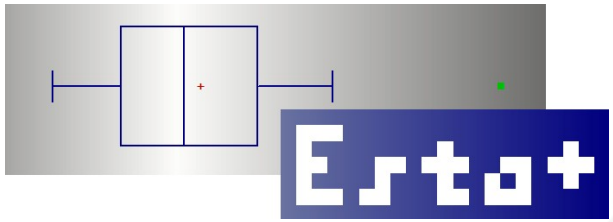
Metodologie ISK

Základy statistického zpracování dat

Ladislava Suchá, 28. dubna 2011

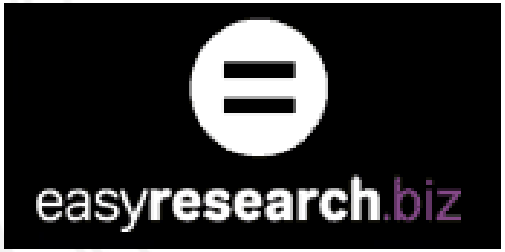
Programy na statistické zpracování dat

SPSS®



Aplikace na online dotazování, které zvládají nějkteré základní i složitější statistické operace

SurveyMonkey.com
because knowledge is everything



Fáze vyhodnocování dat

- Kódování
- Třídění prvního stupně
- (Úpravy znaků)
- Třídění druhého stupně

Kódování

- Jednotlivým variantám znaku jsou přiřazovány symboly (čísla) podle kódovacího čísla
- Kódování často probíhá přímo v terénu nebo ho provádí aplikace

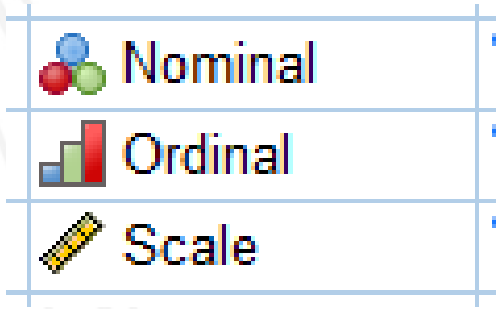
Zápis do matice dat:

- Jednotlivé případy = řádky
- Jednotlivé proměnné = sloupce

Druhy proměnných

Nominální

- Známe hodnoty, ale můžeme o nich říci pouze to, že jsou různé
- Nelze provádět aritmetické operace
- Přiřazení znaku je symbolické



Pořadové

- Můžeme určit pořadí (vzdělání, spokojenost)
- Znaky = míra pořadovosti

Kardinální (intervalové, spojité)

- Můžeme říci, o kolik je jedna hodnota vyšší než druhá (měsíční příjem, počet dětí v domácnosti atd.)
- Přiřazení znaku = reálné číslo

Otázka v dotazníku a její zpracování

Považujete obor Informační studia a knihovnictví za perspektivní?

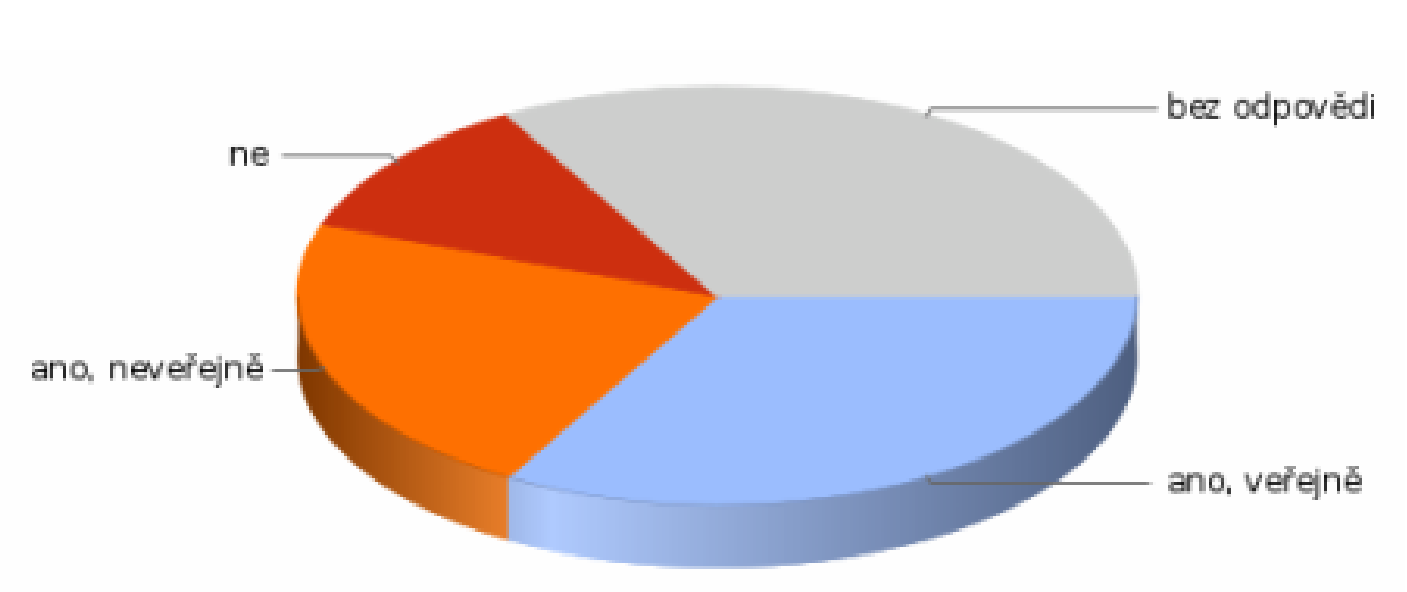
velmi perspektivní	1
spíše perspektivní	2
spíše neperspektivní	3
zcela neperspektivní	4
nevím, nemohu odpovědět	-1
neodpověděl/a	-2

Hodnoty proměnné

Zapisujeme jako „value labels“

Chybějící hodnoty (missing values)

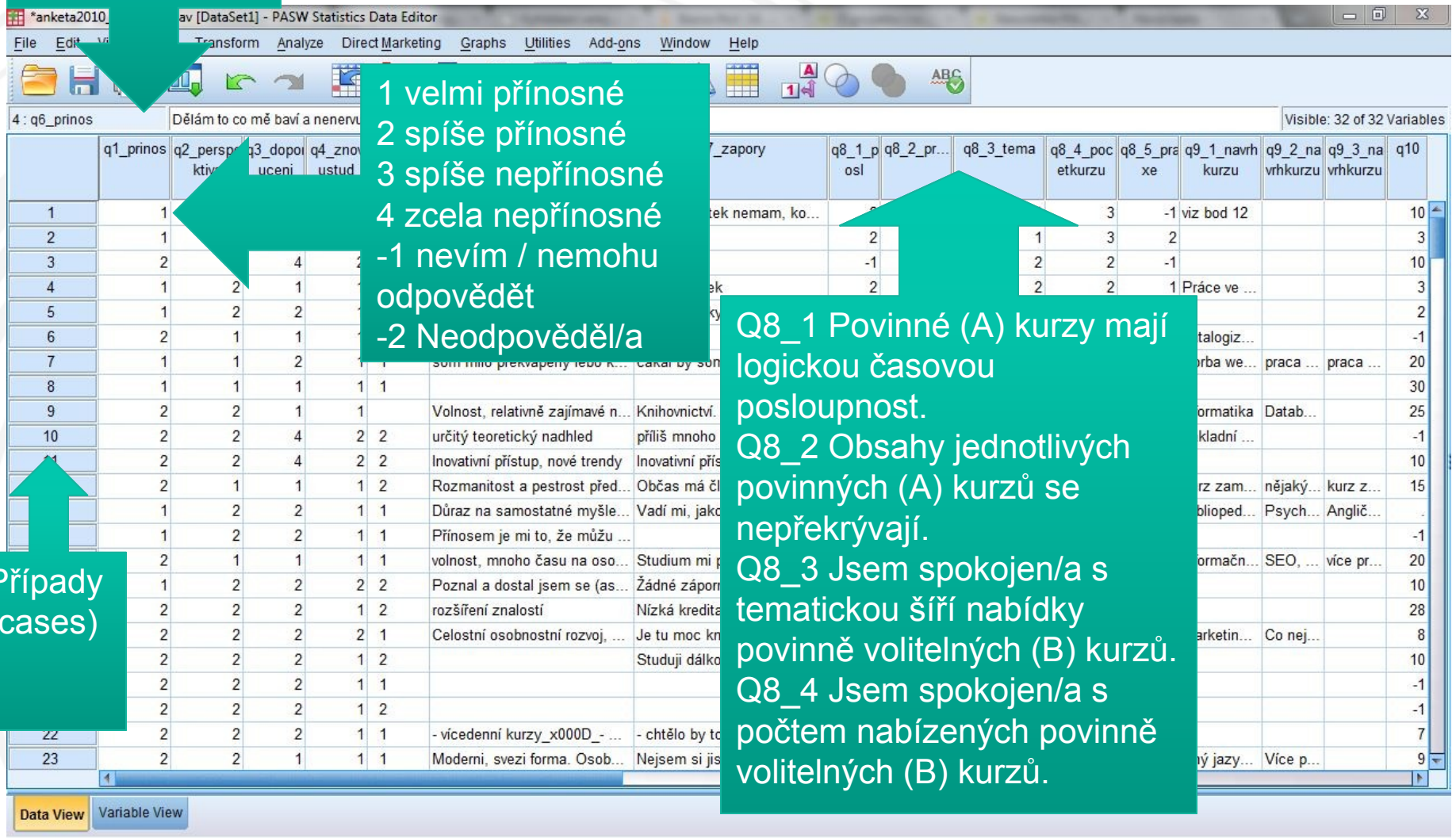
Ukázka – zahrnutí missing values (chyba)



Graf 1: V jakém rozsahu poskytovatelé uchovávají data o uživatelích svých služeb

Q1_prinos Studium na KISK hodnotím jako:

1 velmi přínosné
2 spíše přínosné
3 spíše nepřínosné
4 zcela nepřínosné
-1 nevím / nemohu odpovědět
-2 Neodpověděl/a



Případy (cases)

Q8_1 Povinné (A) kurzy mají logickou časovou posloupnost.
Q8_2 Obsahy jednotlivých povinných (A) kurzů se nepřekrývají.
Q8_3 Jsem spokojen/a s tematickou šíří nabídky povinně volitelných (B) kurzů.
Q8_4 Jsem spokojen/a s počtem nabízených povinně volitelných (B) kurzů.

Definování proměnných

Druh proměnné

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

Zápis baterie otázek

Třídění prvního stupně

- Sleduje se četnost výskytu jednotlivého znaku
  - Kolik je v souboru mužů a žen
  - Kolik je v souboru lidí, kteří chodí do knihovny atd…

- Sledujeme základní **statistické míry znaků**

Třídění prvního stupně

Absolutní četnosti

- Absolutní číslo – kolik případů má danou vlastnost
- Součet absolutních četností u všech hodnot (včetně missing values) = celkový počet respondentů

(V souboru je 71 žen.)

Relativní četnosti

- Jaký podíl (v procentech z výběrového souboru) představují případy s jednotlivou vlastností

(V souboru je 34 % osob se středoškolským vzděláním.)

Kumulativní relativní četnosti

(V souboru je 52 % osob s alespoň středoškolským vzděláním.)

Rozložení hodnot proměnných

Statistics

Považujete za dostačující komunikaci KISKu se studenty?

N	Valid	442
	Missing	3

„Missing“ do grafů nezařazujeme

(počítáme jen s těmi, kteří odpověděli)

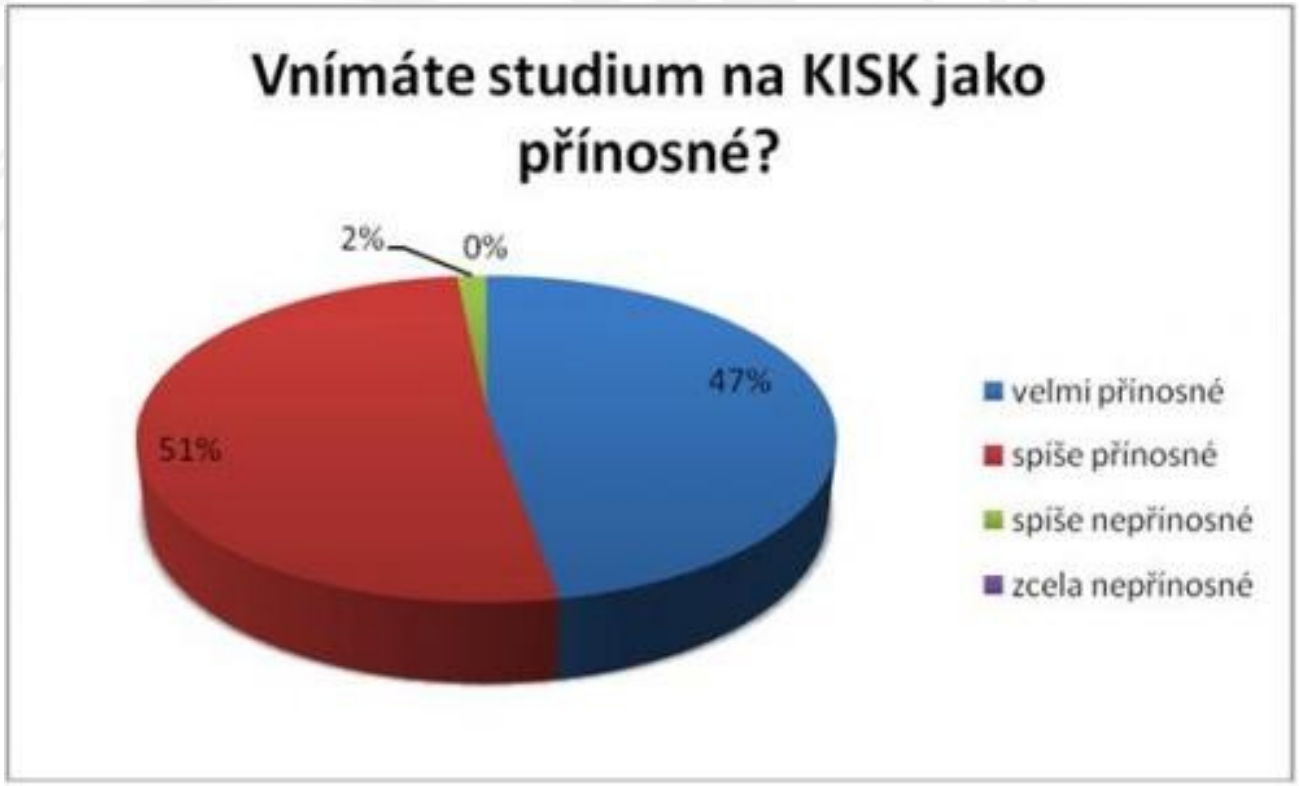
Relativní četnosti bez „missing values“

Považujete za dostačující komunikaci KISKu se studenty?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ano	390	87,6	88,2	88,2
	s výhradami, jak kteří vyučující	22	4,9	5,0	93,2
	ne	30	6,7	6,8	100,0
	Total	442	99,3	100,0	
Missing	neodpověděli	3	,7		
Total		445	100,0		

Zobrazování výsledků

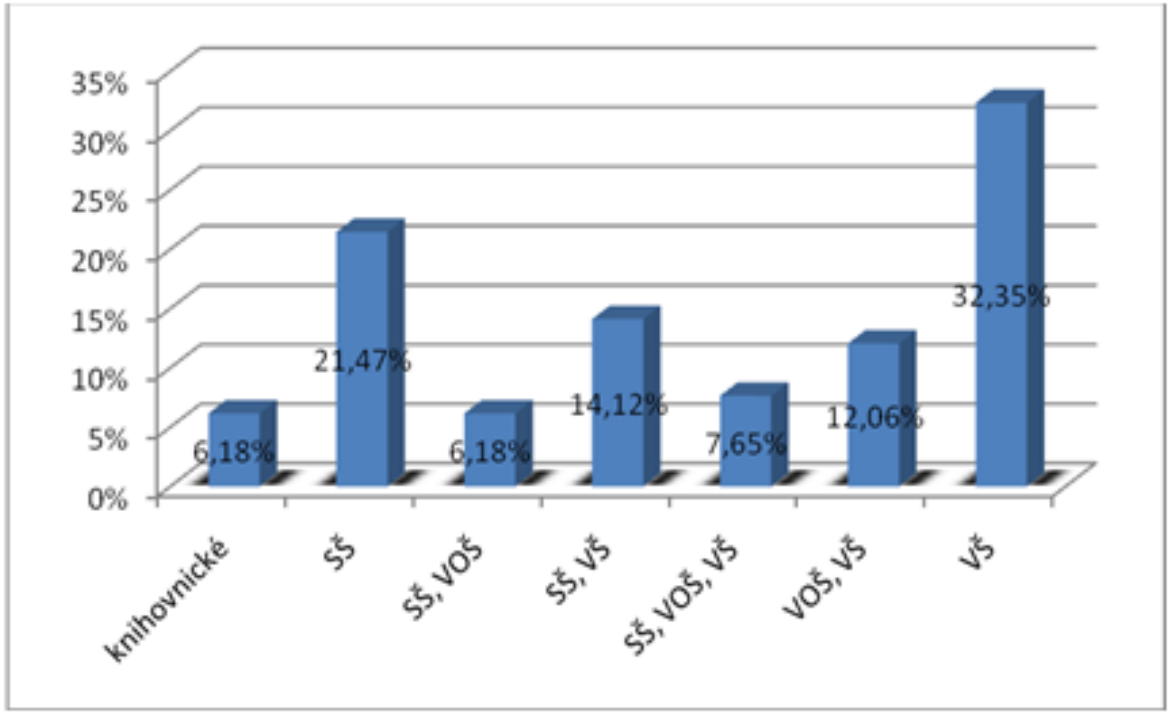
Koláčové, sloupcové grafy



Zobrazování výsledků

Koláčové, sloupcové grafy

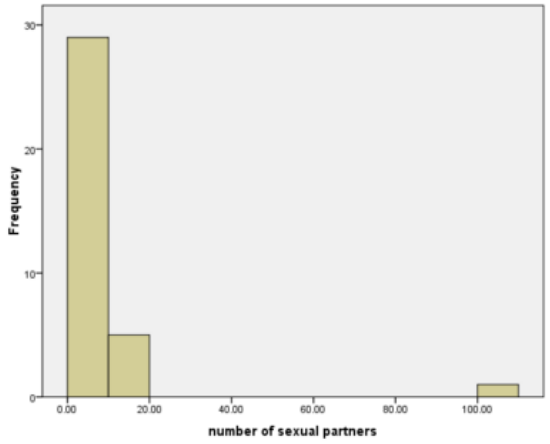
Graf č. 4 – Požadované vzdělání knihovníků

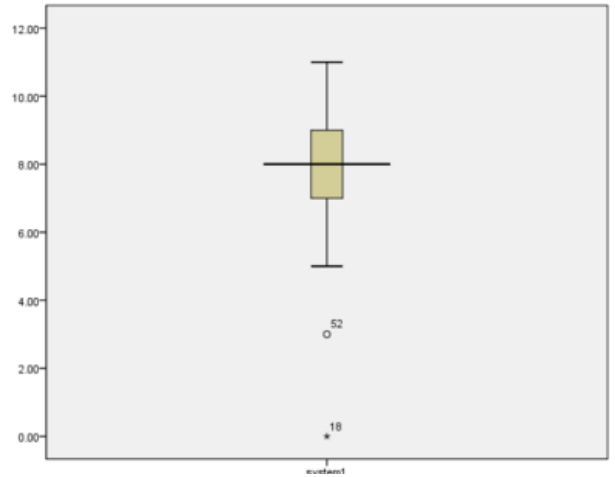


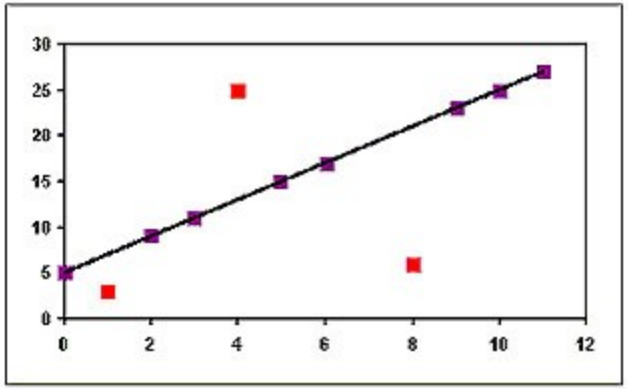
Deskriptvní statistika a čištění dat

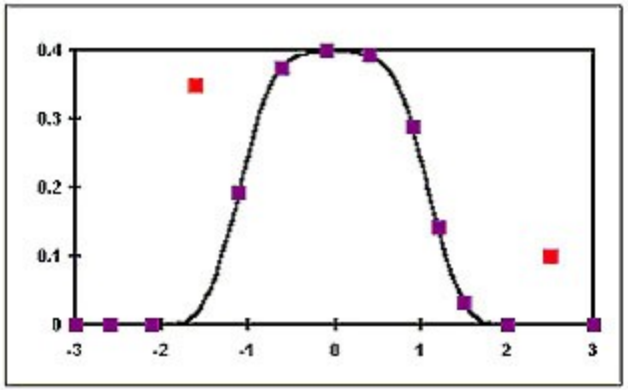
- První krok při každém zpracování dat
- „GIGO“ *(Garbage in, garbage out)*

- **Outliers** (extrémní hodnoty)
  - Podíváme se na nejvyšší a nejnižší hodnoty
    - (SPSS najde automaticky)









Zdroje obrázků: Psychwiki, www.ibm.com

Charakteristiky rozložení proměnné: modus, medián, průměr

MODUS

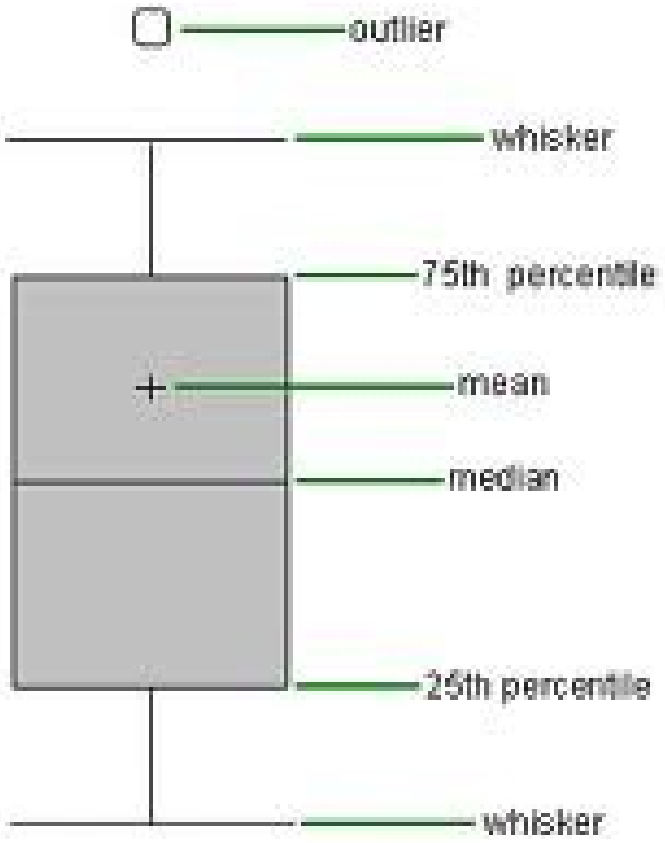
- U **nominálních proměnných**
- Nejčastěji obsazená kategorie/hodnota proměnné

MEDIÁN

- U **nominálních a ordinálních (pořadových) proměnných**
- Nejméně 50 % hodnot je menších nebo rovno mediánu a nejméně 50 % hodnot je větších nebo rovných mediánu
- Není ovlivněn extrémními hodnotami
- Pokud má soubor sudý počet prvků, dvě varianty (rozdílný výklad):
  - za medián označuje aritmetický průměr hodnot na místech $n/2$ a $n/2+1$
  - Medián nelze určit

Charakteristiky rozložení proměnné: modus, medián, průměr

- Medián = kvantil
- Kvartil
- Decil
- Percentil



Charakteristiky rozložení proměnné: modus, medián, průměr

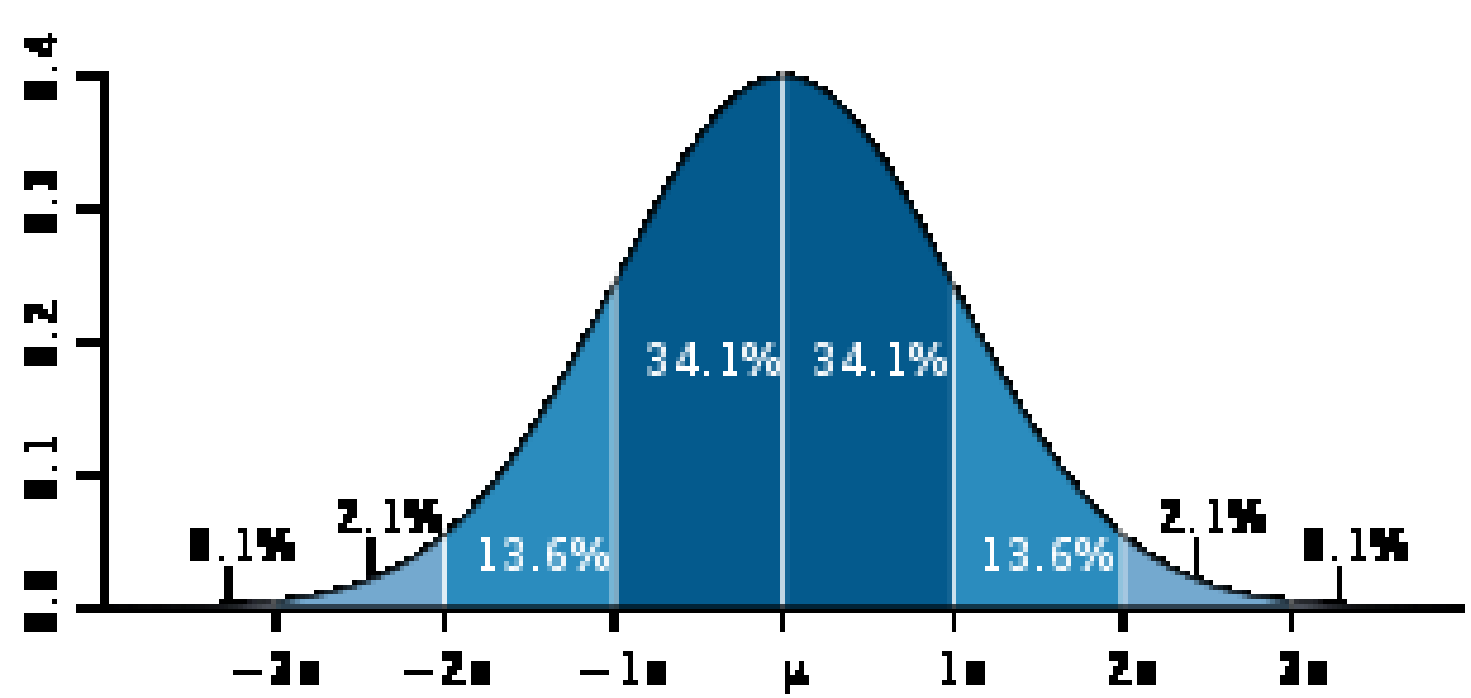
ARITMETICKÝ PRŮMĚR

- Citlivý na extrémní hodnoty
- Aplikovatelná jen u **kardinálních znaků**
- Často udává hodnotu, která se v souboru vůbec nevyskytuje
  - (průměrný Čech navštíví knihovnu 1,12krát za rok)

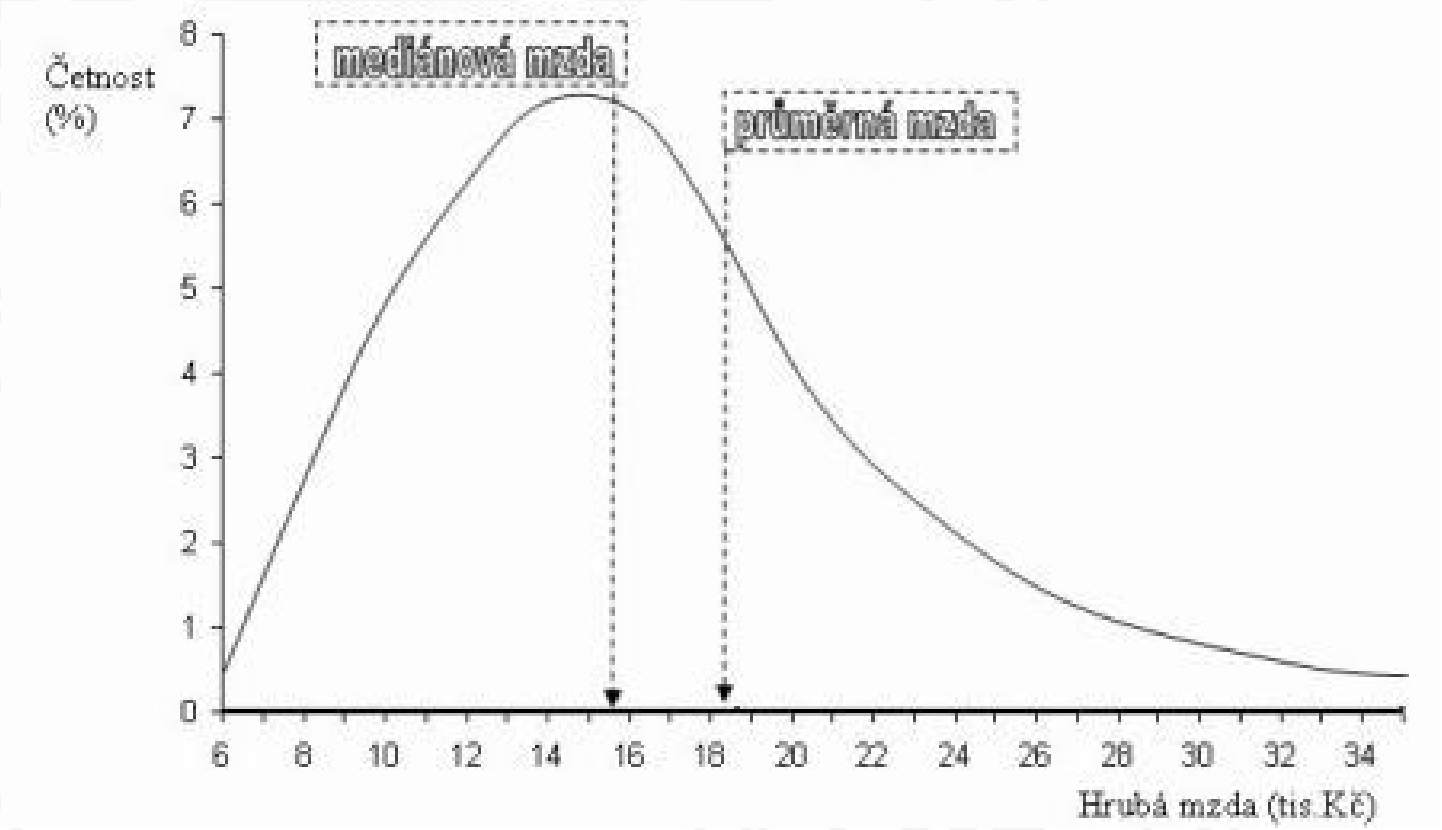
- Kardinální znaky – nemá cenu vytvářet frekvenční tabulku nebo klasické grafy – využívá se **histogram**

Normální rozložení

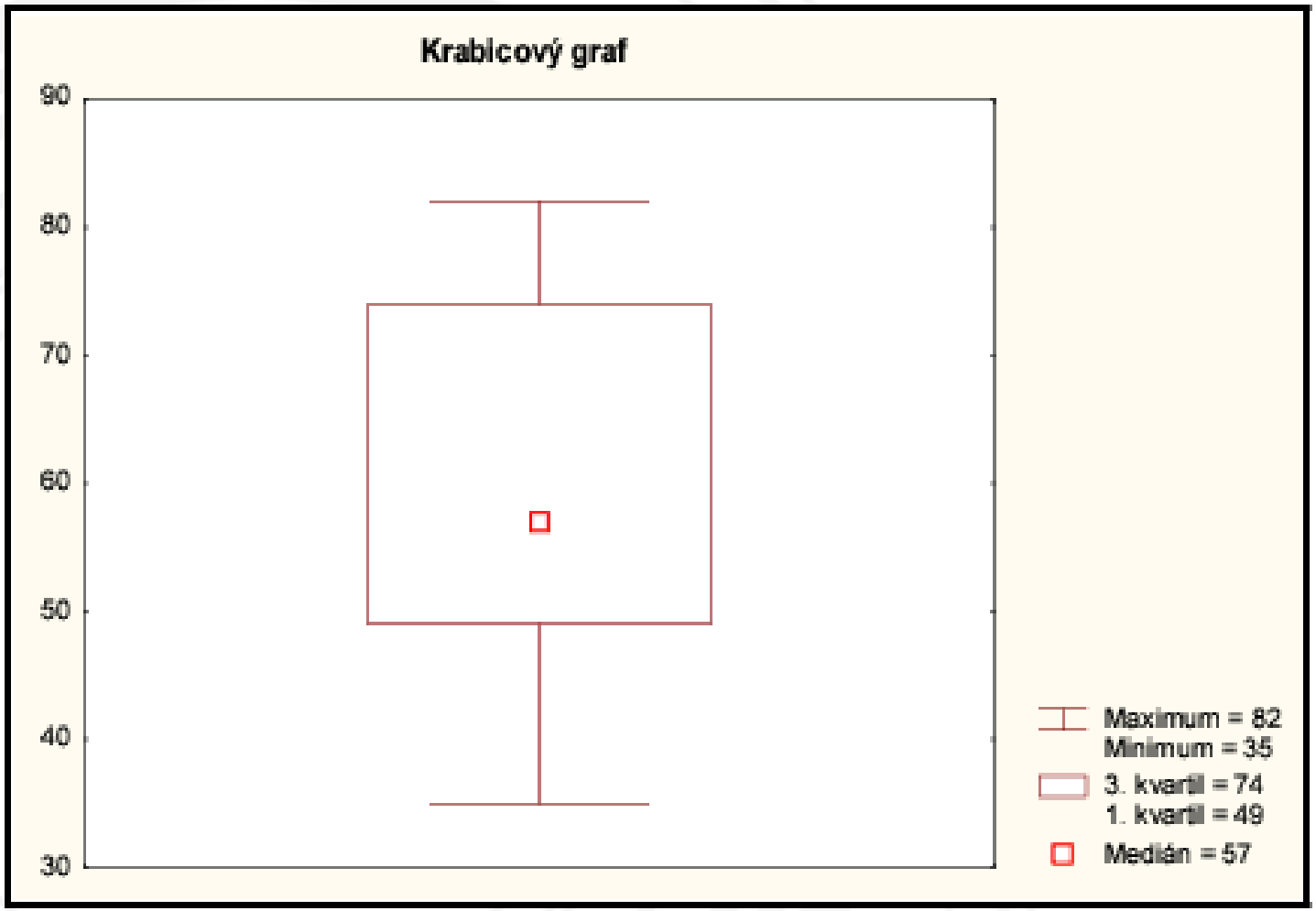
- Normální rozložení: modus = medián = průměr
- Asymetrie rozložení = šikmost



Ukázka šikmého rozložení



Ukázka šikmého rozložení



Rozložení u kardinálních dat

Rozpětí: rozdíl mezi nejmenší a nejvyšší hodnotou

Rozptyl: vypovídá o rozložení hodnot kolem aritmetického průměru

(průměrná čtvercová chyba (ve čtvercích jednotek původní proměnné) – součet druhých mocnin odchylek všech jednotlivých hodnot od průměru dělený rozsahem souboru

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Směrodatná odchylka:

- Druhá odmocnina rozptylu
- ukazuje homogenitu/variabilitu souboru
- čím menší SO je, tím více můžeme věřit aritmetickému průměru

Průměr a standardní odchylka

Zajímavost předmětu	není vůbec zajímavý	..***(X)***..	je velmi zajímavý
Přínosnost předmětu	není vůbec přínosné	.***X(*)**...	je velmi přínosné
Obtížnost obsahu	velmi snadný	.....(.)**X**	velmi obtížný
Náročnost na přípravu	velmi snadný	.....(.)**X**	velmi obtížný
Dostupnost studijních zdrojů	velmi špatně dostupné	.....(.)....X	velmi dobře dostupné
Jak učitel učí	velmi špatný	..***(X)***..	vynikající
Učitel jako odborník	není odborníkem	.....(.)....X	je odborníkem

Zajímavost předmětu	není vůbec zajímavý	.....(.)**X**	je velmi zajímavý
Přínosnost předmětu	není vůbec přínosné	.....(.)..*X*	je velmi přínosné
Obtížnost obsahu	velmi snadný	.....(*)*X**.	velmi obtížný
Náročnost na přípravu	velmi snadný	.....(.)**X**	velmi obtížný
Dostupnost studijních zdrojů	velmi špatně dostupné	.....(.)**X**	velmi dobře dostupné
Jak učitel učí	velmi špatný	.....(.)..*X*	vynikající
Učitel jako odborník	není odborníkem	.....(.)..*X*	je odborníkem

Transformace dat a proměnných

- Kategorizace spojitých proměnných (CATEGORIZE) → vytvoření intervalů

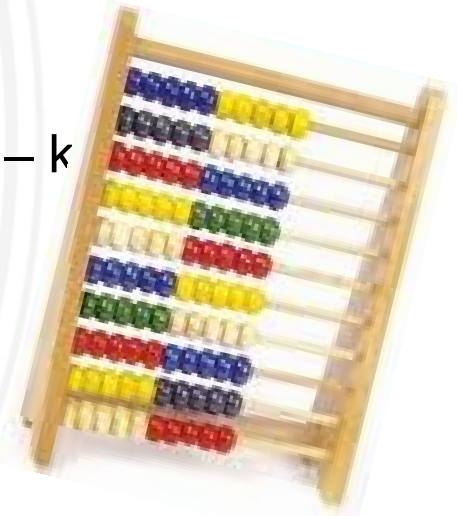
Otázka: Proč je důležité rekategorizovat proměnné?

- Slučování kategorií (spíše spokojen – velice spokojen = spokojen)

Otázka: Kdy je vhodné slučovat proměnné?

Otázka: Lze slučovat i nominální proměnné?

- COUNT – vytváří novou proměnnou (pro sady otázek – k nabízených možností respondent zvolil)



Připomeňme si…

Hypotéza ➡ proměnné ➡ otázky v dotazníku

Hypotéza: Lidé s vyšším vzděláním navštěvují knihovny častěji, než lidé s nižším vzděláním.

Proměnné: vzdělání, frekvence návštěv knihovny

Otázky:

- *Jaké je Vaše nejvyšší ukončené vzdělání?*
- *Jak často navštěvujete knihovnu?*

Třídění druhého stupně

- Porovnání rozložení znaku v podsouborech populace (dle jiného znaku)
- Hypotézy nás vedou v tom, jaké vlastnosti a jejich souvislosti sledovat

Kdy to má smysl:

- Jedná-li se o reprezentativní výběrový soubor (ideálně náhodný výběr)
- Jde-li o nezávislý výběr

Jak statistika vypovídá o základním souboru?

Hlavní roli hraje **směrodatná odchylka / výběrová chyba**:

S 95% jistotou (5% riziko chyby) můžeme tvrdit, že:
průměr základního souboru (parametr)
=
průměr výběrového souboru (statistika)
± 2 směrodatné chyby

S 99% jistotou (1% riziko chyby) můžeme tvrdit, že:
průměr základního souboru (parametr)
=
průměr výběrového souboru (statistika)
± 3 směrodatné chyby

Statistické testování hypotéz

1. *Testování nulové hypotézy o neexistenci vztahu mezi proměnnými*
2. *Hypotéza zamítnuta → testování alternativní hypotézy*

Příklad nulové hypotézy:

Rozložení četností hodnot proměnné (vlastností jednotky), např. příjmu, věku, míry anomie, spokojenosti v životě (atd.) ve výběrovém souboru odpovídá rozložení proměnné v populaci.

Mezi vzděláním a výší příjmu není žádný vztah.

Testy pro statistické testování nulových hypotéz

- T-test o shodě dvou průměrů (parametrický test)
- Man-Whitney test (neparametrický test)

Zlaté pravidlo pro induktivní statistiku:

vysoká hodnota testu signifikance (tj. $\alpha > 0{,}05$) → držíme nulovou hypotézu

nízká hodnota testu signifikance (tj. $\alpha \leq 0{,}05$) → zamítáme nulovou hypotézu

Porovnávání průměrů

Report

Jsem spokojen/a s tematickou šíří nabídky povinně volitelných (B) kurzů.

V jakém stupni stu...	Mean	N	Std. Deviation
bakalářské studium	2,42	229	,821
magisterské studium	2,37	210	,766
Total	2,39	439	,795

Směrodatná odchylka u normálního rozložení:

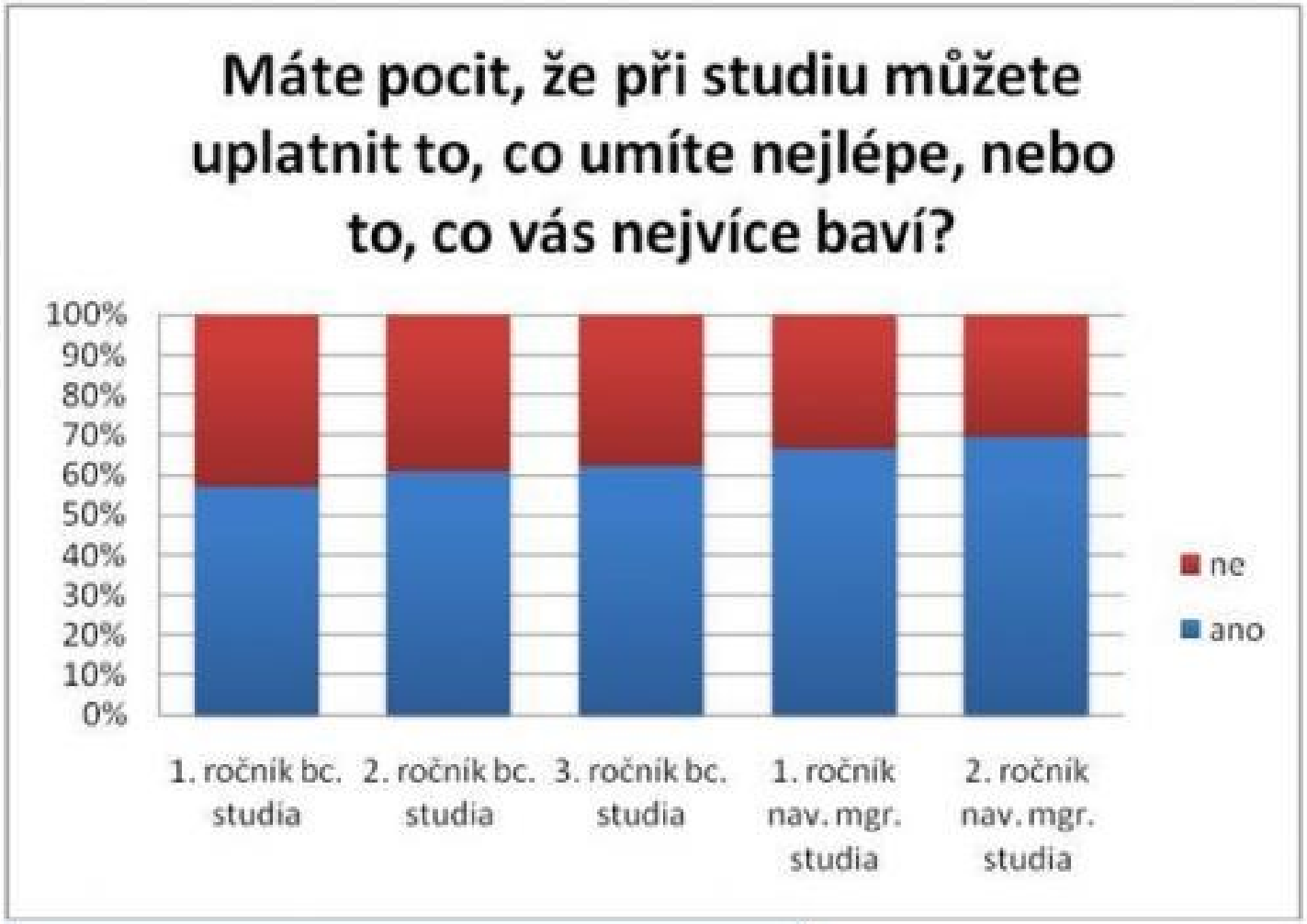
- 68 % případů < 1 směrodatná odchylka
- 95 % případů < 2 směrodatné odchylky
- 99 % případů < 3 směrodatné odchylky

Tabulky rozložení

Vnímáte studium na KISK jako přínosné? * V jaké formě a stupni studia? Crosstabulation

			V jaké formě a stupni studia?				
			bakalářské prezenční	bakalářské kombinované	magisterské prezenční	magisterské kombinované	Total
Vnímáte studium na KISK jako přínosné?	velmi přínosné	Count	59	56	42	56	213
		% within V jaké formě a stupni studia?	45,7%	53,8%	41,2%	51,9%	48,1%
	spíše přínosné	Count	70	48	60	52	230
		% within V jaké formě a stupni studia?	54,3%	46,2%	58,8%	48,1%	51,9%
Total		Count	129	104	102	108	443
		% within V jaké formě a stupni studia?	100,0%	100,0%	100,0%	100,0%	100,0%

Grafy


Máte pocit, že při studiu můžete uplatnit to, co umíte nejlépe, nebo to, co vás nejvíce baví?
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1. ročník bc. studia
2. ročník bc. studia
3. ročník bc. studia
1. ročník nav. mgr. studia
2. ročník nav. mgr. studia
ne
ano