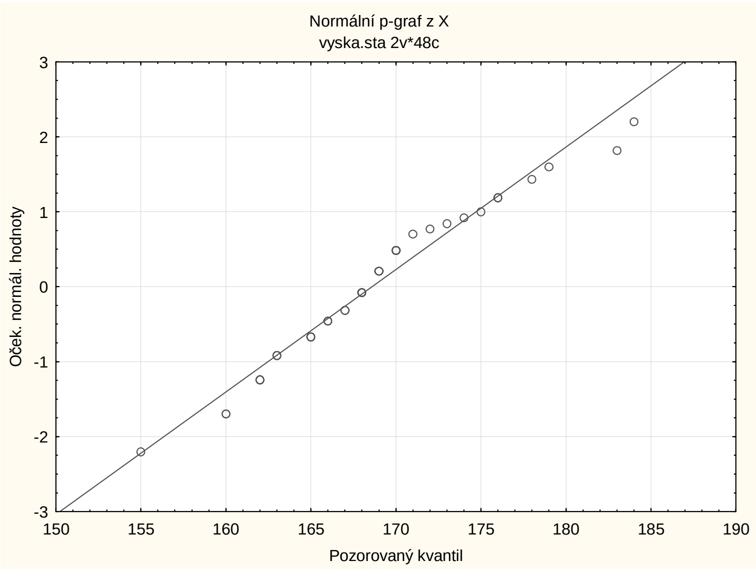
Cvičení 7.: Ověřování normality, úlohy o náhodném výběru z normálního a alternativního rozložení

Úkol 1. : U 45 studentek VŠE v Praze byla zjišťována výška a obor studia (1 – národní hospodářství, 2 – informatika). Hodnoty jsou uloženy v souboru vyska.sta. Pomocí S-W testu testujte na hladině významnosti 0,05 hypotézu, že data pocházejí z normálního rozložení. Pomocí N-P plotu posuďte vizuálně předpoklad normality.
Návod:
Statistiky – Základní statistiky/tabulky – Tabulky četností – OK – Proměnné X – OK – Normalita – zaškrtneme S-W test – Testy normality.

	Testy normality (vyska.sta)		
Proměnná	N	W	p
X: vyska	48	0,965996	0,176031

Výstupní tabulka obsahuje počet pozorování, testovou statistiku S-W testu (W = 0,965996) a odpovídající p-hodnotu (p = 0,176031). Vidíme, že S-W test nezamítá hypotézu o normalitě na hladině významnosti 0,05.

Statistiky – Grafy – 2D grafy – Normální pravděpodobnostní grafy – Proměnné X – OK – odškrtneme Neurčovat průměrnou pozici svázaných pozorování - OK



Tečky se od ideální přímky odchylují jen nepatrně, zřejmě jde o data z normálního rozložení

Samostatný úkol: Testy normality a grafické ověření normality proveďte jak pro výšky studentek oboru národní hospodářství, tak pro výšky studentek oboru informatiky. (Upozornění: Úkol lze provést pomocí filtru nebo pomocí volby Analýza skupin, kde roli skupinové proměnné hraje Z.)

Pro kontrolu:
Výsledky pro obor národní hospodářství:

Proměnná	Testy normality (vyska.sta) Zhrnout podmínku: z=1		
	N	W	p
X: vyska	28	0,970969	0,606793

S-W test hypotézu o normalitě nezamítá na hladině významnosti 0,05 (p-hodnota je větší než 0,05).

Výsledky pro obor informatika:

Proměnná	Testy normality (vyska.sta) Zhrnout podmínku: z=2		
	N	W	p
X: vyska	20	0,922747	0,111924

S-W test hypotézu o normalitě nezamítá na hladině významnosti 0,05.

Úkol 2.: Intervaly spolehlivosti pro parametry μ, σ^2 normálního rozložení

Z populace stejně starých selat téhož plemene bylo vylosováno šest selat a po dobu půl roku jim byla podávána táž výkrmná dieta. Byly zaznamenávány průměrné denní přírůstky hmotnosti v Dg. Z dřívějších pokusů je známo, že v populaci mívají takové přírůstky normální rozložení, avšak střední hodnota i rozptyl se měnívají. Přírůstky v Dg: 62, 54, 55, 60, 53, 58.

a) Najděte 95% empirický interval spolehlivosti pro neznámou střední hodnotu μ při neznámé směrodatné odchylce σ.
b) Najděte 95% empirický interval spolehlivosti pro směrodatnou odchylku σ.

Návod:

Vytvoříme nový datový soubor o jedné proměnné X a 6 případech. Do proměnné X napíšeme dané hodnoty.

Ad a) Statistiky – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme Meze spolehl. prům. (ostatní volby zrušíme) – ponecháme implicitní hodnotu 95,00 – Výpočet.

Proměnná	Popisné statistiky (Tabulka25)	
	Int. spolehl. -95,000%	Int. spolehl. 95,000%
X	53,24542	60,75458

Vidíme, že 53,25 Dg < μ < 60,75 Dg s pravděpodobností aspoň 0,95.

Ad b) Statistiky – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme Meze sp. směr. odch., ponecháme implicitní hodnotu 95,00 – Výpočet.

Proměnná	Popisné statistiky (Tabulka1)	
	Spolehlivost Sm.Odch. -95,000%	Spolehlivost Sm.Odch. +95,000%
X	2,233234	8,774739

Dostáváme výsledek: 2,23 Dg < σ < 8,77 Dg s pravděpodobností aspoň 0,95.

Úkol 3.: Testování hypotézy o parametru μ normálního rozložení
Systematická chyba měřicího přístroje se eliminuje nastavením přístroje a měřením etalonu, jehož správná hodnota je $\mu = 10{,}00$. Nezávislými měřeními za stejných podmínek byly získány hodnoty: 10,24 10,12 9,91 10,19 9,78 10,14 9,86 10,17 10,05, které považujeme za realizace náhodného výběru rozsahu 9 z rozložení $N(\mu, \sigma^2)$. Je možné při riziku 0,05 vysvětlit odchylky od hodnoty 10,00 působením náhodných vlivů?
Návod:
Na hladině významnosti 0,05 testujeme hypotézu H_0: $\mu = 10$ proti oboustranné alternativě H_1: $\mu \neq 10$. Jde o úlohu na jednovýběrový t-test. Ten je ve STATISTICE implementován. Otevřeme datový soubor mereni_etalonu.sta. V Základních statistikách/tabulkách vybereme t-test, samostatný vzorek. Do Referenčních hodnot zapíšeme 10. Ve výstupu se podíváme na hodnotu testového kritéria a na p-hodnotu. Pokud p-hodnota bude menší nebo rovna 0,05, zamítneme hypotézu H_0: $\mu = 10$ ve prospěch oboustranné alternativní hypotézy H_1: $\mu \neq 10$ na hladině významnosti 0,05. V opačném případě H_0 nezamítáme. V našem případě je

Proměnná	Test průměrů vůči referenční konstantě (hodnotě)							
	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Prom1	10,05111	0,162669	9	0,054223	10,00000	0,942611	8	0,373470

Protože p-hodnota 0,373470 > 0,05 nulovou hypotézu nezamítáme na hladině významnosti 0,05. Odchylky od hodnoty 10 lze vysvětlit působením náhodných vlivů.
Všimněme si ještě hodnoty testového kriteria: $t_0 = 0{,}942611$. Kritický obor

$$W = \left(-\infty, -t_{1-\alpha/2}(n-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-1), \infty\right) = \left(-\infty, -t_{0,975}(8)\right\rangle \cup \left\langle t_{0,975}(8), \infty\right) =$$
$$= \left(-\infty, -2{,}306\right\rangle \cup \left\langle 2{,}306, \infty\right)$$

Protože $t_0 \notin W$, nezamítáme na hladině významnosti 0,05 hypotézu H_0 .

Úkol 4.: Interval spolehlivosti pro rozdíl parametrů $\mu_1 - \mu_2$ dvourozměrného rozložení
Bylo vylosováno 6 vrhů selat a z nich vždy dva sourozenci. Jeden z nich vždy dostal náhodně dietu č. 1 a druhý dietu č. 2. Přírůstky v Dg jsou následující: (62,52), (54,56), (55,49), (60,50), (53,51), (58,50). Za předpokladu, že rozdíly uvedených dvojic tvoří náhodný výběr z normálního rozložení se střední hodnotou $\mu_1 - \mu_2$, sestrojte 95% interval spolehlivosti pro rozdíl středních hodnot.
Návod:
Vytvoříme datový soubor o třech proměnných a šesti případech. Do proměnných v1 a v2 zapíšeme naměřené přírůstky, do proměnné v3 uložíme rozdíly v1 - v2.
Ve STATISTICE je implementován výpočet oboustranného intervalu spolehlivosti pro μ, když σ^2 neznáme. Pomocí Popisných statistik zjistíme meze 95% intervalu spolehlivosti pro střední hodnotu proměnné v3 tak, že zaškrtneme Meze spoleh. prům.

Proměnná	Popisné statistiky	
	Int. spolehl. -95,000%	Int. spolehl. +95,000%
Prom3	0,626461	10,70687

Dostaneme výsledek: 0,63 Dg < μ < 10,71 Dg s pravděpodobností aspoň 0,95.

Úkol 5.: Testování hypotézy o rozdílu parametrů $\mu_1 - \mu_2$ dvourozměrného rozložení
Pro data z úkolu 4 testujte na hladině významnosti 0,05 hypotézu, že obě výkrmné diety mají stejný vliv.

Návod:

Označme $\mu = \mu_1 - \mu_2$. Na hladině významnosti 0,05 testujeme hypotézu H_0: $\mu = 0$ proti oboustranné alternativě H_1: $\mu \neq 0$. Jde o úlohu na párový t-test. Ten je ve STATISTICE implementován.Vytvoříme datový soubor o dvou proměnných a šesti případech. Do proměnných v1 a v2 zapíšeme naměřené přírůstky. V menu Základní statistiky/tabulky vybereme t-test, závislé vzorky. Zadáme názvy obou proměnných a ve výstupu se podíváme na hodnotu testového kritéria a na p-hodnotu.

	t-test pro závislé vzorky Označ. rozdíly jsou významné na hlad. p < ,05000							
Proměnná	Průměr	Sm.odch.	N	Rozdíl	Sm.odch. rozdílu	t	sv	p
Prom1	57,00000	3,577709						
Prom2	51,33333	2,503331	6	5,666667	4,802777	2,890087	5	0,034183

Protože p-hodnota 0,034183 < 0,05, zamítáme hypotézu H_0: $\mu = 0$ ve prospěch alternativní hypotézy H_1: $\mu \neq 0$ na hladině významnosti 0,05. Znamená to, že jsme s rizikem omylu nejvýše 5% prokázali rozdíl v účinnosti obou výkrmných diet.

Všimněme si ještě hodnoty testového kriteria: $t_0 = 2,890087$. Kritický obor

$$W = \left(-\infty, -t_{1-\alpha/2}(n-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-1), \infty\right) = \left(-\infty, -t_{0,975}(5)\right\rangle \cup \left\langle t_{0,975}(5), \infty\right) =$$
$$= \left(-\infty, -2,5706\right\rangle \cup \left\langle 2,5706, \infty\right)$$

Protože $t_0 \in W$, zamítáme na hladině významnosti 0,05 hypotézu H_0.

Úkol 6.: Asymptotický interval spolehlivosti pro parametr ϑ alternativního rozložení
Může politická strana, pro niž se v předvolebním průzkumu vyslovilo 60 z 1000 dotázaných osob, očekávat se spolehlivostí aspoň 0,95, že by v této době ve volbách překročila 5% hranici pro vstup do parlamentu?

Návod:
Zavedeme náhodné veličiny $X_1, ..., X_{1000}$, přičemž $X_i = 1$, když i-tá osoba se vysloví pro danou politickou stranu a $X_i = 0$ jinak, i = 1, ..., 1000. Tyto náhodné veličiny tvoří náhodný výběr z rozložení A(ϑ). V tomto případě n = 1000, m = 60/1000 = 0,06, $\alpha = 0,05$, $u_{1-\alpha} = u_{0,95}$ = 1,645.
Ověření podmínky $n\vartheta(1-\vartheta) > 9$: parametr ϑ neznáme, musíme ho nahradit výběrovým průměrem. Pak 1000.0,06.0,94 = 56,4 > 9.
95% levostranný interval spolehlivosti pro ϑ je

$$\left(m - \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha}\ ;\ \infty\right) = \left(0,06 - \sqrt{\frac{0,06(1-0,06)}{1000}}u_{0,95}\ ;\ \infty\right) = (0,0476, \infty)$$

Postup ve STATISTICE:
Statistiky – Analýza síly testu – Odhad intervalu – Jeden podíl, Z, Chí-kvadrát test – OK – Pozorovaný podíl p: 0,06, Velik. vzorku (N): 1000, Spolehlivost: 0,9 – Vypočítat. Dostaneme 0,0476.
S pravděpodobností přibližně 0,95 tedy $\vartheta > 0,047647$. Protože tento interval zahrnuje i hodnoty nižší než 0,05, nelze vyloučit, že strana získá méně než 5 % hlasů.
Upozornění: Spolehlivost volíme 0,9, protože dolní mez 90% oboustranného intervalu spolehlivosti je stejná jako dolní mez levostranného 95% intervalu spolehlivosti.

Úkol 7: Testování hypotézy o parametru ϑ 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 $X_1, ..., X_{150}$ z rozložení A(0,3). Testujeme H_0: $\vartheta = 0,3$ proti levostranné alternativě H_1: $\vartheta < 0,3$. V tomto případě je testovým kritériem statistika

$$T_0 = \frac{M - c}{\sqrt{\frac{c(1-c)}{n}}}$$

, která v případě platnosti nulové hypotézy má asymptoticky rozložení N(0,1)
Musíme ověřit splnění podmínky $n\vartheta(1-\vartheta) > 9$: 150.0,3.0,7 = 31,5 > 9.

Vypočteme realizaci testové statistiky: $t_0 = \frac{m-c}{\sqrt{\frac{c(1-c)}{n}}} = \frac{\frac{38}{150} - 0,3}{\sqrt{\frac{0,3(1-0,3)}{150}}} = -1,24722$.

Kritický obor: $W = \left(-\infty, -u_{1-\alpha}\right\rangle = \left(-\infty, -1,645\right\rangle$.
Protože testová statistika nepatří do kritického oboru, H_0 nezamítáme na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 5 % tedy naše data neprokázala pokles zájmu zákazníků cestovní kanceláře o zemi X.

Postup ve STATISTICE:
Použijeme aplikaci Testy rozdílů:
Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma poměry – do políčka P 1 napíšeme 0,2533 (tj. 35/150), do políčka N1 napíšeme 150, do políčka P 2 napíšeme 0,3, do políčka N2 napíšeme 32767 (větší hodnotu systém neumožní) – zaškrtneme Jednostr. - Výpočet. Dostaneme p-hodnotu 0,1065, tedy nezamítáme nulovou hypotézu na hladině významnosti 0,05.