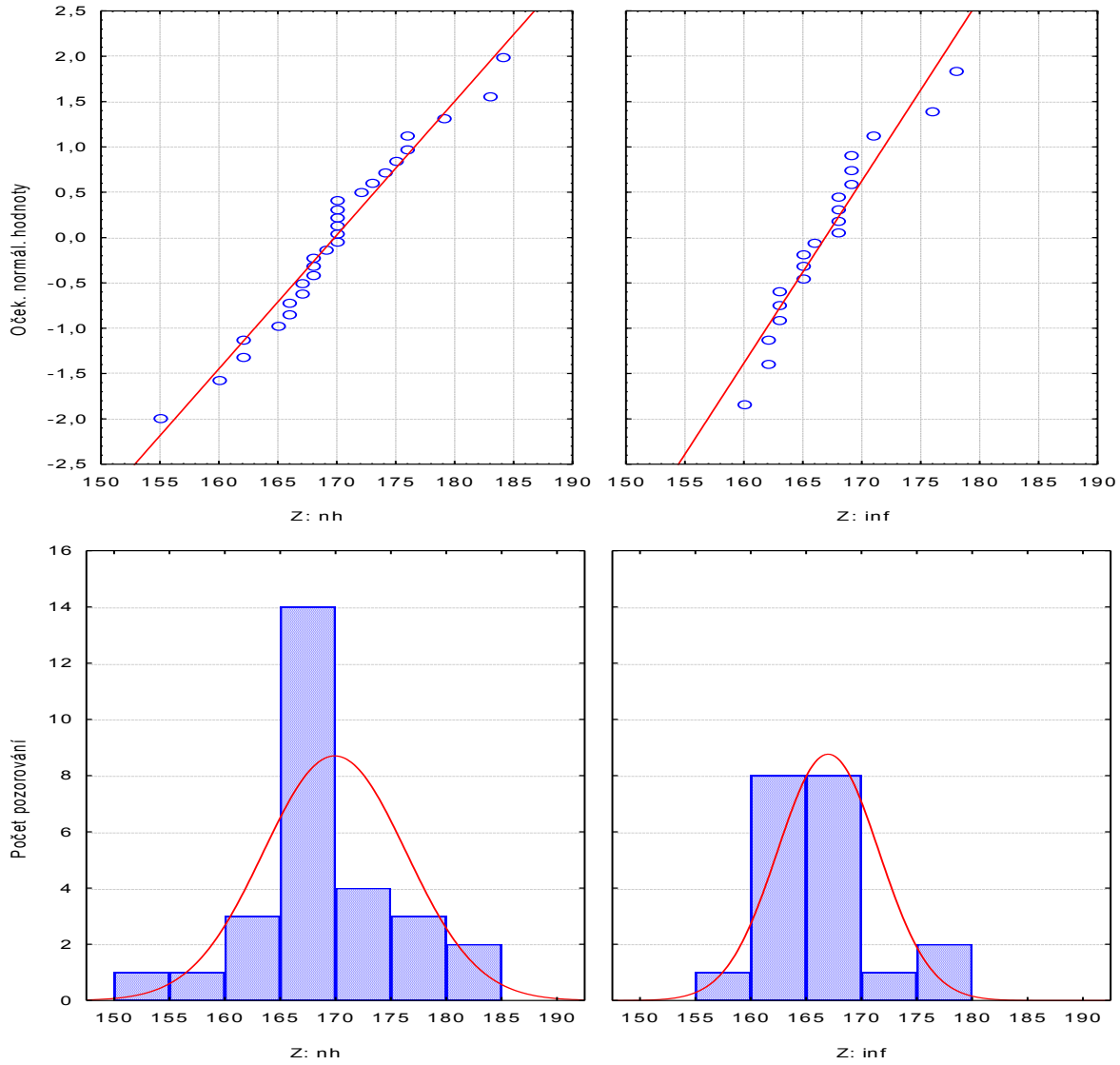
Cvičení 5.: Parametrické úlohy
o dvou nezávislých výběrech z normálních rozložení

Úkol 1.: 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).

Úkol 2.: Orientačně ověřte normalitu výšky ve skupině studentek oboru národní hospodářství a oboru informatika vykreslením N-P plotu a histogramu.

Návod:
Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnné X – na záložce Kategorizovaný zaškrtneme Kategorie X Zapnuto – Změnit proměnnou – Z - OK – OK.
Podobně pro histogram.



Komentář: Grafy svědčí o mírném narušení normality, jedná se o mírné kladné zešikmení.

Nyní provedeme testy normality.
Statistiky – Základní statistiky/tabulky – Tabulky četností – OK – Select cases – Zapnout filtr – některé vybrané pomocí Z=1 – OK – Proměnná X – OK - Normalita - zaškrtneme Liliefors test, Shapiro-Wilk's test - Testy normality. Dostaneme tyto výsledky:

Pro studentky oboru nh

	Testy normality (studentky.sta) Zhrnout podmínku: Z=1				
Proměnná	N	max D	Lilliefors p	W	p
X: vyska	28	0,167473	p < ,05	0,970969	0,606793

Pro studentky oboru inf

	Testy normality (studentky.sta) Zhrnout podmínku: Z=2				
Proměnná	N	max D	Lilliefors p	W	p
X: vyska	20	0,172301	p < ,15	0,922747	0,111924

Komentář: Vypočtenou p-hodnotu porovnáváme se zvolenou hladinou významnosti testu (většinou volíme $\alpha = 0{,}05$). Je-li vypočtená p-hodnota $\leq \alpha$, pak hypotézu o normalitě zamítáme na hladině významnosti α. V našem případě dojde k zamítnutí hypotézy o normalitě výšky na hladině významnosti 0,05 pouze u Lilieforsova testu pro studentky oboru nh.

Úkol 3.: Sestrojte 95% empirický interval spolehlivosti pro střední hodnotu výšky
a) studentek oboru nh,
b) studentek oboru inf.

Návod:
Vzhledem k tomu, že data lze považovat za realizace náhodného výběru z normálního rozložení, můžeme použít postup pro konstrukci intervalu spolehlivosti pro střední hodnotu, když rozptyl neznáme. Výpočet je implementován ve STATISTICE. Meze 95% intervalu spolehlivosti pro střední hodnotu proměnné X zjistíme pomocí Popisných statistik, kde zaškrtneme Meze spoleh. prům.

	Popisné statistiky (studentky.sta) Zhrnout podmínku: Z=1	
Proměnná	Int. spolehl. -95,000%	Int. spolehl. 95,000
X	167,3328	172,3100

	Popisné statistiky (studentky.sta) Zhrnout podmínku: Z=2	
Proměnná	Int. spolehl. -95,000%	Int. spolehl. 95,000
X	164,7693	169,0307

Komentář: S pravděpodobností aspoň 95% lze očekávat, že střední hodnoty výška studentek oboru národní hospodářství leží v intervalu 167,3 cm až 172,3 cm, zatímco u studentek oboru informatika v intervalu 164,8 cm až 169 cm.

Úkol 4.: 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	Popisné statistiky (Tema7) Include condition: z=1	
	N platných	Rozptyl
X	28	41,18915

Proměnná	Popisné statistiky (Tema7) Include condition: z=2	
	N platných	Rozptyl
X	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 5.: 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 $H_0 : \frac{\sigma_1}{\sigma_2} = 1$ proti oboustranné alternativě

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

1. způsob: lze využít výsledku 4. 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 skupn - OK, Proměnné – Závislé proměnné X, Grupovací proměnná Z – OK – Výpočet

Proměnná	t-testy; grupováno: Z: obor studia (Tema7) 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.
S rizikem omylu nanejvýš 5% se tedy neprokázalo, že by rozptyly výšek studentek oborů nh a inf byly odlišné.

Úkol 6.: Sestrojte 95% interval spolehlivosti pro rozdíl středních hodnot výšek studentek oboru nh a inf.

Návod:
K datovému souboru přidáme další dvě proměnné DM1 a HM1 pro výpočet dolní a horní meze intervalu spolehlivosti. Do Dlouhého jména těchto proměnných zapíšeme vzorce pro dolní a horní mez intervalu spolehlivosti pro rozdíl středních hodnot (viz skripta Základní statistické metody, Věta 7.1.2.1., bod 2 (a)). Výběrové průměry a výběrové rozptyly pro první a druhý výběr zjistíme pomocí Popisných statistik.
Oboustranný interval spolehlivosti pro $\mu_1 - \mu_2$, když rozptyly σ_1^2, σ_2^2 neznáme, ale víme, že jsou shodné, je:

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

$s_*^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ je vážený průměr výběrových rozptylů.

Do Dlouhého jména proměnné DM1 napíšeme
=169,8214-166,9-
sqrt((27*41,18915+19*20,72622)/46)*sqrt((1/28)+(1/20))*VStudent(0,975;46)
Do Dlouhého jména proměnné HM1 napíšeme
=169,8214-166,9+
sqrt((27*41,18915+19*20,72622)/46)*sqrt((1/28)+(1/20))*VStudent(0,975;46)
Vyjde DM1 = -0,450446, HM1 = 6,293246
S pravděpodobností aspoň 0,95 tedy -0,45 cm < $\mu_1 - \mu_2$ < 6,29 cm.

Úkol 7.: Na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty výšek studentek oboru nh a inf jsou shodné. Výpočet doplňte krabicovými diagramy.

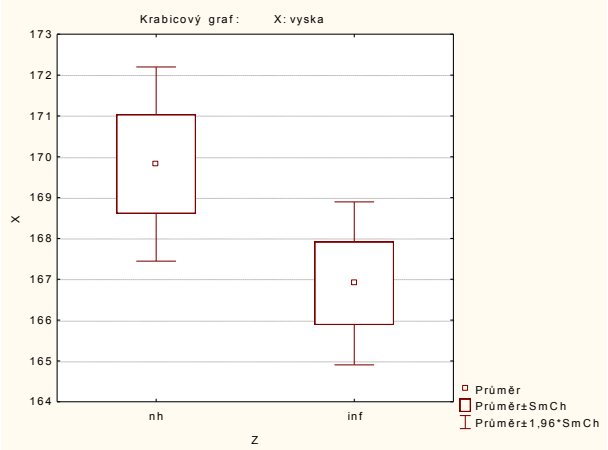
Návod:
Jedná se o dvouvýběrový t-test, kdy testujeme hypotézu $H_0 : \mu_1 - \mu_2 = 0$ proti oboustranné alternativě $H_1 : \mu_1 - \mu_2 \neq 0$

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

	t-testy; grupováno: Z: obor studia (Tema7) Skup. 1: nh: narodni hospodarstvi Skup. 2: inf: informatika										
Proměnná	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. S rizikem omylu nanejvýš 5% se tedy neprokázal rozdíl mezi středními hodnotami výšek studentek oborů nh a inf.

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.)