Metodologie pro Informační studia a knihovnictví 2

Modul 4: Kódování a rekódování. Deskriptivní statistika – popis dat I

Co se dozvíte v tomto modulu?

- Co zjišťujeme u nominálních proměnných?
- Co zjišťujeme u ordinálních proměnných?
- Jak zjistit modus a medián?
- Jak popsat grafy?

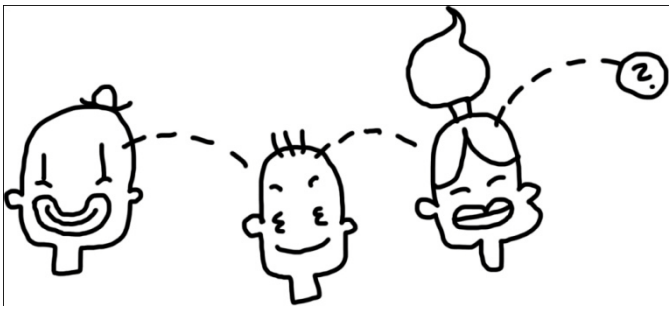
V tomto modulu si ukážeme, jak popsat kategorizovaná data.

Obsah

1 Kódování a rekódování

Náš dataset obsahuje kódy odpovědí. Tedy například místo odpovědi „velmi spokojen/a“ je uveden kód odpovědi „1“. Kódy jsou užitečné z nejen pro statistické softwary a aplikace, ale i pro snížení chybovosti při zápisu dat. Například pokud kombinujete sběr dat online a offline, je výhodné si nechat z online aplikace vygenerovat pouze číselné kódy a odpovědi sesbírané v terénu „doťukat“ do tabulky ve formě kódů. Výrazně se tím šetří čas, i pokud budete dále zpracovávat data pouze v Excelu.

Další práce s datasetem se liší, pokud pracujete v Excelu a ve statistickém software – například v SPSS.



Práce v Excelu

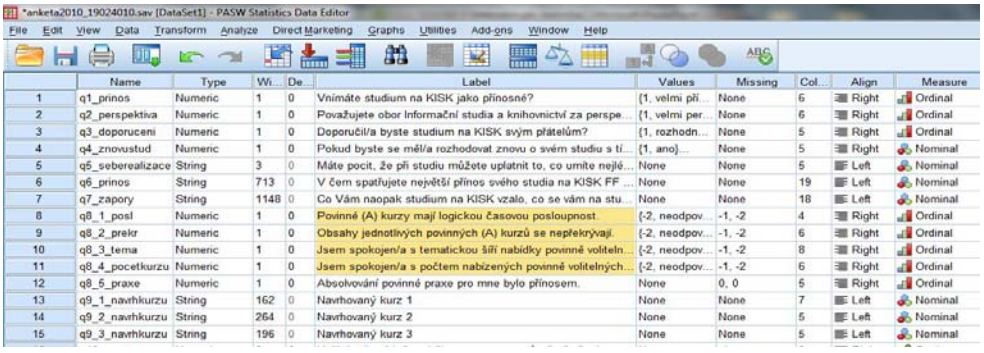
Pokud pracujete v Excelu, je výhodné si data opět překódovat tak, aby se vám ve výsledných tabulkách opět objevovaly celé odpovědi, nikoliv jen kódy. Jednoduše to uděláte pomocí funkce „Najít a nahradit“ (CTRL+H).

> *Protože takovou práci s dokumenty určitě umíte, dataset nemusíte celý překódovat, ale budete jej mít nahraný v ISu již překódovaný.*

Pokud se rozhodnete pracovat s okódovaným souborem, nezapomeňte ve výsledných tabulkách přepsat kódy odpovědí na skutečné odpovědi.

Práce v SPSS

Pro práci v SPSS má smysl kódy ponechat. SPSS pracuje přímo s kódy, kterým přiřazujete popisky („labels“). Jednoduše přepínáte mezi zobrazením responsí a zobrazením proměnných a jejich popisu.


*anketa2010_19024010.sav [DataSet1] - PASW Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Wi...	De...	Label	Values	Missing	Col...	Align	Measure
1	q1_prinos	Numeric	1	0	Vnímáte studium na KISK jako přínosné?	{1, velmi při...	None	6	Right	Ordinal
2	q2_perspektiva	Numeric	1	0	Považujete obor Informační studia a knihovnictví za perspe...	{1, velmi per...	None	6	Right	Ordinal
3	q3_doporuceni	Numeric	1	0	Doporučil/a byste studium na KISK svým přátelům?	{1, rozhodn...	None	5	Right	Ordinal
4	q4_znovustud	Numeric	1	0	Pokud byste se měl/a rozhodovat znovu o svém studiu s tí...	{1, ano}...	None	5	Right	Nominal
5	q5_seberealizace	String	3	0	Máte pocit, že při studiu můžete uplatnit to, co umíte nejlé...	None	None	5	Left	Nominal
6	q6_prinos	String	713	0	V čem spatřujete největší přínos svého studia na KISK FF ...	None	None	19	Left	Nominal
7	q7_zapory	String	1148	0	Co Vám naopak studium na KISK vzalo, co se vám na stu...	None	None	18	Left	Nominal
8	q8_1_posl	Numeric	1	0	Povinné (A) kurzy mají logickou časovou posloupnost.	{-2, neodpov...	-1, -2	4	Right	Ordinal
9	q8_2_prekr	Numeric	1	0	Obsahy jednotlivých povinných (A) kurzů se nepřekrývají.	{-2, neodpov...	-1, -2	6	Right	Ordinal
10	q8_3_tema	Numeric	1	0	Jsem spokojen/a s tematickou šíří nabídky povinně voliteln...	{-2, neodpov...	-1, -2	8	Right	Ordinal
11	q8_4_pocetkurzu	Numeric	1	0	Jsem spokojen/a s počtem nabízených povinně volitelných...	{-2, neodpov...	-1, -2	6	Right	Ordinal
12	q8_5_praxe	Numeric	1	0	Absolvování povinné praxe pro mne bylo přínosem.	None	0, 0	5	Right	Ordinal
13	q9_1_navrhkurzu	String	162	0	Navrhovaný kurz 1	None	None	7	Left	Nominal
14	q9_2_navrhkurzu	String	264	0	Navrhovaný kurz 2	None	None	5	Left	Nominal
15	q9_3_navrhkurzu	String	196	0	Navrhovaný kurz 3	None	None	5	Left	Nominal

Odlišná je i práce s tzv. „missing values“ (chybějícími hodnotami. Zatímco při práci v SPSS nebo statistických softwarech je vhodné je okódovat odlišným způsobem (např. -1 nebo 99) a programu „říci“, že se jedná o chybějící hodnoty, se kterými nemá počítat, při práci v Excelu můžeme ponechat políčka volná, případně i ponechat popisy typu „Neví/neodpověděl“. To, že se tyto hodnoty nezahrnují do analýzy, označujeme až při samotné tvorbě tabulky četností (viz předchozí týden).

2 Deskriptivní statistika (kategorizované proměnné)

Nejprve malé opakování:

- **Deskriptivní statistika** se zabývá popisem dat, jejich sumarizaci a prezentací.
- **Kategorizované proměnné** jsou všechny proměnné, jejichž hodnoty se nacházejí v určitých kategoriích. Jedná se tedy o nominální, ordinální i kardinální proměnné (pouze ale kardinální poměrové).

Různé druhy proměnných umožňují různé druhy popisu.

Popis nominálních proměnných

U nominálních proměnných zjišťujeme:

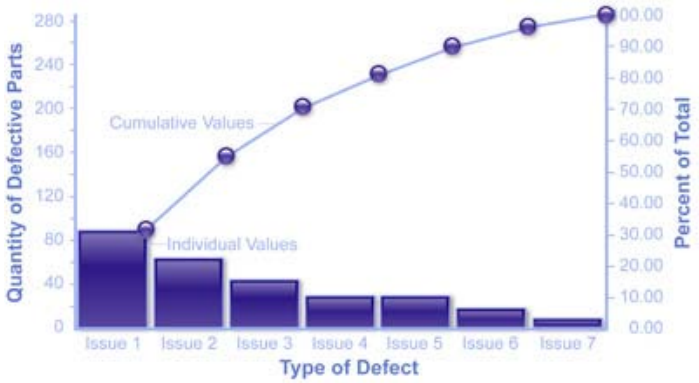
- **rozložení četností** variant znaku (pomocí tabulek četností),
- nejčastěji zastoupenou kategorii – **modus** (modálních kategorií někdy může být více než 1),
- **variační poměr**, který se vypočítá tak, že od jedné odečteme podíl četnosti modální kategorie a velikosti souboru.

Rozložení nominální proměnné můžeme – je-li to vhodné – znázornit i tzv. **Paretovým diagramem**. Paretův diagram (nebo také Paretův graf) kombinuje sloupcový a čárový graf. Sloupci jsou vyznačené četnosti jednotlivých kategorií seřazené podle velikosti, čarou je vyznačená kumulativní četnost. Paretův graf se využívá ve strategickém rozhodování a jako nástroj zlepšování kvality – dokáže velmi účinně zvýraznit důležité kategorie od nedůležitých – tzv. „vital few“ vs. „trivial many“ (Levine & Stephan 2010)

Paretův graf získáme v Excelu z této tabulky:

	Četnost	Kumulativní relativní četnost
Položka A		
Položka B		
Položka C		
Položka D		

Příklad Paretova diagramu:

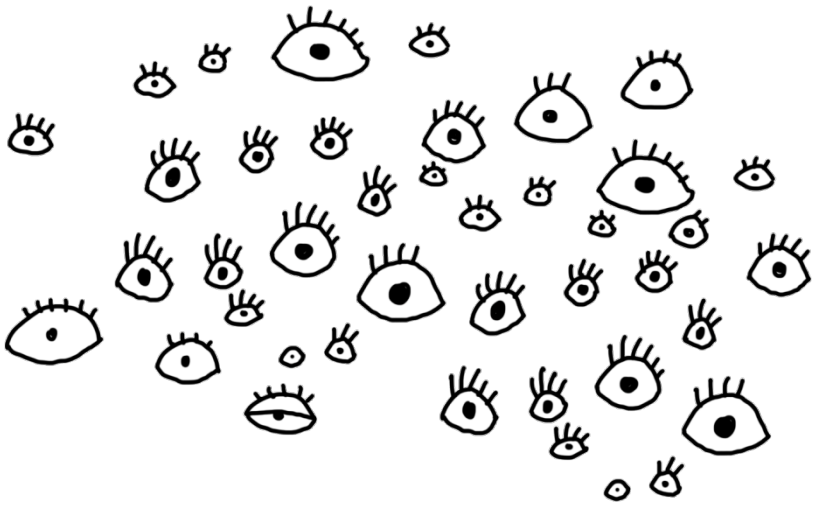


(zdroj: http://www.billiondollargraphics.com/paretochart.html)

Popis ordinálních proměnných

U ordinálních proměnných zjišťujeme:

- rozložení četností variant znaku (pomocí tabulek četností),
- nejčastěji zastoupenou kategorii – **modus** (modálních kategorií někdy může být více než 1),
- **medián** (mediánovou kategorii),
- variační poměr,
- další vlastnosti, jako je ordinální variance či normalizovaná ordinální variance (dorvar – těmi se ale nebudeme dopodrobna zabývat).



Rozložení četností

Zjištění rozložení četností je základní operací popisné statistiky. Ukázali jsme si jej už v minulém modulu. Při popisu rozložení četností vytvoříme vždy:

- tabulku četností,
- graf četností (koláčový či sloupcový).

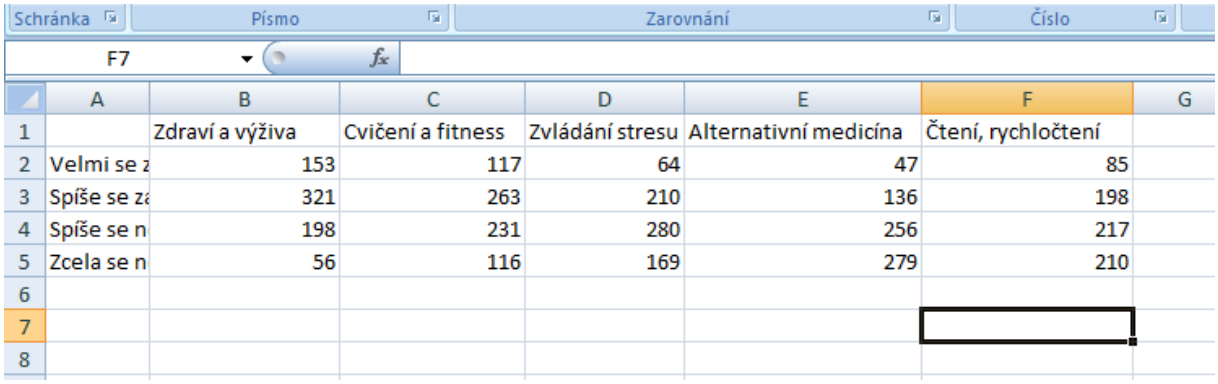
V grafu i v tabulce četností pracujeme vždy s validními četnostmi (tedy nezahrnujeme odpovědi typu „nevím“ nebo „*neodpověděl/a*“.

V případě nominálních proměnných je pro přehlednost vhodné kategorie ve sloupcovém diagramu seřadit dle výskytu od největší po nejmenší.

Porovnání rozložení četností

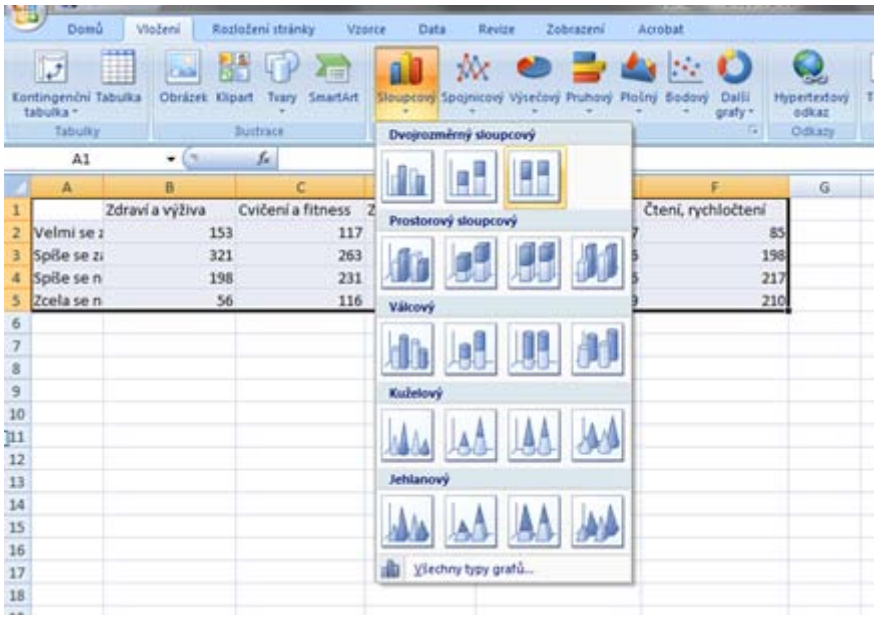
Pro zobrazení porovnání rozložení četností u baterií otázek se používají **skládané sloupcové grafy**.

Skládaný sloupcový graf můžete vytvořit tak, že si připravíte tabulku s absolutními validními četnostmi u jednotlivých kategorií:

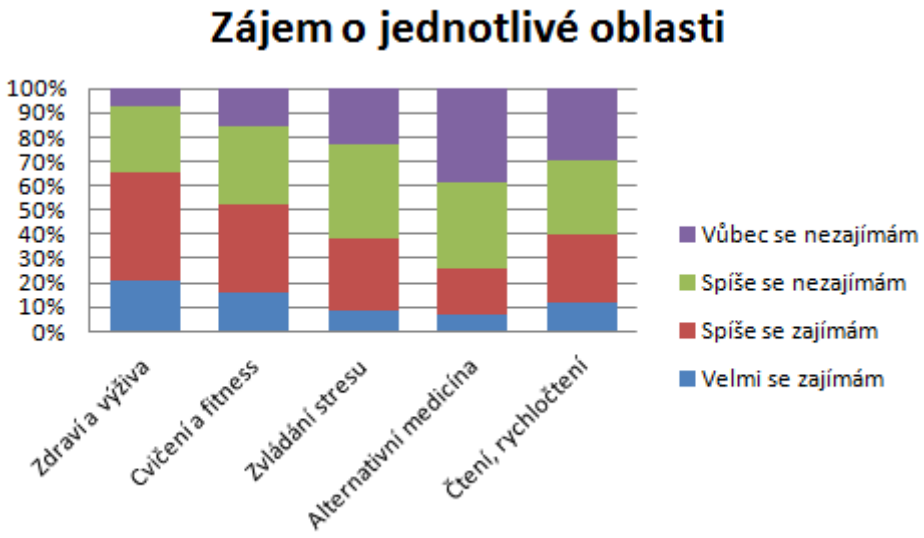


	A	B	C	D	E	F	G
1		Zdraví a výživa	Cvičení a fitness	Zvládání stresu	Alternativní medicína	Čtení, rychločtení	
2	Velmi se z	153	117	64	47	85	
3	Spíše se za	321	263	210	136	198	
4	Spíše se n	198	231	280	256	217	
5	Zcela se n	56	116	169	279	210	
6							
7							
8							

Tabulku si označíte a zvolíte možnost „Vložení“ – „Grafy“ – „Sloupcový“.



Výsledkem je skládaný sloupcový graf, který přehledně ukazuje rozdíly v rozložení jednotlivých proměnných.



Modus a medián

Pro připomenutí z minulého semestru si uveďme, v čem se liší MODUS a MEDIÁN (obě udávají tzv. míry centrální tendence a často se pletou):

> ***MODUS*** *je hodnota, která se v datech vyskytuje nejčastěji.*
> ***MODÁLNÍ KATEGORIE*** *je tedy nejpočetněji zastoupená kategorie.*
>
> ***MEDIÁN*** *dělí řadu výsledků seřazených podle velikosti na dvě stejně početné poloviny.*
> ***MEDIÁNOVÁ KATEGORIE*** *je ta, ve které je dosaženo 50% všech údajů, postupujeme-li od první kategorie výše.*

Jestliže je počet položek ve výzkumném souboru lichý, pak platí:

$$\textbf{Medián} = x_{(n+1)/2}$$

Jestliže je počet položek ve výzkumném souboru sudý, pak platí:

$$\textbf{Medián} = 0{,}5(x_{n/2}+x_{n/2+1})$$

Představte si otázku na počet dětí. Odpovědi respondentů jsou {0,1,1,2,2,3,5}.

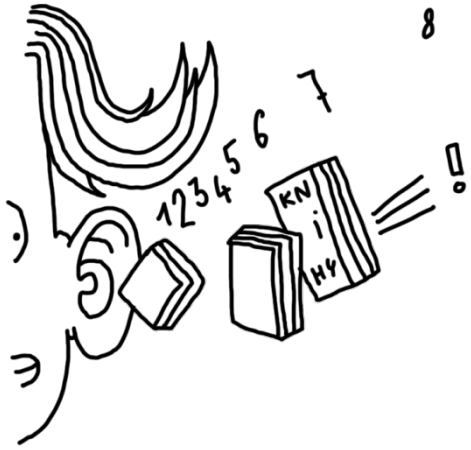
- *V souboru jsou dvě modální kategorie (tedy kategorie s nejvyšším počtem výskytů) – jsou to hodnoty 1 a 2.*
- *Mediánová kategorie je 2. Medián je na rozdíl od aritmetického průměru málo citlivý k odlehlým (extrémním) hodnotám. Pokud by byly odpovědi respondentů {0,1,1,2,2,3,5,10}, medián stale zůstává roven 2.*

Modus a medián v Excelu

V Excelu existují na výpočet mediánu a modu jednoduché příkazy MEDIAN a MODE. Syntaxe zápisu je snadná:

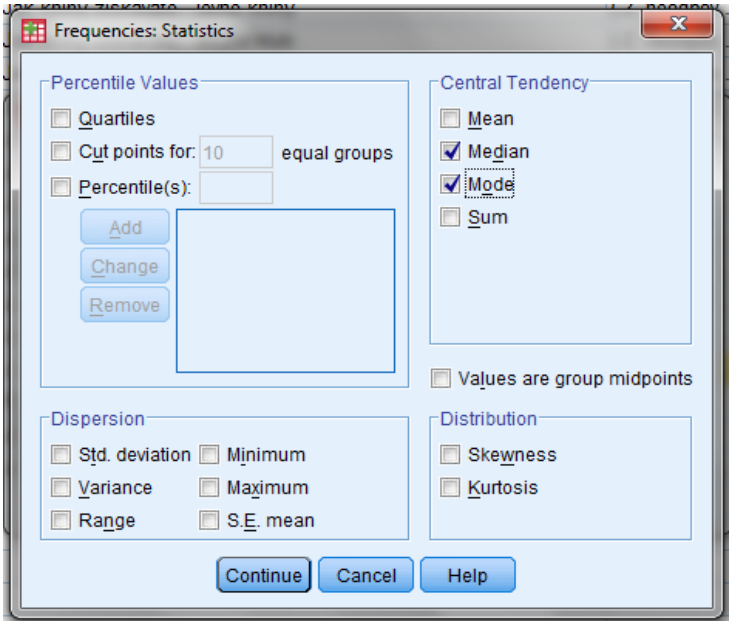
- =MEDIAN(datová oblast) – např. =MEDIAN(A1:A730)
- =MODE(datová oblast) – např. =MODE(A1:A730)

(Příkazy vypočítají medián a modus ze sloupce A, řádků 1-730.)



Modus a medián v SPSS

V SPSS vyberete v nabídce položky Analyze > Descriptive Statistics > Frequencies (zde zvolíte proměnnou) > Statistics > Median, Mode.



3 Tipy pro vytváření grafů

Levine a Stephan (2010) shrnují několik tipů pro prezentaci dat prostřednictvím grafů v akademickém prostředí:

- vždy si vyberte ten nejjednodušší graf,
- vždy používejte popisek grafu,
- popište obě osy,
- vyvarujte se ilustrací a zbytečného používání grafiky na pozadí nebo okrajích grafu,
- vyvarujte se používání módních piktogramů, které by mohly ztížit čitelnost dat,
- vertikální osa by měla začínat nulou (pokud nezačíná negativními hodnotami).

V neakademickém prostředí (např. pro účely marketingu) je využití grafiky vhodné, v prostředí akademickém je na prvním místě čitelnost dat. 3D efekty a vkládání obrázků mohou znemožnit čtení hodnot dat. Další tipy pro vytváření grafů najdete třeba zde.

Literatura

Hendl, J. *Přehled statistických metod analýzy dat*. Praha : Portál 2009

Levine, D. M., & Stephan, D. (2010). *Even you can learn statistics: A guide for everyone who has ever been afraid of statistics*. Upper Saddle River, N.J: FT Press.