol k samostatnému řešení:** Přírůstky cen akcií na burze (v %) u 10 náhodně vybraných společností dosáhly těchto hodnot: 10, 16, 5, 10, 12, 8, 4, 6, 5, 4. Sestrojte 95% asymptotický empirický interval spolehlivosti pro pravděpodobnost, že přírůstek ceny akcie překročí 8,5%.

**Výsledek:**  $0,096 < \vartheta < 0,704$  s pravděpodobností aspoň 0,95.

Znamená to, že pravděpodobnost, že přírůstek ceny akcie překročí 8,5%, je aspoň 9,6% a nanejvýš 70,4% (při spolehlivosti 95%).

### Úkol 5.: Testování hypotézy o parametru $\vartheta$ alternativního rozložení

Určitá cestovní kancelář organizuje zahraniční zájezdy podle individuálních přání zákazníků. Z několika minulých let ví, že 30% všech takto organizovaných zájezdů má za cíl zemi X. Po zhoršení politických podmínek v této zemi se cestovní kancelář obává, že se zájem o tuto zemi mezi zákazníky sníží. Ze 150 náhodně vybraných zákazníků v tomto roce má 38 za cíl právě zemi X. Potvrzují nejnovější data pokles zájmu o tuto zemi? Volte hladinu významnosti 0,05.

**Návod:** Máme náhodný výběr  $X_1, \dots, X_{150}$  z rozložení  $A(0,3)$ . Testujeme  $H_0: \vartheta = 0,3$  proti jednostranné alternativě  $H_1: \vartheta < 0,3$ . V tomto případě je testovým kritériem statistika

$$T_0 = \frac{M - \vartheta}{\sqrt{\frac{\vartheta(1 - \vartheta)}{n}}}, \text{ která v případě platnosti nulové hypotézy má asymptoticky rozložení } N(0,1).$$

Musíme ověřit splnění podmínky  $n\vartheta(1 - \vartheta) > 9$ :  $150 \cdot 0,3 \cdot 0,7 = 31,5 > 9$ . Vypočteme realizaci

$$\text{testového kritéria: } t_0 = \frac{m - \vartheta}{\sqrt{\frac{\vartheta(1 - \vartheta)}{n}}} = \frac{\frac{38}{150} - 0,3}{\sqrt{\frac{0,3(1 - 0,3)}{150}}} = -0,24722. \text{ Kritický obor:}$$

$W = \langle -\infty, -u_{1-\alpha} \rangle = \langle -\infty, -1,645 \rangle$ . Protože testové kritérium nepatří do kritického oboru,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

### Postup ve STATISTICE:

Asymptotický způsob: Vytvoříme datový soubor o dvou proměnných (nazveme je  $t_0$  a kvantil) a jednom případě. Vypočteme realizaci testového kritéria tak, že do Dlouhého jména proměnné  $t_0$  napíšeme

$$=(38/150-0,3)/\text{sqrt}(0,3*0,7/150)$$

Do Dlouhého jména proměnné kvantil napíšeme

$$=\text{VNormal}(0,95;0;1)$$

Tím získáme kvantil  $u_{0,95}$ .

	1	2
	$t_0$	kvantil
1	-1,24721913	1,644854

Jelikož realizace testového kritéria  $t_0 = -1,24721913$  nepatří do kritického oboru

$W = \langle -\infty, -1,644854 \rangle$ ,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

Přibližný způsob: Do nového datového souboru o jedné proměnné  $X$  a 150 případech uložíme 38 jedniček (indikují zájem o danou zemi) a 112 nul (indikují nezáměr o danou zemi).

Statistika – Základní statistiky a tabulky – t-test, samost. vzorek – OK – Proměnné  $X$  – OK,

Test všech průměrů vůči 0,3 – Výpočet.

Proměnná	Test průměrů v úči referenční konstantě (hodnotě) (Tabulka4)							
	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
X	0,253333	0,436377	150	0,035630	0,300000	-1,30976	149	0,192294

Hodnota testové statistiky je při tomto přibližném způsobu -1,30976. Odpovídající p-hodnota je 0,1923, ovšem to je p-hodnota pro oboustranný test. Tuto p-hodnotu tedy musíme dělit dvěma a dostaneme 0,0961. Na asymptotické hladině významnosti 0,05 nelze zamítnout hypotézu, že zájem o danou zemi se nezměnil.