

Základy popisné statistiky



Anotace



- Realitu můžeme popisovat různými typy dat, každý z nich se specifickými vlastnostmi, výhodami, nevýhodami a vlastní sadou využitelných statistických metod - od binárních přes kategoriální, ordinální až po spojitá data roste míra informace v nich obsažené.
- Základním přístupem k popisné analýze dat je tvorba frekvenčních tabulek a jejich grafických reprezentací – histogramů.

Typy proměnných



Kvalitativní (kategoriální) proměnná

- Ize ji řadit do kategorií, ale nelze ji kvantifikovat
- Příklady: pohlaví, HIV status.....

Kvantitativní (numerická) proměnná

- můžeme ji přiřadit číselnou hodnotu
- Příklady: výška, počet hospitalizací....

Kvalitativní znaky



- **Binární znaky:** dvě kategorie, obvykle se kódují pomocí čísel 1 (přítomnost sledovaného znaku) a 0 (nepřítomnost sledovaného znaku)

Příklady: Diabetes (1-ano, 0-ne)

Pohlaví (1-muž, 0-žena)

- **Nominální znaky:** několik kategorií (A,B,C), které nelze uspořádat

Příklad: krevní skupiny (A/B/AB/0)

- **Ordinální znaky:** několik kategorií, které lze vzájemně seřadit, tedy můžeme se ptát, která je větší/menší ($1 < 2 < 3$)

Příklady: stupeň bolesti (mírná/střední/velká)

stadium maligního onemocnění (I/II/III/IV)

Kvantitativní znaky



- **Intervalové znaky:** interpretace rozdílu dvou hodnot (stejný interval mezi jednou a druhou dvojicí hodnot vyjadřuje i stejný rozdíl v intenzitě zkoumané vlastnosti). Společný znak intervalových znaků: nula byla stanovena uměle, tedy pouhou konvencí.

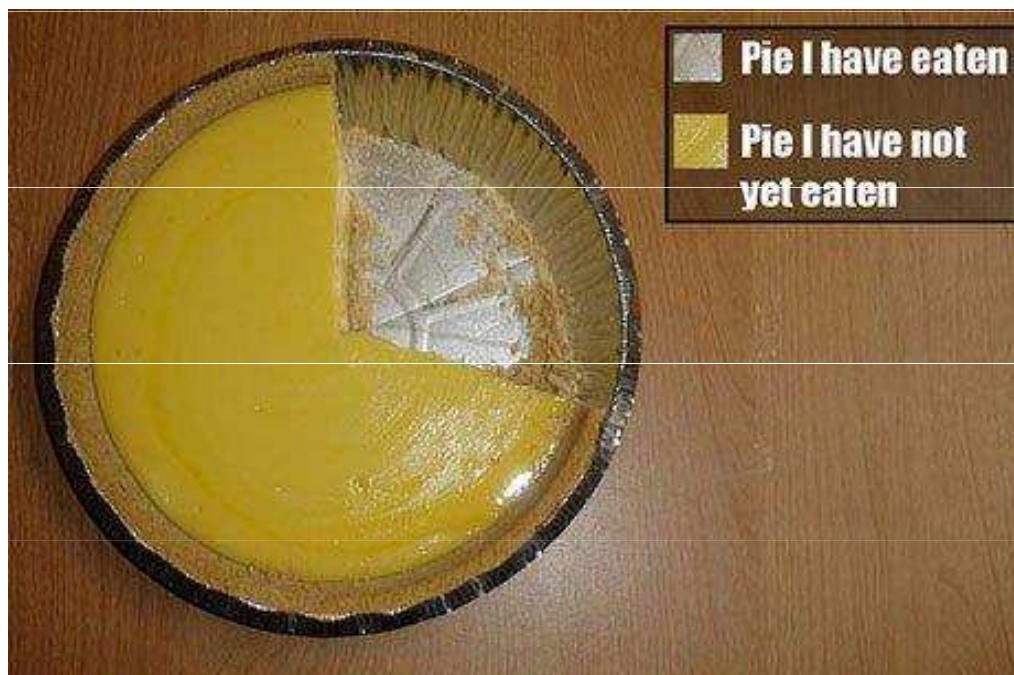
Příklad: teplota měřená ve stupních...

- **Poměrové znaky:** kromě rozdílu interpretujeme i podíl dvou hodnot

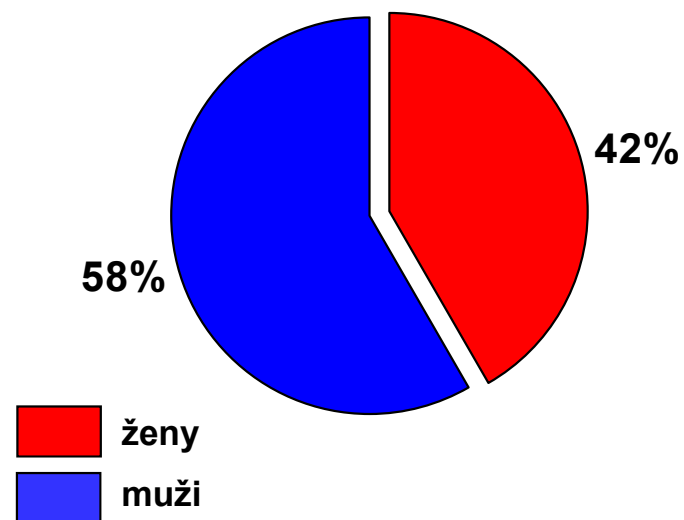
Příklady: výška v cm, váha v kg..

- *Někdy je výhodné **kvantitativní data agregovat do kategorií** (např. věk do 10ti -letých věkových skupin)- tímto krokem však ztrácíme část informace.*

Zobrazení kvalitativních dat: koláčový graf



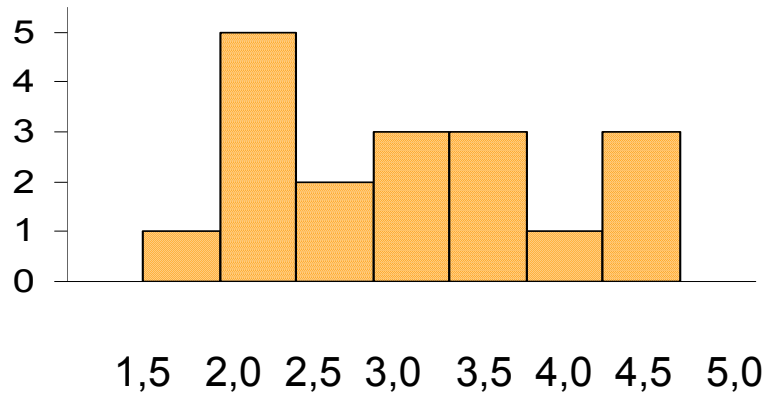
	počet	%
ženy	15	41,7%
muži	21	58,3%



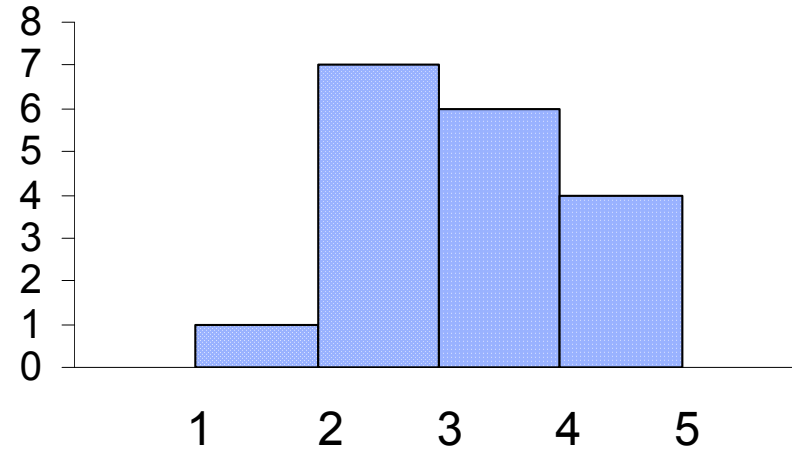
Zobrazení kvantitativních dat: histogram



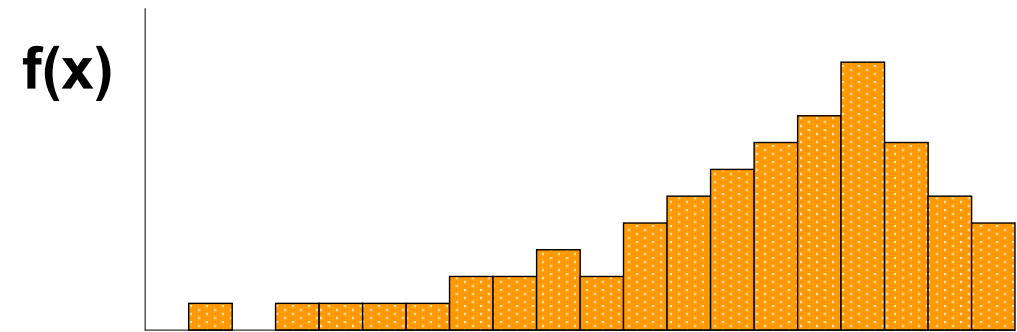
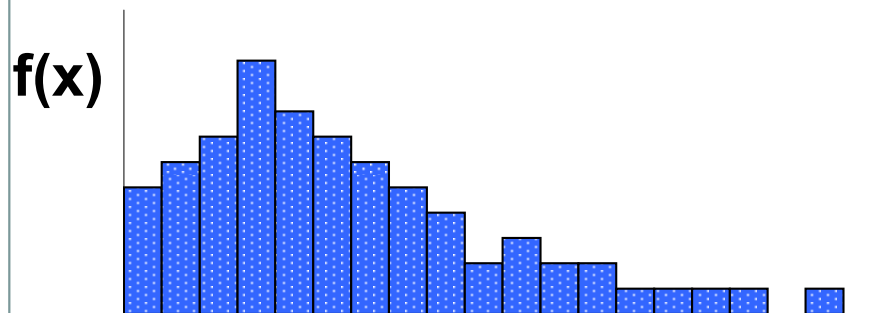
k = 10 tříd



k = 5 tříd



Histogram vyjadřuje tvar výběrového rozložení



Popisné statistiky



Charakteristiky polohy (míry střední hodnoty, míry centrální tendence)

- Udávají, kolem jaké hodnoty se data centrují, resp. které hodnoty jsou nejčastější
- **Aritmetický průměr, medián, modus, geometrický průměr**

Charakteristiky variability (proměnlivosti)

- Zachycují rozptýlení hodnot v souboru (proměnlivost dat)
- **Variační rozpětí, rozptyl, směrodatná odchylka, variační koeficient, střední chyba průměru**

Nominální znaky



Charakteristika polohy

- **Modus**: nejčastěji se vyskytující hodnota proměnné v souboru (hodnota s největší četností). V tabulce rozdělení četností se modus určí jednoduše z hodnoty znaku s největší četností.

Ordinální znaky



Charakteristika polohy

- **α -kvantil**: je-li $\alpha \in (0,1)$, pak α -kvantil x_α je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl α všech dat a na horní úsek obsahující aspoň podíl $1-\alpha$ všech dat.

- Pro speciálně zvolená α užíváme názvů:

$x_{0,50}$ - **medián**, $x_{0,25}$ - **dolní kvartil**, $x_{0,75}$ - **horní kvartil**, $x_{0,1}, \dots, x_{0,9}$ - **decily**

- **Medián** znamená hodnotu, jež dělí řadu podle velikosti seřazených výsledků na dvě stejně početné poloviny. Jestliže n je sudé číslo, pak $\tilde{x} = 0,5(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$
Jestliže n je liché číslo, pak $\tilde{x} = x_{(n+1)/2}$

Intervalové a poměrové znaky



Charakteristika polohy

- **Aritmetický průměr**: je definován jako součet všech naměřených údajů vydělený jejich počtem,

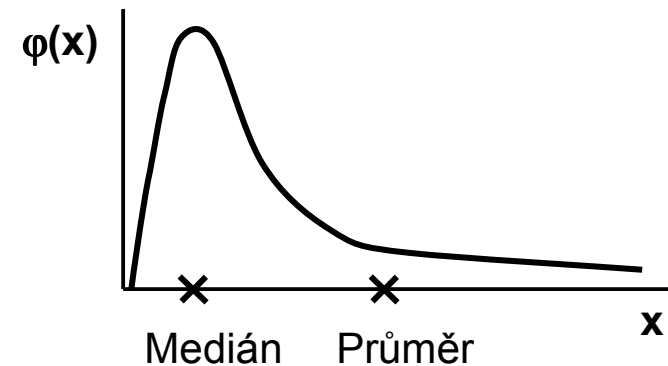
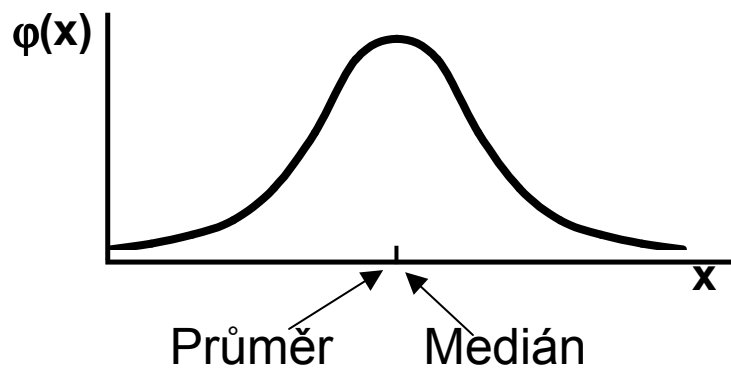
$$E(x) = \bar{x} = \sum_{i=1}^n \frac{x_i}{n} \quad \text{kde } x_i \text{ jsou jednotlivé hodnoty a } n \text{ jejich počet}$$

Průměr vs medián



PAMATUJ:

- Průměr je silně ovlivněn extrémními hodnotami (tzv. odlehlá pozorování), medián není ovlivněn vybočujícími pozorováními
- Průměr je vhodný ukazatel středu u normálního/symetrického rozložení, medián je vhodnou charakteristikou středu souboru i v případě veličin s neznámým rozdělením
- V případě symetrického rozložení jsou jejich hodnoty v podstatě shodné, v případě asymetrického rozložení však nikoliv!



Intervalové a poměrové znaky



Charakteristiky variability

- **Rozptyl (variance)** je ukazatelem šířky rozložení získaný na základě odchylky jednotlivých hodnot od průměru

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

Obdobně jako u průměru je jeho vypovídací schopnost nejvyšší v případě symetrického/normálního rozložení

- **Směrodatná odchylka (SD-standard deviation)** je druhá odmocnina z rozptylu

Další parametry rozložení



- **Počet hodnot** – důležitý ukazatel, znamená jak moc lze na data spoléhat
- **Suma hodnot**
- **Minimum, maximum**
- **Variační rozpětí** – rozdíl mezi největší a nejmenší hodnotou řady
- **Střední chyba průměru (SE)** - měří rozptýlenost vypočítaného aritmetického průměru v různých výběrových souborech vybraných z jednoho základního souboru.