

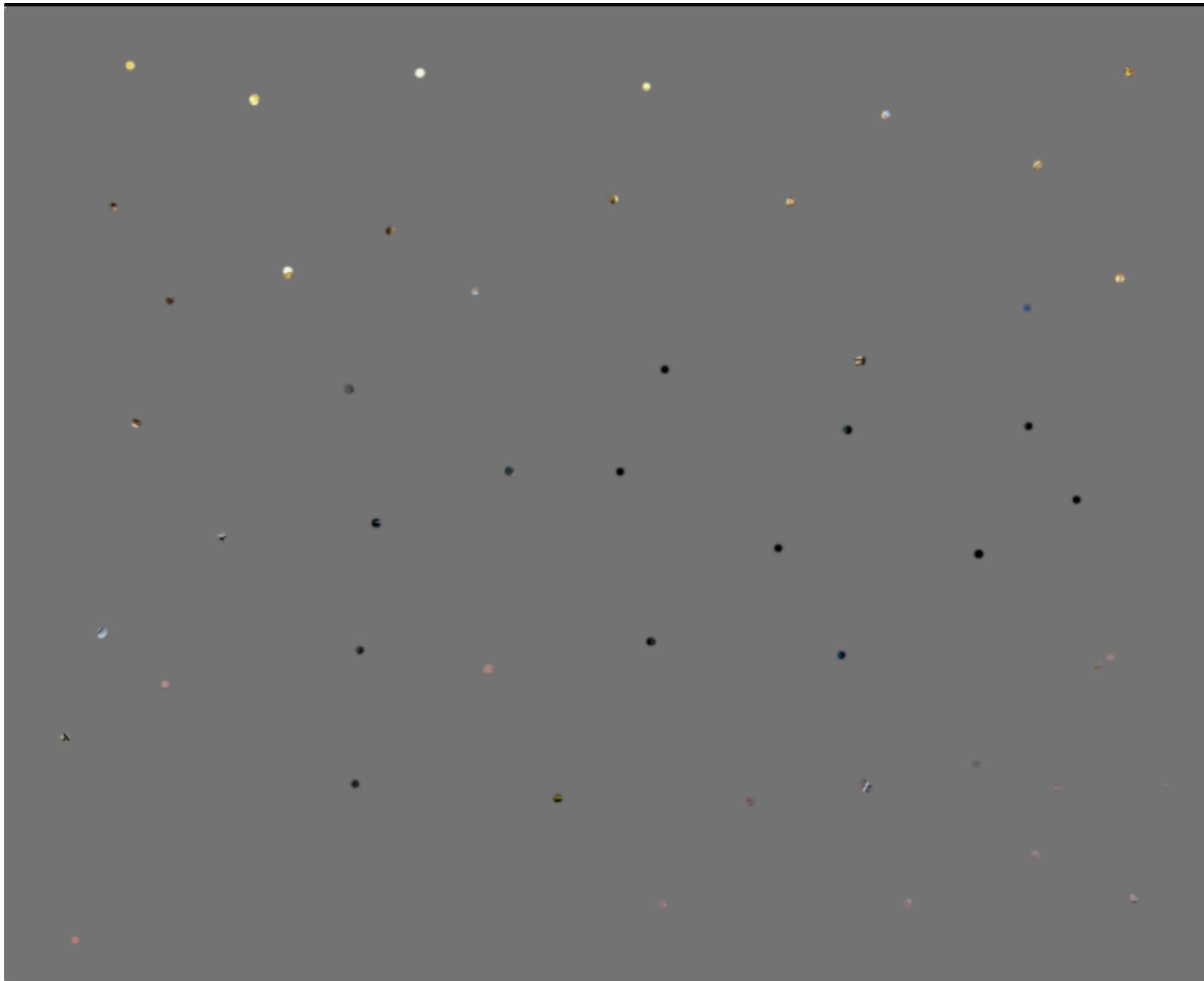
Základy pravděpodobnosti a matematické statistiky

7. Úvod do matematické statistiky

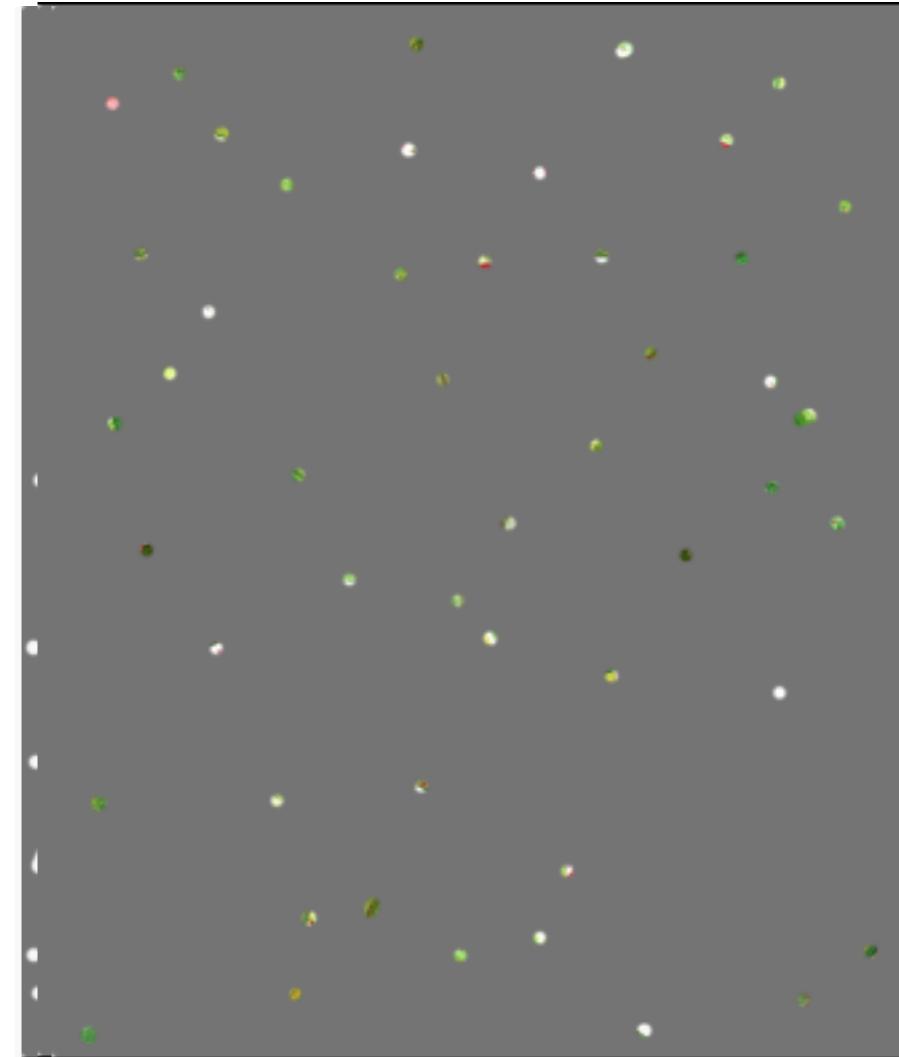
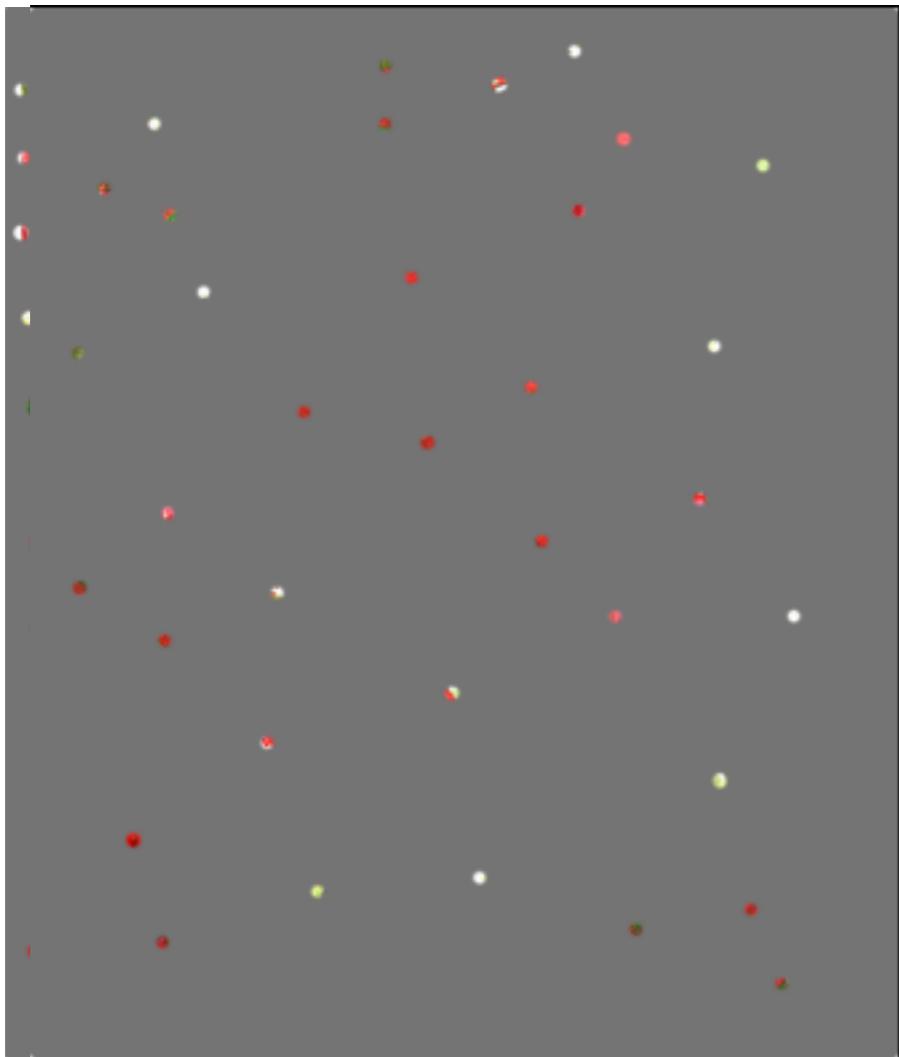


7. Úvod do matematické statistiky

Úloha statistické indukce

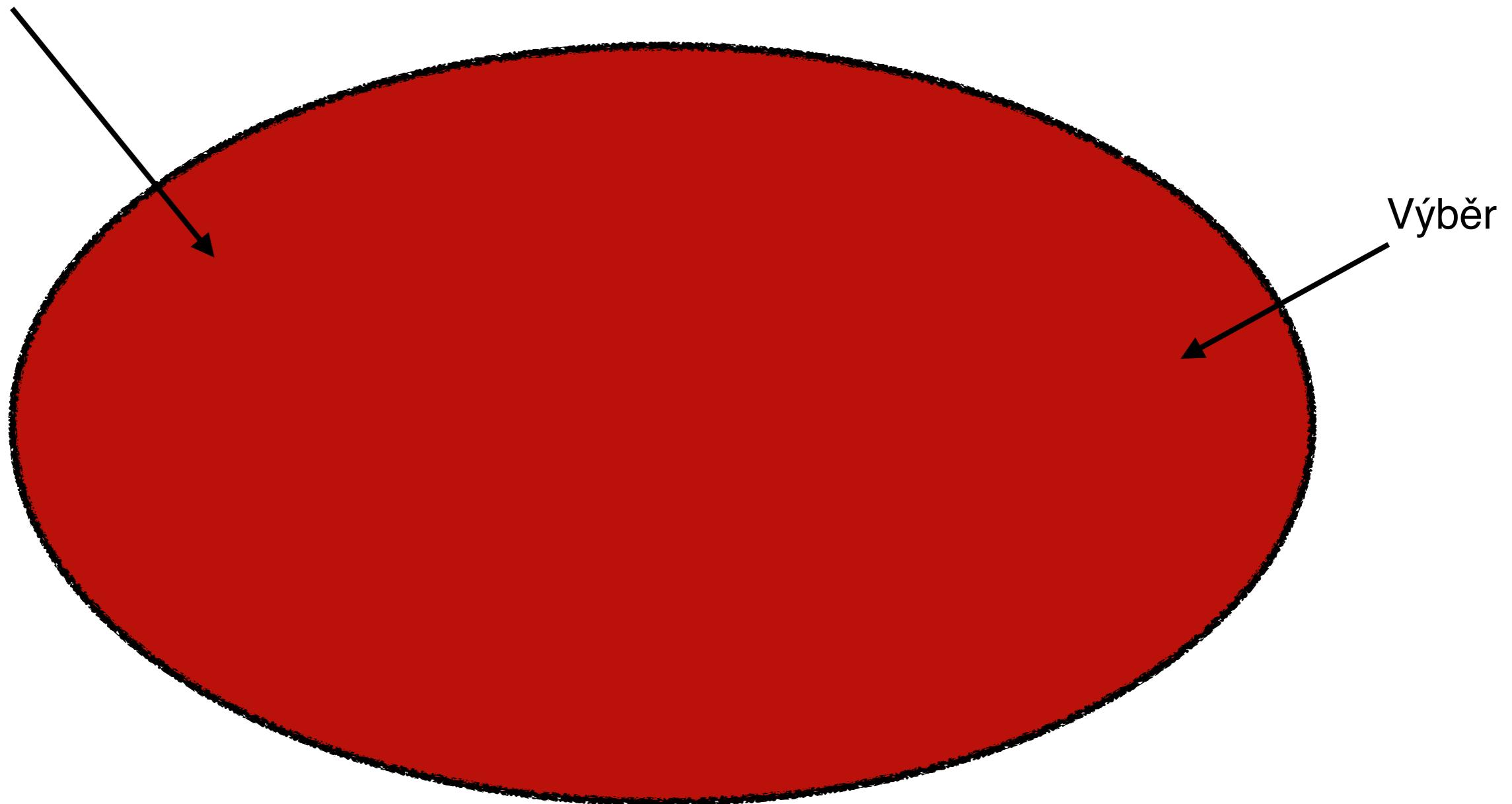


Úloha statistické indukce



Úloha statistické indukce

Základní soubor - nositel sledovaného znaku (veličiny)



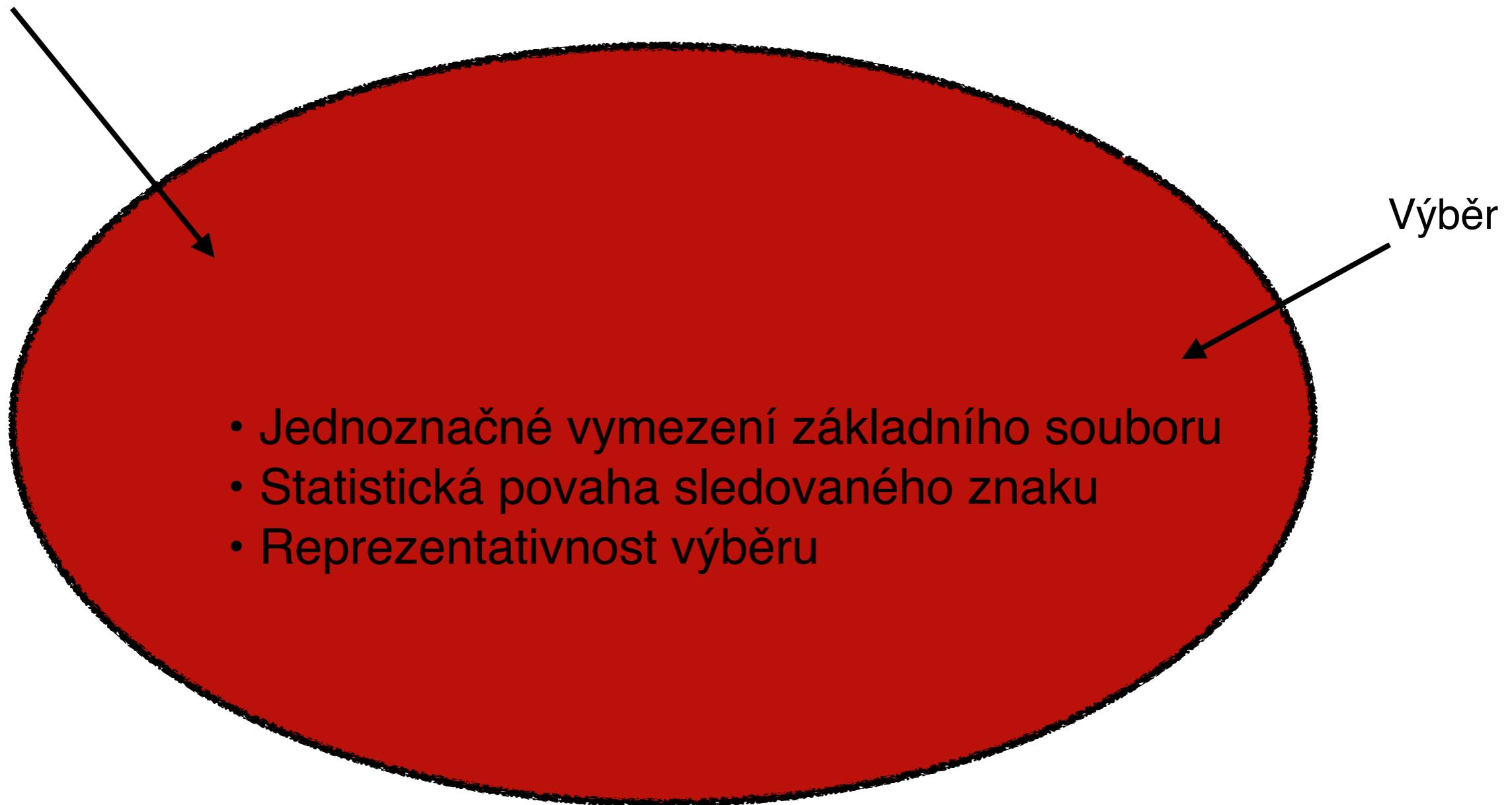
Pozorování výběru (měření sledovaného znaku) => Zjištění vlastností výběru

=> zobecnění na celý základní soubor



Úloha statistické indukce

Základní soubor - nositel sledovaného znaku (veličiny)



Pozorování výběru (měření sledovaného znaku) => Zjištění vlastností výběru

=> zobecnění na celý základní soubor



Statistické charakteristiky

Statistické charakteristiky: jsou spočteny na základě pozorování x_1, x_2, \dots, x_n výběru X_1, X_2, \dots, X_n .

Pravděpodobnostní charakteristiky	Výběrové charakteristiky
Střední hodnota $E(X) = \int_{-\infty}^{\infty} xf(x)dx$	Výběrový průměr $\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$
Momenty $\mu_k(X) = E(X - E(X))^k$	Výběrové momenty $m_k(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^k$
Rozptyl $var(X) = E(X - E(X))^2$	Výběrový rozptyl $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2$
Kvantity $\tilde{x}_{100\alpha}$	Výběrové kvantity $X_{([np]+1)}$



Statistické charakteristiky

Statistické charakteristiky: jsou spočteny na základě pozorování x_1, x_2, \dots, x_n výběru X_1, X_2, \dots, X_n .

Za předpokladu, že náhodný výběr je nezávislý a je z normálního rozdělení, t.j. X_1, X_2, \dots, X_n jsou i.i.d. a $X_k \sim N(\mu, \sigma^2)$, $k = 1, 2, \dots, n$, lze určit rozdělení pravděpodobnosti některých charakteristik:

- Pokud je μ a σ^2 známé, má výběrový průměr \bar{X}_n rozdělení $N(\mu, \sigma^2/n)$
- Pokud μ a σ^2 neznáme, má veličina $T = (X - \bar{X})/s$ tzv. Studentovo neboli t -rozdělení $t(n-1)$
- Veličina $S^2 = (n-1).s^2/\sigma^2$ má $\chi^2(n-1)$ rozdělení (o $n-1$ stupních volnosti)



Statistické charakteristiky

Další důležité výběrové charakteristiky:

- Výběrová šikmost (skewness): $Skew(X) = \frac{m_3(X)}{m_2^{3/2}(X)}$

pro $X \sim N(\mu, \sigma^2)$ je

$$E(Skew(X)) = 0$$

$$var(Skew(X)) = \frac{6(n-2)}{(n+1)(n+3)}$$

- Výběrová špičatost (kurtosis): $Kurt(X) = \frac{m_4(X)}{m_2^2(X)} - 3$

pro $X \sim N(\mu, \sigma^2)$ je

$$E(Kurt(X)) = -\frac{6}{n+1}$$

$$var(Kurt(X)) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

Máme-li dostatečný počet pozorování (řádově stovky), mají statistiky

$$T_3 = \frac{S_{kew}^{norm}}{\sqrt{Var(S_{kew}^{norm})}}$$

$$T_4 = \frac{K_{urt}^{norm} - E(K_{urt}^{norm})}{\sqrt{Var(K_{urt}^{norm})}}$$

přibližně standardní normální rozdělení pravděpodobnosti.



Statistické charakteristiky

Statistické charakteristiky: jsou spočteny na základě pozorování x_1, x_2, \dots, x_n výběru X_1, X_2, \dots, X_n .

- *Uspořádaný výběr:* $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ vznikne z původního výběru X_1, X_2, \dots, X_n uspořádáním podle velikosti pozorovaných hodnot x_1, x_2, \dots, x_n .
- *Pořadová statistika:* $X_{(k)}$ je náhodná veličina X_m , která je k -tá v pořadí podle velikosti pozorovaných hodnot x_1, x_2, \dots, x_n . Index k nazýváme *pořadím veličiny* X_m a zapisujeme to $R_m = k$.
- Statistika $X_{(1)}$ se nazývá minimum, $X_{(n)}$ je maximum
- medián \tilde{x}_{50} : je-li n liché, je roven $X_{([n/2]+1)}$
pro n sudé je roven $(X_{(n/2)} + X_{(n/2+1)})/2$
- dolní kvartil \tilde{x}_{25} : $X_{([n/4]+1)}$ resp. $(X_{(n/4)} + X_{(n/4+1)})/2$
- horní kvartil \tilde{x}_{75} : $X_{([3n/4]+1)}$ resp. $(X_{(3n/4-1)} + X_{(3n/4)})/2$



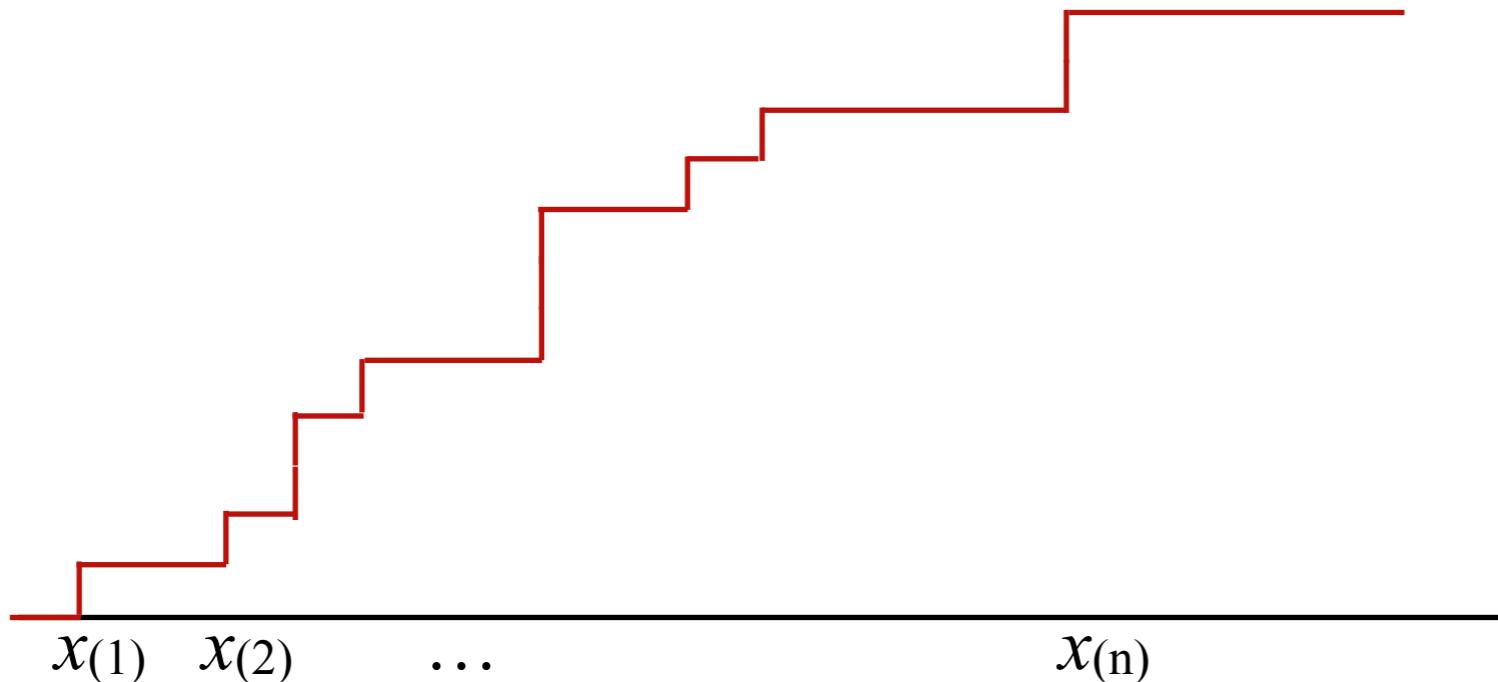
Grafická analýza

Statistické charakteristiky: jsou spočteny na základě pozorování x_1, x_2, \dots, x_n výběru X_1, X_2, \dots, X_n .

Empirická distribuční funkce:

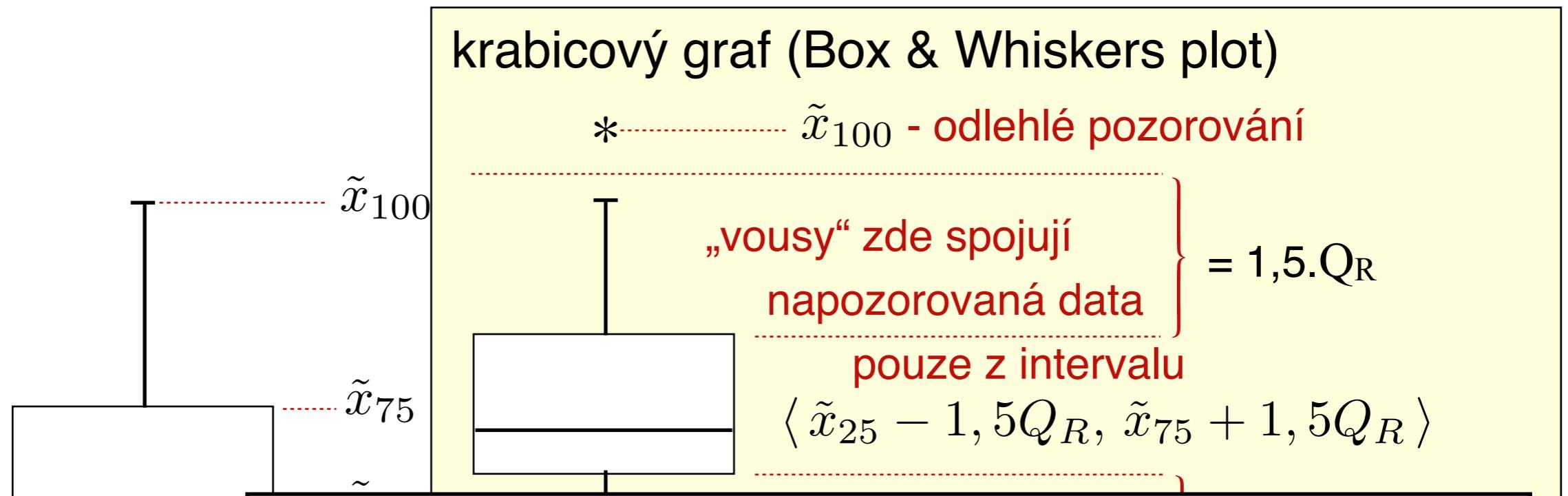
vycházíme z uspořádaného výběru: $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. Potom $F_n(x_{(i)}) = \frac{i}{n}$

a tedy $F_n(x) = \frac{\max\{k : X_{(k)} \leq x\}}{n}, \quad x \in \mathbf{R}$

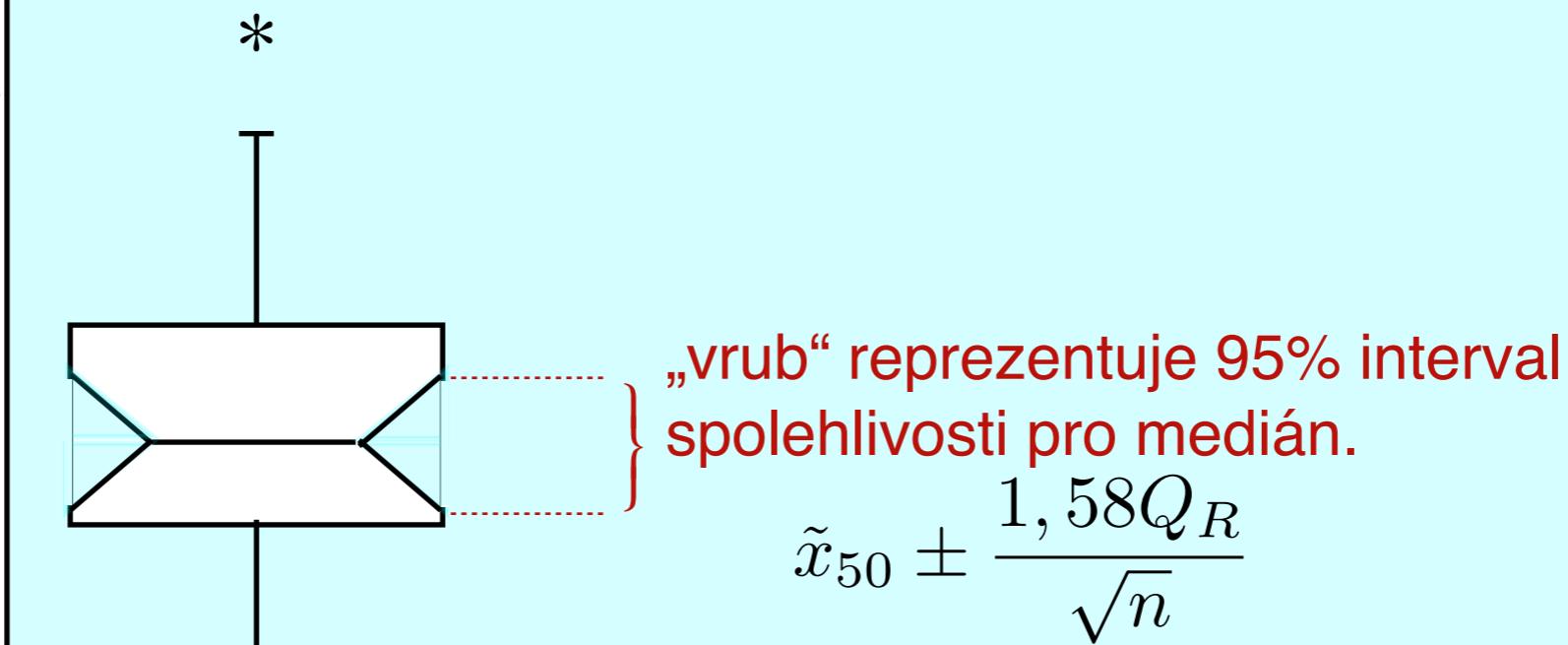


Grafická analýza

krabicový graf (Box & Whiskers plot)

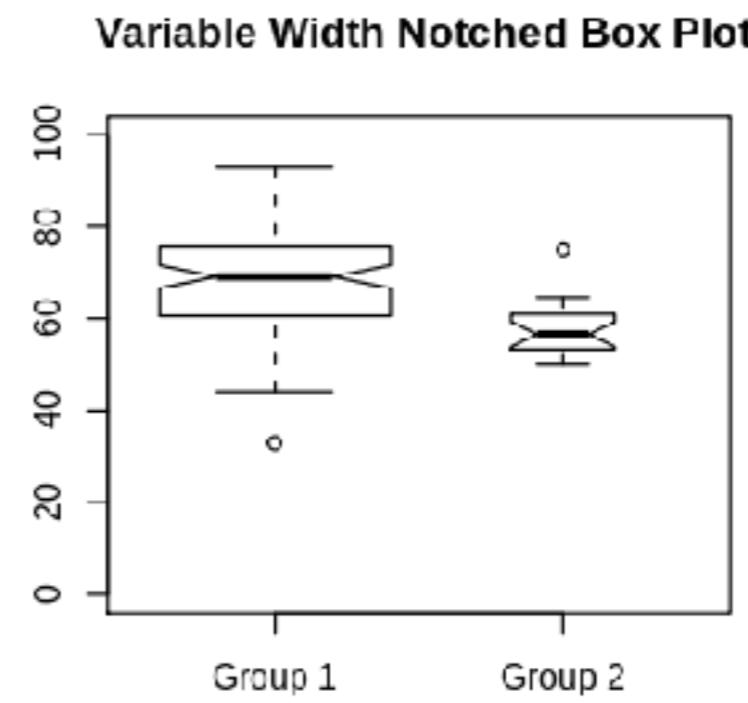
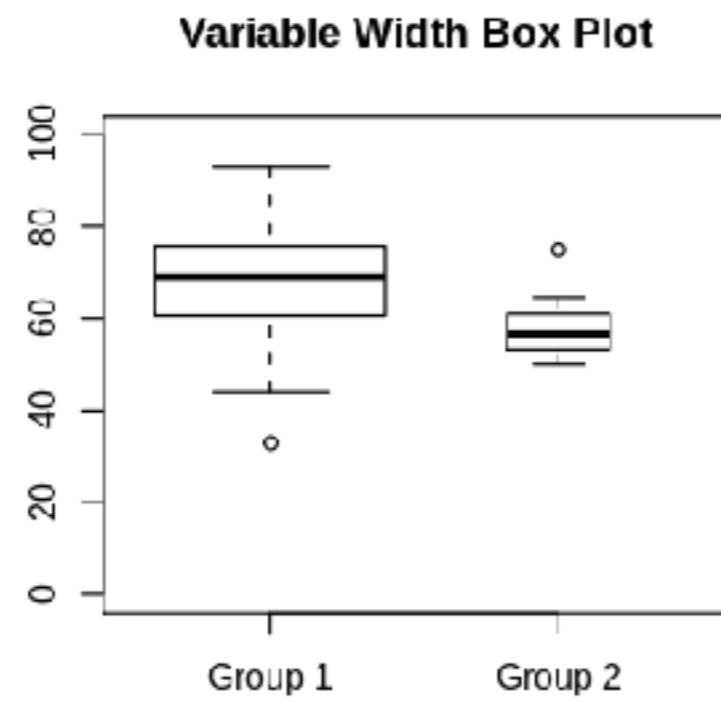
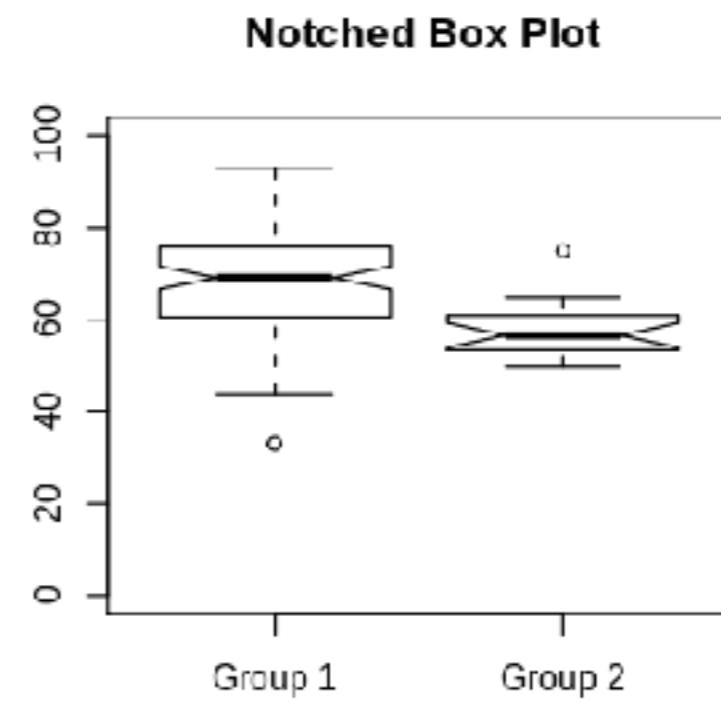
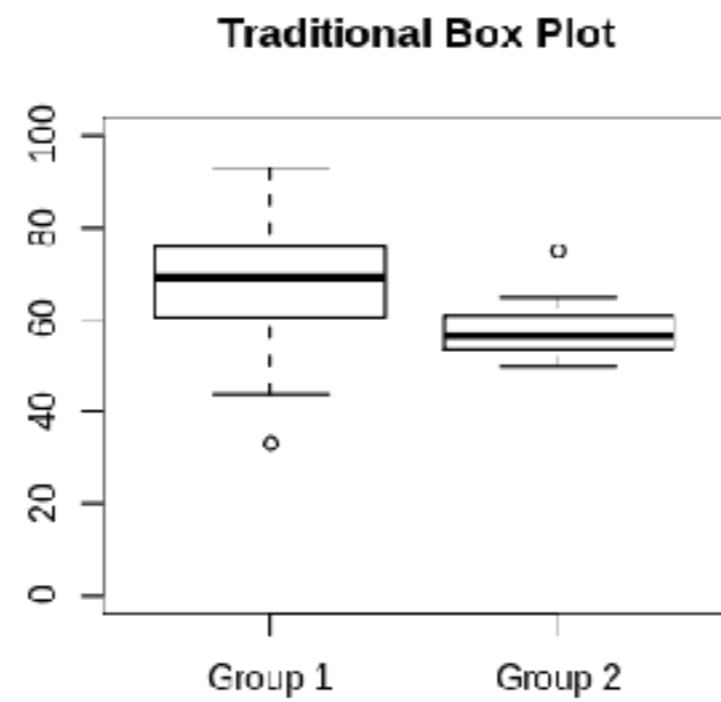


vrubový krabicový graf (notched Box & Whiskers plot)



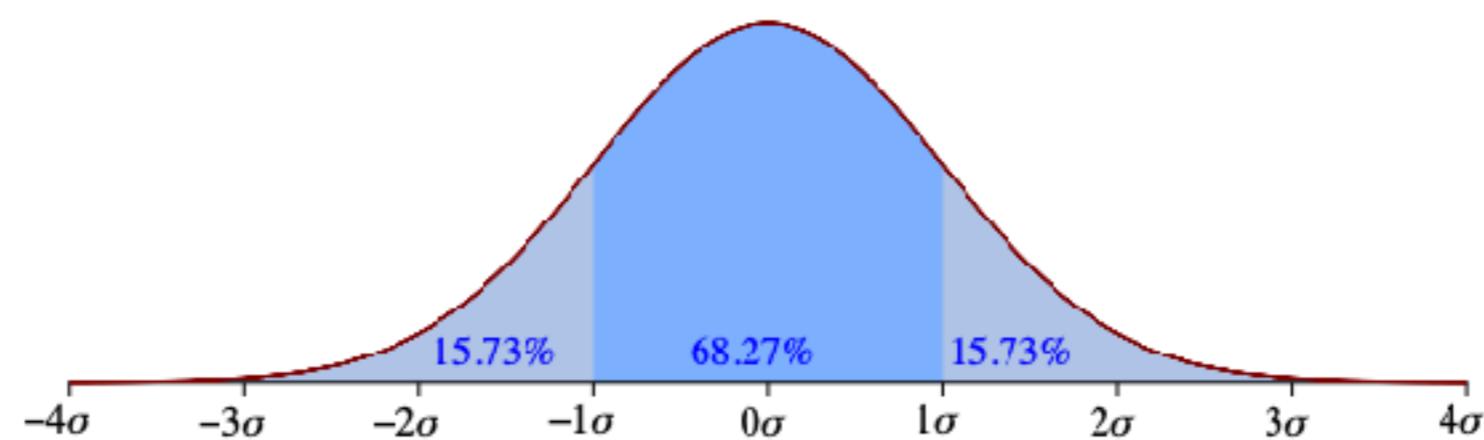
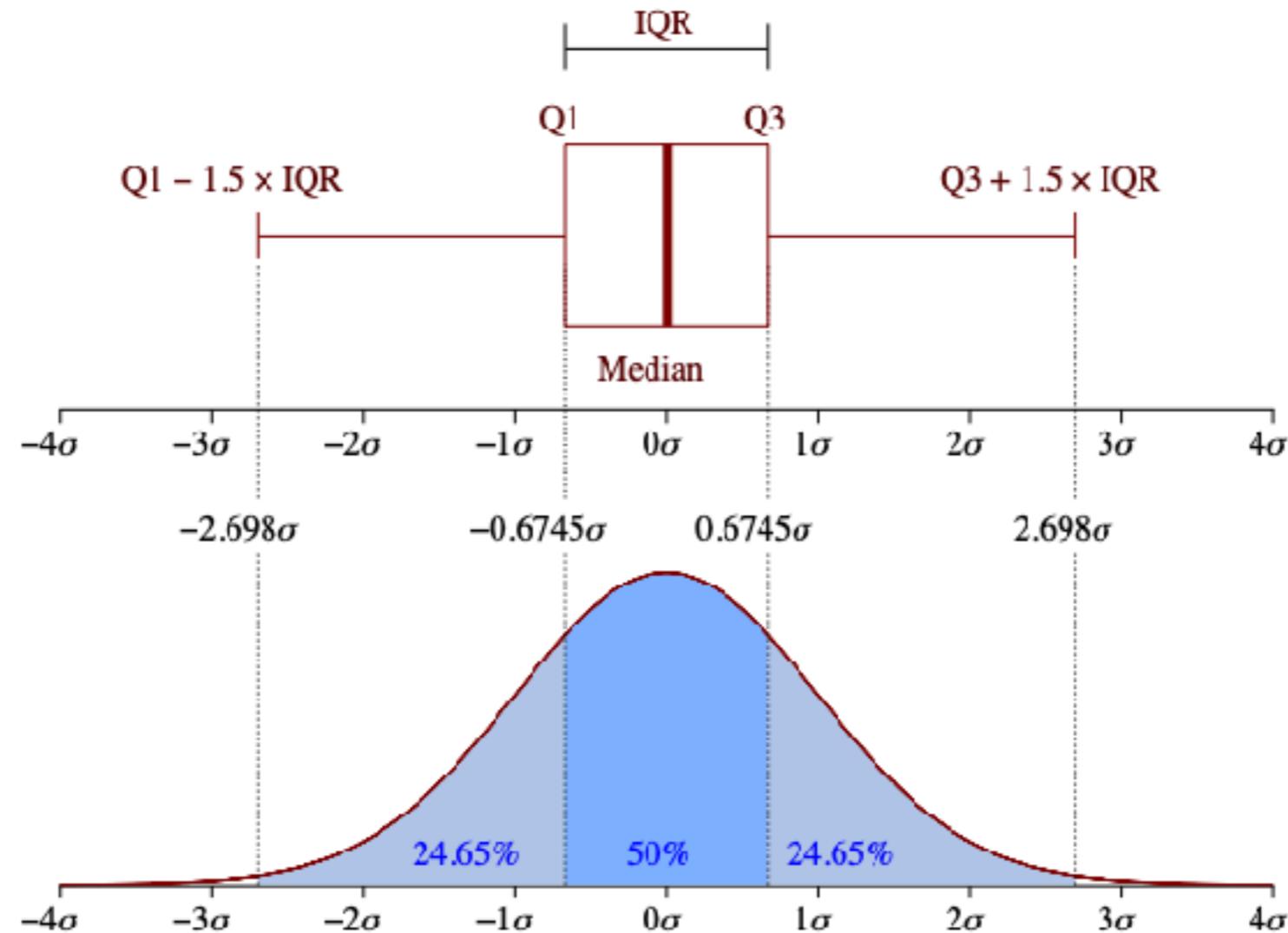
Grafická analýza

krabicový graf (Box & Whiskers plot)



Grafická analýza

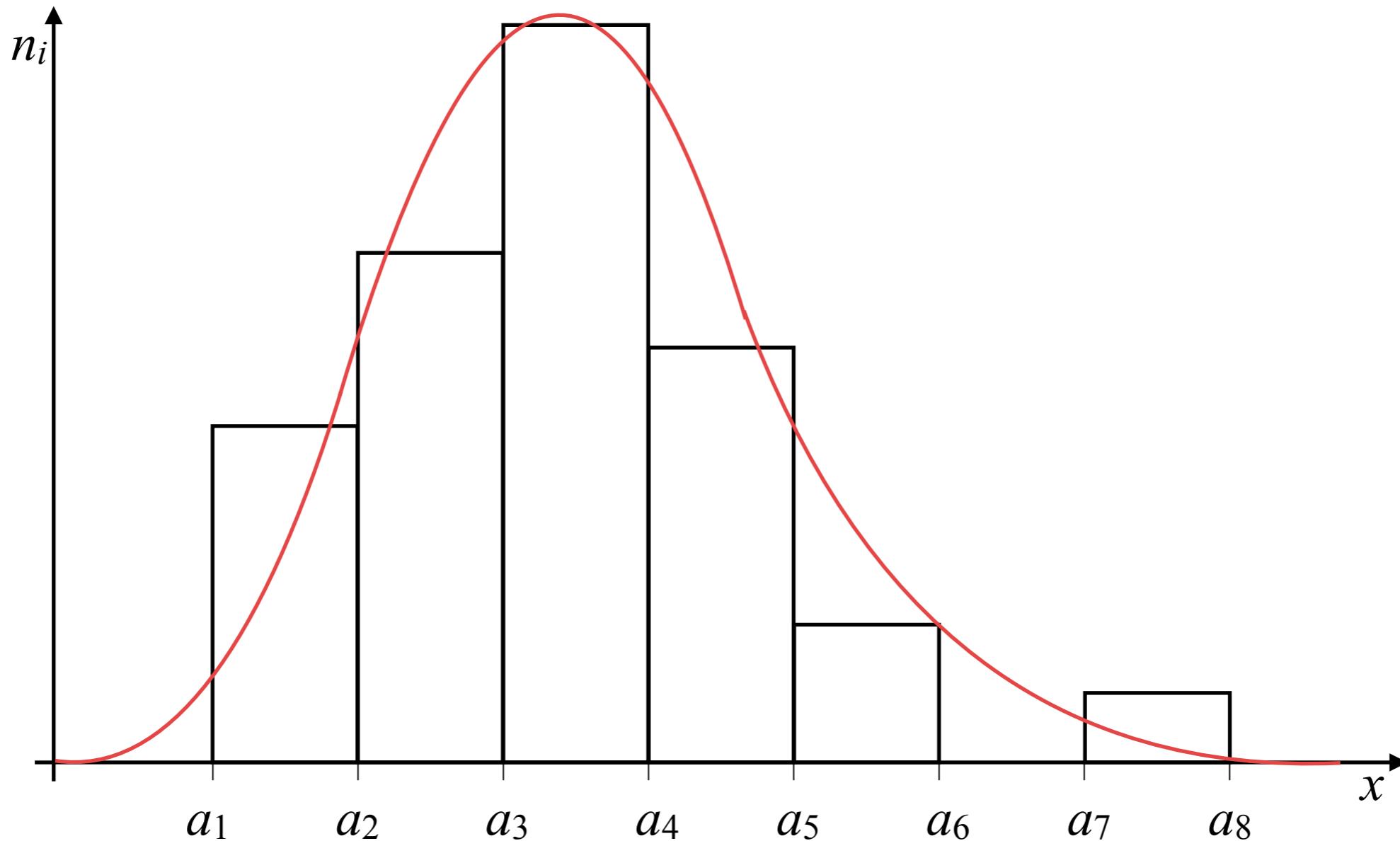
krabicový graf (Box & Whiskers plot)



Frekvenční analýza

Histogram

Máme pozorování x_1, x_2, \dots, x_n náhodného výběru X_1, X_2, \dots, X_n .



Frekvenční analýza

Histogram

Máme pozorování x_1, x_2, \dots, x_n náhodného ýběru X_1, X_2, \dots, X_n .

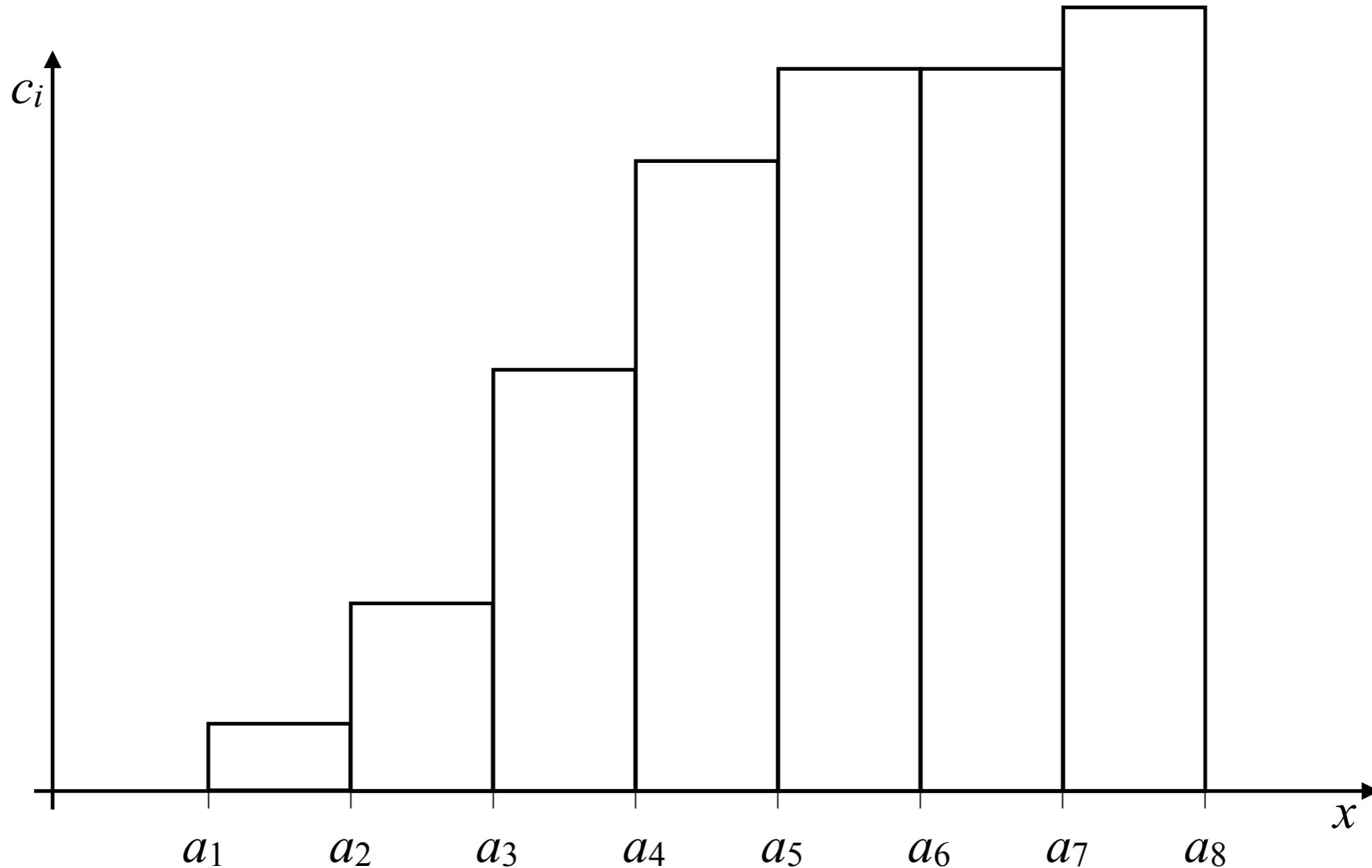
pořadí třídy	třídní intervaly	(prosté) absolutní četnosti	(prosté) relativní četnosti	kumulativní četnosti	kumulativní relativní četnosti
1	$a_2 - a_1$	n_1	$f_1 = n_1/n$	$c_1 = n_1$	$d_1 = c_1/n$
2	$a_3 - a_2$	n_2	$f_2 = n_2/n$	$c_2 = n_1 + n_2$	$d_2 = c_2/n$
:	:	:	:	$c_j = \sum_{i=1}^j n_i$:
k	$a_k - a_{k-1}$	n_k	$f_k = n_k/n$	$c_k = n$	$d_k = c_k/n = 1$



Frekvenční analýza

Histogram

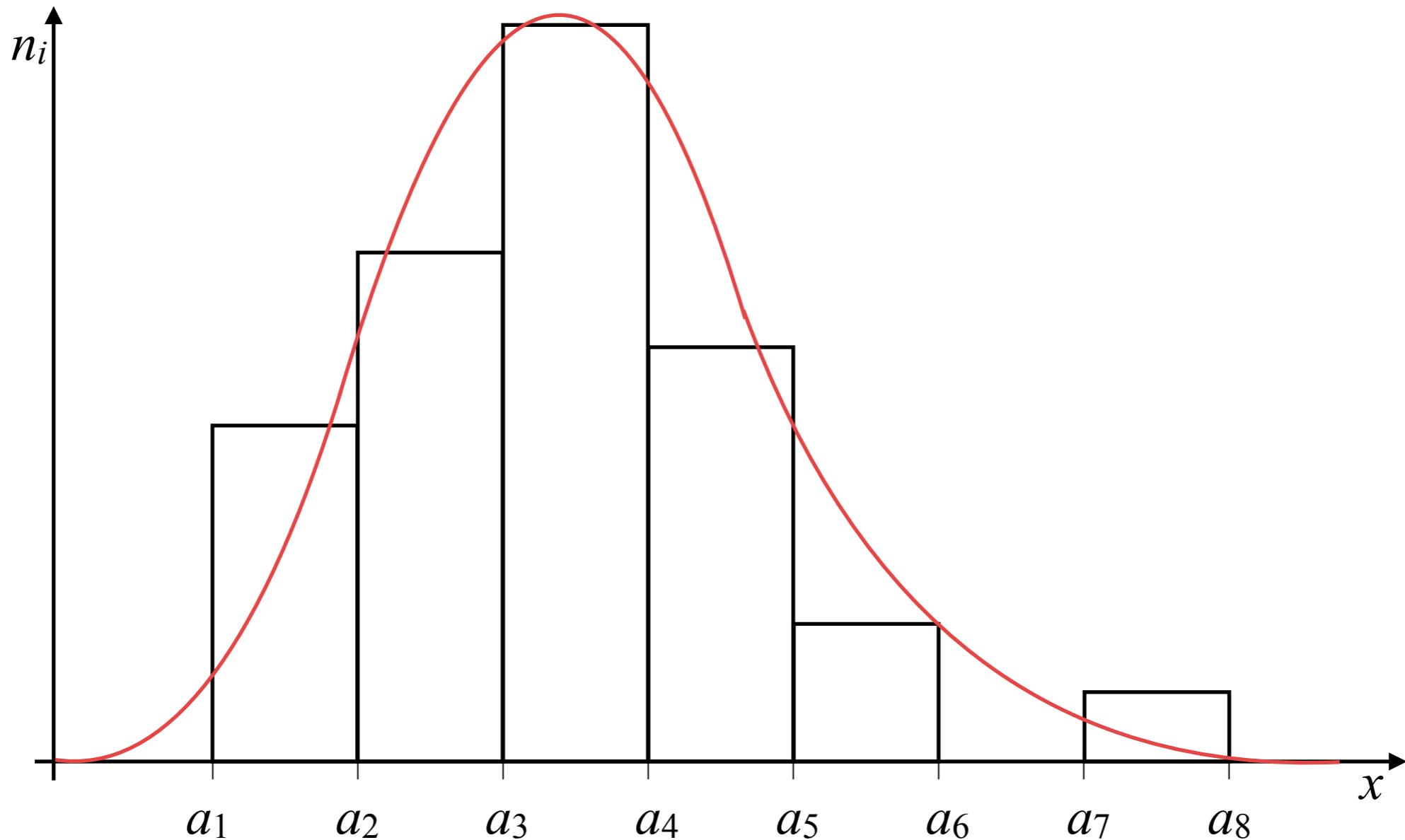
Máme pozorování x_1, x_2, \dots, x_n náhodného výběru X_1, X_2, \dots, X_n .



Frekvenční analýza

Histogram

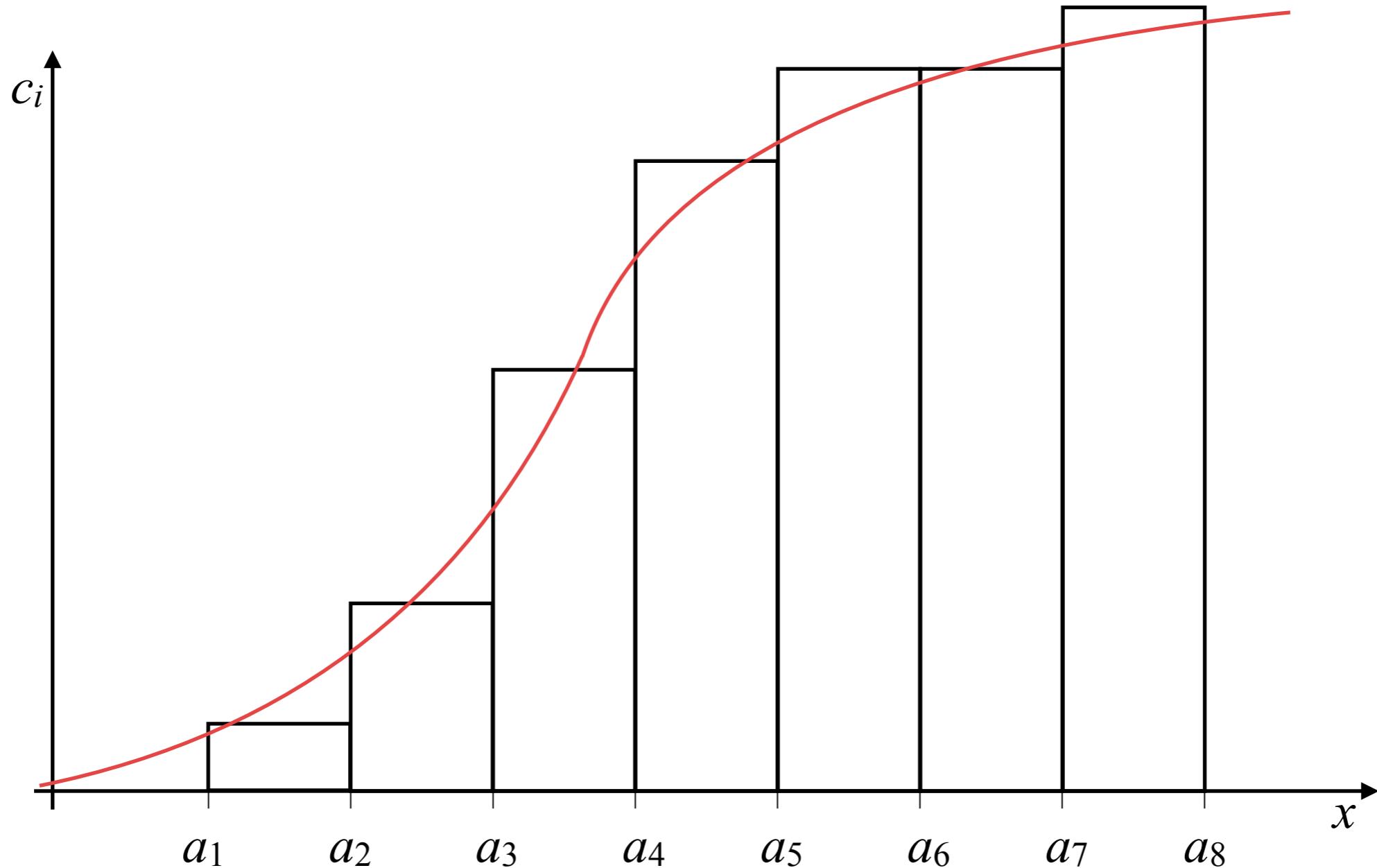
Máme pozorování x_1, x_2, \dots, x_n náhodného výběru X_1, X_2, \dots, X_n .



Frekvenční analýza

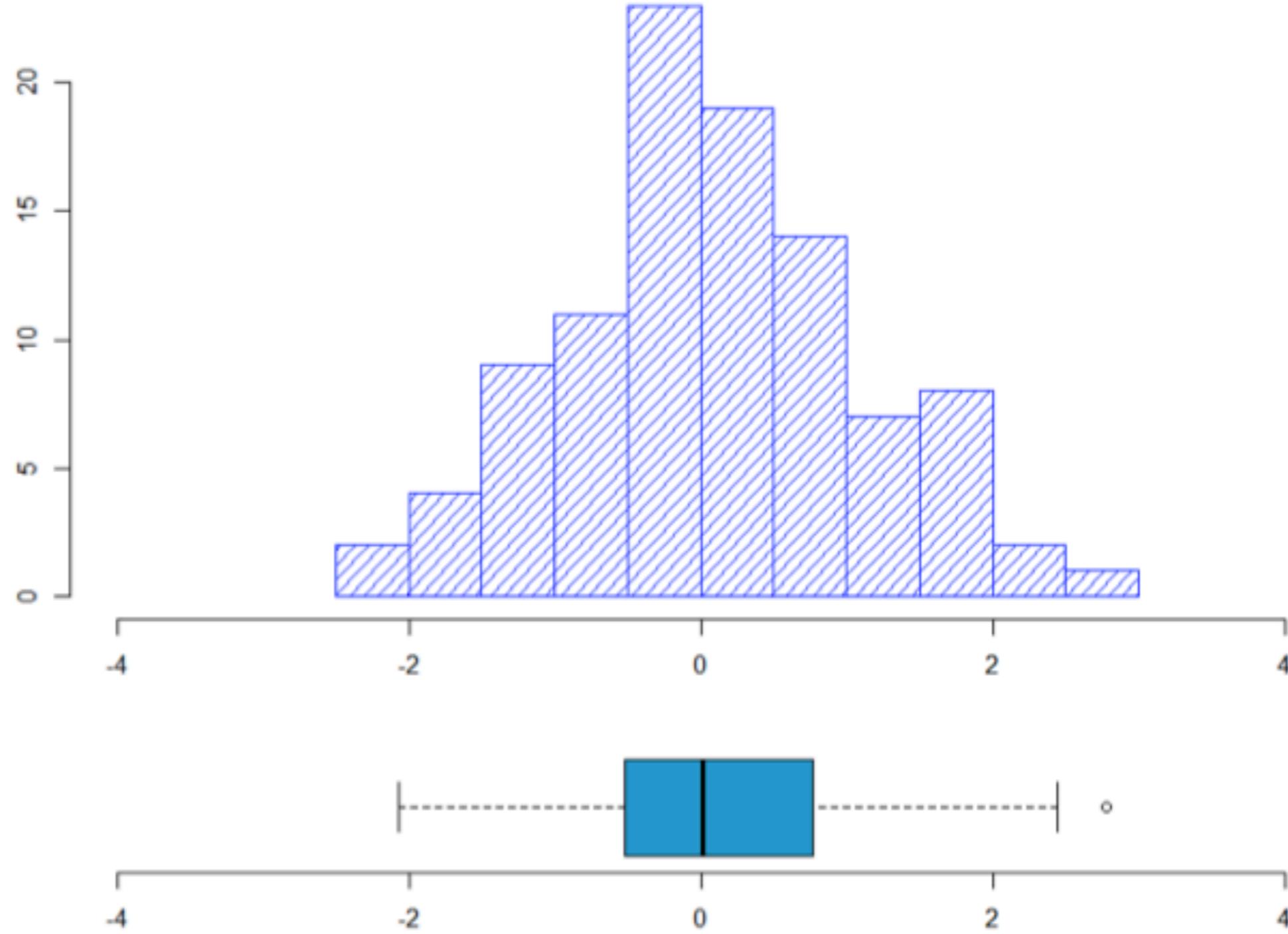
Histogram

Máme pozorování x_1, x_2, \dots, x_n náhodného výběru X_1, X_2, \dots, X_n .

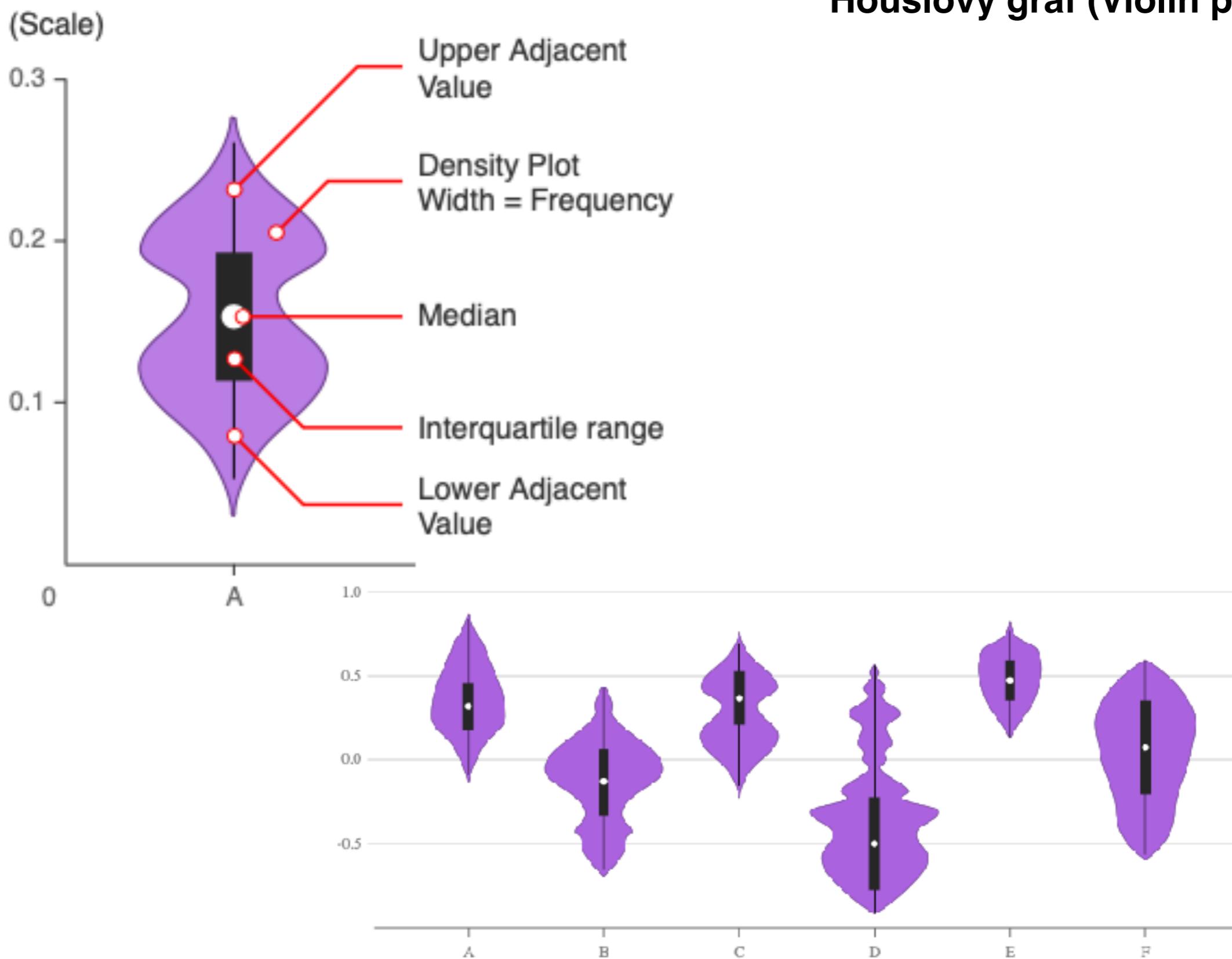


Frekvenční analýza

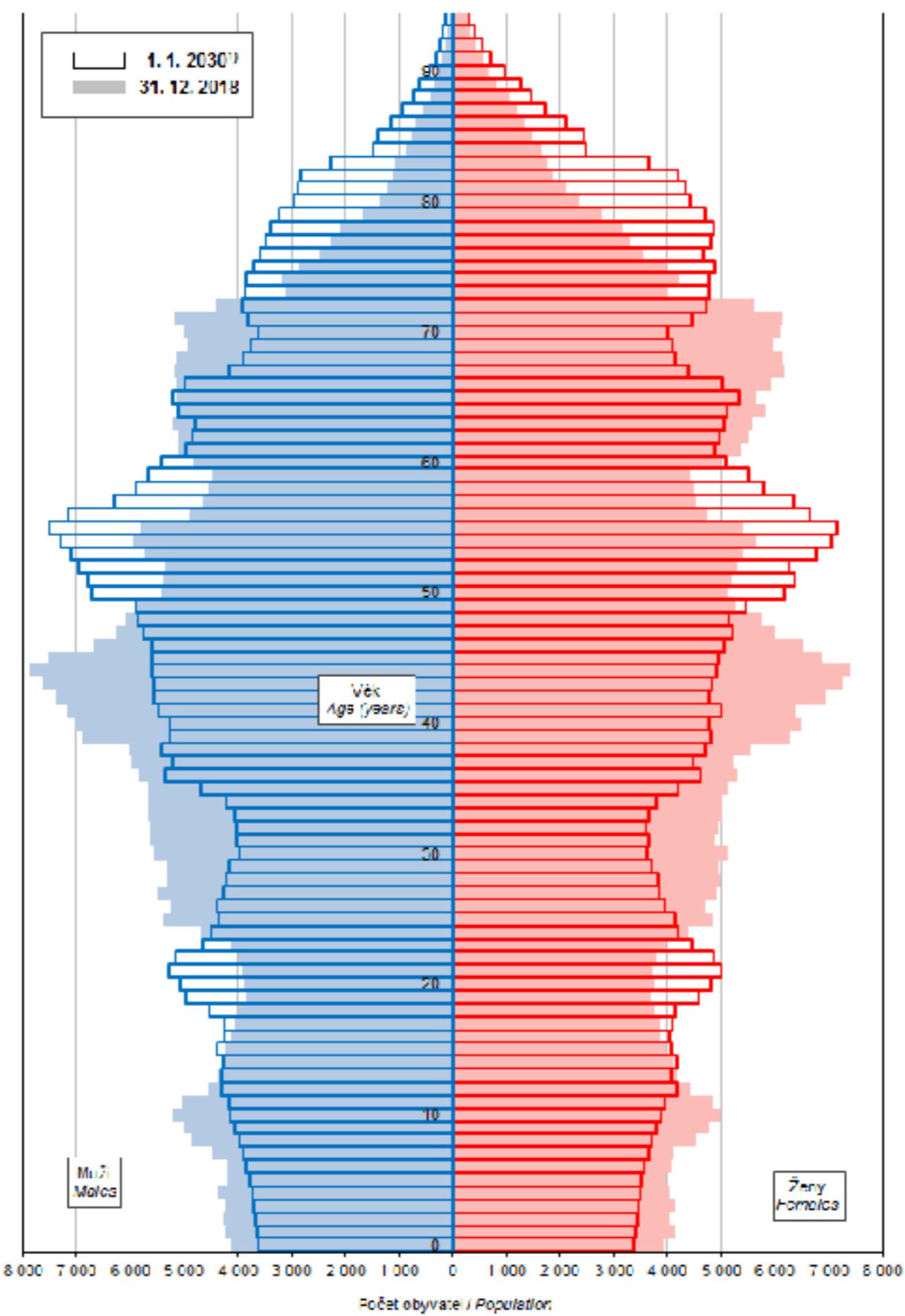
Histogram x Box plot



Frekvenční analýza



Věkové složení obyvatelstva Ústeckého kraje k 31. 12. 2018 a k 1. 1. 2030
Age distribution of the population in the Ústecký Region as at 31 December 2018 and 1 January 2030



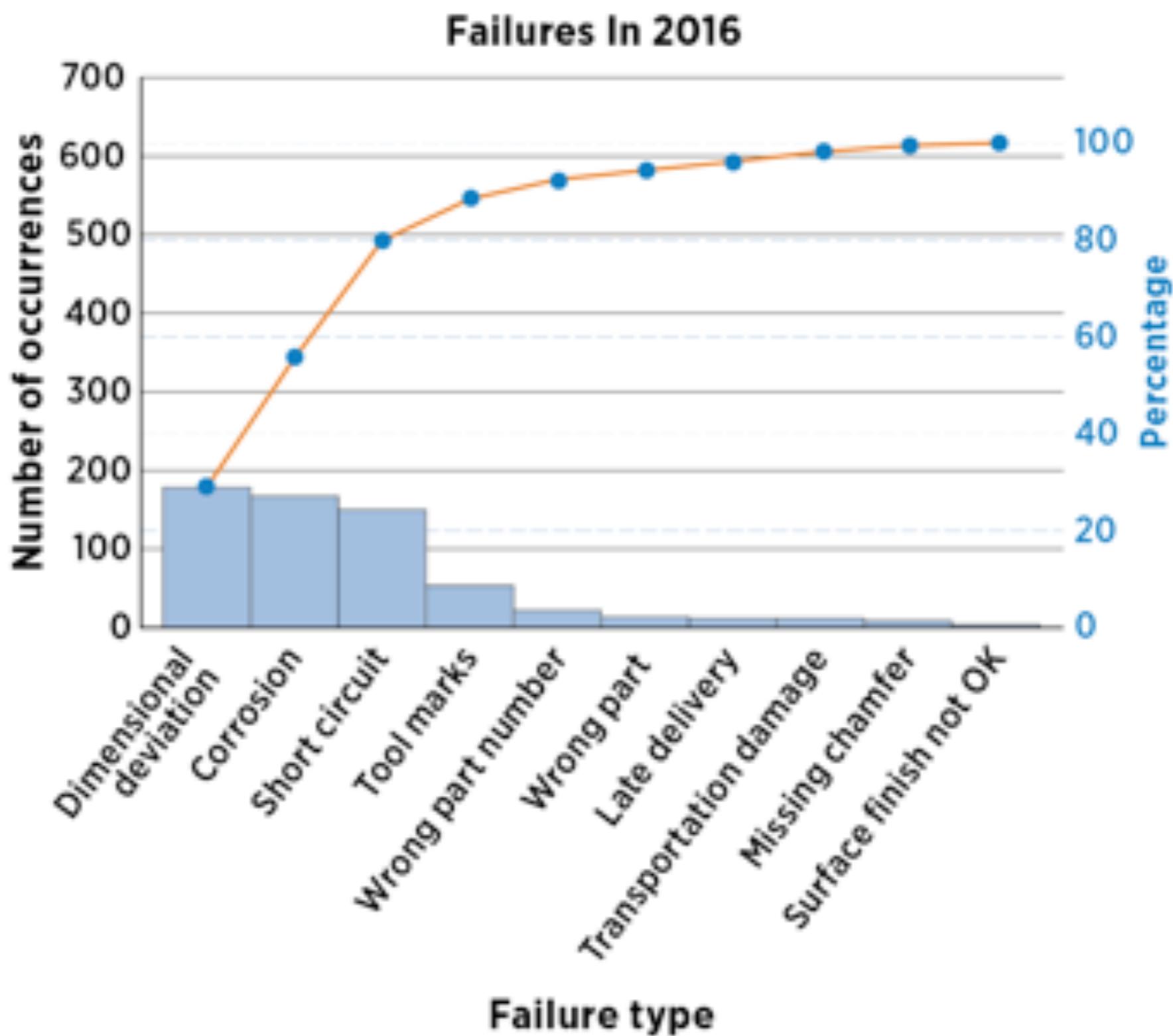
^aZdroj: Projekce obyvatelstva v krajích ČR do roku 2070

^bSource: UZSÚ evaluation "Projekce obyvatelstva v krajích ČR do roku 2070" (Uzsch only)



Frekvenční analýza

Paretův graf



Základní analýza v systému R



R je programovací jazyk, zaměřený na statistické výpočty

<https://www.r-project.org/>

https://wikisofia.cz/wiki/R_-_programovací_jazyk

<https://www.karlin.mff.cuni.cz/~maciak/NMSA230/Rmanual2.pdf>

R - Studio je grafické prostředí pro výpočty v jazyce R



<https://posit.co/download/rstudio-desktop/>

https://www.math.muni.cz/~kolacek/vyuka/vypsyst/navod_R.pdf



Základní analýza v systému R

```
VI  
[1,] 24.52586  
[2,] 24.17119  
[3,] 24.54486  
[4,] 24.44240  
[5,] 23.93455  
[6,] 24.20389  
[7,] 24.19974  
[8,] 24.34851  
[9,] 23.94024  
[10,] 24.21022  
[11,] 24.87474  
[12,] 25.06155  
[13,] 25.48924  
[14,] 25.32572  
[15,] 23.71721  
[16,] 24.61622  
[17,] 25.06676  
[18,] 24.90055  
[19,] 24.36213  
[20,] 24.98580  
[21,] 24.80591  
[22,] 24.20853  
[23,] 24.72623  
[24,] 24.64437  
[25,] 24.70405  
[26,] 23.97645  
[27,] 25.29837  
[28,] 24.46910  
[29,] 24.99453  
[30,] 25.42994  
[31,] 24.66147  
[32,] 24.75773  
[33,] 25.03970  
[34,] 24.44901  
[35,] 25.13285  
[36,] 24.40205  
[37,] 24.78721  
[38,] 23.83656  
[39,] 24.17186  
[40,] 23.65390  
[41,] 24.48244  
[42,] 24.68550  
[43,] 24.22988  
[44,] 23.83956  
[45,] 24.09777  
[46,] 24.52098  
[47,] 24.89240  
[48,] 24.25332  
[49,] 24.14259  
[50,] 25.12906
```

```
# Volání potřebných knihoven  
> library(moments)  
> library(nortest)  
  
#  
# Načtení dat  
> x <- data.matrix(read.table("davkovac.txt"))  
> x  
> summary(x)  
  
VI  
Min. :23.65  
1st Qu.:24.21  
Median :24.52  
Mean :24.55  
3rd Qu.:24.89  
Max. :25.49  
  
> skewness(x)  
0.09303432  
> kurtosis(x)  
2.27457
```

MS Excel

data jsou uložena
v poli A1:A50

=MIN(A1:A50)
=QUARTIL(A1:A50;2)
=MEDIAN(A1:A50)
=PRŮMĚR(A1:A50)
=QUARTIL(A1:A50;3)
=MAX(A1:A50)

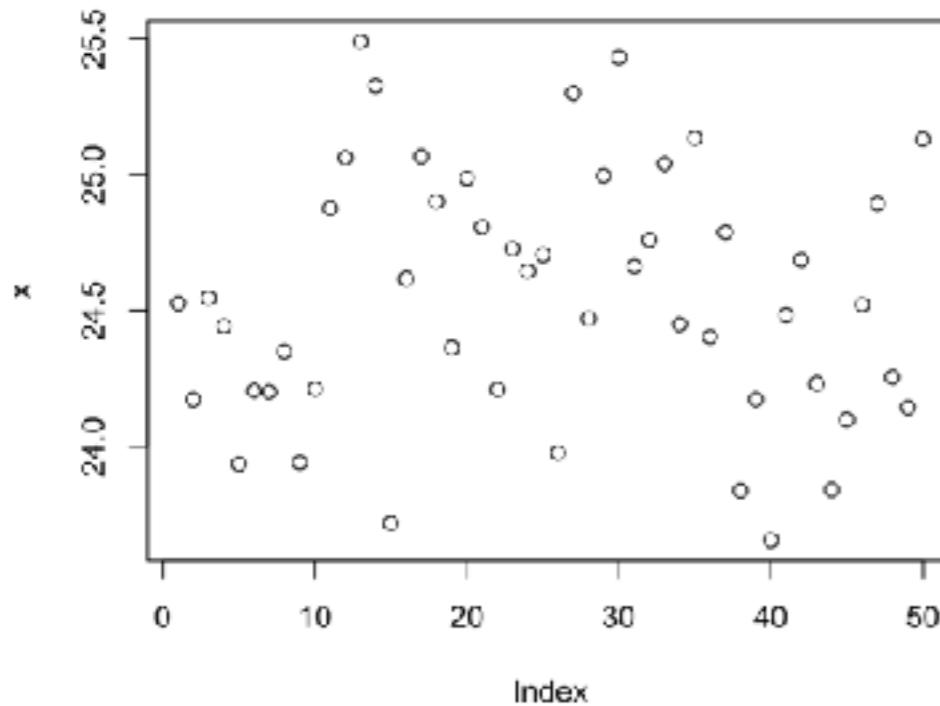
=SKEW(A1:A50)

=KURT(A1:A50) + 3



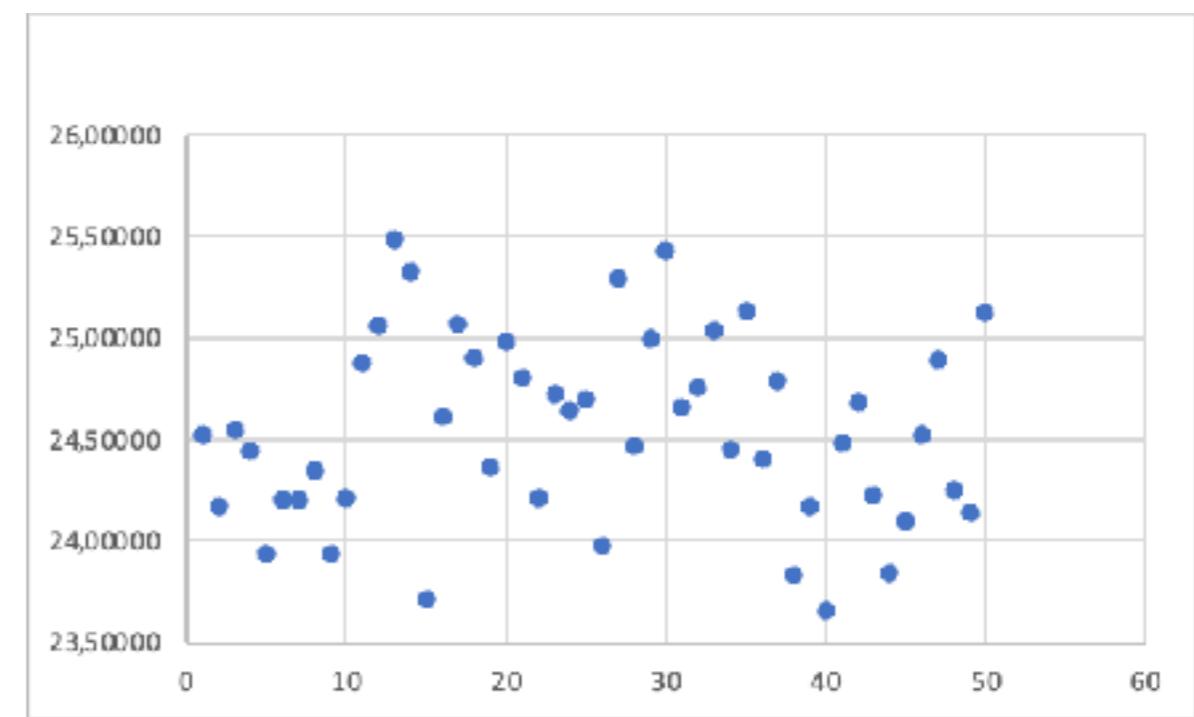
Základní analýza v systému R

> plot(x)



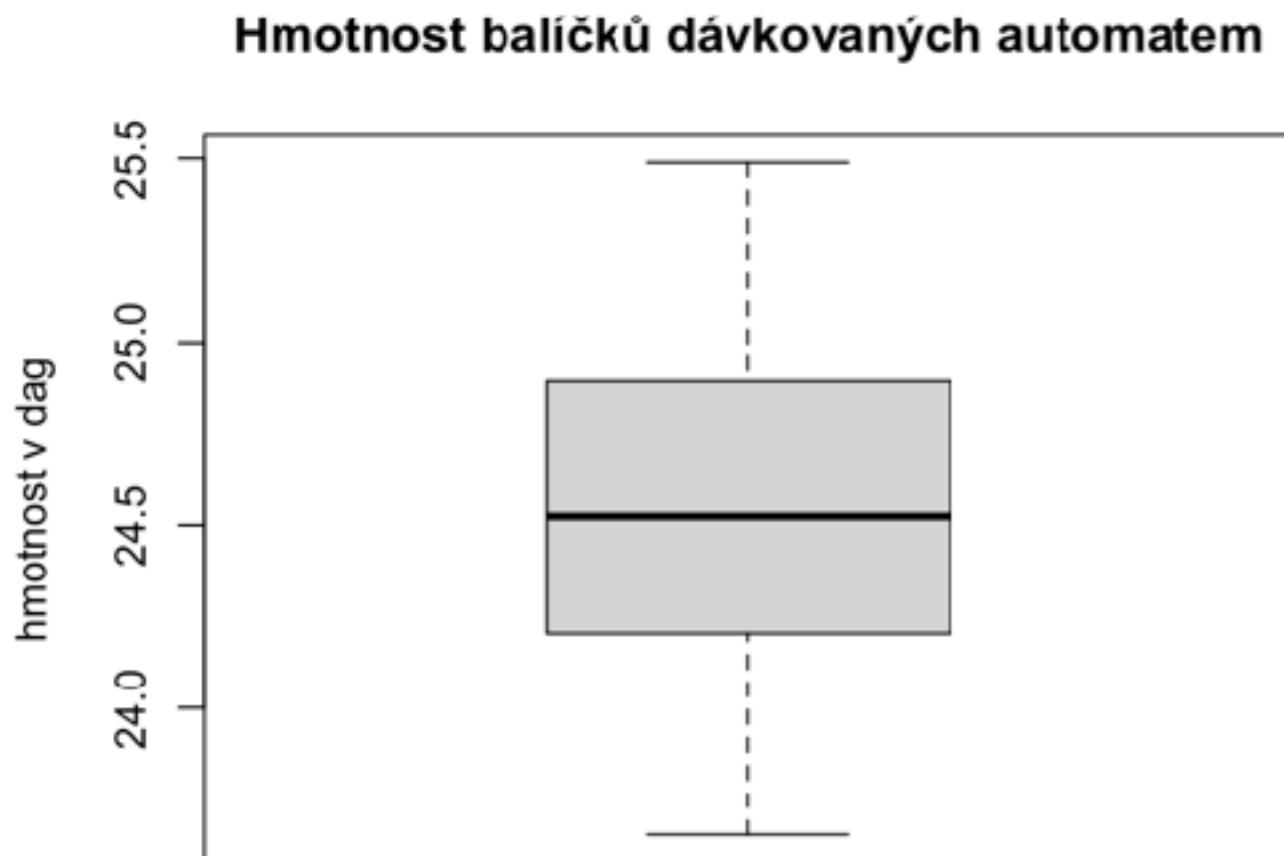
MS Excel

XY bodový graf oblasti A1:A50



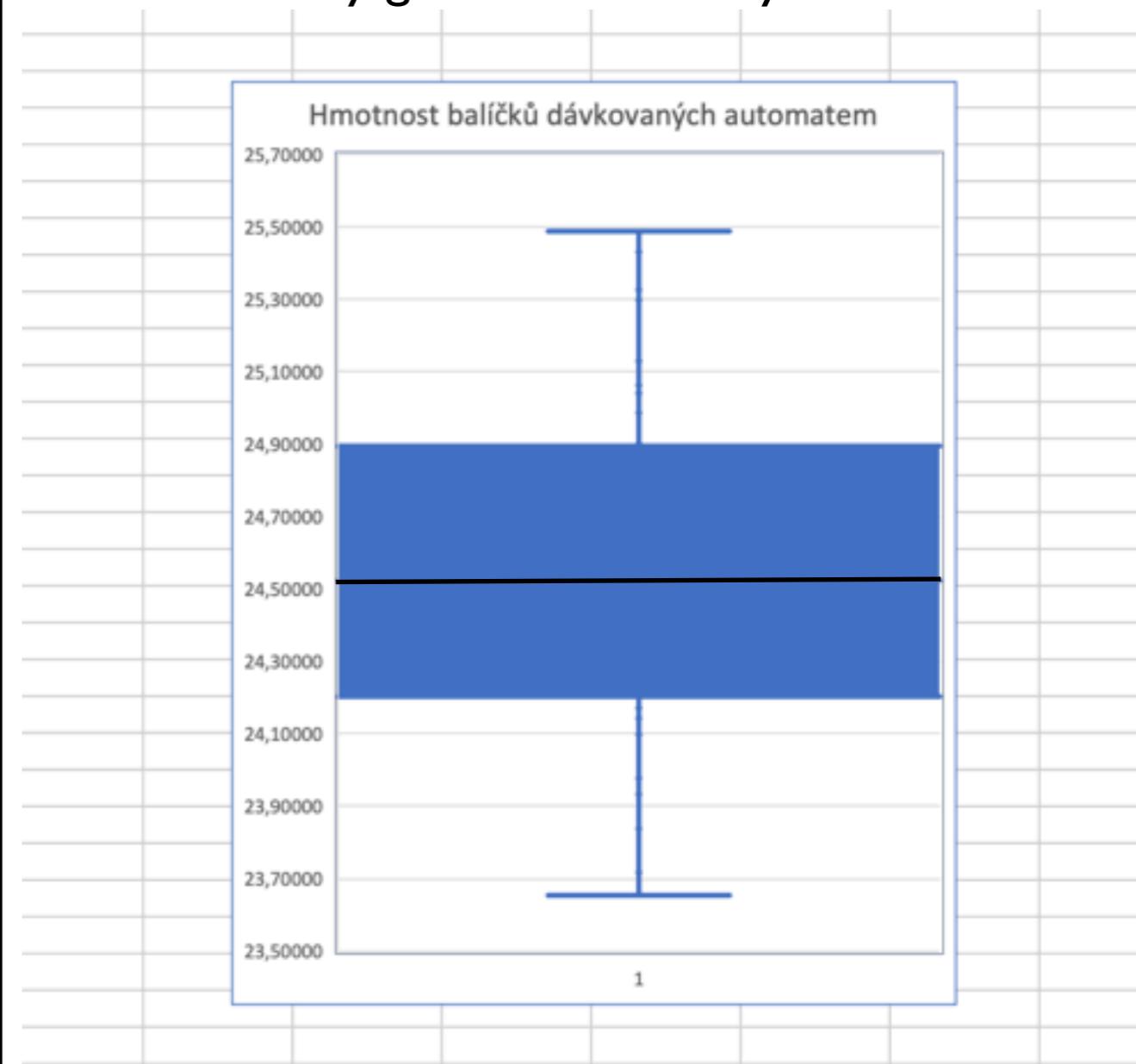
Základní analýza v systému R

> boxplot(x)



MS Excel

statistický graf → Krabicový oblasti A1:A50

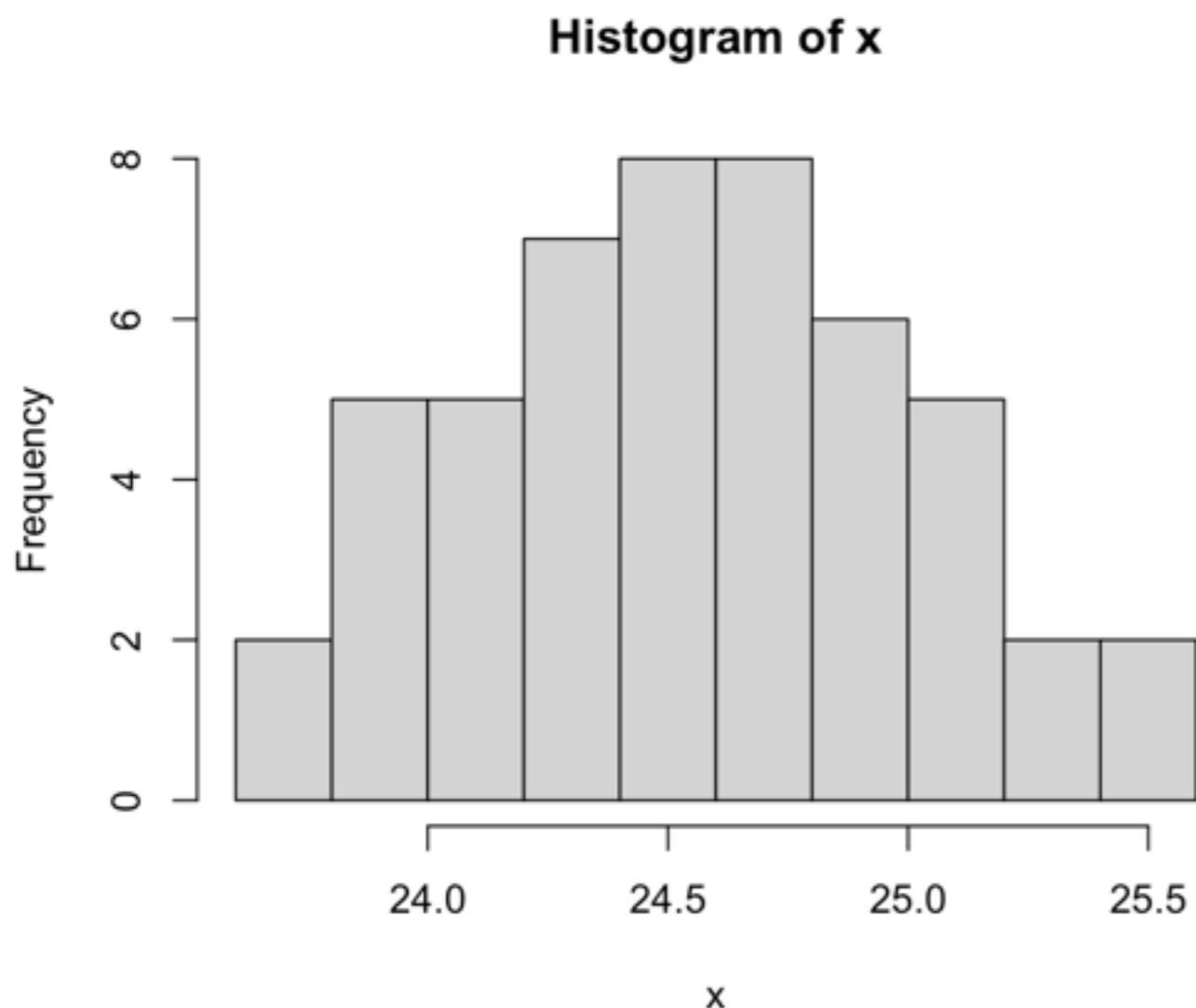


> boxplot(x, main="Hmotnost balíčků dávkovaných automatem", ylab="hmotnost v dag")



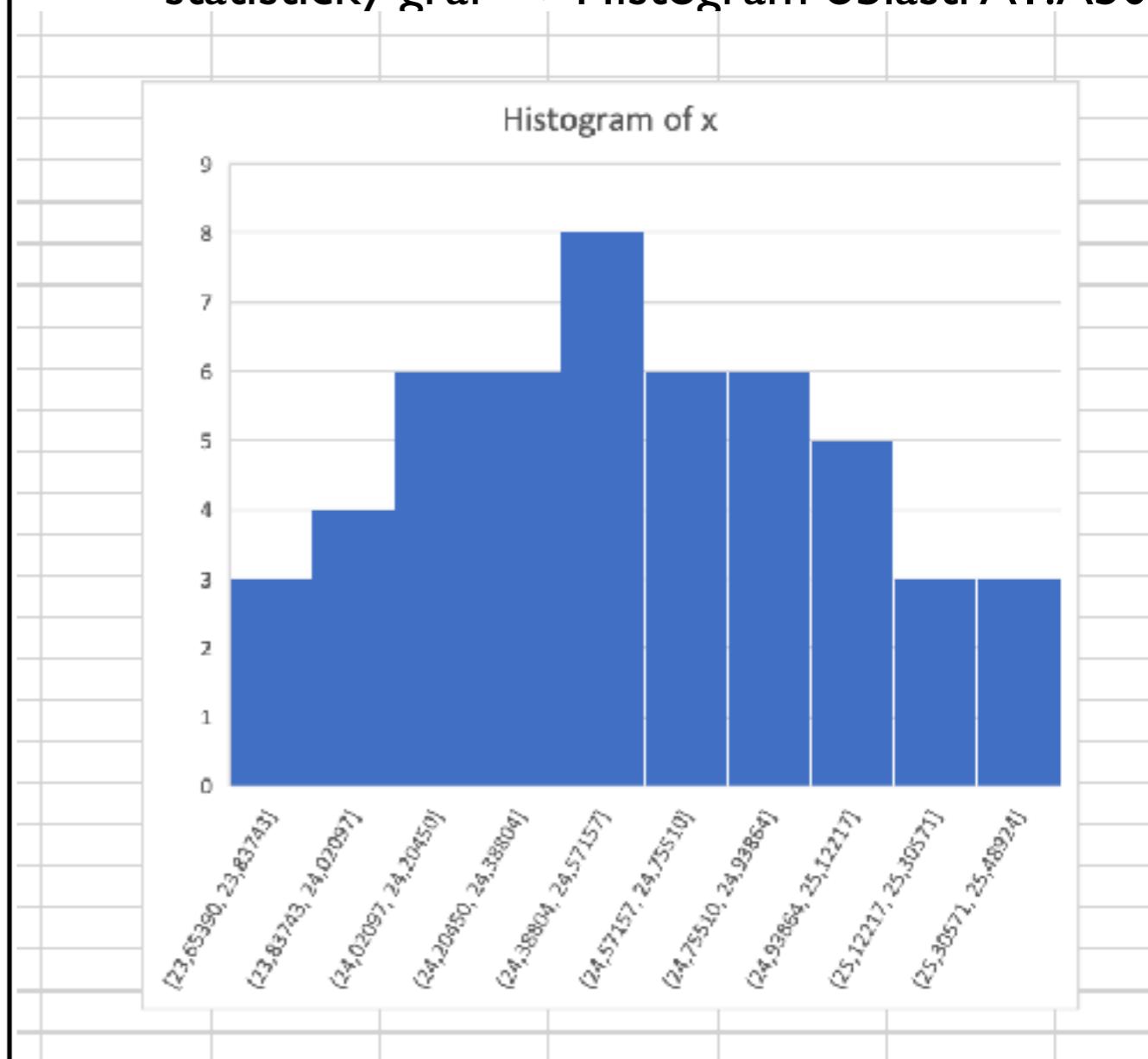
Základní analýza v systému R

> hist(x)



MS Excel

statistický graf → Histogram oblasti A1:A50



Základní analýza v systému R

Histogram

```
> hx <- hist(x)
> print(hx)
> h<-hist(x, main="Balící automat na kávu", > xlab="hmotnost balíčků v dekagramech",
  col="yellow", freq=FALSE, breaks = 8)
> yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
> lines(xfit, yfit, col="red", lwd=2)
> lines(density(x))
```

#Empirická distribuční funkce

```
> ecdf(x)
> plot(ecdf(x), verticals = FALSE, do.points="FALSE")
> xfit<-seq(min(x),max(x),length=40)
> zfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
> lines(xfit, zfit, col="red", lwd=2)
```

#Q-Q graf

```
> qqnorm(x, main='Normal')
> qqline(x)
```



Základní analýza v systému R

#Testy normality

```
> shapiro.test(x)
```

Shapiro-Wilk normality test

```
data: x, W = 0.98406, p-value = 0.7307
```

```
> ad.test(x)
```

Anderson-Darling normality test

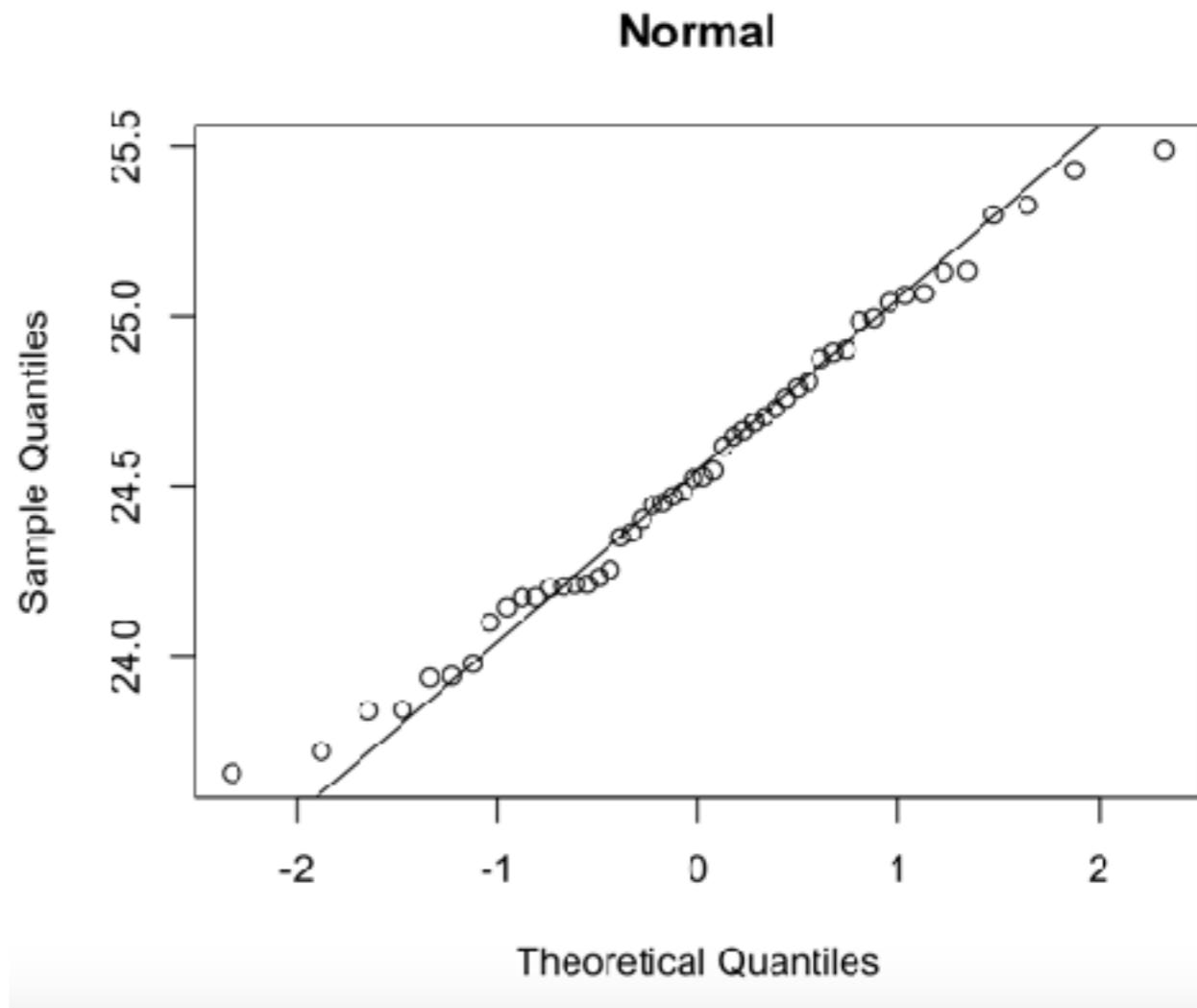
```
data: x, A = 0.19883, p-value = 0.8799
```

```
> jarque.test(x)
```

Jarque-Bera Normality Test

```
data: y, JB = 1.1685, p-value = 0.5575
```

```
alternative hypothesis: greater
```



Malý slovníček statistických pojmu, aneb

co je třeba znát a porozumět tomu:

- **Základní soubor (populace, universum)**: množina objektů, na nichž provádíme statistické zkoumání; musí být přesně specifikována
- **Výběr (ze základního souboru)**: n -tice náhodných veličin X_1, X_2, \dots, X_n odpovídající nezávislým pozorováním vybraných objektů základního souboru na nichž pozorujeme nějakou veličinu X reprezentující určitou měřitelnou (a přesně danou) vlastnost všech objektů základního souboru.
- **Rozsah výběru** je počet objektů n zahrnutých do výběru.
- **Reprezentativnost výběru** je vlastnost výběru, zaručující rovnoměrné zastoupení charakteristických vlastností objektů základního souboru.
- **Náhodný výběr** vznikne tehdy, když každý objekt základního souboru má stejnou pravděpodobnost být zahrnut do výběru.
- **Realizace výběru**: je množina naměřených (napozorovaných) číselných hodnot x_1, x_2, \dots, x_n veličin z výběru.



Malý slovníček statistických pojmů, aneb

co je třeba znát a porozumět tomu:

- **Uspořádaný výběr:** $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ vznikne z původního výběru X_1, X_2, \dots, X_n upořádáním podle velikosti pozorovaných hodnot x_1, x_2, \dots, x_n .
- **Pořadová statistika:** $X_{(k)}$ je náhodná veličina X_m , která je k -tá v pořadí podle velikosti pozorovaných hodnot x_1, x_2, \dots, x_n .
- **Pořadí m -tého pozorování veličiny X_m ve výběru:**
pokud $X_m = X_{(k)} \Rightarrow R_m = k$.
- $X_{(1)}$ se nazývá **(výběrové) minimum**, $X_{(n)}$ je **(výběrové) maximum**
- **medián** je prostřední hodnota ve výběru: je-li n liché, je roven $X_{([n/2]+1)}$
pro n sudé je roven $(X_{(n/2)} + X_{(n/2+1)})/2$
- **dolní kvartil:** $X_{([n/4]+1)}$ resp. $(X_{(n/4)} + X_{(n/4+1)})/2$
- **horní kvartil:** $X_{([3n/4]+1)}$ resp. $(X_{(3n/4-1)} + X_{(3n/4)})/2$
- **Výběrový modus** je nejčastější hodnota, která se vyskytuje v realizaci výběru. Tato hodnota nemusí existovat.



Malý slovníček statistických pojmů, aneb

co je třeba znát a porozumět tomu:

- **Výběrový průměr** nahrazuje neznámou střední hodnotu veličiny X :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- **Výběrový rozptyl** je charakteristika odpovídající rozptylu náhodné veličiny X

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- **Výběrová směrodatná odchylka** s je druhou odmocninou z výběrového roptylu
- **Výběrový index šikmosti** je výběrovou variantou indexu šikmosti a je mírou symetrie pozorované veličiny X

$$Skew(X) = \frac{m_3(X)}{m_2^{3/2}(X)}$$

- **Výběrový index špičatosti** je výběrovou variantou indexu špičatosti a je mírou soustředění hodnot pozorované veličiny X kolem průměru.

$$Kurt(X) = \frac{m_4(X)}{m_2^2(X)} - 3$$



Malý slovníček statistických pojmů, aneb

co je třeba znát a porozumět tomu:

- **Třídní intervaly** rozdělují maximální rozsah pozorovaných hodnot náhodné veličiny (od minima do maxima) na k stejných dílů.
- **(prostá absolutní) četnost i -té třídy** je počet pozorování náhodné veličiny X v i -té třídě, $i = 1, \dots, k$.
- **(prostá) relativní četnost i -té třídy** je poměr počtu pozorování náhodné veličiny X v i -té třídě ku rozsahu výběru n , $i = 1, \dots, k$.
- **kumulativní (absolutní) četnost i -té třídy** je počet pozorování náhodné veličiny X od minima až do i -té třídy včetně, $i = 1, \dots, k$.
- **kumulativní relativní četnost i -té třídy** je součet relativních četností pozorování náhodné veličiny X až do i -té třídy včetně, $i = 1, \dots, k$.
- **Histogram četností** je grafické zobrazení četností ve formě sloupkového grafu. Relativní četnosti lze zobrazovat i ve formě kruhového (koláčového) grafu. Existuje celá řada variant.



Malý slovníček statistických pojmů, aneb

co je třeba znát a porozumět tomu:

- **Krabicový diagram** je grafické zobrazení rozdělení pozorovaných hodnot pomocí pěti (Tukey's) charakteristik: minima, dolního kvartilu, mediánu, horního kvartilu a maxima.
- **Empirická distribuční funkce** je grafické zobrazení realizace výběru formou grafu po částech konstantní funkce

vycházíme z uspořádaného výběru: $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. Potom

$$F_n(x_{(i)}) = \frac{i}{n} \quad \text{a tedy} \quad F_n(x) = \frac{\max\{k : X_{(k)} \leq x\}}{n}, \quad x \in \mathbf{R}$$

