Plant Biosystematics Group

Masaryk University
Brno, Czech Republic

PŘÍRODOVĚDECKÁ FAKULTA · MASARYKOVA UNIVERZITA · BRNO

ÚSTAV BOTANIKY A ZOOLOGIE

Bi6589

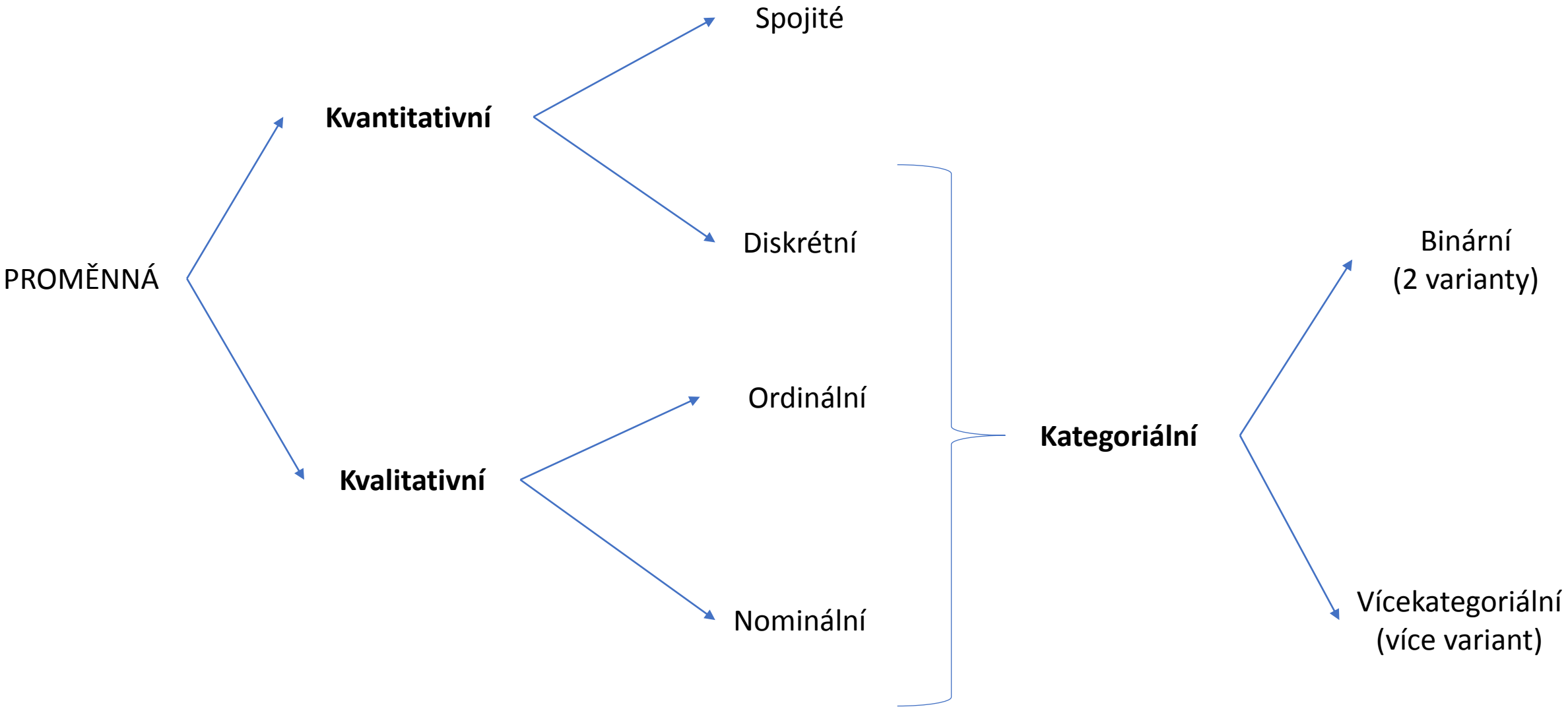
Laboratorní a bioinformatické metody rostlinné biosystematiky

Vyhodnocení a vizualizace dat

Typy proměnných

- Naměřili (pojmenovali) jsme spoustu údajů = PROMĚNNÝCH
- Proměnné potřebujeme vizualizovat a porovnat
- Pro použití správné statistické metody je klíčové správné určení typu proměnné

Typy proměnných



Kvalitativní proměnné

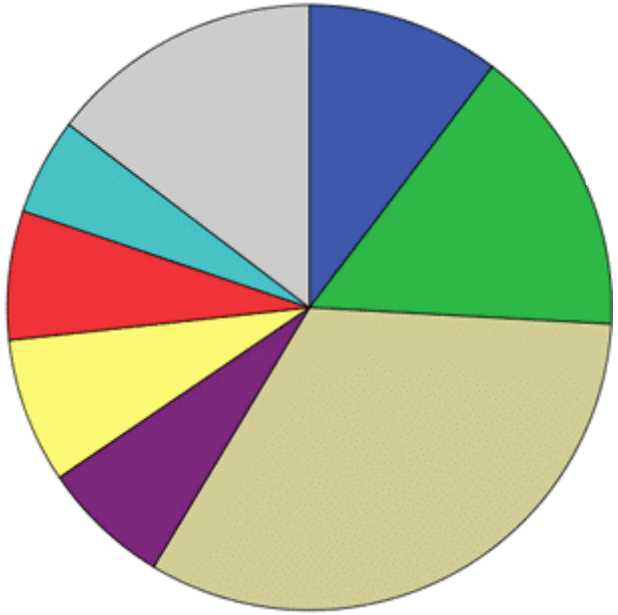
Nominální data NELZE vzájemně seřadit (NELZE rozhodnout, která hodnota je větší/menší).

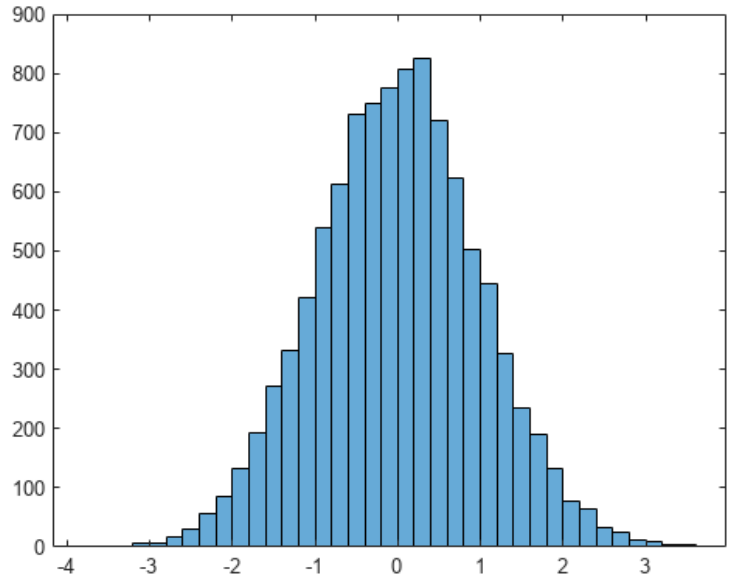
Např. **barva květu**: béžová, červená, žlutá

Ordinální data LZE vzájemně seřadit (má smysl se ptát na relaci větší/menší).

Např. **ostnitost listu**: žádná, mírná, střední, vysoká.

Pro vizualizaci dat lze využít **histogram** (sloupcový graf) nebo **koláčový graf.**





Kvantitativní proměnné

Kvantitativní data = numerické, kardinální

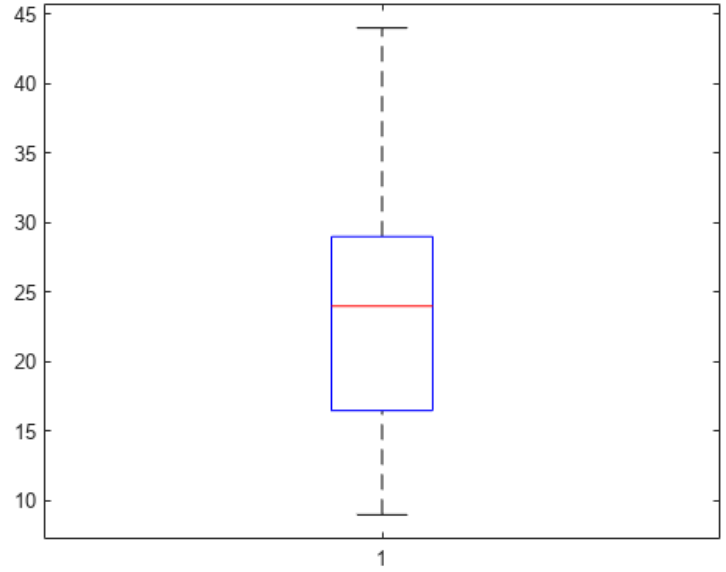
Spojitá data mohou **nabývat jakýchkoliv hodnot** v určitém rozmezí (intervalu). Lze je porovnávat.

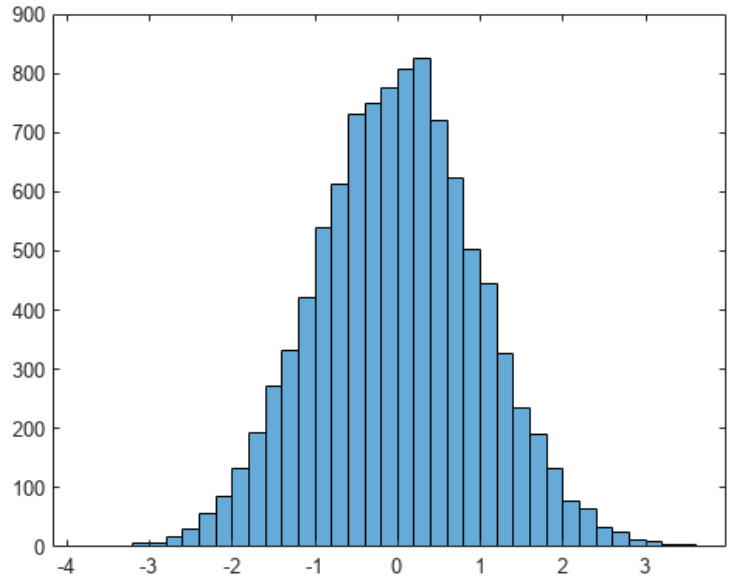
Např. výška rostlin

Diskrétní data mají **konečný počet hodnot/bodů**. Lze je porovnávat.

Např. počet semen v úboru (neexistuje 2,5 semene v úboru)

Pro vizualizaci dat lze využít **krabicový graf** nebo **histogram** (sloupcový graf).





Kategoriální proměnné

Binární (alternativní, dichotomická) data nabývají **pouze dvou hodnot**. Většinou jsou to data typu **ano/ne (1/0)**.
Např. **přítomnost prašníkových trubiček:** mají/nemají

Vícekategoriální (množná, polychotomická) data nabývají **více než dvou hodnot**.
Např. **pohlaví rostlin:** samice, hermafrodit, samec

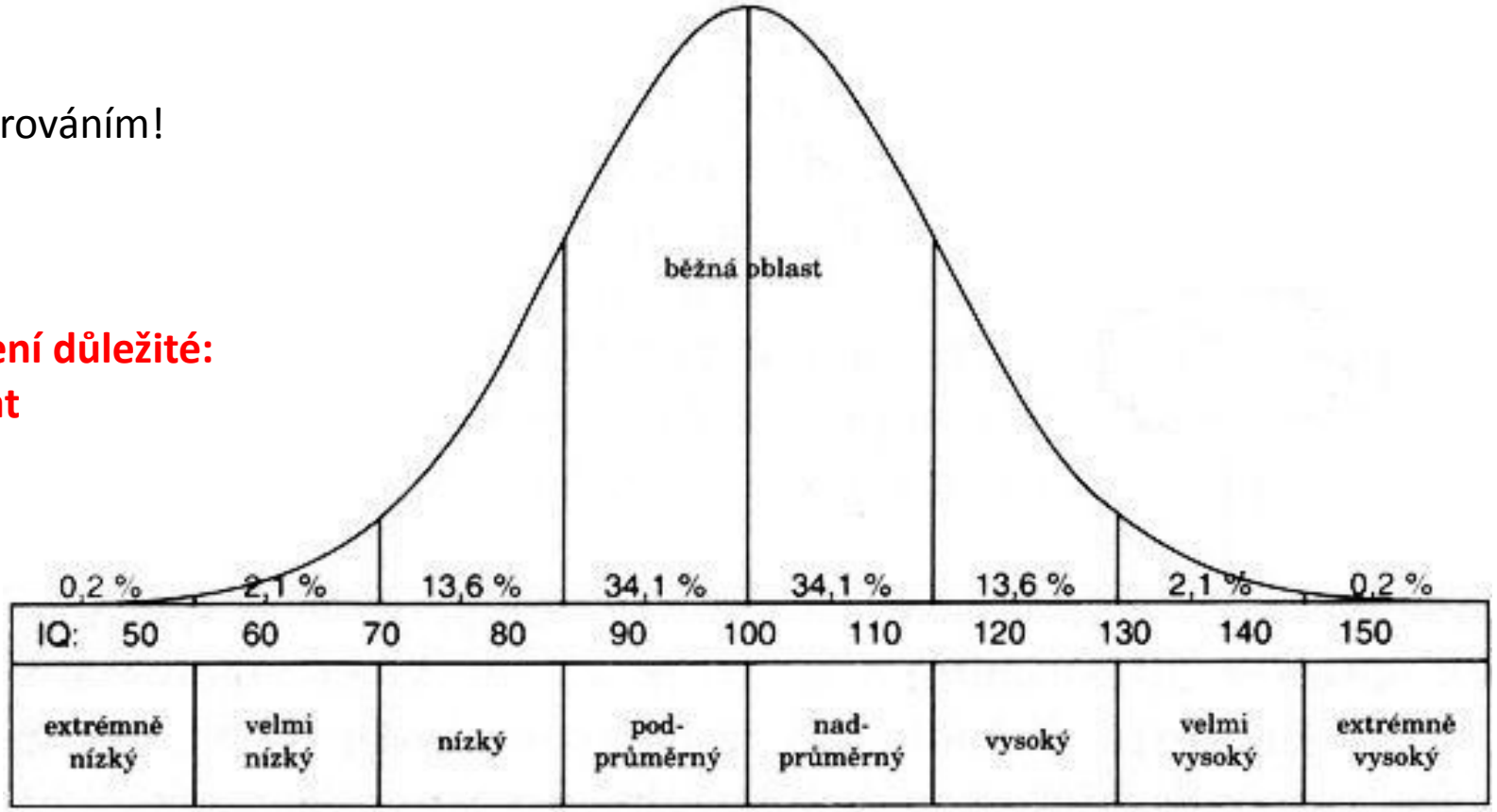
Základní statistické údaje

Míra polohy (centrální tendence)

Průměr = součet všech hodnot vydělený jejich počtem (typická hodnota souboru).
Medián = prostřední hodnota seřazeného souboru (50% dat nad i pod touto hodnotou)
Modus = nejčetnější hodnota v souboru (uni-, bi-, tri- atd. modální soubory dat).

PRŮMĚR vs. MEDIÁN
Průměr není rezistentní vůči odlehlým pozorováním!

Pro základní statistické hodnocení důležité:
normální rozložení dat



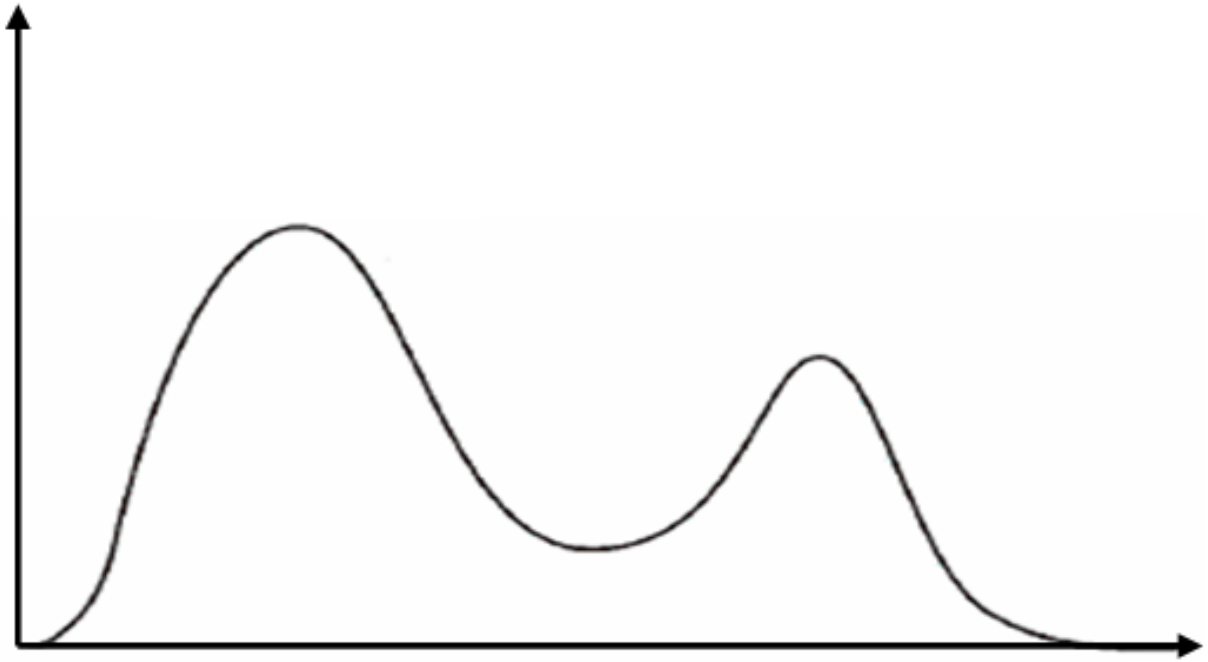
Základní statistické údaje

Míra polohy (centrální tendence)
Průměr = součet všech hodnot vydělený jejich počtem (typická hodnota souboru).
Medián = prostřední hodnota seřazeného souboru (50% dat nad i pod touto hodnotou)
Modus = nejčetnější hodnota v souboru (uni-, bi-, tri- atd. modální soubory dat).

PRŮMĚR vs. MEDIÁN
Průměr není rezistentní vůči odlehlým pozorováním!

Je nesmyslné počítat průměr např. při bimodálním rozložení dat!

Lze testovat: Shapiro-Wilk's test

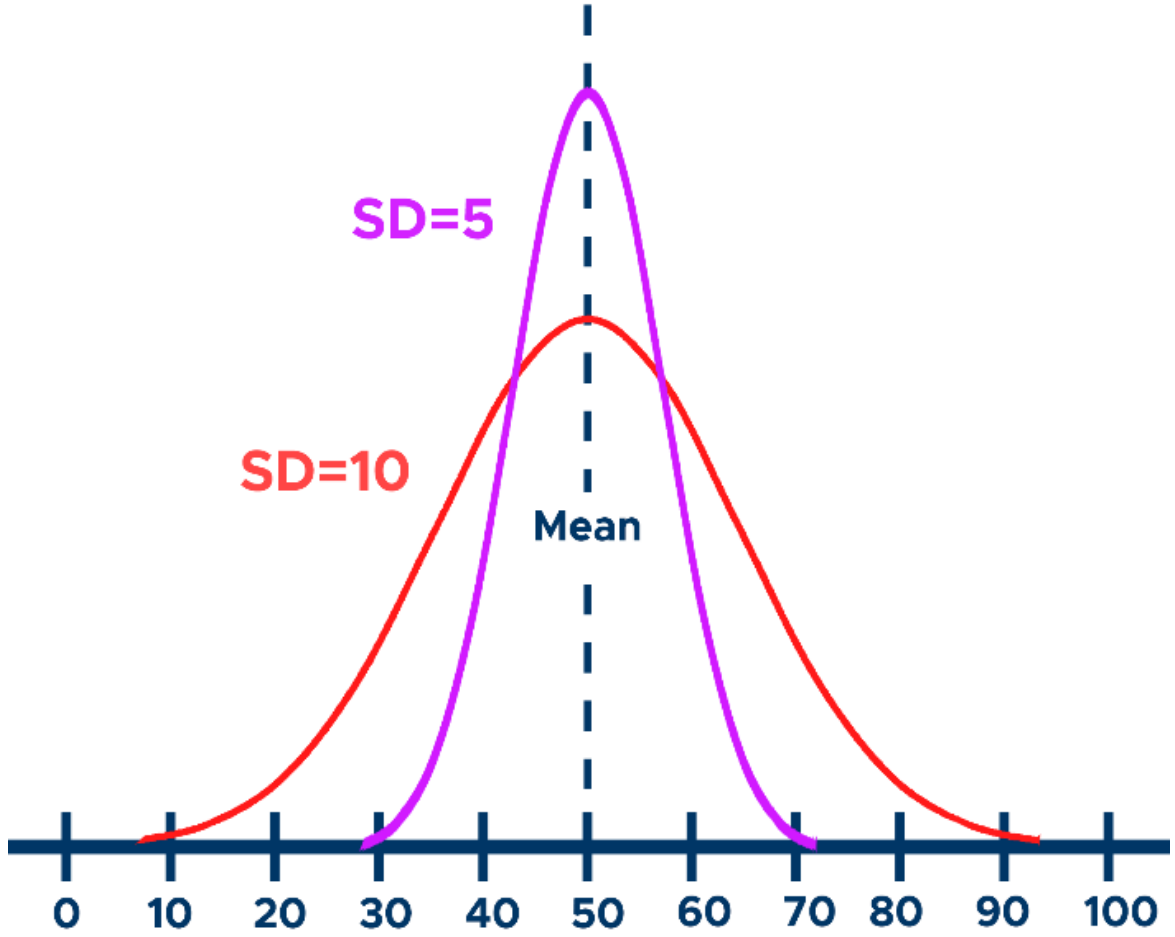


Základní statistické údaje

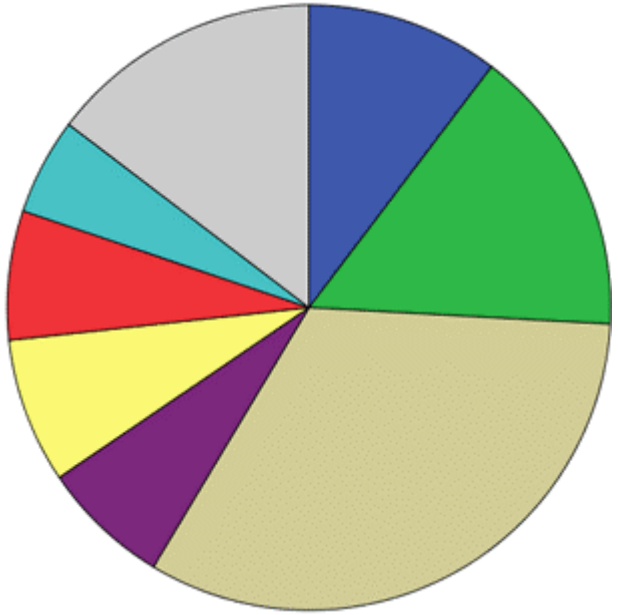
Míry variability souboru

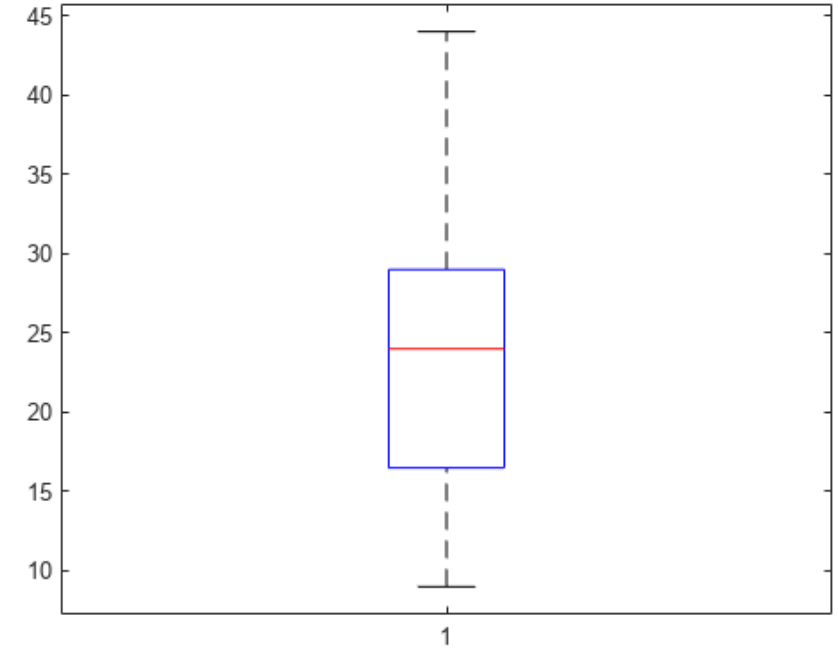
Rozpětí = rozdíl maxima a minima (max – min)
Rozptyl = **průměr čtverců odchylek od průměru** (vyjadřuje variabilitu rozdělení souboru náhodných hodnot kolem její střední hodnoty)
Směrodatná odchylka (standard deviation; SD) = **odmocnina z rozptylu** (vychází v původních jednotkách proměnné a tedy ji můžeme snadno interpretovat jako +- hodnoty, tedy jak daleko každá hodnota leží od průměru).
Variační koeficient (coefficient variation; CV) = směrodatná odchylka podělená aritmetickým průměrem hodnot (lze ji udávat i v procentech).

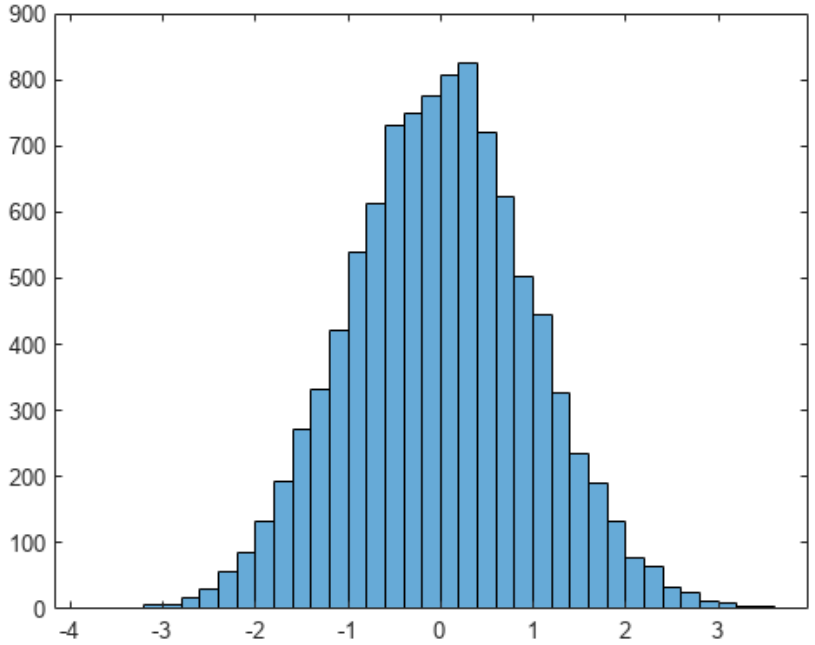
Směrodatná odchylka a **rozptyl** – neumožňují srovnání rozptylu proměnných, které mají různé rozměry (jednotky).
Rozptyl citlivý na odlehlé hodnoty.

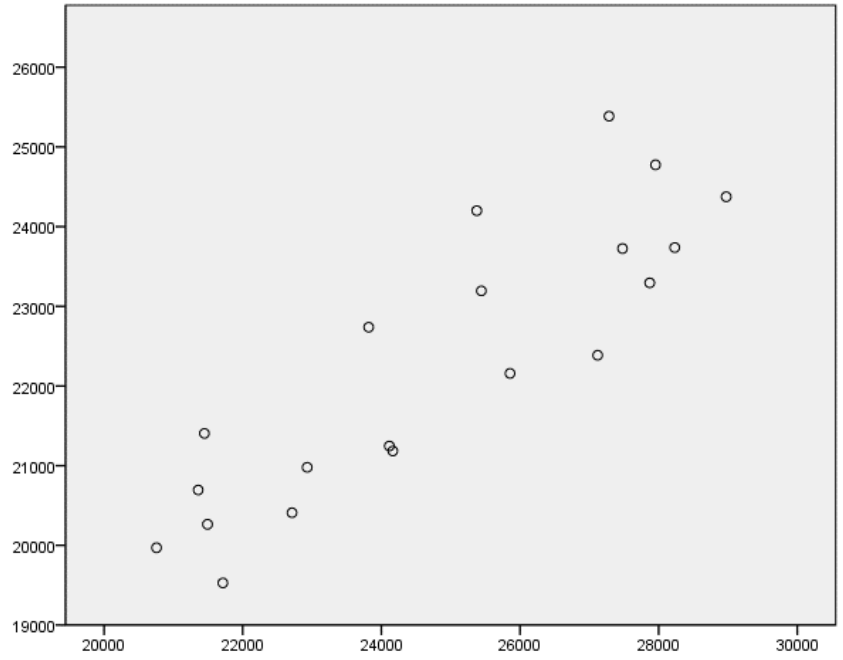


Vizualizace dat



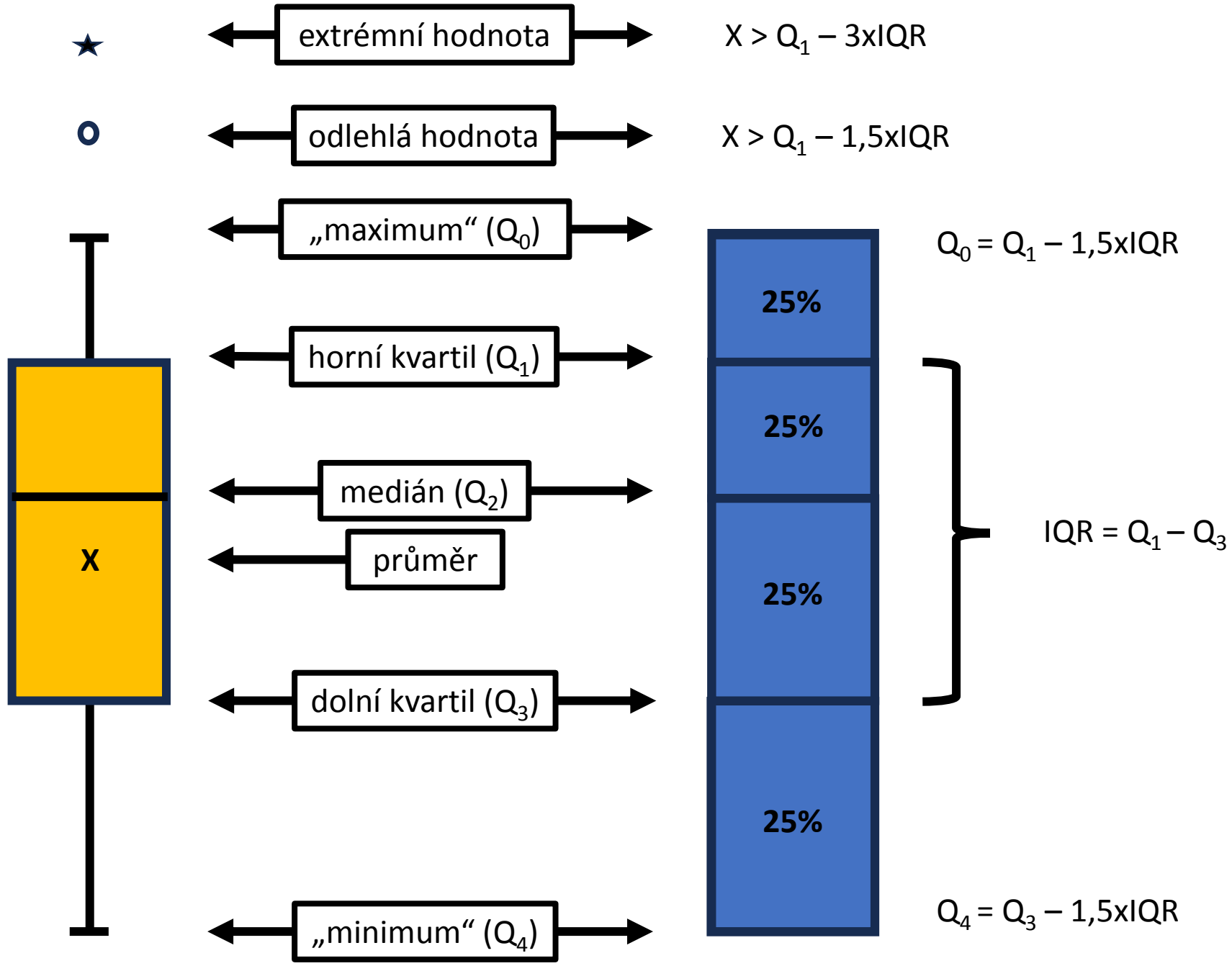






Boxplot

Krabicový graf

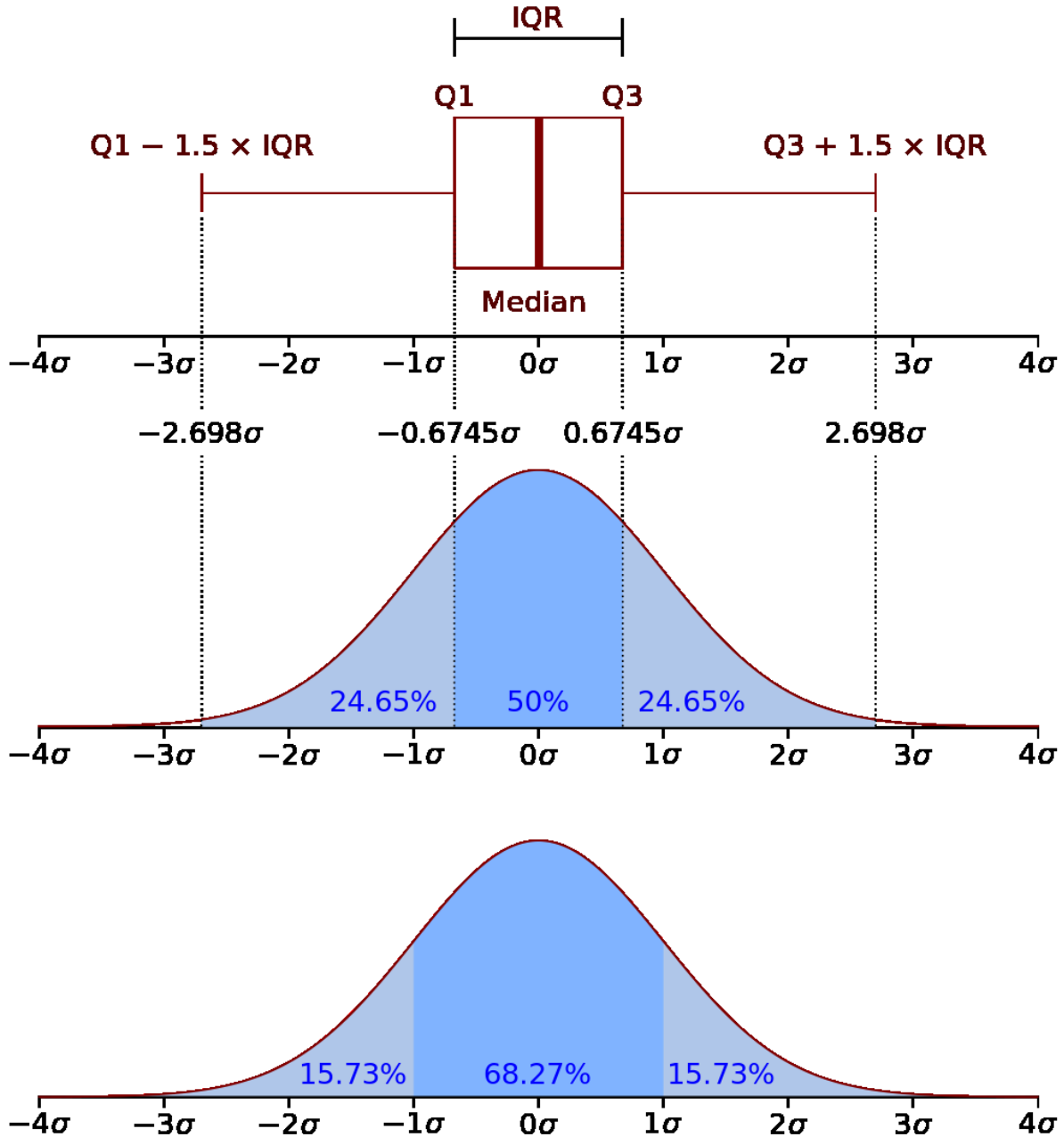


Znázorňuje centrální tendence dat a jejich rozptýlenost - uvádí medián, kvartily a nejmenší a největší hodnoty.

Umožňuje posouzení zešikmení a přítomnost odlehlých hodnot.

Boxplot

Krabicový graf



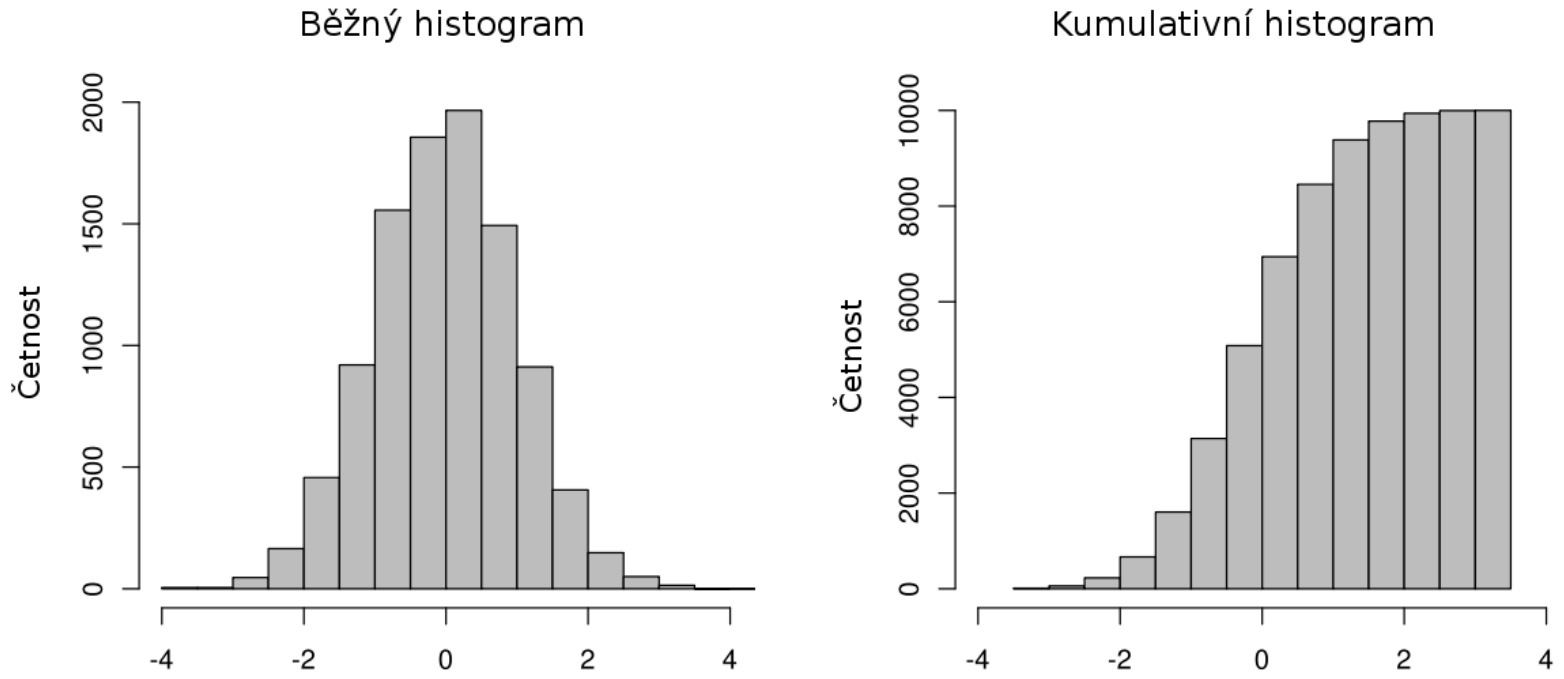
Souvislosti se směrodatnou odchylkou

Histogram

Vlastnosti histogramu

Špičatost
Šikmost (lze usuzovat z rozdílu mezi průměrem a mediánem)

Vhodný pro kategoriální/kvalitativní data.



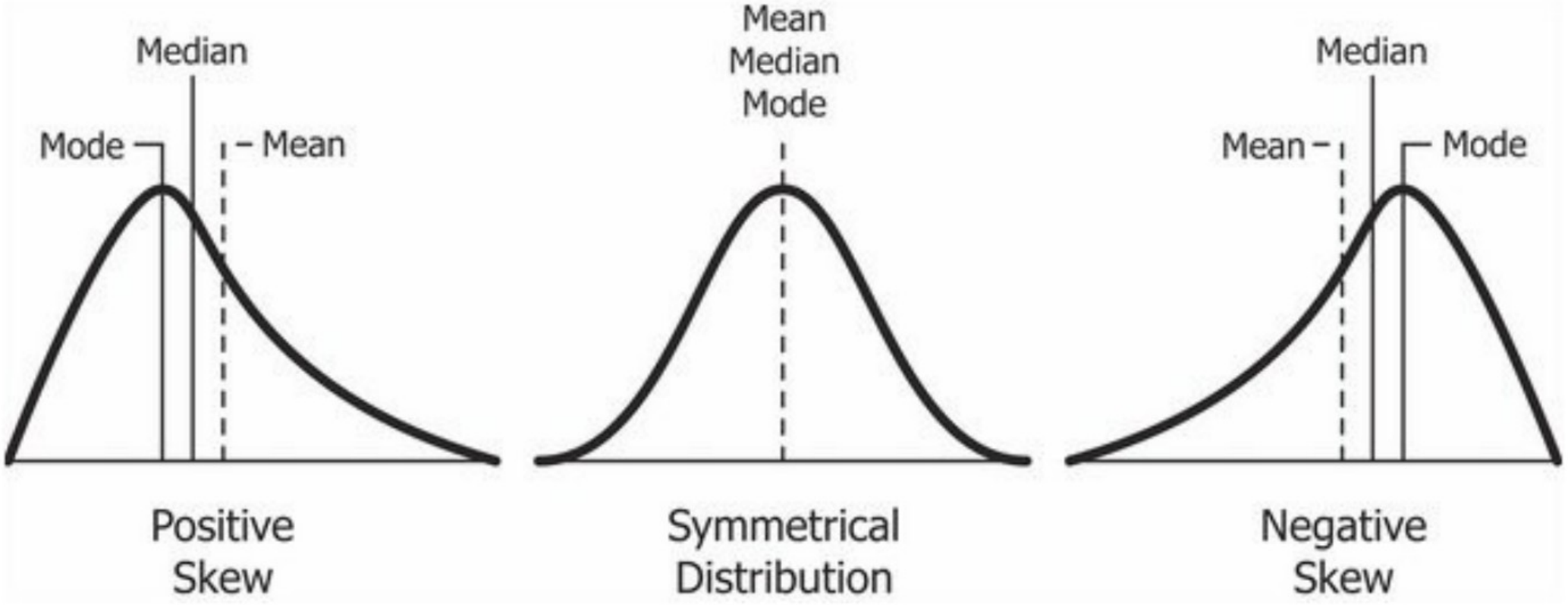
Sloupcový graf =
speciální varianta histogramu

Na ose x jsou možné hodnoty (kategorie, intervaly)

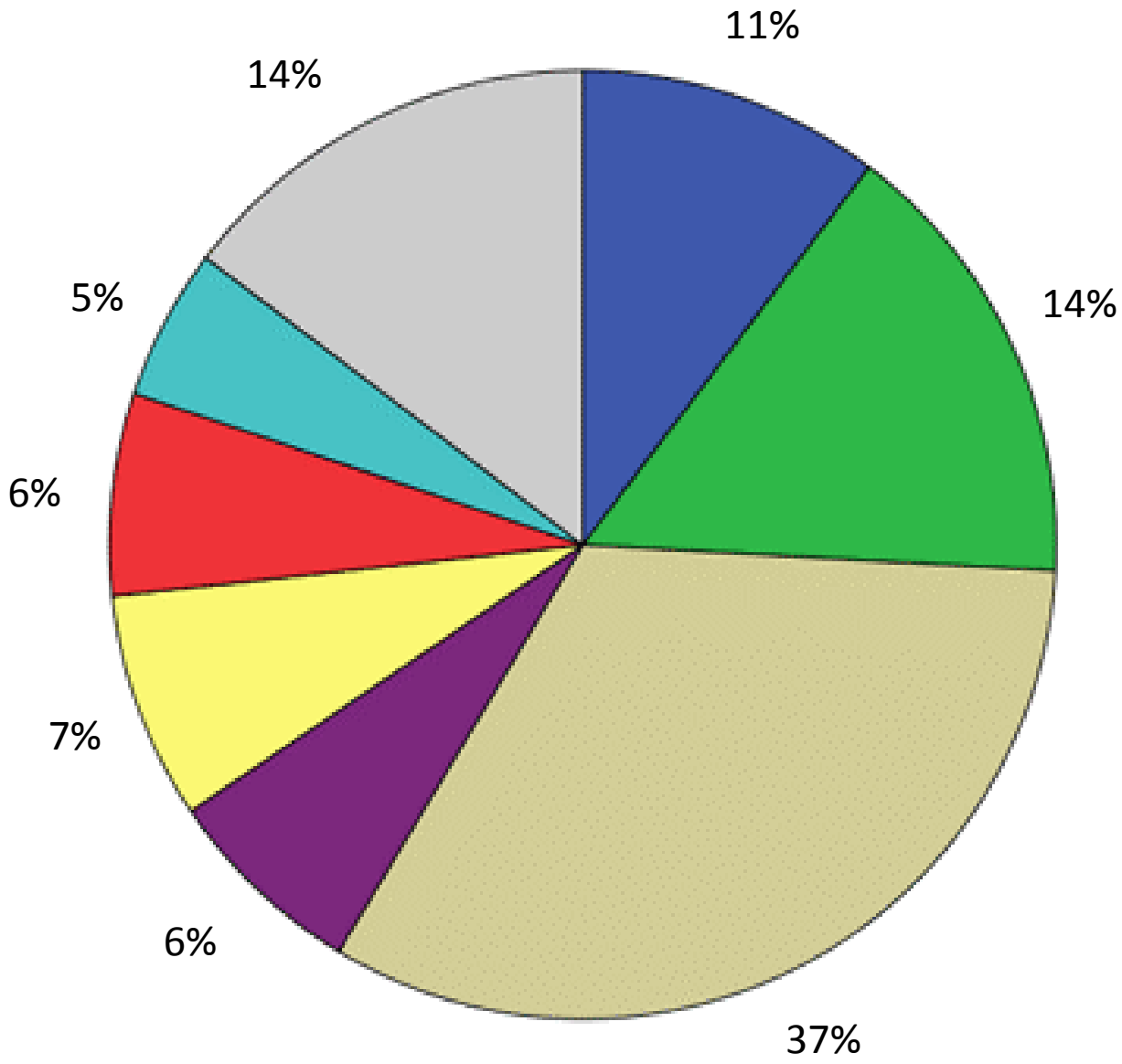
Histogram

Vlastnosti histogramu

Špičatost
Šikmost (lze usuzovat z rozdílu mezi průměrem a mediánem)



Koláčový graf



Znázorňuje rozdělení všech hodnot (100%) do menších částí.

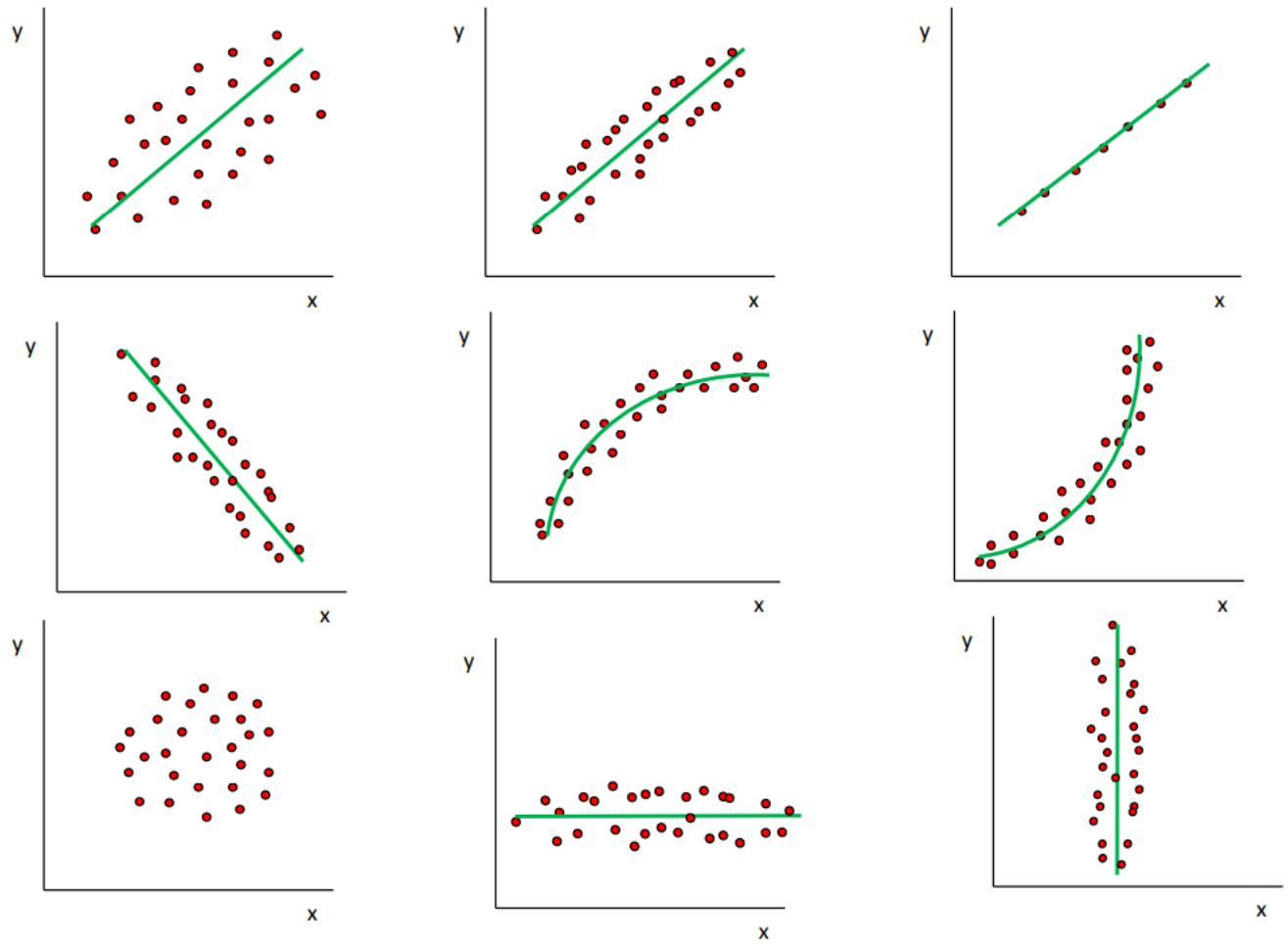
Jednotlivé výseky ukazují proporce dílčí části

Nejvhodnější pro nominální hodnoty reprezentující kvalitativně odlišné kategorie (např. barva květu: béžová, červená, žlutá)

Bodový graf

Vlastnosti bodového grafu

Typ závislosti (tvar křivky)
Směr (sklon křivky)
Těsnost (rozptyl bodů)



Bodový graf slouží pro zobrazení vztahu (závislostí) **dvou sad proměnných**.
Vhodný na korelace a regrese.

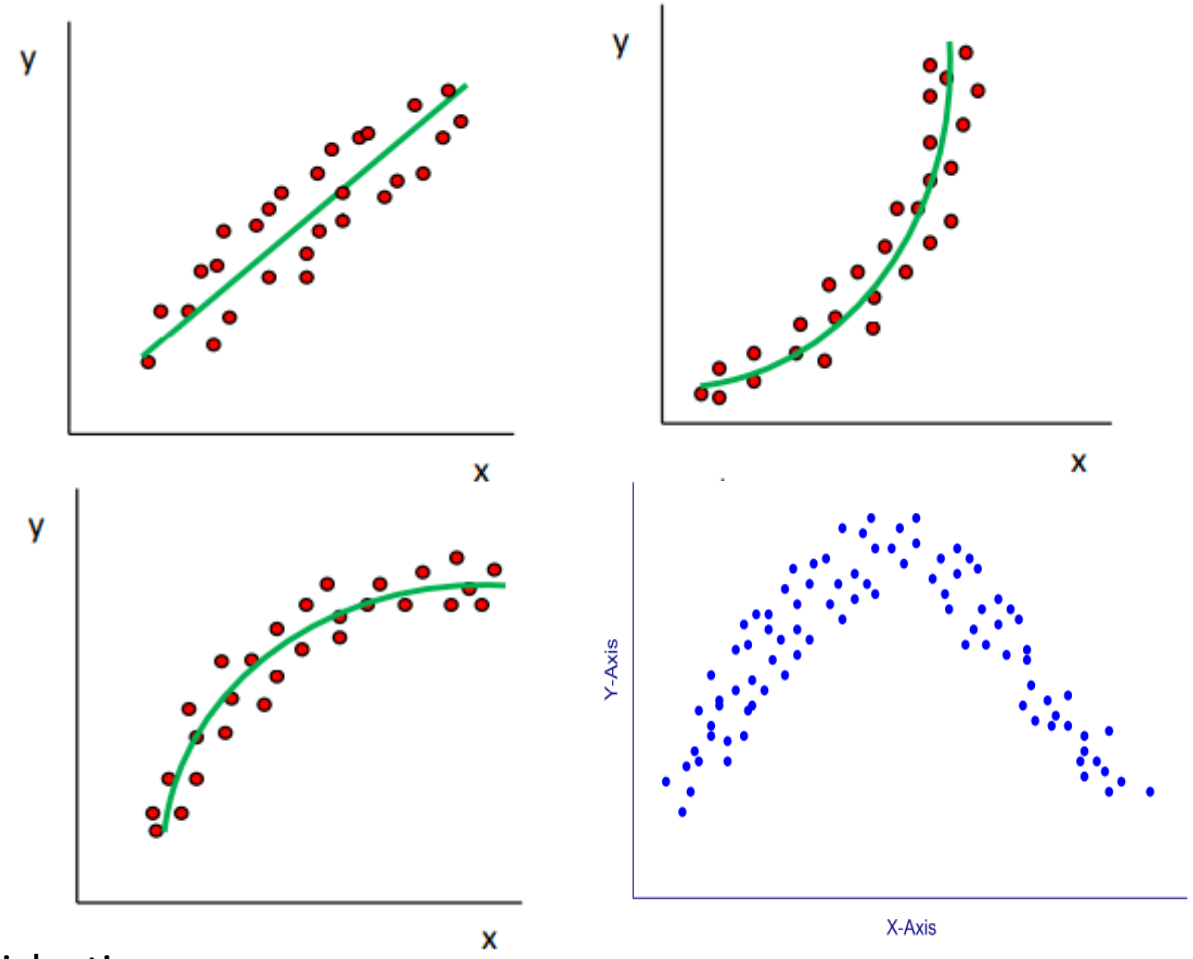
Bodový graf

Vlastnosti bodového grafu

Typ závislosti (tvar křivky)

- Lineární
- Logaritmická
- Exponenciální
- Parabolická

Směr (sklon křivky)
Těsnost (rozptyl bodů)



Lineární závislost = Pearsonův korelační koeficient
Nelineární závislosti= Spearmenův koeficient pořadové závislosti

Bodový graf slouží pro zobrazení vztahu (závislostí) **dvou sad proměnných**.
Vhodný na korelace a regrese.

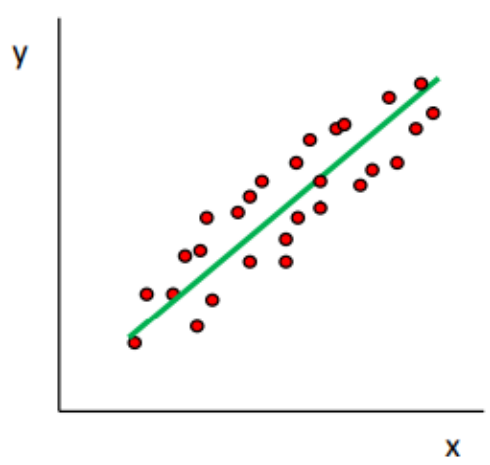
Bodový graf

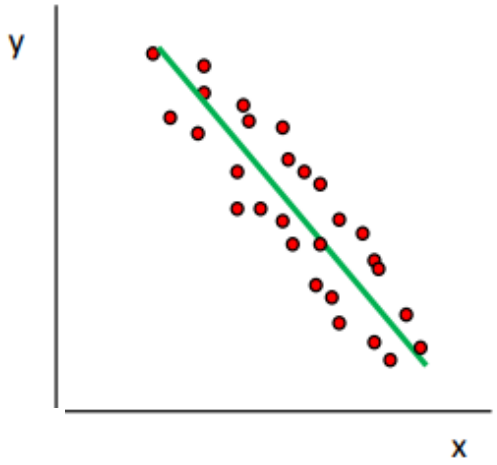
Vlastnosti bodového grafu

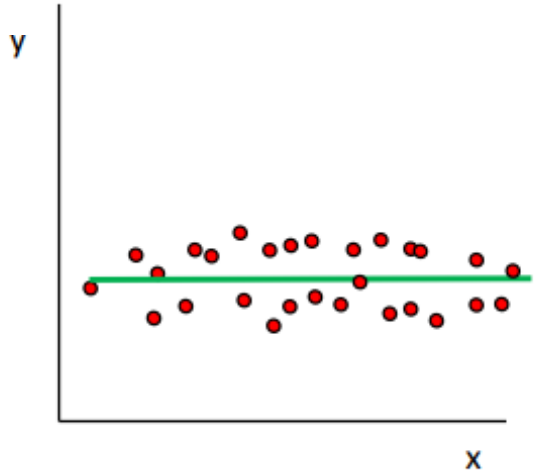
Typ závislosti (tvar křivky)
Směr (sklon křivky)

- Přímá (křivka stoupá)
- Nepřímá (křivka klesá)
- Bez závislosti (křivka vodorovná)

Těsnost (rozptyl bodů)







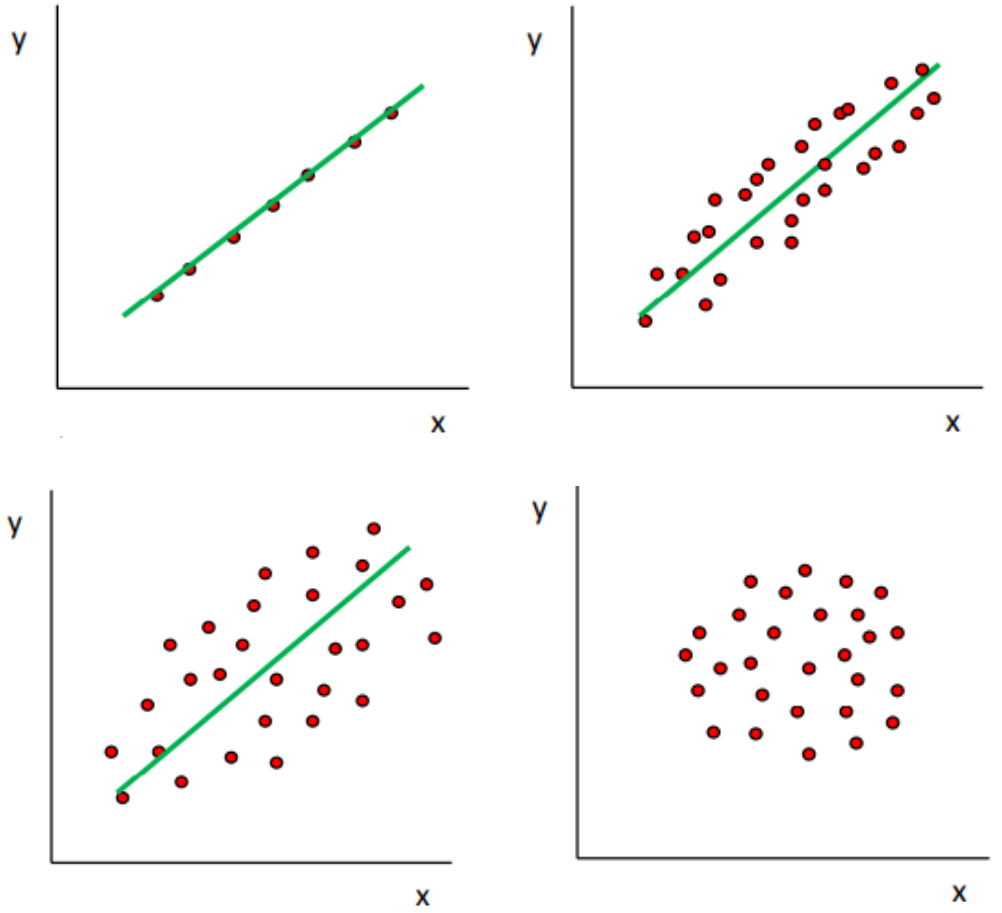
Bodový graf slouží pro zobrazení vztahu (závislostí) **dvou sad proměnných**.
Vhodný na korelace a regrese.

Bodový graf

Vlastnosti bodového grafu

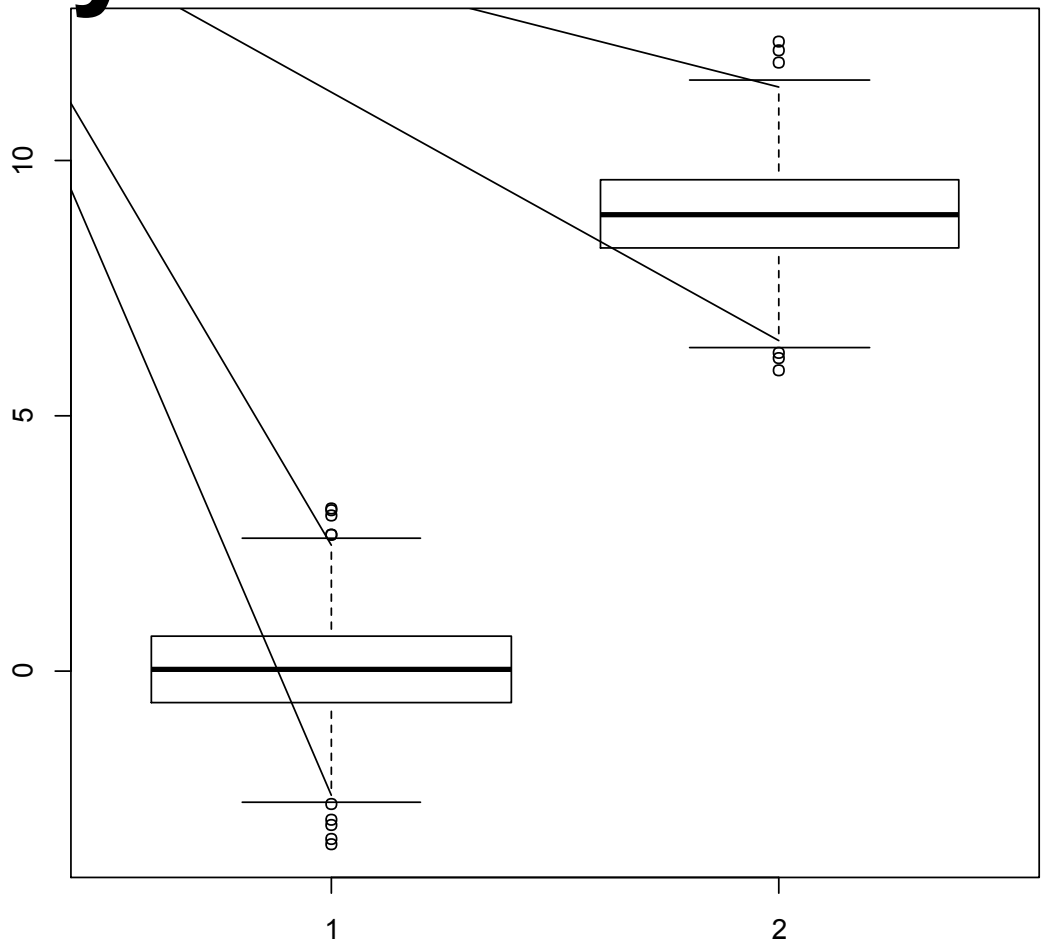
Typ závislosti (tvar křivky)
Směr (přímá, nepřímá)
Těsnost (rozptyl bodů)

- Jak na sobě dvě sady hodnot závisí

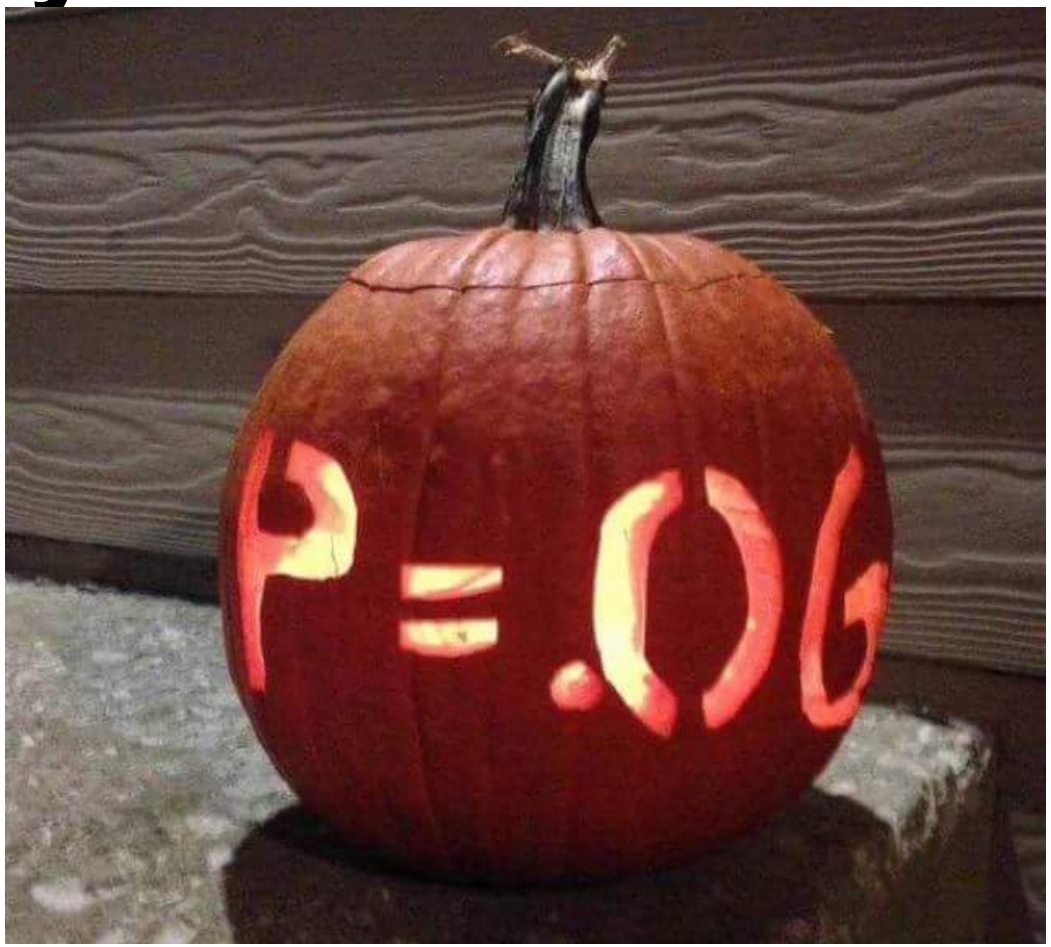


Bodový graf slouží pro zobrazení vztahu (závislostí) **dvou sad proměnných**.
Vhodný na korelace a regrese.

Jak se dva soubory proměnných liší aneb hledáme dobré znaky



Jak se dva soubory proměnných liší aneb hledáme dobré znaky



Signifikance

Signifikance (p-hodnota, p-value)

HYPOTÉZY

Pomocí statistického testu ověřujeme vždy dvě hypotézy

H0: formulovaná negativně („to, co chceme prokázat, **neplatí**“, tj. pozorovaný **efekt vznikl "náhodou**")

HA: formulovaná jako neplatnost H0, tj. že pozorovaný **efekt není náhoda**.

HLADINA VÝZNAMNOSTI α

Číslo od 0-1 (nejčastěji: 0,05; 0,01; 0,001)

PŘIJETÍ x ZAMÍTNUTÍ HYPOTÉZ

Pokud je **p-hodnota < α**, tak platnost H_0 je velmi málo pravděpodobná a potom **zamítáme H_0 na hladině významnosti α a přijímáme H_A.**

CHYBY

Chyba prvního typu – zamítneme-li hypotézu, ačkoliv je správná.

Chyba druhého typu – nezamítneme-li nulovou hypotézu ačkoliv není správná.

Testování odlišnosti dvou souborů

p-value je **vyšší** než 0.05,

hypotéza, že se testovaný (výběrový) soubor liší od konstanty **NEPLATÍ** na hladině významnosti 95%

= průměry testovaného souboru a konstanty se NELIŠÍ

```
One Sample t-test data:

boxplot_data[, 19]
t = -1.1003, df = 9, p-value = 0.2998
alternative hypothesis: true mean is not equal to 19.1308
95 percent confidence interval:
16.45674 20.05480
sample estimates:
mean of x
18.25577
```

Testování odlišnosti dvou souborů

Parametrické testy

Testované soubory musí mít normální rozložení dat a stejný rozptyl. Předpokládám, že testované soubory reprezentují vzorkovanou populaci.

- t-test (Studentův t-test): testuje střední hodnoty souborů (obvykle aritmetický průměr)
- F test: testuje rozdíly dvou rozptylů
- ANOVA

Neparametrické testy

U testovaných souborů nelze předpokládat normální rozdělení pravděpodobností sledovaného znaku. Výpočty u neparametrických testů vycházejí z pořadových čísel jednotlivých hodnot variační řady ("pořadové testy").

- Wilcoxonův test (Mann-Whitney test)
- Chí kvadrát: pro nominální data

Testování odlišnosti dvou souborů

T-test (Studentův t-test)

Předpoklad:

- normální rozložení dat (Shapiro-Wilk test)
- soubory se stejnými rozptyly (F-test)

Jednovýběrový t-test (porovnání základního a výběrového souboru)

Porovnávám aritmetický průměr základního souboru (= konstanta) s aritmetickým průměrem výběrového souboru dat.

Dvouvýběrový t-test (porovnání dvou výběrových souborů)

Ověřujeme, zda *rozdíl* aritmetických průměrů z obou měření je roven určitému číslu (často nule).

- **Párový test = porovnává dvě měření pocházejí z jednoho výběrového souboru** (např. odečet hodnot na stejných rostlinách před a po aplikaci nějakého roztoku do porostu)
- **Nepárový test = porovnává dva nezávislé výběry, které pocházejí ze dvou různých výběrových souborů** (např. pokusná a kontrolní skupina rostlin)

Testování odlišnosti dvou souborů

Mann-Whitney test (Wilcoxonův test)

Mann-Whitney test (porovnává dva nezávislé výběry, které pocházejí ze dvou různých výběrových souborů)

- Neparametrická analogie k nepárovému t-testu

Wilcoxonův test (porovnává dvě měření pocházejí z jednoho výběrového souboru)

- Neparametrická analogie k párovému t-testu (porovnává závislé výběry)

Výhody

Pořadové testy

Vhodné i pro ordinální data.

Lze je použít i pro data s normálním rozdělením

Testy nejsou citlivé na extrémní hodnoty

Nevýhody

Menší statistická síla oproti parametrickým testům

Testování odlišnosti dvou souborů

- Chí-kvadrát test vhodný pro nominální data (např. barva květu, ostnitost)
- Pracuji s četností pozorovaných znaků

- Data: kostka (zda padají čísla se stejnou pravděpodobností/četností)

- X-squared = 13.36, df = 5, p-value = 0.02023

p-value je nižší než 0.05,

Hypotéza, že všechna čísla nepadají se stejnou pravděpodobností platí na hladině významnosti 95%

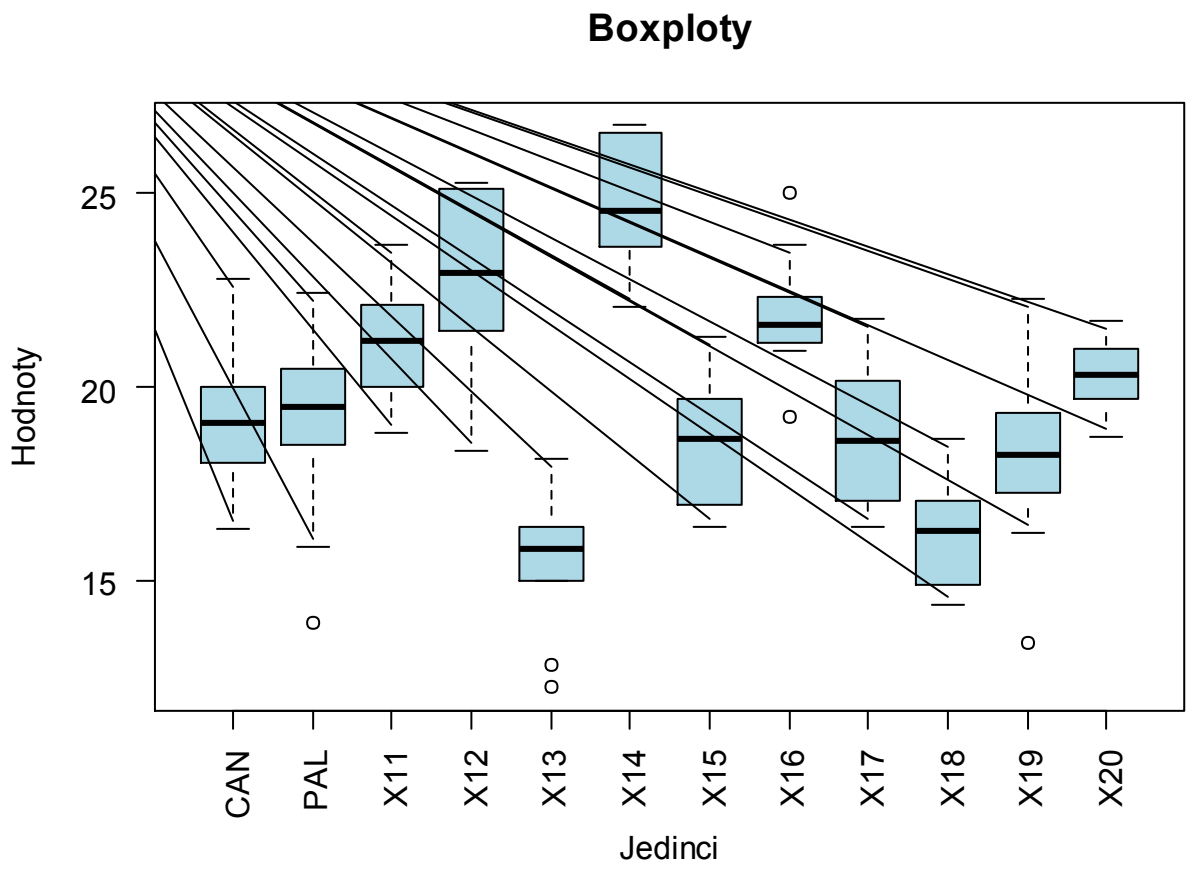
Když budu házet kostkou, pravděpodobnost, že dokážu opak, je cca 2%.
Na hladině významnosti 95% tedy vždy zjistím, že kostka je „cinknutá“.

Vyhodnocení a vizualizace dat

Boxploty
http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization

Testy
http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/76-add-p-values-and-significance-levels-to-ggplots/

AI
Chat GPT 4 - chat s nejdokonalejší umělou inteligencí na trhu (deeply.cz)



Průduchy hybridů mají značný rozptyl

Vysvětlení

1) Mohli by tam být triploidi
2) Jedná se o transgresivní znaky, tedy znaky, které u hybridů (F2, Bc) nabývají výrazně odlišných hodnot než mají rodiče.
3) Je to něco úplně jiného :D

Instrukce

- Do závěrečného "článku" tímto způsobem otestujte všechny znaky.
- Znaky, které se u CAN a OLE signifikantně liší a HYB je v nich intermediární vyneste i jako boxploty, zbytek do článku nedávejte.
- Udělejte tabulku použitelných (signifikantně odlišných) a nepoužitelných (signifikantně neodlišných) znaků. V diskusi se zamyslete, proč studované znaky nevyšly - málo vzorků? opravdu znak není dobrý? heterózní efekt, nebo naopak nižší vitalita hybridů?
- Cytometrická data vyhodnoťte ve zvláštní kapitole Průtoková cytometrie. Nedávejte je k morfometrickým datům.