2 Výpočet číselných charakteristik - OSNOVA

- Minulá hodina → bodové/intervalové rozložení četností.
  - důvod: pilotní analýza; seznámení s daty
- Nová látka: Motivace
  - Karolína s Janou 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. Ve vzorku, který tvořilo 200 studentů (100 mužů a 100 žen) byly snímány dematoglyfy dlaně (Býmová, 1990). Na otiscích bylo hodnoceno zakončení tří hlavních dlaňových linií. Podle vzorce zakončení byly jednotliví studenti rozděleni do tří kategorií: vysoká (Hi), střední (Mi) a nízká (Lo). Současně byla zhodnocena barva vlasů studentů podle standardní Fisher-Sallerové stupnice (Martin a Saller, 1957–1966, s. 391), na zákadě které byli studenti rozděleni do tří skupin: Světlá (LoH), střední (MH) a tmavá (DaH). K dispozici máme početnosti jedinců v jednotlivých kategoriích, zvlášť pro muže a zvlášť pro ženy.

A) Početnosti v jednotlivých kategoriích pro muže

	Hi	Mi	Lo
LiH	6	6	4
MH	20	15	7
DaH	18	12	12

1. Určete modus zakončení dlaňových linií a modus barvy vlasů pro muže.
2. Pomocí Cramérova koeficientu stanovte stupeň závisloti mezi zakončením dlaňových linií a barvy vlasů u mužů.

- = jednotlivé varianty znaku jsou neporovnatelné:
  - zvíře u veterináře: kočka, pes, ara, 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(vysoke=c(6,20,18),
                   stredni=c(6,15,12),
                   nizke=c(4,7,12),
                   row.names=c('svetle', 'stredni', 'tmave'))
apply(data,2,sum)
apply(data,1,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] 0.101
```

Ordinální znaky

Příklad 2.2. Otevřeme datový soubor znamky_me.txt.

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

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

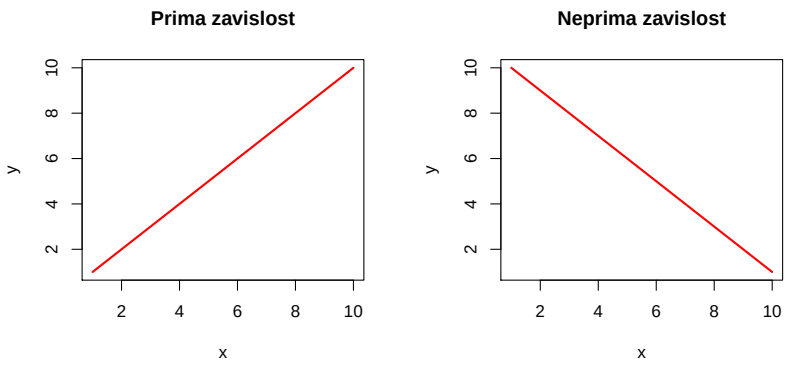
- Hodnoty můžeme porovnávat, ale nemůžeme říci, jaký je mezi nimi rozdíl.
  - 10 pacientů ... pořadí podle závažnosti onemocnění
  - Známky studentů - výborně, chvalitebně, dobře, dostatečně a nedostatečně. Mezi výborně a chvalitebně je jiný rozdíl než mezi dostatečně a nedostatečně.
- Charakteristika polohy
  - α-kvantil ... x_α

    - medián $x_{0.5}$
    - dolní kvartil $x_{0.25}$
    - horní kvartil $x_{0.75}$
  - $n\alpha$ = celé číslo $c \rightarrow x_\alpha = \dfrac{x_{(c)} + x_{(c+1)}}{2}$
  - $n\alpha$ = necelé číslo $\rightarrow$ zaokrouhlíme nahoru na nejbližší celé číslo $c \rightarrow x_\alpha = x_{(c)}$

- Charakteristika variability:
  - kvartilové rozpětí
  - $q = x_{0.75} - x_{0.25}$
  - v intervalu leží 50 % dat.

```
data <- read.delim('znamky_me.txt', sep='\t', dec='.',header=T)
source('AS-funkce.R')

matematika <- data$math
anglictina <- data$english
pohlavi    <- data$sex

qA <- quantile(anglictina, probs=c(0.5,0.25,0.75), type=2) # type=5
iqrA <- qA[3]-qA[2]

(tabA<-data.frame(median=qA[1], kv1=qA[2], kv3=qA[3],
                  IQR=iqrA, row.names='anglictina'))

boxplot(matematika, anglictina, main='Krabicovy graf',
        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
    - $r_S < 0$ ... nepřímá závislost
    - $r_S = 0$ ... nezávislost



```
cor(matematika, anglictina, method='spearman')
```

- Nakreslete tečkový graf

```
dotplot(matematika, anglictina,
        main='Teckovy graf', 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)
```

Příklad 2.3. Vypočítejte medián, dolní a horní kvartil, kvartilovou odchylku a vytvořte krabicový diagram pro známky z angličtiny, když víte, že absolutní četnosti známek byly

známka	1	2	3	4
absolutní četnost	4	4	7	5

```
anglictina <- c(1,1,1,1,2,2,2,2,3,3,3,3,3,3,3,4,4,4,4,4)
q.A <- quantile(anglictina, probs=c(0.5,0.25,0.75), type=2) #type=5
iqrA <- qA[3]-qA[2]

(tabA<-data.frame(median=qA[1], kv1=qA[2], kv3=qA[3],
                  IQR=iqrA, row.names='anglictina'))
```

Intervalové znaky

Příklad 2.4. 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. Nakreslete dvourozměrný tečkový diagram.

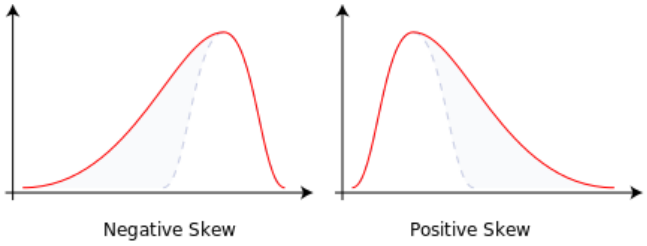
- Hodnoty znaků můžeme nejen vzájemně porovnat, ale můžeme též říci, o kolik se liší:
- Výška/váha dětí, hodnota glukózy v krvi, množství vyplaveného testosteronu, šířka lebky mužů/žen/neandrtálců, ...
- Charakteristika polohy:
  - aritmetický průměr: $m = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - ovlivněn vybočujícími hodnotami $\rightarrow$ vhodný máme-li symetrická data
- Charakteristika polohy:

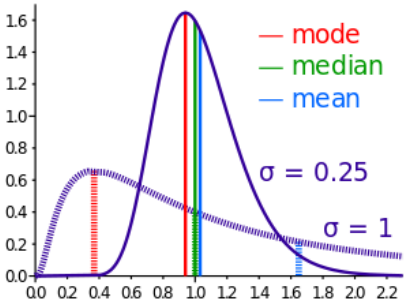
1. rozptyl:
   - $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$
   - průměrná kvadratická odchylka hodnot od jejich aritmetického průměru.
   - $s^2 \geq 0$

  - ovlivněn vybočujícími hodnotami → je vhodný, máme-li symetrická data
  - oproti jednotkám původních dat je rozptyl v jednotkách ∧ 2.
2. směrodatná odchylka
  - $s = \sqrt{s^2}$
  - převádí rozptyl do původních jednotek

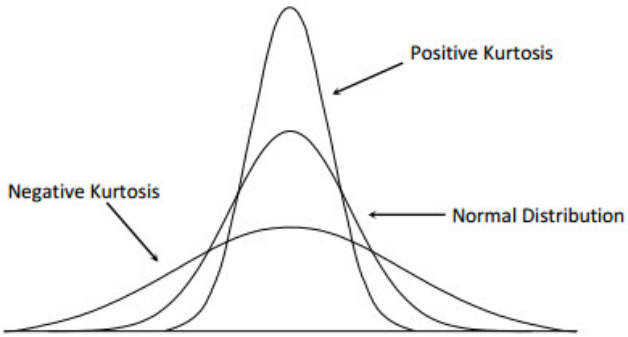
- Charakteristika nesymetrie:

1. šikmost α_3
  - $\alpha_3 = 0 \rightarrow$ symetrické rozdělení dat
  - $\alpha_3 < 0 \rightarrow$ záporně zešikmené rozdělení → prodloužený levý
  - $\alpha_3 > 0 \rightarrow$ kladně zešikmené rozdělení → prodloužený pravý konec





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

mean.D      <- mean(delka.M)
s2.D        <- 1/n*sum((delka.M-mean.D)^2)
s.D         <- sqrt(s2.D)
koef.var.D  <- s.D/mean.D*100
sikmost.D   <- skewness(delka.M, type=2)
spicatost.D <- kurtosis(delka.M, type=2)
tab.D       <- round(data.frame(n=n, prumer=mean.D, rozptyl=s2.D, sm.odch=s.
   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
       * $r_{12} > 0$ ... přímá závislost
       * $r_{12} < 0$ ... nepřímá závislost
       * $r_{12} = 0$ ... 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')

plot(delka.M, sirka.M, main='Teckovy graf', 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)
```