Analýza dat z dotazníkových šetření

Cvičení 3.

- Jednorozměrné třídění

Zdrojová data: dotazník http://www.vyplnto.cz/realizovane-pruzkumy/konzumace-ryb-a-rybich-vyrob/

- 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 minulé hodiny:

Vyhodnoťte poslední otázku: *Podle vašeho názoru je na českém trhu poptávka po rybách a rybích výrobcích….*

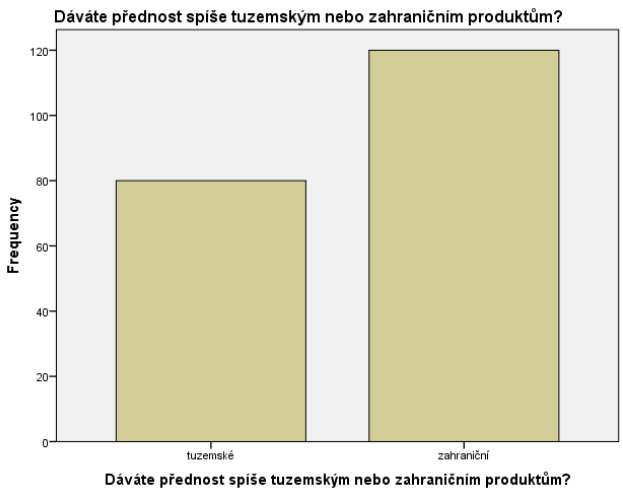
a) Obecně podle názoru respondentů.
b) Podle pohlaví mají stejný názor muži i ženy?
c) Proveďte vyvážení souboru podle pohlaví a proveďte znovu analýzu

Ad a)
Hodnoty správně seřadíme, recode into different variables, která odpověď je modus?
Modus = spíše nízká, 119 odpovědí, percent 59,5 – modální kategorie.

Vyhodnoťte otázku: *Dáváte přednost spíše tuzemským nebo zahraničním produktům?*
Co respondenti upřednostňují?
Jak toto souvisí s cenou za 1 kus ryby? (vyhodnoťte cenu 1 kusu ryby, otázka: *Jakou částku jste ochoten/ochotna za jeden kus ryby utratit?*, v závislosti zda se jedná o tuzemskou nebo zahraniční rybu)

Nejprve vyhodnocení tuzemská versus zahraniční:

Respondenti upřednostňují zahraniční ryby.



Souvislost s cenou ryby?

Dáváte přednost spíše tuzemským nebo zahraničním produktům?

		kusryby			
Valid	tuzemské	80	40,0	40,0	40,0
	zahraniční	120	60,0	60,0	100,0
	Total	200	100,0	100,0	

		Frequency		Percent	
		Dáváte přednost spíše tuzemským nebo zahraničním produktům?		Dáváte přednost spíše tuzemským nebo zahraničním produktům?	
		tuzemské	zahraniční	tuzemské	zahraniční
Valid	50,00	8	11	10,0	9,2
	75,00	6	10	7,5	8,3
	100,00	37	41	46,3	34,2
	150,00	20	31	25,0	25,8
	175,00	1	12	1,3	10,0
	200,00	8	15	10,0	12,5
	Total	80	120	100,0	100,0

Respondenti nejvíce kupují ryby v ceně do 100,-Kč/kus případně do 150,-kč/kus a nedělají rozdíl mezi tuzemskou nebo zahraniční rybou.

1. **Kategoriální proměnná nominální:** (Tabulka a graf četností)

- řádky tabulky představují jednotlivé třídy (kategorie)
- sloupce tabulky vyjadřují četnosti (počty jednotek)

Vraťme se k úloze z předešlého cvičení, na proměnnou jsme se dívali jako na nominální. Spočítejme jednotlivé charakteristiky vhodné pro nominální proměnnou.

Pokud ano, jak často?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	denně	1	,5	,5	,5
	jednou týdně	68	34,0	36,8	37,3
	jednou za měsíc	45	22,5	24,3	61,6
	několikrát týdně	33	16,5	17,8	79,5
	příležitostně	38	19,0	20,5	100,0
	Total	185	92,5	100,0	
Missing	0	15	7,5		
Total		200	100,0		

Jaký je modus u těchto proměnných? *(68)*
Určeme Giniho koeficient (nominální rozptyl): G_{nom}= nomvar = $\sum_{i=1}^{K}(p_i(1-p_i)) \cong \mathbf{0,78}$

Tabulka pomocných výpočtů:

i	p_i	$p_i(1-p_i)$	$p_i \ln p_i$

1	0,005	0,0049	-0,149
2	0,37	0,233	-0,368
3	0,24	0,182	-0343
4	0,18	0,145	-0,309
5	0,21	0,166	-0,328
Součet		0,734	-1,373

všech kategorií) (*ln 5 = 1,61*)

Normalizovaný nominální rozptyl:

$$norm.nomvar = \frac{K}{K-1}\sum_{i=1}^{K}\left(p_i(1-p_i)\right) = \frac{5}{4}\cdot 0{,}734 = 0{,}9175$$

Entropie: $H = -\sum_{i=1}^{k} p_i \ln p_i = -(-1{,}373) \cong 1{,}4$

- min. hodnota 0 (je-li v souboru zastoupena jen jedna kategorie)
- max. hodnota: *ln k* (v případě rovnoměrného zastoupení

Normalizovaná entropie: $H^* = \frac{H}{\ln k} = \frac{1{,}5}{1{,}61} = 0{,}931$

Ordinální proměnná

modální kategorie x_{MO}, mediánová kategorie x_{ME}

Mediánová kategorie je ta kategorie, v níž je dosaženo nebo poprvé překročeno postupným kumulováním relativních četností 50% rozsahu souboru.

Tzn.: $F_{ME-1} < 0{,}5; \quad F_{ME} \leq 0{,}5$

Př.: Určeme mediánovou kategorii pro proměnnou *Útrata za ryby* (z tabulky četností i graficky).

- Jaký podíl jednotek v této kategorii by při rozpůlení souboru náležel k jeho dolní a jaký k jeho horní části?

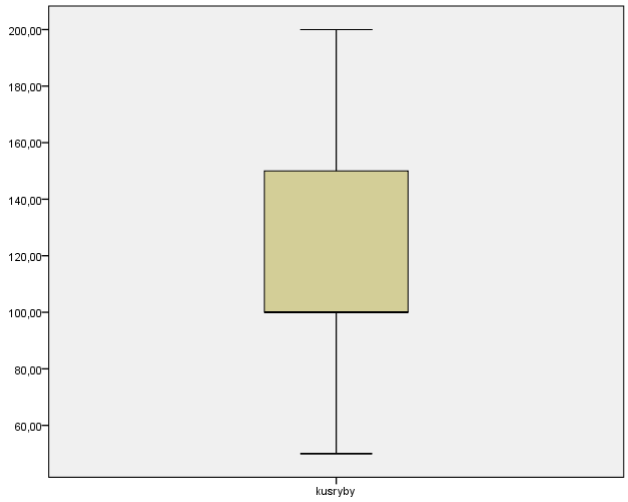
$$\frac{0{,}5 - F_{ME-1}}{p_{ME}} = \frac{0{,}5 - 0{,}175}{0{,}39} = 0{,}8333$$

P_{ME} ... relativní četnost mediánové kategorie

F_{ME-1}… kumulativní relativní četnost kategorie předcházejí mediánové

kusryby

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	50,00	19	9,5	9,5	9,5
	75,00	16	8,0	8,0	17,5
	100,00	78	39,0	39,0	56,5
	150,00	51	25,5	25,5	82,0
	175,00	13	6,5	6,5	88,5
	200,00	23	11,5	11,5	100,0
	Total	200	100,0	100,0	



Ordinální rozptyl (diskrétní ordinální variance, Giniho průměrná diference pořadí)

$$G_{ord} = dorvar = 2 \cdot \sum_{i=1}^{k} F_i(1 - F_i) = 2 \cdot 0{,}726 = 1{,}452$$

Nabývá maxima, právě když u 50% objektů nabývá sledovaná proměnná hodnoty $\boldsymbol{x_1}$ a u zbylých 50% objektů hodnoty $\boldsymbol{x_k}$.

Min. hodnota: 0 (je-li v souboru zastoupena pouze jediná kategorie
Max. hodnota: $\frac{k-1}{2}$ (pokud jsou v souboru zastoupeny pouze krajní hodnoty, a to vždy z jedné poloviny)
Normalizovaný ordinální rozptyl (nabývá hodnot z intervalu od 0 do 1):

$$norm.dorvar = \frac{2}{k-1} \sum_{i=1}^{k} F_i(1 - F_i) = 2 \cdot \frac{1{,}452}{4} = 0{,}726$$

Co nám vyšlo? Kdy mohu s ordinálními proměnnými zacházet jako s měřitelnými?

Tabulka pomocných výpočtů:

i	**F_i**	**$F_i(1-F_i)$**
1	0,095	0,086
2	0,175	0,144
3	0,565	0,246
4	0,82	0,148
5	0,885	0,102
Součet		0,726

Nominální proměnná více hodnotová odpověď.

- Podívejme se do dotazníku na otázku: ***Které faktory Vás nejvíce ovlivňují při koupi ryby?***
  Je zde 5 možností odpovědi, každá pro jiný sloupec: cena, kvalita, původ, druh, jiné

Jak tyto data vyhodnotit? Tímto se zabývá *analýza vícehodnotových odpovědí (Multiple Response Analysis)*. Podle typu proměnných máme dva přístupy:

- **Dichotomické proměnné** (odpověď na otázku jen ano/ne)
- **Vícekategoriální proměnné** (respondent si vybírá, z možné škály odpovědí (cena, kvalita, druh…))

SPSS je možno rovnocenně zpracovat oba přístupy, ale data musí být vždy v numerické podobě.

Př. Proveďme vyhodnocení pro otázku: ***Které faktory Vás nejvíce ovlivňují při koupi ryby?***

1. Překódování proměnných (např. 1-cena, 2-kvalita, 3původ, 4-druh, 5-jiné)
2. Vytvoření balíku odpovědí: *Analyze* -> *Multiple Response* -> *Define Variable Sets* vyberu balík, který chci vyhodnocovat (již překódovaný musí být **NUMERIC**), zaškrtnu *Categories* a vyplním 1 až 5.
3. Pojmenuju mojí novou proměnnou: např. **Faktor_koupe** a přidám proměnnou *$faktor_koupe*
4. Zavřu okno a pokračuji dále, znovu obdobný postup: *Analyze* -> *Multiple Response* -> *Frequencies* a vyberu naší proměnnou

Který faktor ovlivňuje respondenty nejvíce a který naopak nejméně?

Př.: Proveďme vyhodnocení pro otázku: ***Ze kterých důvodů podle vás nemají někteří lidé o ryby zájem?***

1. Překódování proměnných na dichotomické (když je odpověď tak 1 jinak 0) nebo přes *Automatic Recode* (pak pozor jak se překóduje odpověď v našem případě, když tam odpověď je tak číslo 2)
2. Vytvoření balíku odpovědí: *Analyze* -> *Multiple Response* -> *Define Variable Sets* vyberu balík, který chci vyhodnocovat (již překódovaný musí být **NUMERIC**), zaškrtnu *Dichotomies* a vyplním 2 (pokud používám Automatic Recode)
3. Pojmenuju mojí novou proměnnou
4. Zavřu okno a pokračuji dále, znovu obdobný postup: *Analyze* -> *Multiple Response* -> *Frequencies* a vyberu naší proměnnou

Ze kterých důvodů si lidé myslí, že ryby nenakupují a jaké důvody je ovlivňují nejméně?

Př.: Vyhodnoťte otázku „***Poslední otázka: Podle vašeho názoru je na českém trhu poptávka po rybách a rybích výrobcích:“***, přistupujte k ní jako k ordinální proměnné.

Abychom mohli přistupovat k odpovědím jako k ordinálním, musíme nejprve vhodně seřadit. Pozor na odpověď *„nevím“,* podle struktury otázky nebo vhodnosti odpovědí je, buď do odpovědí zařadíme, jako „únikovou“ možnost nebo můžeme „vyhodit“ jako *missing*. Vždy záleží na konkrétním dotazníku a na požadavcích na zpracování, podobně se zachází i s možností „*jiné“*.
V této analýze je zařadíme doprostřed možností a budeme k ní přistupovat jako k plnohodnotné odpovědi.

potávka ryby

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	velmi nízká	10	5,0	5,0	5,0
	spíše nízká	119	59,5	59,5	64,5
	nízká	37	18,5	18,5	83,0
	nevím	17	8,5	8,5	91,5
	5	13	6,5	6,5	98,0
	6	2	1,0	1,0	99,0
	7	2	1,0	1,0	100,0
	Total	200	100,0	100,0	

Podle názorů respondentů je poptávka po rybách spíše nízká 119; 59,5% - mediánová kategorie

Modální kategorie = spíše nízká

Ani při rozdělení souboru na muže a ženy nejsou rozdíly nijak velké. Stále je modus spíše nízká, i modální kategorie zůstává stejná.

potávka ryby

		Frequency		Percent		Cumulative Percent	
		Vaše pohlaví:		Vaše pohlaví:		Vaše pohlaví:	
		muž	žena	muž	žena	muž	žena
Valid	velmi nízká	1	9	2,4	5,7	2,4	5,7
	spíše nízká	27	92	64,3	58,2	66,7	63,9
	nízká	5	32	11,9	20,3	78,6	84,2
	nevím	4	13	9,5	8,2	88,1	92,4
	5	3	10	7,1	6,3	95,2	98,7
	6	1	1	2,4	,6	97,6	99,4
	7	1	1	2,4	,6	100,0	100,0
	Total	42	158	100,0	100,0		

Další úkoly a příklady k procvičení:

- Vypočtěte jednotlivé charakteristiky pro proměnnou poptávka ryby z předchozího příkladu, zkuste se na proměnnou dívat jako na nominální a potom na ordinální. (*G_{nom}, norm.var, Entropie, normalizovaná entropie, G_{ord}, norm.dorvar*)
- Zopakujte si číselné a grafické charakteristiky numerické proměnné (rozptyl, odchylka, kvartily, kvantily, boxplot)