Statistika I

Základy

Obsah

- Typy proměnných
- Populace x výběr
- Střední hodnota, kvantily rozptyl, směrodatná odchylka, šikmost, špičatost
- Typy rozdělení (diskrétní, spojité)
- Grafy

Typy proměnných

- **2 základní otázky:**
  - Můžeme určit pořadí hodnot?
  - Jsou od sebe hodnoty stejně vzdálené?

- **Kategoriální:**
  - **Nominální** – hodnoty nemají jasné pořadí ani stejnou vzdálenost (např. název města původu, pohlaví)
  - **Ordinální** – hodnoty mají jasné pořadí, ale ne stejnou vzdálenost (např. známky ve škole, postojové škály)
- **Kardinální** - hodnoty mají jasné pořadí a stejnou vzdálenost (např. příjem, věk)

Populace vs. Výběr/vzorek

- Populace = všichni lidé, na které naše závěry zobecňujeme
  - Reálně téměř vždy nemožné získat informace o celé populaci
  - Pokud známe informace o všech členech populace jsme schopní spočítat přesné hodnoty parametrů (střední hodnota, rozptyl, atd.)

- Vzorek = lidé, kteří byli dotazování/zkoumáni
  - Vybírán z populace, na kterou chceme naše výsledky zobecnit
  - Výběrové verze parametrů = odhady populačních parametrů
  - (např. výběrová směrodatná odchylka je odhadem reálné směrodatné odchylky)

Základní pojmy – míry polohy

- Střední hodnota (EX) = teoretický střed rozdělení
  - Nemusí být reálně dosažitelná (např. na 6stranné kostce EX = 3,5)
  - U nominálních proměnných vyjádřená modem
  - U ordinálních proměnných indikuje podobu rozdělení
  - U kardinálních proměnných funguje jako průměr/medián
- Minimum – nejmenší hodnota souboru; =min()
- Maximu – největší hodnota souboru; =max()

Základní pojmy – míry polohy

- Kvantily = hodnoty dělící soubor na stejné části
- Obecně pořadí $hodnoty = \frac{počet\ hodnot * hladina\ kvantilu}{100}$
- Různé typy kvantilů:
  - Medián (2 skupiny); =median()
  - Kvartily (4 skupiny) =quartil.inc() NEBO =quartil.exc()
  - Kvintily (5 skupin)
  - Decily (10 skupin)
  - Percentily (100 skupin) =percentil.inc() NEBO quartil.exc()

Základní pojmy – míry variability

- Rozptyl (variance. σ^2) = souhrnná míra rozdílu mezi střední hodnotou a jednotlivými proměnnými
  - Velikost závisí na měřítku (porovnávat jde jen např. u škál se stejným počtem odpovědí
- Směrodatná odchylka (SD, σ) = jednotka rozptylu
  - $= \sqrt{rozptyl}$
  - Dají se pomocí ní porovnávat konkrétní hodnoty
  - Používá se ke standardizaci ($z\ skór = \frac{x_i - EX}{SD}$; převede rozdělení EX=0, SD=1)

Základní pojmy – míry variability + šikmost a špičatost

- Rozpětí = rozdíl maxima a minima
- Mezikvartilové rozpětí (IQR) = Rozdíl 3. a 1. kvartilu

- Šikmost= míra asymetrie rozdělení pravděpodobnosti na obě strany mediánu (jak je nebo není rozdělení symetrické)
  - Různé způsoby zjišťování – grafické, koeficienty špičatosti, stat. Testováním
- Špičatost = míra „těžkosti“ konců rozdělení (jak moc pozorování se nachází na krajích)
  - Zjišťuje se obdobně jako šikmost

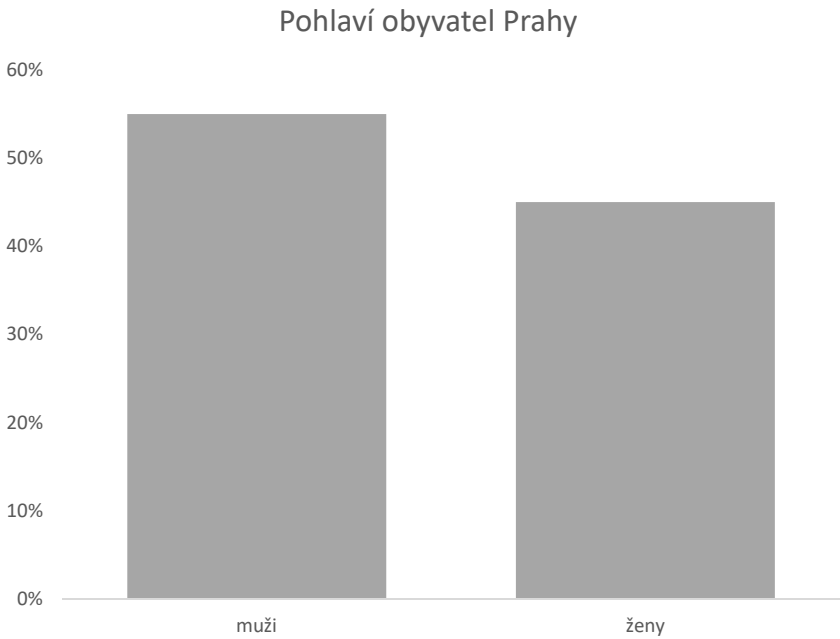
Typy rozdělení

- Diskrétní:
  - Jednotlivé hodnoty jsou od izolované (např. na kostce nikdy nehodíme 3,5)
  - Příklady: Alternativní (Bernoulliho), Rovnoměrné, Binomické, Poissonovo

- Spojité:
  - Mezi každými dvěma hodnotami existují další hodnoty (např. člověk může měřit 168,2 centimetru)
  - Příklady: Normální, Studentovo, Chí-kvadrát

Diskrétní rozdělení

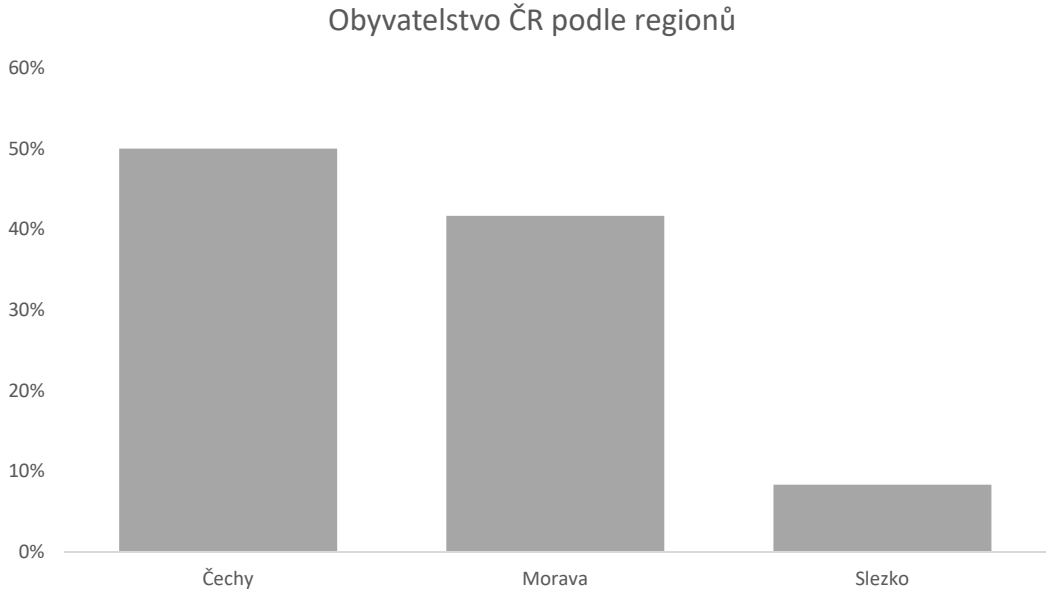
- Alternativní:
  - Pouze dva možné výsledky (úspěch/neúspěch, panna/orel,…)
  - $Střední\ hodnta = EX = \sum(x_i * p_i)$
  - $Rozptyl = \mathrm{var} = \sum[(x_i - EX)^2 * p_i]$
  - Střední hodnota říká, jaká varianta převažuje
  - Rozptyl určuje míru hetero/homogenity
  - Maximální heterogenita pokud v každé kategorii 50%



Muži = 0; ženy = 1
EX = 0,45
Var = 0,248

Diskrétní rozdělení

- Rovnoměrné diskrétní:
  - Rozšíření diskrétního (více možností)
  - $Střední\ hodnta = EX = \sum(x_i * p_i)$
  - $Rozptyl = \mathrm{var} = \sum[(x_i - EX)^2 * p_i]$
  - Střední hodnota značí, ke kterému kraji se rozdělení kloní (smysluplné jen u ordinálních a kardinálních proměnných)
  - Maximální heterogenita pokud jsou krajní kategorie zastoupeny každá z 50%



EX = 1,58
Var = 0,41

Diskrétní rozdělení

- Binomické:
  - Pravděpodobnost určitého počtu úspěchu z *n* pokusů

  - $EX = pravděpodobnost\ úspěchu * počet\ pokusů = p * k$
  - $var = EX * (1 - pravděpodobnost\ úspěchu) = EX * (1 - p)$
  - $Hustota = \binom{n}{k} * p^k * (1-p)^{n-k}$

  - Excel:
  - Dist. fce.: =binom.dist(počet úspěchů;pokusy;pravd.úspěchu;kumulovat)

  - **Binomické rozdělení se při velkém počtu pokusů blíží tvarem normálnímu**

Diskrétní rozdělení

N = 100
P = 51%

EX = 51
Var = 24,99

Pravděpodobnost, že ve vzorku bude mít 50 až 60 příznivců: 58,98%

Spojité rozdělení

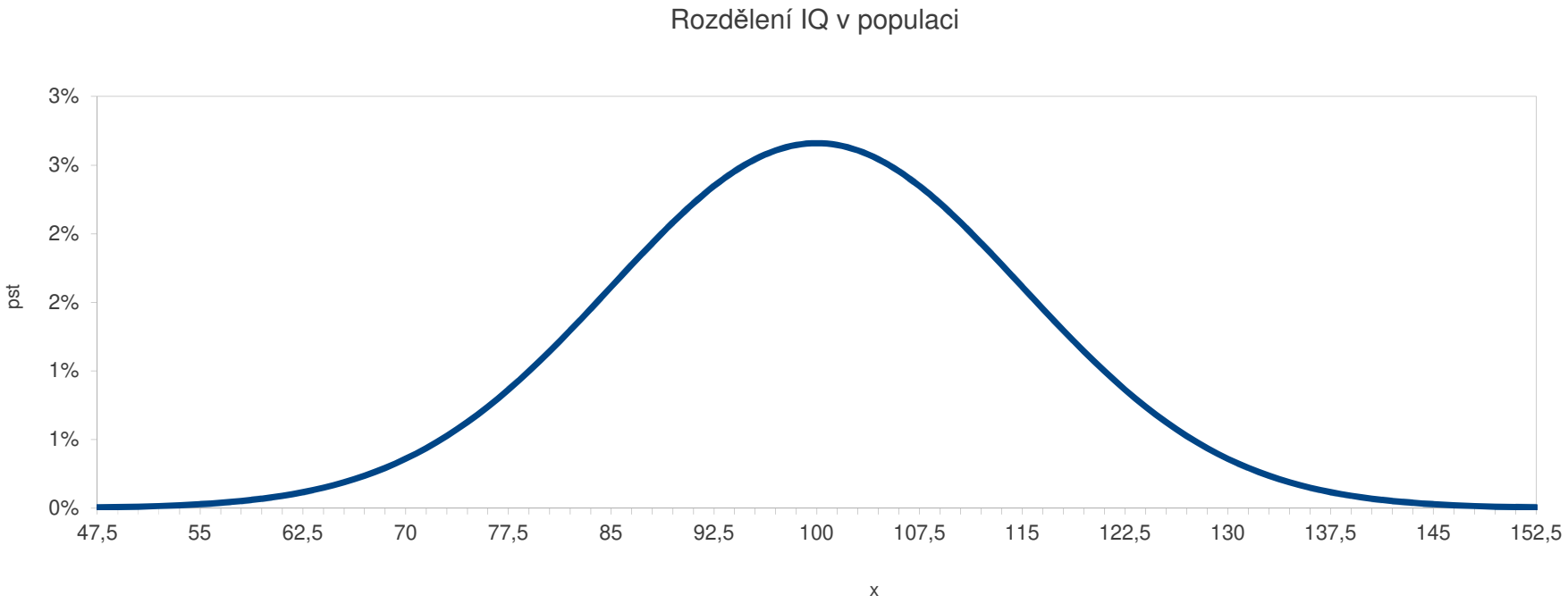
- Spojité rozdělení obecně:
  - EX = aritmetický průměr nebo medián
  - Výběrový rozptyl = $\frac{\sum(x_i - EX)^2}{N-1}$

- V Excelu:
  - EX =průměr() nebo =median()
  - Výb. Rozptyl =var.s()

Spojité rozdělení

- Normální:
  - Jedno u nejdůležitějších vůbec (základ pro statistické testy, transformace, atd.)
  - Definováno střední hodnotou a směrodatnou odchylkou
  - Nulová špičatost a šikmost (průměr se rovná mediánu)
  - Hustota = $\int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}\,\mathrm{d}t.$ v Excelu =norm.dist()
  - Platí pro něj pravidlo 3σ:
    - 68% všech pozorování se nachází v intervalu <-1σ;1σ>
    - 95% všech pozorování se nachází v intervalu <-2σ;2σ>
    - 99% všech pozorování se nachází v intervalu <-3σ;3σ>

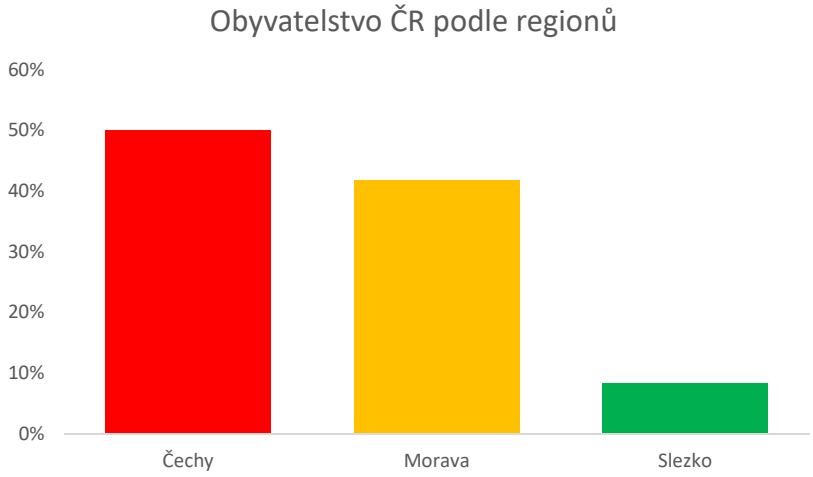
Spojité rozdělení

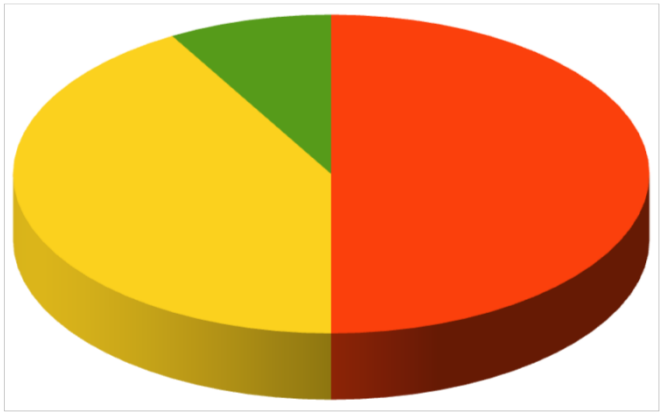


EX = 100
SD = 15

Typy grafů – Barploty a Piecharty

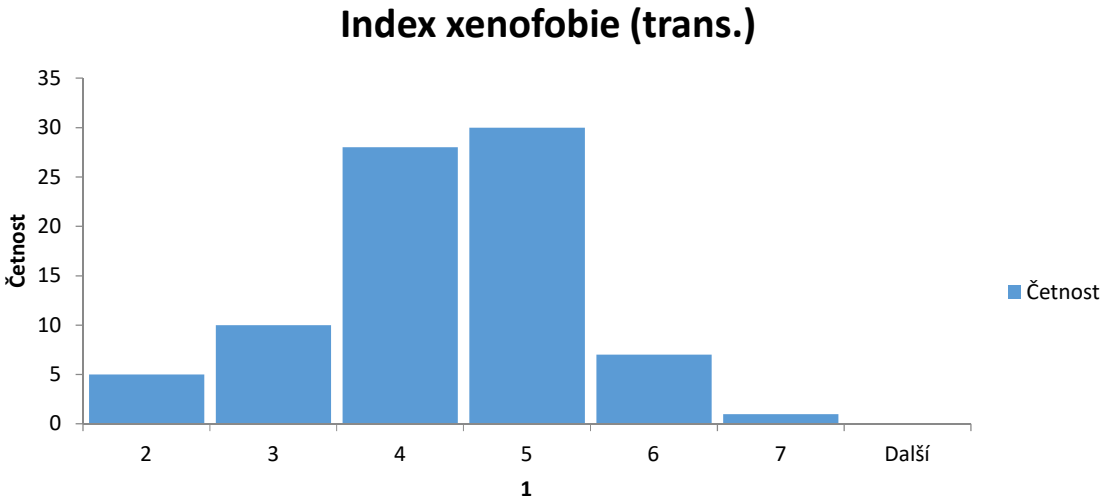
- Vyjadřují četnosti nominálních nebo ordinálních proměnných a diskrétních rozdělení
- Na piecharty pozor – mohou být zavádějící

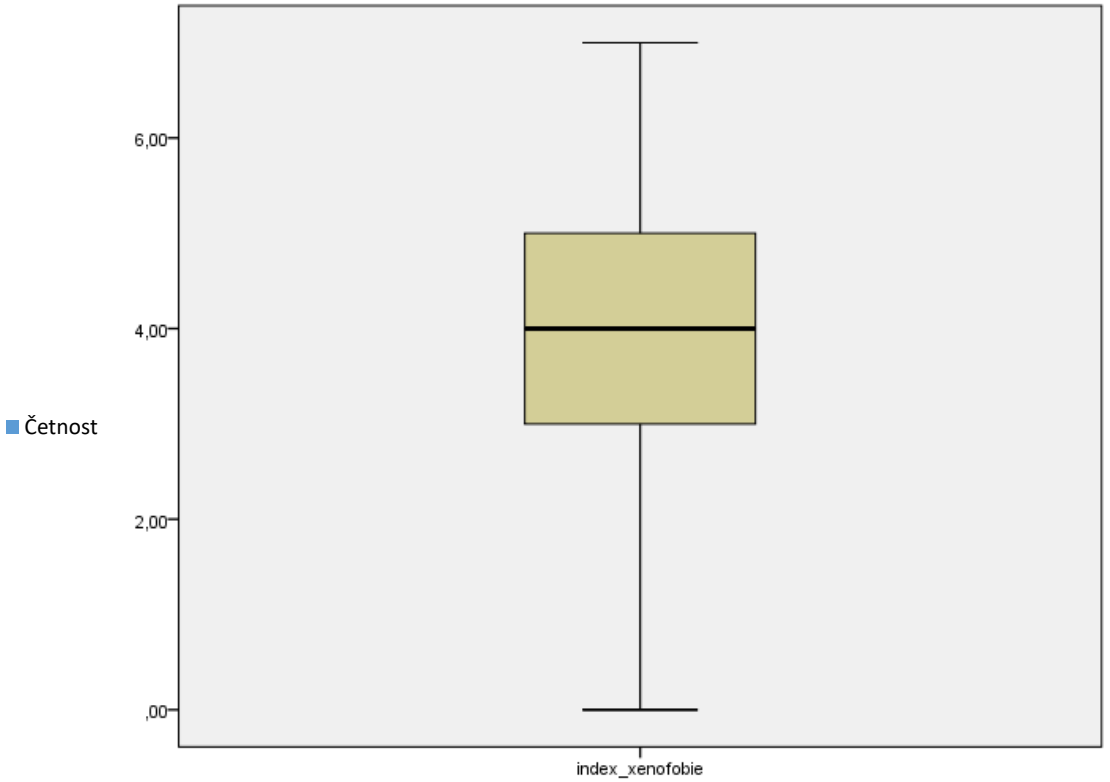




Grafy – Histogramy a boxploty

- Četnosti kardinálních proměnných a spojitých rozdělení
- Nevýhoda boxplotu – neukazuje tvar rozdělení





Grafy – Scatterploty (korelační grafy)

- Zobrazují vztah mezi dvěma kardinálnimi proměnnými (někdy se dá použít i na ordinální)