Analýza dat z dotazníkových šetření

Cvičení 4. + 5.

- Zobecňování výběru na populaci

Zdrojová data: dotazník *http://www.vyplnto.cz/realizovane-pruzkumy/37771/*

- Seznamte se s dotazníkem a strukturou otázek, zamyslete se nad vhodností jednotlivých odpovědí a škál
- Stáhněte si data z dotazníku ve formátu ****.csv*** a otevřete je v aplikaci Excel, zamyslete se, která data by bylo vhodnější upravit přímo v Excelu -> uložte dotazník ve formátu ****.xls*** a otevřete datový soubor v prostředí **SPSS**

Opakování: Nominální proměnná více hodnotová odpověď.

- Podívejme se do dotazníku na otázku: ***Jaké jsou další 2 nejdůležitější kritéria při výběru cestovní kanceláře pro Vaši dovolenou?***
- Je zde několik různých možností odpovědí. Nicméně pro časovou úsporu budeme analyzovat jen první 3 sloupce: *cena, webové stránky, doporučení*

Data musí být vždy v numerické podobě. Provedeme překódování, pro úsporu času překódujeme pomocí funkce *Automatic Recode* (potom máme hodnoty proměnných 1, když je buňka prázdná a 2 když je vypsána odpověď – proměnná dichotomická, počitatelná hodnota 2)

Vytvoříme balík odpovědí: *Analyze* -> *Multiple Response* -> *Define Variable Sets* vyberu balík, který chci vyhodnocovat (již překódovaný musí být **NUMERIC**).

Zavřu okno a pokračuji dále, znovu obdobný postup: *Analyze* -> *Multiple Response* -> *Frequencies* zaškrtnu *Dichotomies* a jakou *Counted Value* napíšu 2.

$kriteria Frequencies

		Responses		
		N	Percent	Percent of Cases
kriteria vyberu dovolene[a]	Jaké jsou další 2 nejdůležitější kritéria při výběru cestovní kanceláře pro Vaši dovolenou? - cena zájezdů/pobytů	35	53,8%	67,3%
	Jaké jsou další 2 nejdůležitější kritéria při výběru cestovní kanceláře pro Vaši dovolenou? - přehledné webové stránky, online podpora	10	15,4%	19,2%
	Jaké jsou další 2 nejdůležitější kritéria při výběru cestovní kanceláře pro Vaši dovolenou? - doporuční	20	30,8%	38,5%
Total		65	100,0%	125,0%

a. Dichotomy group tabulated at value 2.

Opakování minulé hodiny:

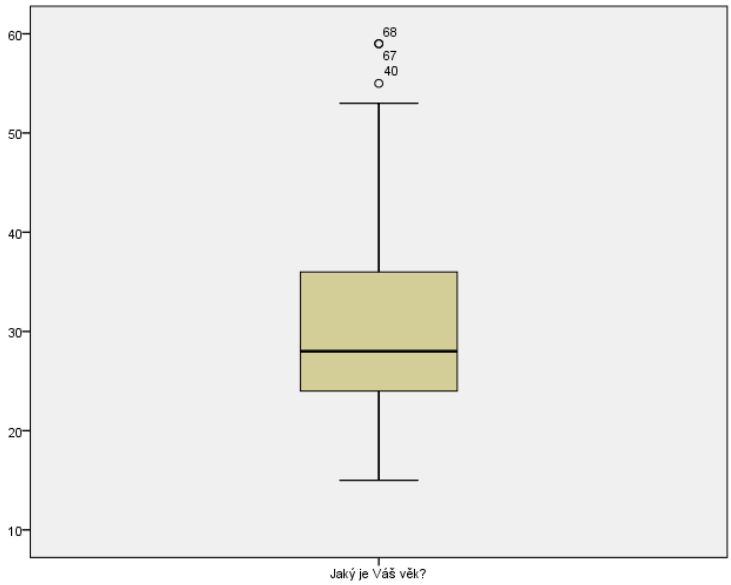
- Zopakujte si číselné a grafické charakteristiky numerické proměnné (rozptyl, odchylka, kvartily, kvantily, boxplot)

Vyhodnoťme všechny charakteristiky pro numerickou proměnnou: *Jaký je váš věk?*

Statistics

Jaký je Váš věk?

N	Valid	102
	Missing	0
Mean		31,06
Median		28,00
Mode		26
Std. Deviation		10,073
Variance		101,462
Minimum		15
Maximum		59
Percentiles	25	24,00
	50	28,00
	75	36,25



- *Určeme 95% interval spolehlivosti pro průměr proměnné Věk*

$$\bar{x} - \frac{\hat{\sigma}}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1) \leq \mu \leq \bar{x} + \frac{\hat{\sigma}}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)$$

$t_{1-\frac{\alpha}{2}}(n-1)$ …..$100 \cdot \left(1-\frac{\alpha}{2}\right)$ procentní kvantil Studentova rozdělení s *(n-1)* stupni volnosti

$\hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ odhad rozptylu základního souboru

Většina hodnot je známa z předchozí tabulky:

$$\bar{x} = 31{,}06;\ \hat{\sigma} = 101{,}46; n = 102; t_{1-\frac{\alpha}{2}}(n-1) = 1{,}66$$

$t_{1-\frac{\alpha}{2}}(n-1)$- kvantil studentova rozdělení zjistíme pomocí SPSS. Spss vždy požaduje, aby byla zadána nějaká buňka v tabulce (existující sloupec). Vytvoříme nový sloupec *t*, jen vytvoříme sloupec, nic dalšího nepotřebujeme, dále pokračujeme:

Transform -> *Compute Variable*

- Jako *Target Variable* zaznačíme náš nově vytvořený sloupec *t*, do kolonky *Numeric Expression* zadáme funkci, která nám vrátí konkrétní výsledek (nyní se jedná o kvantil proto funkce *Inverse DF* a jelikož jde o studentovo rozdělení nebo-li *t-test* výsledná funkce bude *IDF.T(prob. df)*
- Do závorek zadáváme po řadě tyto proměnné *prob* = kolika procentní kvantil požadujeme, *df* = počet stupňů volnosti
- Nyní tedy *IDF.T (0.95, 101)*…. Ve vyjádření funkce vždy píšeme desetinnou tečku!!! Čárkou jsou odděleny od sebe jednotlivé hodnoty.

$$31{,}06 - \frac{101{,}46}{\sqrt{102}} \cdot 1{,}66 \leq \mu \leq 31{,}06 + \frac{101{,}46}{\sqrt{102}} \cdot 1{,}66$$
$$31{,}06 - 10{,}05 \cdot 1{,}66 \leq \mu \leq 31{,}06 + 10{,}05 \cdot 1{,}66$$
$$14{,}38 \leq \mu \leq 47{,}7$$

95% hodnot proměnné věk leží v intervalu 14 až 48 let.

Př.: Z průzkumu bylo zjištěno, že v roce 2013 navštívilo Chorvatsko 62% respondentů. Jaký je intervalový odhad (α=5%), jestliže bylo tázáno 1225 respondentů.

Předpoklad: $n \cdot p \cdot (1-p) \geq 5$

$1225 \cdot 0{,}62 \cdot 0{,}38 \geq 5 => 288{,}6$

$$p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

$z_{1-\frac{\alpha}{2}}$ ….$100 \cdot \left(1-\frac{\alpha}{2}\right)$ procentní kvantil normalizovaného normálního rozdělení

Vytvoříme nový sloupec *norm.chor*, jen vytvoříme sloupec, nic dalšího nepotřebujeme, dále pokračujeme:

Transform -> *Compute Variable* -

- Jako *Target Variable* zaznačíme náš nově vytvořený sloupec *norm.chor*, do kolonky *Numeric Expression* zadáme funkci, která nám vrátí konkrétní výsledek (nyní se jedná o kvantil proto funkce *Inverse DF* a jelikož jde o normální rozdělelní výsledná funkce bude *IDF.NORMAL(prob, mean, stddev)*
- Do závorek zadáváme po řadě tyto proměnné *prob* = kolika procentní kvantil požadujeme, *mean* = střední hodnota (pro normalizované normální rozdělení je střední hodnota 0), *stddev* = směrodatná odchylka (pro normalizované normální rozdělení je směrodatná odchylka 1)
- Nyní tedy *IDF.T (0.62,0,1)*…. Ve vyjádření funkce vždy píšeme desetinnou tečku!!! Čárkou jsou odděleny od sebe jednotlivé hodnoty.

$$0{,}62 - 0{,}31\sqrt{\frac{0{,}62(1-0{,}62)}{1225}} \leq \pi \leq 0{,}62 + 0{,}31\sqrt{\frac{0{,}62(1-0{,}62)}{1225}}$$
$$0{,}62 - 0{,}014 \leq \pi \leq 0{,}62 + 0{,}014 => 0{,}61 \leq \pi \leq 0{,}63$$

95% interval spolehlivosti je mezi 0,61 a 0,63….

Testování hypotéz:

Proti sobě dvě hypotézy nulová H_0 a alternativní H_A :

H_0: $\theta = \theta_0$; H_A: $\theta \neq \theta_0$ … oboustranná alternativa případně existují jen levostranná H_A: $\theta < \theta_0$a pravostranná H_A: $\theta > \theta_0$alternativa

Cílem testování nulové hypotézy je dospět k úsudku, zda můžeme tuto hypotézu zamítnout vzhledem ke stanovené hypotéze alternativní.

- *Pomocí testového kritéria = určitá vhodná statistika, která má při platnosti nulové hypotézy známé pravděpodobnostní rozdělení (def. Obor má dvě části:* ***kritický obor*** *a* ***obor přijetí)***

Kritické hodnoty:- při zjišťování se bere v úvahu stanovená hladina významnosti α:

$$P(a_\alpha \leq T \leq b_\alpha) = 1 - \alpha$$

Leží-li testové kritérium v oboru přijetí, **nezamítneme** $\boldsymbol{H_0}$; rozdíl je vysvětlitelný na dané hladině významnosti náhodností výběru.

Leží-li testové kritérium v kritickém oboru, **zamítneme** $\boldsymbol{H_0}$ (přijmeme H_A); rozdíly považujeme za statisticky významné na zvolené hladině významnosti, tzn. nedají se vysvětlit pouze náhodností výběru.

Dosažená hladina významnosti: *p-hodnota (p_value, p_level)* – je pravděpodobnost, s jakou obdržíme naše data nebo data více extrémní (ještě více odporující nulové hypotéze), za předpokladu, že je nulová hypotéza pravdivá.

Postup vyhodnocení testu:

1. Zvolíme hladinu významnosti α – jistá mezní hodnota.
2. *p-hodnotu* porovnáme s α. Jestliže:
   a. $p < \alpha, H_0$ zamítneme; říkáme, že výsledek je statisticky významný.
   b. $p \geq \alpha, H_0$ nezamítneme; říkáme, že výsledek není statisticky významný. (pozorované rozdíly je možno vysvětlit pomocí náhody).

Testy hypotéz:

Jakých hodnot nabývají populační četnosti?
Jsou mezi nimi rozdíly?
Je správný předpoklad o mediánové kategorii?
Je v rozdělení modus?
Nejedná se o modus majoritní?

Jako úvod krátká prezentace a konkrétní příkald + vyřešení v SPSS.

Transform -> *Compute Variable* -> CDF & NonCentral DF -> cdf.T(-1.59,9)

Binomický test:

Př.: Analyzujte v závislosti na pohlaví jaký typ stravování respondenti na dovolené využívají:

Data -> Split File -> Compare Groups -> Jste?

Analyze -> Ferquencies -> Jakýtypstravovánípreferujete?

Jaký typ stravování preferujete?

		Frequency		Percent		Valid Percent		Cumulative Percent	
		Jste?		Jste?		Jste?		Jste?	
		muž	žena	muž	žena	muž	žena	muž	žena
Valid	all inclusive	17	15	37,8	26,3	37,8	26,3	37,8	26,3
	bez stravy	11	9	24,4	15,8	24,4	15,8	62,2	42,1
	plná penze	1	5	2,2	8,8	2,2	8,8	64,4	50,9
	polopenze	11	24	24,4	42,1	24,4	42,1	88,9	93,0
	snídaně	5	4	11,1	7,0	11,1	7,0	100,0	100,0
	Total	45	57	100,0	100,0	100,0	100,0		

Zaměřme se na plnou penzi, v souboru ji preferuje 5 žen a jeden muž.

a) Můžeme na základě takto malého počtu zamítnout hypotézu o shodě podílů pro dvě sledované kategorie?
b) Můžeme zamítnout $H_0: \pi_{1,0} = 0{,}4$ vůči levostranné alternativní hypotéze?
c) Můžeme zamítnout $H_0: \pi_{1,0} = 0{,}1$ vůči pravostranné alternativní hypotéze?

Ad a)
Náhodná veličina má binomické rozdělení s parametry $\pi_{0,1} = 0{,}5$, $n = 6$. Dále víme, že $n_1=1$ a $n_2=5$. Hladina významnosti se tedy spočte podle vztahu:

$$\alpha' = 2 \cdot \sum_{i=0}^{min\{n_1,n_2\}} \binom{n}{i} (0{,}5)^n = 2 \sum_{i=0}^{1} \binom{6}{i} (0{,}5)^6 = 2(1 \cdot 0{,}5^6 + 6 \cdot 0{,}5^6) = 2 \cdot (0{,}0156 + 0{,}9375) = 0{,}2187$$

Tato hladina významnosti je větší než *0,05* nemůžeme na hladině významnosti *5%* zamítnout nulovou hypotézu o shodě podílů.

Řešení v SPSS:
Musíme si připravit datový soubor:

- překódováním (*Transform, Recode into Different Variables* a pak jen vybráním konkrétních osob (*Data, Select cases, If condition is satisfied*)
- zapsáním výsledků do zvláštních sloupců (vytvoření nové proměnné)

Pro testování zadáme: *Analyze -> Nonparametric Test -> Binomial*

Binomial Test

Jste?			Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
muž	stravovani	Group 1	2	5	,83	,50	,219
		Group 2	1	1	,17		
		Total		6	1,00		

Získáváme absolutní (*N*) a relativní (*Observed Prop.*) četnosti pro obě skupiny (*Group 1*, *Group 2*) a celý soubor (*Total*), zadanou hodnotu $\pi_{1,0}$ (*Test Prop.*) a minimální hladinu významnosti od které zamítáme nulovou hypotézu H_0 (*Exact Sig (2-tailed)*.
Vypočtená hodnota je stejná jako hodnota vrácená pomocí SPSS, tzn. nemůžeme na hladině významnosti 5% zamítnout nulovou hypotézu o shodě podílů.

Ad b) zamítnout $H_0: \pi_{1,0} = 0{,}4$ vůči levostranné alternativní hypotéze ($H_A: \pi_1 < \pi_{1,0}$)

Vzorec pro ruční výpočet viz literatura: Hana Řezáková: Analýza dat z dotazníkových šetření, str.62, vzorec 3.16

Pro testování zadáme: *Analyze -> Nonparametric Test -> Binomial*
- *test proportion = 0,4*
- *cut point = 1*

Binomial Test

Jste?			Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
muž	stravovani	Group 1	<= 1	1	,2	,4	,233[a]
		Group 2	> 1	5	,8		
		Total		6	1,0		

a. Alternative hypothesis states that the proportion of cases in the first group < ,4.

Tato hladina významnosti je větší, než *0,05* nemůžeme na hladině významnosti *5%* zamítnout nulovou hypotézu o shodě podílů.

Ad c) Můžeme zamítnout $H_0: \pi_{1,0} = 0{,}1$ vůči pravostranné alternativní hypotéze? $H_A: \pi_1 > \pi_{1,0}$

Binomial Test

Jste?			Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
muž	stravovani	Group 1	<= 1	1	,2	,1	,469
		Group 2	> 1	5	,8		
		Total		6	1,0		

Tato hladina významnosti je větší, než *0,05* nemůžeme na hladině významnosti *5%* zamítnout nulovou hypotézu o shodě podílů vzhledem k pravostranné alternativě.

Chí kvadrát test dobré shody

Testujeme hypotézu H_0: $\pi_i = \pi_{i,0}$, kde ***i* = *1,2,…,K*** (K je počet kategorii) a $\sum \pi_{i,0} = 1$, vůči alternativní hypotéze H_A: H_0 neplatí. Pokud se konstanty $\pi_{i,0}$ rovnají, pak můžeme nulovou hypotézu vyjádřit jako H_0: $\pi_1 = \pi_2 = \cdots = \pi_K$. Pro $n\pi_{i,0} > 5$ se používá statistika chí-kvadrát daná vztahem:

$$\chi^2 = \sum_{i=1}^{K} \frac{(n_i - n\pi_{i,0})^2}{n\pi_{i,0}}$$

Kde $n\pi_{i,0}$ je teoretická (očekávané) obsazení i-té kategorit při výběru o rozsahu ***n***.
Za předpokladu, že platí H_0 , chí kvadrát rozdělení s *(K-1)* stupni volnost, tj. vypočtenou hodnotu porovnáváme s kvantilem $\chi^2_{1-\alpha}(K-1)$

Př.: Zaměřme se znovu na způsob stravování občanů ČR na dovolené. Podle záznamů cestovní kanceláře očekáváme, že asi 50% bude preferovat (all inclusive nebo plnou penzi), 10% bude preferovat dovolenou bez stravy a 40% bude jezdit na dovolené s polopenzí nebo jen se snídaní. Ověřme, zda jsou zjištěné údaje v souladu s naším předpokladem:

10% plná penze (6)

30% all in (32)

20% bez jídla (20)

30% polopenze (35)

10% snídaně (9)

Jaký typ stravování preferujete?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	all inclusive	32	31,4	31,4	31,4
	bez stravy	20	19,6	19,6	51,0
	plná penze	6	5,9	5,9	56,9
	polopenze	35	34,3	34,3	91,2
	snídaně	9	8,8	8,8	100,0
	Total	102	100,0	100,0	

$$\chi^2 = \frac{(6-10{,}2)^2}{10{,}2} + \frac{(32-30{,}6)^2}{30{,}6} + \frac{(20-20{,}4)^2}{20{,}4} + \frac{(35-30{,}6)^2}{30{,}6} + \frac{(9-10{,}2)^2}{10{,}2} = 2{,}57$$

Pomocí SPSS - *Transform* -> *Compute Variable* -> CDF & NonCentral DF - *CDF.CHISQ(2.57,4)*
Vrátí hodnotu 0,3679 – podle definice výše se hodnota chí kvadrát, ale počítá jako:
1 - α'=0,3679, tzn. $\alpha = 0{,}632$

Hodnota α je větší než 0,05 tedy na hladině významnosti 5% nezamítáme nulovou hypotézu o rozdělení způsobu stravování, můžeme tvrdit, že respondenti volí na dovolené způsob stravování podle předchozích záznamů CKí.

Postup pomocí SPSS

Nonparametric tests – Chi square – vybereme naší proměnnou (POZOR je třeba ji překódovat na numerickou) – zadáme ji do Test Variable List a dále v části Expectes Values vytvoříme seznam očekávaných četností (Zapisujeme jednotlivé hodnoty a přidáváme je do seznamu pomocí *Add*) – přidáváme po řadě tak jak jsme volili proměnné!!!

Výsledkem jsou dvě tabulky, první obsahuje zjištěné hodnoty (*Observed N*) a očekávané hodnoty (*Expected N*) a jejich rozdíly (R*esidulas*). Z druhé tabulky zjistíme hodnotu statistiky chí kvadrát (*Chi-Square*), počet stupňů volnosti (*df)* a minimální hladinu významnosti od které zamítáme hypotézu H_0 (*Asymp. Sig.*).

jidlo

	Observed N	Expected N	Residual
bez stravy	20	20,4	-,4
snídaně	9	10,2	-1,2
polopenze	35	30,6	4,4
plná penze	6	10,2	-4,2
all inclusive	32	30,6	1,4
Total	102		

Test Statistics

	jidlo
Chi-Square	2,575[a]
df	4
Asymp. Sig.	,631

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 10,2.

Znaménkové schéma odchylek
Provádíme tehdy, zamítáme-li H_0 o shodě zjištěných a očekávaných četností, pak můžeme vytvořit znaménkové schéma odchylek, přičemž pro $n \geq 30$ *a* $n\pi_{i.0} > 5$ využíváme aproximaci na normované normální rozdělení.
Vzorce a mezní hodnoty viz přednáška 4.

Příklad:
Na základě údajů získaných o populaci ve městě XY, kde probíhal průzkum na téma dovolená. Předpokládáme, že asi 50% respondentů je bezdětných, 25% má jedno dítě a 25% má 2-3 děti. Ověřme, jestli se tento předpoklad shoduje s našimi daty. Případně, které kategorie se nejvíce odlišují.

Řešení, nejprve je nutno proměnnou máte děti překódovat a dále postupujeme dle návodu uvedeném v příkladu výše:
Nonparametric tests – Chi square – vybereme naší proměnnou (POZOR je třeba ji překódovat na numerickou) – zadáme ji do Test Variable List a dále v části Expectes Values vytvoříme seznam očekávaných četností (Zapisujeme jednotlivé hodnoty a přidáváme je do seznamu pomocí *Add*) – přidáváme po řadě tak jak jsme volili proměnné!!!

deti

	Observed N	Expected N	Residual
nemám	72	51,0	21,0
ano - 1	13	25,5	-12,5
ano - 2-3	17	25,5	-8,5
Total	102		

Test Statistics

	deti
Chi-Square	17,608[a]
df	2
Asymp. Sig.	,000

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 25,5.

Hodnota p_value=0, jelikož je p_hodnota < α, (α=0,05) zamítáme nulovou hypotézu. Nelze tvrdit, že se skutečné a očekávané četnosti rovnají.

Nyní na základě uvedeného vzorce vypočítáme normovanou hodnotu e_i^* a porovnáme s kvantily:

Děti	$\mathbf{n_i}$	$\boldsymbol{n\pi_{i,0}}$	$\boldsymbol{1-\pi_{i,0}}$	$\boldsymbol{e_i^* = \frac{n_i - n\pi_{i,0}}{\sqrt{n\pi_{i,0}(1-\pi_{i,0})}}}$	**schéma**
Nemám	72	51	1-0,5	4,15	+++
Ano – 1	13	25,5	1-0,25	-2,86	--
Ano – 2-3	17	25,5	1-0,25	-1,94	0

Poslední kategorie přibližně odpovídá zjištěným údajům, kategorie s jedním je silně nižší než očekávané hodnoty a kategorie bezdětných má naopak velmi vyšší četnosti než jsou očekávány.