Kapitola 2: Frekvenční tabulky a grafy

- = sumarizace, „uvaření syrových (původních) dat“ do srozumitelné, interpretovatelné, prezentovatelné podoby
- 2 způsoby:
  - Frekvenční/četnostní tabulky
  - Grafy
- Vizualizují data, ukazují prvotní trendy

Frekvenční tabulky 1

- Kolik lidí spadá pod danou kategorii proměnné?=frekvence neboli četnost kategorie
- = shrnují četnosti každé hodnoty proměnné
- Př. Sharon administrativní pracovnice zařízení sociálních služeb chce vědět zda zařízení opravdu slouží starším lidem (+=50)
  - 1) zaznamená si věk všech klientů za měsíc říjen, vznikne seznam 20 klientů, vidí že jen David je +50

„Syrová data“					
Jméno	Věk	Jméno	Věk	jméno	věk
Rashad	32	David	69	Clarisse	37
Rosina	27	Herb	26	Karen	26
Brad	26	Vincent	31	Elwin	49
Chuck	21	Rose	37	Tony	21
Shanti	37	Marguerite	49	Leon	27
Kathy	31	Raquel	31	Mario	31
Antoinette	32	Peter	27		

- Sharon seřadí klienty podle věku a vznikne přehlednější seznam
- Přehlednost je zřetelnější pokud např. N=250 namísto N=20
- Nyní může Sharon sjednocením dat konečně vytvořit první frekvenční tabulku, která je ještě přehlednější

„Syrová data“ seřazená podle velikosti					
Jméno	Věk	Jméno	Věk	jméno	věk
Chuck	21	Leon	27	Shanti	37
Tony	21	Kathy	31	Rose	37
Brad	26	Vincent	31	Clarisse	37
Herb	26	Raquel	31	Marguerite	49
Karen	26	Mario	31	Elwin	49
Rosina	27	Rashad	32	David	69
Peter	27	Antoinette	32		

- =tabulka absolutních četností
- Každé hodnotě proměnné je přiřazeno číslo, podle toho kolikrát se vyskytuje
- Lze sestrojit pro každý typ proměnné
- Př. 3 z 20 klientů mají 37 let

Tabulka 1: absolutních četností (proměnná věk)

Jméno	Věk	Abs. četnost
Chuck+Tony	21	2
Brad+Herb+Karen	26	3
Rosina+Peter+Leon	27	3
Kathy+Vincent +Raquel+Mario	31	4
Rashad+Antoinette	32	2
Shanti+Rose+Clarisse	37	3
Marguerite+Elwin	49	2
David	69	1
Total		20

- = tabulka kumulativních četností
- = kumulativní četnost hodnoty X je rovna součtu všech absolutních četností hodnot <=X

  Př. 17 klientů (2+3+3+4+2+3) z 20 má 37 let a méně

  (smysl interpretace pouze u nejméně ordinálních proměnných)
- Poslední řádek kumulativní četnosti = celkový počet případů

Tabulka 2: kumulativních četností (proměnná věk)

Jméno	Věk	Abs. četnost	Kum. četnost
Chuck+Tony	21	2	2
Brad+Herb+Karen	26	3	5
Rosina+Peter+Leon	27	3	8
Kathy+Vincent +Raquel+Mario	31	4	12
Rashad+Antoinette	32	2	14
Shanti+Rose+Clarisse	37	3	17
Marguerite+Elwin	49	2	19
David	69	1	20
Total		20	

- = tabulka relativních četností
- máme-li 20 lidí ve vzorku, pak každý člen reprezentuje 5% (100/20) – viz David
- Př. 15 % lidí ve vzorku má 37 let
  - důkaz (3/20)*100=15%

Tabulka 3: relativních četností (proměnná věk)

Jméno	Věk	Abs. četnost	Rel. četnost
Chuck+Tony	21	2	10
Brad+Herb+Karen	26	3	15
Rosina+Peter+Leon	27	3	15
Kathy+Vincent +Raquel+Mario	31	4	20
Rashad+Antoinette	32	2	10
Shanti+Rose+Clarisse	37	3	15
Marguerite+Elwin	49	2	10
David	69	1	5
Total		20	100

- =tabulka relativních kumulativních četností
- Relativní kumulativní četnost pro hodnotu X je rovna součtu všech relativních četností hodnot <= X
- Př. 85% klientů má 37 let a méně (10+15+15+20+10+15=85)
- Užitečné chcem-li znát relativní pozici určité hondoty vzhledem k ostatním v datech (viz též percentil)

Tabulka 4: relativních kumulativních četností

Jméno	Věk	Abs. četnost	Rel. četnost	Rel. Kum. četnost
Chuck+Tony	**21**	2	**10**	10
Brad+Herb+Karen	**26**	3	**15**	25
Rosina+Peter+Leon	**27**	3	**15**	40
Kathy+Vincent +Raquel+Mario	**31**	4	**20**	60
Rashad+Antoinette	**32**	2	**10**	70
Shanti+Rose+Clarisse	**37**	3	**15**	**85**
Marguerite+Elwin	49	2	10	95
David	69	1	5	100
Total		20	100	

- Pokud mnoho dat (zvláště spojitých), je přehlednější vytvořit skupiny (intervaly)
- Jak široké mají intervaly být?
  - 2 hlediska:
    - a) dostatečný počet případů v každé skupině
    - b) logika skupin (homogenita uvnitř skupiny)
  - Pokud jsou hodnoty rovnoměrně rozložené, pak v každém intervalu stejně případů.
  - Pokud ne, pak snaha o smysluplné skupiny zhledem k určité vlivné proměnné
- Každý případ musí spadat pouze do jednoho intervalu
- Pokud je případ na hranici např. 29.6 pak zaokrouhlujeme (dolní hranice intervalu 30-39 =29.5, horní hranice=39.49)

Tabulka 5: seskupených relativních kumulativní četností (proměnná věk)

Věkové skupiny	Abs.Rel. četnost	Kum.rel. četnost
20-29	40	40
30-39	45	85
40-49	10	95
50-59	0	95
60-69	5	100
Total	100	

Užití frekvenčních tabulek v analýze 1

- Př. Jennifer chce prozkoumat neomluvené absence personálu, zda neexistuje nějaké sezónní vzorce, které by mohly být odstraněny přizpůsobením politiky dovolených.
- Vidí, že zatímco na jaře (duben+květen) bylo zaznamenáno 70 (35 %) případů, tak v létě (červen+červenec) jich bylo 130 (65 %)

Počet denních absencí personálu podle měsíců (N=200) (celkem v měsíci)		
měsíc	Abs. četnost	Kum.rel. četnost
Duben	30	15
Květen	40	35
Červen	60	65
Červenec	70	100

- Užitečné chceme-li srovnat měření z dvou rozdílných skupin nebo databází
- Př. SoPka Sue vytvořila studijní příručku - je příručka efektivní?

  Jak zjistit? Srovnat zkouškové body lidí kteří použili příručku (experimentální skupina X) a kteří ji nepoužili (kontrolní skupina C)

  Vidí že:
  - 20 % lidí v X skórovalo nad 90 zatímco v C pouze 5 %
  - v X skórovalo pouze 10 % pod 70, zatímco v C relativně 2x více (20 %) – příručka zdá se pomohla!
- Na základě rel.kum.četn. lze určit pořadí – tzv. percentil
  - = percentilové pořadí určuje procento případů jejichž hodnota je nižší než příslušná hodnota
  - Př. Clarice skórovala 90 bodů – skončila lépe než nejméně 80 % všech lidí v X skupině
- Všimnětě si různé velikosti skupin (N=200 a N=300) – použití procent umožňuje srovnání skupin

Skóre experimentální skupiny (X) (N=300)		
Skóre	Abs.Rel.četnost	Kum.rel. četnost
50-59	0	0
60-69	10	10
70-79	40	50
80-89	30	80
90-100	20	100
Skóre kontrolní skupiny (C) (N=200)		
Skóre	Abs.Rel.četnost	Kum.rel. četnost
50-59	5	5
60-69	15	20
70-79	40	60
80-89	35	95
90-100	5	100

Chybná prezentace výsledků 1

- Př. Personální pracovnice Emma pyšně sděluje řediteli firmy, že její snaha zaměstnat více žen byla velmi úspěšná
- Emma: „v 5 ze 6 sektorů jsem přijala relativně (vyšší procento) více žen než mužů“
- Pravda ale: celkově přijala jen 21% žen vs. 78 % mužů
- Chyba Emmy: používala nestejně velké skupiny (D>A+B+C+E+F)
- Obecně je lepší používat procenta pouze u vysokých četností jako 146 z 411, u četností jako 3 z 5 lépe používat absolutní hodnoty

Tabulka náborů 2001 2002 ve firmě XYZ podle pohlaví

Klasifikace pracovního místa	Muži		ženy	
	Počet	Rel. Četnost	počet	Rel. četnost
A	3 ze 6	50	4 ze 6	67
B	1 ze 3	33	1 ze 2	50
C	0 z 1	0	1 z 10	10
D	85 ze 100	85	2 ze 40	5
E	2 ze 3	67	2 ze 2	100
F	3 ze 7	43	4 ze 7	57
Celkem	94 ze 120	78	14 ze 67	21

- Způsobuje chudoba delikvenci?
- 1. tabulka: nedostatek evidence – třeba porovnat s bohatými (2.tabulka)

	Chudí
Delikventní	7
Nedelikventní	93
	100 (400)

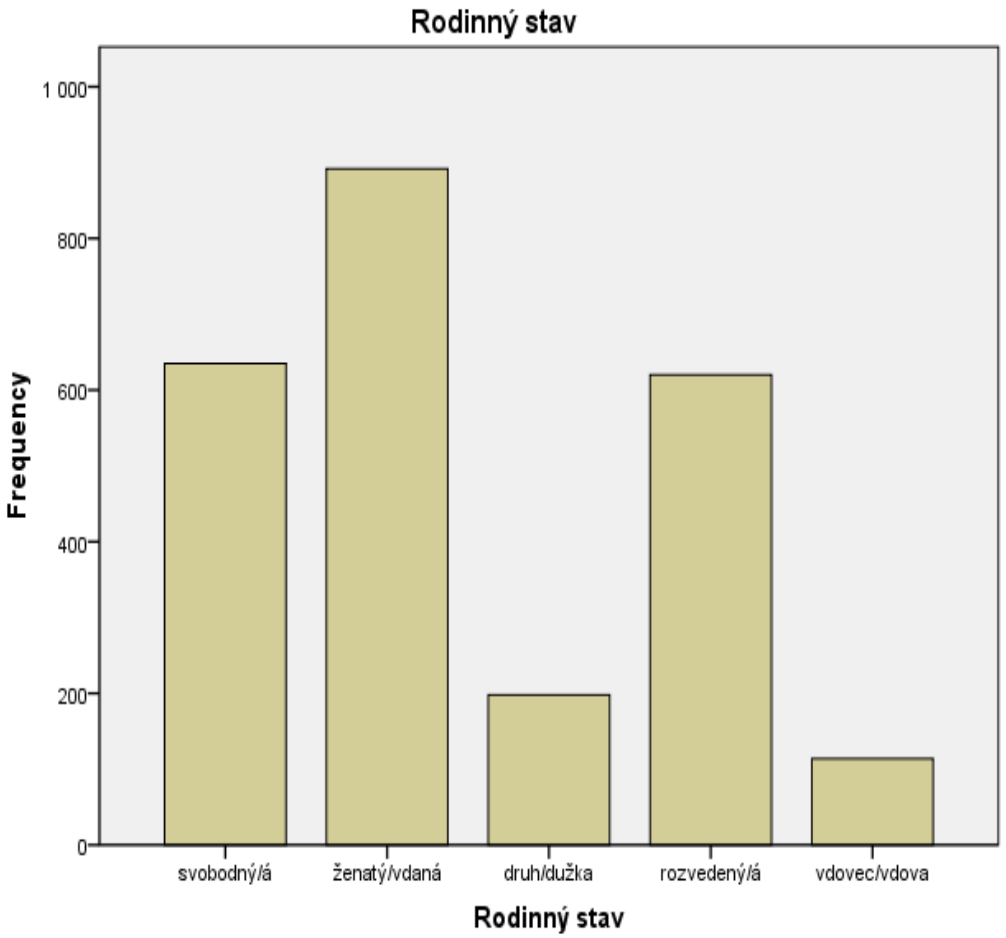
	Chudí	Bohatí
Delikventní	7	7
Nedelikventní	93	93
	100 (400)	100 (654)

Grafická prezentace dat

- Grafická prezentace obětuje přesnost (detail) za komunikativnost sdělení o distribuci hodnot proměnné
- vhodné pokud publikum není např. vědecká rada
- Existuje mnoho grafů – který použít?
  - a) jasnost prezentace
  - b) úroveň měření proměnné
- Většina grafů používá osy x(pro hodnoty proměnné) a y(pro četnost)

Sloupcové/čárové diagramy/grafy (bar/line graphs/charts)

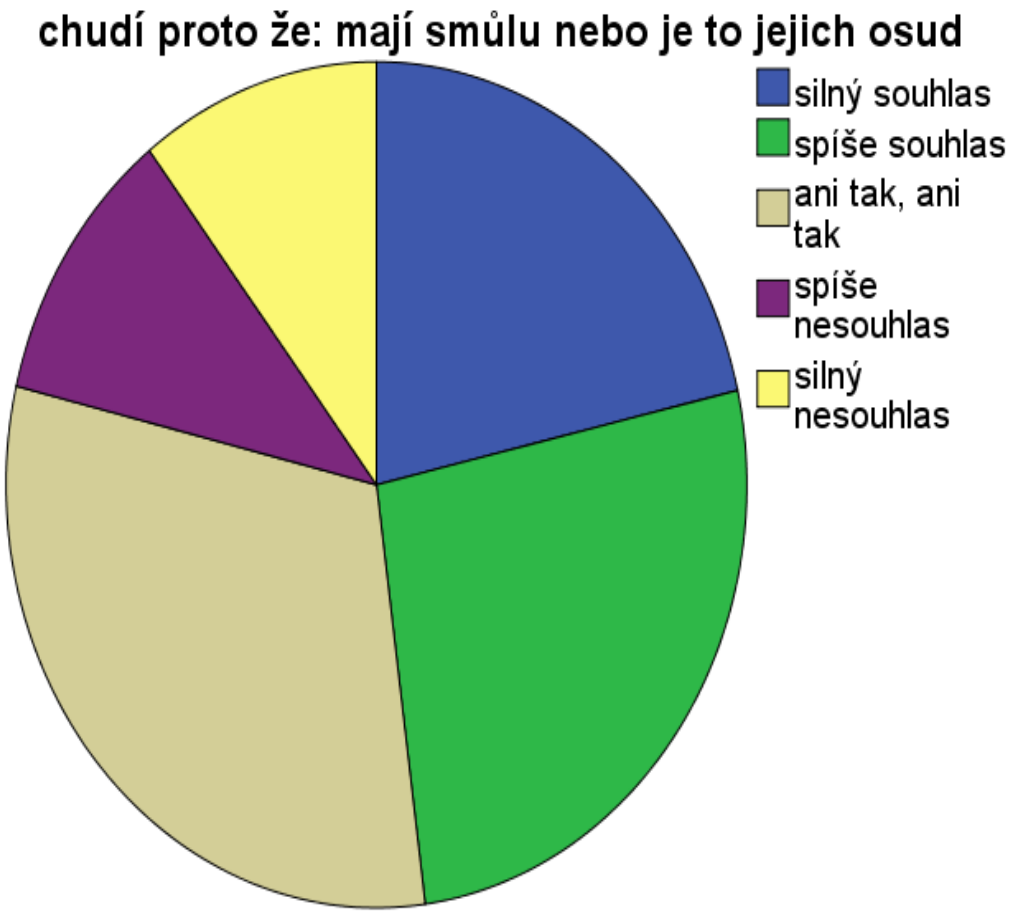
- Nominální data
- Každý sloupec stejná šířka
- Pořadí sloupců nehraje roli (mezi kategoriemi pouze kvalitativní rozdíly)
- Sloupce se nedotýkají (nespojitost)
- Výška sloupce reflektuje četnost hodnoty
- Př. 2x vyšší sloupec=2x vyšší četnost



Koláčové diagramy/grafy (pie graphs/charts)

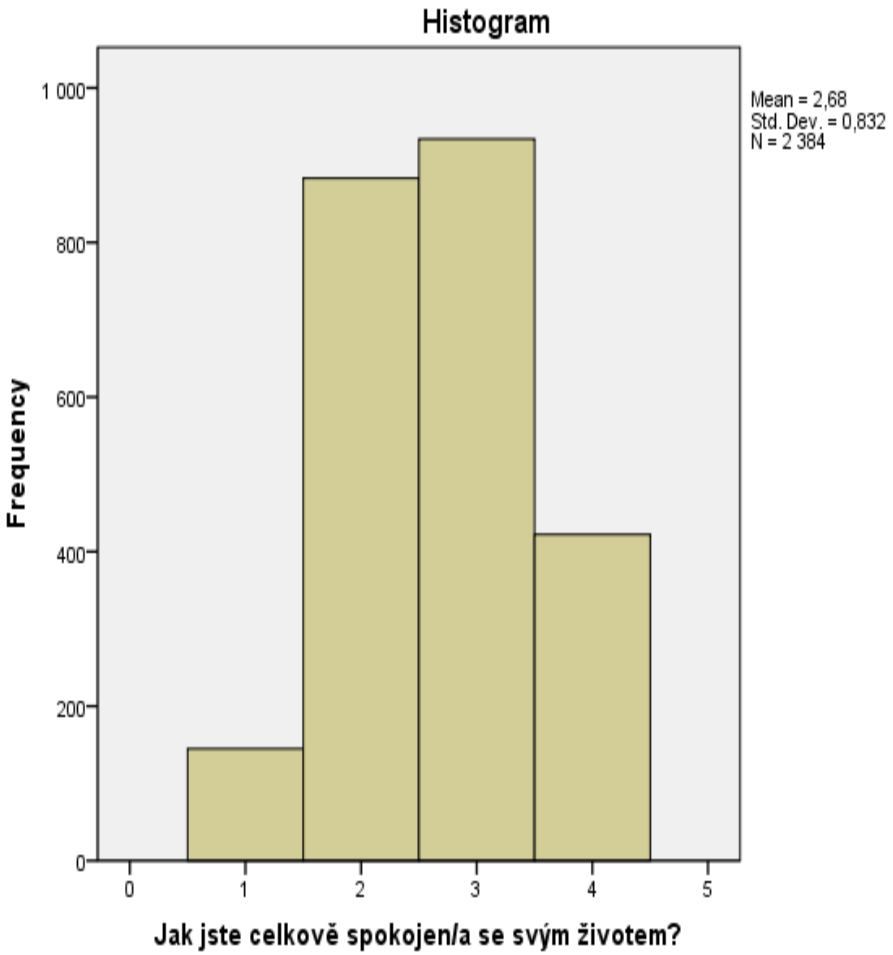
Celý koláč = proměnná = 100 %

- Jednotlivé porce (trojúhelníčky) = kategorie proměnné
- Čím větší porce, tím větší četnost
- Výhoda: nabízí okamžitý pohled na distribuci proměnné
- Nevýhoda: nehodí se na proměnné s více kategoriemi - nepřehledné



histogramy

- Podobné sloupcovým grafům
- Rozdíly:
  - šířka jednotlivých sloupců=šířka intervalu (intervalové/poměrové proměnné seskupené do intervalů)
  - Pořadí sloupců=pořadí kategorií (ordinální proměnná)

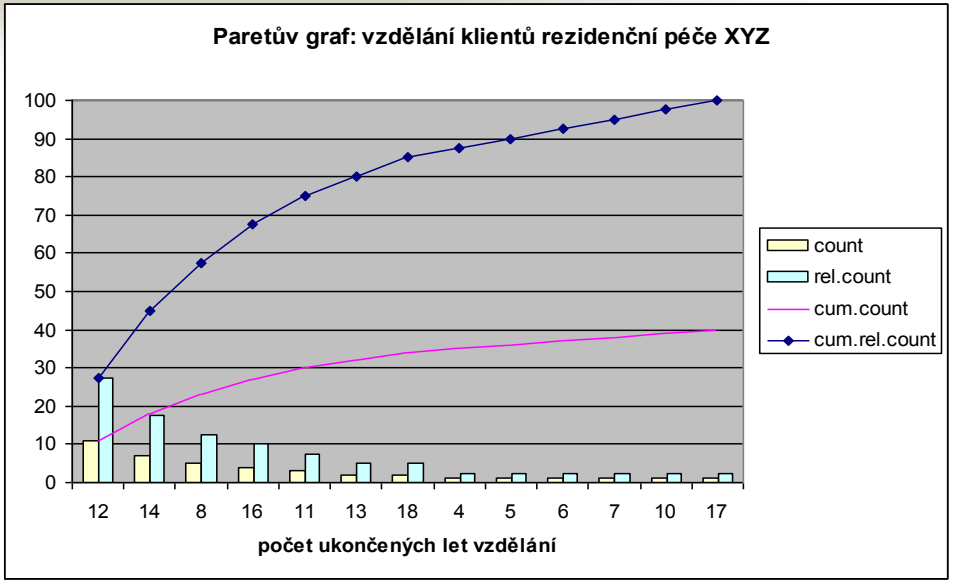


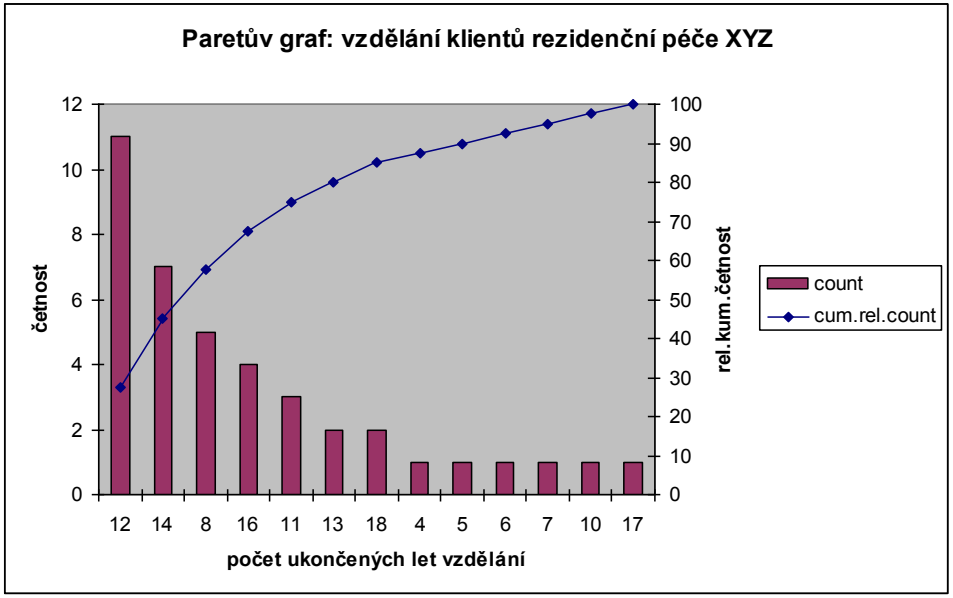
Frekveční mnohoúhelník (polygon)

- Totéž jako histogram
- Rozdíl: namísto sloupců linka spojující středy vrcholů každého sloupce histogramu vytvářející polygon

Paretův diagram (Pareto chart)

- Řadí hodnoty proměnné podle četnosti s klesající tendencí (nejčetnější vlevo)
- Čára nad četnostními sloupci představuje jak kumulativní četnost (levá vertikála) tak relativní kumulativní četnost (pravá vertikála)
- viz alternativní formy vyrobené v excelu





Graf stonků a listů (Stem and leaf plot)

- Stonek=první číslo hodnoty proměnné
- List=poslední číslo hodnoty proměnné

Graf stonků a listů: věk klientů rezidenční péče N=40

Frekv.	Stonek	listy
1	5	9
2	6	24
6	6	566889
4	7	1144
16	7	5577777777788999
6	8	014444
3	8	558
1	9	5
1	10	3

- Př. Jeden klient má 62 let a druhý 64 let