MUNI
ECON

Organizace a vizualizace dat

Povinná literatura: Mann (2016), Kapitola 2

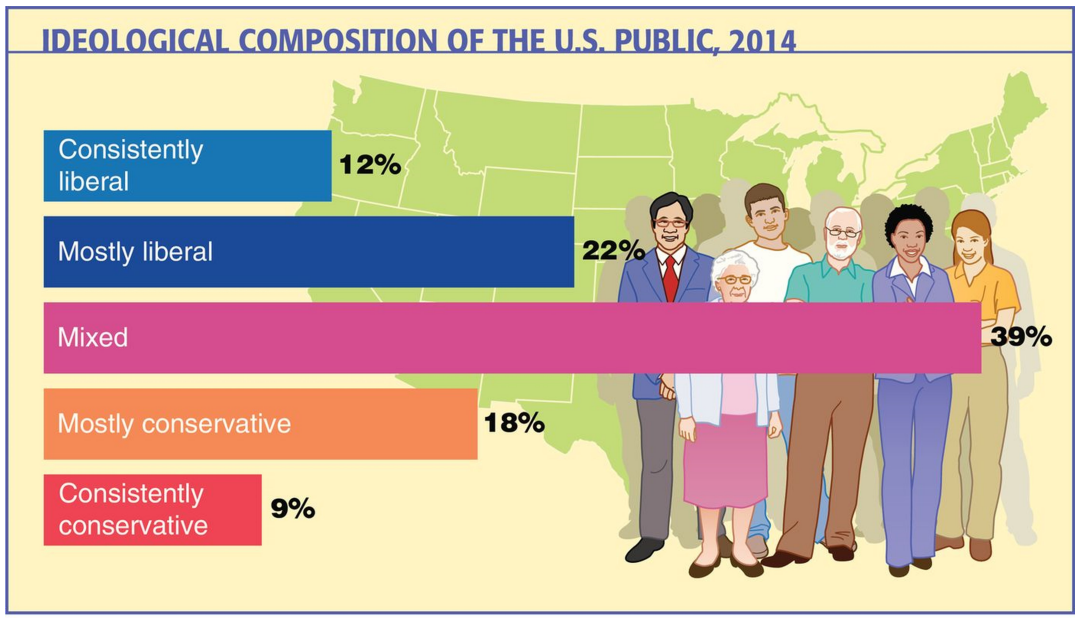
Z čeho studovat druhou lekci?

Povinná literatura:	Mann (2016), Kapitola 2
Příprava na cvičení:	Leaflet 03 Koncepty a procedury, cv.03, kap. 02
Příprava na zkoušku:	Mann (2016), kap. 02 Leaflet 03 Sbírka úloh, kap. 02 Koncepty a procedury, cv.03, kap. 02

Obsah

Ideologické složení americké veřejnosti, 2014

Pew Research Center provedlo národní průzkum mezi 10 013 dospělými od 23. ledna do 16. března 2014, aby zjistilo politické názory dospělých ve Spojených státech.

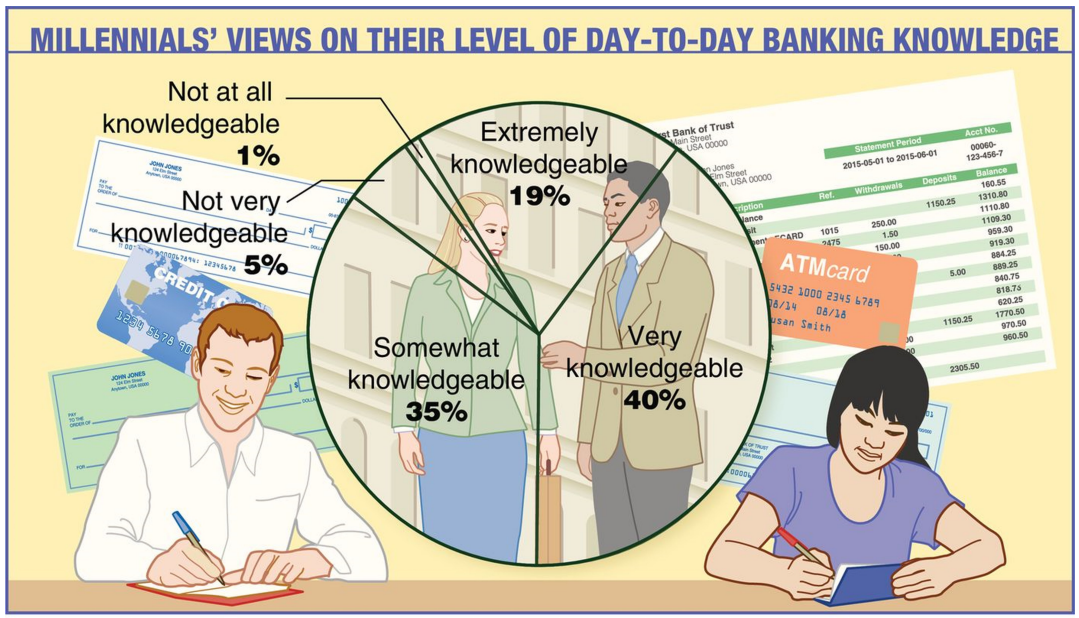


Data source: Pew Research Center

Jak ukazuje výše uvedený sloupcový graf, 12 % dotázaných dospělých uvedlo, že jsou konzistentně liberální, 22 % uvedlo, že jsou převážně liberální, a podobně.

Znalosti mileniálů v oblasti bankovnictví

TD Bank provedla průzkum mezi mileniály (ve věku 18-34 let) od 28. ledna do 10. února 2014, se zaměřením na pochopení jejich bankovního chování a preferencí.

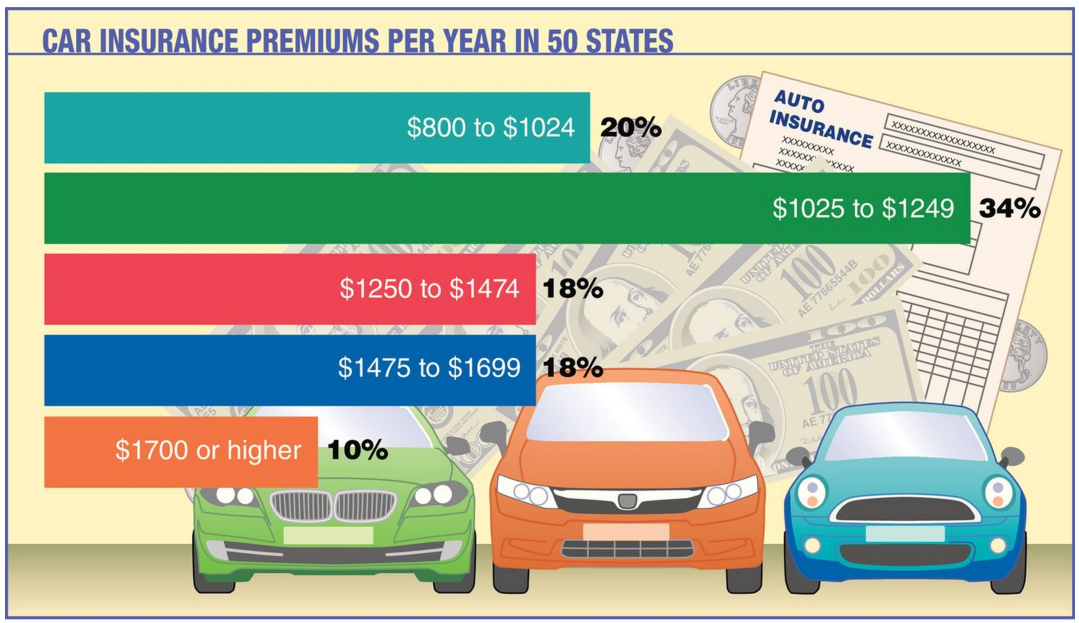


Data source: TD Bank: The Millennial Financial Behaviors & Needs Survey

Jak ukazuje graf, 19 % mileniálů uvedlo, že mají excelentní znalosti o bankovnictví, 40 % uvedlo, že mají velmi dobré znalosti, 35 % uvedlo, že mají částečné znalosti, 5 % přiznalo, že nemají příliš dobré znalosti, a 1 % uvedlo, že nemají žádné znalosti.

Roční pojistné na auto v 50 státech USA

Uvedený graf ukazuje procentuální rozdělení ročních pojistných na auto v 50 státech USA. Data použita k vytvoření tohoto grafu jsou založena na odhadech provedeného společností insure.com.

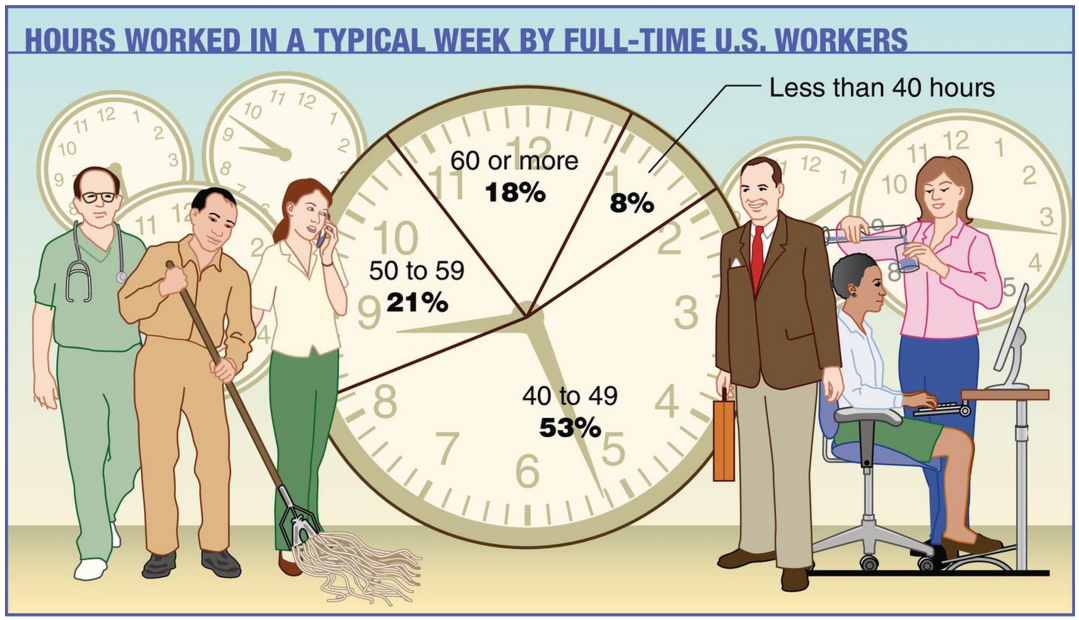


Data source: www.insure.com

Jak ukazuje graf, v 20 % států se sazby za pojištění auta pohybovaly v rozmezí $800 až $1024, a podobně. Všimněte si, že poslední třída ($1700 a více) nemá horní limit. Taková třída se nazývá otevřená třída.

Pracovní doba v typickém týdnu

Uvedený výsečový graf ukazuje procentuální rozdělení hodin odpracovaných v typickém týdnu plně zaměstnanými pracovníky ve Spojených státech.



Data source: www.gallup.com

Jak ukazují čísla ve výsečovém grafu, 8 % těchto pracovníků uvedlo, že pracují méně než 40 hodin týdně, 53 % pracuje 40 až 49 hodin týdně, a podobně. Jak můžete pozorovat, dvě z tříd v tomto grafu jsou otevřené třídy.

Kolik šálků kávy vypijete denně?

V průzkumu Gallup, provedeném telefonicky ve dnech 9.–12. července 2012, byli dotázáni dospělí Američané ve věku 18 a více let: „Kolik šálků kávy vypijete v průměru za den?“

Data source: Gallup poll of U.S. adults aged 18 and older conducted July 9–12, 2012

Podle výsledků průzkumu 36 % těchto dospělých uvedlo, že nepijí žádnou kávu a podobně. Poslední třída je otevřená třída, která ukazuje, že 10 % těchto dospělých vypije čtyři nebo více šálků kávy denně. Tato třída nemá horní limit.

Obsah

Organizace a vizualizace kvalitativních dat

Data zaznamenaná v pořadí, v jakém jsou sbírána, a před tím, než jsou zpracována nebo seřazena, jsou nazývána **surová data (Raw data)**.

Surová kvalitativní data lze poté organizovat a vizualizovat pomocí následujících metod analýzy a prezentace dat:

- Rozložení četností
- Relativní četnosti a procentního zastoupení
- Grafická prezentace kvalitativních dat

Rozložení četností

Rozložení četností kvalitativní proměnné zaznamenává počet prvků, které spadají do každé z kategorií dané proměnné.

Variable → | **Response** | **Number of Adults** ← Frequency column

Response	Number of Adults
Very worried	162
Moderately worried	203
Not too worried	305
Not worried at all	325
Others	20
	Sum = 1015

Category → Not too worried

305 ← Frequency

Příklad

Příklad: Vzorek 30 osob, které často konzumují koblihy, byl dotázán na to, který typ koblih je jejich oblíbený. Odpovědi těchto 30 osob jsou následující:

glazed	filled	other	plain	glazed	other
frosted	filled	filled	glazed	other	frosted
glazed	plain	other	glazed	glazed	filled
frosted	plain	other	other	frosted	filled
filled	other	frosted	glazed	glazed	filled

Příklad: Řešení

Tabulka: Rozložení četností oblíbených typů koblih

Typ	Zaznamenáno	Četnost (f)
glazed	|||||||	8
filled	||||||	7
other	||||||	7
plain	|||	3
frosted	|||||	5
		Suma = 30

Relativní četnost a procentní zastoupení

Výpočet **relativní četnosti kategorie**

$$\text{Relativní četnost kategorie} = \frac{\text{počet výskytů v kategorii}}{\text{počet všech výskytů}}$$

Výpočet **procentní zastoupení kategorie**

$$\text{Procentní zastoupení kategorie} = \text{relativní četnost kategorie} * 100$$

Relativní četnost a procentní zastoupení

Příklad: Určete relativní četnost a procentní zastoupení kategorií pro data v četnostní tabulce oblíbených typů koblih.

Řešení:

Tabulka: Relativní četnost a procentní rozdělení oblíbených typů koblih

Typ	Relativní četnost	Procentní zastoupení
glazed	8/30 = 0.267	0.267*100 = 26.7
filled	7/30 = 0.233	0.233*100 = 23.3
other	7/30 = 0.233	0.233*100 = 23.3
plain	3/30 = 0.100	0.100*100 = 10.0
frosted	5/30 = 0.167	0.167*100 = 16.7
	Suma = 1	Suma = 100 %

Grafická prezentace kvalitativních dat

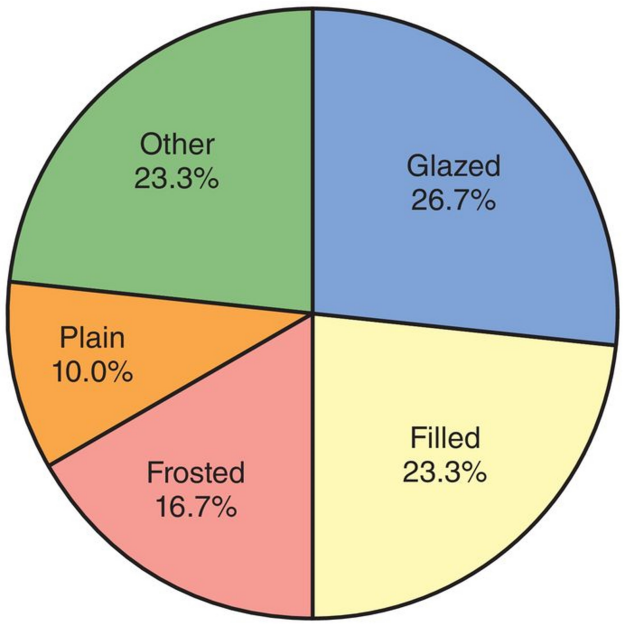
Kruh rozdělený na části, které představují relativní četnosti různých kategorií zastoupených ve vzorku, nebo procenta populace, se nazývá **koláčový graf**.

Graf tvořený sloupci, jejichž výšky reprezentují četnosti příslušných kategorií, se nazývá **sloupcový graf**.

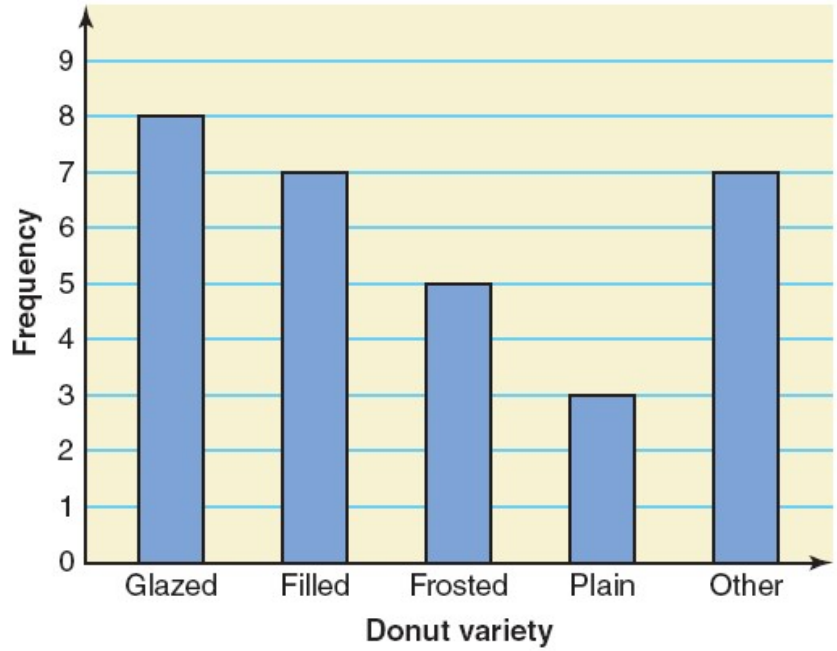
Paretův graf je speciální případ sloupcového grafu, který má sloupce uspořádané podle svých výšek sestupně. Abyste vytvořili Paretův diagram, uspořádejte sloupce podle jejich výšek tak, že sloupec s největší výškou se objeví jako první na levé straně, a následující sloupce jsou uspořádány sestupně, přičemž sloupec s nejmenší výškou se objeví jako poslední na pravé straně.

Koláčový graf vs. sloupcový graf

Obrázek: Koláčový graf

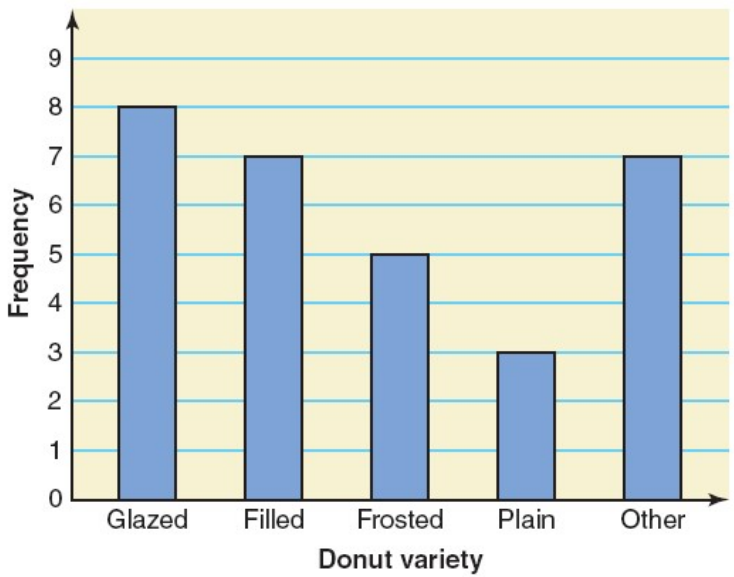


Obrázek: Sloupcový graf

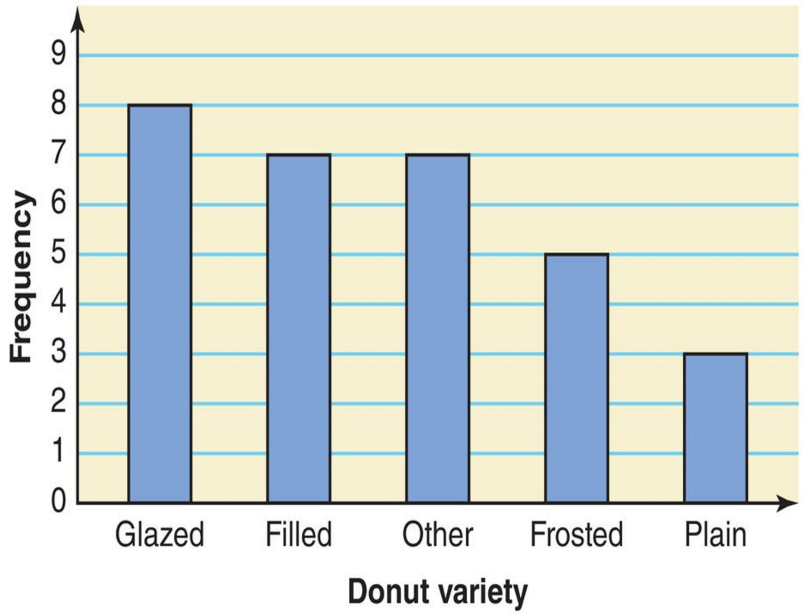


Sloupcový graf vs. Paretův graf

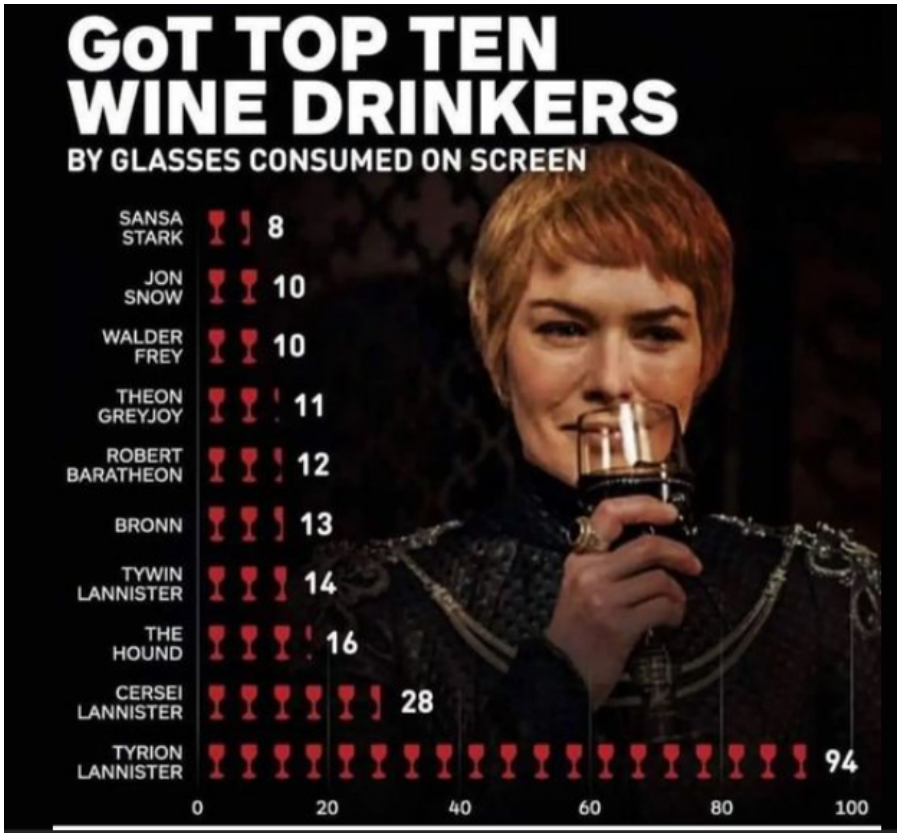
Obrázek: Sloupcový graf



Obrázek: Paretův graf



Příklad vizualizace kvalitativních dat na závěr sekce



Source: 1st Loan Advisor

- Graf ukazuje postavy, které na obrazovce vypily nejvíce sklenic vína.
- Tyrion Lannister je zdaleka největším „pijákem“ s 94 sklenicemi, což je výrazně více než kterýkoliv jiný charakter.
- Graf humorně zvýrazňuje charakteristické chování postav, přičemž ironicky zdůrazňuje, kolik vína vypily během seriálu, což často odráží jejich emocionální stavy nebo společenskou pozici v příběhu.

Obsah

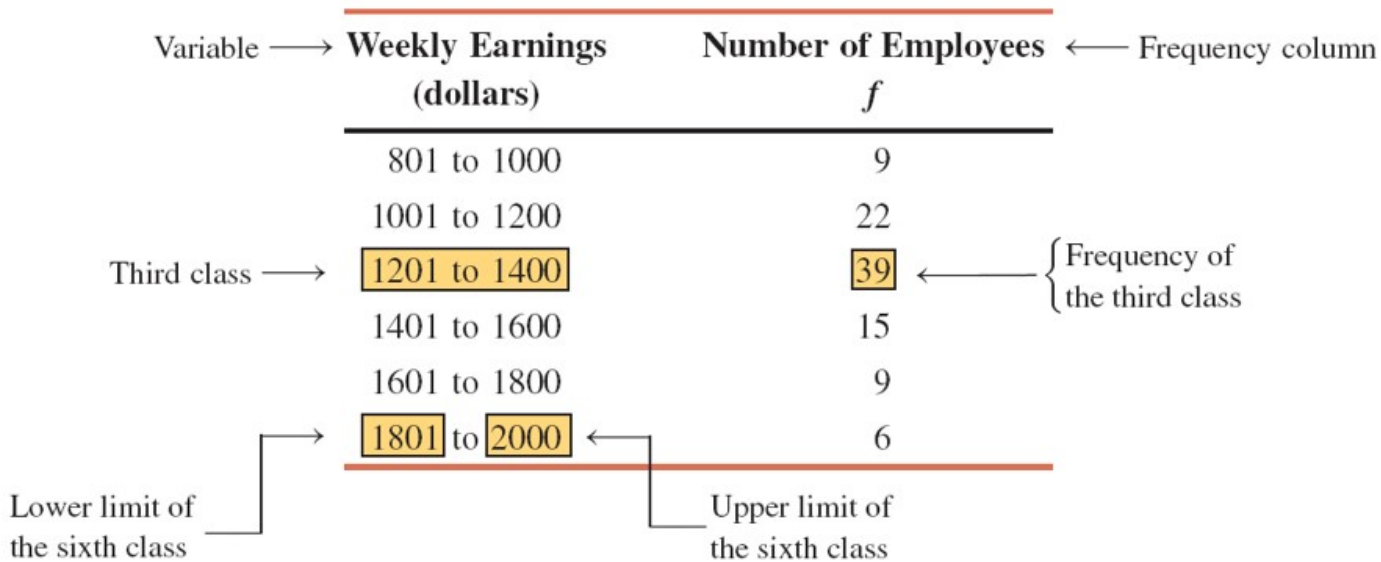
Organizace a vizualizace kvantitativních dat

Surová kvantitativní data lze organizovat a vizualizovat pomocí následujících metod analýzy a prezentace dat:

- Rozložení četností
- Relativní četnosti a procentní zastoupení
- Grafická prezentace seskupených dat

Rozložení četností kvantitativních dat

Rozložení četností kvantitativních dat se provádí skrz stanovení tříd (intervalů) a zaznamenání počtu hodnot, které patří do každé třídy (intervalu). Data prezentovaná ve formě rozložení četností se nazývají **seskupená data**.



Princip stanovení četností u kvantitativních dat

Teoreticky:

1. **Vypočítání šířky třídy:**
   Šířka třídy = Dolní mez následující třídy - Dolní mez aktuální třídy
2. **Výpočet středu třídy:**
   Střed třídy = $\frac{\text{Dolní mez třídy} + \text{Horní mez třídy}}{2}$

Jak ovšem nalézt šířku třídy?

Přibližná šířka třídy $= \frac{\text{Nejvyšší hodnota ve vzorku} - \text{Nejmenší hodnota ve vzorku}}{\text{Uvažovaný počet tříd}}$

Příklad seskupení dat

Příklad: Následující tabulka uvádí hodnotu (v milionech dolarů) každého z 30 baseballových týmů, jak je odhaduje časopis Forbes (zdroj: časopis Forbes, 13. dubna 2015). Sestavte tabulku četnostního rozdělení.

Team	Value (millions of dollars)	Team	Value (millions of dollars)
Arizona Diamondbacks	840	Milwaukee Brewers	875
Atlanta Braves	1150	Minnesota Twins	895
Baltimore Orioles	1000	New York Mets	1350
Boston Red Sox	2100	New York Yankees	3200
Chicago Cubs	1800	Oakland Athletics	725
Chicago White Sox	975	Philadelphia Phillies	1250
Cincinnati Reds	885	Pittsburgh Pirates	900
Cleveland Indians	825	San Diego Padres	890
Colorado Rockies	855	San Francisco Giants	2000
Detroit Tigers	1125	Seattle Mariners	1100
Houston Astros	800	St. Louis Cardinals	1400
Kansas City Royals	700	Tampa Bay Rays	605
Los Angeles Angels of Anaheim	1300	Texas Rangers	1220
Los Angeles Dodgers	2400	Toronto Blue Jays	870
Miami Marlins	650	Washington Nationals	1280

Příklad seskupení dat - nalezení třídy

Řešení: Nejmenší hodnota je 605 a nejvyšší hodnota je 3200. Uvažujme, že se rozhodneme tyto údaje seskupit do šesti tříd o stejné šířce. Potom přibližná šířka každé třídy = $\frac{3200-605}{6} = 432.5$

Nyní zaokrouhlíme tuto přibližnou šířku na vhodné číslo, řekněme 450. Dolní mez první třídy může být vzata jako 605 nebo libovolné číslo menší než 605.

Předpokládejme, že vezmeme 601 jako dolní mez první třídy. Pak naše třídy budou:

601-1050, 1051-1500, 1501-1950, 1951-2400, 2401-2850, 2851-3300.

Příklad seskupení dat - četnostní rozdělení

Tabulka: Četnostní rozdělení hodnoty basebalových týmů

Hodnota týmu (v mil. $)	Zastoupení	Počet týmů (f)
601–1050	||||||||||||||||	16
1051–1500	|||||||||	9
1501–1950	|	1
1951–2400	|||	3
2401–2850		0
2851–3300	|	1
		Suma = 30

Příklad seskupení dat - relativní rozložení četností

Výpočet **relativní četnosti třídy**

$$\text{Relativní četnost třídy} = \frac{\text{Četnost této třídy}}{\text{Součet všech četností}} = \frac{f}{\sum f}$$

Výpočet **procentního zastoupení třídy**

$$\text{Procentní zastoupení třídy} = (\text{Relativní četnost třídy}) \cdot 100\%$$

Příklad seskupení dat - relativní rozložení četností

Tabulka: Relativní rozložení četností a procentní zastoupení hodnoty týmů

Hodnota (v mil. $)	Relativní rozložení čet.	Procentní zastoupení
601–1050	16/30 = 0.533	53.3
1051–1500	9/30 = 0.300	30.0
1501–1950	1/30 = 0.033	3.3
1951–2400	3/30 = 0.100	10.0
2401–2850	0/30 = 0.000	0.0
2851–3300	1/30 = 0.033	3.3
	Suma = 1	Suma = 100 %

Vizualizace kvantitativních (seskupených) dat

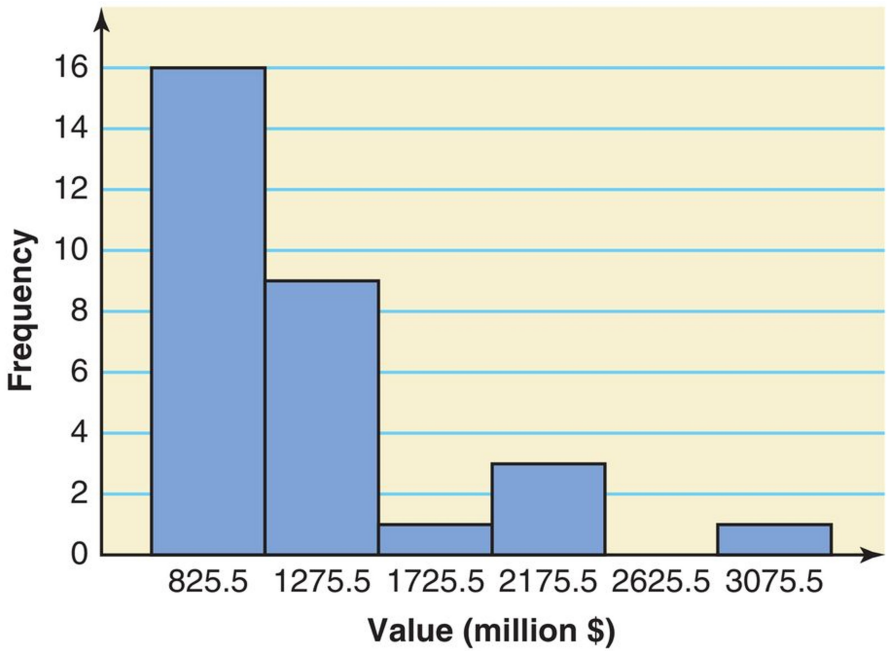
Histogram je graf, ve kterém jsou třídy označeny na vodorovné ose a četnosti, relativní četnosti nebo procenta jsou označeny na svislé ose. Četnosti, relativní četnosti nebo procenta jsou reprezentovány výškou sloupců. V histogramu jsou sloupce vykresleny vedle sebe.

Polygon je graf vytvořený spojením středů vrcholů po sobě jdoucích sloupců v histogramu pomocí přímých čar.

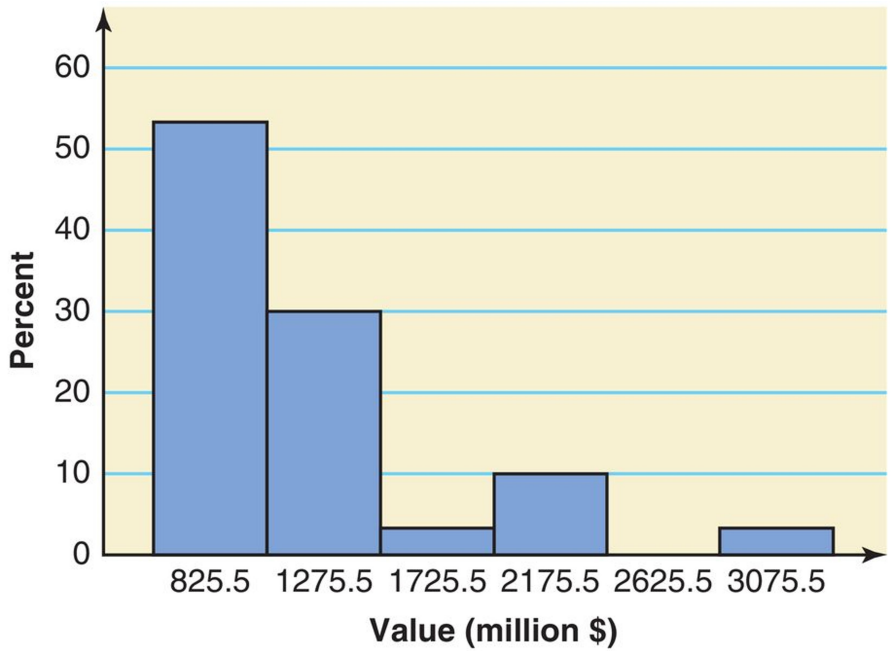
Pro velké datové vzorky, při kterých lze navýšit množství tříd (a snížit šířku jejich intervalů) se může polygon stát hladkou křivkou .Tuto křivku poté nazýváme **frekvenční křivka**.

Histogram vs. Relativní histogram

Obrázek: Histogram (četnost)

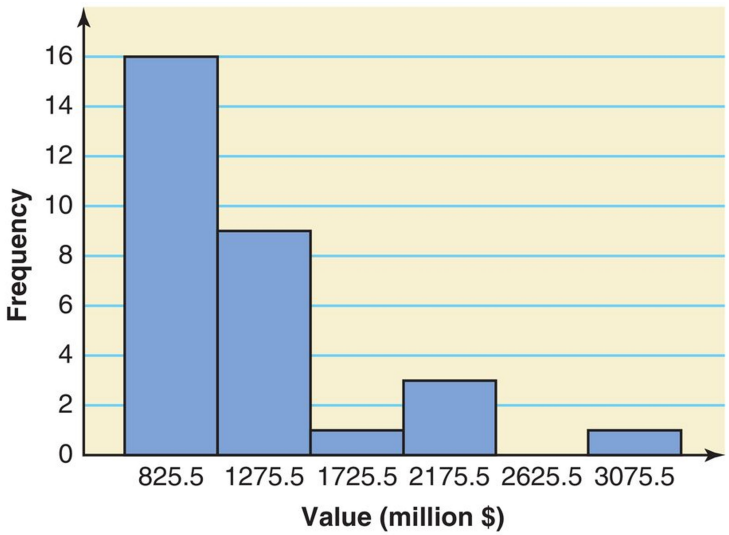


Obrázek: Histogram (% zastoupení)

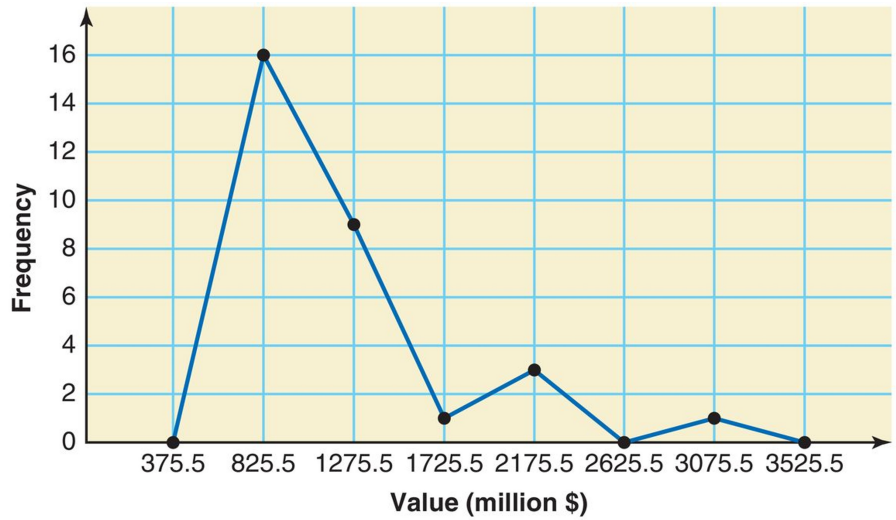


Histogram vs. Polygon

Obrázek: Histogram

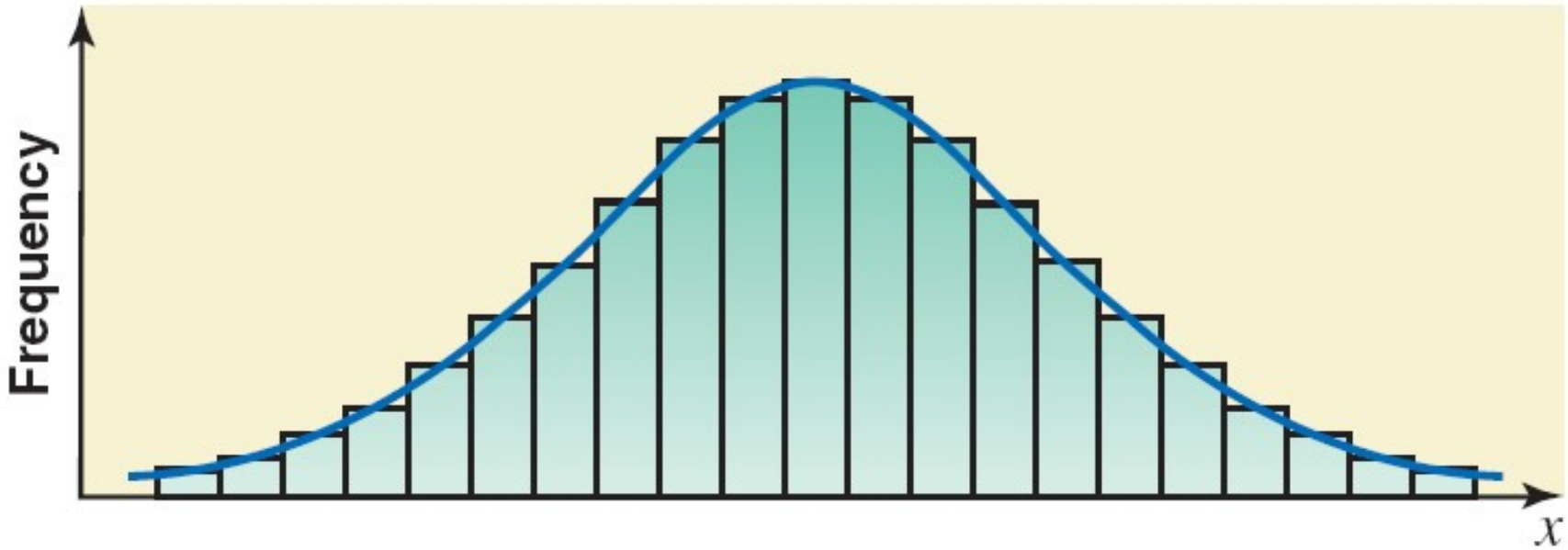


Obrázek: Polygon



Frekvenční křivka

Obrázek: Frekvenční křivka

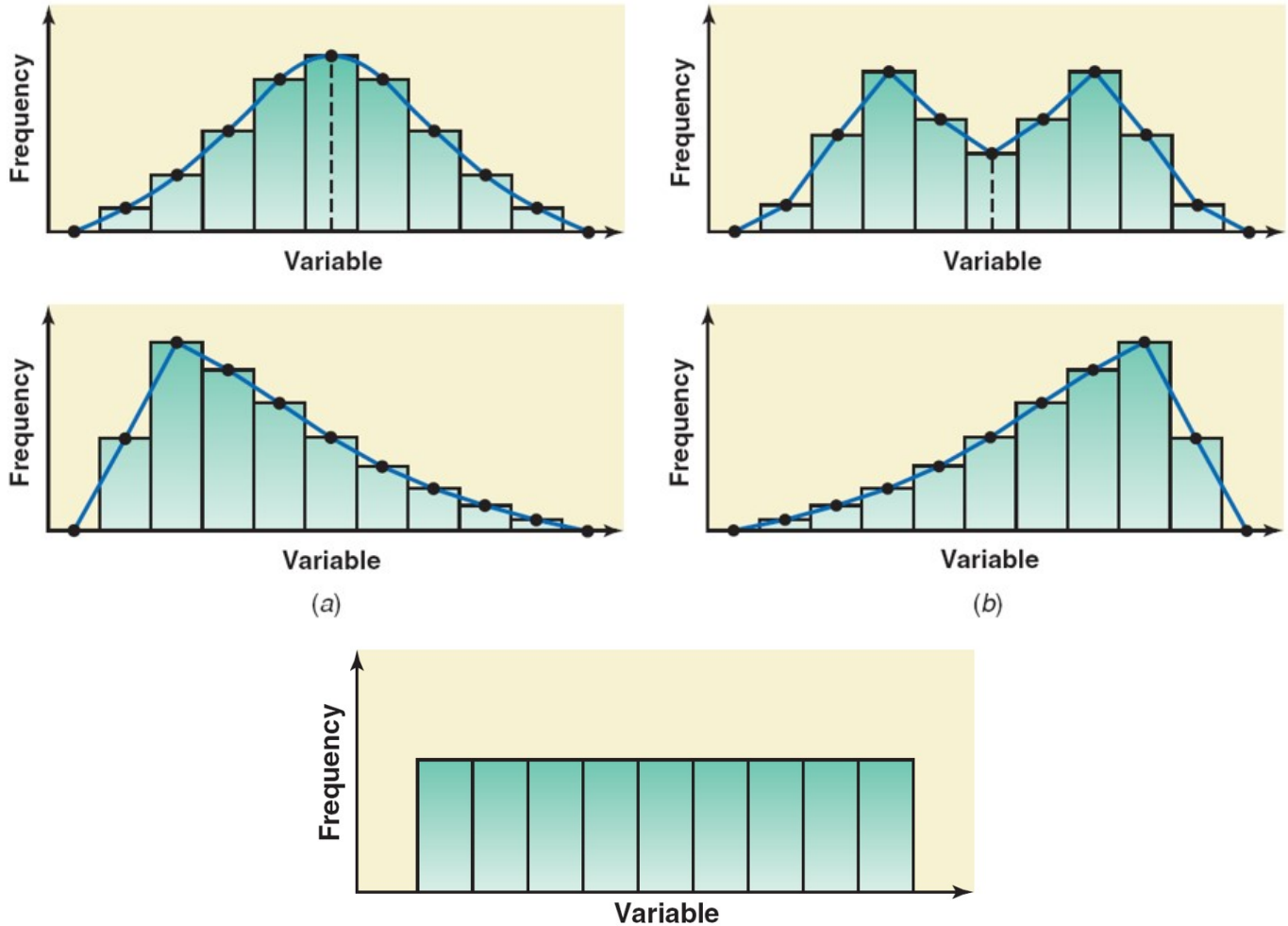


Tvary histogramů

1. **Symetrický:** Data mají symetrické rozložení kolem středové hodnoty.
2. **Zešimený:** Data mají nesymetrické rozložení s vyšší koncentrací hodnot na jedné straně histogramu než na druhé.
3. **Rovnoměrný nebo obdélníkový:** Všechny třídy mají podobné zastoupení četnosti, histogram má podobu obdélníku bez jasně výrazných vrcholů nebo sklonů.

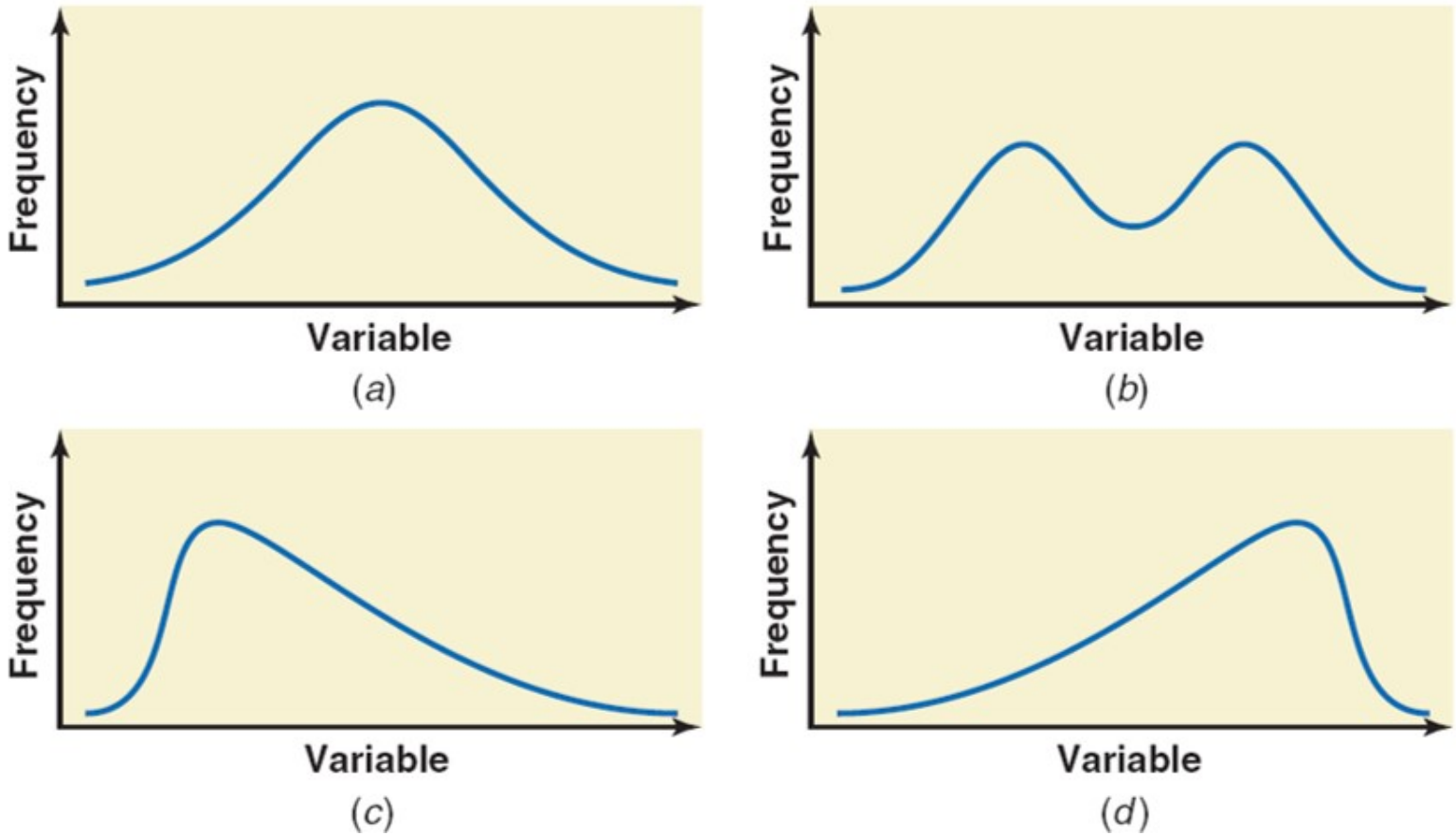
=> Při velkém datovém vzorku lze získat také frekvenční křivky. Tvary frekvenčních křivek lze pojmenovat podobně jako tvary histogramů.

Tvary histogramů



Tvary frekvenčních křivek

Obrázek: (a) a (b) Symetrické frekvenční křivky; (c) Frekvenční křivka zešikmená doprava; (d) Frekvenční křivka zešikmená doleva



Příklad

Zadání: Administrace velkého města chtěla znát rozložení počtu vozidel vlastněných domácnostmi v tomto městě. Výběrový vzorek 40 náhodně vybraných domácností z tohoto města poskytl následující údaje o počtu vlastněných vozidel:

5	1	1	2	0	1	1	2	1	1
1	3	3	0	2	5	1	2	3	4
2	1	2	2	1	2	1	1	1	1
4	2	1	1	2	1	4	1	3	3

Sestavte tabulku rozdělení četností pro tato data.

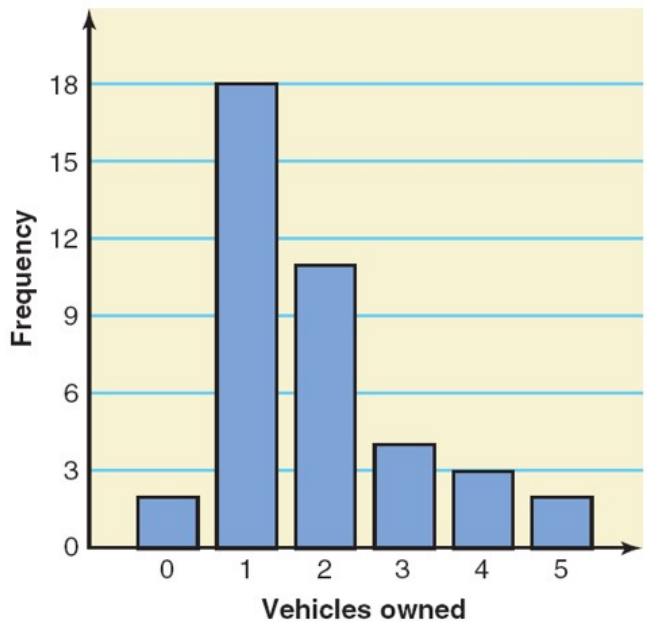
Příklad: Řešení

Pozorování předpokládají pouze šest odlišných hodnot: 0, 1, 2, 3, 4 a 5. Každá z těchto šesti hodnot je použita jako třída v rozdělení četností v následující tabulce a sloupcovém grafu.

Obrázek: Rozdělení četností

Vehicles Owned	Number of Households (f)
0	2
1	18
2	11
3	4
4	3
5	2
	$\Sigma f = 40$

Obrázek: Sloupcový graf



Kumulativní četnost a rozdělení četnosti

Kumulativní četnost udává celkový počet hodnot, které jsou menší než horní hranice intervalu zvolené třídy.

Kumulativní rozdělení četnosti udává postupné sčítání četností v každé třídě. Tímto způsobem se vytváří postupné rozložení četností v rámci celého rozsahu dat.

Kumulativní relativní četnost udává celkový počet hodnot, které jsou menší než horní hranice intervalu zvolené třídy děleno počtem všech hodnot.

Kumulativní procentní zastoupení udává celkový počet hodnot, které jsou menší než horní hranice intervalu zvolené třídy děleno počtem všech hodnot. Tato hodnota je poté vynásobena stem.

Příklad - Kumulativní rozdělení četnosti

Tabulka: Kumulativní rozdělení četnosti hodnoty basebalových týmů

Hodnota týmu (v mil. $)	Počet týmů (f)	Kumulativní četnost
601–1050	16	16
1051–1500	9	16 + 9 = 25
1501–1950	1	16 + 9 + 1 = 26
1951–2400	3	16 + 9 + 1 + 3 = 29
2401–2850	0	16 + 9 + 1 + 3 + 0 = 29
2851–3300	1	16 + 9 + 1 +3 + 0 + 1 = 30

Příklad - Kumulativní relativní četnost a procentní zastoupení

Tabulka: Kumulativní relativní četnost a procentní zastoupení

Hodnota týmu (v mil. $)	Kumulativní relativní četnost týmů	Kumulativní procentní zastoupení
601–1050	16 / 30 = 0.5333	53.33
1051–1500	25 / 30 = 0.8333	83.33
1501–1950	26 / 30 = 0.8667	86.67
1951–2400	29 / 30 = 0.9667	96.67
2401–2850	29 / 30 = 0.9667	96.67
2851–3300	30 / 30 = 1.0000	100.00

Ořezání os

Ořezání os se používá ve statistice a datové vizualizaci k popisu postupu, při kterém jsou na grafu oříznuty osy, aby se zvýraznily určité oblasti nebo trendy dat a potlačily extrémní hodnoty.

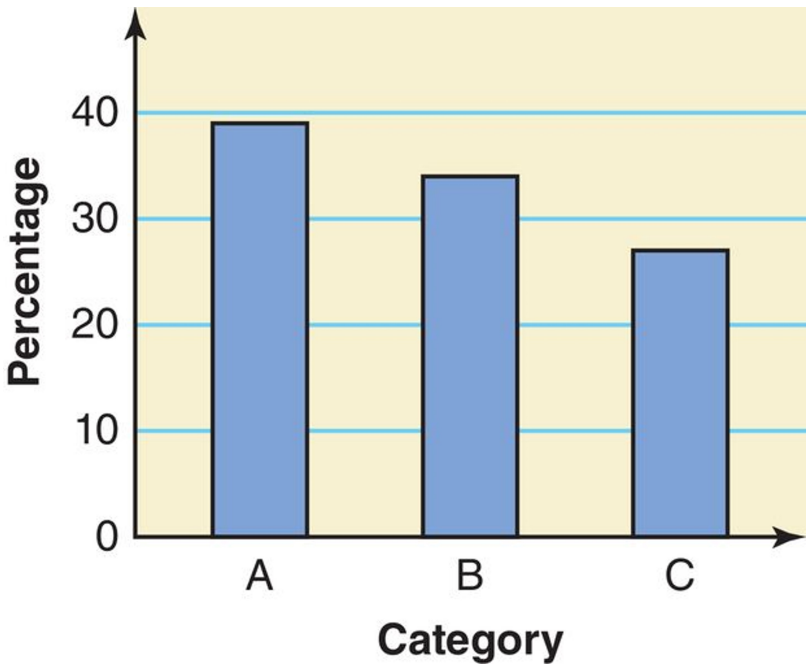
1. Změna měřítka buď na jedné nebo na obou osách, tj. zkrácení nebo protažení jedné nebo obou os.
2. Oříznutí osy frekvencí, tj. zahájení osy frekvencí číslem větším než nula.

Při využívání těchto metod je nutné být obezřetný, aby výsledný graf nebyl vzhledem ke čtenáři zavádějící, či manipulativní.

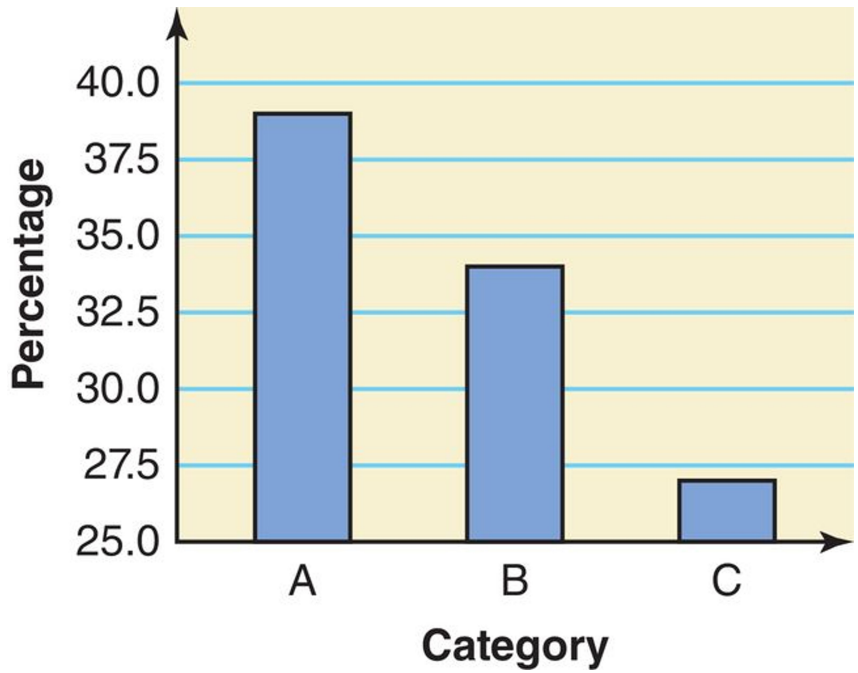
Ořezání os - příklad

Otázka: Líbí se vám kurz Statistika pro ekonomy?

Obrázek: Bez ořezání

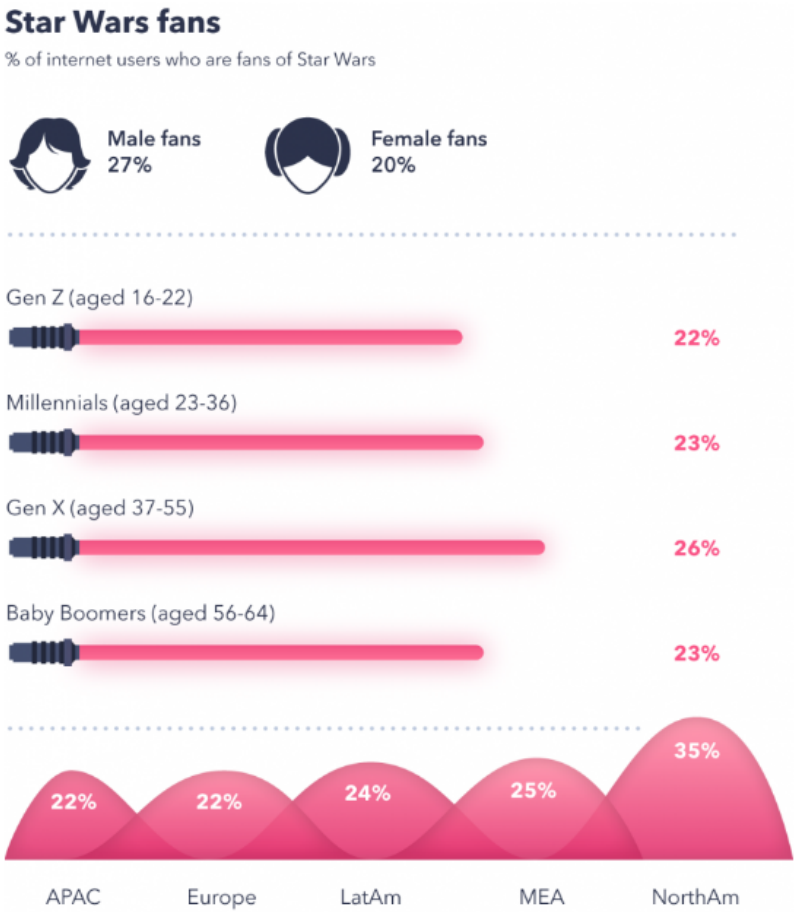


Obrázek: S ořezání



Legenda: A - Souhlasím, B - Nemám názor, C - Nesouhlasím

Příklad vizualizace kvantitativních dat na závěr sekce



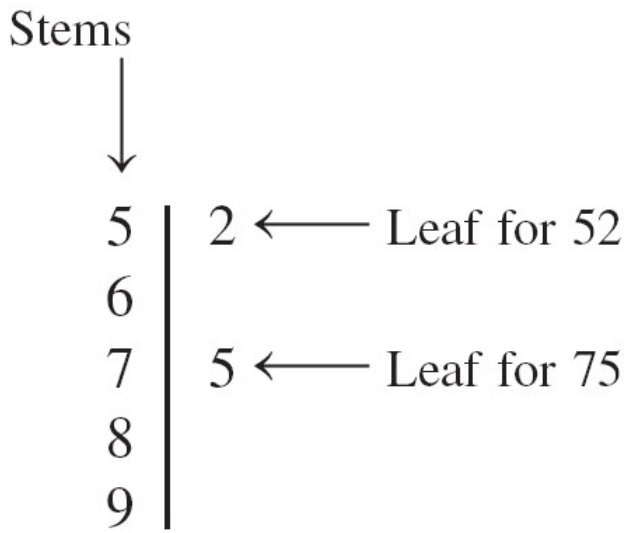
Source: GlobalWebIndex Q3 2019

- Graf ukazuje, že oslovení muži odpovídají častěji, že jsou fanoušky Star Wars než oslovené ženy.
- Z hlediska věkových skupin má Star Wars největší popularitu mezi generací X (37-55 let).
- Geograficky je nejvyšší koncentrace fanoušků v Severní Americe, kde 35 % internetových uživatelů se identifikuje jako fanoušci Star Wars. V ostatních regionech se tato hodnota pohybuje kolem 22-25 %.

Obsah

Stem-and-Leaf zobrazení dat

V **Stem-and-Leaf zobrazení** kvantitativních dat je každá hodnota rozdělena na dvě části - stonek a list. Listy pro každý stvol jsou zobrazeny samostatně v diagramu.



Příklad

Mějme skóre 30 vysokoškolských studentů na statistickém testu:

75	69	83	52	72	84	80	81	77	96
61	64	65	76	71	79	86	87	71	79
72	87	68	92	93	50	57	95	92	98

Vytvořte Stem-and-Leaf zobrazení dat a seřazené Stem-and-Leaf zobrazení dat.

Příklad: Řešení

Pro vytvoření Stem-and-Leaf zobrazení rozdělujeme každé skóre na dvě části. První část obsahuje první číslici, která se nazývá stvol. Druhá část obsahuje druhou číslici, která se nazývá list. Z dat pozorujeme, že stvoly pro všechny skóre jsou 5, 6, 7, 8 a 9, protože všechna skóre leží v rozmezí od 50 do 98. Stem-and-Leaf zobrazení pro skóre je zobrazeno na obrázku vlevo, seřazené Stem-and-Leaf zobrazení vpravo.

Obrázek: Steam and leaf zobrazení

5	2 0 7
6	5 9 1 8 4
7	5 9 1 2 6 9 7 1 2
8	0 7 1 6 3 4 7
9	6 3 5 2 2 8

Obrázek: Steam and leaf (seřazeno)

5	0 2 7
6	1 4 5 8 9
7	1 1 2 2 5 6 7 9 9
8	0 1 3 4 6 7 7
9	2 2 3 5 6 8

Obsah

Bodový graf

Bodový graf je jednoduchá grafická metoda, která zobrazuje jednotlivé pozorování nebo hodnoty dat bodem na řádku podle jejich hodnoty. Tento typ grafu je užitečný pro vizualizaci distribuce jednoduchých datových sad a umožňuje snadné porovnání jednotlivých hodnot.

Bodový graf nám může pomoci identifikovat hodnoty, které jsou velmi malé nebo velmi velké ve srovnání s většinou hodnot v datovém souboru. Tyto hodnoty se nazývají **odlehlé hodnoty** nebo **extrémní hodnoty**.

Příklad

Statistický kurz, který probíhá jednou týdně večer, má 33 studentů. Následující data udávají věk těchto studentů (v letech).

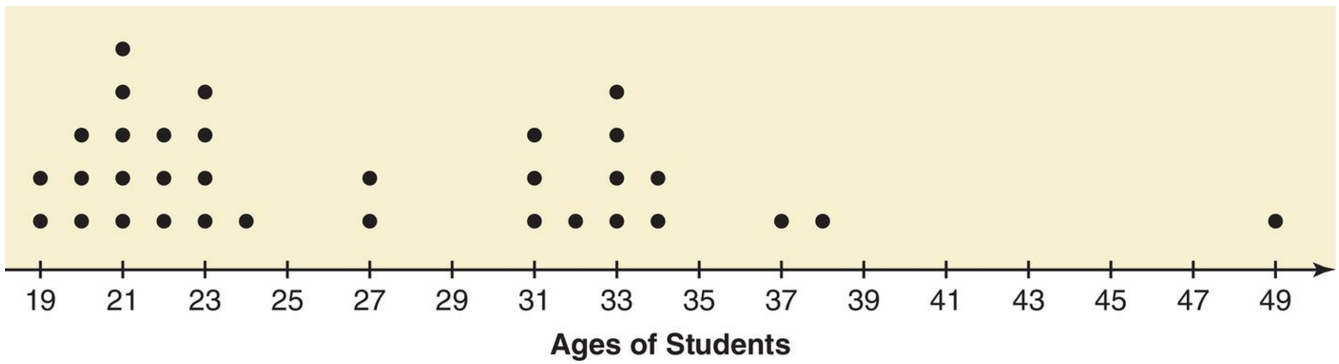
34	21	49	37	23	22	33	23	21	20	19
33	23	38	32	31	22	20	24	27	33	19
23	21	31	31	22	20	34	21	33	27	21

Vytvořte bodový graf pro tato data.

Příklad: Řešení

Krok 1. Nakreslete vodorovnou čáru s čísly, která pokrývá zadaná data.

Krok 2. Umístěte nad hodnotou na číselné ose tečku, která reprezentuje každý z věků uvedených výše. Po umístění všech teček získáme kompletní bodový graf.



Krok 3. Při zkoumání bodového grafu si všímáme, že existují dva shluky dat. Osmnáct z 33 studentů (což je téměř 55 %) je ve věku 19 až 24 let, a 10 z 33 studentů (což je asi 30 %) je ve věku 31 až 34 let. Existuje jeden student, který je 49 let starý a je odlehlou hodnotou.

Shrnutí přednášky:

Vizualizace případových studií

Organizace a vizualizace kvalitativních dat

Organizace a vizualizace kvantitativních dat

Stem-and-Leaf zobrazení dat

Bodový graf

Co si nastudovat na následující týden?

Příprava na cvičení: Leaflet 03
Koncepty a procedury 03

Povinná literatura: Mann (2016), Kapitola 3

Děkuji za pozornost!