

aktualizované  
02-10-2024

**KMA/PAS & KMA/PST**

Pravděpodobnost  
& statistika

ZS 2024/2025

© Martin Bod'a

[martin.boda@outlook.com](mailto:martin.boda@outlook.com)

<https://orcid.org/0000-0002-7503-6898>

# Podstata štatistiky

Štatistika sa zamestnáva napospol s **hromadnými** javmi a poskytuje ich opis, resp. umožňuje úsudky o nich. Pod hromadným javom sa rozumie (obyčajne náhodná) udalosť, ktorá sa opakuje a vyskytuje (nezávisle) „mnoho-krát“ a viedie k realizácii číselnej alebo nečíselnej hodnoty. Hromadnosť javu spôsobuje, že tieto hodnoty sa vyskytujú s určitou zákonitosťou / frekvenciou (a opakujú sa, resp. takmer sa opakujú). Následne sa hovorí o **frekvenčnom rozdelení** alebo o **rozdelení početnosti**. Štatistika sa potom sústredí uje na frekvenčné rozdelenie sledovaných premenných a jej úlohou je ho v nejakej podobe opísat', stanoviť či identifikovať, a to bez ohľadu na to, či ide o deskriptívnu alebo induktívnu štatistiku.

## Príklad – hromadnosť javov

Príkladom pre numerickú premennú je „cena jedného kopčeka obyčajnej zmrzliny v Ústí nad Labem v letnej sezóne roku 2020“. Táto sa mohla pohybovať, povedzme, od 15 do 25 Kč a závisela od prevádzky. Každá prevádzka uplatňovala samostatnú cenu a tieto ceny sa s určitou pravidelnosťou opakovali.

hromadnosť = to, že bolo viacero prevádzok s vlastnou cenou

frekvencia = kol'kokrát sa opakovali tie isté ceny

Príkladom pre nominálnu premennú je „preferovaná príchut“ (preferovaný druh) zmrzliny kupovaný v Ústí nad Labem v letnej sezóne roku 2020“. Závisí od každého konzumenta, či uprednostnil čokoládovú, šmolkovú, nutellovú, ananásovú, jogurtovú alebo inú príchut“. Tieto prichute sa objavovali medzi konzumentmi niekol'kokrát.

hromadnosť = to, že bolo viacero konzumentov s preferenciou

frekvencia = kol'kokrát sa opakovali tie isté preferencie chute

# Frekvenčné rozdelenie a frekvenčná krivka

O frekvenčnom rozdelení sa hovorí pri numerických premených aj kategoriálnych premenných, ale o frekvenčnej krivke sa hovorí iba pri numerických premenných na spojitej škále.

## Numerické premenné (kvantitatívne znaky):

**Frekvenčné rozdelenie (rozdelenie početnosti)** je zákonitosť (pozorovaná pravidelnosť), ktorá určuje, kol'kokrát sa hodnoty premennej opakujú (pri diskrétnych premenných), alebo ktorá určuje rozdelenie hodnôt cez postupné rovnako široké číselné intervaly (pri spojitých premenných).

**Frekvenčná krivka** je ideálne matematické vyjadrenie (matematický funkčný predpis) frekvenčného rozdelenia spojitej numerickej premennej.

## Ordinálne/nominálne premenné (kvalitatívne znaky):

**Frekvenčné rozdelenie (rozdelenie početnosti)** je zákonitosť (pozorovaná pravidelnosť), ktorá určuje, kol'kokrát sa hodnoty sledovanej premennej opakujú.

# Príklad 1

Denné výnosy záverečných cien akcie spoločnosti General Electric Company obchodovanej na NYSE od 03-01-1969 do 31-12-1998. Celkovo ide o 2528 pozorovaní uvedených v percentách p.d.

? hromadnosť

? frekvenčné rozdelenie

? frekvenčná krivka

Time Series:

Start = c(1969, 1)

End = c(1975, 338)

Frequency = 365

```
[1] -1.6760  1.7045 -0.2793  0.0000  0.0000 -0.5602  0.0000  1.4085
[9]  0.0000 -0.2778  0.8357  1.3812 -0.2725 -0.5464 -0.5495  1.3812
[17]  0.0000  2.1798  1.0667  1.3193  0.7812 -0.2584 -0.2591 -0.5195
[25]  0.2611  0.7812 -1.8088 -1.0526 -1.5957 -0.2703 -1.0840  1.3699
[33] -0.5405  1.9022  0.2667 -2.9255  1.0959 -1.8970  1.6575 -0.5435
```

## Príklad 1 (./.)

2528 denných výnosov akcie spoločnosti General Electric Company roztriedených do vhodne zvolených intervalov rovnakej šírky:

Frekvenčné rozdelenie sa charakterizuje početnosťami a rozlišujú sa štyri typy početností:

\*\* **n(i)** - **bežná absolútна početnosť**  
(kol'kokrát sa vyskytlo)

\*\* **N(i)** - **kumulatívna absolútna početnosť**  
(kol'kokrát sa vyskytlo „do“)

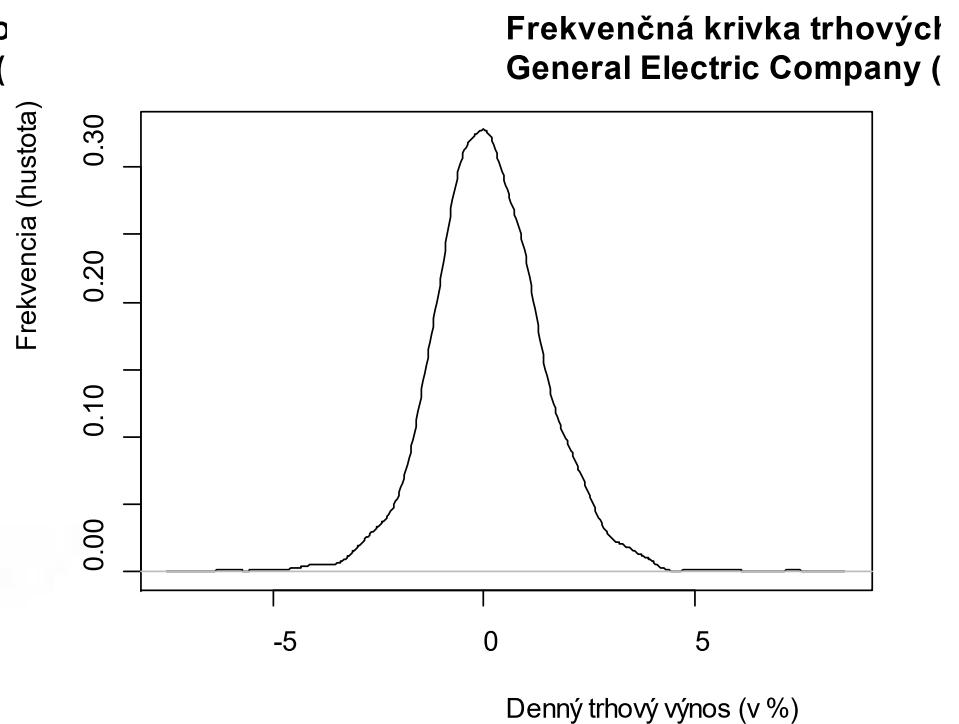
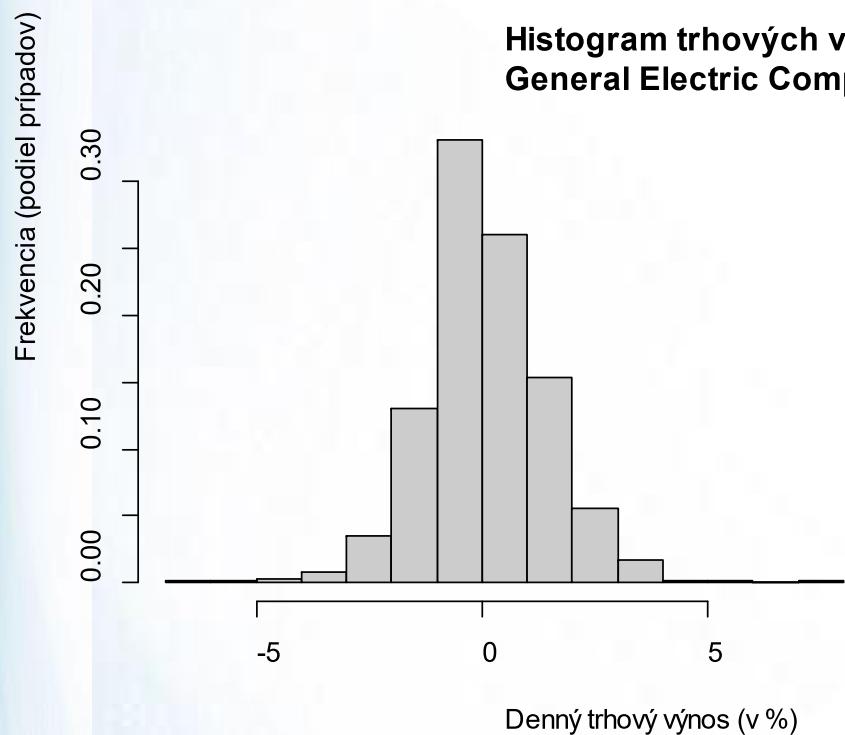
\*\* **f(i)** - **bežná relatívna početnosť**  
(v kol'ko percent prípadov sa vyskytlo)

\*\* **F(i)** - **kumulatívna relatívna početnosť**  
(v kol'ko percent prípadov sa vyskytlo „do“)

	n(i)	N(i)	f(i)	F(i)
(-7, -6]	4	4	0.002	0.002
(-6, -5]	4	8	0.002	0.004
(-5, -4]	6	14	0.002	0.006
(-4, -3]	20	34	0.008	0.014
(-3, -2]	90	124	0.036	0.050
(-2, -1]	328	452	0.130	0.180
(-1, 0]	835	1287	0.330	0.510
(0, 1]	657	1944	0.260	0.770
(1, 2]	387	2331	0.153	0.923
(2, 3]	142	2473	0.056	0.979
(3, 4]	44	2517	0.017	0.996
(4, 5]	3	2520	0.001	0.997
(5, 6]	4	2524	0.002	0.999
(6, 7]	1	2525	0.000	0.999
(7, 8]	3	2528	0.001	1.000
		2528		1.000

## Príklad 1 (./.)

Na grafické znázornenie frekvenčného rozdelenia spojitej premennej slúži histogram (používajú sa bežné absolútne alebo relatívne početnosti). Jeho vyhľadením (a zmenšovaním šírky triediaceho intervalu) vzniká frekvenčná krivka.



## Príklad 2

Ročné tržby obchodov so značkovým oblečením v Holandsku zaznamenaných v roku 1990. K dispozícii je 400 hodnôt ročných tržieb v 10000-násobkoch holandských guldenov.

? hromadnosť

? frekvenčné rozdelenie

? frekvenčná krivka

[1]	75.0000	192.6395	125.0000	69.4227	75.0000	40.0000	130.0000
[8]	49.5340	120.0000	49.5340	91.1315	23.1000	49.2033	150.0000
[15]	33.0000	87.0000	69.4227	24.5228	37.5000	33.0000	30.1133
[22]	97.6817	150.0000	7.7369	47.1000	60.1000	97.6817	85.0000
[29]	49.5340	69.4227	15.6168	49.5340	100.0000	69.4227	97.6817
[36]	15.6168	40.0000	17.9655	87.6000	30.1133	49.5340	30.1133
[43]	85.3000	97.6817	189.4931	97.6817	69.4227	49.5340	69.4227
[50]	97.6817	97.6817	81.5000	49.5340	171.8000	87.5000	79.0000

## Príklad 2 (./.)

ročné tržby 400 predajní so značkovým oblečením vytriedené do intervalov rovnakej šírky:

	n(i)	N(i)	f(i)	F(i)
(0, 50]	130	130	0.325	0.325
(50, 100]	182	312	0.455	0.780
(100, 150]	52	364	0.130	0.910
(150, 200]	27	391	0.068	0.978
(200, 250]	2	393	0.005	0.983
(250, 300]	3	396	0.008	0.991
(300, 350]	0	396	0.000	0.991
(350, 400]	3	399	0.008	0.999
(400, 450]	0	399	0.000	0.999
(450, 500]	1	400	0.001	1.001
	400		1.000	

\*\* n(i) - bežná absolútна početnosť

(kol'kokrát sa vyskytlo)

\*\* N(i) - kumulatívna absolútna početnosť

(kol'kokrát sa vyskytlo „do“)

\*\* f(i) - bežná relatívna početnosť

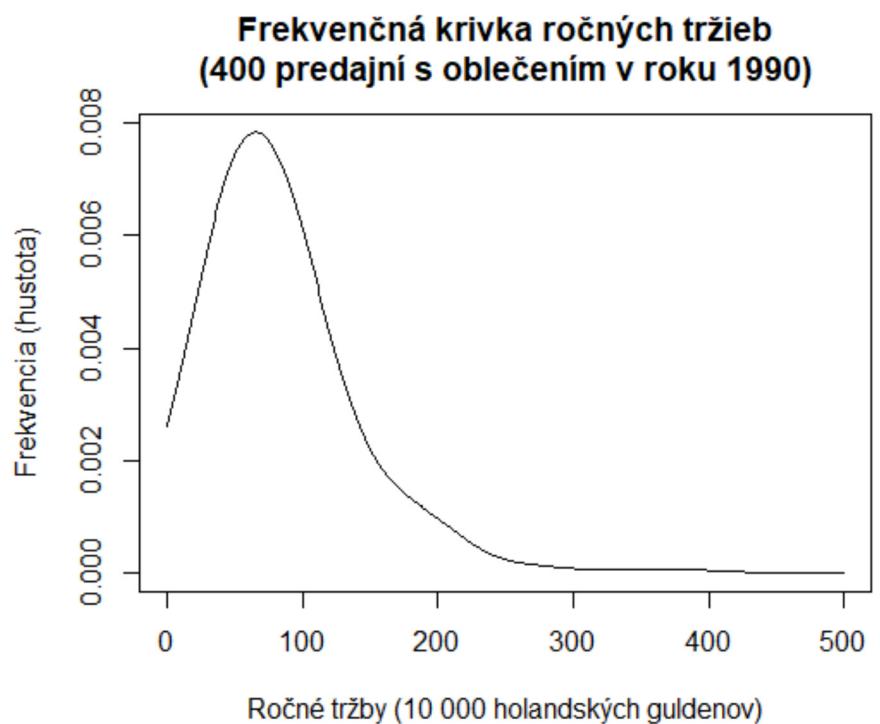
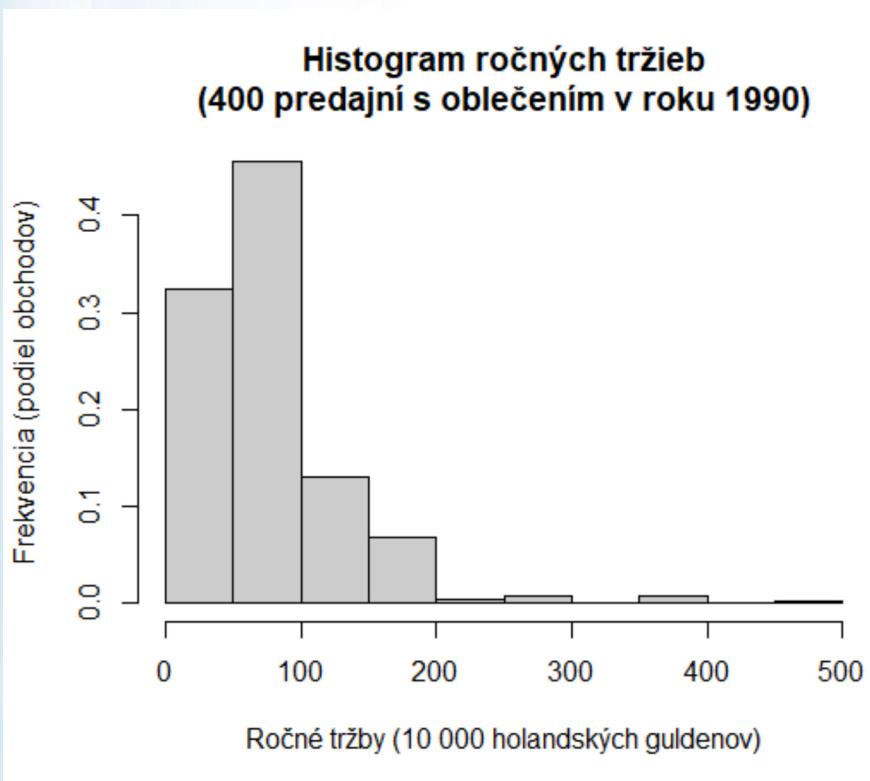
(v kol'ko percent prípadov sa vyskytlo)

\*\* F(i) - kumulatívna relatívna početnosť

(v kol'ko percent prípadov sa vyskytlo „do“)

## Príklad 2 (./.)

Frekvenčné rozdelenie / frekvenčná krivka ukazuje pre dátu o ročných tržbách iný charakteristický tvar a iné vlastnosti ako v predošлом príklade.



## Príklad 3

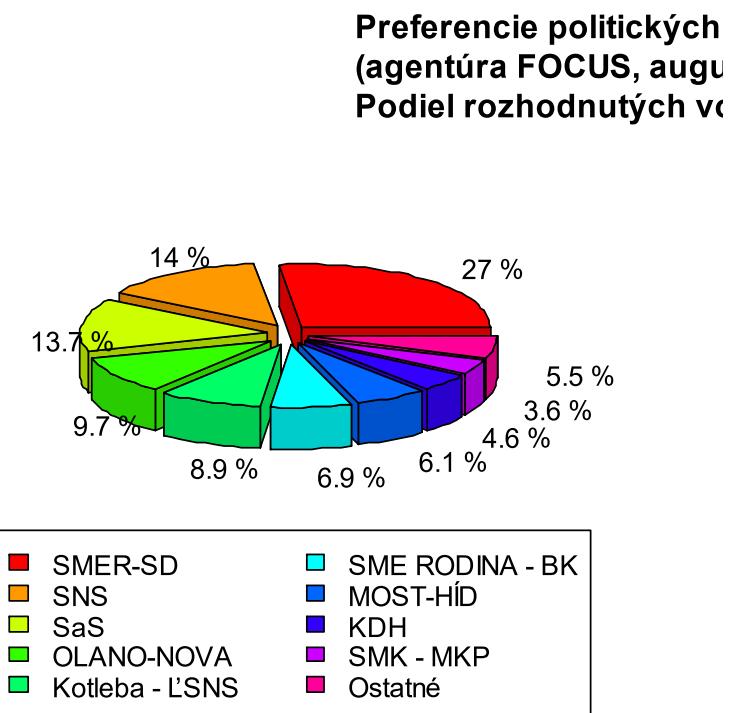
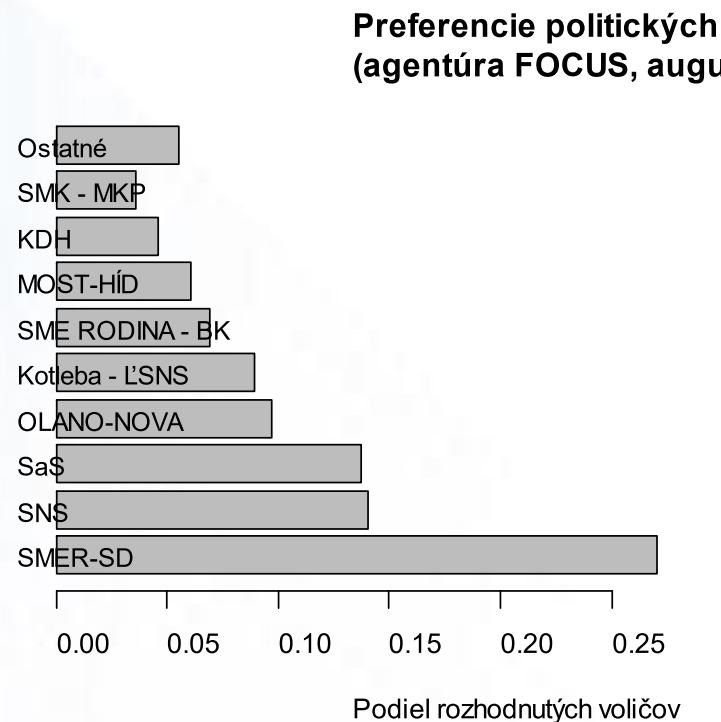
### Volebné preferencie politických strán - august 2016

Agentúra FOCUS uskutočnila v dňoch 11.8. - 16.8.2016 prieskum verejnej mienky formou osobného opytovania. Výberovú vzorku tvorilo 1002 respondentov, ktorí reprezentujú populáciu SR vo veku nad 18 rokov z hľadiska pohlavia, veku, vzdelania, národnosti, veľkostných kategórií sídiel a krajského členenia.

Respondentom bola položená nasledujúca otázka: "Teraz si, prosím, predstavte, že by sa parlamentné voľby konali nasledujúci víkend a kandidovali by nasledujúce politické strany a hnutia. Ktorej strane alebo hnutiu by ste dali svoj hlas?"

[http://www.focus-research.sk/files/  
215\\_Volebne%20preferencie%20politickych%20stran\\_august%202016.pdf](http://www.focus-research.sk/files/215_Volebne%20preferencie%20politickych%20stran_august%202016.pdf)

# Príklad 3 (./.)



# Základné typy frekvenčných kriviek

Pri konfrontovaní analýz spojитých numerických premenných v rôznych oblastiach spoločenských a prírodných vied (napr. rôzne ekonomické, fyzikálne, biometrické, meteorologické dát) sa ukázalo, že niektoré charakteristické reprezentívne tvary/typy frekvenčných rozdelení sa zvyknú opakovat' a na základe toho boli definované niektoré ideálne frekvenčné krivky (v induktívnej štatistike [--> d'alšia časť predmetu] sa nazývajú hustotami).

V zásade možno na úvodné účely rozlíšiť 4 základné typy frekvenčných kriviek:

1. **symetrické rozdelenie**,
2. **mierne asymetrické rozdelenie**,
3. **extrémne asymetrické rozdelenie (rozdelenie tvaru J)**,
4. **rozdelenie tvaru U**.

# Základné typy frekvenčných kriviek (./.)

symetrické rozdelenie	najmä biometrické a antropometrické dáta; v dokonalej podobe sa vyskytuje zriedka
mierne asymetrické rozdelenie	najfrekventovanejší typ vo všetkých oblastiach
extrémne asymetrické rozdelenie	v ekonómii sa vyskytuje často (napr. príjmy, výnosové a nákladové položky)
rozdelenie tvaru U	obvykle sa vyskytuje pri podieloch a dátach vyjadrených v percentách

# Symetrické rozdelenie

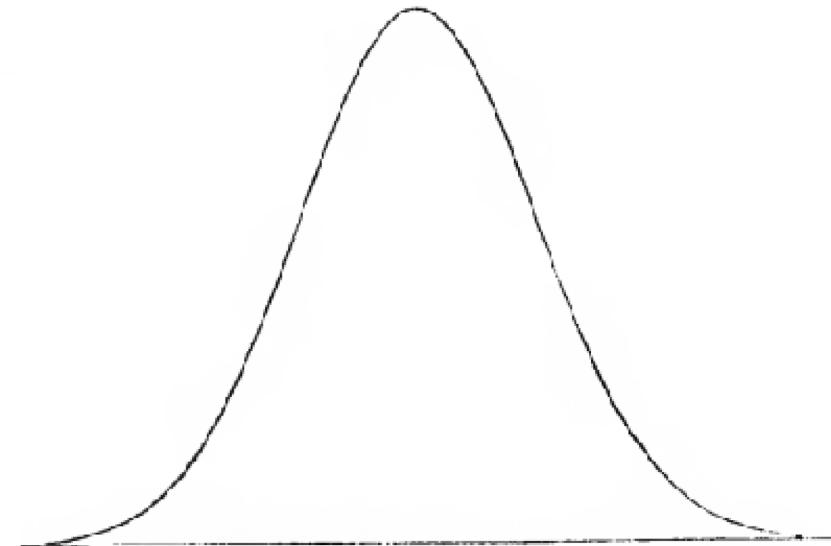


FIG. 5.—An ideal symmetrical Frequency-distribution.

Height without shoes, Inches.	Total.
57-	2
58-	4
59-	14
60-	41
61-	83
62-	169
63-	394
64-	669
65-	990
66-	1223
67-	1329
68-	1230
69-	1063
70-	646
71-	392
72-	202
73-	79
74-	32
75-	16
76-	5
77-	2
<b>Total</b>	<b>8585</b>

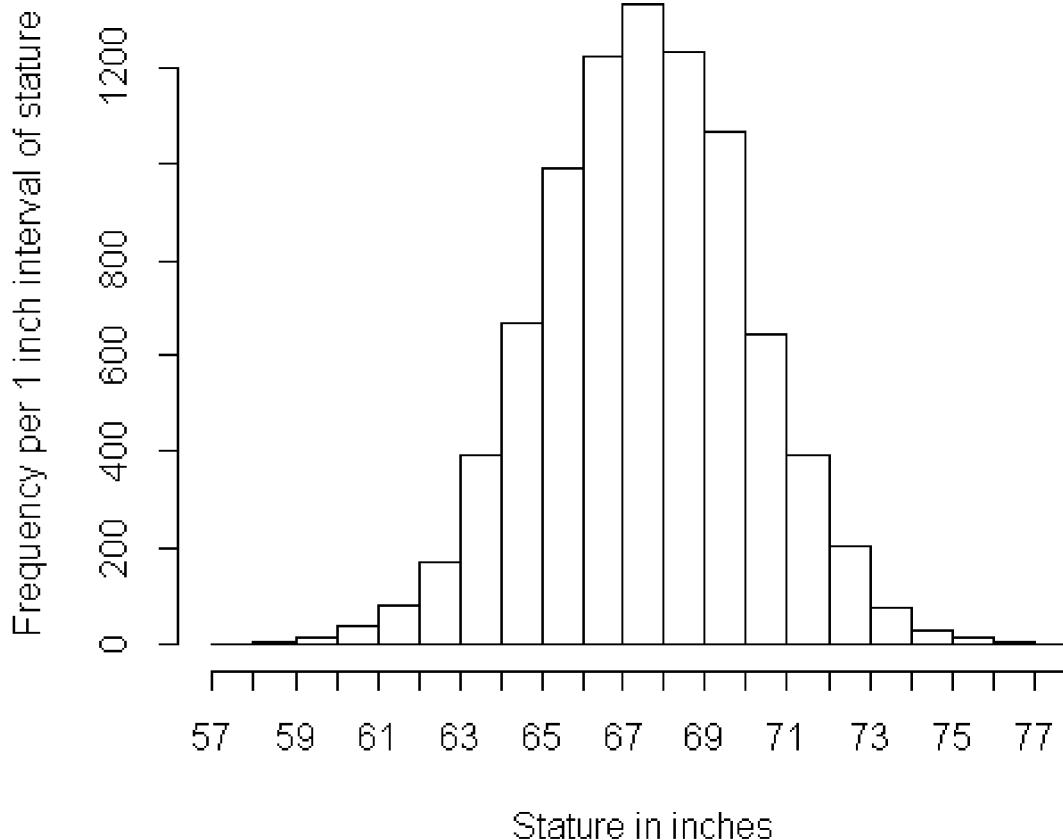


FIG. 6.—Frequency-distribution of Stature for 8585 Adnl Males born in the British Isles. (Table VI.)

(Zdroj:  
Yule,  
1924,  
s. 88-89)

# Mierne asymetrické rozdelenie

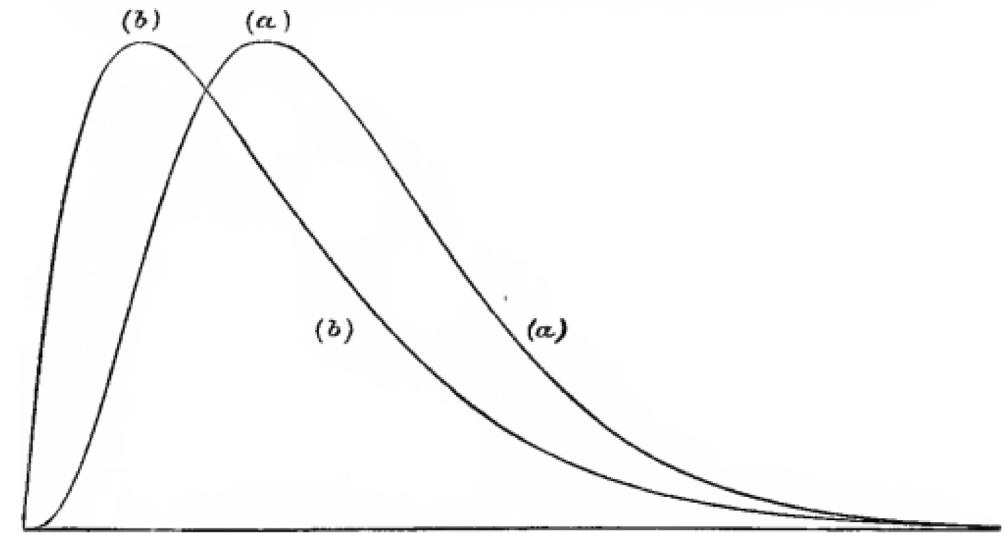


FIG. 9.—Ideal distributions of the moderately asymmetrical form.

Percentage of the Population in receipt of Relief.	Number of Unions with given Percentage in receipt of Relief.
0·75-1·25	18
1·25-1·75	48
1·75-2·25	72
2·25-2·75	89
2·75-3·25	100
3·25-3·75	90
3·75-4·25	75
4·25-4·75	60
4·75-5·25	40
5·25-5·75	21
5·75-6·25	11
6·25-6·75	5
6·75-7·25	1
7·25-7·75	1
7·75-8·25	0
8·25-8·75	1
Total	632

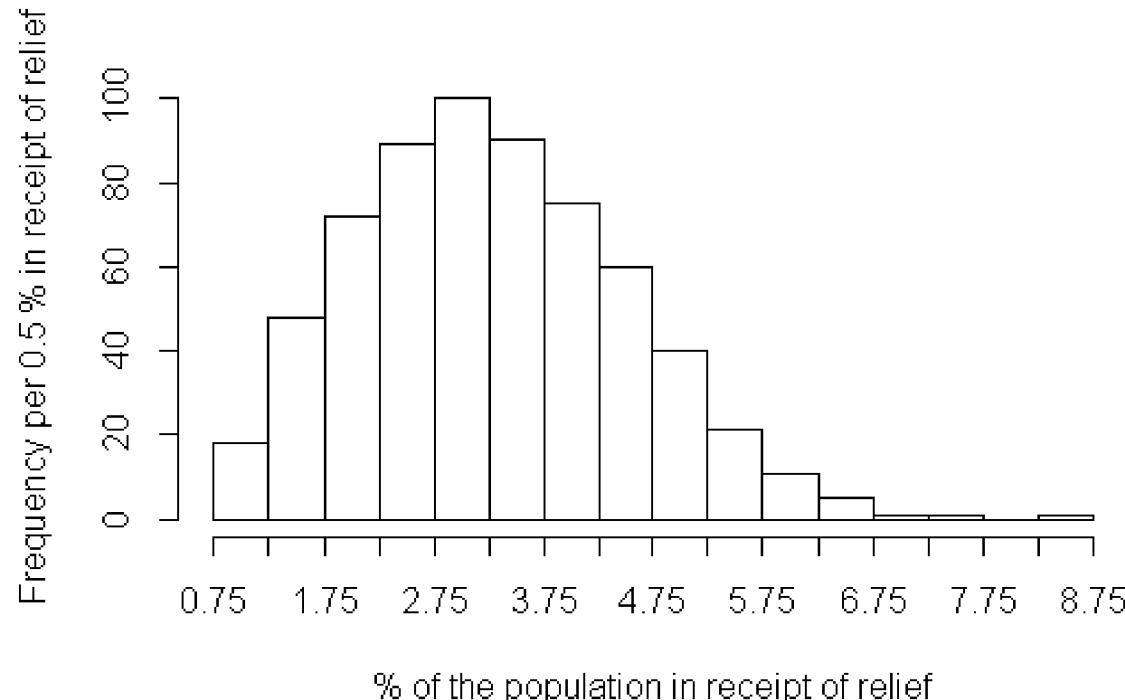


FIG. 10.—Frequency-distribution of Pauperism (Percentage of the Population in Receipt of Poor-law Relief) on 1st January 1891 in the Registration Districts of England and Wales: 632 Districts. (Table VIII.)

(Zdroj:  
Yule,  
1924,  
s. 92-93)

# Extrémne asymetrické rozdelenie

(Zdroj:  
Yule,  
1924,  
s. 99-101)

Annual Value in £100.	Number of Estates.	Annual Value in £100.	Number of Estates.
0- 1	1726·5	17-18	1
1- 2	280	18-19	—
2- 3	140·5	20-21	4
3- 4	87	21-22	1
4- 5	46·5	22-23	1
5- 6	42·5	23-24	1
6- 7	29·5	25-26	—
7- 8	25·5	27-28	2
8- 9	18·5	29-30	—
9-10	21	31-32	1
10-11	11·5	33-34	—
11-12	9·5	35-36	1
12-13	4	37-38	—
13-14	3·5	39-40	1
14-15	8	41-42	—
15-16	3	43-44	1
16-17	5	45-46	—
Total		2476	

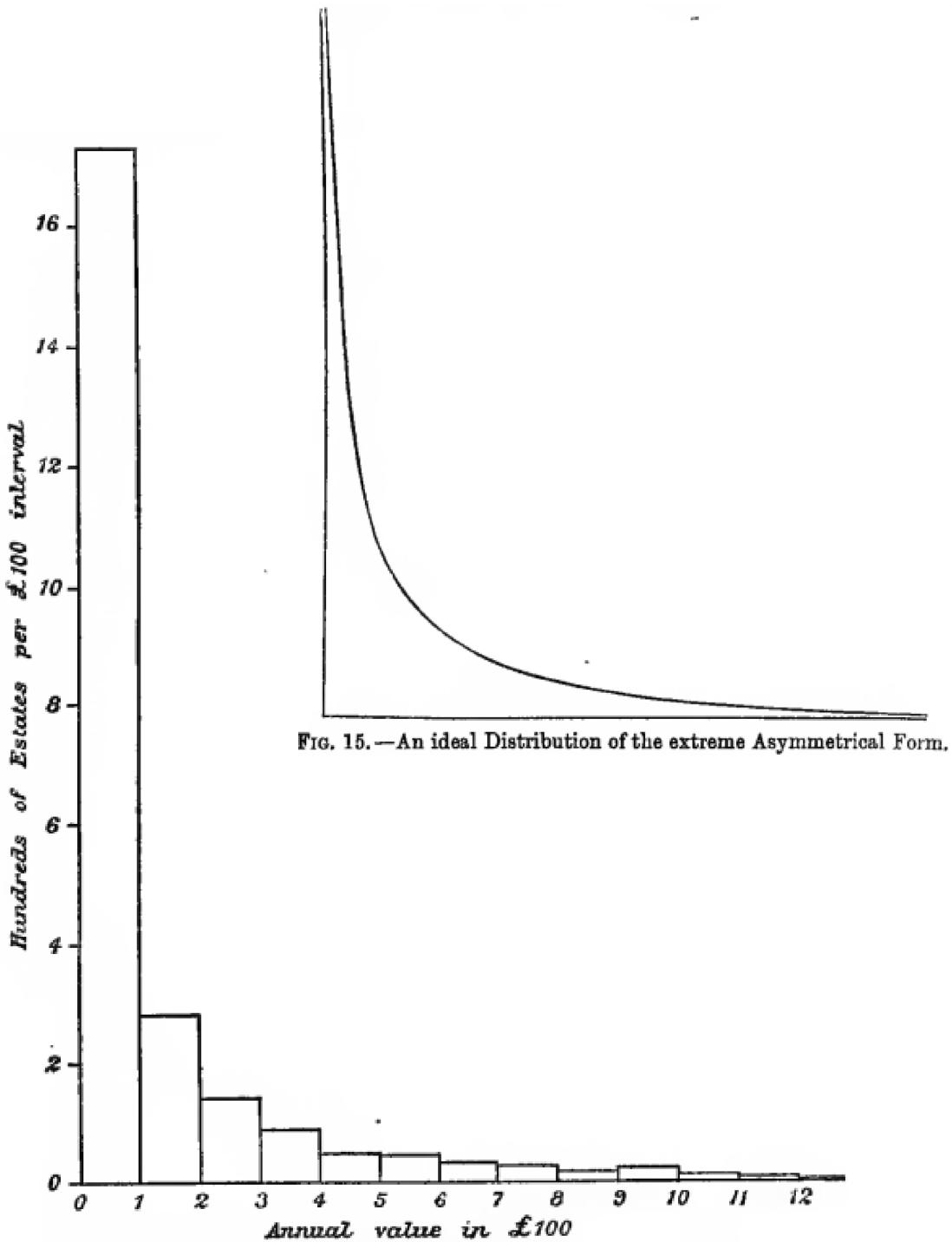


FIG. 15.—An ideal Distribution of the extreme Asymmetrical Form.

FIG. 16.—Frequency-distribution of the Annual Values of certain Estates in England in 1715 : 2476 Estates. (Table XIII.)

# Rozdelenie tvaru U

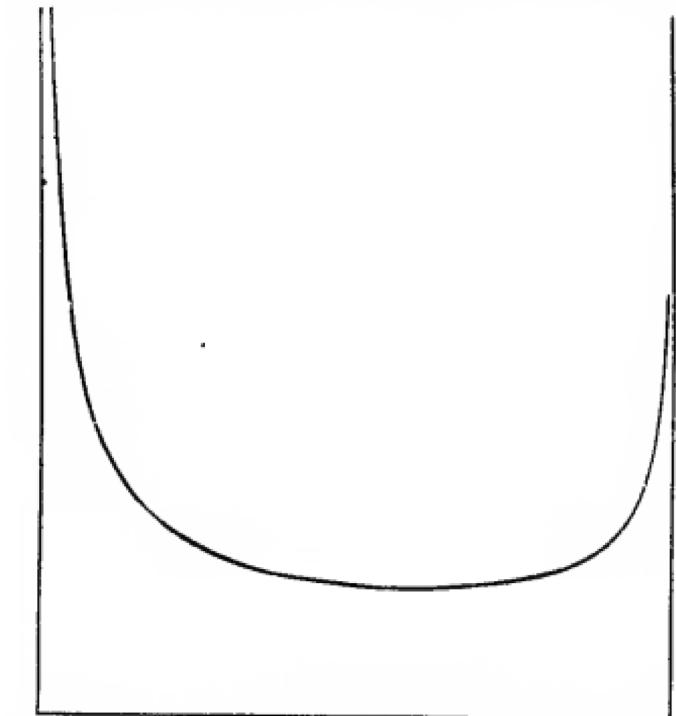


FIG. 18.—An ideal Distribution of the U-shaped Form.

(Zdroj:

Yule,

1924,

s. 103-104)

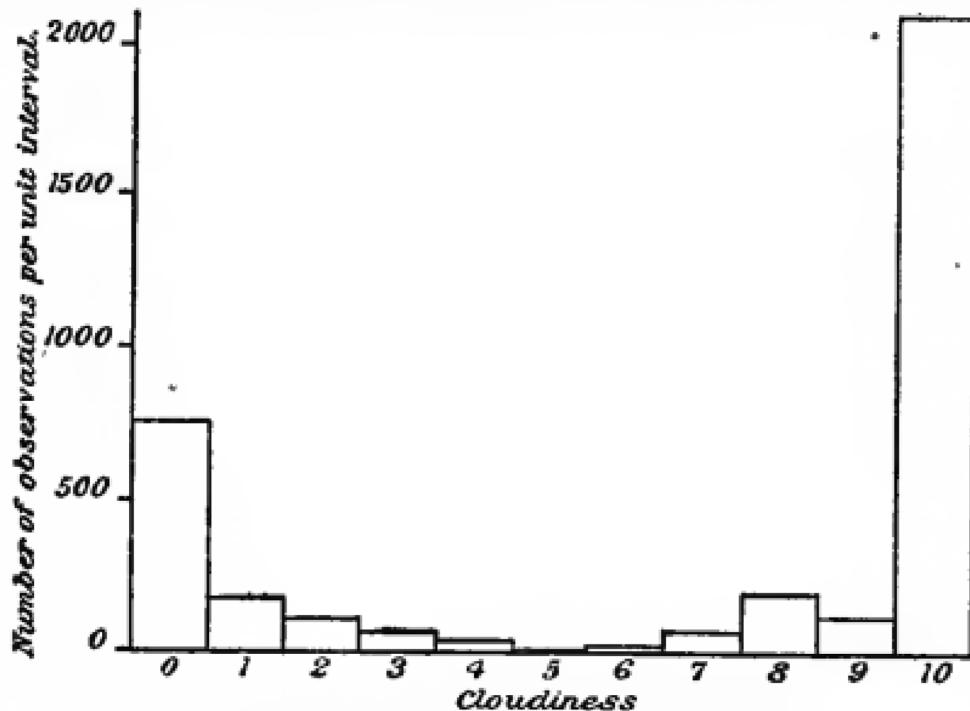


FIG. 19.—Frequency-distribution of Degrees of Cloudiness at Breslau  
1876-85 : 3653 observations. (Table XV.)

Cloudiness.	Frequency.	Cloudiness.	Frequency.
0	751	6	21
1	179	7	71
2	107	8	194
3	69	9	117
4	46	10	2089
5	9		
		Total	3653

# VŠEOBECNÝ POSTUP ŠTATISTICKEJ ANALÝZY

univerzálny pre prierezové dáta, časové rady,  
opakované merania aj panelové dáta

# Všeobecný postup štatistického spracovania a vyhodnotenia dát

1. Stanovenie cieľa a rozsahu štatistickej analýzy.
2. Získanie dát.

Ak je cieľ štatistická inferencia, musí íst o náhodný výber (alebo experimentálne dát).

3. Vizualizácia a grafická explorácia dát

S cieľom slovne (p)opísat' dátovú vzorku, identifikovať odľahlé pozorovania (outliery) a na základe toho zvoliť ďalší postup spracovania dát.

4. Štatistické spracovanie dát.
5. Interpretácia výsledkov.

# TRIEDENIE

numerických aj kategoriálnych premenných

# Pojem triedenia

Triedenie je usporiadanie súboru (dátovej vzorky) do viac alebo menej homogénnych skupín (tried) podľa hodnôt alebo obmien určitej triediacej premennej alebo premenných.

## \* podľa kategoriálnych premenných

znamená pre každú obmenu premennej zistit' počet jej výskytov v súbore, tzn. pre každú obmenu stanoviť bežné početnosti

## \* podľa numerických premenných

znamená pre každú hodnotu alebo rozpätie hodnôt premennej počet výskytov v súbore, tzn. pre každú hodnotu stanoviť bežné početnosti. toto triedenie môže byť diskrétné (ak premenná znak nadobúda malý počet hodnôt) alebo do intervalov (intervalové; ak premenná nadobúda veľký počet heterogénnych hodnôt).

# Pojem triedenia (./.)

\* triedenie podľa jednej premennej (jednostupňové)

\*\* kategoriálne premenné \*\*

Pohlavie respondenta

Pohlavie	Počet respondentov
muž	29
žena	72
Spolu	101

\*\* numerické premenné \*\*

\*\* diskrétné triedenie \*\*

Počet dní prípravy na test  
(ca. 8 hodín denne)

Dni	Počet žiakov
1	35
2	41
3	68
4	14
5	6
Spolu	164

\*\* numerické premenné \*\*

\*\* intervalové triedenie \*\*

Počet bodov z testu

Body	Počet študentov
<= 50	35
<= 60	41
<= 70	35
<= 80	37
<= 90	10
<= 100	6
Spolu	164

Pozn.: Sú uplatňované rôzne pravidlá pre intervalové triedenie numerických premenných na stanovenie počtu tried a ich rozpäťia (šírky), napr. Sturgesovo pravidlo, Scottovo pravidlo.

# Pojem triedenia (./.)

\* triedenie podľa dvoch premenných (dvojstupňové)

\*\* kategoriálne premenné \*\*  
kategoriálna vs. kategoriálna  
**KONTINGENČNÁ TABUĽKA**

Kraj	Pohlavie		Spolu
	muž	žena	
BA	35	29	64
BB	21	16	37
KE	26	19	45
NR	13	10	23
PO	6	6	12
TN	8	9	17
TT	15	6	21
ZA	21	22	43
Spolu	145	117	262

\*\* numerické premenné \*\*  
numerická vs. numerická  
**KORELAČNÁ TABUĽKA**

Body	Počet dní prípravy na test (à 8 h)					Spolu
	1	2	3	4	5	
<= 50	35					35
<= 60		41				41
<= 70			35			35
<= 80				33	4	37
<= 90					10	10
<= 100					6	6
Spolu	35	41	68	14	6	164

\*!\* kontingenčná tabuľka: dvojstupňové triedenie: kategoriálna vs. kategoriálna premenná

\*!\* korelačná tabuľka: triedenie: numerická vs. numerická premenná

# Korelačná tabuľka „v praxi“

**III.2–9. Sobáše podľa veku snúbencov v roku 2003**  
 Marriages by age of engaged couples in 2003



Vek ženicha Age of groom	Spolu Total	Vek nevesty										Age of bride			
		– 19	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49	50 – 54	55 – 59	60 +				
– 19	499	310	151	34	4	–	–	–	–	–	–				
20 – 24	6 706	1 209	4 368	1 013	95	17	–	1	3	–	–				
25 – 29	10 659	516	4 932	4 594	517	71	23	5	1	–	–				
30 – 34	4 065	84	1 052	1 927	819	131	39	10	3	–	–				
35 – 39	1 677	17	265	626	436	243	65	19	5	1	–				
40 – 44	885	7	78	203	201	202	127	48	15	3	1				
45 – 49	531	3	36	71	94	94	99	87	36	8	3				
50 – 54	422	–	18	35	45	59	77	103	70	10	5				
55 – 59	241	–	9	16	10	16	36	53	62	29	10				
60 +	317	–	2	14	6	11	20	41	57	73	93				
<b>Spolu</b>															
<b>Total</b>	<b>26 002</b>	<b>2 146</b>	<b>10 911</b>	<b>8 533</b>	<b>2 227</b>	<b>844</b>	<b>486</b>	<b>367</b>	<b>252</b>	<b>124</b>	<b>112</b>				

Zdroj: Štatistická ročenka Slovenskej republiky 2004

# Pojem početnosti

Početnosti: bežné & kumulatívne.

Početnosti: absolútne & relatívne.

\* pre všetky (typy) premenných (numerické aj kategoriálne)

**Bežná absolútna početnosť** -  $n_i$  - počet jednotiek, ktoré nadobúdajú  $i$ -tu obmenu (hodnotu) premennej  $x_i$  (pre všetky  $i \in \{1, \dots, m\}$ ).

**Bežná relatívna početnosť** -  $f_i$  - podiel jednotiek, ktoré nadobúdajú  $i$ -tu obmenu (hodnotu) premennej  $x_i$  (pre všetky  $i \in \{1, \dots, m\}$ ).

## Pojem početnosti (./.)

\* pre všetky ordinálne a numerické premenné

**Kumulatívna absolútна početnosť** -  $N_i$  - počet jednotiek, ktoré nadobúdajú najviac  $i$ -tu obmenu premennej  $x_i$  (pre všetky  $i \in \{1, \dots, m\}$ ).

**Kumulatívna relatívna početnosť** -  $F_i$  - podiel jednotiek, ktoré nadobúdajú najviac  $i$ -tu obmenu premennej  $x_i$  (pre všetky  $i \in \{1, \dots, m\}$ ).

Výsledkom (jednostupňového) triedenia je vytvorenie  $m$  tried s priradenými absolútymi početnosťami:

$x_1, x_2, \dots, x_m$  -  $n_1, n_2, \dots, n_m$  - ( $n_1 + n_2 + \dots + n_m = n$ ),  
resp. s relatívnymi početnosťami

$x_1, x_2, \dots, x_m$  -  $f_1, f_2, \dots, f_m$  - ( $f_1 + f_2 + \dots + f_m = 1$ ).

# ANALÝZA NUMERICKÝCH PREMENNÝCH

# Symbolika

Numerické premenné sa v deskriptívnej štatistike obvykle označujú veľkými písmenami od konca abecedy ( $X, Y, Z$ , alt. aj  $U, V, W$ ) a hodnoty (realizácie), ktoré nadobúdajú, sa označujú zodpovedajúcimi malými písmenami ( $x, y, z, u, v, w$ ). Jednotlivé hodnoty (realizácie) sa dôsledne označujú dolnou indexáciou.

Obvyklá symbolika:

+  $n$  - rozsah súboru (dátovej vzorky)

+  $X$  - premenná

+  $x_i$  - (dajaká)  $i$ -ta hodnota premennej  $X$  (kde  $i \in \{1, \dots, n\}$ )

+  $x_1, x_2, \dots, x_n$  - jednotlivé hodnoty premennej  $X$

+  $x_{(s)}$  - hodnota premennej  $X$ , ktorá je v usporiadanom súbore (dátovej vzorke) na  $s$ -tom mieste

## Príklad

Počas 33 po sebe idúcich prevádzkových dní v letných mesiacoch roka 2013 boli v predajni TESCO Expres Úsvit v Banskej Bystrici zaznamenané nasledovné denné tržby (tis. euro).

```
[1] 4.74 4.03 4.30 4.81 4.75 4.29 3.90 4.00 4.68 4.82 4.29  
[12] 4.06 4.23 5.25 5.99 5.82 5.24 5.55 6.18 5.17 5.38 5.44  
[23] 4.71 5.52 5.09 4.78 5.20 5.84 5.87 5.53 5.83 5.51 5.60
```

? populácia ? súbor ? (štatistická) jednotka ? premenná  
? (náhodný výber / nenáhodný výber) ? rozsah súboru ?  
zápis dát

# Všeobecný postup štatistickej analýzy dát

1. Stanovenie cieľa a rozsahu štatistickej analýzy.
2. Získanie dát.

Ak je cieľ štatistická inferencia, musí íst o náhodný výber (alebo experimentálne dát).

3. Vizualizácia a grafická explorácia dát

S cieľom slovne (p)opísat' dátovú vzorku, identifikovať odľahlé pozorovania (outliery) a na základe toho zvoliť ďalší postup spracovania dát.

4. Štatistické spracovanie dát.
5. Interpretácia výsledkov.

## Vizualizácia (exploračná analýza) dát

pre prípad jednej numerickej premennej

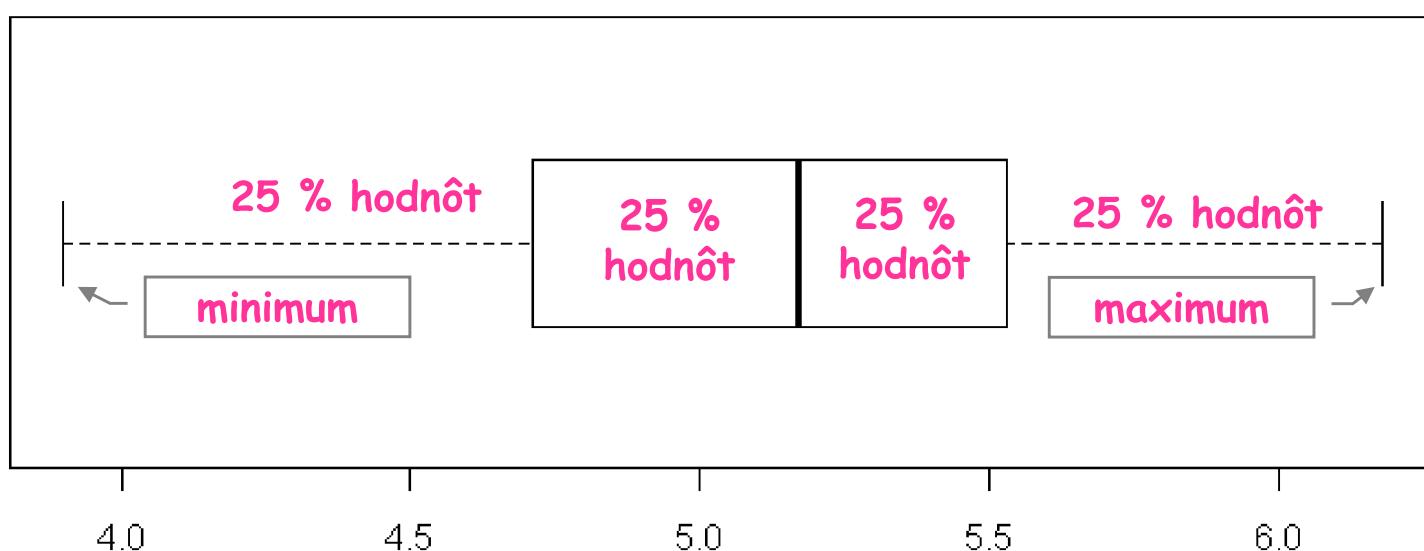
# Význam vizualizácie a grafickej explorácie dát

Úlohou je (možno) neexaktne stanoviť základné vlastnosti analyzovaného dátového súboru, v rámci čoho sa získané dáta vhodným a kombinovaným spôsobom zakresľujú a graficky znázorňujú. Aj v závislosti od zvoleného grafického prostriedku sa pritom sústredíme na určenie

- > minimálnej hodnoty a maximálnej hodnoty v dátovom súbore a rozpäťia dát,
  - > charakteristickej úrovne (umiestnenia) dát,
  - > rozptylenia a rovnomernosti rozmiestnenia dát,
  - > symetrie/zošikmenia dát,
- a snažíme sa posúdiť dátový súbor aj z hľadiska
- > výskytu anomálnych/atypických hodnôt.

# Box-plot

Nazýva sa tiež box-and-whisker plot, resp. v poslovenčenom ekvivalente krabiový graf. Uvedená ukážka predkladá box-plot bez existencie odľahlých hodnôt. Box-plot identifikuje rozvnhutie veľkostne usporiadaných štvrtín dátového súboru na číselnej osi. [Spravidla sa však kombinuje (zriedka však adekvátne) s procedúrou na identifikáciu odľahlých pozorovaní, ktoré sa znázorňujú hviezdičkami.]



Denné tržby predajne TESCO Express Úsvit v Banskej Bystrici (tis. eur)

## Box-plot (./.)

Box-plot je grafickým zobrazením tzv. Tukeyho 5 čísel (Tukey five numbers), ktoré pozostávajú z piatich deskriptívnych charakteristík:

- \* minimum (najmenšie pozorovanie),
  - \* dolný kvartil (resp. prvý kvartil),
  - \* medián (resp. stredný kvartil alebo druhý kvartil),
  - \* horný kvartil (resp. tretí kvartil),
  - \* maximum (najväčšie pozorovanie)
- a ktoré určujú rovnomernosť/nerovnomernosť rozdenenia hodnôt v štatistickom súbore.

Box-plot dáva informáciu o úrovni (polohe) hodnôt, ich rozpätí, dĺžke chvosta hodnôt a (v konvenčnej verzii) aj o odľahlých hodnotách.

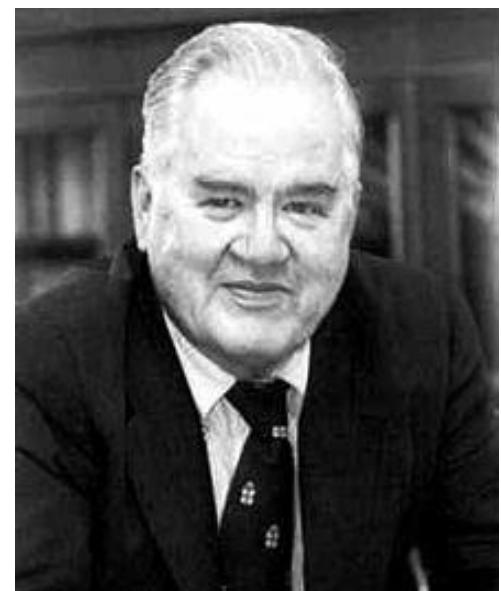
# Box-plot (./.)

Symbolika pre premennú  $X$ :

- +  $x_{\min}$  - minimum
- +  $x_{0.25}$  (alt.  ${}_xQ_1^4$ ) - dolný kvartil
- +  $x_{0.50}$  (alt.  ${}_xQ_2^4$  alebo  $Me_x$ , alebo  $\tilde{x}$ ) - medián
- +  $x_{0.75}$  (alt.  ${}_xQ_3^4$ ) - horný kvartil
- +  $x_{\max}$  - maximum (najväčšie pozorovanie)

**John Wilder Tukey** [tju:ki] (June 16, 1915 – July 26, 2000) was an American statistician best known for development of the FFT algorithm and box-plot.

Zdroj: <http://en.wikipedia.org/wiki/Tukey>

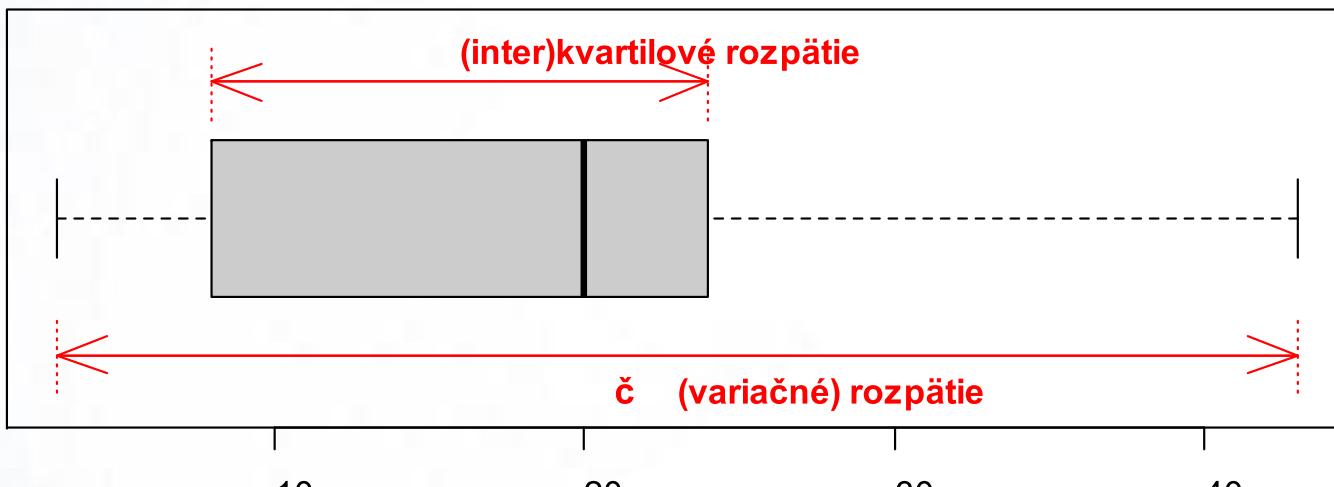


# Poznámka k interpretácii box-plotu

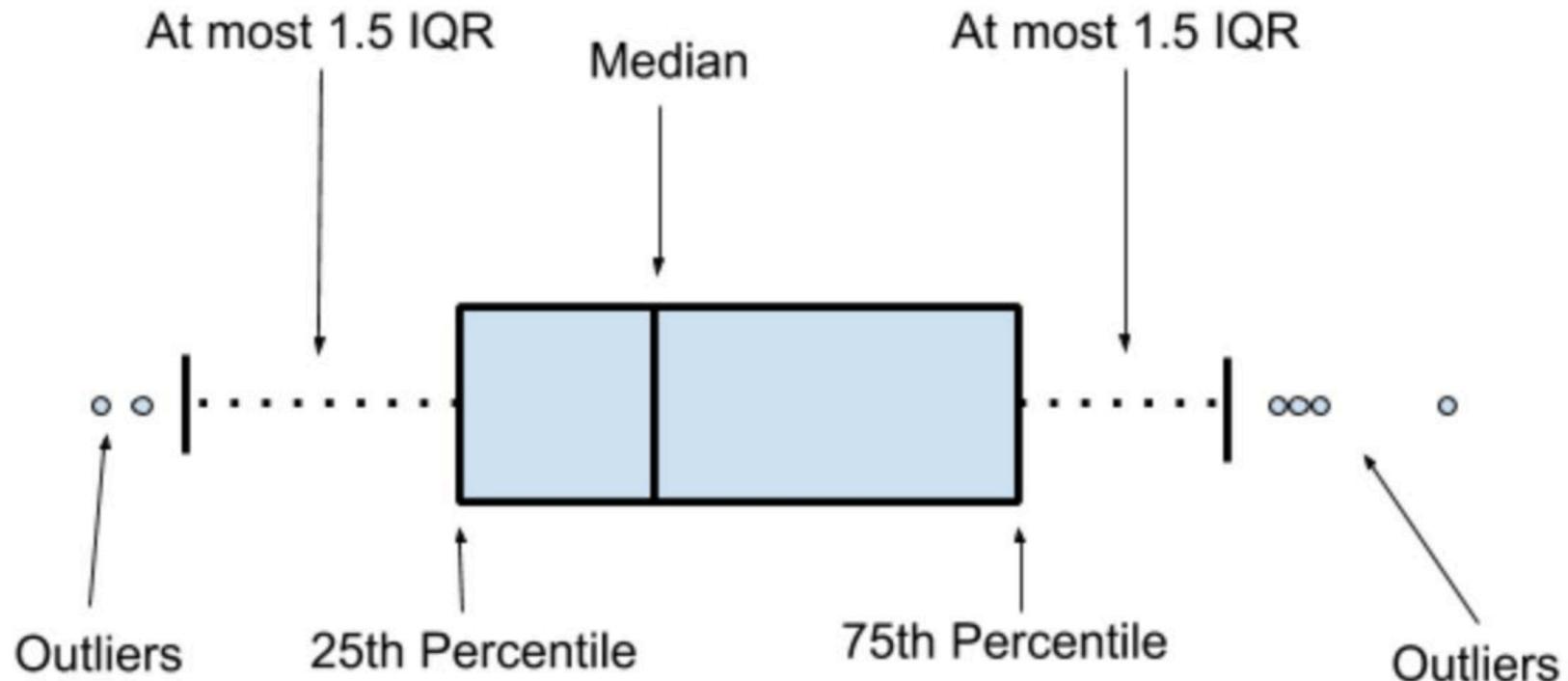
Rozdiel medzi maximálnou a minimálnou hodnotou sa nazýva **(variačné) rozpätie** a určuje ju rozpätie fúzikov. Udáva, ako sa najviac odlišujú napozorované hodnoty sledovanej premennej. Podobne, rozdiel medzi horným kvartilom a dolným kvartilom sa označuje ako **(inter)kvartilové rozpätie** a určuje najväčší rozdiel medzi (pro)strednými zhruba 50 % napozorovaných hodnôt danej premennej.

č

(Variačné) rozpätie [= šírka boxu] & (inter)



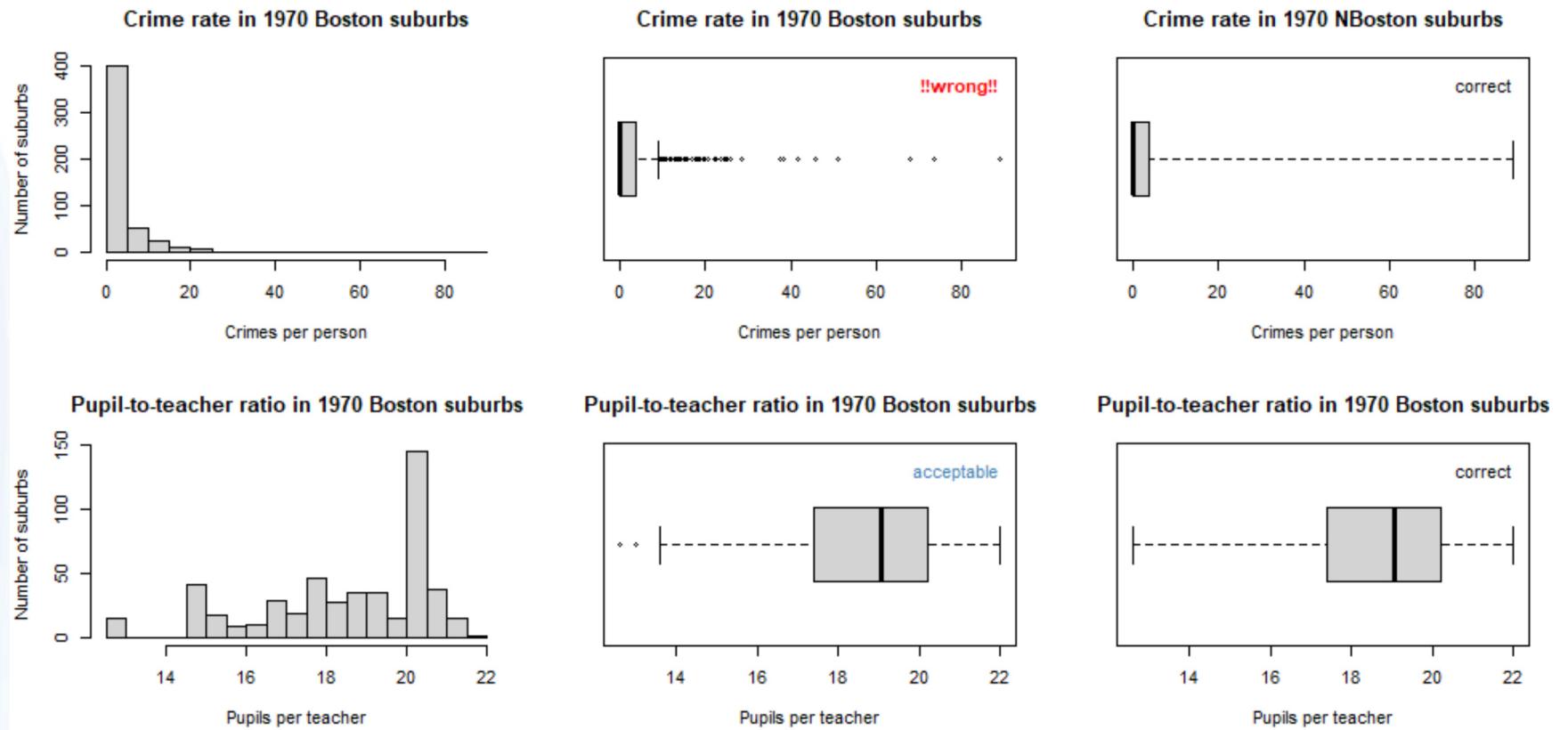
# Box-plot s odľahlými hodnotami



Tradičný prístup k identifikácii odľahlých hodnôt a ich zobrazovanie v box-plote je adekvátny výlučne pre symetrické dátá, resp. aspoň približne symetrické dátá. Pre asymetrické dátá je nutné buď zvoliť inú procedúru identifikácie odľahlých hodnôt, alebo klasifikáciu outlierov „vypnúť“.



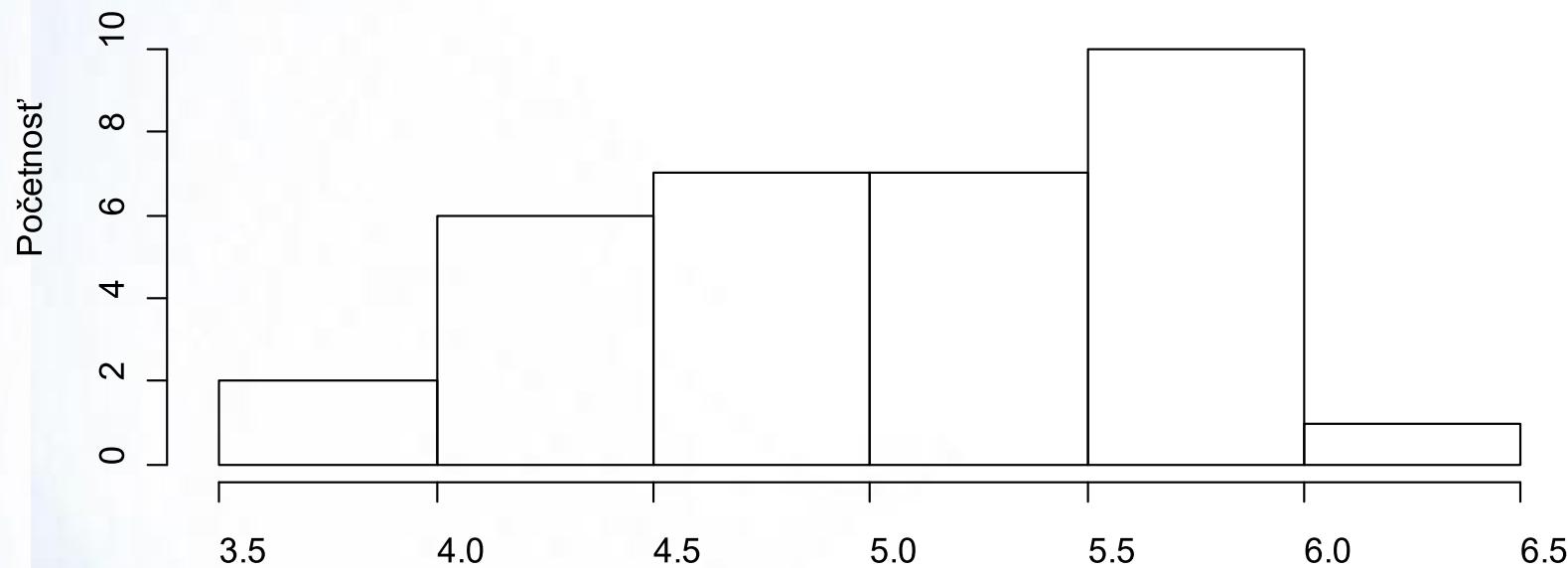
# Box-plot s odľahlými hodnotami



Konvenčná procedúra pre identifikáciu outlierov predpokladá symetrické rozdelenie a konfrontuje hodnoty vzdialené od centra s normálnym zákonom (gaussovskou frekvenčnou krivkou). V prvom prípade je takáto identifikácia a kreslenie outlierov nenáležité, v druhom prípade poruchy symetrie nie sú také drastické, čím má aj box-plot s identifikovaným outlierom výpovednú hodnotu.

# Histogram

Histogram je grafickým výrazovým prostriedkom pre získanie vizuálneho dojmu o rozdelení dát. Používa obdĺžniky pre znázornenie početnosti dát prislúchajúcich po sebe nasledujúcim intervalom (spravidla) rovnejšej šírky.



Denné tržby predajne TESCO Express Úsvit v Banske

## Histogram (./.)

Pri jeho konštrukcii sa zvolí vhodne počiatok  $x_0$  (origin, offset) splňujúci  $x_0 \leq x_{\min}$  a stanoví sa šírka okna (binu)  $h$  (binwidth). Konštruuje sa séria neprekryvajúcich sa intervalov o dĺžke  $h$ :

$$[x_0, x_0 + h), [x_0 + h, x_0 + 2h), [x_0 + 2h, x_0 + 3h), \dots$$

a histogram je schodíkovitým znázornením rozdelenia početnosti sledovaného znaku  $X$  daný vztahom

$$H(x) = \#\{x_i \text{ v rovnakom bine ako } x\}$$

(absolútne početnosti) alebo vztahom

$$H(x) = \frac{1}{n} \#\{x_i \text{ v rovnakom bine ako } x\}$$

(relatívne početnosti).

Poznámka: niekedy sa volí  $x_0 = x_{\min}$  a týmito intervalmi sa pokryje celé rozpätie  $[x_{\min}, x_{\max}]$ .

# Histogram (./.)

Výhodou histogramu je názornosť, nenáročnosť a jednoduchá konštrukcia, nevýhodou je najmä **senzitivita na voľbu offsetu ( $x_0$ ) a šírky okna ( $h$ )**. V každom prípade patrí spolu s box-plotom k najfrekventovanejším grafickým prostriedkom prezentácie dát používaným pri spracovaní dát (v štatistike aj data miningu).

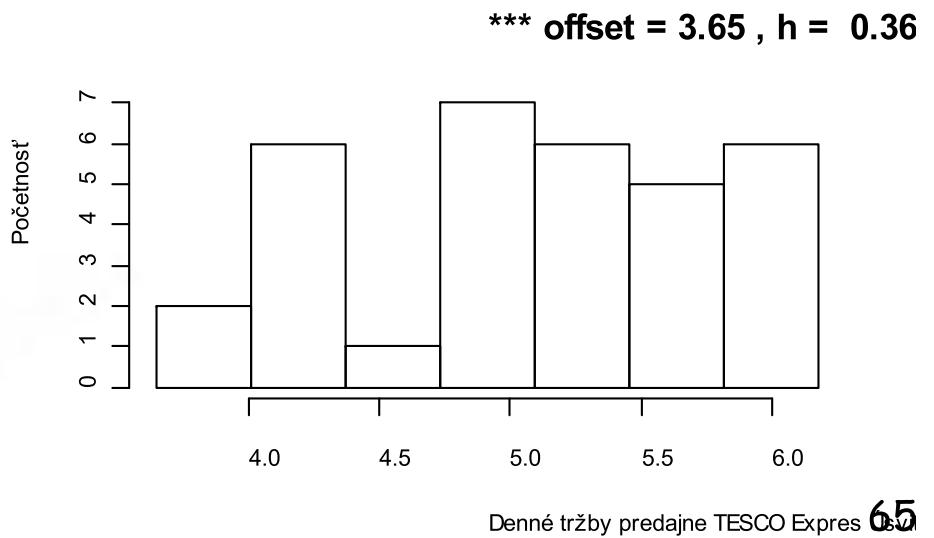
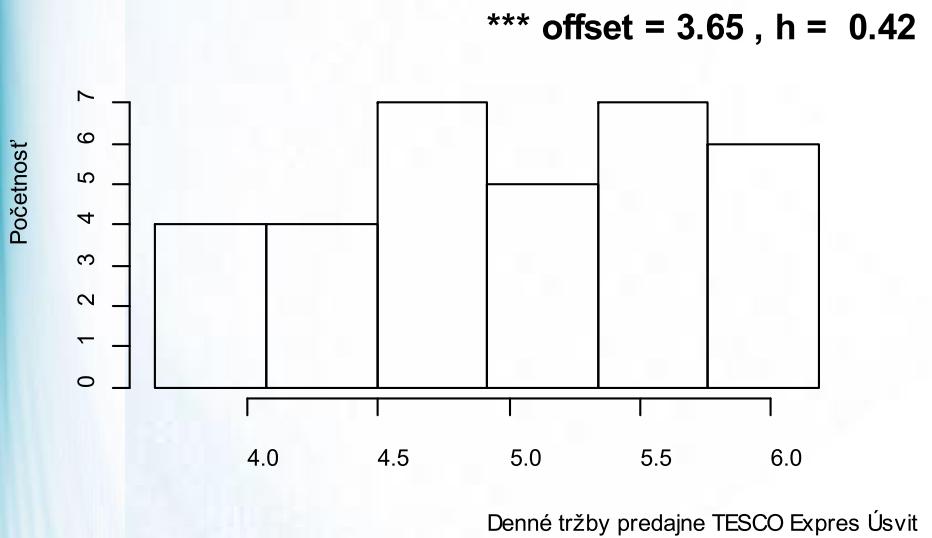
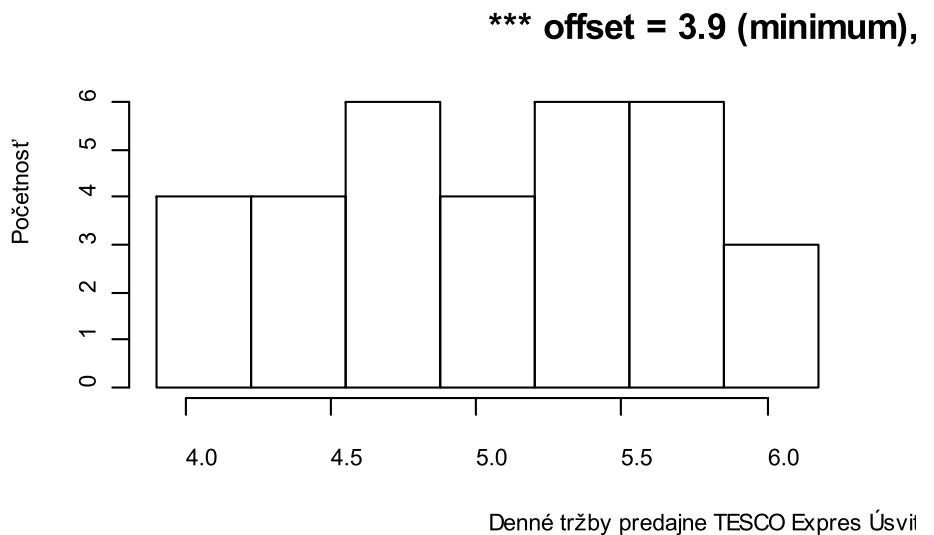
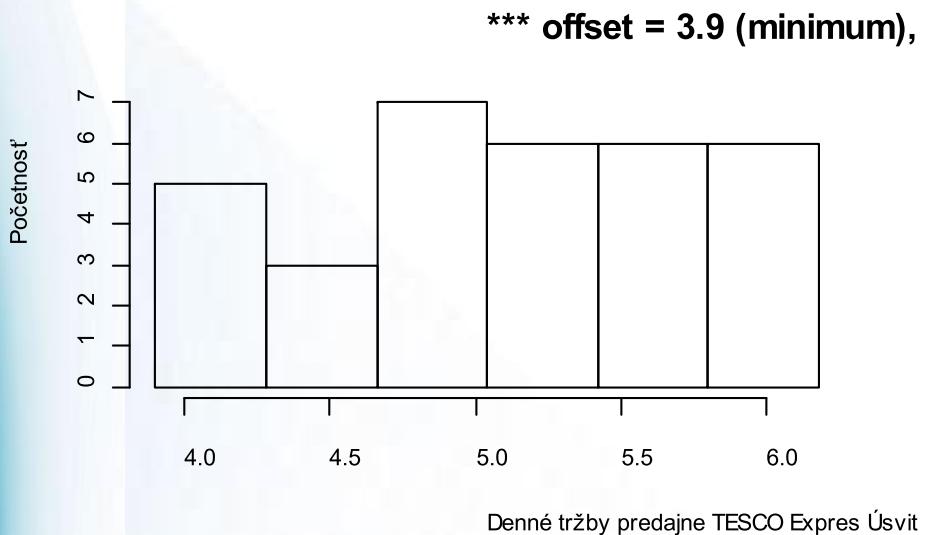
It [...] was first introduced by Karl Pearson.  
Karl Pearson (27 March 1857 - 27 April 1936) was an influential English mathematician who has been credited with establishing the discipline of mathematical statistics.

Zdroj: <http://en.wikipedia.org/wiki/Histogram>, Karl\_Pearson}



# Histogram (./.)

## Senzitivita na voľbu offsetu ( $x_0$ ) a šírky okna ( $h$ )



## Histogram (./.)

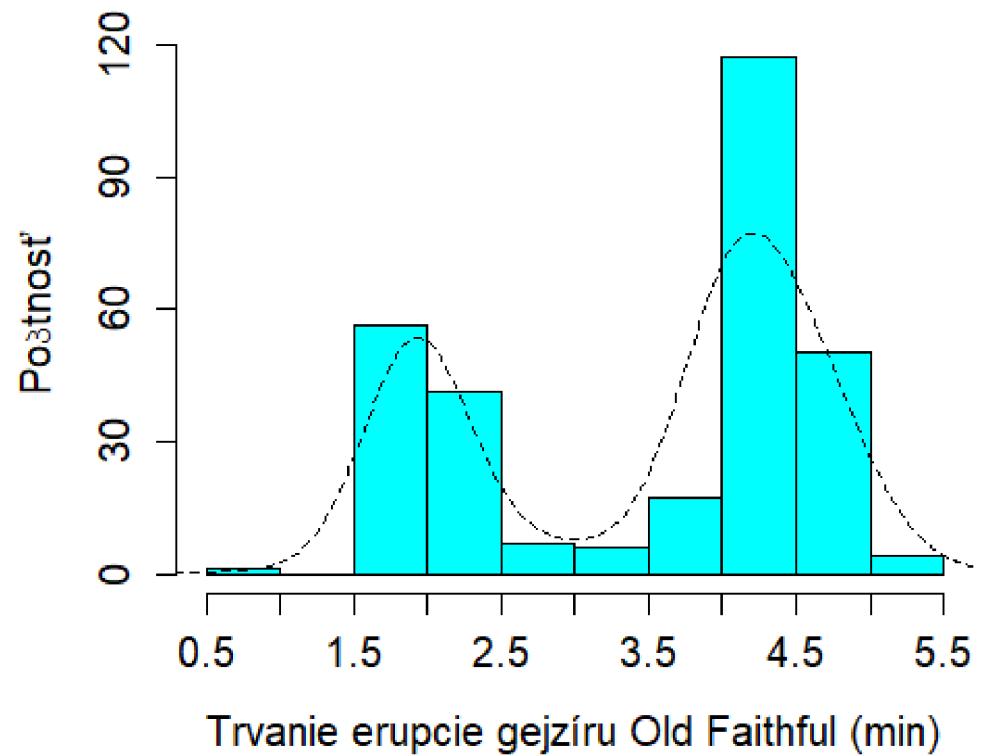
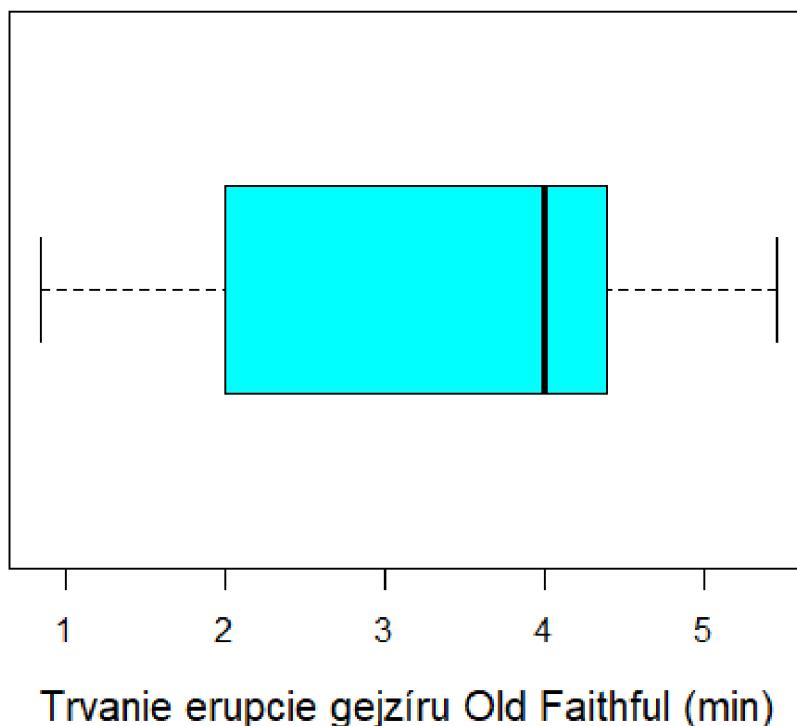
Sú rôzne pravidlá pre „optimálnu“ tvorbu intervalov, pre selekciu offsetu ( $x_0$ ) a šírky okna ( $h$ ), napr. Sturgesovo pravidlo, Scottovo pravidlo alebo pravidlo Freedmana & Diaconisa. Scottovo pravidlo napríklad odporúča optimálnu šírku okna určenú vztahom

$$h_{opt} = \left( \frac{24\sqrt{\pi}}{n} \right)^{\frac{1}{3}} s_x,$$

kde  $s_x$  je výberová smerodajná odcylka. Spravidla je celý proces tvorby histogramu ponechaný na softvér aj s defaultnými nastaveniami. Pri väčšom počte kompaktných dát vizuálny dojem je v zásade rovnaký pri rôznej vol'be offsetu a širky okna.

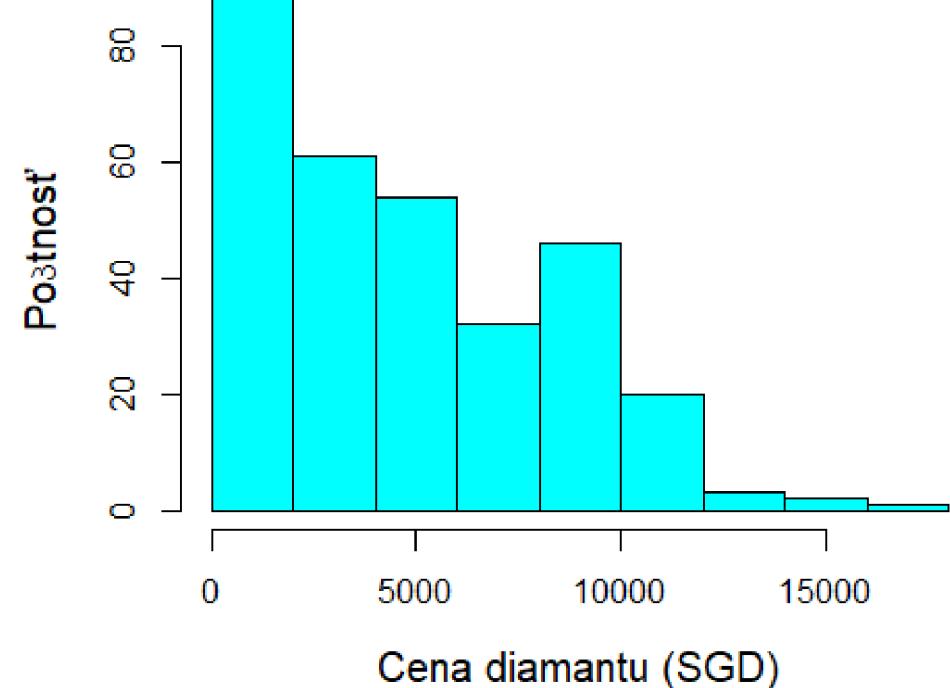
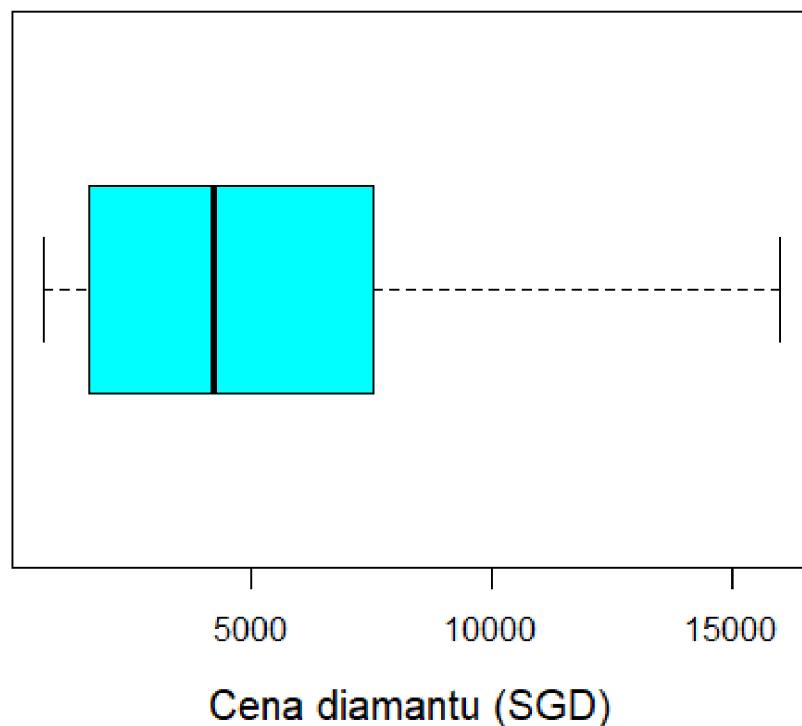
# Box-plot a histogram spolu

Príklad: **Old Faithful Geyser Data.** The eruptions data from the 'Old Faithful' geyser in Yellowstone National Park, Wyoming, the sample consists of 299 observations of continuous measurement from August 1 to August 15, 1985. **duration** --> a numeric variable describing the eruption time in minutes.



# Box-plot a histogram spolu (./.)

Príklad: **Diamond Pricing the C's of Diamond Stones.** Cross-section data from 2000 made up of 308 observations on the price of diamonds in Singapore outlets. Price is in Singapore \\$. (The C's are four: Carat, Clarity, Colour and Cut.)



## Polygón početnosti (frekvenčný polygón)

Používa sa pri dátach, ktoré reprezentujú merania diskrétnej numerickej premennej s opakujúcimi sa (nie veľmi mnohými) hodnotami rozptýlenými na relatívne malom úseku na reálnej osi. V takomto prípade sa nespracovaný súbor  $x_1, x_2, \dots, x_n$  vytriedi, teda sa pre každú jedinečnú napozorovanú obmenu  $x_1, x_2, \dots, x_m$  určí počet jej výskytov  $n_1, n_2, \dots, n_m$  (čím sa stanovia bežné absolútne početnosti).

Polygón početnosti sa nakreslí do ortogonálnej sústavy súradníc tak, že na horizontálnu os sa zaznačia jednotlivé hodnoty premennej  $x_1, x_2, \dots, x_m$  a na vertikálnu os sa nanesú (absolútne) počty ich výskytu  $n_1, n_2, \dots, n_m$ . Takto vzniknuté body sa spoja úsečkami. Iný variant polygónu početnosti vznikne pri použití relatívnych početností  $f_1, f_2, \dots, f_m$ .

## Polygón početnosti (frekvenčný polygón) (./.)

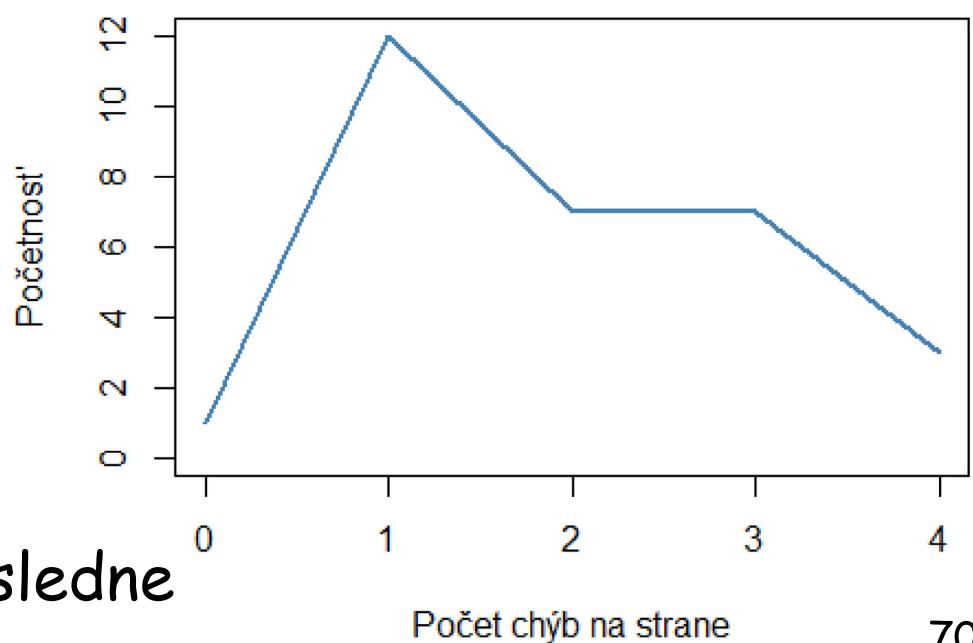
Ako príklad môže poslúžiť počet štylistických a gramatických chýb, ktoré sa vyskytli pri vydávaní posledného čísla 30-stranového časopisu na jednotlivých stranach pred jeho jazykovou korektúrou. Pôvodné hodnoty

0 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 3 4 4 4

po vytriedení sú ( $x_i$  - počet chýb na strane a  $n_i$  - počet výskytov  $x_i$ )

Frekvenčný polygón počtu chýb na strane  
v 30-stranovom časopise pred jazykovou korektúrou

$x_i$	$n_i$
0	1
1	12
2	7
3	7
4	3



a frekvenčný polygón je následne