Popisné statistiky konečného souboru, zde nepracuji s odhady.

Charakteristiky polohy (míry polohy, centrální tendence)

Popisují typickou hodnotu datového souboru, kde data leží na číselné ose.

Popisné statistiky konečného souboru, zde nepracuji s odhady.

Charakteristiky polohy (míry polohy, centrální tendence)

Popisují typickou hodnotu datového souboru, kde data leží na číselné ose.

1/ Minimum a maximum [minimum and maximum] (min, max)

= nejmenší a největší hodnota souboru.

x_{min}

x_{max}

- kvantitativní data (intervalová a poměrová stupnice) a ordinální stupnice
- pro nominální data nemá smysl (je menší červená nebo modrá barva?)

- Značení ve smyslu uspořádaných hodnot:

$x_{min} = x_{(1)}$ $x_{max} = x_{(N)}$, případně $x_{(n)}$

Příklad: výšky 12 náhodně vybraných desetiletých dívek

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
135	141	143	131	146	141	142	132	141	151	146	141

$x_{(1)} = 131 = x_{min}$ $x_{(12)} = 151 = x_{max}$

2/ Aritmetický průměr [arithmetic mean] (mean)

základní soubor: výběrový soubor:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \xrightarrow{\textbf{výběrová verze}} \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Populační průměr [population mean], výběrový průměr [sample mean]

- Jen kvantitativní data (intervalová a poměrová stupnice)

Poznámka: **μ** – čti [mí], označuje skutečný parametr základního (∞) souboru; **μ** většinou nazýváme *střední hodnota* (bude později), ale může označovat i populační průměr.

Příklad: výšky 12 náhodně vybraných desetiletých dívek

$$\bar{x} = \frac{1}{12}(135 + 141 + 143 + 131 + 146 + 141 + 151 + 132 + 141 + 142 + 146 + 141)$$

$$= \mathbf{140,83}$$

3/ Modus [mode] (*mode* v R má jiný význam)

= nejčastěji se vyskytující hodnota

- všechny typy dat
- označení $\hat{x}$, ale i jinak

Příklad: výšky 12 náhodně vybraných desetiletých dívek – uspořádané:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$
131	132	135	141	141	141	141	142	143	146	146	151

$$\hat{x} = \mathbf{141}$$

Poznámka: mohou být dvě (a více) stejně „nejpočetnějších“ hodnot či kategorií.

Poznámka: unimodální a bimodální rozdělení má souvislost právě s počtem modů v (teoretických) datech.

4/ Medián [median] (median)

= označuje „prostřední“ hodnotu, tedy hodnotu v polovině uspořádaného souboru: polovina všech hodnot je menší než hodnota mediánu a polovina je větší než hodnota mediánu

- časté označení $\tilde{x}$
- data kvantitativní (intervalová a poměrová stupnice) a data uspořádaná (ordinální stupnice)

Lichý počet hodnot: $\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$

5 hodnot => prostřední je 3. hodnota
(5+1)/2 = 3

Sudý počet hodnot: $\tilde{x} = \frac{1}{2}\left\{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right\}$

Příklad: výšky 12 náhodně vybraných desetiletých dívek – uspořádané:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$
131	132	135	141	141	141	141	142	143	146	146	151

$\tilde{x} = 141$

ad 4/ Kvartil, kvantil [quartile, quantile] (quantile)

Medián ~ padesátiprocentní kvantil, Q_2

prostřední hodnota, dělí soubor na 50 % – 50 %

Můžeme se ptát také na ***čtvrtiny***. Označujeme je
dolní kvartil, Q_1, 25% kvantil: dělí soubor na 25 % – 75 %
horní kvartil, Q_3, 75% kvantil: dělí soubor na 75 % – 25 %

Obecně např. **30% kvantil:** dělí soubor na 30 % – 70 %
atd.

Poznámka: výpočty kvantilů se mohou v různých softwarech lišit.

Příklad

výšky 12 náhodně vybraných desetiletých dívek – uspořádané:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$
131	132	135	141	141	141	141	142	143	146	146	151

Někdy je užitečné popsat soubor takto uspořádanými charakteristikami:

minimum	131
první kvartil	138
medián	141
průměr	140,83
třetí kvartil	144,5
maximum	151

Můžeme takto popsat i více souborů, čtenář pak porovnává hodnoty mezi soubory.

A další, například:

5/ Geometrický průměr [geometric mean]

$$GM = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 x_2 x_3 \cdots x_n}$$

Čti příklad v *M.J.Crawley*, str.28.
Pojem „centrální tendence“.

- Procesy, kde se hodnoty mění spíše násobně než aditivně, př. 10-1 – 1000 – 10 – 1.
- Data na poměrové stupnici, nesmí obsahovat nulu.

5/ Harmonický průměr [harmonic mean]

$$HM = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

Čti příklad v *M.J.Crawley*, str.30,
slon a průměrná rychlost.

- Data na poměrové stupnici, nesmí obsahovat nulu.
- Například průměr z několika rychlostí.

Charakteristiky rozptylu, variability

- Snaží se popsat rozptýlenost, proměnlivost souboru, „kolik prostoru“ na číselné ose hodnoty zabírají

1/ Rozsah, rozpětí [range] (range)

= rozdíl mezi největší a nejmenší hodnotou souboru

$$\boldsymbol{rozsah} = \boldsymbol{x_{max}} - \boldsymbol{x_{min}}$$

Příklad dívky: $\text{rozsah} = 151 - 131 = 20$

- Data na intervalové a poměrové stupnici
- Charakteristika je ovlivněna netypickými (odlehlými, extrémními) hodnotami, proto se používá zřídka
- ! Odhad rozsahu hodnot v celé populaci na základě výběru: se zvětšováním výběru většinou roste také rozsah, proto se rozsah hodnot celé populace (základního souboru) nedá dobře odhadnout jen z výběrového souboru!
- Lépe bude fungovat následující charakteristika:

2/ Mezikvartilové rozpětí [interquartile range] (IQR)

= vyjadřuje šířku intervalu, ve kterém leží „prostřední“ polovina hodnot

$$\boldsymbol{interkvart.rozpeti = Q_3 - Q_1}$$

- Data na intervalové a poměrové stupnici
- Charakteristika není tolik ovlivněna odlehlými hodnotami

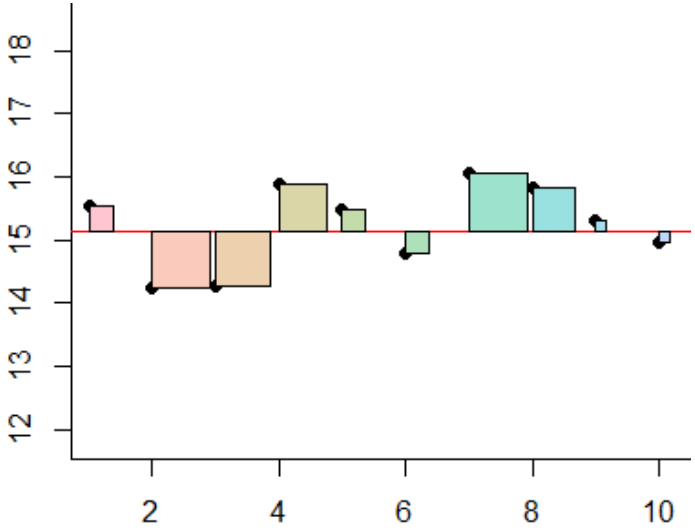
Náš příklad: $Q_3 - Q_1 =$

$$= 144{,}5 - 138 = 6{,}5$$

3/ Rozptyl [variance] (var)

= popisuje, jak jsou hodnoty „rozptýleny“ kolem průměru

$$s_X^2 = VAR(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$



- Nejužívanější charakteristika
- Pro kvan**tita**tivní data.
- Dále bude Entropie pro kvalitativní data.
- Definována jako (téměř) průměrná plocha čtverce odchylky od průměru
- Ve starší literatuře může být jiný vzoreček: $\frac{\sum(x_i-\bar{x})^2}{n}$
- První verze vzorečku má lepší vlastnosti (bude později)
- Další označení: **populační rozptyl = σ^2**. Takto označujeme skutečný parametr základního souboru, který většinou neznáme. Výše uvedeným vzorcem počítáme jeho odhad a označujeme s^2.

4/ Směrodatná odchylka [standard deviation] (sd)

- Odmocnina rozptylu => délka strany průměrného čtverce odchylky.
- Má stejný fyzikální rozměr, jako naměřené hodnoty: Rozptyl má jiný fyzikální rozměr, je totiž umocněn na druhou.

$$s_X = SD(X) = \sqrt{s^2_X}$$

Příklad dívky: $s = 5,8$

5/ Variační koeficient [coefficient of variation]

- Poměr směrodatné odchylky a průměru

$$CV_X = \frac{s_X}{\bar{x}}$$

- (fyzikálně) bezrozměrná hodnota
- Pro data na poměrové stupnici
- Používá se k porovnání variability souborů s nestejnými průměry

Příklad: $CV = \frac{*}{*} = 0,041$

6/ Entropie [entropy] (ekologické indexy v balíku *vegan*)

- neuspořádanost
- Popisuje „rozptyl“ dat s nominálním a ordinálním měřítkem

$$H = -\sum_{j=1}^{m} \frac{n_j}{n} \cdot \ln\left(\frac{n_j}{n}\right)$$

Ln je přirozený logaritmus (o základu ***e***)
nomin. a ordin. data třídíme do kategorií
m počet kategorií, n_j počet hodnot v j-té kategorii
n = počet všech hodnot v souboru

- Entropie je nulová, je-li $n_1 = n$, tedy všechny hodnoty jsou stejné.
- Velké hodnoty entropie dostaneme, máme-li hodně různých kategorií, tedy velké m.
- Pro dané m dosáhne entropie maximální možné hodnoty v případě, že jsou všechny četnosti $n_1, n_2, \cdots, n_m$ stejné.

- Další charakteristiky: Shannonova entropie, Simpsonův index. (Hledejte kapitolu Náhodná veličina.)

7/ Z-skóry [z-score]

= normované hodnoty, tj. upravené (transformované) tak, že potom celý soubor z-skórů má dohromady průměr = 0 a rozptyl = 1 (nulový průměr a jednotkový rozptyl).

$$z_i = \frac{x_i - \overline{x}}{s_X}$$

- Použití při dalších vzorcích a postupech; jen kvantitativní data.

Příklad

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$
131	132	135	141	141	141	141	142	143	146	146	151

$$\frac{x_i - \overline{140,83}}{5,8}$$

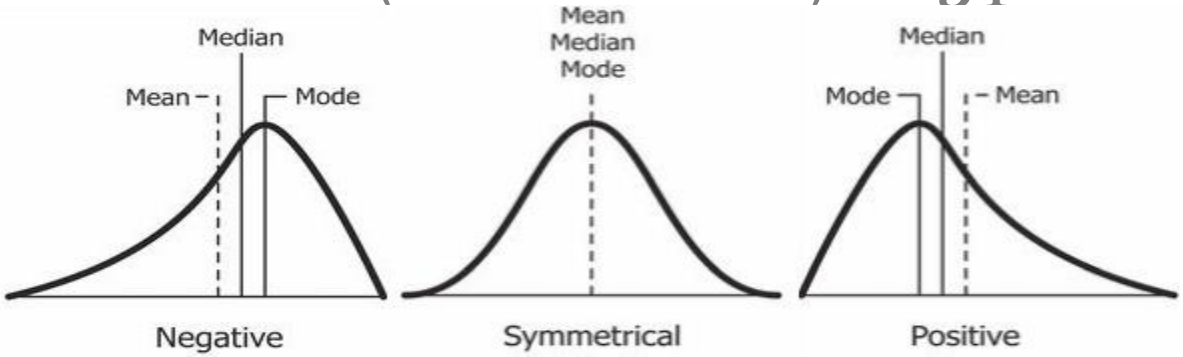
$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$
-1,7	-1,5	-1,0	0,03	0,03	0,03	0,03	0,20	0,37	0,89	0,89	1,75

8/ Šikmost [skewness]

= vyjadřuje symetrii rozložení hodnot kolem průměrné hodnoty

$$g_1 = \frac{1}{n}\sum_{i=1}^{n}(z_i)^3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)^3$$

- Je to průměr ze 3. mocnin normovaných hodnot
- Bezrozměrná charakteristika
- Histogram zešikmený doprava má kladnou g_1, tj. $g_1 > 0$ [positively skewed, right skewed]
- Histogram zešikmený doleva má negativní g_1, tj. $g_1 < 0$ [negatively skewed, left skewed]
- Symetrické rozdělení (Gaussova křivka) má g_1 blízké nule

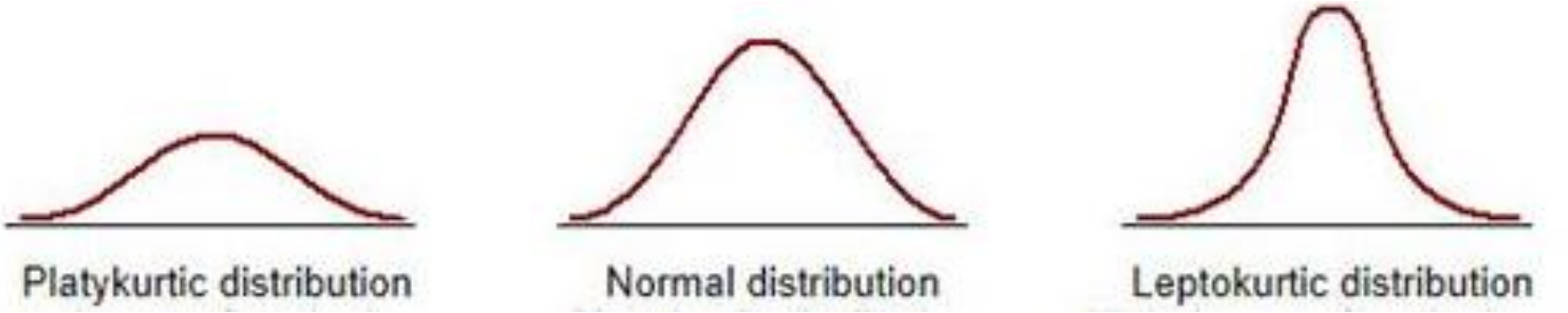


9/ Špičatost [kurtosis]

- Interpretace nesnadná

$$g_2 = \frac{1}{n}\sum_{i=1}^{n}(z_i)^4 - 3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)^4 - 3$$

- Upravený průměr ze 4. mocnin normovaných hodnot
- Bezrozměrná charakteristika
- Špičatý tvar: $g_2 > 0$ [leptokurtic], všechny hodnoty blízko průměru
- Plochý tvar: $g_2 < 0$ [platykurtic], mnohé hodnoty daleko od prům.
- Gaussova křivka (normální rozdělení) má $g_2 \approx 0$ [mesokurtic]



Terminologická vsuvka:

10/ Centrální momenty [central moments]

$$\kappa_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^k$$ … k-tý centrální moment

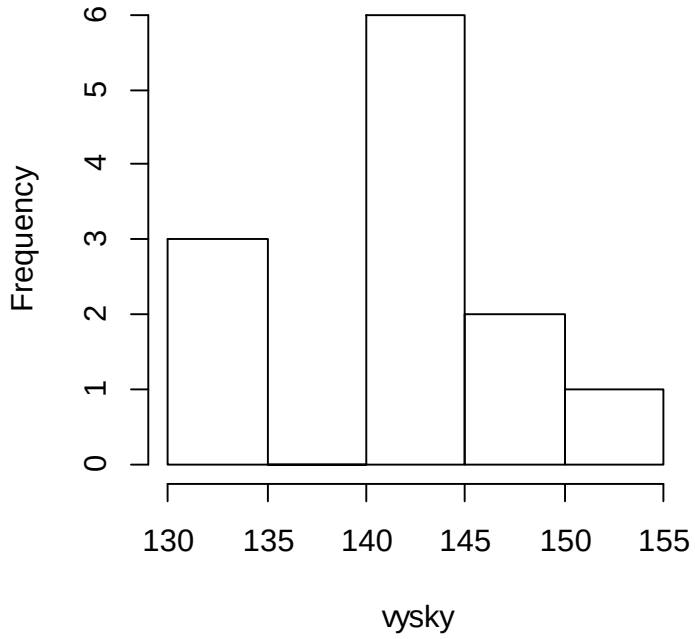
μ je střední hodnota ~ populační průměr

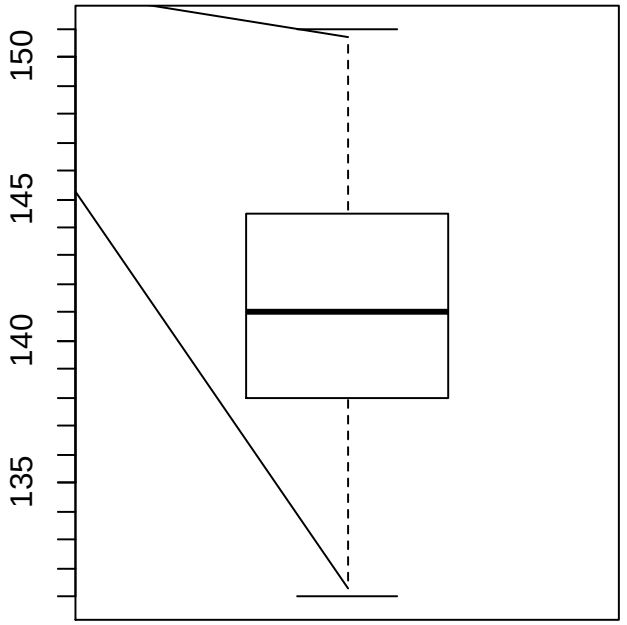
- $\kappa_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)$ … skoro průměr
- $\kappa_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ … skoro rozptyl
- $\kappa_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3$ … skoro šikmost
- $\kappa_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4$ … skoro špičatost

- Další teorie např. na wikipedii

Grafické shrnutí datového souboru

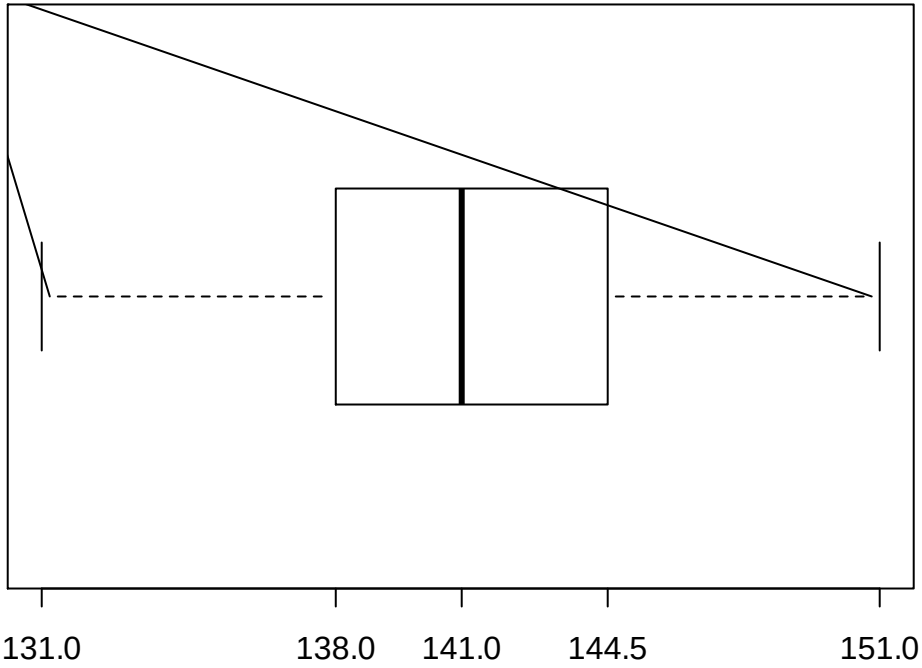
- dobrý graf řekne o datech více než čísla sumární charakteristiky
- EDA = exploratory data analysis = moderní odnož popisné statistiky, znázorňuje předchozí charakteristiky graficky
- ! V různých softwarech jsou odchylky ve výpočtech. Potom stejně vypadající graf může reprezentovat jiné charakteristiky. Proto vždy čtěte komentáře ve zvoleném softwaru.





Krabicový diagram [box-and-whisker plot] (boxplot)

Příklad:



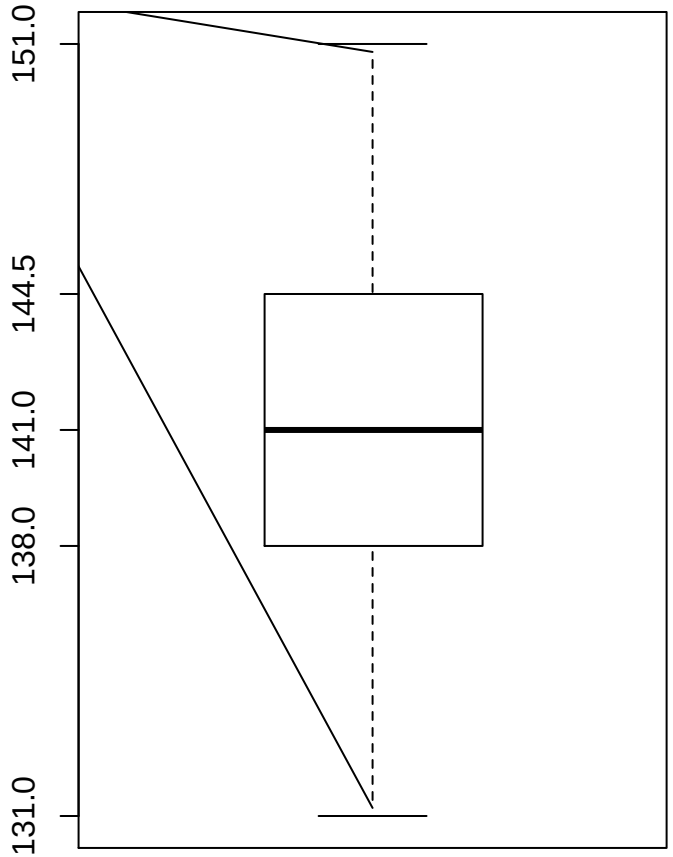
minimum	131
první kvartil	138
medián	141
průměr	140,83
třetí kvartil	144,5
maximum	151

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$
131	132	135	141	141	141	141	142	143	146	146	151

Krabicový diagram

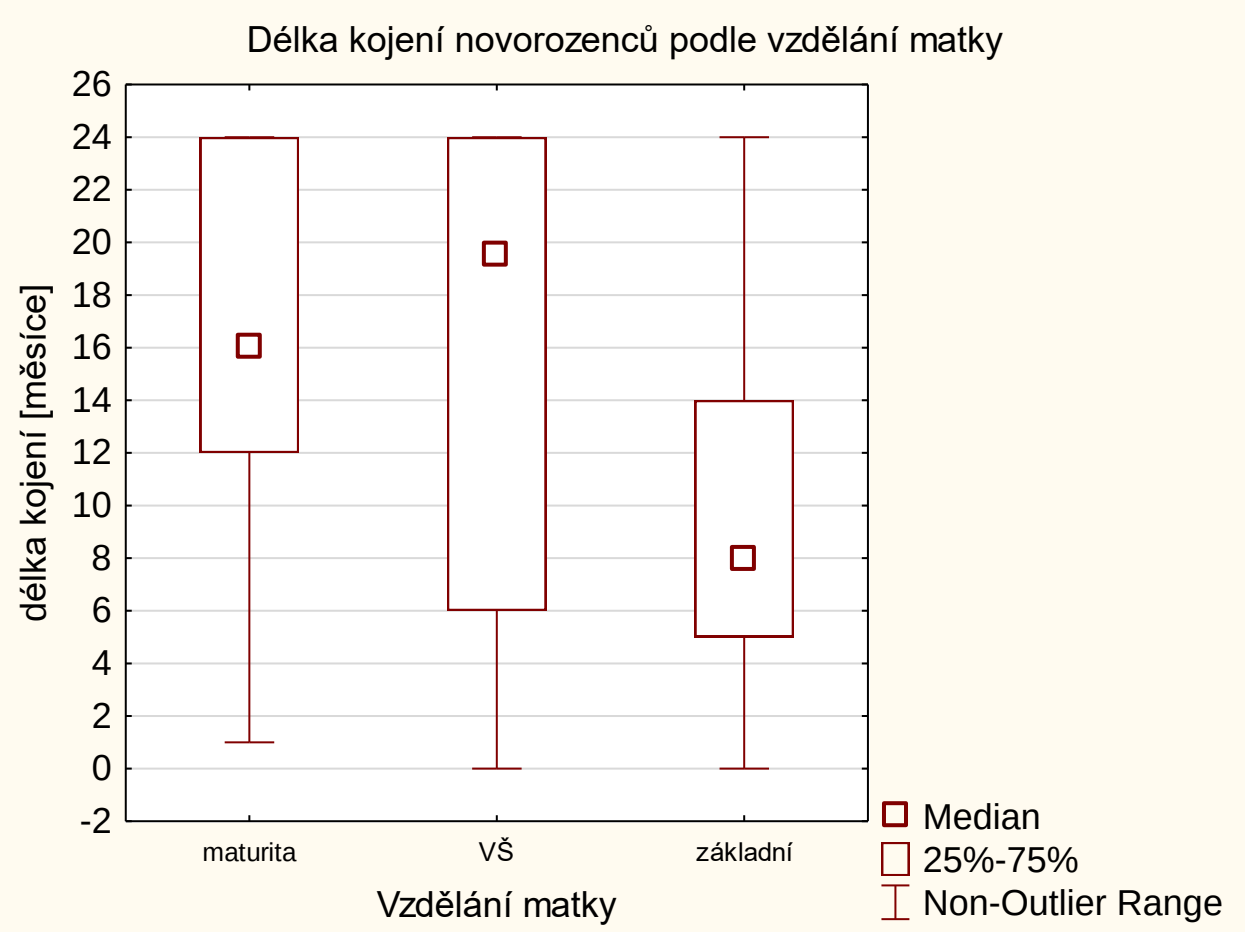
- Nevyčtu počet hodnot (pozorování), ale mohu si udělat představu o symetričnosti rozložení dat kolem mediánu.
- Někdy je možné měnit šířku krabice podle počtu hodnot (R soft.). To má smysl, když porovnáváme několik souborů s různým počtem pozorování.
- STATISTICA má základně nastaveno, že se zobrazuje aritmetický průměr a $\pm$ směrodatná odchylka. To je vhodné pro data se symetrickým rozložením hodnot (např. Gaussova křivka).

- Vždy uvádějte v popisu grafu, které charakteristiky jsou zobrazeny!



Krabicový diagram

Několik výběrů

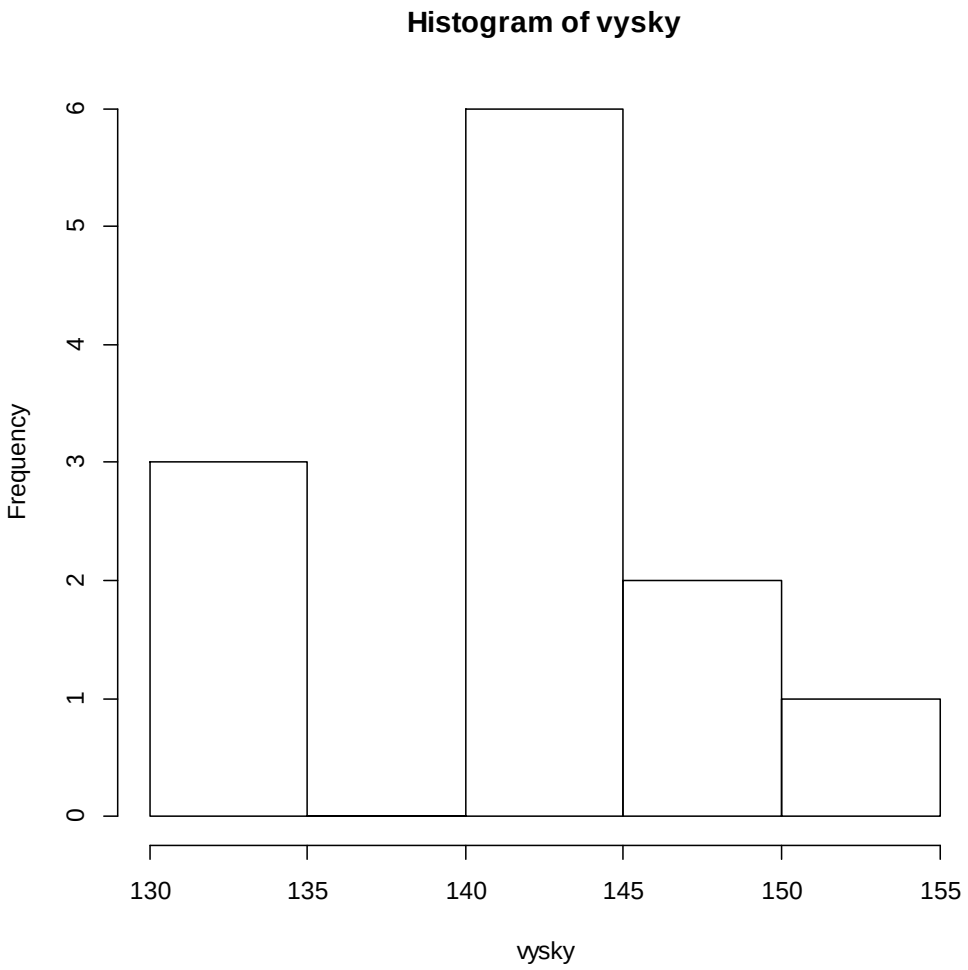


Histogram četností [frequency histogram] (hist)

- Histogram je tabulka četností převedená do grafické podoby.
- Četnost [frequency] = kolikrát se ta která hodnota vyskytuje.
- Kvantitativní data => intervaly
- (Kvalitativní data => kategorie, pro které se ale lépe hodí sloupcový graf – vizte dále.)
- Každý interval může být reprezentován jednou „typickou“ hodnotou, označme ji x_j^*, a k ní přiřadíme počet hodnot, které do intervalu patří:

	x_1^*	x_2^*	x_3^*	x_4^*	x_5^*
x_j^*	132,5	137,5	142,5	147,5	152,5
n_j	3	0	6	2	1

- Toto je tabulka četností.
- $(130,135\rangle$ - kam patří hraniční hodnoty
- Stejná šířka intervalů.
- Změnou šířky intervalů měním i tvar histogramu.



Relativní četnost [relative frequency] = převádí (absolutní) četnost do rozmezí 0 až 1, případně 0 až 100 jako procenta. Takto:

$$(n_j^*) = \frac{n_j}{n}$$ … tedy jakou část z celkového počtu hodnot tvoří hodnoty v kategorii /intervalu ***j***

x_j^*	132,5	137,5	142,5	147,5	152,5	← typické hodnoty
četnost	3	0	6	2	1	součet = 12
relativní četnost	$\frac{3}{12} = 0{,}25$	$\frac{0}{12} = 0$	$\frac{6}{12} = 0{,}5$	$\frac{2}{12} = 0{,}17$	$\frac{1}{12} = 0{,}08$	součet = 1

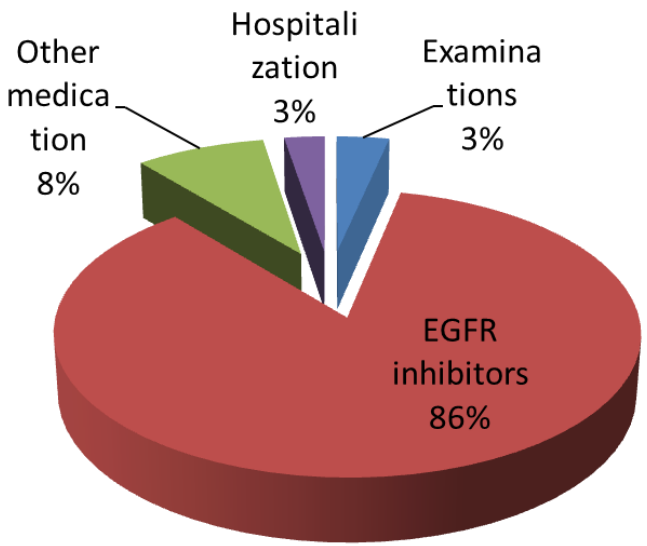
- Kontrola: součet všech relativních četností je roven 1.

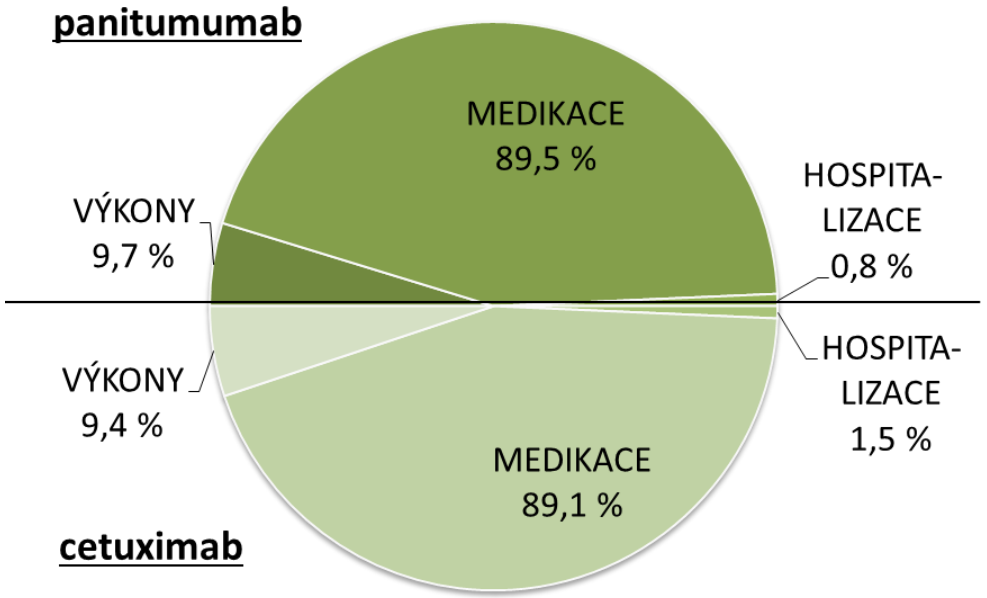
$$\sum_{j=1}^{m} n_j^* = 0{,}25 + 0 + 0{,}5 + 0{,}17 + 0{,}08 = 1$$

- Vyjádření jako procenta: $0{,}25 \rightarrow 25\ \%$
- Součet je potom = 100 %
- Histogram z relativních četností má stejný tvar, změní se měřítko.

Výsečový diagram [pie chart] (pie(relativní četnosti))

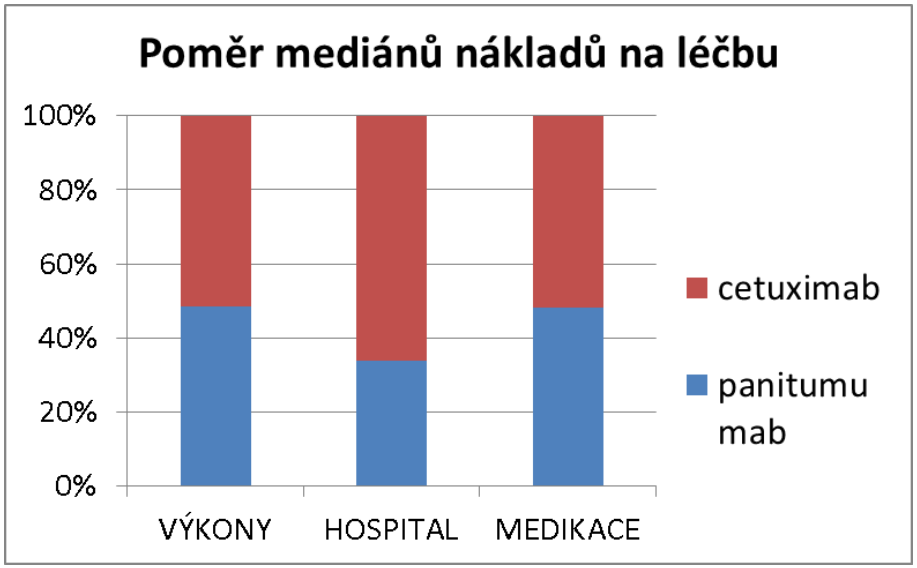
Také koláčový graf

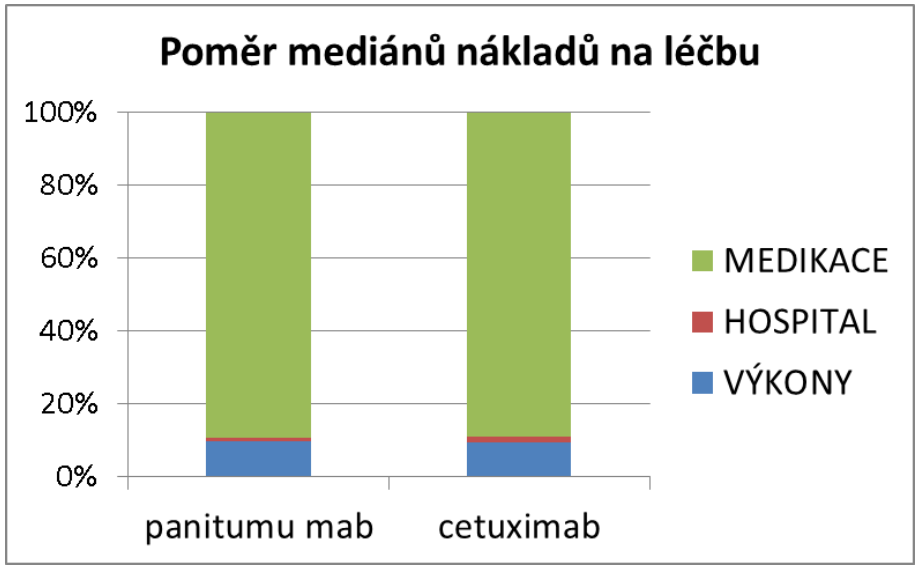




- Vhodný pro kvalitativní data (nominál. a ordinál. stupnice).
- Konstruovaný z relativních četností, software si četnosti většinou počítá sám.
- Není důležité měřítko (jednotky), vynikne jenom poměr velikosti kategorií.
- **Zkušení nedoporučují, z grafu není VIDĚT ZŘETELNĚ informace o množství. Naše oko je dobré v porovnávání LINEÁRNÍCH VZDÁLENOSTÍ, ale rozdíl v ploše porovnává špatně. DOPORUČENÝ je SLOUPCOVÝ DIAGRAM.**

Sloupcový diagram [bar chart] (barplot)





- Všechny typy dat, ale kvalitativní data musím zadat jako četnosti v kategoriích.
- Zakreslí dané hodnoty jako sloupec o odpovídající výšce.
- Chci-li data zobrazit jako relativní čísla, musím do R-příkazu zadat výsledné relativní četnosti (vs. Excel přepočítá sám).