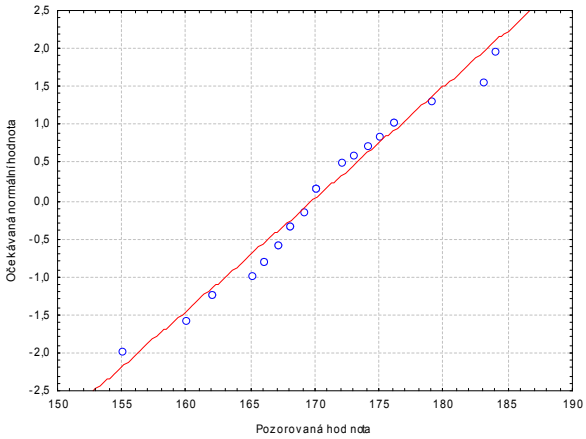
Téma 7: Parametrické úlohy o dvou nezávislých náhodných výběrech

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).

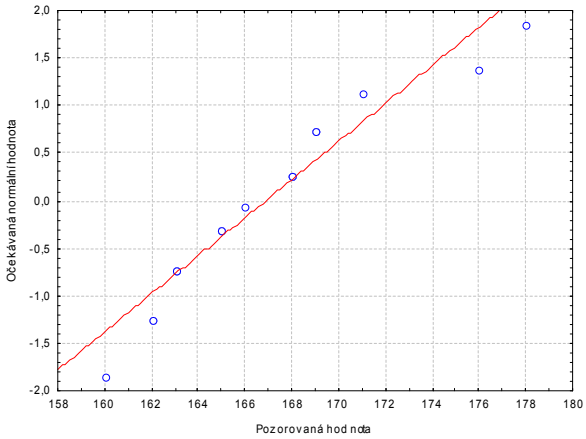
Úkol 2.: Orientačně ověřte normalitu výšky ve skupině studentek oboru národní hospodářství a oboru informatika vykreslením N-P plotu a histogramu.

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

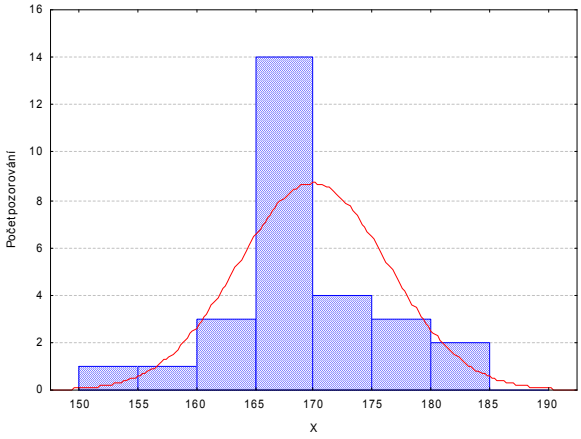
N-P plot výšky pro studentky nh



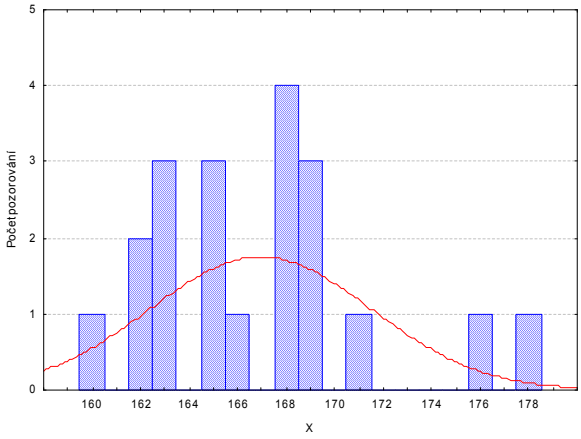
N-P plot výšky pro studentky inf



Histogram výšky pro studentky nh



Histogram výšky pro studentky inf



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

Proměnná	Testy normality (Tema7) Include condition: z=1				
	N	max D	Lilliefors p	W	p
X: vyska	28	0,167473	p < ,05	0,970969	0,606793

Pro studentky oboru inf

Proměnná	Testy normality (Tema7) Include condition: z=2				
	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.

Proměnná	Popisné statistiky (Tema7) Include condition: z=1	
	Int. spolehl. -95,000%	Int. spolehl. +95,000%
X	167,3328	172,3100

Proměnná	Popisné statistiky (Tema7) Include condition: z=2	
	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 $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ proti oboustranné alternativě

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

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.

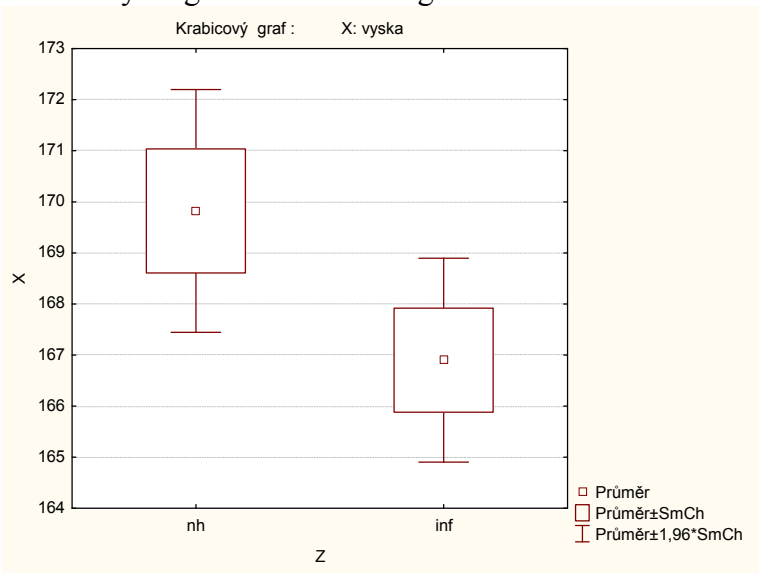
Návod:
Jedná se o dvouvýběrový t-test, kdy testujeme hypotézu $H_0 : \mu_1 - \mu_2 = 0$ proti oboustranné alternativě $H_1 : \mu_1 - \mu_2 \neq 0$

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 8.: Sestrojte 95% asymptotický interval spolehlivosti pro podíl studentek, které mají z matematiky trojku, a to
a) pro studentky oboru nh,
b) pro studentky oboru inf .

Návod:
Používáme vzorce pro dolní a horní mez intervalu spolehlivosti pro parametr ϑ alternativního rozložení (viz skripta Základní statistické metody, Důsledek 6.3.2.2.). Meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametr ϑ jsou:

$$d = m - \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2}, h = m + \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2}.$$

Výběrové průměry pro první a druhý výběr zjistíme pomocí Tabulek četností

	Tabulka četností:Y (Tema7) Include condition: z=1	
Kategorie	Četnost	Kumulativní četnost
vyborne	1	1
velmi dobre	10	11
dobre	17	28

	Tabulka četností:Y (Tema7) Include condition: z=2	
Kategorie	Četnost	Kumulativní četnost
vyborne	6	6
velmi dobre	8	14
dobre	6	20

Z tabulek plyne, že $m_1 = \frac{17}{28}$; $n_1 = 28$; $m_2 = \frac{6}{20}$; $n_2 = 20$.

K datovému souboru přidáme čtyři nové proměnné DM2, HM2, DM3, HM3
Do Dlouhého jména DM2 napíšeme =17/28-sqrt((17/28)*(1-17/28)/28)*VNormal(0,975;0;1)
Do Dlouhého jména HM2 napíšeme =17/28+sqrt((17/28)*(1-17/28)/28)*VNormal(0,975;0;1)
Do Dlouhého jména DM3 napíšeme =6/20-sqrt((6/20)*(1-6/20)/20)*VNormal(0,975;0;1)
Do Dlouhého jména HM3 napíšeme =6/20+sqrt((6/20)*(1-6/20)/20)*VNormal(0,975;0;1)
Vyjde: DM2 = 0,426246, HM2 = 0,78804, DM3 = 0,099163, HM3 = 0,500837.

Komentář: Znamená to tedy, že podíl studentek oboru nh (resp. inf), které mají trojku z matematiky, se s pravděpodobností aspoň 0,95 pohybuje od 42,6% do 78,8% (resp. od 9,9% do 50,1%).

Úkol 9.: Sestrojte 95% asymptotický interval spolehlivosti pro rozdíl podílu studentek, které mají z matematiky trojku, a to pro studentky oboru nh a inf ($0{,}038 < \vartheta_1 - \vartheta_2 < 0{,}584$).

Návod:
K datovému souboru přidáme další dvě proměnné DM4 a HM4 pro výpočet dolní a horní meze intervalu spolehlivosti. Do LongName těchto proměnných zapíšeme vzorce pro dolní a horní mez intervalu spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$ (viz skripta Základní statistické metody, Důsledek 7.2.2.2.). Výběrové průměry pro první a druhý výběr máme zjištěné již z úkolu 8.
Meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro $\vartheta_1 - \vartheta_2$ jsou:

$$d = m_1 - m_2 - \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\, u_{1-\alpha/2},$$

$$h = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\, u_{1-\alpha/2}$$

Do Dlouhého jména DM4 napíšeme:
=17/28-6/20-sqrt((17/28)*(1-17/28)/28+(6/20)*(1-6/20)/20)*VNormal(0,975;0;1)
Do Dlouhého jména HM4 napíšeme:
=17/28-6/20+sqrt((17/28)*(1-17/28)/28+(6/20)*(1-6/20)/20)*VNormal(0,975;0;1)
Vyjde: DM4 = 0,036848, HM4 = 0,577437.

Komentář: Rozdíl podílu studentek oborů nh a inf, které mají z matematiky trojku, se s pravděpodobností aspoň 0,95 pohybuje od 3,7% do 57,7%.

Úkol 10.: Na asymptotické hladině významnosti 0,05 testujte hypotézu, že podíl studentek, které mají z matematiky trojku, je stejný pro studentky oboru nh a inf.

Návod:
Na asymptotické hladině významnosti α testujeme nulovou hypotézu H_0: $\vartheta_1 - \vartheta_2 = c$ proti oboustranné alternativě H_1: $\vartheta_1 - \vartheta_2 \neq c$, kde $c = 0$.
1. způsob: lze využít výsledku 9. úkolu. 95% asymptotický interval spolehlivosti pro rozdíl parametrů $\vartheta_1 - \vartheta_2$ neobsahuje číslo 0, tedy hypotézu o shodě parametrů ϑ_1, ϑ_2 zamítáme na asymptotické hladině významnosti 0,05.

2. způsob: lze využít kritického oboru (viz skripta Základní statistické metody, Upozornění na str. 92). Protože c = 0, označme $M_* = \frac{n_1 M_1 + n_2 M_2}{n_1 + n_2}$ vážený průměr výběrových rozptylů.

Jako testová statistika slouží $T_0 = \frac{M_1 - M_2}{\sqrt{M_*(1-M_*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, která v případě platnosti nulové hypotézy má asymptoticky rozložení N(0,1). Kritický obor má tvar

$W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$

Do datového souboru přidáme další proměnné M* a T. Jak jsme zjistili v 8. úkolu,

$n_1 = 28$, $m_1 = \frac{17}{28}$, $n_2 = 20$, $m_2 = \frac{6}{20}$, tedy $n_1 m_1 + n_2 m_2 = 23$.

Do Dlouhého jména proměnné M* napíšeme
=23/48. Vyjde $m_* = 0{,}479167$.
Do Dlouhého jména proměnné T napíšeme
=(17/28-6/20)/sqrt(0,479167*(1-0,479167)*(1/28+1/20)) . Vyjde $t_0 = 2{,}100009$.

Kritický obor je $W = (-\infty;-1{,}96\rangle \cup \langle 1{,}96;\infty)$. Protože $t_0 \in W$, zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.

3. způsob: systém STATISTICA umožňuje provádět testy rozdílů mezi dvěma korelačními koeficienty, dvěma průměry či podíly. V našem případě se jedná o test rozdílu mezi dvěma podíly. Stačí znát $m_1 = \frac{17}{28} = 0{,}6071$, $n_1 = 28$, $m_2 = \frac{6}{20} = 0{,}3$, $n_2 = 20$.

Statistiky – Základní statistiky/tabulky – Testy rozdílů: r, %, průměry – Rozdíl mezi dvěma poměry – vyplníme příslušná políčka. Dostaneme p-hodnotu 0,0413, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu, že podíl studentek, které mají z matematiky trojku, je stejný pro studentky oboru nh a inf.