

# 1 1 - Bodové a intervalové rozložení četností - OSNOVA

## Úvod

- jde o tzv. pilotní analýzu
- Motivace: Někdo nám poskytne data → chceme se s nimi seznámit, pochopit jejich podstatu, nějak si je graficky znázornit a získat nad nimi nadhled
- podle typu dat volíme různé způsoby reprezentace a vizualizace dat (jiné grafy používáme pro diskrétní data (známky, pohlaví, typ pracího prášku, apod.) a jiné pro spojitá data (výška a váha člověka, krevní tlak, přesný věk pacienta, apod.)
- záleží také, zda zkoumáme pouze jednu vlastnost, nebo více vlastností najednou
- Podle toho, jaká data máme, používáme buď jednorozměrné/vícerozměrné bodové/intervalové rozdělení četností
- orientace v přehledu typů rozdělení

## Jednorozměrné bodové rozdělení četností

**Příklad 1.1.** Načtete soubor `znamky.txt`. Znakům X, Y, Z vytvořte návěští (X - známka z matematiky, Y - známka z angličtiny, Z - pohlaví studenta). Popište, co znamenají jednotlivé varianty (u znaků X a Y: 1 - výborně, 2 - chvalitebně, 3 - dobře, 4 - neprospěl, u znaku Z: 0 - žena, 1 - muž).

```
data <- read.delim('znamky.txt', sep=' ', dec='')
```

```
##   V1 V2 V3
##  1  2  2  0
##  2  1  3  1
##  3  4  3  1
##  4  1  1  0
##  5  1  2  1
##  6  4  4  1
```

- Popis tabulky:
  - Dvaceti žáků jsme se zeptali, jakou dostali na konci roku známku z matematiky (1.sloupec) a z angličtiny (2.sloupec) a zaznamenali jsme si jejich pohlaví (3.sloupec)
  - v každém jednom řádku jsou informace o jednom konkrétním žákovi
  - žák ... **objekt** našeho zkoumání
  - známka z matematiky/angličtiny a pohlaví ... **znaky** každého objektu (žáka)
  - znaku můžeme přiřadit **konkrétní číslo**, které má samo o sobě výpovědní hodnotu (známka z předmětu), nebo jde o **kódování** jisté vlastnosti: např. 0-žena, 1-muž
    - \* znak *pohlaví* je příklad kódování 0 = ženy, 1 = muži
    - \* 1 - výborně, 2 - chvalitebně, 3 - dobře, 4 - neprospěl

- známky 1 - výborně, 2 - chvalitebně, 3 - dobře, 4 - neprospěl ... **varianty** znaku známka

```
data <- read.table('znamky.txt', sep='\t', dec='.')
head(data)

f1 <- factor(data$matematika, levels=c(1,2,3,4),
             labels=c('vyborne','chvalitebne','dobre','nedostatecne'))
f2 <- factor(data$anglictina, levels=c(1,2,3,4),
             labels=c('vyborne','chvalitebne','dobre','nedostatecne'))
f3 <- factor(data$pohlavi, levels=c(0,1), labels=c('zena','muz'))
data2 <- data.frame(f1, f2, f3)
names(data2) = c('matematika','anglictina','pohlavi')
head(data2)
```

### Příklad 1.2. Vytvořte

- variační řadu známek z **matematiky** a angličtiny;
- sloupkový diagram absolutních četností znaků **X=Matematika** a **Y=Angličtina**;
- polygon absolutních četností znaků **X=Matematika** a **Y=Angličtina**.

- pro každou variantu můžeme stanovit její

- absolutní četnost  $n_j$ 
  - \* kolik žáků mělo známku 2
- relativní četnost  $p_j$ 
  - \* poměr žáků, kteří měli z matiky 2 ku celkovému počtu žáků
  - \*  $p_j * 100$  - kolik % žáků mělo známku 2
- absolutní kumulativní četnost  $N_j$ 
  - \* kolik žáků mělo známku  $\leq 2$
- relativní kumulativní četnost  $F_j$ 
  - \* poměr žáků, kteří měli známku z matiky  $\leq 2$  vzhledem k celkovému počtu žáků
  - \*  $F_j * 100$  - kolik % žáků mělo známku  $\leq 2$
- všechny výše zmíněné četnosti můžeme zapsat do přehledné tabulky ... **variační řady**

- teď si naprogramujeme

```
– variační řadu pro známky z matematiky
matematika <- data2$matematika
n1 <- sum(matematika=='vyborne')
n2 <- sum(matematika=='chvalitebne')
n3 <- sum(matematika=='dobre')
n4 <- sum(matematika=='nedostatecne')

nj <- c(n1, n2, n3, n4)
n <- sum(nj)
pj <- nj/n
Nj <- cumsum(nj)
```

```

Fj <- cumsum(pj)

nazvy.znamek <- c('vyborne', 'chvalitebne', 'dobre', 'nedostatecne')
variacni.rada <- data.frame(nj=nj, Nj=Nj, pj=pj, Fj=Fj, row.names =
  nazvy.znamek)
variacni.rada

(VR.Mat <- variacni_rada(X=matematika, nazvy=nazvy.znamek)

```

– Sloupkový diagram

```

# Matematika
barplot(VR.Mat$nj, col='white', border='white', axes=T,
  xlab='Znamka', ylab='Pocet_pozorovani', names=nazvy.znamek,
  main='Sloupkovy_diagram_pro_predmet_matematika')
abline(h=0:9, col='grey80', lty=2)
barplot(VR.Mat$nj, col='blue', axes=F, density=20, border='darkblue',
  add=T)

```

– Polygon četností

```

plot(1:4, VR.Mat$nj, type='n', xlim=c(0.5,4.5), ylim=c(1,9),
  xlab='Znamka', ylab='Absolutni_cetnost',
  main='Polygon_cetnosti_pro_predmet_matematika', axes=F)
abline(h=0:9, col='grey80', lty=2)
abline(v=0:9, col='grey80', lty=2)

lines(1:4, VR.Mat$nj, col='darkblue', lwd=2 )
points(1:4, VR.Mat$nj, col='darkblue', pch=20, cex=1.2)
axis(1, at=0:5, lab=c('', nazvy.znamek, ''))
axis(2, at=0:10)

```

**Příklad 1.3.** Vytvořte variační řady známek z matematiky a angličtiny pouze

a) pro ženy,

```

pohlavi <- data2$pohlavi
variacni_rada(X=matematika[pohlavi=='zena'], nazvy=nazvy.znamek)

```

b) pro muže.

## Dvourozměrné bodové rozložení četností

**Příklad 1.4.** Nadále budeme pracovat s celým datovým souborem. Vytvoříme kontingenční tabulku simultánních absolutních četností znaků **X=Matematika** a **Y=Angličtina**.

- vezměme si nyní z datové tabulky známky z matematiky (znak X) a z angličtiny (znak Y)
- dvourozměrný datový soubor;  $X \dots 4$  varianty;  $Y \dots 4$  varianty
- pro každou dvojici variant (celkem 16 dvojic) můžeme stanovit
  - $n_{jk} \dots$  **simultánní absolutní četnost** dvojice znaků  $x_{[j]}$  a  $y_{[k]}$ 
    - \*  $n_{jk} = \text{pocet}(X = x_{[j]} \text{ a } Y = y_{[k]})$
    - \* počet studentů, kteří měli z matiky 1 a z angličtiny 1, ...

- $n_{.j}$  ... **marginální absolutní četnost varianty  $x_{[j]}$** 
  - $n_{.j} = n_{j1} + \dots + n_{j4}$
  - počet studentů, kteří měli z matiky 1 bez ohledu na to, co měli z angličtiny
- $n_{.k}$  ... **marginální absolutní četnost varianty  $y_{[k]}$** 
  - $n_{.k} = n_{1k} + \dots + n_{4k}$
  - počet studentů, kteří měli z angličtiny 1 bez ohledu na to, co měli z matematiky

```
K.Tab <- table(matematika, anglictina)
K.Tab2 <- cbind(K.Tab, suma=apply(K.Tab, 1, sum))
(K.Tab3 <- rbind(K.Tab2, suma=apply(K.Tab2, 2, sum)))
```

**Příklad 1.5.** Vytvořte kontingenční tabulku řádkově a sloupcově podmíněných relativních četností znaků  $X$ =Matematika a  $Y$ =Angličtina.

- $p_{k(j)}$  ... **řádkově podmíněná relativní četnost varianty  $y_{[k]}$  za předpokladu  $x_{[j]}$** 
  - $p_{k(j)} = \frac{n_{jk}}{n_{.j}}$
  - poměr počtu studentů, kteří měli z matiky 1 a z angličtiny 1 vzhledem k počtu studentů, kteří měli z matiky 1
- $p_{j(k)}$  ... **sloupcově podmíněná relativní četnost varianty  $x_{[j]}$  za předpokladu  $y_{[k]}$** 
  - $p_{j(k)} = \frac{n_{jk}}{n_{.k}}$
  - poměr počtu studentů, kteří měli z matematiky 1 a z angličtiny 1 vzhledem k počtu studentů, kteří měli z angličtiny 1

```
Tab <- table(matematika, anglictina)
# Radkove podmínene relativni cetnosti
round(prop.table(Tab, margin=1), digits=3)

# Sloupcove podmínene relativni cetnosti
round(prop.table(Tab, margin=2), digits=3)
```

## Intervalové rozdělení četností

Práci s intervalovým rozložením četností si ukážeme na datovém souboru lebky.txt.

**Popis datového souboru:** Máme k dispozici údaje o rozměrech lebek staroegyptské populace. Jedná se o 216 mužů a 109 žen. Znak X ... největší délka mozkovny v mm Znak Y ... největší šířka mozkovny v mm Znak Z ... pohlaví osoby (1–muž, 0–žena)

**Příklad 1.6.** Načtete soubor lebky.txt. Podle Sturgersova pravidla najděte optimální počet třídících intervalů pro znaky X a Y a vhodně stanovte meze třídících intervalů, a to zvlášť **pro muže** a zvlášť pro ženy.

- spojitá data  $\rightarrow$  třídíme je do intervalů  $(\infty; u_1)$ ,  $(u_1; u_2)$ ,  $\dots$ ,  $(u_r; u_{r+1})$ ,  $(u_{r+1}; \infty)$

- $(u_j; u_{j+1})$  ...  $j$ -tý třídící interval
- třídící intervalu vyvolíme stejně dlouhé
- Sturgesovo pravidlo  
 $r \approx 1 + 3.3 \log_{10} n$

```
data      <- read.delim('lebky.txt', sep='\t', dec='.', header=F)
names(data) <- c('delka', 'sirka', 'pohlavi')
head(data)

# Muži
data.M    <- data[data$pohlavi=='muz',]
n.M       <- dim(data.M)[1]
(Sturges.M <- round(1+3.3*log10(n.M), digits=0))

delka.M   <- data.M$delka
range(delka.M)
max(delka.M) - min(delka.M)
round((max(delka.M) - min(delka.M))/Sturges.M, digits=0)
```

**Příklad 1.7.** Vytvořte histogram pro  $X$  a pro  $Y$  (s uvedenými absolutními a relativními četnostmi jednotlivých třídících intervalů), a to zvlášť **pro muže** a zvlášť pro ženy.

```
hist(delka.M, breaks=seq(163, 199, by=4), ylim=c(0,52),
     main='Histogram', xlab='Delka┐lebky', ylab='Pocetnosti',
     col='white', border='white', density=20, axes=F)
abline(h=seq( 0, 60, by=10), col='grey80', lty=2)
hist(delka.M, breaks=seq(163, 199, by=4),
     col='blue', border='darkblue', density=20, add=T)
axis(1, at=seq(165, 197, by=4))
axis(2, at=seq( 0, 50, by=10))
box(bty='l')
```