Krabicový graf (Box-Plot) ve statistické analýze

Jednou z možností, jak přehledně zobrazit data ve statistické analýze je použití krabicového grafu neboli Box-Plotu

Existují různé typy krabicových grafů

Ukážeme si krabicový graf, pro jehož konstrukci jsou potřebné kvartily, průměrná hodnota, minimum a maximum

Analogicky lze vytvořit graf s kvantily

Kvantily

- Kvantil je hodnota, která slouží k popisu dat
- Obecně můžeme kvantil označit Q_p, kde p je počet procent
- Hodnota Q_p je hodnota, která odděluje p % dat od $(1 - p)$ % dat
- Kvantil, který rozděluje statistický soubor na dvě poloviny se nazývá medián
- Kvantily, které rozdělují statistický soubor na čtvrtiny se nazývají kvartily
- Kvantily, které rozdělují statistický soubor na desetiny se nazývají decily
- Dalšími speciálními kvantily jsou tercily (1/3), kvintily (1/5), percentily (1/100)
- Často používaný krabicový graf znázorňuje rozložení dat pomocí kvartilů, Q_0, Q_{25}, Q_{50}, Q_{75}, Q_{100}. Někdy se označují jako Q_0, Q_1, Q_2, Q_3, Q_4 resp. minimum (0. kvartil), 1. kvartil (dolní kvartil), 2. kvartil (medián), 3. kvartil (horní kvartil), maximum (4. kvartil)

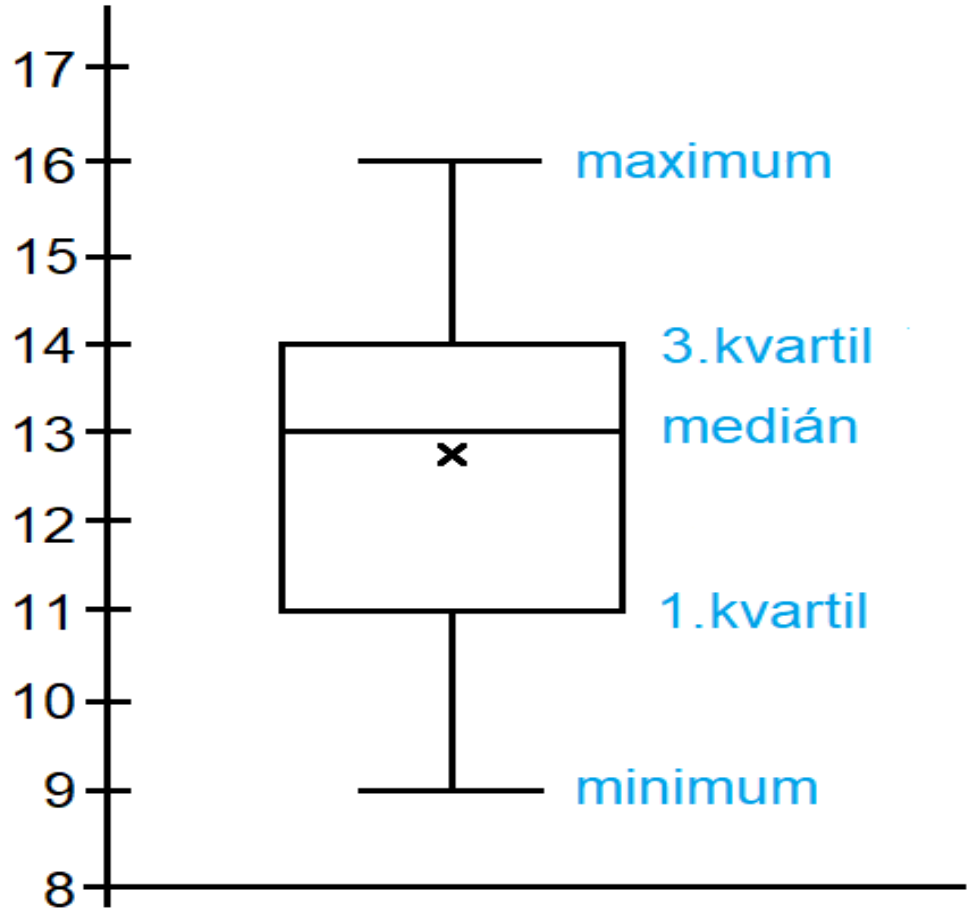
Výpočet kvartilů

Pořadí	1	2	3	4	5	6	7	8	9	10	11	12
Hodnota	9	10	11	12	12	13	13	14	14	14	15	16

- Kvartil Q_{50} (medián) je prostřední hodnota ze souboru hodnot seřazených podle velikosti
- Pokud má soubor sudý počet hodnot, je to průměr dvou prostředních hodnot
- 12 hodnot, prostřední hodnoty jsou 6. a 7., tedy $Q_{50} = 13$
- V některých případech je lepší použít medián místo průměru. Příkladem může být mzda u skupiny pracovníků, kde jednotlivci mají mimořádně velkou mzdu a většina nízkou. Pokud mzdu zprůměrujeme, může vyjít všem vysoká mzda. Na medián mimořádně velká mzda jednotlivců ve velké skupině nemá vliv
- Kvartil Q_{25} (1. kvartil) je v našem případě 3. hodnota ze souboru hodnot seřazených podle velikosti, tedy $Q_{25} = 11$. Tato hodnota odděluje 25 % dat od 75 % dat
- Kvartil Q_{75} (3. kvartil) je v našem případě 9. hodnota ze souboru hodnot seřazených podle velikosti, $Q_{75} = 14$. Tato hodnota odděluje 75 % dat od zbývajících 25 % dat
- Dalším důležitým parametrem je mezikvartilové rozpětí, rozdíl horního a dolního kvartilu $QR = Q_{75} - Q_{25}$
- V našem případě je mezikvartilové rozpětí $Q_{75} - Q_{25} = 14 - 11 = 3$. Pokud se rozhodneme uvádět medián místo průměru, je výhodné místo rozptylu udávat mezikvartilové rozpětí jako míru variability dat

KVARTILY V KRABICOVÉM GRAFU

- V krabicovém grafu je kvartilové rozpětí výškou krabice
- kvartil X_{25} vymezuje spodní hranu krabice
- kvartil X_{75} vymezuje horní hranu krabice
- V grafu jsou pomocí úseček kolmých ke hraně krabice (tzv. vousy) vyznačeny koncové body
- Koncové body se počítají podle vzorců
  $x_{25} - 1{,}5 \cdot \mathrm{QR}$ (spodní bod)
  $x_{75} + 1{,}5 \cdot \mathrm{QR}$ (horní bod)
- V některých případech se koncové body nepočítají podle uvedených vzorců, ale místo nich se vynáší minimum a maximum
- Uvnitř krabice je vodorovná čára, která vymezuje kvartil X_{50} (medián) a je vyznačen bod, který udává střední hodnotu



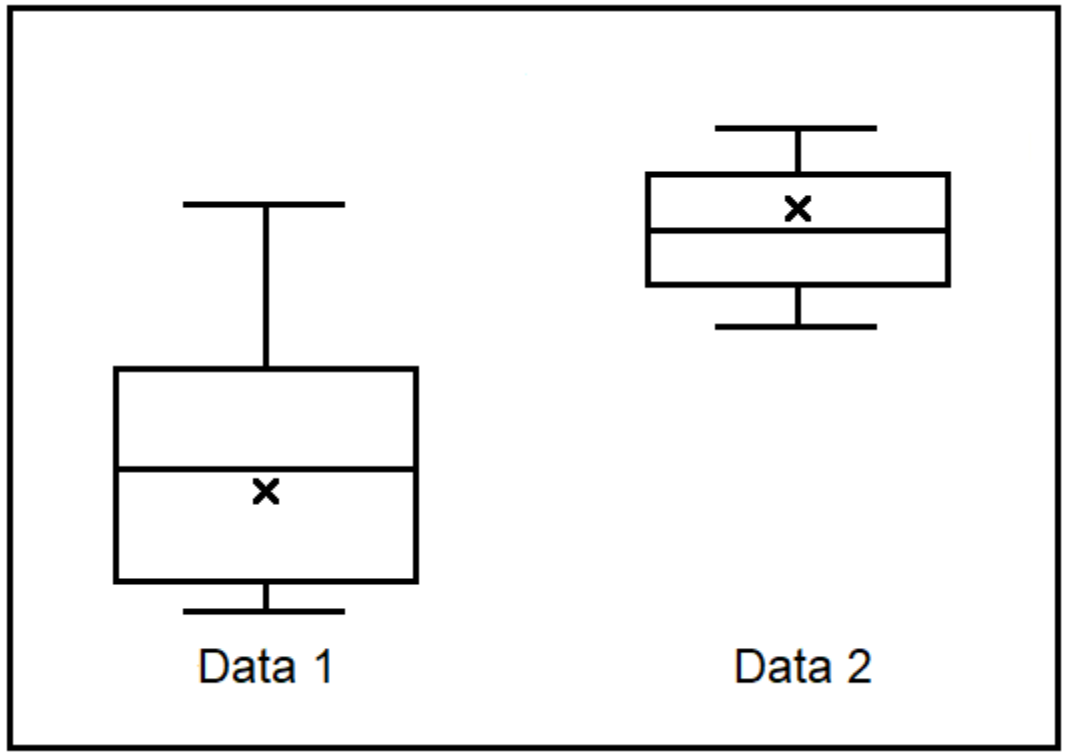
Statistika	minimum	1.kvartil	medián	průměr	3.kvartil	maximum
Hodnota	9	11	13	12,75	14	16

ČTENÍ DAT V KRABICOVÉM GRAFU

- Pokud mají data normální rozdělení neboli Gaussovo rozdělení, je čára, která označuje medián, uprostřed krabice
- Pokud je čára blízko 1. nebo 3. kvartilu, může to naznačovat, že data mají jiné než normální rozdělení
- Pokud je rozdělní symetrické, pak průměr a medián splývají, ale nemusí to platit obráceně
- Pokud průměr a medián jsou stejné, nemusí být ještě rozdělení symetrické

- V krabici se nachází 50 % hodnot dat
- Na obr. nabývá polovina dat hodnoty od 11 do 14, čtvrtina dat nabývá hodnot od 9 a je menších než 11 a čtvrtina dat nabývá hodnot větších než 14 a menších než 16
- Čím je výška krabice větší, tím větší je rozptyl hodnot, které leží mezi 1. a 3. kvartilem (polovina dat)
- Rozptyl si můžeme představit jako průměr míry vzdálenosti jednotlivých dat od průměru
- Zatímco mezikvartilové rozpětí si můžeme představit jako rozmezí možných hodnot, které nabývá prostředních 50 % dat
- Na obrázku je oblast mezi 1. kvartilem a mediánem větší než mezi mediánem a 3. kvartilem. Znamená to, že data v 3. kvartilu jsou méně rozptýlená než data v 2. kvartilu
- Můžeme si všimnout, že v třetím kvartilu nabývají pouze dvou hodnot a to 13 a 14. Průměrná hodnota leží pod mediánem. Medián je hodnota, pro kterou platí, že polovina hodnot je menších než medián a polovina hodnot je větších než medián. To znamená, že více jak polovina hodnot bude větších než průměr

Porovnání dvou souborů dat



- Data 2 nabývají data vyšších hodnot, větší je průměrná hodnota i medián
- Data 1 větší rozptyl dat je v prvním souboru
- Data 1 je minimum a maximum od sebe hodně vzdálené, to znamená, že hodnoty se budou v souboru dat hodně lišit, data 1 větší mezikvartilové rozpětí
- Data 1 maximum vysoko nad horním kvantilem, v souboru budou data, která nabývají mnohem větších hodnot než průměr
- Data 1 minimum nízko pod dolním kvartilem, to znamená, že v oblasti mezi minimem a 1. kvartilem mají hodnoty malé rozpětí
- Data 1 nabývají průměrně menší hodnoty a jsou více rozptýlená
- Data 2 nabývají data průměrně vyšší hodnoty a jsou více koncentrovaná kolem svého mediánu

KRABICOVÝ GRAF A EXCEL

- Krabicový graf je možné vytvořit pomocí Excelu
- Nalezneme ho v nabídce grafů
- Kliknutím na graf se formátuje datová řada
- Pro výpočet kvantilu se používá inkluzivní nebo exkluzivní medián
- V případě inkluzivního mediánu se do výpočtu kvantilů zahrnuje medián, v případě exkluzivního mediánu se medián do výpočtu kvartilů nezahrnuje
- V Excelu se koncové body nepočítají, v grafu se zobrazuje minimum a maximum, ale v případě, že některé hodnoty hodně vybočují, nejsou zahrnuty do výběru hodnot pro maximum a minimum a jsou zobrazeny jako osamocené body nad minimem nebo nad maximem

ZDROJE

- *Jak vytvořit krabicový graf*, 2020 [online]. [cit. 2020-27-4]. Microsoft Office. Dostupné z: https://support.office.com/cs-cz/article/jak-vytvo%c5%99it-krabicov%c3%bd-graf-62f4219f-db4b-4754-aca8-4743f6190f0d