Vysoká škola báňská - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Průvodce k programu Statgraphics

Část 2

Lenka Šimonová

Ostrava, 2006

Průvodce k programu Statgraphics vznikl pro potřeby výuky předmětu Statistika I. na FEI VŠB-TU Ostrava, jak v prezenční tak v kombinované formě jako doplněk základní studijní opory, kterou je skriptum *Briš R., Litschmannová M.: Statistika I. pro kombinované a distanční studium, Ostrava 2004*.

Průvodce k programu Statgraphics ilustruje na příkladech řešených programem *Statgraphics* použití standardních statistických metod probíraných v předmětu Statistika I.. Podrobnější zdůvodnění použití odpovídajících statistických metod a vysvětlení jejich teoretického základu najde čtenář v již zmíněných skriptech *Briš R., Litschmannová M.: Statistika I. pro kombinované a distanční studium, Ostrava 2004* resp. jiné statistické literatuře – viz uvedený seznam literatury v závěru textu. *Průvodce k programu Statgraphics* není úplným manuálem k programu *Statgraphics*. Program *Statgraphics* obsahuje řadu dalších procedur, např. časové řady, které již nejsou v náplni předmětu Statistika I, tudíž nejsou ani zařazeny do tohoto textu.

Zdroje dat: v 1. a 2. kapitole jsou použita fiktivní data, v 3. kapitole jsou vyhodnocena data z balíku DataFile programu *Statgraphics*, 4. a 5. kapitola: Litschmannová M.: Statistika I. - Příklady, Ostrava 2000, 6. část vygenerovaná náhodná čísla programem *Statgraphics*, dále modifikovaná data z použité literatury, 7. kapitola ANOVA, příklady 1. a 2. : Friedrich V. : Statistika 1., Vysokoškolská učebnice pro distanční studium, Západočeská Univerzita, Plzeň 2002, 8. kapitola Regrese, příklad 1.: Novovičová J. : Pravděpodobnost a základy matematické statistiky, ČVUT Praha, 2002. Ostatní zdroje dat pro zpracování úloh ve Statgraphicsu byly internetové stránky statistického úřadu.

Průvodce k programu Statgraphics část 1 obsahuje explorační analýzu dat a metody statistické dedukce, tj. hledání hodnot pravděpodobnostních, distribučních funkcí a kvantilů u daných typů rozdělení. *Průvodce k programu Statgraphics část 2* obsahuje metody statistické indukce, konkrétně testovaní parametrických a neparametrických hypotéz, konstrukce intervalových odhadů, jednofaktorovou analýzu rozptylu ANOVA a jednoduchou lineární regresi.

Autorka přeje studentům příjemné, ničím nerušené, studium předmětu Statistika I.

V Ostravě, 7.6.2006 Mgr. Lenka Šimonová

5. Testování hypotéz a intervalové odhady

V celé této kapitole budeme předpokládat, pokud nebude řečeno jinak, že zadaná data lze považovat za náhodný výběr z normálního rozdělení, dále u testování rovnosti středních hodnot, že mezi rozptyly obou základních souborů není statisticky významný rozdíl. Pokud by tyto předpoklady nebyly splněny, museli bychom přistoupit k neparametrickým testům – viz kapitola 6.

5.1. Jednovýběrové testy

Jednovýběrové testy používáme v případě, kdy chceme pomocí náhodného výběru otestovat, zda je parametr populace roven nějaké (obvykle standardizované – udávaná norma hmotnosti, celorepublikový platový průměr, pokrytí u mobilního operátora, …) hodnotě.

Testování střední hodnoty při známém rozptylu σ^2

Příklad 5.1. *Odběratel s dodavatelem uzavřeli smlouvu o dodávce pytlů obilí. Při známém rozptylu* $\sigma^2 = 0{,}1$ *plnícího stroje má být střední hodnota hmotnosti pytlů 10 kg. Pro ověření skutečnosti, že plnící stroj pracuje dobře, bylo náhodně vybráno 40 pytlů a získán průměr jejich hmotnosti* $\bar{x} = 9{,}8$ *kg. Rozhodněte, zda dodavatel dodržuje stanovenou střední hodnotu hmotnosti.*

Řešení: Testujeme hypotézu o střední hodnotě

$H_0 : \mu = 10 \text{ kg}$
$H_A : \mu < 10 \text{ kg}$

přičemž známe rozptyl $\sigma^2 = 0{,}1$ základního souboru. Jelikož známe rozptyl základního souboru, volili bychom při „ručním výpočtu“ testovací statistiku Z:

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1).$$

Zkusme si nejprve výpočet provést „ručně“ a poté pomocí *Statgraphicsu*.

„Ruční výpočet“ tedy pomocí kalkulačky a tabulek distribučních funkcí, resp. kvantilů:

$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$, kde $n = 40$, $\bar{x} = 9{,}8$ kg jsou údaje týkající se výběrového souboru.

Vypočteme tzv. **p-hodnotu** neboli **p-value**, která je definována následujícím způsobem podle tvaru alternativní hypotézy.

1. $H_A : \theta < \theta_0 \Rightarrow \; p - value = F(x_{obs})$

2. $H_A : \theta > \theta_0 \Rightarrow \; p - value = 1 - F(x_{obs})$

3. $H_A : \theta \neq \theta_0 \Rightarrow \; p - value = 2 \cdot \min\{F(x_{obs}), 1 - F(x_{obs})\}$

Tedy v našem případě

$$x_{obs} = Z\big/_{H_0} = \frac{9{,}8 - 10}{\sqrt{0{,}1}} . \sqrt{40} = -4$$

$$p - value = \Theta(x_{obs}) = \Theta(\frac{9{,}8 - 10}{\sqrt{0{,}1}} \cdot \sqrt{40}) = \Theta(-4) = 0 < 0{,}05 \Rightarrow \text{ Hypotéza } H_0 \text{ se zamítá..}$$

Hodnotu distribuční funkce $\Theta(-4) = 0$ jsme vyhledali v tabulkách distribuční funkce normovaného normálního rozdělení.

Závěr: Hypotéza H_0 se zamítá. Data nám dávají dostatek argumentů k tomu, abychom konstatovali, že střední hodnota hmotnosti dodávaných pytlů nedosahuje 10-ti kg, tj. že dodavatel nedodržuje stanovenou střední hodnotu hmotnosti.

Řešení pomocí *Statgraphicsu*:

Statgraphics nemá zaved způsob testování hypotéz pomocí Z – testu, výpočet tedy lze provést pouze přibližně s pomocí T- Testu, kdy ztotožníme směrodatnou odchylku základního souboru s výběrovou směrodatnou odchylkou. Pokud je rozsah výběru velký (cca n>30), lze tyto dva testy ztotožnit (t-rozdělení, které se používá u t-testu, pro n>30 téměř splývá s normálním rozdělením, které se používá u z-testu) bez velkého rizika chybovosti. Pokud je n<30, je na to třeba dávat pozor, neboť bychom mohli dostat chybné závěry.

Tedy testujeme hypotézu

$H_0 : \mu = 10 \text{ kg}$
$H_A : \mu < 10 \text{ kg}$

pomocí jednovýběrového t-testu:

$$T = \frac{\overline{X} - \mu}{s} . \sqrt{n} \sim t(n-1).$$

Ve *Statgraphicsu* použijeme proceduru:

Menu Desribe/HypothesisTests

Doplníme: 1) údaje týkající se nulové hypotézy:

Null Hypothesis (nulová hypotéza) … 10 (kg)

2) údaje týkající se výběrového souboru:

Sample Mean (průměr) ... 9,8 (kg)
Sample Sigma (směrodatná odchylka) … 0,32 (kg)
Sample Size (velikost výběru) … 40

Dostaneme následující výstup ve *Statgraphicsu*:

Hypothesis Tests

Sample mean = 9,8
Sample standard deviation = 0,32
Sample size = 40

95,0% confidence interval for mean: 9,8 +/- 0,102341 [9,69766;9,90234]

Null Hypothesis: mean = 10,0
Alternative: not equal
Computed t statistic = -3,95285
P-Value = 0,000315452
Reject the null hypothesis for alpha = 0,05.

Jelikož máme zformulovánu jednostrannou alternativu, je třeba tento výstup upravit. Kliknutím pravým tlačítkem myši zobrazíme ve vzniklém výstupu proceduru

Analysis Options. Zvolíme **Less Than** (menší než):

Hypothesis Tests

Sample mean = 9,8
Sample standard deviation = 0,32
Sample size = 40

95,0% upper confidence bound for mean: 9,8 + 0,0852489 [9,88525]

Null Hypothesis: mean = 10,0
Alternative: less than
Computed t statistic = -3,95285
P-Value = 0,000157726
Reject the null hypothesis for alpha = 0,05.

Hodnota testovacího kritéria t-testu (***Computed t statistic***) je rovna:

$$x_{obs} = T \big/ H_0 = \frac{9{,}8 - 10}{0{,}32} . \sqrt{40} = -3{,}95,$$

hodnota $p - value = F(x_{obs}) = F(-3{,}95) \approx 0{,}0002 << 0{,}05.$

Závěr: Zamítáme (***reject***) hypotézu H_0 na 5 % hladině významnosti ve prospěch alternativy H_A. Na základě uvedených údajů lze konstatovat, že existuje odůvodněné podezření z nedodržování stanovené váhy. Potravinářská inspekce může na základě tohoto testu stíhat Mlýny Hrušov pro nedodržení udávané váhy výrobku.

Testování střední hodnoty při neznámém rozptylu

Příklad 5.2. *Mlýny Hrušov udávají na balících mouky váhu 1kg. Potravinářská inspekce otestovala 20 balíků mouky a zjistila váhu (0,95 ± 0,08)kg. Může potravinářská inspekce na základě tohoto testu stíhat Mlýny Hrušov pro nedodržení udávané váhy výrobku? Ověřte čistým testem významnosti.*

Řešení: V zadání příkladu máme uvedeny následující údaje týkající se výběrového souboru: výběrový průměr $\bar{x} = 0{,}95$ kg, výběrovou směrodatnou odchylku s=0,08 kg a počet vzorků ve výběru n=20.

Váha jednoho balíčku mouky má být podle normy 1 kg. Budeme tedy testovat, zda střední (průměrná) hodnota váhy libovolného balíčku mouky vyrobeného firmou Mlýny Hrušov je rovna jednomu kilogramu, tj. zda $\mu_0 = 1\text{kg}$.

Testujeme tedy hypotézu $H_0 : \mu = 1\,\text{kg}$,
oproti alternativě $H_A : \mu \neq 1$ kg.

Jelikož neznáme směrodatnou odchylku základního souboru použijeme k otestování dané hypotézy jednovýběrový t - test:

$$T = \frac{\overline{X} - \mu}{s}.\sqrt{n}$$

Menu Desribe/HypothesisTests

Doplníme: 1) údaje týkající se nulové hypotézy:

Null Hypothesis (nulová hypotéza) … 1 (kg)

2) údaje týkající se výběrového souboru:

Sample Mean (průměr) ... 0,95 (kg)
Sample Sigma (směrodatná odchylka) … 0,08 (kg)
Sample Size (velikost výběru) … 20

Dostaneme následující výstup ve *Statgraphicsu*:

Null Hypothesis: mean = 1,0
Alternative: not equal
Computed t statistic = -2,79508
P-Value = 0,0115468

Reject the null hypothesis for alpha = 0,05.
Hodnota testovacího kritéria t-testu (***Computed t statistic***) je rovna:

$$x_{obs} = T \big/ H_0 = \frac{0{,}95 - 0}{0{,}08} . \sqrt{20} = -2{,}795.$$

Hodnota $p - value = F(x_{obs}) = F(-2{,}795) = 0{,}012 < 0{,}05.$

Závěr: Zamítáme (***reject***) hypotézu H_0 ve prospěch alternativy H_A.

Na základě uvedených údajů lze konstatovat, že existuje odůvodněné podezření z nedodržování stanovené váhy. Potravinářská inspekce může na základě tohoto testu stíhat Mlýny Hrušov pro nedodržení udávané váhy výrobku.

Příklad 5.3. *Balíčky soli mají mít hmotnost 1 kg. Bylo zváženo 10 balíčků a zjištěny odchylky od váhy 1 kg:*

-1,2 0,5 -0,6 -0,3 0,2 -1,0 0,4 -0,8 0,5 -0,4 g .

a) Najděte 95% interval spolehlivosti pro střední hodnotu odchylky hmotnosti balíčku soli od 1 kg.
b) Zjistěte, zda lze na základě zjištěných hodnot konstatovat, že průměrná hmotnost jednoho balíčku nedosahuje 1 kg.

Řešení: V tomto příkladě je nutné nejprve určit výběrový průměr a směrodatnou odchylku:

1. Zadáme data.

2. Provedeme explorační analýzu dat:

Menu Describe/Numeric Data/OneVariableAnalysis

V levém dolním okně se zobrazí:

Count=10 … počet pozorování
Average= - 0,27 … výběrový průměr
Standard deviation=0,637791 … výběrová směrodatná odchylka

3. Samotné řešení úlohy:

Balíčky soli mají mít hmotnost 1 kg. Uvedené údaje se týkají odchylky od 1 kg Budeme tedy testovat, zda je střední hodnota odchylky váhy od 1 kg rovna 0, tj. zda $\mu_0 = 0$ g. Jelikož v úloze řešíme intervalový odhad i otestování dané hypotézy vyřešíme obě úlohy najednou:

Testujeme hypotézu $H_0 : \mu = 0$ g,
Oproti alternativě $H_A : \mu \neq 0$ g.

Opět použijeme k otestování dané hypotézy jednovýběrový t - test:

$$T = \frac{\overline{X} - \mu}{s}\sqrt{n} \sim t(n-1).$$

Menu Desribe/HypothesisTests

Doplníme:

1) údaje týkající se nulové hypotézy:

Null Hypothesis (nulová hypotéza)… 0 (g)

2) údaje týkající se výběrového souboru:

Sample Mean (průměr)…. - 0,27 (g)
Sample Sigma (směrodatná odchylka).. . . 0,64 (g)
Sample Size (velikost výběru).… 10

Intervalový odhad pro střední odchylku váhy jednoho balíčku soli od 1 kg:

95,0% confidence interval for mean: - 0,27 +/- 0,45783 [- 0,72783 ; 0,18783]

$$P\left(t_{0,025} \langle \frac{\overline{X} - \mu}{s}.\sqrt{n} \langle t_{0,975}\right) = 0{,}95 \ , \quad P\left(-2{,}26 \langle \frac{-0{,}27 - \mu}{0{,}64}.\sqrt{10} \langle 2{,}26\right) = 0{,}95$$

$$P(-0{,}72 \langle \mu \langle 0{,}19) = 0{,}95$$

Tedy 95 % interval spolehlivosti pro střední váhu jednoho balíčku soli μ je interval
- 0,27 g ± 0,46 g = (-0,72 ; 0,19) gramů.

Pro otestování hypotézy, zda je střední hodnota odchylky váhy od 1 kg rovna 0 použijeme testovací kritérium:

$$T = \frac{\overline{X} - \mu}{s}\sqrt{n} \sim t(n-1).$$

Hodnota testovacího kritéria je rovna (***Computed t statistic***):

$$x_{obs} = T \big/ H_0 = \frac{-0{,}27 - 0}{0{,}64}.\sqrt{10} = -1{,}33$$

P-Value = 0,214947

$$p - value = F(x_{obs}) = F(-1{,}33) = 0{,}1 > 0{,}05.$$

Do not reject the null hypothesis for alpha = 0,05.

Závěr: Nezamítáme (**do not reject**) hypotézu H_0. Na základě uvedených údajů nelze konstatovat, že průměrná hmotnost jednoho balíčku nedosahuje 1 kg.

Testování směrodatné odchylky nebo rozptylu

Příklad 5.4. *Automat vyrábí pístové kroužky o daném průměru. Výrobce udává, že směrodatná odchylka průměru kroužku je 0,05mm. K ověření této informace bylo náhodně vybráno 80 kroužků a vypočtena směrodatná odchylka jejich průměru 0,04mm. Lze tento rozdíl považovat za významný ve smyslu zlepšení kvality produkce?*

Řešení: Výrobce udává, že směrodatná odchylka průměru kroužku je 0,05mm. Budeme tedy testovat, zda směrodatná odchylka průměru libovolného kroužku je 0,05mm .

Testujeme tedy hypotézu $H_0 : \sigma = 0{,}05$ mm,
oproti alternativě $H_A : \sigma < 0{,}05$ mm.

K ověření platnosti této hypotézy bylo náhodně vybráno 80 kroužků a vypočtena směrodatná odchylka průměru jednotlivého kroužku 0,04mm.

K otestování hypotézy o rozptylu resp. směrodatné odchylky použijeme χ^2 - test (čti chí-kvadrát test):

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1).$$

K řešení pomocí *Statgraphicsu* použijeme opět proceduru **Menu Desribe/HypothesisTests,**

ovšem tentokrát zvolíme **Parameter: Normal Sigma**

Doplníme

1) údaje týkající se nulové hypotézy:

Null Hypothesis (nulová hypotéza) … 0,05 (kg)

2) údaje týkající se výběrového souboru:

Sample Sigma (směrodatná odchylka výběru) … 0,04 (kg)
Sample Size (velikost výběru) … 80

Hodnota testovacího kritéria χ^2 - testu: ***Computed chi-squared statistic = 50,56***

a příslušná p- hodnota: ***P-Value = 0,010588*** < 0,05.

Závěr*:* Zamítáme (***reject***) nulovou hypotézu na 5 % hladině významnosti ve prospěch alternativy.

Na základě zjištěných údajů lze konstatovat, že se jedná o statisticky významné zlepšení kvality produkce.

Příklad 5.5. *Výběrovým šetřením bychom chtěli odhadnout průměrnou mzdu pracovníků určitého výrobního odvětví. Z vyčerpávajícího šetření, které probíhalo před několika měsíci, víme, že směrodatná odchylka mezd byla 750,-Kč. Odhad chceme provést s 95% spolehlivosti a jsme ochotni připustit maximální chybu ve výši 50,-Kč. Jak velký musíme provést výběr, abychom zajistili požadovanou přesnost a spolehlivost?*

Řešení: Ve všech předchozích příkladech, kde jsme znali rozsah výběru a další výběrové charakteristiky a určovali jsme intervaly spolehlivosti, resp. jsme testovali zadané hypotézy. Nyní máme de facto „opačný úkol". Máme zadanou přípustnou maximální chybu intervalového odhadu a máme určit, jak velký má být rozsah výběrového souboru, abychom dodrželi zvolenou přesnost. Ukážeme si zde pouze „zrychlené" řešení pomocí *Statgraphicsu*. Zkuste si rozmyslet, jak by se tato úloha řešila „ručně".

Menu Describe/Sample-SizeDetermination

Chceme odhadnout průměrnou mzdu pracovníků, tedy budeme testovat střední hodnotu a proto volíme: **Normal Mean**

Dále zvolíme

Hypothesized Sigma … 750
Absolute Error … 50

Dostaneme následující výstup:

Sample-Size Determination

Parameter to be estimated: normal mean
Desired tolerance: +- 50,0
Confidence level: 95,0%
Assumed sigma: 750,0

The required sample size is n=865 observations.

Závěr*: Abychom zajistili požadovanou přesnost a spolehlivost při určení průměrné mzdy pracovníků daného výrobního odvětví, bychom měli provést náhodný výběr o rozsahu 865 pracovníků.*

Testování relativní četnosti

Příklad 5.6. *V náhodném výběru čipů vyráběných velkou světovou společnosti 10 % čipů nevyhovuje novým požadavkům na kvalitu. Sestrojte 95 % interval spolehlivosti pro podíl nevyhovujících čipů v celém základním souboru, jestliže rozsah náhodného výběru je:*

a) n = 100,

b) n = 1000.

Řešení: V zadání příkladu máme uveden následující údaj týkající se výběrového souboru:

výběrový podíl: $p_0 = 0{,}1$ (10 % nevyhovujících čipů ve výběru).

Jelikož zkoumáme relativní četnost prvků v základním souboru mající danou vlastnost zvolíme testovací kritérium *P*:

$$P = \frac{\pi - p_0}{\sqrt{p_0(1-p_0)}}.\sqrt{n} \sim \mathrm{N}(0,1)$$

Ve *Statgraphicsu* použijeme opět proceduru **Menu Desribe/HypothesisTests**

a tentokrát zvolíme **Parameter: Binomial Proportion.**

Null Hypothesis … např. 0,2 (volba hodnoty pro nulovou hypotézu není při určování intervalů spolehlivosti podstatná, můžeme klidně nechat nastavenou hodnotu 0,5)

Sample Proportion … 0,1

Ad a) **Sample Size** … 100

Dostaneme následující výstup:

Hypothesis Tests

Sample proportion = 0,1
Sample size = 100
Approximate 95,0% confidence interval for p: [0,0490047;0,176223]

$$P\left(x_{0,025} < \frac{\pi - 0{,}1}{\sqrt{0{,}1.0{,}9}}.\sqrt{100} < x_{0,975}\right) = 0{,}95 \quad , \quad P\left(-1{,}96 < \frac{\pi - 0{,}1}{\sqrt{0{,}09}}.\sqrt{100} < 1{,}96\right) = 0{,}95$$

Závěr: 95,0 % intervalem spolehlivosti pro podíl nevyhovujících čipů v celém základním souboru je interval (4,9 % ; 17,6 %) - za předpokladu, že jsme při náhodném testování 100 čipů zjistili 10 % nevyhovujících čipů.

Ad b) **Sample Size** … 1 000

Dostaneme následující výstup:

Hypothesis Tests

Sample proportion = 0,1
Sample size = 1000
Approximate 95,0% confidence interval for p: [0,0828229;0,119053]

$$P\left(x_{0,025} \langle \frac{\pi - 0{,}1}{\sqrt{0{,}1.0{,}9}}.\sqrt{1000} \langle x_{0,975} \right) = 0{,}95 \quad , \quad P\left(-1{,}96 \langle \frac{\pi - 0{,}1}{\sqrt{0{,}09}}.\sqrt{1000} \langle 1{,}96 \right) = 0{,}95$$

Závěr: 95,0 % intervalem spolehlivosti pro podíl nevyhovujících čipů v celém základním souboru je interval (8,3 % ; 11,9 %) - za předpokladu, že jsme při náhodném testování 1 000 čipů zjistili 10 % nevyhovujících čipů.

Neboli pokud jsme náhodným výběrem zjistili, že mezi 1 000 výrobků je 10 % vadných, pak s 95% jistotou můžeme říct, že procento vadných v celém základním souboru nebude menší než 8 % a větší než 12 %.

Příklad 5.7. *V předvolební kampani si politická strana XYZ chce nechat ověřit své preference a nechá si udělat předvolební průzkum. Na anketu odpoví 200 potenciálních voličů a z nich 106 preferuje stranu XYZ.*
a) Zaručuje tento výsledek straně XYZ nadpoloviční většinu u skutečných voleb? (rozhodněte na základě pravostranného 95 % intervalu spolehlivosti)
b) Kolik bychom museli oslovit respondentů, aby chyba odhadu činila maximálně 2 % ?

Řešení: Ad a) Hledáme jednostranný interval spolehlivosti pro podíl, přičemž procentuální podíl ve vzorku je

$$p_0 = \frac{106}{200} = 0{,}53 = 53\%$$

Jelikož zkoumáme relativní četnost prvků v základním souboru mající danou vlastnost zvolíme testovací kritérium *P*:

$$P = \frac{\pi - p_0}{\sqrt{p_0(1-p_0)}}.\sqrt{n} \sim \mathrm{N}(0,1)$$

Ve *Statgraphicsu* použijeme opět proceduru **Menu Desribe/HypothesisTests**

Zvolíme **Parameter: Binomial Proportion**

Null Hypothesis … 0,5 (předpokládáme přesně 50 %)

Sample Proportion … 0,53 (zadaný údaj týkající se vzorku)

Sample Size … 200

Hledáme pravostranný interval spolehlivosti pro podíl, tedy je ještě třeba změnit nastavení z oboustranného intervalu spolehlivosti na pravostranný interval:
Pravým tlačítkem myši ve vzniklém výstupu:

Analysis Options … Less Than (menší než)

Dostaneme následující výstup:

Hypothesis Tests

Sample proportion = 0,53
Sample size = 200

Approximate 95,0% upper confidence bound for p: [0,589919]

Null Hypothesis: proportion = 0,5
Alternative: less than
P-Value = 0,821015
Do not reject the null hypothesis for alpha = 0,05.

$$P\left(\frac{\pi - 0{,}53}{\sqrt{0{,}53.(1-0{,}53)}}.\sqrt{200} \langle x_{0,95} \right) = 0{,}95 \quad \text{tudíž} \quad P(\pi \langle 0{,}59) = 0{,}95$$

Závěr: Pravostranným intervalem spolehlivosti pro podíl je tedy interval (0 %, 59 %). Nulovou hypotézu (50 % podporu) nelze na 5 % hladině významnosti zamítnout. Nelze říci, že pokud v předvolebním průzkumu dostala strana XYZ 53 % hlasů, pak dostane u skutečných voleb nadpoloviční (více než 50 %) většinu hlasů.

Ad b) Chceme odhadnout počet respondentů, aby chyba odhadu činila maximálně 2 %

Menu Describe/Sample-SizeDetermination

Binomial Proportion … Hypothesized Proportion … 0,5 (50 %)

Absolute Error … 0,02 (2 %)

Dostaneme výstup („ruční výpočet“ si proveďte sami):

Sample-Size Determination

Parameter to be estimated: binomial parameter
Desired tolerance: +- 0,02 when proportion = 0,5
Confidence level: 95,0%
The required sample size is n=2599 observations.

Závěr: Doporučuje se oslovit 2599 respondentů, abychom dostali odhad s chybou maximálně 2 %.

5.2. Dvouvýběrové testy

Dvouvýběrové testy používáme tehdy, chceme porovnat dva základní soubory a máme k dispozici dva výběrové soubory (z každého základního souboru jeden). Níže uvedené testy se dají použít v případě, že náhodné výběry jsou nezávislé.

Testování rozdílu středních hodnot

Příklad 5.8. *U 12-ti náhodně vybraných rodin se 2-mi dětmi byly zjištěny roční výdaje na průmyslové zboží (v tisících Kč):*

41,2 39,4 36,3 38,7 39,9 38,3 40,6 41,5 37,4 43,1 35,7 35,8.

Obdobně u šesti náhodně vybraných rodin se 4-mi dětmi byly údaje následující:

39,2 43,8 38,9 44,3 41,2 44,1.

Zjistěte, zda se střední hodnota ročních výdajů na průmyslové zboží liší u rodin se 2-mi a 4-mi dětmi.

Řešení: Nejprve je třeba provést explorační analýzu obou datových souborů.

Menu Describe/Numeric Data/OneVariableAnalysis

Dostaneme následující charakteristiky proměnné „dve_deti"

Summary Statistics for dve_deti
Count (počet) **= *12***
Average (výběrový průměr) **= *38,9917***
Standard deviation (výběrová směrodatná odchylka) **= *2,39297***

a charakteristiky proměnné „ctyri_deti"

Summary Statistics for ctyri_deti
Count (počet) **= *6***
Average (výběrový průměr) **= *41,9167***
Standard deviation (výběrová směrodatná odchylka) **= *2,48951***

Testujeme hypotézu o rovnosti středních hodnot:

$$H_0 : \mu_1 = \mu_2$$

oproti alternativě

$$H_A : \mu_1 \neq \mu_2$$

Neznáme σ_1, σ_2 - rozptyly základních souborů, proto volíme výběrovou charakteristiku T_2 s $(n_1 + n_2 - 2)$ stupni volnosti neboli dvouvýběrový t-test:

$$T_2 = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2), \text{ kde}$$

$$s_P = \frac{\sqrt{(n_1 - 1).s_1^2 + (n_2 - 1).s_2^2}}{n_1 + n_2 - 2} = \sqrt{\frac{11.2{,}39^2 + 5.2{,}49^2}{16}} = 2{,}42.$$

Hodnotu testovacího kritéria a následně hodnotu p-value najdeme ve *Statgraphicsu* pomocí procedury **Compare** (srovnávat):

Menu Compare/TwoSamples/HypothesisTests

Zvolíme:

Null Hypothesis for Difference of Means … 0,0

(neexistuje žádný rozdíl mezi středními hodnotami výdajů rodin se 2-mi a 4-mi dětmi)

Sample 1 Mean … 38,9917
Sample 1 Sigma … 2,39297
Sample 1 Size … 12

Sample 2 Mean ... 41,9167
Sample 2 Sigma … 2,48951
Sample 2 Size … 6

Hodnota testovacího kritéria T_2 (***Computed t statistic***) je rovna:

$$x_{obs} = T_2 \big/ H_0 = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(38{,}99 - 41{,}92) - 0}{2{,}42\sqrt{\frac{1}{12} + \frac{1}{6}}} = -2{,}41$$

a hodnota ***P-Value = 0,0281345***

$$p - value = 2 \cdot \min\{F(x_{obs}), 1 - F(x_{obs})\} = 2 \cdot \min\{F(-2{,}42), 1 - F(-2{,}42)\} = 0{,}028 < 0{,}05.$$

Reject the null hypothesis for alpha = 0,05.

Závěr: Zamítáme (***reject***) nulovou hypotézu na 5 % hladině významnosti ve prospěch alternativy. Existuje statisticky významný rozdíl mezi výdaji rodin se 2-mi a 4-mi dětmi (neboli rodiny se 4-mi dětmi mají významně vyšší výdaje než rodiny se 2-mi dětmi, rozmyslete si proč).

Příklad 5.9. Testujte dvouvýběrovým t-testem hypotézu o rovnosti středních hodnot dvou základních souborů, pokud víte, že pro odpovídající výběrové charakteristiky platí:
$\overline{x_1} = 33{,}77$, $s_1 = 3{,}83$, $n_1 = 14$, $\overline{x_2} = 33{,}17$, $s_2 = 8{,}90$, $n_2 = 14$.

Řešení: Testujeme hypotézu o rovnosti středních hodnot dvou základních souborů

Menu Compare/TwoSamples/HypothesisTests

Hypothesis Tests

Sample means = 33,77 and 33,17
Sample standard deviations = 3,83 and 8,9
Sample sizes = 14 and 14

95,0% confidence interval for difference between means: 0,6 +/- 5,32285
[-4,72285;5,92285]

Null Hypothesis: difference between means = 0,0
Alternative: not equal
Computed t statistic = 0,231703
P-Value = 0,818583
Do not reject the null hypothesis for alpha = 0,05.

(Equal variances assumed): upozornění na nesplnění předpokladu rovnosti rozptylů základních souborů. Mezi rozptyly základních souborů existuje statisticky významný rozdíl tudíž nelze testovat rovnost středních hodnot pomocí dvouvýběrového t-testu.

V následující kapitole Testování hypotéz - neparametrické testy si ukážeme jednu z možností jak lze podobné problémy nesplnění předpokladů standardních testů řešit.

Testování rozdílu relativních četností

Příklad 5.10. *TV stanice zjišťuje sledovanost určitého pořadu a zajímá ji, zda u dospělých osob do 25 let („mladší osoby“) je tato sledovanost vyšší, než u věkově starších osob. Daný pořad sledovalo 80 z 500 náhodně vybraných mladších osob a 100 z 1000 náhodně vybraných starších osob.*
a) Najděte 99% interval spolehlivosti pro rozdíl podílů sledovanosti uvedeného pořadu u těchto dvou věkových skupin .
b) Otestujte danou hypotézu.

Řešení: Ad a) Jelikož zkoumáme relativní četnost prvků v základním souboru mající danou vlastnost, zvolíme testovací kritérium P_2:

$$P_2 = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{p.(1-p).\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)\text{, kde } p_1 = \frac{x_1}{n_1} = \frac{80}{500} = 0{,}16,\quad p_2 = \frac{x_2}{n_2} = \frac{100}{1000} = 0{,}1,$$

$$p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{180}{1500} = 0{,}12,\quad \alpha = 0{,}01 \Rightarrow \frac{\alpha}{2} = 0{,}005$$

K výpočtu použijeme proceduru

Menu Compare/TwoSamples/HypothesisTests

Zvolíme: **Binomial Proportions**

Null Hypothesis for Difference of Proportions … 0.0

(neexistuje žádný rozdíl mezí sledovaností TV pořadu u jednotlivých skupin)

Sample 1 Proportion … 0,16
Sample 2 Proportion … 0,1

Sample Size … 500
Sample Size … 1 000

Pravým tlačítkem myši změníme ve vzniklém výstupu standardně stanovený 95 % interval spolehlivosti na 99 % interval spolehlivosti:

Analysis Options … Alpha … 1 (%)

Pravostranným intervalem spolehlivosti pro rozdíl podílů je interval:

$$P\left(z_{0,005} < \frac{0{,}16 - 0{,}1 - (\pi_1 - \pi_2)}{\sqrt{0{,}12.0{,}88.\left(\frac{1}{500} + \frac{1}{1000}\right)}} < z_{0,995}\right) = 0{,}99 \Rightarrow P\left(-2{,}55 < \frac{0{,}06 - (\pi_1 - \pi_2)}{\sqrt{0{,}12.0{,}88.\left(\frac{1}{500} + \frac{1}{1000}\right)}} < 2{,}55\right) = 0{,}99 \Rightarrow$$

Approximate 99,0% confidence interval for difference between proportions: [0,0112085;0,108791]

$$P(0{,}011 < \pi_1 - \pi_2 < 0{,}109) = 0{,}99$$
$$\pi_1 - \pi_2 \in (0{,}011; 0{,}109)$$

Závěr: 99,0 % intervalem spolehlivosti pro rozdíl podílů sledovanosti u daného TV pořadu mezi dospělými osobami do 25 let a věkově staršími osobami je interval (1,1 % ; 10,9 %).

Ad b) Testujeme hypotézu o rovnosti podílů sledovanosti TV pořadu u jednotlivých skupin:

$$H_0 : \pi_1 = \pi_2$$

oproti alternativě

$$H_A : \pi_1 > \pi_2$$

K testování této hypotézy použijeme již výše uvedené testovací kritérium P_2:

$$P_2 = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{p.(1-p).\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1).$$

K nastavení jednostranné alternativy použijeme pravé tlačítko myši a zvolíme

Analysis Options … Greather Than

Ve výstupu se nám typ zvolené alternativy zobrazí na řádku hned pod nulovou hypotézou pod označením ***Alternative***:

Null Hypothesis: difference between proportions = 0,0
Alternative: greater than

Hodnotu testovacího kritéria a hodnotu p-value jsme vypočetli pomocí Statgraphicsu ***Computed z statistic = 3,371***:

$$x_{obs} = P_2 \big/ H_0 = \frac{0{,}16 - 0{,}1 - (\pi_1 - \pi_2)}{\sqrt{0{,}12.0{,}88.\left(\frac{1}{500} + \frac{1}{1000}\right)}} = 3{,}371$$

P-Value = 0,000374533

$p - value = 1 - F(x_{obs}) = 1 - F(3{,}37) = 0{,}00037 < 0{,}01 \Rightarrow H_0$ se zamítá na 1 % hladině významnosti..

Reject the null hypothesis for alpha = 0,01.

Závěr: Zamítáme ***(reject)*** nulovou hypotézu na hladině významnosti $\alpha = 0{,}01$ ve prospěch alternativy.

Sledovanost daného TV pořadu u dospělých osob do 25 let se statisticky významně liší od sledovanosti u věkově starších osob. U mladších osob je sledovanost statisticky významně vyšší než u věkově starších osob.

6. Testování hypotéz – neparametrické testy

6.1. Testování normality dat

Příklad 6.1. *Zjistěte, zda data vytvořená v 10. příkladě 4. části tohoto textu lze považovat za náhodný výběr z normálního rozdělení (vygenerovaný soubor náhodných čísel „random_numbers“).*

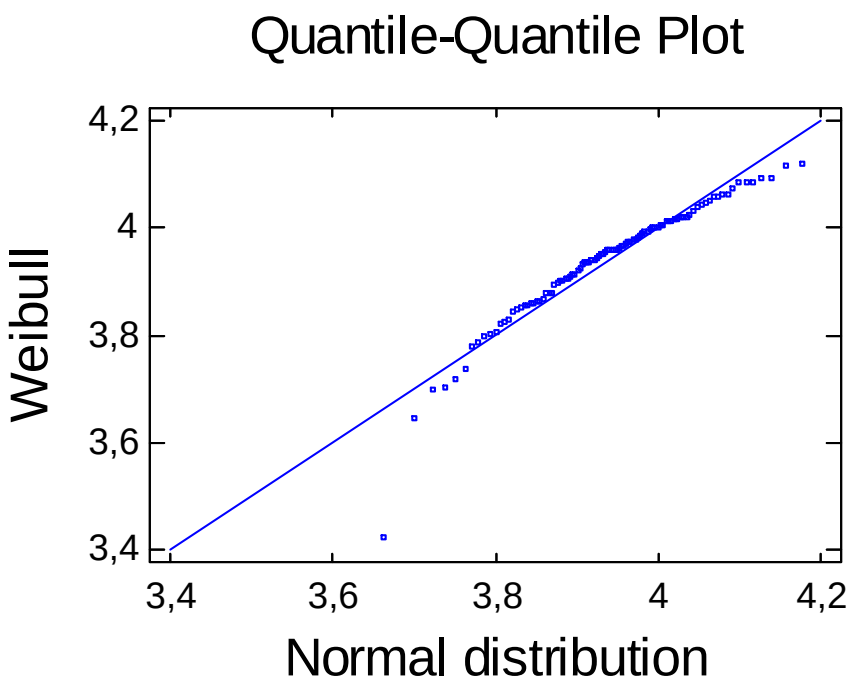
Řešení: Program Statgraphics umožňuje testovat normalitu dat několika způsoby:

a) Ověření normality pomocí Q-Q grafu

Menu Desribe/Distributions/Distribution Fitting (Uncensored Data)

Proceduru provedeme pro proměnnou označenou “Weibull”

Graphical Options (3. ikona): zvolíme **Density Trace** (křivka hustoty rozdělení) a **Quantille-Quantille Plot** (Q-Q graf vyjadřující poměr mezi kvantily zadaného rozdělení a normálního rozdělení):



Grafické vyhodnocení normality pomocí Q-Q grafu: pokud body leží na zobrazené přímce, lze data považovat za výběr z normálního rozdělení, pokud „utíkají“ dál od přímky, vzdalují se data od předpokládané normality. V našem případě nelze náhodná čísla „Weibull“ považovat za přibližně normální (dvě nejnižší hodnoty kvantilů jsou hodně vzdáleny od odpovídajících kvantilů normálního rozdělení).

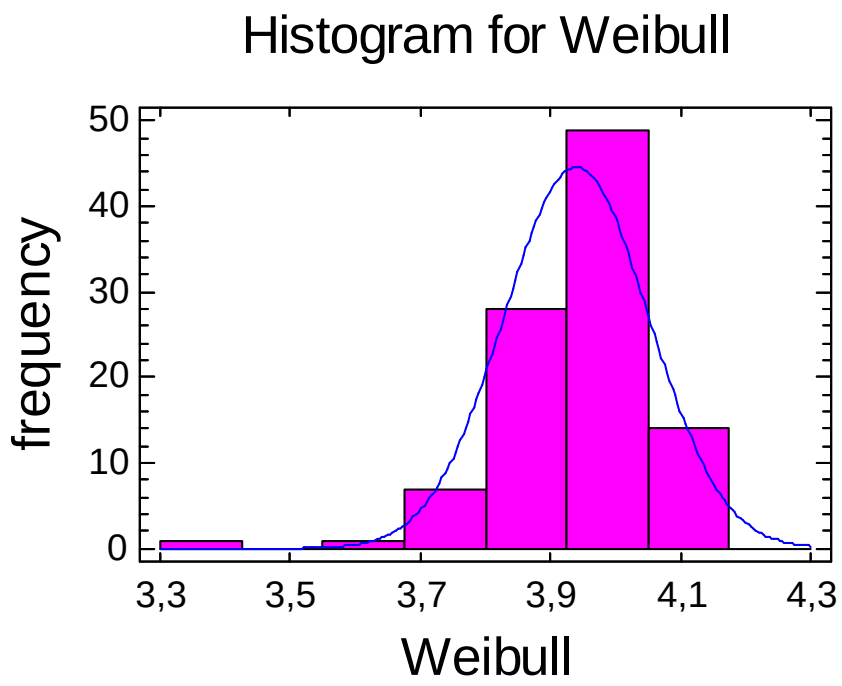
b) Ověření normality pomocí šikmosti a špičatosti

Menu Desribe/NumericData/OneVariableAnalysis

Zobrazí se (volba dalších kvantilů: **Pane Options**):

Count = 100
Average = 3,93958
Median = 3,95812
Variance = 0,0125302
Standard deviation = 0,111938
Minimum = 3,42096
Maximum = 4,11969
Range = 0,69873
Lower quartile = 3,87897
Upper quartile = 4,01307
Skewness (**šikmost**)= -1,34783
Stnd. skewness = -5,50248
Kurtosis (**špičatost **)= 3,86857
Stnd. kurtosis= 7,89669

Pokud je šikmost i špičatost blízko nuly, jedná se o výběr z normálního rozdělení.
V našem případě se jedná se o záporně sešikmená data - **Skewness (**šikmost**)= -1,34783** (data se nacházejí převážně v pravé části intervalu) s velkou špičatostí **Kurtosis (**špičatost **)= 3,86857** (prostřední sloupec je vyšší, než by odpovídal normálnímu rozdělení – viz zakreslená Gaussova křivka).



Zkusíme zjistit, zda porušení normality nebylo způsobeno odlehlým pozorováním (výše uvedené grafické výstupy toto signalizují):

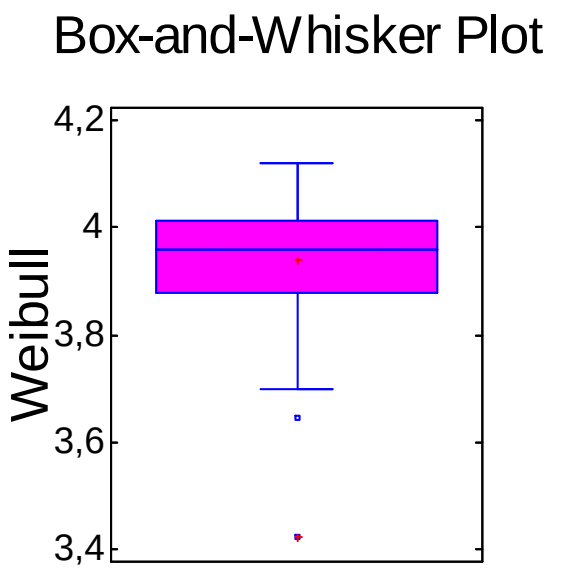
Menu Desribe/NumericData/OutlierIdentification

Dostaneme tabulku hodnot (v levém dolním okně) a jejich mediánové souřadnice (v posledním sloupci):

Sorted Values

Row	*Value*	*Studentized Values Without Deletion*	*Modified With Deletion*	*MAD Z-Score*
73	*3,42096*	*-4,63308*	*-5,26875*	*-5,94733*
97	*3,64393*	*-2,64118*	*-2,75417*	*-3,47862*
16	*3,69909*	*-2,14841*	*-2,21183*	*-2,8679*
90	*3,7039*	*-2,10544*	*-2,16547*	*-2,81464*
66	*3,71555*	*-2,00136*	*-2,05375*	*-2,68565*
...				
78	*4,09253*	*1,36638*	*1,38647*	*1,48823*
55	*4,09383*	*1,378*	*1,39848*	*1,50262*
38	*4,11614*	*1,5773*	*1,60568*	*1,74964*
91	*4,11916*	*1,60428*	*1,63388*	*1,78307*
92	*4,11969*	*1,60902*	*1,63883*	*1,78894*

Hodnoty na 73. a 97. řádku mají mediánovou souřadnici v absolutní hodnotě větší než tři, tedy v souboru se nacházejí 2 odlehlá pozorování (viz následující obrázek):



Nyní ověříme, zda soubor po vyjmutí těchto dvou odlehlých pozorování již bude splňovat podmínky normality. Vyjmeme tedy ze souboru dat (původní tabulka) 73. a 97. statistickou jednotku (řádek) a provedeme znova vyhodnocení normality. Tentokrát pomocí přesnějšího Pearsonova χ^2 - testu dobré shody.

c) Testování normality pomocí testů dobré shody (Goodness Fit Tests)

Zformulujeme nulovou a alternativní hypotézu:

H_0 … jedná se o náhodný výběr z normálního rozdělení
H_A … nejedná se o náhodný výběr z normálního rozdělení

Menu Desribe/Distributions/Distribution Fitting (Uncensored Data)

V levém dolním okně se nám zobrazí test dobré shody (χ^2 - test)

Goodness-of-Fit Tests for Weibull

Chi-Square Test

Lower Limit	*Upper Limit*	*Observed Frequency*	*Expected Frequency*	*Chi-Square*
at or below	*3,83848*	*12*	*12,25*	*0,01*
3,83848	*3,88374*	*13*	*12,25*	*0,05*
3,88374	*3,91758*	*9*	*12,25*	*0,86*
3,91758	*3,94789*	*10*	*12,25*	*0,41*
3,94789	*3,97819*	*15*	*12,25*	*0,62*
3,97819	*4,01204*	*13*	*12,25*	*0,05*
4,01204	*4,05729*	*13*	*12,25*	*0,05*
above	*4,05729*	*13*	*12,25*	*0,05*

Chi-Square = 2,0816 with 5 d.f. P-Value = 0,837743 *(Pearsonův χ^2 - test)*

Since the smallest P-value amongst the tests performed is greater than or equal to 0.10, we can not reject the idea that Weibull comes from a normal distribution with 90% or higher confidence.

Závěr: H_0 nelze zamítnout, tedy upravená data „Weibull“ (po vyjmutí 2 odlehlých pozorování) lze považovat za náhodný výběr z normálního rozdělení.

Příklad 6.2. *Pomocí χ^2 - testu testujte normalitu proměnné „Exp“ v souboru „random_numbers“. Dále tuto proměnnou transformujte pomocí funkce LOG a potvrďte, že tato transformovaná data lze již považovat za výběr z normálního rozdělení.*

Řešení: Otevřete si vytvořený soubor „numbers“ a ověřte, zda proměnnou „Exp“ lze považovat za náhodný výběr z normálního rozdělení:

Menu Desribe/Distributions/Distribution Fitting (Uncensored Data)

Podle p-value Pearsonova χ^2 - testu vyslovíme závěr: p-value = $6{,}98413 \cdot 10^{-9} < < 0{,}05$, tedy H_0 se zamítá. Data nelze považovat za výběr z normálního rozdělení.

V tomto případě zkusíme transformovat zadaná data:

Menu Edit/GenerateData

operator … LOG(?) - dvojité kliknutí myší na LOG(?)

variable … Exp – dvojité kliknutí na Exp

V okně se zobrazí **LOG(Exp)**. Potvrdíme Ok a nově vytvořený sloupec nazveme “log”.

Zkusme nyní vyhodnotit normalitu transformovaných dat:

Mann-Whitney W test
Podle p-value Pearsonova χ^2 - testu vyslovíme závěr: p-value = 0,047 leží v “hraničním pásmu” 0,01 až 0,05, tedy zde si dovolíme H_0 nezamítnout, i když už je to tzv. na hraně.

Závěr: Nezamítneme hypotézu o normalitě transformovaných dat (odklon od normality není tak velký jako u původních dat).

Pozn.1. Tuto proceduru - transformaci dat na logaritmická data můžeme často použít, pokud data nesplňují předpoklad normality. V praxi se často vyskytují lognormální data, tedy teprve jejich logaritmy splňují předpoklad normality. Následné testování hypotéz pak budeme provádět již s logaritmovanými hodnotami.

Pozn.2. Pokud p-value leží v “hraničním pásmu” 0,01 až 0,05 obvykle H_0 zamítáme i když ne s takovou důrazností jako v případě p-value < 0,01. Doporučuje se zopakovat náhodný výběr pokud možno s větším rozsahem výběru.

6.2. Kolmogorovův-Smirnovův test

Kolmogorovův-Smirnovův test se používá pro ověření hypotézy, zda data pocházejí z daného rozdělení se spojitou distribuční funkcí F_0. Podmínkou tohoto testu je znalost všech parametrů testovaného rozdělení. Používá se obvykle pokud je rozsah výběru malý a tudíž není vhodně použití χ^2 - testu dobré shody.

H_0 : základní soubor lze charakterizovat distribuční funkcí F_0

H_A : neplatí H_0

Příklad 6.3. *Ověřte Kolmogorovým-Smirnovovým testem, že náhodná čísla "Normal" vygenerovaná v souboru „numbers" odpovídají normálnímu rozdělení s parametry* $\mu = 8, \sigma = 3$.

Řešení: Otevřeme soubor „numbers" vytvořených náhodných čísel.

Menu Desribe/Distributions/Distribution Fitting (Uncensored Data)

Jako data zvolíme: „Normal"

Statgraphics automaticky nastaví teoretickou distribuční funkci normálního rozdělení s danými parametry.

V dolní části levého okna se nám zobrazí výsledek Kolmogorova-Smirnovova testu:

EDF Statistic	*Value*	*Modified Form*	*P-Value*
Kolmogorov-Smirnov D	*0,0732829*	*0,738325*	*>=0.10**
Anderson-Darling A^2	*0,425133*	*0,428417*	*0,3107**

Indicates that the P-Value has been compared to tables of critical values specially constructed for fitting the currently selected distribution. Other P-values are based on general tables and may be very conservative.

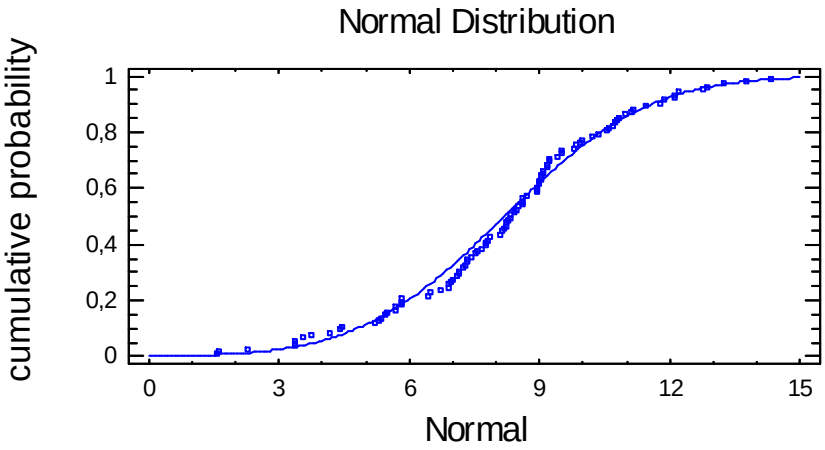
The StatAdvisor

Kolmogorov-Smirnov test computes the maximum distance between the cumulative distribution of Normal and the CDF of the fitted normal distribution. In this case, the maximum distance is 0,0732829. The other EDF statistics compare the empirical distribution function to the fitted CDF in different ways.

Since the smallest P-value amongst the tests performed is greater than or equal to 0.10, we can not reject the idea that Normal comes from a normal distribution with 90% or higher confidence.

Grafický výstup: body odpovídají zadaným hodnotám v souboru – jejich kumulativním relativním četnostem (empirická distribuční funkce), křivka je teoretická distribuční funkce normálního rozdělení s parametry $\mu = 8, \sigma = 3$:

Graphic Options - zvolíme **Quantile Plot**



Závěr: Nelze zamítnout hypotézu, že teoretická distribuční funkce základního souboru je distribuční funkce normálního rozdělení s parametry $\mu = 8, \sigma = 3$, tedy že data pocházejí z normálního rozdělení s parametry $\mu = 8, \sigma = 3$.

6.3. Test nezávislosti dvou kategoriálních proměnných

Příklad 6.4. *Zjistěte, zda počet válců automobilů závisí za zemi původu (soubor Cardata).*

Řešení: Otevřeme soubor „CarData“ (hledáme v adresáři Statgra/TestData).

Menu Describe/CategoricalData/Crosstabulation

Row Variable … **origin**

Column Variable … **cylinders**

V dolním levém okně se zobrazí tabulka relativních četností počtu válců u jednotlivých zemí původu:

Frequency Table for origin by cylinders

	3	4	5	6	8	Row Total
1	0	44	0	24	17	85
	0,00%	28,39%	0,00%	15,48%	10,97%	54,84%
2	0	20	3	3	0	26
	0,00%	12,90%	1,94%	1,94%	0,00%	16,77%
3	1	40	0	3	0	44
	0,65%	25,81%	0,00%	1,94%	0,00%	28,39%
Column Total	1	104	3	30	17	155
	0,65%	67,10%	1,94%	19,35%	10,97%	100,00%

Nyní budeme testovat, zda počet válců závisí na zemi původu. Nulovou hypotézu vyslovíme vždy ve tvaru nezávislosti, i když chceme potvrdit závislost (všechny indicie ukazují na platnost H_A):

H_0 : počet válců nezávisí na zemi výroby

H_A : neplatí H_0

χ^2 - test (žlutá ikona **Tabular Options** přidat **Chi-Square Test**):

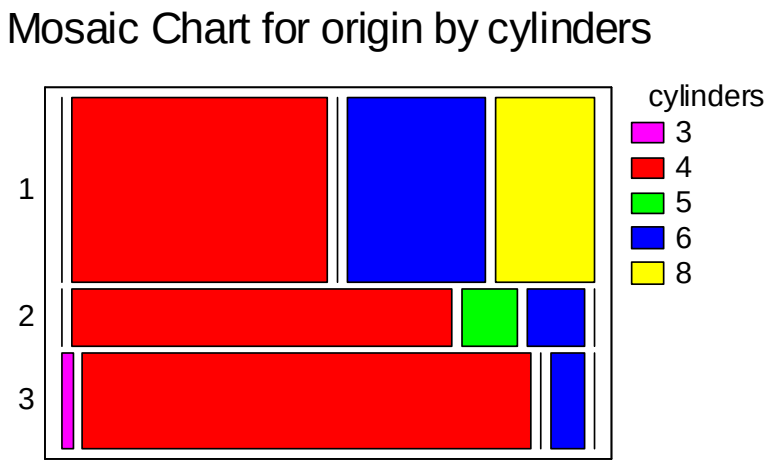
Chi-Square Test

Chi-Square	*Df*	*P-Value*
46,33	*8*	*0,0000*

Warning: some cell counts < 5.

Zobrazilo se varování, že počty teoretických (očekávaných) četností jsou v některých buňkách tabulky menší než pět. Tedy v tomto případě nejsou korektně splněny předpoklady tohoto χ^2 - testu (počet pozorování v každé sdružené kategorii teoretických četností by měl být větší než pět).

Jelikož nejsou splněny předpoklady použití χ^2 - testu , provedeme vyhodnocení pouze na základě mozaikového grafu, tedy pomocí prostředků explorační analýzy.



Na první pohled vidíme rozdíl v umístění barev v jednotlivých pásech. Kdyby mozaikový graf obsahoval pouze svislé barevné pásy, tedy rozmístěné barev by bylo nezávislé na jednotlivých řádcích, znamenalo by to, že kategoriální proměnné jsou nezávislé. V našem případě vidíme, že proměnná **cylinders** závisí na proměnné **origin**.

Tedy počet válců u auta závisí na zemi výroby. Americká auta (1) mají asi poloviční podíl čtyřválců a čtvrtinové podíly šesti a osmiválců, kdežto Evropská (2) a Japonská (3) auta mají jiné zastoupení aut co se týče počtu válců – viz zobrazený mozaikový graf.

Příklad 6.5. *Zjistěte, zda existuje statisticky významná závislost mezi typem absolvované střední školy a známkou ze statistiky. Data jsou zadána ve formě kontingenční tabulky:*

Známka ze statistiky

Střední škola	**výborně**	**velmi dobře**	**dobře**
Gymnázium	9	28	15
SPŠ, SOŠ	20	66	50
SOU	8	25	26

Řešení: Zadáme data do *Statgraphicsu* (pozor! 1. sloupec je kategoriální proměnná (nazvěme „střední škola“); 2.- 4. je numerická proměnná, 2. sloupec nazvěme „vyborne“, 3. sloupec „velmi dobre“, 4. sloupec „dobre“).

Nulovou hypotézu opět vyslovíme vždy ve tvaru nezávislosti.

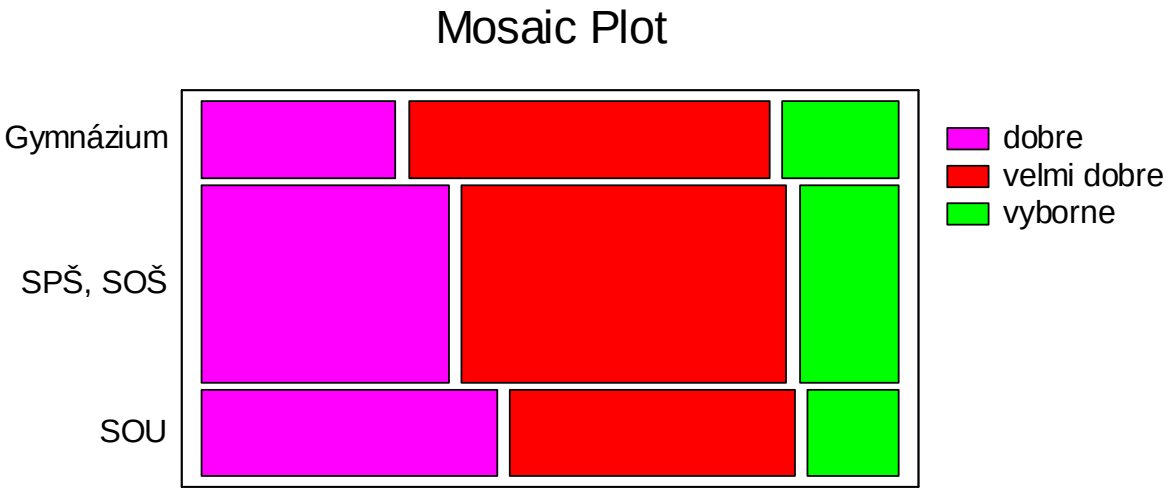
H_0: známka ze statistiky nezávisí na typu střední školy

H_A: neplatí H_0

Menu Describe/CategoricalData/ContingencyTables

Columns … „vyborne“, „velmi dobre“, „dobre“

Labels … „střední škola“



Opticky z mozaikového grafu vidíme, že rozmístěné barev je téměř stejné v jednotlivých řádcích, což signalizuje nezávislost kategoriálních proměnných. Přesný závěr vyslovíme pomocí p-value χ^2 - testu (předpoklad, aby v každé buňce tabulky teoretických četností, byly hodnoty větší než pět, je tentokrát splněn):

Chi-Square Test

Chi-Square	*Df*	*P-Value*
2,78	*4*	*0,5960*

The StatAdvisor

Since the P-value is greater than or equal to 0.10, we cannot reject the hypothesis that rows and columns are independent.

Závěr: p-value χ^2 - testu vyšlo 0,6166 > 0,05 tedy H_0 nelze zamítnout.

Nebyla potvrzena existence statisticky významné závislosti mezi známkou ze statistiky a typem absolvované střední školy.

6. 4. Pořadové testy

Pořadové testy se obvykle používají, pokud data nesplňují předpoklad normality. Jedná se o neparametrické testy, které nekladou požadavky na normalitu dat. Tedy místo testování střední hodnoty normálního rozdělení (parametrický test, předpokládá normalitu) testujeme medián základního souboru (neparametrický test, nepožaduje normalitu dat), místo rovnosti středních hodnot dvou základních souborů testujeme rovnost mediánů těchto základních souborů.

Jednovýběrový test mediánu

Příklad 6.6. *U „stogramových" balíčků kávy byly zjištěny následující hmotnosti:*

99, 103, 102, 97, 97, 98, 101, 102, 100, 100, 100, 95, 104, 98, 101, 103, 97, 96, 98, 100 g.

Určete pomocí znaménkového testu, zda medián hmotností balíčků kávy daného výrobce je roven 100 gramů.

Řešení: Zformulujeme nulovou a alternativní hypotézu:

$H_0 : x_{0,5} = 100$

$H_A : x_{0,5} \neq 100$

Tuto hypotézu budeme testovat pomocí tzv. **znaménkového testu** (**Sign Test**).

Menu Describe/NumericData/One-VariableAnalysis

TabularOptions (žlutá ikona), zvolíme **Hypothesis Tests**

Pravým tlačítkem myši: **Pane Options**: změníme: Mean=100

sign test

Null hypothesis: median = 100,0
Alternative: not equal

Number of values below hypothesized median: 9
Number of values above hypothesized median: 7

Large sample test statistic = 0,25 (continuity correction applied)
P-Value = 0,802583

Do not reject the null hypothesis for alpha = 0,05.

Závěr: Nelze zamítnout hypotézu, že medián hmotnosti balíčků kávy je roven 100 g.

Pozn: Tento test má velice malou sílu testu, tedy vypovídací schopnost. Může být velmi nepřesný. Existuje velké riziko chyby II. druhu, tj. že přijmeme platnost nulové hypotézy (nezamítneme H_0), i když ve skutečnosti neplatí.

Dvouvýběrový test mediánů

Příklad 6.7. *Za účelem analýzy hrubé měsíční mzdy bylo dotázáno 20 osob v jeden den v určitém městě v ČR (zkrácená verze datového souboru uveřejněného na internetových stránkách ČSÚ):*

Pohlaví	**Věk**	**Hrubá mzda (tis. Kč)**
M	29	12
Z	33	15,6
M	26	20
M	31	23,1
M	40	24
M	52	23,6
Z	38	19,5
M	19	22
Z	41	18,5
Z	55	23,8
M	40	18
Z	58	10,5
Z	21	13,4
Z	29	17,5
Z	31	18,5
Z	44	16,9
Z	46	17,1
M	39	27
M	32	19,9
Z	30	17,3

Zjistěte, zda výše hrubé mzdy v daném městě závisí na pohlaví. Použijte Mann Whitneyho test.

Řešení: Nejprve otestujeme normalitu dat v obou výběrových souborech:

Menu Desribe/Distributions/Distribution Fitting (Uncensored Data)

Jako Data zvolíme „Hrubá mzda“ a jako Select postupně : „Pohlaví=1“, „Pohlaví =2“

(nejprve je třeba transformovat proměnnou pohlaví na číselnou proměnnou, přiřadíme např. M=1, Z=2).

Normalita sice v obou výběrech „prošla“, tedy je zde možné použít pro srovnání obou základních souborů buď středních hodnot, anebo mediánů. Zvolíme z ilustrativních důvodů 2. metodu. Obecně pokud data „projdou normalitou“ je lepší testovat porovnání dvou základních souborů pomocí středních hodnot nikoli pomocí mediánů, neboť 1. metoda dává přesnější výsledky. 2. metoda má velkou chybu II. druhu, tudíž je vhodné použít ji pouze v případě, kdy není splněna normalita a 1. metodu použít nelze .

Zformulujeme nulovou a alternativní hypotézu:

$H_0 : x_{0,5}^1 = x_{0,5}^2$

$H_A : x_{0,5}^1 \neq x_{0,5}^2$

K řešení použijme Mann-Whitney W test pro srovnání mediánů dvou základních souborů:

Menu Compare/Two Samples/TwoSamplesComparison

V zobrazeném okně Two Sample Comparison v kolonce Input (vpravo dole) přepneme na „Data and Code Columns“. Zadáme

Data … Hrubá mzda,

Sample Code … Pohlaví.

V zobrazeném výstupu klikneme na

TabularOptions (žlutá ikona) a zvolíme **Comparison of Medians**

Comparison of Medians for Hrubá mzda

Median of sample 1: 22,0
Median of sample 2: 17,3

Mann-Whitney (Wilcoxon) W test to compare medians

Null hypothesis: median1 = median2
Alt. hypothesis: median1 NE median2

Average rank of sample 1: 14,0
Average rank of sample 2: 7,63636

W = 18,0 P-value = 0,0184693

The StatAdvisor

This option runs a Mann-Whitney W test to compare the medians of the two samples. This test is constructed by combining the two samples, sorting the data from smallest to largest, and comparing the average ranks of the two samples in the combined data. Since the P-value is less than 0,05, there is a statistically significant difference between the medians at the 95,0% confidence level.Median of sample 1: 57,5 Median of sample 2: 62,5

Závěr: p-value Mann-Whitney W testu vyšlo 0,0184693< 0,05, tedy H_0 se zamítá. Zamítá se hypotéza, že mediány obou základních souborů jsou stejné.

Existuje statisticky významný rozdíl mezi výší hrubé mzdy u mužů a žen v daném městě. Výše hrubé mzdy v daném městě závisí na pohlaví.

Párový test pro srovnání mediánů dvou základních souborů

Příklad 6.8. *Vedení záchranné služby chce prokázat, že noční směna ošetří více případů, než denní směna. Údaje v tabulce odpovídají měsíci duben 2007 (počet případů ošetřených záchrannou službou v určitý den v měsíci rozdělený podle denních a nočních směn):*

Denní	*55*	*48*	*51*	*58*	*63*	*57*	*62*	*64*	*50*	*61*	*58*	*60*	*70*	*73*	*49*
Noční	*62*	*37*	*51*	*56*	*68*	*63*	*54*	*64*	*56*	*64*	*57*	*48*	*72*	*73*	*56*

Denní	*65*	*56*	*65*	*49*	*50*	*63*	*49*	*57*	*49*	*56*	*68*	*69*	*57*	*57*	*58*
Noční	*69*	*65*	*73*	*49*	*57*	*69*	*52*	*68*	*59*	*56*	*72*	*64*	*68*	*64*	*60*

Otestujte pomocí párového testu mediánů.

Řešení: Hodnoty týkající se denní a noční směny jsou vždy vázány k určitému dni. Je nebytné tedy zde tedy použít párový test, který nevyhodnocuje mediány obou souborů nezávisle na sobě, ale vyhodnocuje vždy hodnoty v páru, tedy současné hodnoty denní i ranní směny za konkrétní den. Vyhodnocení se provádí pomocí vyhodnocení diferencí *d* u každého jednotlivého páru.

Denní	*55*	*48*	*51*	*58*	*63*	*57*	*62*	*64*	*50*	*61*	*58*	*60*	*70*	*73*	*49*
Noční	*62*	*37*	*51*	*56*	*68*	*63*	*54*	*64*	*56*	*64*	*57*	*48*	*72*	*73*	*56*
d	*-7*	*11*	*0*	*2*	*-5*	*-6*	*8*	*0*	*-6*	*-3*	*1*	*12*	*-2*	*0*	*-7*

Denní	*65*	*56*	*65*	*49*	*50*	*63*	*49*	*57*	*49*	*56*	*68*	*69*	*57*	*57*	*58*
Noční	*69*	*65*	*73*	*49*	*57*	*69*	*52*	*68*	*59*	*56*	*72*	*64*	*68*	*64*	*60*
d	*-4*	*-9*	*-8*	*0*	*-7*	*-6*	*-3*	*-11*	*-10*	*0*	*-4*	*5*	*-11*	*-7*	*-2*

Dvourozměrnou úlohu srovnávání mediánů dvou souborů tedy převádíme na jednorozměrnou úlohu testování diferencí. Jelikož není splněna normalita dat, použijeme znaménkový test pro otestování hypotézy, zda je medián diferencí roven 0 (pokud by data splňovala podmínky normality, použili bychom raději přesnější t-test pro otestování hypotézy, zda je střední hodnota mediánů rovna 0).

$H_0 : x_{0,5}^d = 0$

$H_A : x_{0,5}^d \neq 0$

Menu Compare/Two Samples/Paired-SamplesComparison

Zvolíme:

Sample1 … denní

Sample2 … noční

TabularOptions (žlutá ikona), zvolíme **Hypothesis Tests:**

sign test

Null hypothesis: median = 0,0
Alternative: not equal

Number of values below hypothesized median: 19
Number of values above hypothesized median: 6

Large sample test statistic = 2,4 (continuity correction applied)
P-Value = 0,016395

Reject the null hypothesis for alpha = 0,05.

Závěr: Zamítáme H_0. Existuje statisticky významný rozdíl mezi počty případů ošetřených noční a denní směnou.

Pozor! Pokud bychom tento příklad chtěli řešit předchozí metodou porovnání mediánů dvou nezávislých základních souborů a formálně bychom použili Mann-Whitney W test, vyšlo by nám, že neexistuje statisticky významný rozdíl mezi počty případů ošetřených noční a denní směnou. Tedy úplně opačný závěr, než závěr získaný párovým testem. Kde je tedy chyba? Chyba je v použití Mann-Whitney W testu, neboť náhodné výběry nejsou nezávislé. Pokud jsou hodnoty zadané „v páru“, jedná se o závislé náhodné výběry a je třeba použít párový test.

7. Jednofaktorová analýza rozptylu - ANOVA

Výchozí předpoklady pro analýzu rozptylu ANOVA můžeme shrnout do následujících tří bodů:

a) všechny výběry (tzv. třídy) lze považovat za náhodný výběr z normálního rozdělení,
b) je splněna podmínka tzv. homoskedasticity, tj. neexistuje statisticky významný rozdíl mezi rozptyly jednotlivých tříd,
c) jednotlivé náhodné výběry jsou nezávislé.

Příklad 7.1. *Na 48-mi pozemcích rozmístěných ve zkoumané zemědělské oblasti byly provedeny pokusy se 4-mi druhy olejnatých rostlin. Výsledky udávající množství získaného oleje v tunách na hektar jsou uvedeny podle jednotlivých druhů rostlin:*

Hořčice	**řepka**	**lnička**	**Sója**
0,188	0,415	0,382	0,227
0,067	0,261	0,199	0,357
0,232	0,113	0,473	0,402
0,124	0,114	0,262	0,267
0,285	0,062	0,152	0,017
0,300	0,270	0,293	0,240
0,387	0,068	0,428	0,167
0,155	0,1960	0,241	0,321
0,031	0,308	0,195	0,179
	0,365		0,020
	0,230		0,280
	0,262		0,384
	0,050		0,214
	0,127		0,168
	0,078		0,086

Určete, zda existuje závislost mezi množstvím získaného oleje a druhem rostliny, ze které byl olej získán (použijte jednofaktorovou analýzu roztylu ANOVA).

Řešení*:* Ověření předpokladů:

a) Otestujeme normalitu dat v každé třídě:

Nulová a alternativní hypotéza pro ověření normality dat:

H_0 … hodnoty proměnné „hořčice“ lze považovat za náhodný výběr z normálního rozdělení

H_A … hodnoty proměnné „hořčice“ nelze považovat za náhodný výběr z normálního rozdělení

Menu Desribe/Distributions/Distribution Fitting (Uncensored Data)

Zvolíme Data … hořčice

a postupně provedeme ověření normality i pro zbylé tři třídy:

Distributoin Fitting (1. ikona – červená), volíme postupně dále ... řepka, lnička, sója

P-value všech testů vyšlo větší než 0,05, tedy nezamítáme nulovou hypotézu, neboli výběr ve všech třídách lze považovat za náhodný výběr z normálního rozdělení.

b) Dále potvrdíme rovnost rozptylů, resp. směrodatných odchylek (tzv homoskedasticitu):

Testujeme hypotézu

$$H_0 : \sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$$

oproti alternativě

$$H_A : \text{neplatí } H_0,$$

Ve *Statgraphicsu* použijeme proceduru, která se týká již samotné analýzy rozptylu ANOVA:

Menu Compare/MultipleSamples/ Multiple-Sample Comparison

Input: Multiple Data Colums (jednotlivé třídy jsou zadány ve sloupcích)

Tabular Options (žlutá ikona) … **Variance Check**

Jelikož nebyla porušena podmínka normality, použijeme k vyhodnocení Bartlettův test.
P-value Bartletova testu vyšlo větší než 0,05, tedy nezamítáme nulovou hypotézu, neboli nebyla porušena podmínka homoskedasticity.

c) Ze zadání je zřejmé, že jednotlivé náhodné výběry jsou nezávislé.

Nyní tedy již můžeme přistoupit k samotné analýze rozptylu ANOVA.

Při analýze rozptylu provádíme testování hypotézy o rovnosti středních hodnot:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

oproti alternativě

$$H_A : \text{neplatí } H_0,$$

kde μ_1 je střední hodnota množství získaného oleje u hořčice,

μ_2 je střední hodnota množství získaného oleje u řepky,

μ_3 je střední hodnota množství získaného oleje u lničky,

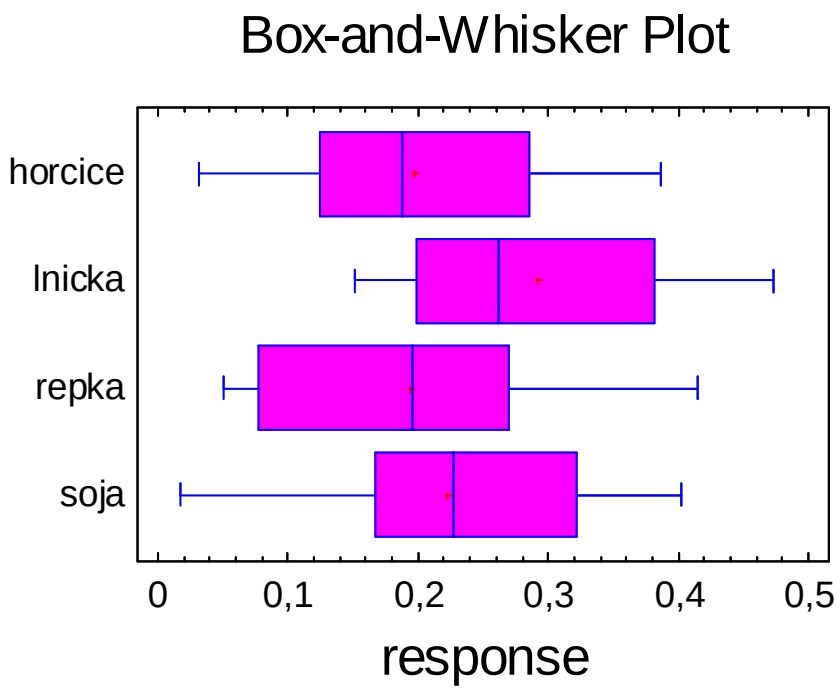
μ_4 je střední hodnota množství získaného oleje u sóji,

Budeme tedy srovnávat střední hodnoty množství získaného oleje u uvedených čtyř druhů rostlin (tzv. tříd). Použijeme proceduru **Compare** (srovnávat). Srovnáváme více než dvě třídy, tedy **Multiple-Sample Comparison:**

Menu Compare/MultipleSamples/ Multiple-Sample Comparison

Input: Multiple Data Colums (jednotlivé třídy jsou zadány ve sloupcích)

Grafický výstup (vícenásobný krabicový graf):



Tabulka ANOVA:

ANOVA Table

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Between groups	*0,0607652*	*3*	*0,0202551*	*1,48*	*0,2327*
Within groups	*0,601613*	*44*	*0,013673*		
Total (Corr.)	*0,662378*	*47*			

V tabulce ANOVA jsou již uvedeny všechny údaje, které potřebujeme k vyhodnocení testované hypotézy:

p-value F-testu je větší než 0,05, tedy neexistuje statisticky významný rozdíl mezi středními hodnotami jednotlivých tříd (***Since the P-value of the F-test is greater than or equal to 0,05, there is not a statistically significant difference between the means***).

Shrnutí: p-value vyšlo 0,2327, což je více než 0,05, tedy H_0 nelze zamítnout.

Neexistuje statisticky významný rozdíl mezi středními hodnotami množství získaného oleje zmíněných druhů olejnatých rostlin.

Závěr: Na základě uvedených údajů nebyla potvrzena závislost mezi množstvím získaného oleje a druhem rostliny, ze které byl olej získán.

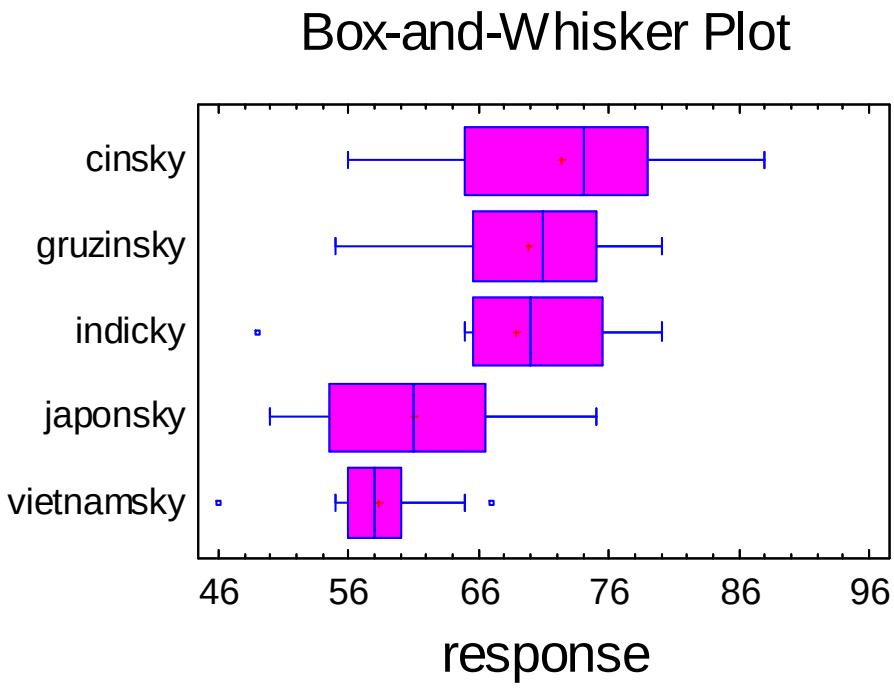
Příklad 7.2. *Majitel čajovny nabízí hostům různé čaje: čínský, indický, japonský, gruzínský a vietnamský. Rád by věděl, zda jsou všechny skupiny čajů stejně oblíbené. Proto požádal náhodně vybrané zákazníky, aby zhodnotili jednotlivé druhy čajů na žebříčku od 0 do 100. Získal tyto výsledky:*

čínský	**indický**	**Japonský**	**gruzínský**	**vietnamský**
60	49	51	55	46
68	72	68	71	65
88	77	70	78	67
79	65	75	71	58
72	68	57	65	60
65	66	62	66	56
56	74	60	80	58
76	80	65	72	55
77		58		60
83		63		
		50		
		52		

Lze na základě uvedených údajů tvrdit, že jsou všechny skupiny čajů zákazníky hodnoceny stejně, tedy stejně oblíbené?

***Řešení**:* Pokud bychom postupovali stejně jako u předchozího příkladu, dostali bychom se k následujícímu grafickému výstupu:

Vícenásobný krabicový graf s odlehlým pozorováním:



Když si jej pozorně prohlédnete, něco Vám na něm nebude sedět. Znázorněné tečky, tj. odlehlá pozorování. Mechanickým ověřením normality by předpoklad normality prošel ve všech třídách, i homoskedasticita, ale pokud jsou v souboru evidentní odlehlá pozorování, může F – test dávat zkreslené výsledky. Jak si s touto situací poradit? Máme dvě možnosti: Buď data ponecháme v původní podobě a analýzu rozptylu provedeme v neparametrické podobě (postup viz následující př. 7.3), ale toto testování má obecně slabší sílu testu, anebo z každé třídy vyloučíme odlehlá pozorování a budeme pracovat s upravenou verzí dat (Třetí možností by bylo „opravit“ odlehlé hodnoty tak, aby již nebyly odlehlé, např. v případě přepisu mohl vzniknout překlep atd.).

Identifikujme odlehlá pozorování v každé třídě a tuto hodnotu vyjměme z dalšího zpracování:

Menu Desribe/NumericData/Outlier Identification

Statgraphics identifikoval v souboru 3 odlehlá pozorování:první hodnotu (49) u indického čaje a první (46) a třetí (67) hodnotu u vietnamského čaje. Z tabulky odstraníme tyto 3 hodnoty, neboť nepředpokládáme, že tato úprava podstatně ovlivní výsledek.

Takže pracujeme s novou tabulkou dat:

Čínský	**indický**	**Japonský**	**gruzínský**	**vietnamský**
60	-	51	55	-
68	72	68	71	65
88	77	70	78	-
79	65	75	71	58
72	68	57	65	60
65	66	62	66	56
56	74	60	80	58
76	80	65	72	55
77		58		60
83		63		
		50		
		52		

Ověření předpokladů:

a) Otestujeme normalitu dat v každé třídě:

Nulová a alternativní hypotéza pro ověření normality dat:

H_0 … hodnoty proměnné „čínský“ (čaj) lze považovat za náhodný výběr z normálního rozdělení
H_A … hodnoty proměnné „čínský“ (čaj) nelze považovat za náhodný výběr z normálního rozdělení

Menu Desribe/Distributions/Distribution Fitting (Uncensored Data)

Zvolíme Data … čínský

a postupně provedeme ověření normality i pro zbylé čtyři druhy čajů.

P-value všech testů vyšlo větší než 0,05, tedy nezamítáme nulovou hypotézu, neboli výběr ve všech třídách lze považovat za náhodný výběr z normálního rozdělení.

b) Dále ověříme rovnost rozptylů, resp.směrodatných odchylek (homoskedasticitu):

Testujeme hypotézu $\mathrm{H}_0 : \sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5$

oproti alternativě H_A : neplatí H_0,

Ve *Statgraphicsu* použijeme proceduru, která se týká již samotné analýzy rozptylu ANOVA:

Menu Compare/MultipleSamples/ Multiple-Sample Comparison

Input: Multiple Data Colums (jednotlivé třídy jsou zadány ve sloupcích)

Tabular Options (žlutá ikona) … **Variance Check**

Jelikož nebyla porušena podmínka normality, použijeme k vyhodnocení Bartlettův test. P-value Bartletova testu vyšlo větší než 0,05, tedy nezamítáme nulovou hypotézu, neboli nebyla porušena podmínka homoskedasticity.

c) Ze zadání je zřejmé, že jednotlivé náhodné výběry jsou nezávislé.

Po provedené korekci jsou opět splněny všechny předpoklady, můžeme tedy přistoupit k parametrické podobě analýzy rozptylu ANOVA:
Testujeme hypotézu:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

oproti alternativě

H_A : neplatí H_0,

Kde μ_1 je střední hodnota bodového hodnocení čínského čaje,
μ_2 je střední hodnota bodového hodnocení indického čaje,
μ_3 je střední hodnota bodového hodnocení japonského čaje,
μ_4 je střední hodnota bodového hodnocení gruzínského čaje,
μ_5 je střední hodnota bodového hodnocení vietnamského čaje.

Budeme srovnávat uvedené druhy čajů podle bodového hodnocení zákazníků. Použijeme proceduru **Compare** (srovnávat). Srovnáváme více než dvě třídy, tedy **Multiple-Sample Comparison:**

Menu Compare/MultipleSamples/ Multiple-Sample Comparison

Input: Multiple Data Colums (jednotlivé třídy jsou zadány ve sloupcích)

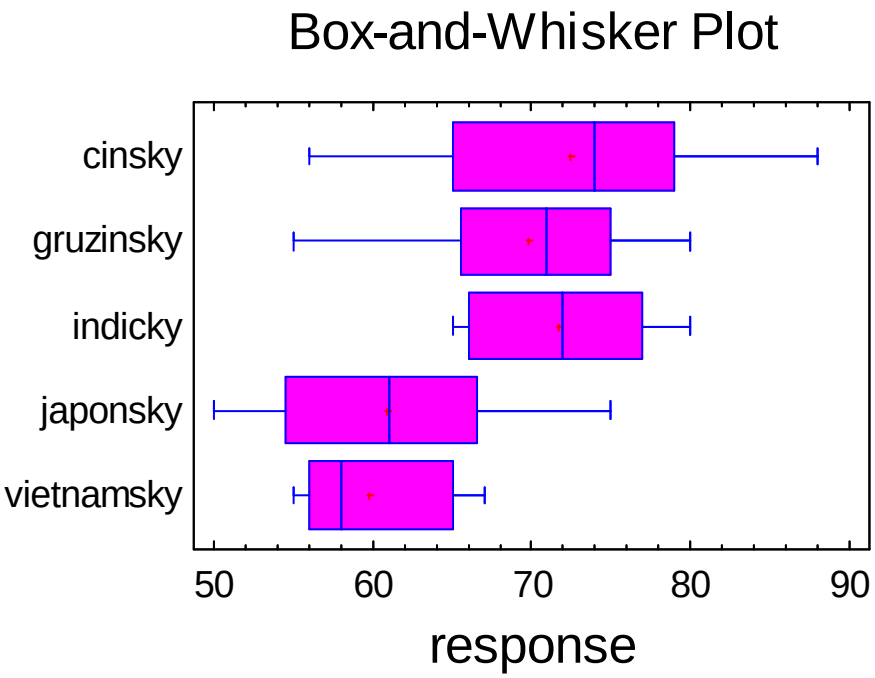
Tabulka ANOVA:

ANOVA Table

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Between groups	*1304,44*	*4*	*326,111*	*5,40*	*0,0015*
Within groups	*2357,1*	*39*	*60,4385*		
Total (Corr.)	*3661,55*	*43*			

Grafický výstup:



V tabulce ANOVA je uvedeno p-value, které potřebujeme k vyhodnocení hypotézy H_0:

p-value F-testu je menší než 0,05, tedy existuje statisticky významný rozdíl mezi středními hodnotami jednotlivých tříd (***Since the P-value of the F-test is less than 0,05, there is a statistically significant difference between the means of the 5 variables at the 95,0% confidence level***).

Jelikož existuje statisticky významný rozdíl mezi středními hodnotami, budeme dále pokračovat v tzv. **Analýze Post-Hoc** (chceme specifikovat bližší určení rozdílů mezi jednotlivými třídami):

Klikneme na „žlutou ikonu“ a aktivujeme **Multiple Range Tests:**

Multiple Range Tests

Method: 95,0 percent LSD

	Count	Mean	Homogeneous Groups
vietnamsky	7	59,8571	X
japonsky	12	60,9167	X
gruzinsky	8	69,75	X
indicky	7	71,7143	X
cinsky	10	72,4	X

Contrast	Difference	+/- Limits
cinsky - gruzinsky	2,65	7,45897
cinsky - indicky	0,685714	7,74931
cinsky - japonsky	*11,4833	6,73299
cinsky - vietnamsky	*12,5429	7,74931
gruzinsky - indicky	-1,96429	8,1384
gruzinsky - japonsky	*8,83333	7,17739
gruzinsky - vietnamsky	*9,89286	8,1384
indicky - japonsky	*10,7976	7,47867
indicky - vietnamsky	*11,8571	8,4053
japonsky - vietnamsky	1,05952	7,47867

** denotes a statistically significant difference.*

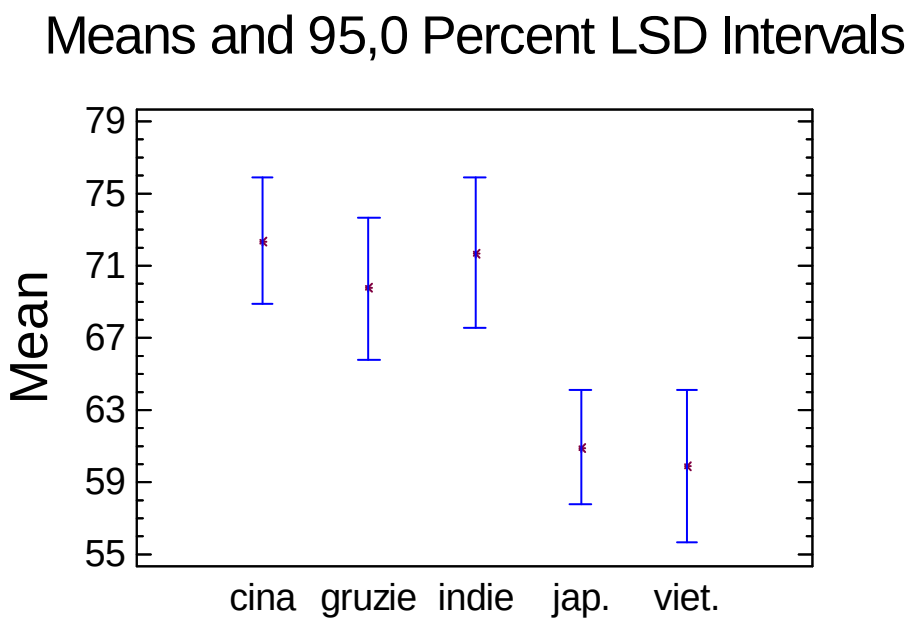
indicky - japonsky	*7,95833	7,75854
indicky - vietnamsky	*10,5417	8,2596
japonsky - vietnamsky	2,58333	7,49546

** denotes a statistically significant difference.*

(***** označuje třídy, mezi kterými je statisticky významný rozdíl)

Dále klikněme na třetí ikonu (**Graphical Options**) a aktivujme **Means Plot**

(dostáváme grafický výstup - srovnávání 95 % intervalů spolehlivosti pro jednotlivé střední hodnoty):



V grafickém výstupu přehledněji vidíme znázornění předchozího slovního výstupu:

Třídy čínský, gruzínský a indický můžeme sloučit v jednu skupinu (vzájemně se příliš mezi sebou neliší) a japonský s vietnamským, kdežto mezi sebou tyto dvě podskupiny vykazují statisticky významný rozdíl mezi středními hodnotami.

Shrnutí: p-value vyšlo 0,0015, což je méně než 0,05, tedy H_0 se zamítá. Existuje statisticky významný rozdíl mezi středními hodnotami bodového hodnocení jednotlivých druhů čajů.

Závěr: Na základě uvedených údajů byla potvrzena závislost mezi bodového hodnocení čajů a druhem čaje. Lze tvrdit, že všechny skupiny čajů nejsou zákazníky hodnoceny stejně. Nejvíce jsou oblíbeny čaje čínský, gruzínský a indický (jejich oblíbenost je zhruba stejná), kdežto méně oblíbeny jsou japonský a vietnamský čaj (oblíbenost japonského a vietnamského čaje je také zhruba stejná).

Poznámka: Post Hoc analýzu jsme mohli vyhodnotit i LSD testem s Bonifferoniho korekcí, resp. obecně platným Scheffého testem.

Příklad 7.3. *Otestujte na souboru* ***Cardata ,*** *zda výkon automobilu je závislý na místu výroby automobilu.*

Řešení: Otevřeme soubor **Cardata** z datového balíku, který patří k softwaru *Statgraphics*, tj. hledáme v adresáři **Statgra** na lokálním disku:

Menu File/Open/OpenDataFile/TestData/Cardata.sf.

Nejprve opět ověříme předpoklady:

a) Nejprve otestujeme normalitu dat v každé třídě:

Nulová a alternativní hypotéza pro ověření normality dat:

H_0 … hodnoty proměnné „horsepower|Origin=1“ (výkon automobilů vyrobených v Americe) lze považovat za náhodný výběr z normálního rozdělení

H_A … hodnoty proměnné „horsepower|Origin=1“ (výkon automobilů vyrobených v Americe) nelze považovat za náhodný výběr z normálního rozdělení

Menu Desribe/Distributions/Distribution Fitting (Uncensored Data)

Zvolíme Data … horsepower

a postupně provedeme testování pro všechny tři třídy

Select … Origin=1 (=2, =3)

Zjistili jsme, že existuje alespoň jedna třída, v níž data nelze považovat za náhodný výběr z normálního rozdělení, tedy při samotné ANOVě zvolíme její neparametrickou podobu.

b) Dále otestujeme rovnost rozptylů, resp. směrodatných odchylek (homoskedasticitu):

Testujeme hypotézu

$$\mathrm{H}_0 : \sigma_1 = \sigma_2 = \sigma_3$$

oproti alternativě

H_A : neplatí H_0,

Ve *Statgraphicsu* použijeme proceduru, která se týká již samotné analýzy rozptylu ANOVA:

Jelikož data máme zadána v jiném formátu, než v předchozích příkladech, postupujeme následovně:

Menu Compare/AnalysisOfVariance/One Way ANOVA

Dependent Variable … horsepower (výkon motoru)

Factor ... origin (země původu – 1… Amerika, 2… Evropa, 3… Japonsko)

Poznámka: Faktor musí být numerická proměnná !!! (i když se de facto jedná o kategoriální. Pokud je proměnná zadána jako kategoriální, musí se převést na odpovídající numerickou proměnnou.)

Testování rovnosti rozptylů neboli homoskedasticitu provedeme opět pomocí procedury

Tabular Options (žlutá ikona) … **Variance Check**

Jelikož nebyla splněna podmínka normality, použijeme k otestování rovnosti rozptylů Leveneův test. P-value Leveneova testu vyšlo větší než 0,05, tedy nezamítáme nulovou hypotézu, neboli nebyla porušena podmínka homoskedasticity. Pozn. Hartleyův ani Cochranův test ani v jednom z uvedených příkladů nemůžeme použít, neboť se nejedná o vyvážené třídění.

c) Ze zadání je zřejmé, že jednotlivé náhodné výběry jsou nezávislé.

Při testování normality jsme zjistili, že nelze ve všech třídách považovat data za náhodný výběr z normálního rozdělení. Použijeme tudíž neparametrickou podobu analýzy ANOVA tzv. Kruskal - Wallisův test.

Testujeme hypotézu oproti alternativě

$H_0 : x^1_{0,5} = x^2_{0,5} = x^3_{0,5}$ H_A : neplatí H_0,

kde $x^1_{0,5}$ je medián výkonu automobilů vyrobených v Americe,

$x^2_{0,5}$ je medián výkonu automobilů vyrobených v Evropě,

$x^3_{0,5}$ je medián výkonu automobilů vyrobených v Japonsku.

Budeme tedy srovnávat výkon motoru automobilů podle zemí výroby automobilu.

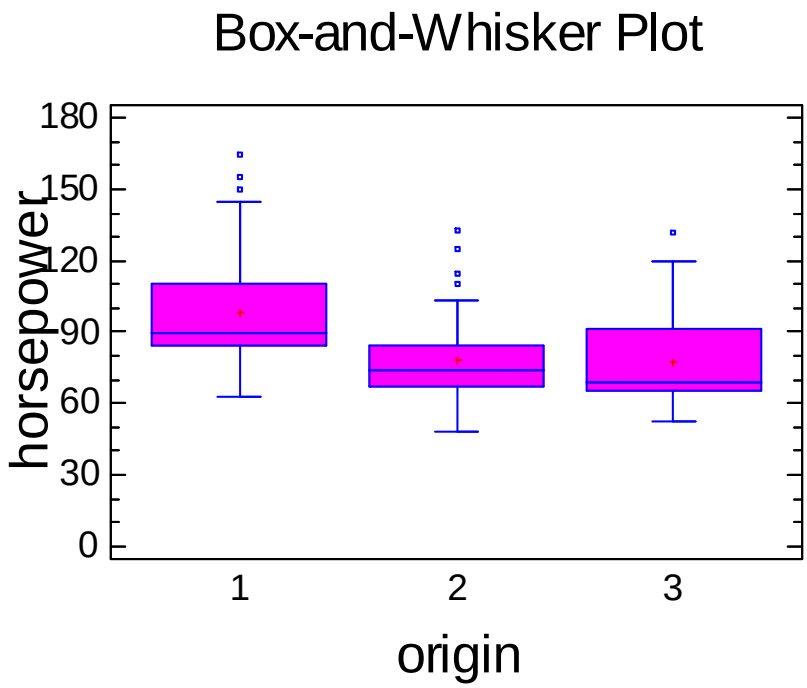
Použijeme již otevřenou proceduru **Compare** (srovnávat).

Menu Compare/AnalysisOfVariance/One Way ANOVA

Dependent Variable … horsepower (výkon motoru)

Factor ... origin (země původu – 1… Amerika, 2… Evropa, 3… Japonsko)

Jako grafický výstup dostaneme opět vícenásobný krabicový graf (překlopili jsme jej do vertikálního tvaru):



Textový výstup je výstupem Kruskal – Wallisova testu:

Kruskal – Wallis Test … Tabular Options (žlutá ikona)

Kruskal-Wallis Test for horsepower by origin

origin	*Sample Size*	*Average Rank*
1	*83*	*93,6265*
2	*24*	*56,8958*
3	*44*	*53,1705*

Test statistic = 30,0951 P-Value = 2,91698E-7

p-value je menší než 0,05, tedy H_0 se zamítá. Existuje statisticky významný rozdíl mezi mediány (***Since the P-value is less than 0,05, there is a statistically significant difference amongst the medians at the 95,0% confidence level***) výkonu motorů automobilů vzhledem k místu výroby.

Jelikož existuje statisticky významný rozdíl mezi mediány, budeme dále pokračovat v tzv. **Analýze Post-Hoc** (chceme specifikovat bližší určení rozdílů mezi jednotlivými třídami) a zvolíme v tomto případě neparametrickou podoby tohoto testu – Tukeyho metodu:

Klikneme na „žlutou ikonu“ a aktivujeme **Multiple Range Tests** a pravým tlačítkem myši zvolíme Pane Options **Tukey HSD:**

Dostaneme textový výstup:

Multiple Range Tests for horsepower by origin

Method: 95,0 percent Tukey HSD

origin	*Count*	*Mean*	*Homogeneous Groups*
3	*44*	*77,1591*	*X*
2	*24*	*78,4583*	*X*
1	*83*	*98,3253*	*X*

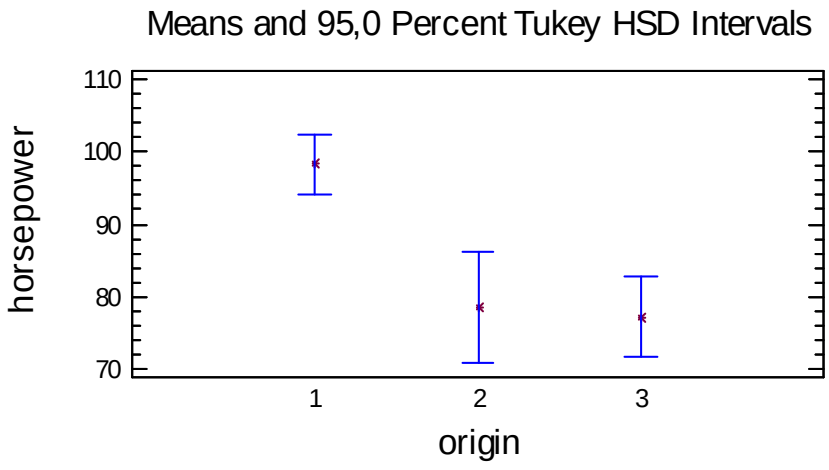
Contrast	*Difference*	*+/- Limits*
1 - 2	**19,867*	*12,2015*
1 - 3	**21,1662*	*9,81754*
2 - 3	*1,29924*	*13,3595*

**** denotes a statistically significant difference.***

a grafický výstup:

Graphical Options … **Means Plot**

Klikneme pravým tlačítkem myši na zobrazený grafický výstup, zvolíme **Pane Options**, a upravíme na **Tukey HSD intervals**):



Ze slovního popisu i obrázku vidíme, že kategorie 1-2, 1-3 se liší, kdežto 2-3 nevykazují odlišnost.

Závěr: Na základě uvedených údajů byla potvrzena závislost mezi výkonem motoru sledovaných typů automobilů a zemí výroby. Výkon motoru je nejvyšší u Amerických automobilů, nižší je u Evropských a Japonských vozů, jejichž výkon můžeme považovat za srovnatelný.

Poznámka 1: Jelikož jsme stejně použili k výpočtu Kruskal-Wallisův test, nemusíme se tolik znepokojovat skutečností, že v souboru jsou odlehlá pozorování. Pokud bychom je vyňali, stejně bychom „nespravili“ normalitu, tudíž by bylo použití Kruskal-Wallisova testu opět nezbytné. Zkuste si sami provést tuto úpravu a následné vyhodnocení a výsledky porovnejte.

Poznámka 2: Post Hoc analýzu jsme mohli vyhodnotit i obecně platným Scheffého testem.

8. Jednoduchá lineární regrese

Příklad 8.1. *Následující data byla získána v autobazaru ABC a týkají se stáří (v letech) a ceny (v tisících Kč) ojetých aut značky Škoda (Fabia, Felicia):*

Auto	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
Stáří	5	4	6	5	5	5	6	6	2	7	7
Cena	85	103	70	82	89	98	66	95	169	70	48

Testujeme závislost ceny na stáří ojetého auta.

a) Najděte rovnici vyrovnávací přímky vyjadřující závislost ceny na stáří ojetého auta a znázorněte ji .
b) Proveďte dílčí t-testy pro testování významnosti parametrů vyrovnávací přímky.
c) Verifikujte zvolený lineární model.
d) Ověřte kvalitu modelu.
e) Pomocí nalezené závislosti zkuste predikovat (předpovědět) cenu 3 a 4 roky starého automobilu, včetně intervalu spolehlivosti a intervalu predikce.
f) Lze pomocí nalezené regresní funkce predikovat cenu 11 let starého automobilu?

Řešení: Do *Statgraphicsu* zadáme data (samozřejmě do sloupců!). Proměnné označme „Stáří“ a „Cena“

Hledáme závislost mezi dvěma proměnnými, tedy volíme proceduru **Relate** (relace=vztah):

Menu Relate/SimpleRegression

Zvolíme:

Y … Cena

X … Stáří

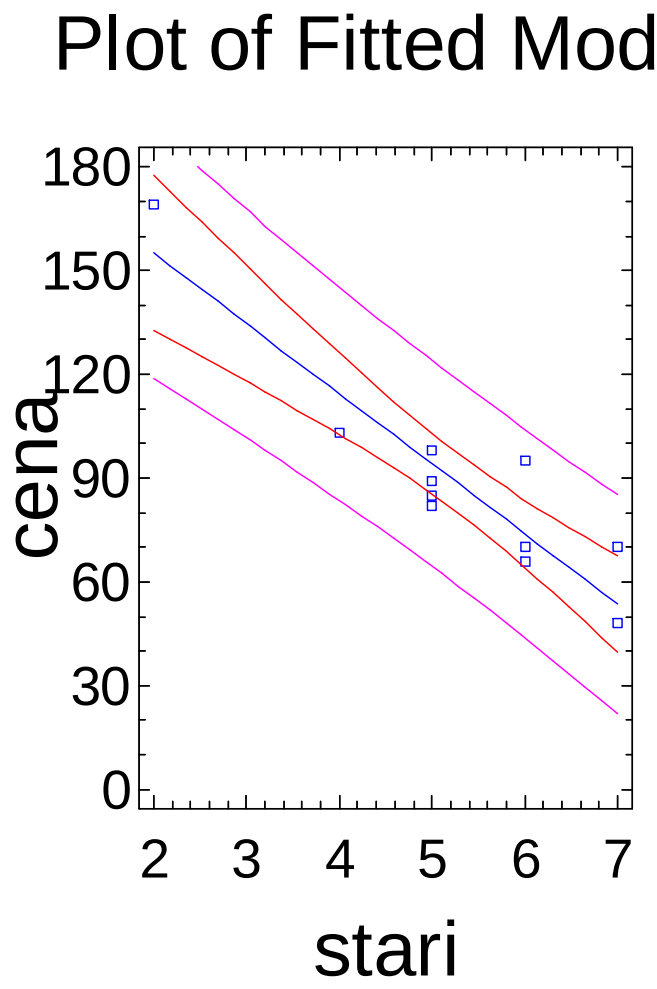
Zobrazí se nám slovní komentář a grafický výstup:

a) Rovnici vyrovnávací přímky najdeme v levém dolním okně **The StatAdvisor**:

The StatAdvisor
***---------- ***
The output shows the results of fitting a linear model to describe the relationship between stari and cena. The equation of the fitted model is

cena = 195,468 - 20,2613*stari.

Grafické znázornění zkoumané závislosti vidíme v pravém okně:



b) Dílčí t-testy pro testování významnosti parametrů vyrovnávací přímky

Skutečná regresní přímka: $\boldsymbol{Y} = \alpha + \beta * \boldsymbol{X}$.

Výběrová regresní přímka (neboli vyrovnávací přímka): $\boldsymbol{Y} = a + b * \boldsymbol{X}$.

Při dílčích t-testech testujeme parametry α, β skutečné regresní přímky

1) $H_0 : \alpha = 0$ 2) $H_0 : \beta = 0$

$H_A : \alpha \neq 0$ $H_A : \beta \neq 0$

Odhady *a, b* parametrů α, β regresní přímky nalezneme ve výstupu ze *Stagraphicsu* (jejich bodové odhady jsme již vlastně uvedli v rovnici vyrovnávací přímky) a dále zde nalezneme p-value výše uvedených testů (v 1. řádku pro 1. test, tedy pro parametr α a v 2. řádku pro 2. test, tedy pro parametr β):

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	*195,468*	*15,2403*	*12,8257*	*0,0000*
Slope	*-20,2613*	*2,79951*	*-7,23743*	*0,0000*

Obě hodnoty p-value jsou menší než 0,05, tedy v obou případech se H_0 zamítá. Neboli ani jeden z parametrů *a, b* nelze z modelu vypustit.

c) Verifikace (ověření správnosti použití) zvoleného lineárního modelu.

Potvrzení závislosti veličin X a Y:

Abychom mohli zkoumat závislost mezi veličinami X a Y musí být zřejmé, že tato závislost existuje. Tedy, že kolísání hodnot veličiny Y není dáno pouze náhodně, ale že y-ové hodnoty jsou nějakým způsobem provázány s x-ovými hodnotami.

Tabulka ANOVA, charakterizující závislost nebo nezávislost mezi zvolenými proměnnými:

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	*8285,01*	*1*	*8285,01*	*52,38*	*0,0000*
Residual	*1423,53*	*9*	*158,17*		
Total (Corr.)	*9708,55*	*10*			

Pokud by v tabulce ANOVA vyšlo p-value vyšší než 0,05, nemělo by smysl vůbec žádnou regresi provádět, neboť proměnné X a Y by byly nezávislé. Regresní křivka specifikuje typ závislosti mezi veličinami. Nemá tedy smysl ji dělat pro nezávislé veličiny.

p-value < 0,05, tedy mezi veličinami X a Y existuje závislost.

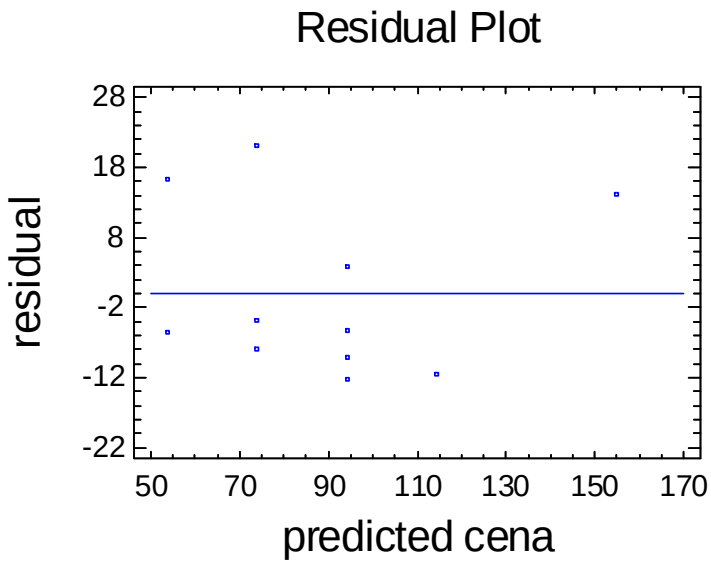
Ověření normality reziduí

Nejprve vytvoříme novou proměnnou, do které uložíme rezidua (chyby nahrazení skutečné y-ové hodnoty hodnotou na vyrovnávací přímce), tj. hodnoty

$z_i = y_i - (a + b \cdot x_i)$.

Graphical Options : Rezidual versus Predicted

Pravým tlačítkem myši (**Pane Options**) změníme „Studentized residuals“ na „Residuals“. Dostaneme následující grafický výstup:



Nyní aktivujme proceduru

Save Results Options (4. ikona, s obrázkem diskety)

a zvolme Save Residuals.

Nová proměnná se nám zobrazí v původní tabulce hodnot (nikoli v grafickém výstupu). Standardním způsobem budeme testovat normalitu reziduí:

Menu Describe/Distibutions/DistributionsFitting(Uncensored Data)

Zvolíme: Residuals

Pomocí p-value 0,034 testu dobré shody nezamítneme na 1 % hladině významnosti nulovou hypotézu normality reziduí (pokud vyjde p-value ležící v „hraničním pásmu“, tj. mezi 0,01 a 0,05 doporučuje se opakovat náhodný výběr a test provést znova. Jelikož zde tímto testem nerozhodujeme samotný výsledek testu, ale pouze testujeme předpoklady, můžeme si dovolit být mírní a normalitu „pustit“).

Testování nulovosti střední hodnoty reziduí

Dále testujeme nulovost střední hodnoty reziduí, tj. hypotézu:

$H_0 : \mu_z = 0$ oproti alternativě

$H_A : \mu_z \neq 0$.

Menu Describe/Numeric Data/One-Variable Analysis

Zvolíme Residuals

Ve výstupu aktivujeme proceduru

Tabular Options: Hypothesis Tests

a pomocí p-value 0,99 t-testu nezamítneme nulovou hypotézu.

d) Ověření kvality modelu

Index determinace (***R-squared***) vyšel 85 % :

R-squared = 85,3373 percent.

Koeficient ***R-squared*** udává jak těsná je závislost mezi proměnnými X, Y, resp. jak přesné je nahrazení závislosti zvolenou regresní křivkou (někdy závislost dvou proměnných nejlépe popisuje přímka, jindy parabola či exponenciála, nebo jiná funkce). Čím je ***R-squared*** blíže jedničce, o tím přesnější aproximaci se jedná. Pokud je ***R-squared*** blízko nule, je zvolená funkce nevhodná pro aproximaci dané závislosti anebo mezi uvažovanými veličinami vůbec neexistuje těsná závislost.

Koeficient R-Squared = 85 %, který je dosti vysoký („blízko“ 100%), vyjadřuje skutečnost, že typ regresní závislosti (v našem lineární) je vhodně zvolen.

e) Cena 3, resp 4 roky starého automobilu (predikce hodnoty y pro x ležící v intervalu zadaných hodnot):

Bodový odhad ceny 3 roky starého automobilu:

cena = 195,468 - 20,2613*(stari=3)=195,468 - 20,2613*3=134,69 (tis Kč).

Bodový odhad ceny 4 roky starého automobilu:

cena = 195,468 - 20,2613*(stari=4)=195,468 - 20,2613*4=114,43 (tis Kč).

Tedy 3 roky starý automobil značky Škoda bude v autobazaru ABC pravděpodobně stát 134 690 Kč a 4 roky starý automobil bude stát 114 430.

Statgraphics nám kromě bodového odhadu umožňuje nalézt i intervalový odhad

pro střední hodnotu Y při dané hladině x, tzv interval spolehlivosti (**confidence limits**) a

pro individuální hodnotu Y při dané hodnotě x, tzv. interval predikce (**prediction limits**).

Najděme tyto intervaly pro 3 a 4 roky staré auto:

Tabular Options (žlutá ikona): **Forecasts**

Pravým tlačítkem myši (**Pane Options**) upravíme vstupní hodnoty na 3 a 4:

Predicted Values

		95,00% Prediction Limits		95,00% Confidence Limits	
X	**Predicted Y**	**Lower**	**Upper**	**Lower**	**Upper**
3,0	**134,685**	**101,667**	**167,702**	**117,929**	**151,44**
4,0	**114,423**	**83,6344**	**145,212**	**102,653**	**126,194**

Tedy pro x = 3 roky staré auto jsme dostali:

95 % intervalem spolehlivosti (**confidence limits**) pro střední hodnotu Y při dané hladině x=3 je interval 117 až 151 tisíc (neboli 95 % středních hodnot cen 3 roky starých aut leží v tomto intervalu),

95 % intervalem predikce (**prediction limits**) pro individuální hodnotu Y při dané hodnotě x=3 je interval 101 až 167 tisíc tisíc (neboli 95 % cen 3 roky starých aut leží v tomto intervalu).

Pro x = 4 roky staré auto:

95 % intervalem spolehlivosti (**confidence limits**) pro střední hodnotu Y při dané hladině x=4 je interval 102 až 126 tisíc,

95 % intervalem predikce (**prediction limits**) pro individuální hodnotu Y při dané hodnotě x=4 je interval 83 až 145 tisíc.

Interval predikce pro individuální hodnotu je vždy širší než interval spolehlivosti pro střední hodnotu, neboť chyba v odhadu střední hodnoty např. ceny všech 3 roky starých aut je způsobena tím, že teoretická regresní přímka je odhadnuta pomocí výběrové regresní přímky. Zatímco chyba v predikci ceny náhodně vybraného 3 roky starého auta je způsobena již zmíněnou chybou v odhadu střední hodnoty a ještě navíc variabilitou mezi cenami všech 3 roky starých aut.

V grafickém výstupu je červeně znázorněn pás spolehlivosti a fialově pás predikce pro spojitě se měnící x.

f) Cena 11 let starého automobilu (extrapolace):

Kdybychom postupovali obdobným způsobem jako v části e), dostali bychom nemyslný výsledek:

Cena 11 let starého automobilu:

cena = 195,468 - 20,2613*stari.=195,468 - 20,2613*11= - 27,99 (tis Kč).

Neboli zmíněný model by nám říkal, že nám ještě někdo zaplatí 27 990, abychom si vzali jeho 11 let staré auto.

Predikci podle nalezené regresní závislosti lze provádět pouze uvnitř intervalu zadaných hodnot, tedy v našem případě pouze pro stáří 2 až 7 let. Pro auta starší než 7 let a „mladší“ než 2 roky predikci pomocí vytvořeného modelu provádět nelze! Vztah mezi stářím a cenou auta můžeme považovat za přibližně lineární pouze v intervalu 2 až 7 let. Mimo tento interval již nemusí existovat buď vůbec žádná závislost mezi stářím a cenou auta (cenu třeba vice než stáří ovlivňuje stav opotřebovanosti automobilu, najeté kilometry atd.) anebo závislost může být např. kvadratická, resp. exponenciální, což již není možné ze zadaných dat určit.

Závěr: Pomocí vytvořeného modelu není vhodné odhadovat cenu 11 let starého automobilu.

Příklad 8.2. *Testujeme závislost mezi veličinami* ***horsepower*** *(výkon automobilu) a* ***mpg*** *(počet ujetých mil na galon pohonných hmot) zadanými v souboru* ***CarData*** *.*

a) Najděte rovnici vyrovnávací přímky vyjadřující závislost výkonu automobilu na počtu ujetých mil na galon pohonných hmot .
b) Proveďte dílčí t-testy pro testování významnosti parametrů vyrovnávací přímky.
c) Verifikujte zvolený lineární model.
d) Ověřte kvalitu modelu.

Řešení*:* Otevřeme soubor **CarData** z datového balíku, který patří k softwaru *Statgraphics*, tj. hledáme v adresáři **Statgra** na lokálním disku:

Menu File/Open/OpenDataFile/TestData/CarData.sf.

Hledáme závislost mezi dvěma proměnnými, tedy volíme proceduru **Relate:**
Menu Relate/SimpleRegression

Zvolíme:

Y … mpg

X … horsepower

Zobrazí se nám slovní komentář a grafický výstup:

Regression Analysis - Linear model: Y = a + b*X

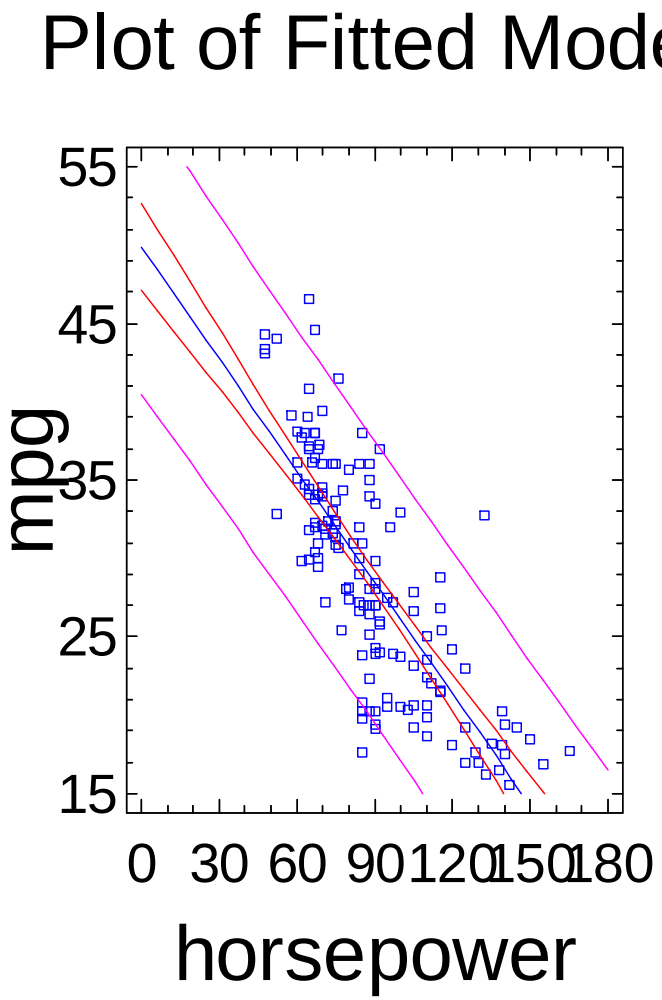
Dependent variable: mpg (závislá proměnná … Y)
Independent variable: horsepower (nezávislá proměnná … X)
----

a) Rovnice vyrovnávací přímky a grafické znázornění:

The StatAdvisor

The output shows the results of fitting a linear model to describe the relationship between mpg and horsepower. The equation of the fitted model is

mpg = 49,8706 - 0,237707*horsepower



b) Testování významnosti parametrů vyrovnávací přímky:

Testujeme hypotézy

1) $H_0 : \alpha = 0$ 2) $H_0 : \beta = 0$

$H_A : \alpha \neq 0$ $H_A : \beta \neq 0$

Parameter	*Estimate*	*Standard Error*	*T Statistic*	*P-Value*
Intercept	*49,8706*	*1,40312*	*35,5426*	*0,0000*
Slope	*-0,237707*	*0,0152283*	*-15,6096*	*0,0000*

Obě hodnoty p-value jsou menší než 0,05, tedy v obou případech se H_0 zamítá. Neboli ani jeden z parametrů *a, b* nelze z modelu vypustit (oba jsou statisticky významné).

Kontrolní otázka: bylo by možné z modelu vypustit koeficient *b* (tedy mohlo by vyjít, že $H_0 : \beta = 0$ nelze zamítnout)?

c) Verifikace modelu.

Potvrzení závislosti X a Y:

Tabulka ANOVA, charakterizující závislost nebo nezávislost mezi zvolenými proměnnými:

Analysis of Variance

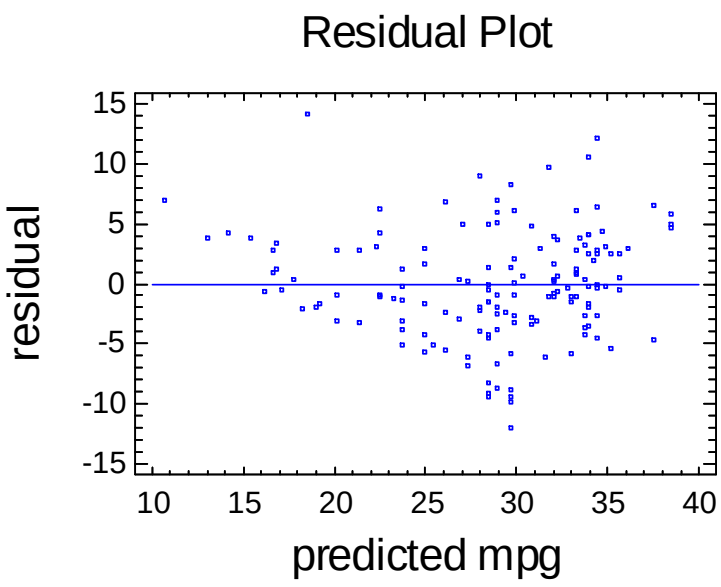
Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	***5030,95***	***1***	***5030,95***	***243,66***	***0,0000***
Residual	***3055,83***	***148***	***20,6475***		
Total (Corr.)	***8086,77***	***149***			

Hodnotou p-value 0,0000 byla potvrzena závislost mezi zkoumanými veličinami.

Nyní vygenerujeme hodnoty reziduí a otestujeme jejich normalitu:

Graphical Options : Rezidual versus Predicted

Pravým tlačítkem myši (**Pane Options**) změníme „Studentized residuals“ na „Residuals“. Dostaneme následující grafický výstup:



Nyní aktivujme proceduru

Save Results Options (4. ikona, s obrázkem diskety)

a zvolme Save Residuals.

Nová proměnná se nám zobrazí v původní tabulce hodnot (nikoli v grafickém výstupu). Standardním způsobem budeme testovat normalitu reziduí:

Menu Describe/Distibutions/DistributionsFitting(Uncensored Data)

Zvolíme: Residuals

Pomocí p-value 0,74 testu dobré shody nezamítáme na 5 % hladině významnosti nulovou hypotézu normality reziduí.

Testování nulovosti střední hodnoty reziduí

Dále testujeme nulovost střední hodnoty reziduí, tj. hypotézu:

$H_0 : \mu_z = 0$ oproti alternativě

$H_A : \mu_z \neq 0$.

Menu Describe/Numeric Data/One-Variable Analysis

Zvolíme Residuals

Ve výstupu aktivujeme proceduru

Tabular Options: Hypothesis Tests

a pomocí p-value 0,99 t-testu nezamítneme nulovou hypotézu.

d) Ověření kvality modelu:

R-squared = 62,212 percent

Koeficient R-Squared = 62 % vyjadřuje skutečnost, že typ regresní závislosti (v našem lineární) je ještě celkem dobře zvolen. V tomto případě bývá obvykle vhodné podívat se na indexy determinace jiných modelů. Mohou být vyšší, ale nemusí. Pokud je index determinace příliš vzdálený od 100 %, může to být způsobeno buď nevhodně zvoleným modelem anebo slabou závislostí mezi zkoumanými veličinami.

Jak zvolit jiný model v závislosti na hodnotě indexu determinace si uvedeme v následujícím příkladě (Vždy je nutno provést tuto volbu ručně, žádný statistický software tuto volbu za vás neprovede).

Příklad 8.3. *Najděte typ regresní křivky nejlépe aproximující závislost mezi spotřebou a objemem motoru (srovnejte hodnoty koeficientů R-Squared pro možné volby typů regresních křivek a tím zdůvodněte, proč je pro popis závislosti mezi spotřebou a objemem motoru nejvhodnější logaritmický model). Použijte data uvedená v následující tabulce:*

Automobil	objem motoru [cm^3]	spotřeba [l/100km]
Alfa Romeo 147 2,0 Selespd	1970	8,9
Audi A3 2,0 FSI	1984	6,9
Mini Cooper	1598	6,7
Škoda Fabia RS	1896	5,4
BMW 545i	4398	10,9
Volvo S80 T6	2922	11,3
Ford Mondeo ST 220	2967	10,4
Mercedes-Benz C 320	3199	10,9
Volkswagen Golf 2,0 FSI	1984	7,6
Fiat Seicento Sporting	1108	6
Opel Corsa GSI	1796	7,6
Honda Civic Type-R	1998	8,9
Subaru Forester XT	1994	9,8
Jaguar XK8	4196	11,3
Saab 9-5 Aero	2290	9,1

Řešení:

Menu Relate/SimpleRegression

Y … spotřeba
X … objem motoru

1) Potvrzení závislosti spotřeby na objemu motoru:

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	*39,3526*	*1*	*39,3526*	*32,11*	*0,0001*
Residual	*15,9314*	*13*	*1,22549*		
Total (Corr.)	*55,284*	*14*			

Hodnotou p-value 0,0001 byla potvrzena závislost mezi spotřebou a objemem motoru.

2) Volba vhodného modelu

Porovnáme koeficienty R-Squared u možných modelů pro popis regresních závislosti mezi objemem motoru a spotřebou daného typu automobilu (žlutá ikona – **Comparison of Alternative Models**):

Comparison of Alternative Models

Model	*Correlation*	*R-Squared*
Logarithmic-X	***0,8437***	***71,18 %***
Square root-X	***0,8324***	***69,28 %***
Reciprocal-X	***-0,8220***	***67,58 %***
Multiplicative	***0,8151***	***66,44 %***
Linear	***0,8106***	***65,70 %***
S-curve	***-0,8049***	***64,79 %***
Square root-Y	***0,7936***	***62,98 %***
Double reciprocal	***0,7775***	***60,45 %***
Exponential	***0,7744***	***59,97 %***
Reciprocal-Y	***-0,7298***	***53,27 %***

Podle porovnání hodnot ***R-Squared*** je patrné, že nejlepší model je ***Logarithmic-X,*** protože ***R-Squared*** se u něj nejvíce blíží 100%.

Změníme nastavení z lineární závislosti na logaritmickou:

Analysis Options … Logarithmic-X

Zobrazí se nám slovní komentář a grafický výstup:

Regression Analysis - Logarithmic-X model: Y = a + b*ln(X)

Dependent variable: spotreba
Independent variable: objem

Parameter	*Estimate*	*Standard Error*	*T Statistic*	*P-Value*
Intercept	***-26,5512***	***6,24142***	***-4,25404***	***0,0009***
Slope	***4,57224***	***0,806859***	***5,66672***	***0,0001***

Z výše uvedených výstupů vidíme, že závislost budeme popisovat logaritmickou funkcí, jejíž obecný tvar je ***Y = a + b*ln(X)***.

Výsledky dílčích t-testů pro parametry *a, b* ukázaly, že ani jeden nelze v modelu vynechat (oba jsou statisticky významné).

Rovnici hledané závislosti můžeme tedy vyjádřit ve tvaru:

spotreba = -26,5512 + 4,57224*ln(objem).

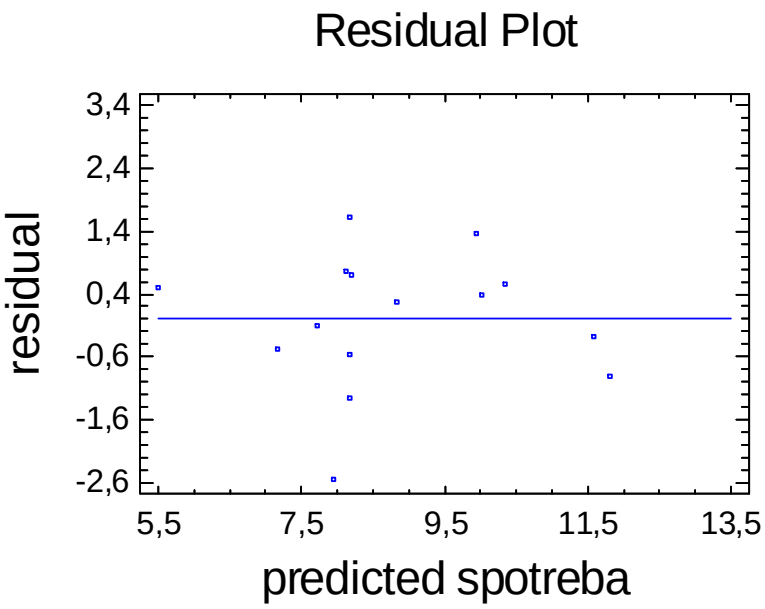
3) Dále budeme verifikovat zvolený logaritmický model

Otestujeme normalitu reziduí a nulovost střední hodnoty reziduí (vše provádíme již pro logaritmickou závislost!):

Vygenerujeme hodnoty reziduí a otestujeme jejich normalitu:

Graphical Options : Rezidual versus Predicted

Pravým tlačítkem myši (**Pane Options**) změníme „Studentized residuals“ na „Residuals“. Dostaneme následující grafický výstup:



Nyní aktivujme proceduru

Save Results Options (4. ikona, s obrázkem diskety)

a zvolme Save Residuals.

Nová proměnná se nám zobrazí v původní tabulce hodnot (nikoli v grafickém výstupu). Standardním způsobem budeme testovat normalitu reziduí:

Menu Describe/Distibutions/DistributionsFitting(Uncensored Data)

Zvolíme: Residuals

Pomocí p-value 0,10 testu dobré shody nezamítáme na 5 % hladině významnosti nulovou hypotézu normality reziduí.

Testování nulovosti střední hodnoty reziduí

Dále testujeme nulovost střední hodnoty reziduí, tj. hypotézu:

$H_0 : \mu_z = 0$ oproti alternativě $H_A : \mu_z \neq 0$.

Menu Describe/Numeric Data/One-Variable Analysis

Zvolíme Residuals

Ve výstupu aktivujeme proceduru

Tabular Options: Hypothesis Tests

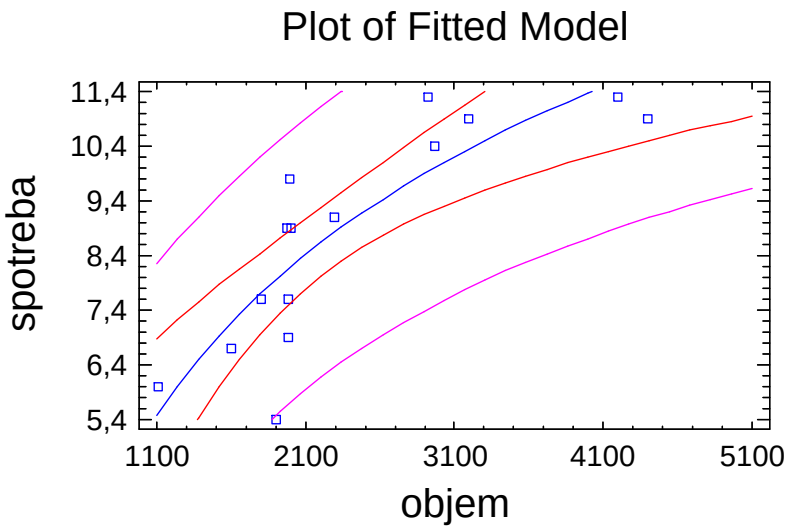
a pomocí p-value 0,99 t-testu nezamítneme nulovou hypotézu.

Tedy jsou splněny všechny předpoklady pro použití logaritmického modelu.

Závěr: Uvedené výstupy ukazují vhodnost použití logaritmického-X modelu pro popis závislosti mezi spotřebou a objemem. Rovnici hledané závislosti můžeme vyjádřit ve tvaru:

spotreba = -26,5512 + 4,57224*ln(objem),

přičemž oba uvedené parametry jsou statisticky významné, nelze je tedy z modelu vypustit.



Graf 3. Model Logarithmic-X

9. Některé další procedury ve *Statgraphicsu*

9.1. Paretová analýza pro ordinální proměnnou

Menu Special/QualityControl/ParetoAnalysis

Zkuste si tuto proceduru například na souboru **CarData** z datového balíku, který patří k softwaru *Statgraphics* (hledáme jej v adresáři **Statgra** na lokálním disku). Zvolte proměnnou „cylinders“.

9.2. Souhrnná analýza pro jednu proměnnou, včetně testování normality dat

Menu Snapstats/OneSampleAnalysis

Zkuste si tuto proceduru například na souboru **CarData** z datového balíku, který patří k softwaru *Statgraphics* (hledáme jej v adresáři **Statgra** na lokálním disku). Zvolte proměnnou „horsepower“.

9.3. Souhrnná analýza pro dvě proměnné

Menu Snapstats/TwoSampleComparison

Zkuste si tuto proceduru například na souboru **CarData** z datového balíku, který patří k softwaru *Statgraphics* (hledáme jej v adresáři **Statgra** na lokálním disku). Zvolte proměnné „horsepower“ a „mpg“.

9.4. Souhrnná analýza pro více než dvě proměnné

Menu Snapstats/MultipleSampleComparison

Zkuste si tuto proceduru například na souboru **CarData** z datového balíku, který patří k softwaru *Statgraphics* (hledáme jej v adresáři **Statgra** na lokálním disku). Zvolte proměnnou „horsepower“ a faktor „origin“. Pozor! Data jsou zadána v jednom sloupci („horsepower“) a v druhém sloupci („origin“) je zadán kód jednotlivých tříd, tedy při zadávání neodklikneme **Multiple Data Columns,** který je automaticky nastaven, ale **Data and Code Colums.**

Příklady k procvičení ke kapitolám 5 až 7

1. Pracovníci obchodní inspekce kontrolují váhu porce masa v určitém výrobku konzervárenského průmyslu. Technologická norma konzervy a tomu odpovídající cenová kalkulace udávají váhu masa v konzervě 90 g. Inspekce vyhodnotila 15 výrobků s těmito výsledky:

90 87 88 90 85 88 86 90 89 88 92 87 87 90 89 g.

Najděte 95% interval spolehlivosti pro střední hodnotu váhy porce masa.

2. Ze základního souboru 10.000 automaticky balených sáčků piškotů bylo vybráno 1% sáčků a zjištěna průměrná váha 15,8g a směrodatná odchylka 4,8g. Určete se spolehlivosti 0,99, v jakých mezích lze očekávat průměrnou váhu balíčků piškotů.

3. Při kontrole data spotřeby určitého druhu masové konzervy ve skladech produktů masného průmyslu bylo náhodně vybráno 320 konzerv a zjištěno, že 59 z nich má prošlou záruční lhůtu. Stanovte 95% interval spolehlivosti pro odhad procenta konzerv s prošlou záruční lhůtou.

4. Automat vyrábí pístové kroužky o daném průměru. Výrobce udává, že směrodatná odchylka průměru kroužku je 0,05mm. K ověření této informace bylo náhodně vybráno 80 kroužků a vypočtena směrodatná odchylka jejich průměru 0,04mm. Lze tento rozdíl považovat za významný ve smyslu zlepšení kvality produkce?

5. Tabáková firma TAB prohlašuje, že jejich cigarety mají nižší obsah nikotinu než cigarety NIK. Pro ověření tohoto prohlášení bylo náhodně vybráno z produkce TAB 20 krabiček cigaret (po 20-ti kusech) a v nich bylo zjištěno ($42{,}6 \pm 3{,}7$) mg nikotinu (v jediné cigaretě). Ve 25-ti krabičkách cigaret NIK (po 20-ti kusech) bylo zjištěno ($48{,}9 \pm 4{,}3$) mg nikotinu na cigaretu.

a) Ověřte tvrzení firmy TAB čistým testem významnosti.
b) Nalezněte 95% interval spolehlivosti pro obsah nikotinu v cigaretách TAB.

6. Data uvádějí množství zahraničních návštěvníků ČR v jednotlivých měsících a způsob jejich dopravy do ČR:

	Leden	Únor	Březen	Duben	Květen	Červen	Červenec	Srpen	Září	Říjen	Listopad	Prosinec
Road	6492	6464	6887	7854	8911	8589	9917	10497	8121	8443	7044	7273
Rail	340	301	359	396	415	455	494	522	423	443	356	366
Air	95	87	111	131	138	139	152	146	140	135	115	92
Celkem	6927	6852	7357	8381	9464	9183	10563	11165	8684	9021	7515	7731

a) Proveďte analýzu rozptylu ANOVA pro srovnání středních hodnot počtů zahraničních turistů podle způsobu jejich dopravy do ČR (Road-silnice, Rail-železnice, Air-letecká linka).

b) Najděte nejvhodnější typ regresní závislosti (pokud taková existuje) celkového počtu zahraničních turistů na jednotlivých měsících (měsíce nejprve převeďte na numerickou proměnnou).

7. Následující data byla převzata z informačního serveru BusinessInfo.cz a reprezentují nezaměstnanost a volná pracovní místa v ČR v období 2004-2005.

Měsíc	**PN 2004**	**PN 2005**	**VPM 2004**	**VPM 2005**
1.	569,5	561,7	41,7	54,2
2.	570,8	555,0	43,9	56,0
3.	559,8	540,5	42,4	55,9
4.	535,1	512,6	42,7	55,9
5.	520,4	494,6	44,7	57,2
6.	517,5	489,7	45,4	57,0
7.	532,1	500,3	45,7	56,8
8.	536,0	505,3	48,5	59,3
9.	530,2	503,4	47,1	55,8
10.	517,8	491,9	49,0	55,1
11.	517,7	490,8	50,3	53,0
12.	541,7	510,4	51,2	52,2

Legenda:

PN 2004 … počet nezaměstnaných v roce 2004 (v tisících)
PN 2005 … počet nezaměstnaných v roce 2005 (v tisících)
VPM 2004 … počet volných pracovních míst v roce 2004 (v tisících)
VPM 2005 … počet volných pracovních míst v roce 2005 (v tisících)

a) Srovnejte střední hodnotu počtu nezaměstnaných v roce 2004 a v roce 2005.

b) Zakreslete regresní křivku zachycující vývoj nezaměstnanosti na jednotlivých měsících, v roce 2004, resp. v roce 2005 (použijte kvadratickou funkci).

c) Zjistěte, zda existuje závislost mezi počtem volných pracovních míst a počtem nezaměstnaných osob v roce 2004, resp. 2005. Pokud ano, specifikujte ji.

8. Za účelem analýzy hrubé měsíční mzdy bylo dotázáno 20 osob v jeden den v určitém městě v ČR (zkrácená verze datového souboru uveřejněného na internetových stránkách ČSÚ):

Pohlaví	**Věk**	**Hrubá mzda (tis. Kč)**
M	29	12
Z	33	15,6
M	26	20
M	31	23,1
M	40	24
M	52	23,6
Z	38	19,5
M	19	22
Z	41	18,5
Z	55	23,8
M	40	18
Z	58	10,5
Z	21	13,4
Z	29	17,5
Z	31	18,5
Z	44	16,9
Z	46	17,1
M	39	27
M	32	19,9
Z	30	17,3

a) Zjistěte, zda se střední hodnota hrubé mzdy v daném městě liší od uváděného celorepublikového průměru 18 900 Kč.

b) Zjistěte, zda výše hrubé mzdy v daném městě závisí na pohlaví osoby.

c) Zjistěte, zda výše hrubé mzdy v daném městě závisí na věku.

Slovníček některých anglických termínů

Anglicky	**Česky**
Variable	Proměnná
Observation	Pozorování
Plot	Graf, vykreslit
Scaterplot	Bodový graf
Describe	Popsat
Compare	Porovnat
Relation	Závislost
Relate	Najít závislost
Simple regression	Jednoduchá regrese
Frequency	Četnost
Frequency histogram	Histogram rozdělení čeností
Average, SampleMean	Průměr
Standard deviation (St Dev)	Směrodatná odchylka
Count	Počet
Skewness	Šikmost
Kurtosis	Špičatost
z-score	z-souřadnice
Reject H_0	Zamítáme H_0
Do not reject H_0	Nezamítáme H_0

Literatura

1. Anděl J. : Matematická statistika, Praha, SNTL, 1978

2. Briš R., Litschmannová M. : Statistika I. Pro kombinované a distanční studium, VŠB-TU Ostrava, 2004,

3. Cyhelský L., Kalounová J., Hindls R. : Elementární statistická analýza, Management Press Praha, 1996,

4. Dupač V., Hušková M. : Pravděpodobnost a matematická statistika, Karolinum, Praha, 2001

5. Dummer M. : Introduction to Satistical Sience, VŠB-TU Ostrava, 1998,

6. Dummer M., Klímková M. : Statistika I. (cvičení), VŠB-TU Ostrava, 1997,

7. Friedrich V. : Statistika 1., Vysokoškolská učebnice pro distanční studium, Západočeská Univerzita, Plzeň 2002,

8. Hebák P., Kahounová J. : Počet pravděpodobnosti v příkladech, SNTL Praha, 1988

9. Hebák P., Hustopecký J., Jarošová E., Pecáková I. : Vícerozměrné statistické metody (1), (2), (3), Informatorium Praha, 2004

10. Hindls R., Hronová S., Seger J. : Statistika pro ekonomy, Professional Publishing Praha, 2004

11. Kunderová P.: Úvod do teorie pravděpodobnosti a matematické statistiky, Olomouc, 1997,

12. Křivý I. : Úvod do teorie pravděpodobnosti, Ostravská Univerzita, 1983,

13. Křivý I. : Základy matematické statistiky, Ostravská Univerzita, 1985,

14. Likeš J., Cyhelský L., Hindls R. : Úvod do statistiky a pravděpodobnosti, VŠE Praha, 1994

15. Likeš J., Machek J. :Počet pravděpodobnosti, SNTL Praha, 1982,

16. Likeš J., Machek J. : Matematická statistika, SNTL Praha, 1988,

17. Litschmannová M. : Statistika I. - příklady, VŠB-TU Ostrava, 2000,

18. Novovičová J. : Pravděpodobnost a základy matematické statistiky, ČVUT Praha, 2002

19. Riečan B. : Pravděpodobnost a matematická statistika, Bratislava

20. Riečan B, Neubrunn T. : Teória miery, Bratislava, 1992