Základy statistiky pro biology

Kontakt: Kateřina Kintrová

kintrova@sci.muni.cz

místnost: 242 (koridor), konzultace pondělí + úterý od 10:30

Osnova přednášky

1. Základní popis datového souboru;
2. Náhodná veličina;
3. Některá rozdělení pravděpodobností;
4. Odhady výběrových parametrů, statistické hypotézy;
5. Jednovýběrové testy, parametrické a neparametrické metody;
6. Testování předpokladů, dvouvýběrové testy;
7. Chi-kvadrat testy, kontingenční tabulky;
8. Několik výběrů, ANOVA;
9. Lineární regrese a korelace;
10. Poznámky k designu experimentů a pozorování.

Literatura

Zvára Karel: Základy statistiky v prostředí R. Edice: Biomedicínská statistika IV. Karolinum, 2013.

Lepš Jan a Šmilauer Petr: Biostatistika.
Episteme, nakladatelství Jihočeské univerzity v Českých Budějovicích, 2016.

Zvára Karel: Biostatistika. Karolinum, 2004 (vyprodaná, jen knihovny).

Pavlík Tomáš a Dušek Ladislav: Biostatistika. IBA MU, 2012.

https://www.iba.muni.cz/res/file/ucebnice/pavlik-biostatistika-v2.pdf

Zar, J.H.: Biostatistical analysis. Prentice Hall, London, několik vydání.

Sokal R. R. a Rohlf F. J.: Biometry (The principles and practice of statistics in biological research). W. H. Freeman, několik vydání.

Typy biologických dat - příklady

Botanik studuje 3 typy společenstev na škále vlhkosti. Pro každý snímek zapisuje:

- datum snímkování
- úroveň vlhkosti
- typ společenstva
- počet druhů
- pokryvnost
- hmotnost biomasy

Michal Hájek, www.botanickafotogalerie.cz

Typy biologických dat

Zoolog sleduje populaci hraboše během roku. Ke každému chycenému jedinci zapíše:

- datum
- teplota vzduchu
- pohlaví
- mládě - dospělec
- váha
- délka
- zdravotní stav

Data na NOMINÁLNÍ stupnici, KATEGORIE

(data na jmenovitá škále, měřítku)
[nominal scale, categorical data, categorial data, factors -> levels of factor]

příklad: BARVA OČÍ -> úrovně: černá, hnědá, modrá;
pohlaví, druh, očkování ano/ne, kosení ano/ne

- jsou to vlastnosti, kvalitativní data
- vyjadřujeme většinou slovně, ale můžeme kódovat čísly
- úrovně vlastnosti nelze seřadit ve smyslu větší – menší, nelze tedy ani počítat rozdíly mezi úrovněmi
- 0 – 1 kódování = binární stupnice (typicky ano/ne)
- !! Hodnoty lze kódovat čísly (černá = 1, hnědá = 2, modrá = 3), ale z těchto čísel nemůžeme počítat průměr apod. Takové číslo nemá žádnou interpretaci, vysvětlení.
- variabilitu vyjadřujeme jako entropii

Data na ORDINÁLNÍ stupnici, POŘADÍ

(data na pořadové škále, měřítku)
[ordinal scale]

příklad: VZDĚLÁNÍ -> úrovně: základní, střední, vysoké;
klasifikační stupně, zdravotní stav, stupeň znečištění, stupeň vlhkosti

- opět to jsou vlastnosti, kvalitativní data
- úrovně znaku můžeme seřadit ve smyslu větší – menší, ale přírůstek není konstantní, možná se nedá ani změřit
- data můžeme kódovat čísly, ale opět s nimi nelze přímo počítat
- variabilitu vyjadřujeme jako entropii

Data na INTERVALOVÉ stupnici

[data on the interval scale]

příklad: teplota

- hodnoty vyjadřujeme čísly, KVANTITATIVNĚ, data mají většinou fyzikální rozměr (°C, °F)
- mezi hodnotami jsou stejné vzdálenosti (konstantní přírůstek)
- nula je na dohodnutém místě (arbitrární nula) a nemusí znamenat neexistenci měřené vlastnosti, záporné hodnoty mají smysl
- ptáme se na rozdíly hodnot, protože poměry hodnot nemají smysl
- DATA NA CIRKULÁRNÍ STUPNICI [circular statistics]
  - příklad: dny roku, hodiny dne, azimut
  - zvláštní případ dat na intervalové stupnici – maximum sousedí s minimem => speciální analýzy těchto dat

Data na POMĚROVÉ stupnici

[ratio scale]

příklad: výška, váha, počet

- hodnoty vyjadřujeme čísly, KVANTITATIVNĚ, data mají většinou fyzikální rozměr
- známe jednotkové množství, máme konstantní přírůstek
- přirozená nula = absolutní neexistence měřené vlastnosti
- ptáme se nejen na rozdíl hodnot, ale také na jejich podíl, poměr (kolikrát je A větší než B)

Shrnutí typů dat:

Data
- Kvalita => nominální + ordinální stupnice
- Kvantita => intervalová + poměrová stupnice

Data SPOJITÁ versus DISKRÉTNÍ.

Typy biologických dat – příklady

Botanik studuje 3 typy společenstev na škále vlhkosti. Pro každý snímek zapisuje:

- datum snímkování = intervalová stupnice (cirkulární data)
- úroveň vlhkosti = ordinální (pořadí)
- typ společenstva = nominální (kategorie)
- počet druhů = poměrová
- pokryvnost = poměrová
- hmotnost biomasy = poměrová stupnice

Nominální – ordinální – intervalová – poměrová stupnice.

Příklad:

Zoolog sleduje populaci hraboše během roku. Ke každému chycenému jedinci zapíše:

- datum = intervalová stupnice
- teplota vzduchu = intervalová
- pohlaví = nominální (kategorie)
- mládě - dospělec = ordinální (mládě < dospělec)
- váha = poměrová
- délka = poměrová
- zdravotní stav = ordinální

Nominální – ordinální – intervalová – poměrová stupnice.

Poznámky k typům dat

- Charakteristiky měříme nebo odhadujeme
  (výška stromu, počet krvinek, …)
- Poznámky o přesnosti měření
  - výška stromu s přesností na 1 metr,
  - počet krvinek s přesností na 1000 krvinek
- Rozlišujeme data diskrétní a spojitá
- Jenže ne vždy to jde dodržet, jindy není třeba tak přísně rozlišovat

Statistika popisná a induktivní

Statistika popisná popisuje datový soubor zcela, pracuje s údaji o všech uvažovaných subjektech, základní soubor je konečný.
Příklad: všechny nemocnice v ČR, všichni studenti biologických oborů…

Statistika induktivní nemá údaje o všech subjektech a statistický popis odvozuje (indukuje) z vybrané skupiny subjektů.
Příklad: populace netopýrů v jeskyni, charakteristika Pcháče osetu, …

Dvojice pojmů:

základní soubor (populace) – výběrový soubor
parametr – odhad
populační průměr – výběrový průměr
pravděpodobnost – relativní četnost

Výběr ze základního souboru

Potřebujeme reprezentativní vzorek, který splňuje určitá pravidla. Potom můžeme použít statistické metody a výsledky budou mít smysl.

Pravidla: všechny subjekty mají stejnou pravděpodobnost, že budou vybrány, a výběr jednoho nezávisí na tom, který byl vybrán dříve. [iid. = independent and identically distributed]

Je tedy rozdíl mezi výběrem z konečného a nekonečného základního souboru! Prakticky: pokud je výběr velmi malou částí konečného základního souboru – do 5 %, potom lze zpravidla použít metody výběru pro nekonečný základní soubor.

Příklad (výběr „velmi malé části“):

Základní soubor má 100 jedinců/subjektů. Pravděpodobnost výběru:

1. subjekt 1/100 = 0,0**10**0
2. subjekt 1/99 = 0,0**10**1
3. subjekt 1/98 = 0,0**10**2
5. subjekt 1/96 = 0,0**10**4
6. subjekt 1/95 = 0,0**10**5
11. subjekt 1/90 = 0,0**11**1

Základní soubor má 1000 subjektů. Prst. výběru:

1. subjekt 1/1000 = 0,001**00**0
20. subjekt 1/981 = 0,001**01**9367
30. subjekt 1/971 = 0,001**02**9866
48. subjekt 1/953 = 0,001**04**9317
50. subjekt 1/951 = 0,001**05**1524

Poznámky ke způsobu výběru:

- ideálně: očíslovat všechny jedince a generátorem náhodných čísel vybrat výběrový soubor
- ! subjektivní výběr typu „jdu loukou a občas vyberu rostlinu“ není náhodný!
- Prakticky mnoho problémů – je třeba znát biologii sledovaných druhů a konzultovat se školiteli. Data sebraná špatným způsobem nelze interpretovat!!!
- Rostliny v ploše: vytvořím systém pravoúhlých souřadnic, v počítači generuji náhodné středy pokusných ploch.
- Rostliny shlukovitě vs. solitéry: pozor, NEFUNGUJE výběr jedince nejbližšího náhodnému bodu, protože solitéry mají vyšší prst. výběru a vychylují výsledek!
- Hraboši do pastí: zkušenější jedinci budou chybět.
- Netopýři v jeskyni: nedosáhnu všude…
- Celý soubor rozdělím na homogenní podsoubory a z těch vybírám
- Laboratorní zvířata = hypotetický základní soubor: reprezentují skupinu stejně starých, stejně živených, stejně … jedinců.

Slovníček

základní soubor – výběrový soubor (výběr)
(statistical) population – (random) sample

parametr – odhad
parameter – estimate, estimation

populační průměr – výběrový průměr
population mean – sample mean

populační rozptyl – výběrový rozptyl
population variability – sample variability

pravděpodobnost – relativní četnost
probability – relative frequency

Jedinec, subjekt, objekt, pozorování, měření, hodnota, …

Matematické značení:

N – počet subjektů v základním souboru, většinou ∞

n – počet subjektů ve výběrovém souboru

Výběrový soubor: $(x_1, x_2, x_3, \ldots, x_n)$, také x_i pro $i = 1, \ldots, n$

Uspořádaný seznam: $x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}$, kde $x_{(1)}$ je min. hodnota $x_{(n)}$ je max. hodnota.

Pořadí hodnot: $r_1, r_2, r_3, \ldots,$ tedy $r_1 = r_{x_1}$ je pořadí první hodnoty.

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	151	132	141	142	146	141

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

r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}	r_{12}
3	5,5	9	1	10,5	5,5	12	2	5,5	8	10,5	5,5

Matematické značení:

x_i - naměřená hodnota z výběrového souboru
X_i - označení pro teoretickou hodnotu ze základního souboru (=náhodna veličina)

α, β, γ, μ, σ – řecká písmena označují skutečné parametry
a, b, g, $\bar{x}, \tilde{x}, \hat{x}, var, S^2$ - latinka označují naše odhady parametrů

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

čti: „suma x í od 1 do n“

$$\sum_{i=1}^{3} \sum_{j=1}^{i} x_{ij} = x_{11} + x_{21} + x_{22} + x_{31} + x_{32} + x_{33}$$

$$\prod_{k=1}^{n} y_k = y_1 \cdot y_2 \cdot \cdots \cdot y_n$$

čti: „součin y ká od 1 do n“

Popisné statistiky konečného souboru, zde nepracuji s odhady.

Charakteristiky polohy (míry polohy)

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

1/ Minimum a maximum [minimum and maximum]

= nejmenší a největší hodnota souboru.

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

- kvantitativní data (intervalová a poměrová stupnice) a ordinální stupnice

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	151	132	141	142	146	141

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

$x_{min} = 131$
$x_{max} = 151$

2/ Aritmetický průměr [arithmetic mean]

konečný soubor: výběrový soubor:

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

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

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

$$= 140{,}83$$

3/ Modus [mode]

= 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

$$\hat{x} = 141$$

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]

= 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/ Populační medián [population median]

Stále označuje „prostřední“ hodnotu, ale populace může být až nekonečná, nejsme tedy schopni jedince očíslovat.

Proto je definován pomocí tzv. *kvantilové funkce* (později). Stále platí, že 50 % všech hodnot je menších a 50 % větších než medián, ale definice je vedena přes pravděpodobnost: náhodně vybraná hodnota je s pravděpodobností 50 % menší než medián a s prstí. 50 % je větší než medián.

Rozšířené názvosloví:

Medián ~ padesátiprocentní kvantil, 50% kvantil, Q_2

Můžeme se ptát také na 25% či 75% kvantily. Označujeme je

dolní kvartil = Q_1 = 25% kvartil [quartile]

horní kvartil = Q_3 = 75% kvartil [quartile]

A také **30% kvantil** [quantile].

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]

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

- Data na poměrové stupnici, nesmí obsahovat nulu.

5/ Harmonický průměr [harmonic mean]

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

- Data na poměrové stupnici, nesmí obsahovat nulu.

Charakteristiky rozptylu, variability

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

1/ Rozsah, rozpětí [range]

= rozdíl mezi největší a nejmenší hodnotou souboru

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

Náš příklad: $\text{rozsah} = 151 - 131 = 20$

- 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]

= 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]

= popisuje, jak jsou hodnoty „rozptýleny“ kolem průměru

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

- Nejužívanější charakteristika; náš příklad: $s^2 = 33,79$
- Pro kvantitativní data. Dále bude Entropie pro kvalitativní data.
- Definována jako (téměř) průměrný čtverec 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/ Entropie [entropy]

- 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.)

5/ Směrodatná odchylka [standard deviation]

- Odmocnina rozptylu
- 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: $s = 5,8$

6/ 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: $CV = \frac{*}{*} = 0,041$

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

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]
- Normální 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.
- Normální rozdělení (Gaussova křivka) 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

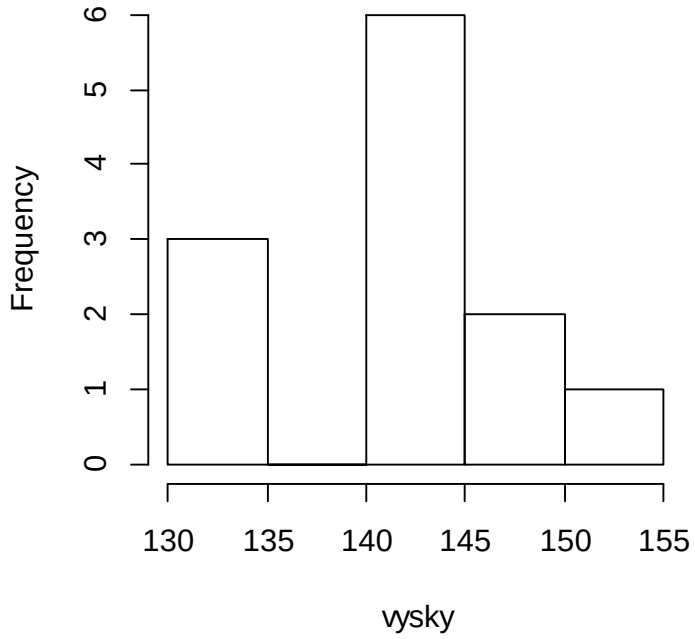
μ je střední hodnota ~ populační průměr

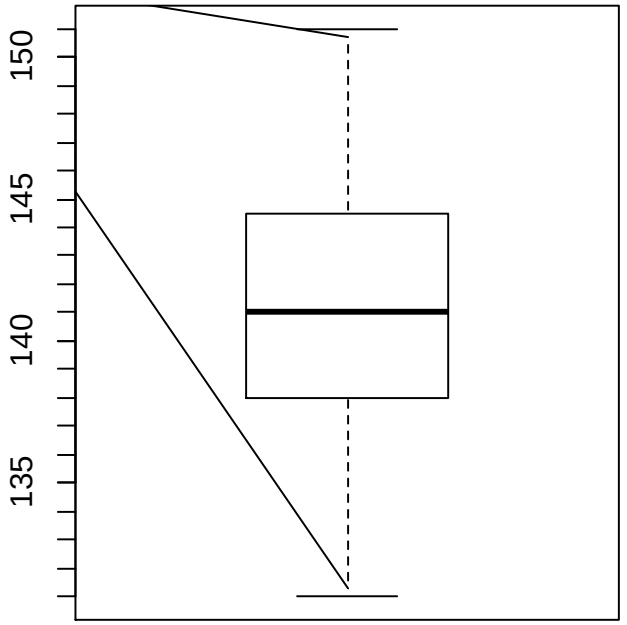
- $\kappa_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ … skoro rozptyl
- $\kappa_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3$ … skoro šikmost
- $\kappa_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4$ … skoro špičatost

- Další teorie na wikipedii

Grafické shrnutí datového souboru

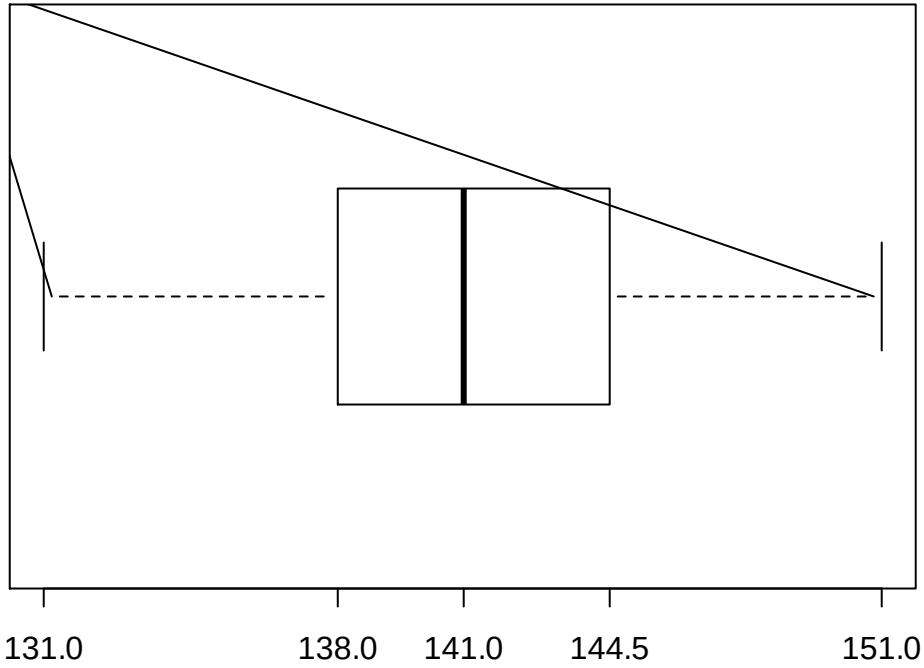
- dobrý graf řekne o datech více než čísla sumární charakteristiky
- EDA = exploratory data analysis = moderní odnož popisné statistiky, znázorňuje předchozí charakteristiky graficky
- ! V různých softwarech jsou odchylky ve výpočtech. Potom stejně vypadající graf může reprezentovat jiné charakteristiky.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





Krabicový diagram [box-and-whisker plot]

Příklad:



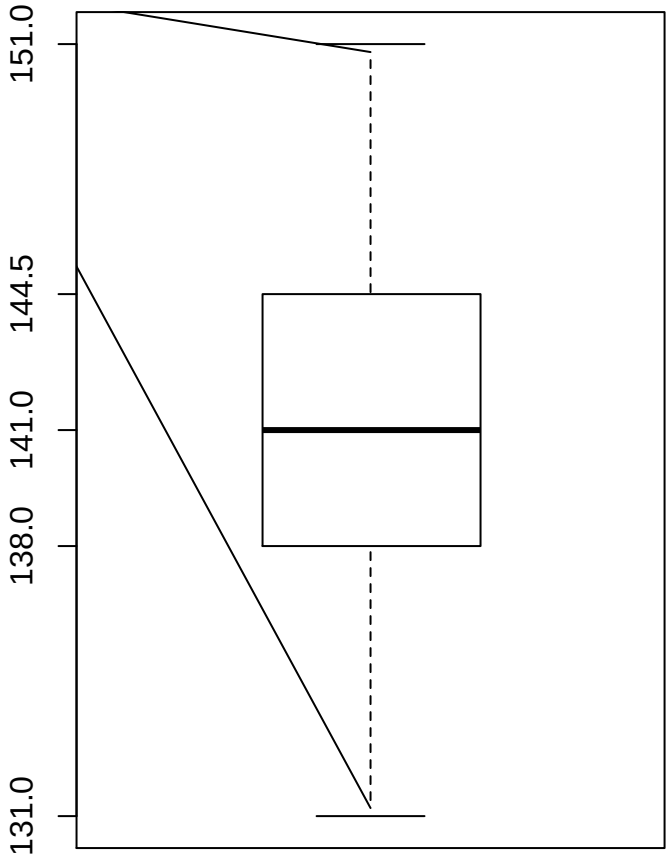
minimum	131
první kvartil	138
medián	141
průměr	140,83
třetí kvartil	144,5
maximum	151

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

Krabicový diagram

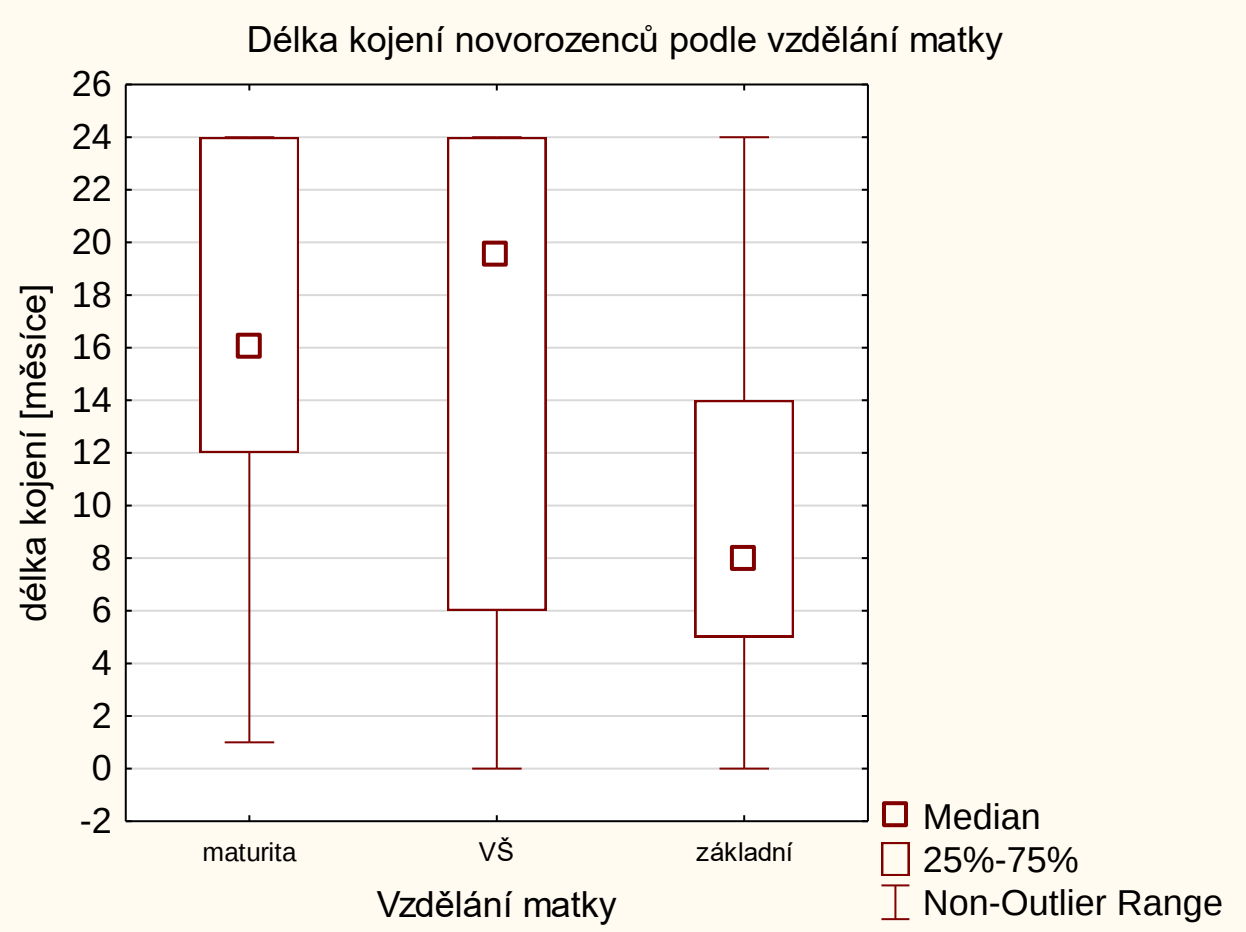
- Nevyčtu počet hodnot (pozorování), ale mohu si udělat představu o symetričnosti rozložení dat kolem mediánu.
- Někdy je možné měnit šířku krabice podle počtu hodnot (R soft.). To má smysl, když porovnáváme několik souborů s různým počtem pozorování.
- STATISTICA má základně nastaveno, že se zobrazuje aritmetický průměr a ± 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

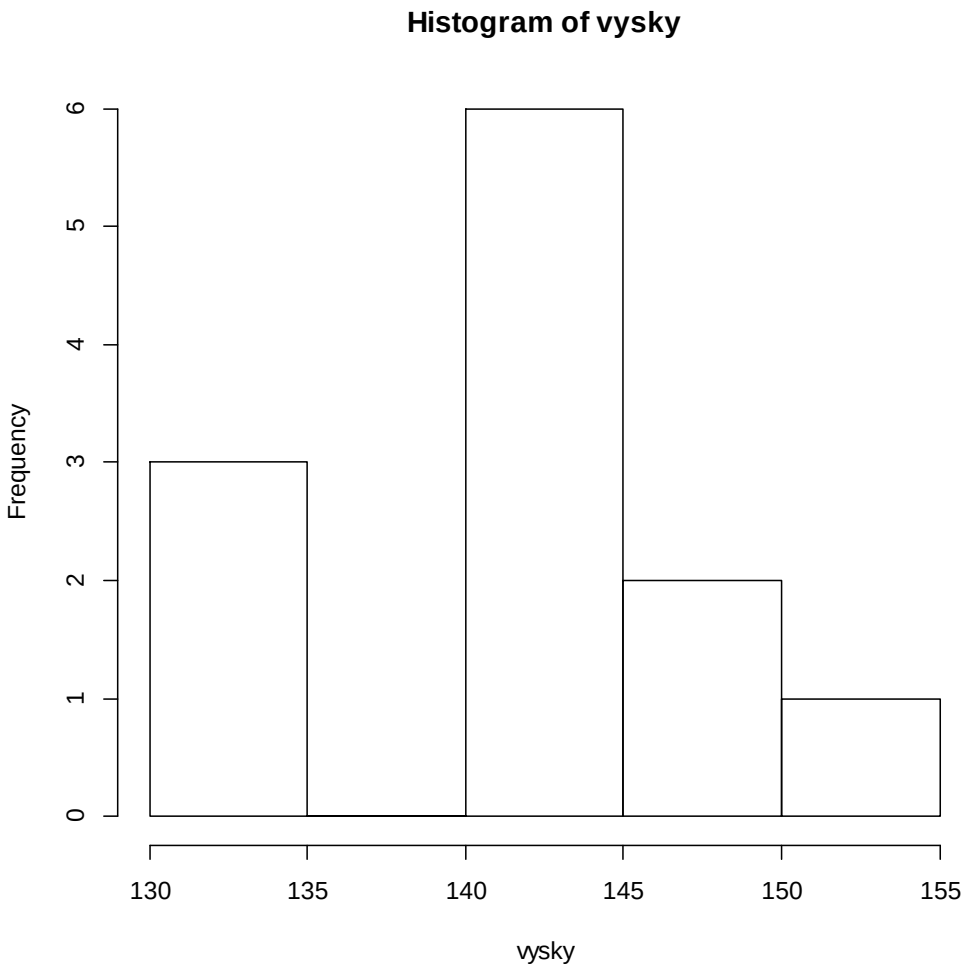


Histogram četností [frequency histogram]

- 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

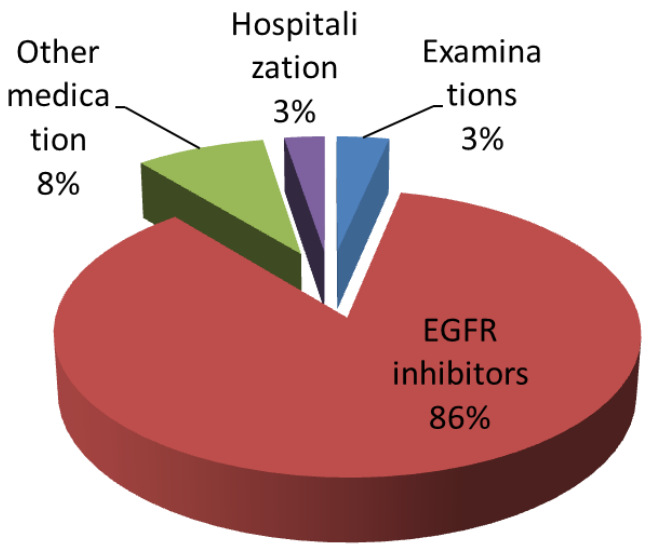
- Kontrola: součet všech relativních četností je roven 1.

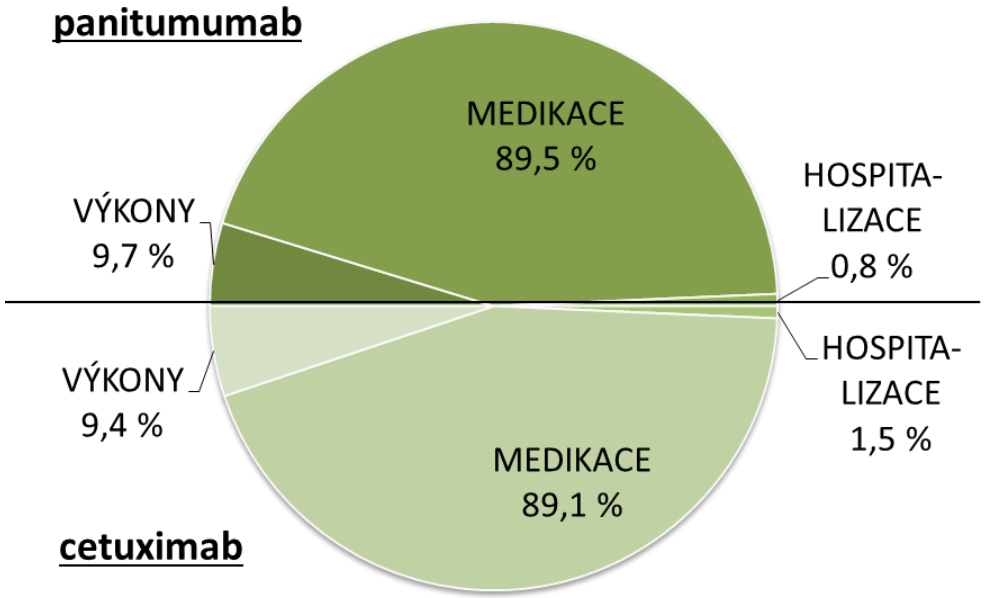
$$\sum_{j=1}^{m} n_j{}^* = 0{,}25 + 0 + 0{,}5 + 0{,}17 + 0{,}08 = 1$$

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

Výsečový diagram

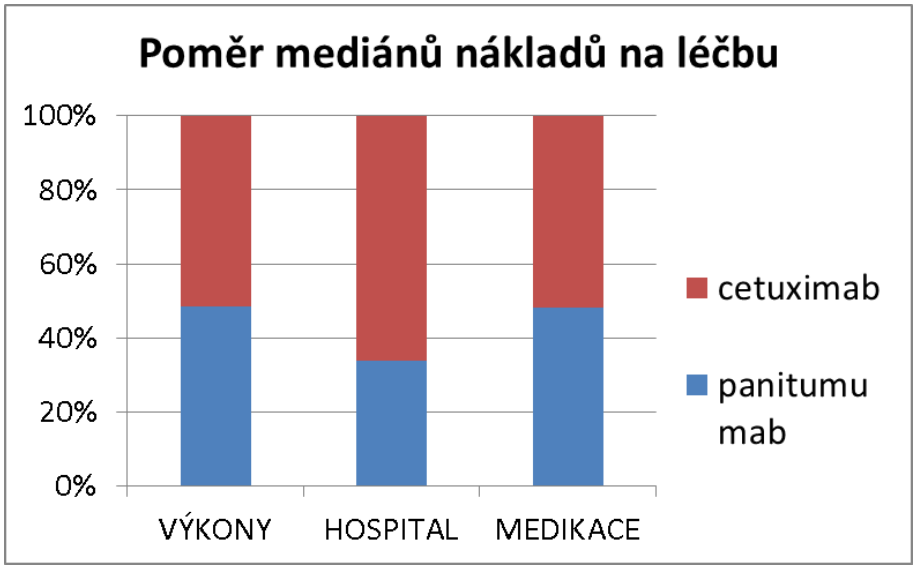
Také koláčový graf

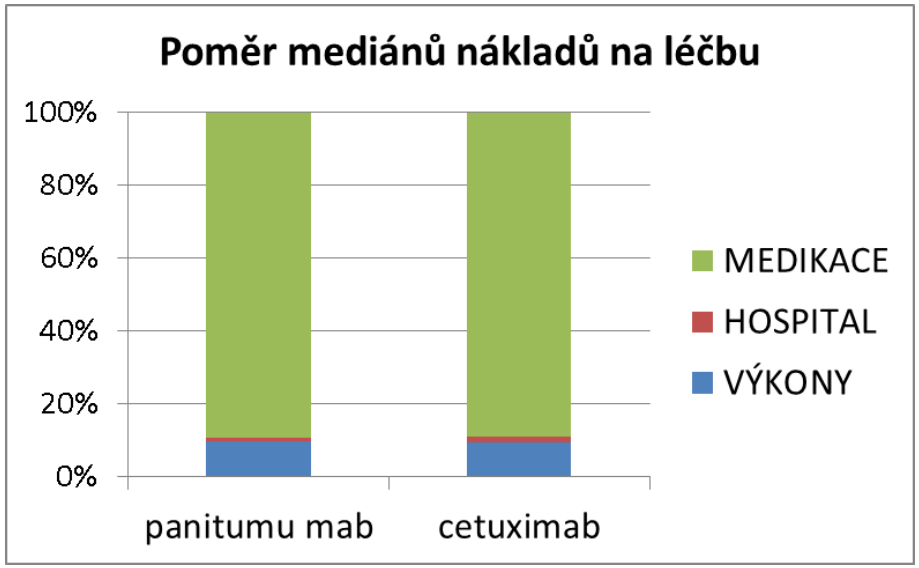




- Vhodný zejména pro kvalitativní data (nominál. a ordinál. stupnice)
- Konstruovaný z relativních četností, software si četnosti většinou počítá sám.
- Není důležité měřítko, vynikne jenom poměr velikosti kategorií.

Sloupcový diagram





- Všechny typy dat.
- Porovnává spočítané charakteristiky několika souborů (typicky mediány nebo průměry)
- Charakteristiky můžeme zobrazit jako absolutní nebo relativní čísla