2 Výpočet číselných charakteristik - OSNOVA

- Minulá hodina → bodové/intervalové rozložení četností.
  - důvod: pilotní analýza; seznámení s daty
- Nová látka: Motivace
  - Karolína s Markétou se domluví na výzkumu. Půjdou na dvě různé školy → 20 žáků → u každého zjistí známku z matiky a anginy → výsledky roztřídí do variační tabulky → 2 variační řady → porovnávání absolutních četností pro každou dvojici známek? ... nepřehledné a neefektivní.
- Potřebujeme jednodušší charakteristiky, které nám řeknou o datech ty nejdůležitější informace a budou dostatečně jednoduché na to, aby se dali snadno vypočítat a interpretovat.
- Různá data → různé charakteristiky:
- Typy dat:
  - Nomiální
  - Ordinální
  - Intervalová
- Tři základní typy charakteristik:
  - polohy
  - variability
  - závislosti
  - + nesymetrie (intervalové znaky)

Nominální znaky

Příklad 2.1. U 100 náhodně vybraných domácností byl zjišťován způsob zásobování bramborami (znak X, varianty 1 = vlastní sklep, 2 = jinde, 3 = nákup) a bydliště (znak Y, varianty 1 = velké město, 2 = malé město, 3 = vesnice).

- = jednotlivé varianty znaku jsou neporovnatelné:
  - zvíře u veterináře: kočka, pes, papoušek, želva
  - oblast výzkumu: dolní věstonice, pohansko, klášterec
  - barva očí: modrá, zelená, hnědá
- Charakteristika polohy
  - varianty jsou navzájem neporovnatelné → můžeme vybrat pouze nejčetnější variantu ... *modus*.

```
(data <- data.frame(velke.mesto=c(13,11,19), male.mesto=c(15,7,9),
   vesnice=c(14,2,10),
                     row.names=c('sklep','jinde','nakup')))

apply(data,1,sum)
apply(data,2,sum)
```

- Charakteristika závislosti
  - Cramérův koeficient r_C - slouží k určení těsnosti závislosti u nominálních veličin
  - $r_C \in \langle 0; 1 \rangle$.

```
library(lsr)
round(cramersV(data), digits=3)
[1] 0.179
```

Ordinální znaky

Příklad 2.2. Otevřeme datový soubor znamky.txt.

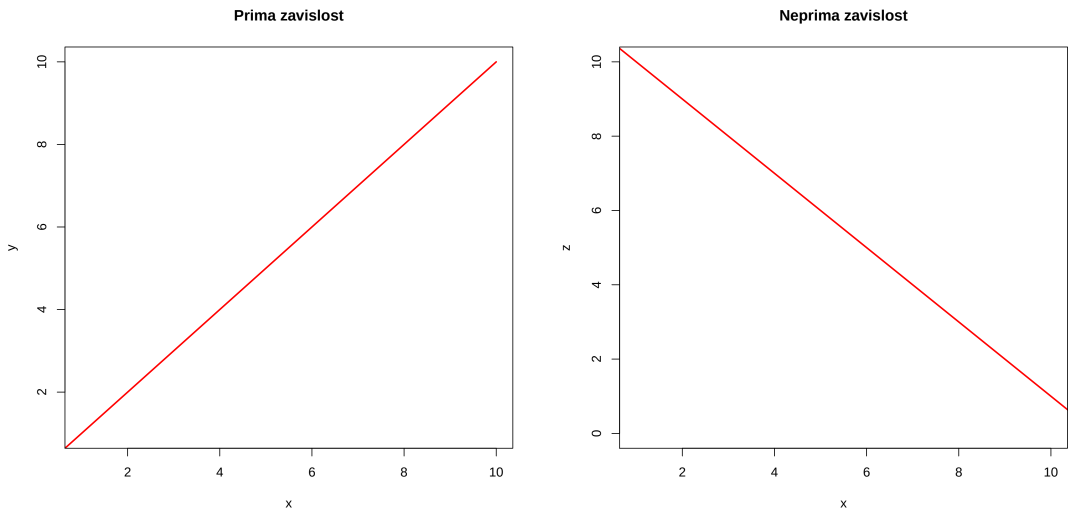
a) Pro známky z **matematiky** a angličtiny vypočteme medián, dolní a horní kvartil, kvartilovou odchylku a vytvoříme krabicový diagram.

b) Vypočteme **Spearmanův korelační koeficient** známek z matematiky a angličtiny pro všechny studenty.

- Získaná data můžeme porovnávat, ale nemůžeme říci, jaký je mezi nimi rozdíl.
  - 10 pacientů . . . pořadí podle závažnosti onemocnění
  - Známky studentů - výborně, chvalitebně, dobře, dostatečně a nedostatečně. Mezi výborně a chvalitebně je jiný rozdíl než mezi dostatečně a nedostatečně.
- Charakteristika polohy
  - α-kvantil . . . x_α
    - medián $x_{0.5}$
    - dolní kvartil $x_{0.25}$
    - horní kvartil $x_{0.75}$
  - $n\alpha =$ celé číslo $c \to x_\alpha = \dfrac{x_{(c)} + x_{(c+1)}}{2}$
  - $n\alpha =$ necelé číslo $\to$ zaokrouhlíme nahoru na nejbližší celé číslo $c \to x_\alpha = x_{(c)}$
- Charakteristika variability:
  - kvartilové rozpětí
  - $q = x_{0.75} - x_{0.25}$
  - v intervalu leží 50 % dat.

```
data <- read.delim('znamky.txt', sep='\t', dec='.',header=F)
source('AS-funkce.R')
head(data)
names(data) <- c('matematika', 'anglictina', 'pohlavi')
f3    <- factor(data$pohlavi, levels=c(0,1), labels=c('zena','muz'))
data[,3] <- f3
head(data)

matematika <- data$matematika
anglictina <- data$anglictina
pohlavi    <- data$pohlavi

q.M <- quantile(matematika, probs=c(0.5,0.25,0.75), type=2) #type=5
iqr.M <- q.M[3]-q.M[2]

(tabulka<-data.frame(median=q.M[1], kv1=q.M[2], kv3=q.M[3],
                     IQR=iqr.M, row.names='matematika'))

boxplot(matematika, anglictina, main='Krabicovy graf dvou promennych',
        names=c('matematika','anglictina'), ylab='znamka', ylim=c(0,5),
        border='darkgreen', col='darkolivegreen1')
```

- Charakteristika závislosti:
  - Spearmanův koeficient pořadové korelace r_S
  - máme dva znaky: X - známka z matematiky, Y známka z angličtiny
  - existuje mezi znaky X a Y závislost a když, jak silná?
  - $r_S \in \langle -1; 1 \rangle$.
    - $r_S > 0$ ...přímá závislost (s rostoucí hodnotou znaku X roste i hodnota znaku Y)
    - $r_S < 0$ ...nepřímá závislost (s rostoucí hodnotou znaku X hodnota znaku Y klesá)
    - $r_S = 0$ ...nezávislost




```
cor(matematika, anglictina, method='spearman')
cor(matematika[pohlavi=='zena'], anglictina[pohlavi=='zena'], method='
   spearman')
```

- Nakreslete tečkový graf

```
dotplot(matematika[pohlavi=='zena'], anglictina[pohlavi=='zena'],
        main='Teckovy␣graf␣znamek␣-␣Zeny', xlab='matematika', ylab='
            anglictina',
        col='darkgreen', bg='darkolivegreen1', xlim=c(1,4), ylim=c(1,4))
abline(v=seq(1,4,by=0.5), col='grey80', lty=2)
abline(h=seq(1,4,by=0.5), col='grey80', lty=2)
```

Intervalové znaky

Příklad 2.3. Otevřeme datový soubor lebky.txt.

a) Pro největší délku a největší šířku mozkovny mužů vypočteme aritmetický průměr, rozptyl, směrodatnou odchylku, koeficient variace, šikmost a špičatost.

b) Vypočítejte Pearsonův koeficient korelace největší délky a největší šířky mozkovny mužů. Dále vypočtěte kovarianci těchto dvou znaků a nakreslete dvourozměrný tečkový diagram.

- Hodnoty znaků můžeme nejen vzájemně porovnat, ale můžeme též říci, o kolik se liší:
- Výška/váha dětí, věk pacienta, hodnota glukózy v krvi, množství vyplaveného testosteronu, šířka lebky mužů/žen/neandrtálců, ...
- Charakteristika polohy:
  - aritmetický průměr: $m = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - součet podprůměrných hodnot je stejný, jako součet nadprůměrných hodnot
  - silně ovlivněn vybočujícími hodnotami → vhodný máme-li symetrická data
- Charakteristika polohy:
  1. rozptyl:
     - $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$
     - průměrná kvadratická odchylka hodnot od jejich aritmetického průměru.
     - $s^2 \geq 0$
     - je ovlivněn vybočujícími hodnotami → je vhodný, máme-li symetrická data
     - oproti jednotkám původních dat tato data jsou v jednotkách na druhou.
  2. směrodatná odchylka
     - $s = \sqrt{s^2}$
     - převádí rozptyl do původních jednotek
- Charakteristika nesymetrie:
  1. šikmost α_3
     - $\alpha_3 = 0 \rightarrow$ rozložení dat je symetrické
     - $\alpha_3 < 0 \rightarrow$ záporně zešikmené rozložení $\rightarrow$ prosloužený levý
     - $\alpha_3 > 0 \rightarrow$ kladně zešikmené rozložení $\rightarrow$ prosloužený pravý konec

2. špičatost α_4
   - $\alpha_4 = 0 \rightarrow$ normální rozložení dat
   - $\alpha_4 > 0 \rightarrow$ strmé rozložení dat
   - $\alpha_4 < 0 \rightarrow$ ploché rozložení dat (Říp)

```
library(e1071)
data        <- read.delim('lebky.txt', sep='\t', dec='.', header=F)
names(data) <- c('delka', 'sirka', 'pohlavi')
head(data)
delka.M   <- data$delka[data$pohlavi=='muz']
n         <- length(delka.M)

prumer.D    <- mean(delka.M)
rozptyl.D   <- 1/n*sum((delka.M-prumer.D)^2)
sm.odch.D   <- sqrt(rozptyl.D)
koef.var.D  <- sm.odch.D/mean(delka.M)*100
sikmost.D   <- skewness(delka.M, type=2)
spicatost.D <- kurtosis(delka.M, type=2)
(tab.D      <- round(data.frame(n=n, prumer=prumer.D, rozptyl=rozptyl.D, sm.
   odch=sm.odch.D,
                    koef.var=koef.var.D, sikmost=sikmost.D, spicatost=
                       spicatost.D), digits=4))
```

- Charakteristika těsnosti závislosti:
  - máme dva intervalové znaky – existuje mezi nimi nějaká závislost a když, tak jak silná?
    1. Pearsonův koeficient korelace
       * $r_{12} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - m_1}{s1}\frac{y_i - m_2}{s_2}$
       * nabývá hodnot mezi -1 a 1
       * $r_{12} > 0 \ldots$ přímá závislost
       * $r_{12} < 0 \ldots$ nepřímá závislost
       * $r_{12} = 0 \ldots$ nezávislost
    2. kovariance
       * $s_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_1)(y_i - m_2)$

```
cor(delka.M, sirka.M, method='pearson')

kovariance <- sum((delka.M-prumer.D)*(sirka.M-prumer.S))/n
round(kovariance, 4)

plot(delka.M, sirka.M, main='Teckovy␣graf␣delky␣a␣sirky␣lebky␣muzu', pch=21,
        xlab='delka␣lebky', ylab='sirka␣lebky', col='darkgreen', bg='
           darkolivegreen1')
abline(v=seq(160,200,by=5), col='grey80', lty=2)
abline(h=seq(120,145,by=5), col='grey80', lty=2)
```