Cvičení 5.: Parametrické úlohy
o dvou nezávislých výběrech z normálních rozložení

Do programu STATISTICA načtěte soubor studentky.sta, který obsahuje údaje o 48 náhodně vybraných studentkách VŠE v Praze:
1. sloupec – výška, 2. sloupec – známka z matematiky v 1. semestru, 3. sloupec – obor studia (1 – národní hospodářství, 2 – informatika).
Ověření normality výšky ve skupině studentek oboru národní hospodářství a oboru informatika bylo provedeno ve cvičení 4..

Úkol 1.: Sestrojte 95% interval spolehlivosti pro podíl rozptylů výšek studentek oboru nh a inf.

Návod:
K datovému souboru přidáme další dvě proměnné DM a HM pro výpočet dolní a horní meze intervalu spolehlivosti. Do Dlouhého jména těchto proměnných zapíšeme vzorce pro dolní a horní mez intervalu spolehlivosti pro podíl rozptylů (viz skripta Základní statistické metody, Věta 7.1.2.1., bod 4 (a)). Výběrové rozptyly pro 1. a 2. výběr zjistíme pomocí Popisných statistik.
Interval spolehlivosti je

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

přičemž první výběr tvoří studentky nh, druhý výběr studentky inf.

Proměnná	Souhrnné výsledky Popisné statistiky (vyska)		
	Z	N platných	Rozptyl
X	nh	28	41,18915
X	inf	20	20,72632

Do Dlouhého jména proměnné DM napíšeme:
=(41,18915/20,72622)/VF(0,975;27;19)
(Funkce VF(x;ný;omega) počítá x-kvantil Fisherova – Snedecorova rozložení F(ný, omega).)
Do Dlouhého jména proměnné HM napíšeme:
=(41,18915/20,72622)/VF(0,025;27;19)
Vyjde DM = 0,821186, HM = 4,513831.
S pravděpodobností aspoň 0,95 tedy platí: $0{,}821 < \sigma_1^2 / \sigma_2^2 < 4{,}514$.

Úkol 2.: Na hladině významnosti 0,05 testujte hypotézu, že rozptyly výšek studentek oboru nh a inf jsou shodné.

Návod:

Jedná se o F-test, kdy testujeme hypotézu $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ proti oboustranné alternativě

$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

1. způsob: lze využít výsledku 1. úkolu. 95% interval spolehlivosti pro podíl rozptylů obsahuje číslo 1, tedy hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05.

2. způsob: F-test je implementován ve STATISTICE.
Statistiky – Základní statistiky/tabulky – t-test, nezávislé, podle skupin - OK, Proměnné – Závislé proměnné X, Grupovací proměnná Z – OK – Výpočet

Proměnná	t-testy; grupováno: Z: obor studia (vyska) Skup. 1: nh: narodni hospodarstvi Skup. 2: inf: informatika Průměr nh	Průměr inf	t	sv	p	Poč.plat nh	Poč.plat. inf	Sm.odch. nh	Sm.odch. inf	F-poměr Rozptyly	p Rozptyly
X	169,8214	166,9000	1,744008	46	0,087837	28	20	6,417878	4,552616	1,987288	0,124925

Komentář: Ve výstupní tabulce nás zajímá hodnota testové statistiky F-testu (v našem případě 1,987288) a odpovídající p-hodnota: 0,124925. Protože p-hodnota je větší než hladina významnosti $\alpha = 0,05$, nelze na hladině významnosti 0,05 zamítnout nulovou hypotézu.

Úkol 3.: Sestrojte 95% interval spolehlivosti pro rozdíl středních hodnot výšek studentek oboru nh a inf.

Návod:
Meze intervalu spolehlivosti pro rozdíl středních hodnot lze získat pomocí aplikace pro dvouvýběrový t-test.
Statistiky – Základní statistiky/tabulky – t-test, nezávislé, podle skupin - OK, Proměnné – Závislé proměnné X, Grupovací proměnná Z – OK – na záložce Možnosti zaškrtneme Meze spol. pro odhady, ponecháme implicitní spolehlivost 0,95 – Výpočet. V posledních dvou sloupcích výstupní tabulky jsou uvedeny meze intervalu spolehlivosti. Vidíme, že s pravděpodobností aspoň 0,95 platí, že $-0,45 \text{ cm} < \mu_1 - \mu_2 < 6,29 \text{ cm}$.

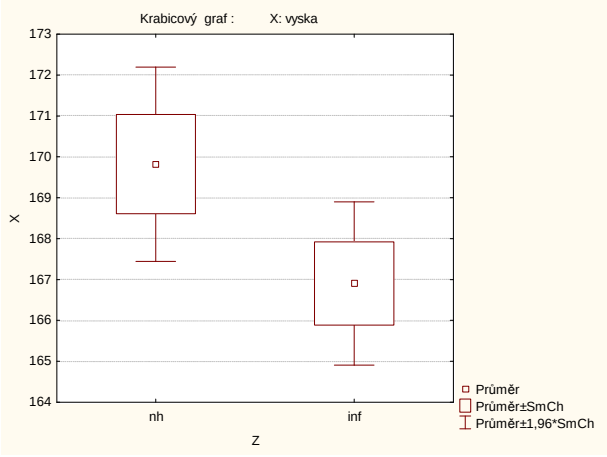
Úkol 4.: Na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty výšek studentek oboru nh a inf jsou shodné. Výpočet doplňte krabicovými diagramy.

Návod:
Jedná se o dvouvýběrový t-test, kdy testujeme hypotézu $H_0 : \mu_1 - \mu_2 = 0$ proti oboustranné alternativě $H_1 : \mu_1 - \mu_2 \neq 0$
1. způsob: lze využít výsledku 6. úkolu. 95% interval spolehlivosti pro rozdíl středních hodnot obsahuje číslo 0, tedy hypotézu o shodě středních hodnot nezamítáme na hladině významnosti 0,05.
2. způsob: dvouvýběrový t-test je implementován ve STATISTICE.
Statistiky – Základní statistiky/tabulky – t-test, nezávislé, podle skupin - OK, Proměnné – Závislé proměnné X, Grupovací proměnná Z – OK – Výpočet

Proměnná	t-testy; grupováno: Z: obor studia (vyska) Skup. 1: nh: narodni hospodarstvi Skup. 2: inf: informatika Průměr nh	Průměr inf	t	sv	p	Poč.plat nh	Poč.plat. inf	Sm.odch. nh	Sm.odch. inf	F-poměr Rozptyly	p Rozptyly
X	169,8214	166,9000	1,744008	46	0,087837	28	20	6,417878	4,552616	1,987288	0,124925

Komentář: Ve výstupní tabulce najdeme hodnotu testového kritéria ($t_0 = 1,744006$) a odpovídající p-hodnotu. Protože p-hodnota = 0,087837 je větší než hladina významnosti 0,05, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Konstrukce krabicových diagramů: V tabulce t-test, nezávislé, podle skupin zvolíme Krabicový diagram. Dostaneme graf:



Komentář: Ze vzhledu krabicových diagramů je vidět, že rozložení výšek v obou skupinách je vcelku symetrické kolem průměru, odlehlé ani extrémní hodnoty se nevyskytují, variabilita vyjádřená směrodatnou odchylkou se liší jen nepatrně a průměrná výška ve skupině studentek oboru inf je o něco menší než ve skupině studentek oboru nh.

Poznámka: Protože F-test neprokázal odlišnost rozptylů, mohli jsme ve STATISTICE použít variantu dvouvýběrového t-testu se shodnými rozptyly. Pokud by však F-test zamítl na dané hladině významnosti hypotézu o shodě rozptylů, museli bychom zvolit variantu dvouvýběrového t-testu se separovanými odhady rozptylů.

Úkol k samostatnému řešení: Hejtman Jihomoravského kraje chtěl porovnat situaci svého kraje s ostatními moravskými kraji vzhledem ke znečištění ovzduší oxidem siřičitým, oxidy dusíku a oxidem uhelnatým. Požádal proto Stranu zelených, aby na základě údajů ze Statistické ročenky ČSÚ za léta 2000 až 2006 její experti provedli příslušnou analýzu. Roční měrné emise jsou uvedeny v tunách na km^2. Data jsou uložena v souboru znecisteni.sta. Vaším úkolem bude provést srovnání středních hodnot znečištění oxidem siřičitým v Jihomoravském kraji a Olomouckém kraji. Na hladině významnosti 0,05 ověřte normalitu dat, homogenitu rozptylů a proveďte test shody středních hodnot. Výpočty doplňte krabicovými grafy a rovněž vypočtěte Cohenův koeficient věcného účinku.

Výsledek:

Průměrné znečištění oxidem siřičitým v Jihomoravském kraji v letech 2000 – 2006 je 0,51, v Olomouckém 1,23. Testová statistika pro test shody rozptylů se realizuje hodnotou 1,94117, odpovídající p-hodnota je 0,4397, 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 na záložce Možnosti zaškrtnout volbu Test se samostatnými odhady rozptylu.)

Testová statistika pro test shody středních hodnot se realizuje hodnotou -12,247, počet stupňů volnosti je 12, odpovídající p-hodnota je velmi blízká 0, tedy hypotézu o shodě středních hodnot zamítáme na hladině významnosti 0,05. Znamená to, že s rizikem omylu nejvýše 5% se prokázal rozdíl ve středních hodnotách znečištění oxidem siřičitým v Jihomoravském a Olomouckém kraji.

Cohenův koeficient nabyl hodnoty 6,55, vliv kraje na velikost znečištění oxidem siřičitým je tedy velký. (Výpočet Cohenova koeficientu je možno provést pomocí programu Cohen.svb.)