# Práce se systémem STATISTICA

## Téma: Diagnostické grafy, testování normality

Vedení pojišťovny (zaměřené na pojištění automobilů) požádalo manažera oddělení marketingového výzkumu o provedení průzkumu, který by ukázal názory zákazníků na uvažovaný nový systém pojištění aut.

Náhodně bylo vybráno 110 současných zákazníků pojišťovny a ti byli telefonicky seznámeni s následujícím textem:

„Naše pojišťovna nabízí nový systém pojištění aut výhradně pro cesty nad 300 km. Za roční poplatek 12 tisíc Kč budete pojištěni pro případ libovolných potíží s autem při všech cestách nad 300 km. V případě nehody pojišťovna uhradí opravu, cestovní náklady a popř. i některé další výlohy, jako je ubytování a stravování v hotelu, telefon atd.

Stupnicí od 1 (jednoznačný nezájem) do 5 (jednoznačný zájem) laskavě vyjádřete svůj postoj k nabízenému novému typu pojištění. Dále uveďte svůj věk, počet cest nad 300 km v loňském roce, stáří vašeho auta a váš rodinný stav. Děkujeme."

Získané odpovědi byly zaznamenány do datového souboru a zakódovány takto:
POSTOJ ... postoj k novému typu pojištění (ne = 1, asi ne = 2, nevím = 3, asi ano = 4, ano = 5).
RODSTAV ... rodinný stav (svobodný = 1, rozvedený, ovdovělý = 2, ženatý = 3).
VEK ... věk v dokončených letech.
STARIAUT ... stáří auta v letech.
CESTY ... počet cest nad 300 km v předešlém roce.

**Úkol 1.** Datový soubor pojist.sta načtěte do systému STATISTICA. Všem proměnným vytvořte návěští a popište význam jednotlivých variant proměnných POSTOJ a RODSTAV.
**Návod**: File – Open – Soubory typu Data Files – pojist.sta – Otevřít.
Názvy a vlastnosti proměnných se upravují v okně, do něhož vstoupíte, když 2x kliknete myší na název proměnné. Návěští se píše do Long Name, význam variant do Text labels.

**Úkol 2.** Zjistěte absolutní a relativní četnosti a absolutní a relativní kumulativní četnosti proměnných POSTOJ a RODSTAV.
**Návod**: Statistics – Basic Statistics/Tables – Frequency Tables – OK – Variables POSTOJ, RODSTAV – OK – Summary Frequency tables. Tabulky se uloží do workbooku, listovat v nich můžete pomocí stromové struktury v levém okně.

**Úkol 3.** Vytvořte histogram proměnné VEK se šesti třídicími intervaly <23,29>, (29,35>, (35,41>, (41,47>, (47,53>, (53,59>.
**Návod**: Graphs – Histograms – Variables VEK, OK, Advanced – zaškrtněte Boundaries – Specify Boundaries – Enter Upper Boundaries 29 35 41 47 53 59, OK.

**Úkol 4.** Vytvořte kategorizovaný histogram proměnné VEK podle proměnné RODSTAV.
**Návod**: Postupujte stejně jako v předešlém případě a zvolte Categorized – X-categoried ON – Change Variable RODSTAV, OK, Codes – Specify Codes All, OK, OK.

**Úkol 5.** Sestrojte krabicový diagram proměnné CESTY. S jeho pomocí zjistěte, zda proměnná CESTY obsahuje odlehlé či extrémní hodnoty.
**Návod**: Graphs – 2D Graphs – Box Plots – Variables – Dependent variable CESTY – OK – OK.

Interpretace: Medián je posunut k dolnímu kvartilu, což svědčí o kladně zešikmeném rozložení. Vyskytují se odlehlé i extrémní hodnoty, jedná se tedy o špičaté rozložení.

**Úkol 6.** Pro proměnnou STARIAUT sestrojte NP plot a s jeho pomocí posuďte normalitu této proměnné.
**Návod**: Graphs – 2D Graphs – Normal Probability Plots – Variables STARIAUT – OK. Interpretace: Vzhled NP plot svědčí o kladně zešikmeném rozložení, nejedná se tedy o normální rozložení.

**Úkol 7.** Pro proměnnou STARIAUT nakreslete histogram s proloženou hustotou normálního rozložení. Ponechejte implicitní počet třídicích intervalů.
**Návod**: Graphs – Histograms – Variables STARIAUT – OK.
Interpretace: Tvar histogramu svědčí o kladně zešikmeném rozložení, jehož hustota neodpovídá hustotě normálního rozložení.

**Úkol 8.** Rozhodněte pomocí K-S testu a S-W testu na hladině významnosti 0,05, zda lze údaje o věku zákazníků považovat za realizace náhodného výběru z normálního rozložení.
**Návod**: Statistics – Basic Statistics / Tables – Descriptive statistics – OK – Variables VEK, OK – Normality – zaškrtněte Kolmogorov – Smirnov & Liliefors test for normality a Shapiro - Wilk's W test – Frequency tables. Ve výstupu se objeví tabulka, v níž je uvedena hodnota testové statistiky pro K-S test ($d = 0{,}11222$) a S-W test ($W = 0{,}96695$) a odpovídající p-hodnoty. U K-S testu uvažujte Liliefors p, které je počítáno na základě parametrů odhadnutých z dat. V našem případě $p < 0{,}01$ a pro S-W test $p = 0{,}00783$, tedy oba testy zamítají na hladině významnosti 0,05 hypotézu o normalitě. Výpočet je vhodné doplnit N-P plotem.

# Téma: Úlohy o jednom náhodném výběru z normálního rozložení

**Úkol 1.**: **Vlastnosti výběrového průměru z normálního rozložení**
Předpokládejme, že velký ročník na vysoké škole má výsledky ze statistiky normálně rozloženy kolem střední hodnoty 72 bodů se směrodatnou odchylkou 9 bodů. Najděte pravděpodobnost, že průměr výsledků náhodného výběru 10 studentů bude větší než 80 bodů.
**Návod**: $X_1, ..., X_{10}$ je náhodný výběr z N(72, 81). Počítáme P(M > 80), přičemž výběrový průměr M má normální rozložení se střední hodnotou 72 a rozptylem 81/10. Tedy $P(M > 80) = 1 - P(M \leq 80) = 1 - \Phi(80)$, kde $\Phi(80)$ je hodnota distribuční funkce rozložení N(72; 8,1) v bodě 80. Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Long Name této proměnné napíšeme =1 – INormal(80;72;sqrt(8,1)). Zjistíme, že $1 - \Phi(80) = 0{,}00248$. Funkce INormal(x;μ;σ) počítá hodnotu distribuční funkce rozložení $N(\mu,\sigma^2)$ v bodě x.

**Úkol 2.**: **Intervaly spolehlivosti pro parametry $\mu,\sigma^2$ normálního rozložení**
Z populace stejně starých selat téhož plemene bylo vylosováno 6 selat a po dobu půl roku jim byla podávána táž výkrmná dieta. Byly zaznamenávány průměrné denní přírůstky hmotnosti v Dg. Z dřívějších pokusů je známo , že v populaci mívají takové přírůstky normální rozložení, avšak střední hodnota i rozptyl se měnívají. Přírůstky v Dg: 62, 54, 55, 60, 53, 58. Při riziku $\alpha = 0{,}05$ odvoďte:
a) levostranný interval spolehlivosti pro neznámou střední hodnotu μ při neznámé směrodatné odchylce σ.
b) interval spolehlivosti pro směrodatnou odchylku σ.
**Návod**: Vytvoříme datový soubor o 5 proměnných a 6 případech. První proměnnou nazveme hmotnost, druhou dm1, třetí dm2 a čtvrtou hm2. Do proměnné hmotnost zapíšeme zjištěné údaje. Pomocí Descriptive Statistics zjistíme výběrový průměr a výběrovou směrodatnou odchylku. Do LongName proměnné dm1 zapíšeme výraz = m – s* VStudent(0,95;5)/√6, kde za m a s dosadíme vypočtené hodnoty. Funkce VStudent(x;df) počítá x-kvantil rozložení t(df). Dostaneme výsledek 54,06, tedy μ > 54,06 Dg s pravděpodobností aspoň 0,95.
Do Long Name proměnné dm2 zapíšeme výraz =s*sqrt(5)/sqrt(VChi2(0,975;5)), kde za s dosadíme vypočtenou směrodatnou odchylku. Podobně do LongName proměnné hm2 zapíšeme výraz =s*sqrt(5)/sqrt(VChi2(0,025;5)), kde za s dosadíme vypočtenou směrodatnou odchylku. Funkce VChi2(x;nu) počítá x-kvantil rozložení $\chi^2$(nu).
Výsledek: 2,23 g < σ < 8,77 g s pravděpodobností aspoň 0,95.

**Úkol 3.**: **Interval spolehlivosti pro rozdíl parametrů $\mu_1$-$\mu_2$ dvourozměrného normálního rozložení**
Bylo vylosováno 6 vrhů selat a z nich vždy dva sourozenci. Jeden z nich vždy dostal náhodně dietu č. 1 a druhý dietu č. 2. Přírůstky v Dg jsou následující: (62,52), (54,56), (55,49), (60,50), (53,51), (58,50). Za předpokladu, že uvedené dvojice tvoří náhodný výběr z dvourozměrného normálního rozložení s vektorem středních hodnot ($\mu_1$, $\mu_2$), sestrojte 95% interval spolehlivosti pro rozdíl středních hodnot.
**Návod**: Vytvoříme datový soubor o třech proměnných a 6 případech. Do proměnných v1 a v21 zapíšeme naměřené přírůstky, do proměnné v3 uložíme rozdíly v1-v2.Pomocí Descriptive Statistics zjistíme meze 95% intervalu spolehlivosti pro střední hodnotu proměnné v3 tak, že zaškrtneme Conf. limits for mean. Dostaneme výsledek: 0,63 Dg < μ < 10,71 Dg s pravděpodobností aspoň 0,95.

**Úkol 4.**: **Testování hypotézy o parametru μ normálního rozložení**
Systematická chyba měřicího přístroje se eliminuje nastavením přístroje a měřením etalonu, jehož správná hodnota je $\mu = 10{,}00$. Nezávislými měřeními za stejných podmínek byly získány hodnoty: 10,24 10,12 9,91 10,19 9,78 10,14 9,86 10,17 10,05, které považujeme za realizace náhodného výběru rozsahu 9 z rozložení $N(\mu, \sigma^2)$. Je možné při riziku 0,05 vysvětlit odchylky od hodnoty 10,00 působením náhodných vlivů?
**Návod**: Jde o úlohu na jednovýběrový t-test. Vytvoříme datový soubor o jedné proměnné a 9 případech, kam zapíšeme naměřené hodnoty. V Descriptive Statistics vybereme t-test, single sample. Do Reference values zapíšeme 10. Ve výstupu se podíváme na hodnotu testového kritéria a na p-hodnotu. Pokud p-hodnota bude menší nebo rovna 0,05, zamítneme hypotézu $H_0$: $\mu = 0$ ve prospěch alternativní hypotézy $H_1$: $\mu \neq 0$ na hladině významnosti 0,05. V opačném případě $H_0$ nezamítáme. V našem případě nulovou hypotézu nezamítáme na hladině významnosti 0,05.

**Úkol 5.**: **Testování hypotézy o rozdíl parametrů $\mu_1$-$\mu_2$ dvourozměrného normálního rozložení**
Pro data z úkolu 3 testujte na hladině významnosti 0,05 hypotézu, že obě výkrmné diety mají stejný vliv.
**Návod**: Jde o úlohu na párový t-test. Vytvoříme datový soubor o dvou proměnných a šesti případech. V Descriptive Statistics vybereme t-test, dependent samples. Zadáme názvy obou proměnných a ve výstupu se podíváme na hodnotu testového kritéria a na p-hodnotu. Pokud p-hodnota bude menší nebo rovna 0,05, zamítneme hypotézu $H_0$: $\mu = 0$ ve prospěch alternativní hypotézy $H_1$: $\mu \neq 0$ na hladině významnosti 0,05. V opačném případě $H_0$ nezamítáme. V našem případě nulovou hypotézu zamítáme na hladině významnosti 0,05.

# Téma: Úlohy o dvou nezávislých náhodných výběrech z normálních rozložení

Do programu STATISTICA načtěte ASCII soubor studentky.dat, který obsahuje údaje o 48 náhodně vybraných studentkách VŠE v Praze. 1. sloupec – výška, 2. sloupec – známka z matematiky v 1. semestru, 3. sloupec – obor studia (1 – národní hospodářství, 2 – informatika). Tyto tři proměnné nazvěte X,Y,Z a vytvořte jim návěští.
Před prováděním následujících úkolů pomocí diagnostických grafů orientačně ověřte normalitu dat.

**Úkol 1.** Sestrojte 95% interval spolehlivosti pro střední hodnotu výšky studentek oboru nh a studentek oboru inf.
**Návod**: Meze 95% intervalu spolehlivosti pro střední hodnotu proměnné X zjistíme pomocí Descriptive Statistics, kde zaškrtneme Conf. limits for mean. Výsledek pro studentky oboru nh: $167,3 < \mu < 172,3$ s pravděpodobností aspoň 0,95, pro studentky oboru inf: $164,8 < \mu < 169,0$ s pravděpodobností aspoň 0,95.

**Úkol 2.** Sestrojte 95% interval spolehlivosti pro podíl rozptylů výšek studentek oboru nh a inf.
**Návod**: K datovému souboru přidáme další dvě proměnné DM a HM pro výpočet dolní a horní meze intervalu spolehlivosti. Do LongName těchto proměnných zapíšeme vzorce pro dolní a horní mez intervalu spolehlivosti pro podíl rozptylů. Výběrové rozptyly pro 1. a 2. výběr zjistíme pomocí Descriptive Statistics. Výsledek: $0,821 < \sigma_1^2/\sigma_2^2 < 4,513$ s pravděpodobností aspoň 0,95.

**Úkol 3.** Na hladině významnosti 0,05 testujte hypotézu, že rozptyly výšek studentek oboru nh a inf jsou shodné.
**Návod**: Lze využít výsledku 2. úkolu. 95% interval spolehlivosti pro podíl rozptylů obsahuje číslo 1, tedy hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05.

**Úkol 4.** Sestrojte 95% interval spolehlivosti pro rozdíl středních hodnot výšek studentek oboru nh a inf.
**Návod**: K datovému souboru přidáme další dvě proměnné DM a HM pro výpočet dolní a horní meze intervalu spolehlivosti. Do LongName těchto proměnných zapíšeme vzorce pro dolní a horní mez intervalu spolehlivosti pro rozdíl středních hodnot. Výběrové průměry a výběrové rozptyly pro 1. a 2. výběr zjistíme pomocí Descriptive Statistics. Výsledek: $-0,452 < \mu_1 - \mu_2 < 6,292$ s pravděpodobností aspoň 0,95.

**Úkol 5.** Na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty výšek studentek oboru nh a inf jsou shodné.
**Návod**:
1. způsob: lze využít výsledku 4. úkolu. 95% interval spolehlivosti pro rozdíl středních hodnot obsahuje číslo 0, tedy hypotézu o shodě středních hodnot nezamítáme na hladině významnosti 0,05.
2. způsob: úloha vede na dvouvýběrový t-test. Statistics – Basic Statistics – t-test, independent, by groups – OK, Variables – Dependent X, grouping Z – OK – Summary. Ve výstupní tabulce najdeme hodnotu testového kritéria a p-hodnotu. Protože p-hodnota = 0,087837 je větší než hladina významnosti 0,05, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

## Téma: Analýza rozptylu jednoduchého třídění

**Úkol:** V jisté továrně se měřil čas, který potřeboval každý ze tří dělníků k uskutečnění téhož pracovního úkonu. Čas v minutách:
1. dělník: 3,6 3,8 3,7 3,5,
2. dělník: 4,3 3,9 4,2 3,9 4,4 4,7,
3. dělník: 4,2 4,5 4,0 4,1 4,5 4,4.
Na hladině významnosti 0,05 testujte hypotézu, že výkony těchto tří dělníků jsou stejné. Zamítnete-li nulovou hypotézu, určete, výkony kterých dělníků se liší na dané hladině významnosti.
**Návod**: Vytvořte datový soubor se dvěma proměnnými (X a ID) a 16 případy. Do 1. sloupce napište změřené časy, do 2. sloupce dejte čtyřikrát jedničku, šestkrát dvojku a šestkrát trojku. Statistics - Basic Statistics and Tables - Breakdown & one-way ANOVA - Variables Dependent X, Grouping ID, OK, Codes for grouping variables – All, OK, Quick – Summary: Table of statistics (zobrazí se průměry, směrodatné odchylky a rozsahy všech tří výběrů) – návrat do Statistics by Groups – Categorized box & whisker plot (současné zobrazení krabicových diagramů pro všechny tři výběry – změna typu krabicového diagramu se provede po dvojnásobném kliknutí myší na graf v menu Plot:Box/Whisker) – návrat do Statistics by Groups – ANOVA & tests – Categorized normal prob. plots (vizuální posouzení normality všech výběrů) – návrat do Statistics by Groups – Levene tests (testování homogenity rozptylů všech tří výběrů - p-hodnota = 0,256, tedy na hladině významnosti 0,05 se nezamítá hypotézu o shodě rozptylů) – návrat do Statistics by Groups – Analysis of Variance (provední analýzy rozptylu). Ve výstupní tabulce je použito následující označení: SS Effect … skupinový součet čtverců $S_A$, MS Effect … $S_A/(r-1)$, SS Error … reziduální součet čtverců $S_E$, MS Error … $S_E/(n-r)$. Protože p-hodnota = 0,00268, zamítá se na hladině významnosti 0,05 hypotézu o shodě středních hodnot. Návrat do Statistics by Groups – Post- hoc – Scheffé test. Výsledek Scheffého metody ukazuje, že na hladině významnosti 0,05 se liší výkony dělníků (1,2), (1,3) a neliší se (2,3).

# Téma: Pořadové testyy o mediánech

**Úkol 1**.: **Jednovýběrový Wilcoxonův test**
Vyráběné ocelové tyče mají kolísavou délku s předpokládanou hodnotou mediánu 10 m. Náhodný výběr 10 tyčí poskytl tyto výsledky:
9,83 10,10 9,72 9,91 10,04 9,95 9,82 9,73 9,81 9,90
Na hladině významnosti 0,05 testujte hypotézu, že předpoklad o mediánu délky tyčí je oprávněný.
Návod: Vytvořte datový soubor se dvěma proměnnými X a Y a 10 případy. Do proměnné X napište změřené hodnoty, proměnná Y bude obsahovat konstantu 10. Statistics –Nonparametrics - Comparing two dependent samples(variables) – OK – First variable list X, Second variable list Y – Wilcoxon matched pair test - OK. Wilcoxonův test poskytne p-hodnotu = 0,024933, tedy nulová hypotéza se zamítá na hladině významnosti 0,05.

**Úkol 2**.: **Párový Wilcoxonův test**
Při zjišťování kvality jedné složky půdy se používají dvě metody označené A a B. Výsledky:

| Vzorek | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0,275 | 0,312 | 0,284 | 0,3 | 0,365 | 0,298 | 0,312 | 0,315 | 0,242 | 0,321 | 0,335 | 0,307 |
| B | 0,28 | 0,312 | 0,288 | 0,298 | 0,361 | 0,307 | 0,319 | 0,315 | 0,242 | 0,323 | 0,341 | 0,315 |

Na hladině významnosti 0,05 testujte hypotézu, že metody A a B dávají stejné výsledky.
**Návod**: Vytvořte datový soubor se dvěma proměnnými A a B a 12 případy. Statistics –Nonparametrics - Comparing two dependent samples(variables) – OK – First variable list A, Second variable list B – OK – Wilcoxon matched pair test. Výstupní tabulka poskytne hodnotu testové statistiky (ozn. T), hodnotu asymptotické testové statistiky $U_0$ a p-hodnotu pro $U_0$. (STATISTICA tedy nezohledňuje omezení $n \geq 30$ pro použití $U_0$.) V tomto případě je p-hodnota 0,038153, tedy nulová hypotéza se zamítá na hladině významnosti 0,05.
Grafické znázornění výsledků: Návrat do Comparing two variables - Box & Whisker plots for all variables – OK - Box & Whisker Type: Median/Quart/Range – OK. Z krabicových diagramů je vidět, že obě metody se poněkud liší v úrovni, ale neliší se ve variabilitě.

**Úkol 3**.: **Dvouvýběrový Wilcoxonův test**
Bylo vybráno 10 polí stejné kvality. Na čtyřech z nich se zkoušel nový způsob hnojení, zbylých šest bylo ošetřeno starým způsobem. Pole byla oseta pšenicí a sledoval se její hektarový výnos. Je třeba testovat na hladině významnosti 0,05, zda nový způsob hnojení má týž vliv na průměrné hektarové výnosy pšenice jako starý způsob hnojení.
hektarové výnosy při novém způsobu: 51 52 49 55
hektarové výnosy při starém způsobu: 45 54 48 44 53 50
**Návod**: Vytvořte datový soubor o dvou proměnných (X a ID) a deseti případech. Do X napište výnosy pšenice při obou způsobech hnojení, do ID čtyřikrát jedničku a šestkrát dvojku. Statistics – Nonparametric – Comparing two independent samples (groups) – OK - Dependent variable X, Grouping variable ID, OK - Mann – Whitney U test. Ve výstupní tabulce jsou součty pořadí $T_1$, $T_2$, hodnota testové statistiky $\min(U_1, U_2)$ ozn. U, hodnota asymptotické testové statistiky $U_0$ (ozn. Z), p-hodnota pro $U_0$ a přesná p-hodnota (ozn. 2*1 sided exact p – ta se používá pro rozsahy výběrů pod 30). V našem případě přesná p-hodnota = 0,352381, tedy $H_0$ nezamítáme na hladině významnosti 0,05. Výpočet je vhodné doplnit krabicovým diagramem typu Median/Quart/Range.

**Úkol 4**.: **Kruskalův – Wallisův test a mediánový test**

Voda po holení jisté značky se prodává ve čtyřech různých lahvičkách stejného obsahu. Údaje o počtu prodaných lahviček za týden v různých obchodech:

1.typ: 50 35 43 30 62 52 43 57 33 70 64 58 53 65 39
2.typ: 31 37 59 67 44 49 54 62 34 42 40
3.typ: 27 19 32 20 18 23
4.typ: 35 39 37 38 28 33.

Posuďte na 5% hladině významnosti, zda typ lahvičky ovlivňuje úroveň prodeje vyjádřenou mediánem.

**Návod**: Vytvořte nový datový soubor o dvou proměnných X a ID a 38 případech. Do proměnné X napište zjištěné údaje o prodeji, do proměnné ID 15 x jedničku, 11 x dvojku, 6 x trojku a 6 x čtyřku. Statistics – Nonparametrics – Comparing multiple independent samples(groups) – OK – Dependent varaible VÝKON, Grouping varaible SKUPINA – OK – Summary: Kruskal-Wallis ANOVA & Median test. Ve dvou výstupních tabulkách se objeví výsledky mediánového testu a K-W testu. Oba testy zamítají hypotézu o shodě mediánů v daných čtyřech skupinách, ale K-W test je poněkud silnější (p-hodnota = 0,0003, zatímco p-hodnota pro mediánový test je 0,0005). Grafické znázornění výsledků: návrat do Kruskal-Wallis ANOVA & Median test – Box & Whisker – Select variable X – OK - Box & Whisker Type: Median/Quart/Range – OK. Je vidět, že úroveň prodeje pro 1. typ je nevyšší, zatímco pro 3. typ nejnižší. Dále je možno vytvořit histogramy proměnné X ve všech čtyřech skupinách: návrat do Kruskal-Wallis ANOVA & Median test – Categorized histogram - Select variable X – OK.

Pomocí metody mnohonásobného porovnávání lze zjistit, že na hladině významnosti 0,05 se liší 1. a 3. typ, 1. a 4. typ a 2. a 3. typ.

# Téma: Analýza závislosti dvou veličin

**Úkol 1.**: **Testování nezávislosti nominálních veličin**
V roce 1950 zkoumali Yule a Kendall barvu očí a vlasů u 6800 mužů.

| Barva očí | Barva vlasů | | | |
|---|---|---|---|---|
| | světlá | kaštanová | černá | rezavá |
| modrá | 1768 | 807 | 180 | 47 |
| šedá nebo zelená | 946 | 1387 | 746 | 53 |
| hnědá | 115 | 438 | 288 | 16 |

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti barvy očí a barvy vlasů. Vypočtěte Cramérův koeficient. Simultánní četnosti znázorněte graficky.
**Návod**: Vytvořte nový datový soubor o 12 případech a třech proměnných (OCI, VLASY, CETNOST). Do proměnné OCI napište varianty barvy očí $x_{[1]} = 1$ (modrá), $x_{[2]} = 2$ (šedá nebo zelená), $x_{[3]} = 3$ (hnědá), přičemž každá varianta se objeví čtyřikrát pod sebou. Do proměnné VLASY napište třikrát pod sebe všechny varianty $y_{[1]} = 1$ (světlá), $y_{[2]} = 2$ (kaštanová), $y_{[3]} = 3$ (černá), $y_{[4]} = 4$ (rezavá). Statistics - Basic Statistics/Tables - Tables and Banners - Specify Tables – List 1 OCI, List 2 VLASY, OK, Weight - CETNOST Status On, OK – Statistics for two way tables - zaškrtněte Pearson & M-L Chi -square, Phi & Cramer's V – Advanced - Detailed two-way tables. Ve výstupní tabulce najdete mj. hodnotu testové stastistiky (Chi-square = 1073,51) s počtem stupňů volnosti (df = 6) a odpovídající p-hodnotou (p = 0,0000) i Cramérův koeficient (V = 0,281). Pro grafické znázornění četností se vraťte do Crosstabulation Table Results – Advanced – 3D histograms. Po vytvoření grafu je nutné manuálně zvětšit rozsah zobrazovaných hodnot na osách x a y.
Pomocí STATISTIKY je možno lehce ověřit splnění podmínek dobré aproximace (tzn., že teoretické četností mají být aspoň v 80% případů větší než 5 a ve zbylých 20% případů nemají klesnout pod 2. Teoretické četnosti se vypočítají tak, že v Options zaškrtneme Expected frequencies. V našem případě jsou podmínky dobré aproximace splněny.

**Úkol 2.**: **Fisherův faktoriálový test**
100 náhodně vybraných mužů a žen bylo dotázáno, zda dávají přednost nealkoholickému nápoji A či B. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

| pohlaví | nápoj | |
|---|---|---|
| | A | B |
| muž | 20 | 30 |
| žena | 30 | 20 |

Na hladině významnosti 0,05 testujte pomocí Fisherova faktoriálového testu hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.
**Návod**: Podle návodu z předešlého příkladu vytvořte datový soubor. Proměnné budou mít názvy POHLAVI, NAPOJ a CETNOST. Statistics - Basic Statistics/Tables - Tables and Banners - Specify Tables – List 1 POHLAVI, List 2 NAPOJ, OK, Weight - CETNOST Status On, OK – Options - Statistics for two way tables - zaškrtněte Fisher exact, Yates, McNemar (2x2) – Advanced - Detailed two-way tables. Ve výstupní tabulce je mimo jiné uvedena p-hodnota pro oboustranný (Fisher exact, two tailed) a jednostranný test (Fisher exact, one tailed). V našem případě se jedná o jednostrannou závislost, zajímáme se tedy o Fisher exact, one tailed. Ta je 0,03567. Protože p-hodnota je menší nebo rovna 0,05, zamí-

táme na hladině významnosti hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

**Úkol 3.**: **Podíl šancí**
18 mužů onemocnělo určitou chorobou. Někteří z nich se léčili, jiní ne. Někteří se uzdravili, jiní zemřeli. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

| přežití | léčení | |
|---|---|---|
| | ano | ne |
| ano | 5 | 3 |
| ne | 6 | 4 |

Vypočtěte podíl šancí a sestrojte 95% asymptotický interval spolehlivosti pro podíl šancí. Pomocí tohoto intervalu spolehlivosti testujte na asymptotické hladině významnosti 0,05 hypotézu, že přežití nezávisí na léčení.

**Návod**: Podíl šancí vypočteme ručně. $\mathrm{OR} = \frac{ac}{bd} = \frac{5 \cdot 4}{3 \cdot 6} = \frac{20}{18} = \frac{10}{9} = 1,\bar{1}$. (Protože podíl šancí je větší než 1, je zřejmě výhodnější se nechat léčit.) Dolní a horní mez intervalu spolehlivosti pro OR zjistíme pomocí STATISTIKY. Vytvoříme datový soubor o dvou proměnných DM a HM a dvou případech. Do Long Name proměnné DM napíšeme vzorec pro dolní mez: =exp(log(10/9)-sqrt((1/5+1/3+1/6+1/4)*VNormal(0,975;0;1)) a analogicky zjistíme horní mez. Výsledek: 0,1645 < OR < 7,506 s pravděpodobností přibližně 0,95. Protože tento interval spolehlivosti obsahuje 1, nelze na asymptotické hladině významnosti 0,05 zamítnout hypotézu, že přežití nezávisí na léčení.

**Úkol 4.**: **Testování nezávislosti ordinálních veličin**
Dva lékaři hodnotili stav sedmi pacientů po témž chirurgickém zákroku tak, aby nejvyšší pořadí měl nejtěžší případ.

| č. pacienta | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1. lékař | 4 | 1 | 6 | 5 | 3 | 2 | 7 |
| 2. lékař | 4 | 2 | 5 | 6 | 1 | 3 | 7 |

Vypočtěte Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou lékařů jsou nezávislá.
**Návod**: Vytvořte datový soubor o sedmi případech a dvou proměnných X a Y. Statistics - Nonparametrics - Correlations – First variable list X, Second variable list Y, OK, Spearman R. Ve výstupní tabulce najdete Spearmanův koeficient a p-hodnotu. Nulová hypotéza se zamítá na hladině významnosti 0,05, protože p-hodnota = 0,013697 ≤ 0,05.

**Úkol 5.**: **Testování nezávislosti intervalových a poměrových veličin**
Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.
Počet bodů z 1. testu: 80 50 36 58 72 60 56 68
Počet bodů z 2. testu: 65 60 35 39 48 44 48 61
Nakreslete dvourozměrný tečkový diagram, vypočtěte výběrový korelační koeficient a na hladině významnosti 0,05 testujte hypotézu o nezávislosti výsledků obou testů.
**Návod**: Vytvořte datový soubor o dvou proměnných X a Y a osmi případech. Obvyklým způsobem zobrazte dvourozměrný tečkový diagram, s jehož pomocí posoudíte dvourozměrnou normalitu dat.

Testování hypotézy o nezávislosti: Statistics - Basic Statistics /Tables - Correlation matrices – OK - One variable list X,Y, OK – OK - Display r, p-levels and N's - Summary. Ve výstupní tabulce je hodnotu výběrového korelačního koeficientu $R_{12}$ (r=0,6264, tzn. že mezi X a Y existuje nepříliš silná přímá lineární závislost) a p-hodnotu pro test hypotézy o nezávislosti (p=0,097, $H_0$ tedy nelze zamítnout na hladině významnosti 0,05).
**Poznámka:** Pokud známe výběrový koeficient korelace a rozsah výběru, můžeme test nezávislosti veličin X, Y provést pomocí Probabilty Calculator. Statistics – Probability Calculator – Correlation – zadáme n a r, zaškrtneme Compute ρ from r – Compute.