Přednáška č. 5

SOUBOR A VÝBĚR, BODOVÉ ODHADY PARAMETRŮ, LIMITNÍ VĚTY

18. 3. 2021

P. Pecherková

Soubor, výběr

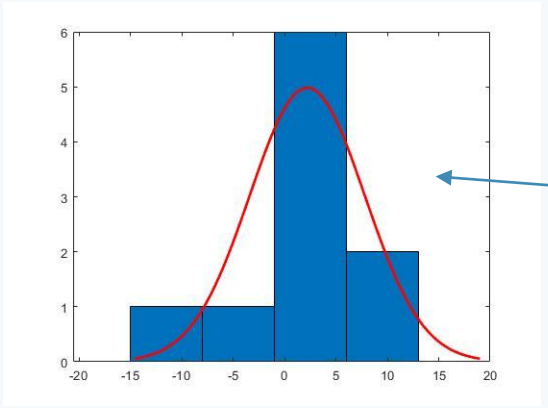
- Náhodná veličina
  - všechny možné výsledky náhodného pokusu,
  - Diskrétní nebo spojitá náhodná veličina
  - úplný popis je distribuční funkce (pravděpodobnostní funkce / hustota pravděpodobnosti)
  - Známe rozdělení
  - Problém : v praxi není možné / efektivní měřit všechny možné hodnoty

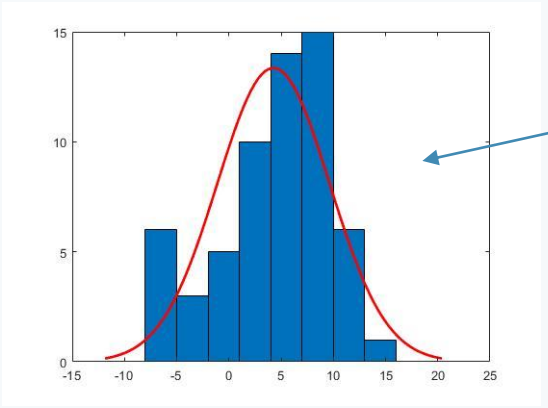
Příklad:

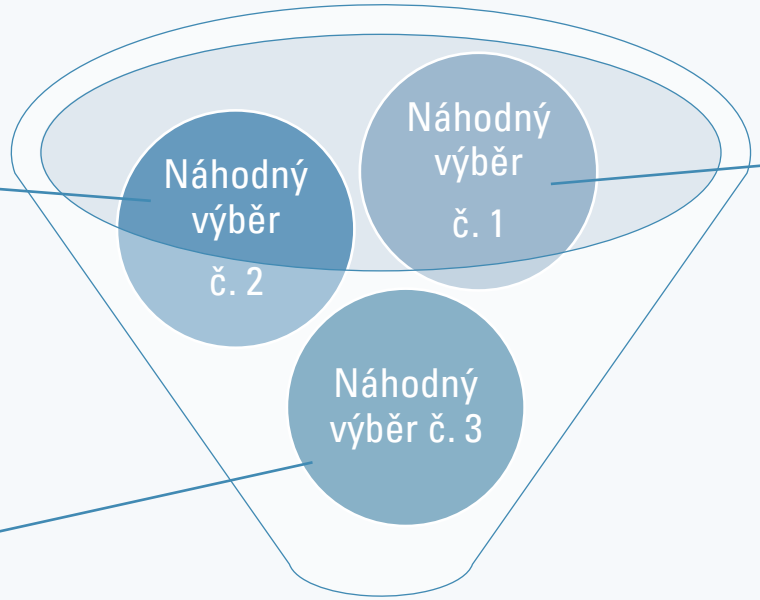
- Průměrný věk v EU (nemožné)
- Zjišťovat preferovaný dopravní prostředek na cestě do práce / školy

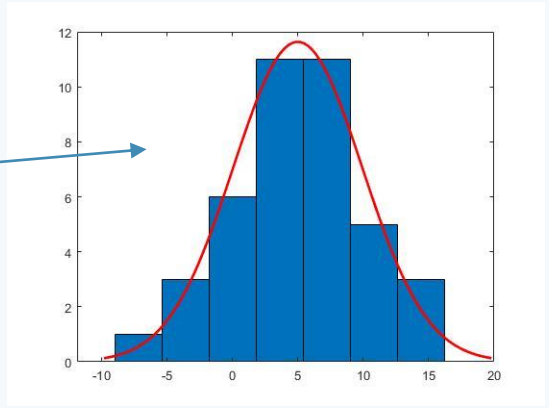
Část souboru, různě velké, náhodné, nezávislé, **stejně rozdělené**

Populace, všechny možnosti, neznáme přesně









Jak vypadá soubor, ze kterého pochází?

Odhadneme:

1. Normální rozdělení
2. E[X]=?
3. D[X]=?

Statistika

Statistika (bodový odhad)

- funkce výběru, hodnota která nám dá bodový odhad parametru

$$T(X) = \hat{\theta}$$

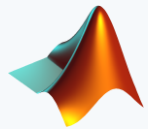
- Minulý příklad: lze odhadnout buď parametr μ nebo σ
- Konstrukce statistiky:
  1. Metoda momentů – jednodušší, ale větší omezení
  2. Metoda maximální věrohodnosti – složitější, často používaná pro svou univerzalitu

Více o obou metodách je ve skriptech. My budeme využívat již odvozené (vypočtené) statistiky.

- Výběrový průměr
  - Je to obdoba střední hodnoty, proto nejlepším bodovým odhadem pro střední hodnotu je průměr (viz konstrukce statistiky)
  - Výběrovým průměrem náhodného výběru je

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - Vhodné pro binomické, Poissonovo, normální či exponenciální rozdělení
  - Střední hodnota výběrového průměru je rovna střední hodnota souboru
  - Rozptyl výběrového průměru je roven podílu rozptylu a počtu dat $\left(\frac{\sigma^2}{n}\right)$
    - to znamená, že čím víc máme dat ve výběru, tím menší je rozptyl výběrového průměru



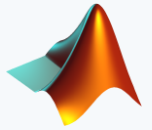
- Výběrový rozptyl
  - Je to obdoba souborového rozptylu, proto se používá jako nejlepší bodový odhad pro souborový rozptyl
  - Již jsme o něm hovořili a je i téměř podobný souborovému (rozdíl je v jmenovateli pod sumou)

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

  - Střední hodnota výběrového rozptylu je

$$E[s^2] = \sigma^2$$

  - U normálního rozdělení

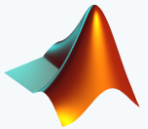


- Výběrový podíl
  - je to obdoba souborového podílu, proto se používá jako nejlepší bodový odhad pro souborový podíl

$$p = \frac{n^+}{n}$$

kde n^+ je počet úspěchů

  - u alternativního rozdělení



Vlastnosti statistiky

Abychom bodový odhad (statistiku) určili jako vhodnou, měla by splnit následující vlastnosti:

1. Nestrannost
2. Konzistence
3. Vydatnost

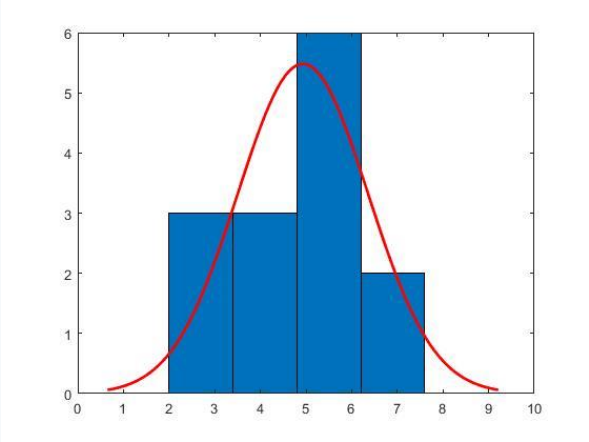
Může se stát, že odhad nebude mít jednu ze zadaných vlastností?

ANO, ale ideální je mít takovou statistiku, která splňuje všechny vlastnosti.

- Nestrannost
  - Statistika T poskytuje nestranný bodový odhad parametru, jestliže její střední hodnota se rovná tomuto parametru

$$E[T] = \theta$$

  - Střední hodnotu $E[T]$ je možné si představit jako průměr přes všechny možné výběry, resp. přes všechny bodové odhady, které z těchto výběrů uděláme.
  - *Příklad: výběrový průměr*

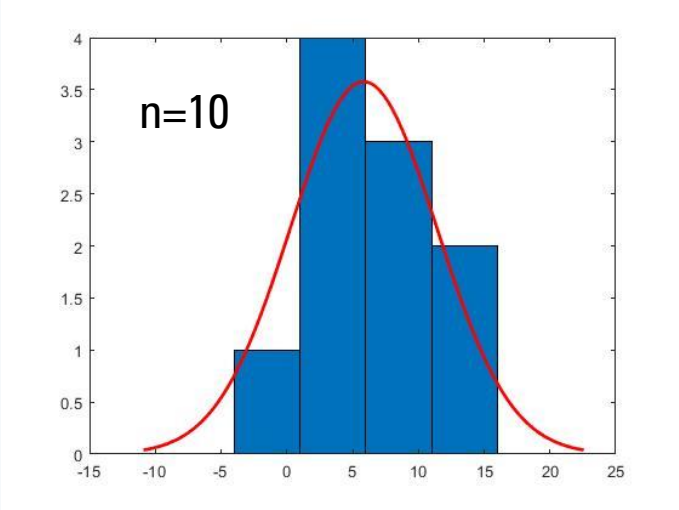


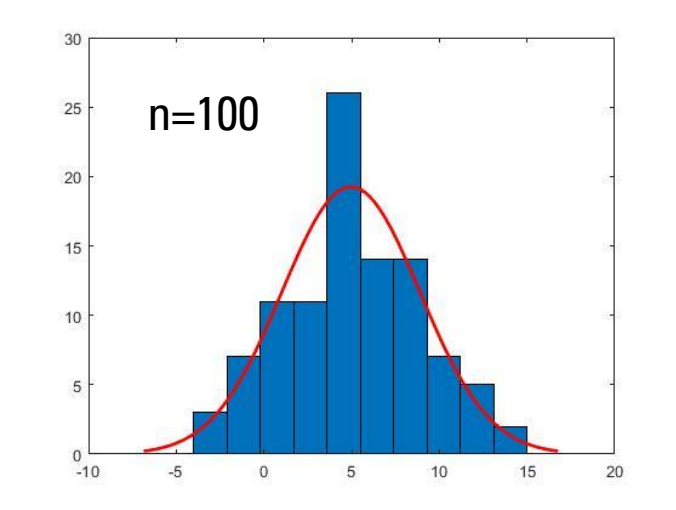
Pokud je odhad nestranný, neznamená to, že je automaticky dobrý, jen se vylučuje systematická chyba

- Konzistence
  - Statistika T dává konzistentní bodový odhad parametru θ, jestliže pro rostoucí rozsah výběru se hodnota statistiky (v pravděpodobnosti) neomezeně blíží skutečnému parametru

$$\lim_{n\to\infty} P(|T-\theta|<\epsilon)=1, \qquad \forall\epsilon>0$$

  - Tato vlastnost statistiky říká, že s rostoucím rozsahem výběru $n \to \infty$ rozptyl statistiky klesá $D[T] \to 0$





- Vydatnost
  - Jak již bylo řečeno, neznamená nestrannost to, že je odhad dobrý. V ideálním případě budou bodové odhady rozmístěny těsně kolem odhadovaného parametru, tedy s co nejmenším rozptylem. Na tomto principu je založena vydatnost.
  - Pro dvě nestranné statistiky T a U definujeme jako vydatnější tu z nich, která má menší rozptyl

$$D[T] < D[U] \Rightarrow \text{T je vydatnější než U}$$

  - Pozor, tato vlastnost platí pouze nestranné statistiky. Pokud nemáme nestrannou statistiku, je třeba vydatnost porovnávat pomocí MSE (mean square error). Pokud Vám přijde tento termín povědomý, již byl použit u regrese, jako RMSE (rozptyl na jeden vzorek).

Limitní věty

Limitní věty jsou důležité, protože nám umožňují pracovat s daty efektivněji, jednodušeji a můžeme využít dobré vlastnosti jiných rozdělení.

Existuje velké množství limitních vět, my si uvedeme jen ty dvě základní.

Zákon velkých čísel

- Jsou-li X_1, X_2 ... nezávislé a stejně rozdělené náhodné veličiny (výsledky náhodných a nezávislých pokusů) s konečnou střední hodnotou blíží k teoretickým charakteristikám souboru.
- Při velkém rozsahu výběru $n \to \infty$ se hodnoty výběrových charakteristik se neomezeně blíží k charakteristikám souboru.
- Tento zákon byl již využit u vlastnosti statistiky – nestrannost.
- Ve statistice o dostatečném rozsahu hovoří u výběru o rozsahu $n = 30$.

- Centrální limitní věta
  - Moiverova-Laplaceova věta
    - Náhodná veličina X s binomickým rozdělením, která vznikla jako součet velkého množství nezávislých náhodných veličin X_i s alternativním rozdělením (Alt(π)), konverguje k normálnímu rozdělení.
    - *Příklad: Pravděpodobnost, že automobil na křižovatce pojede rovně je 0,9. Jaká je pravděpodobnost, že při sledování 300 automobilů se bude průjezd přibližně rovnat střední hodnotě počtu automobilů jedoucích rovně.*
  - Lévyho-Lindebergova věta
    - Náhodná veličina X, která vznikla jako součet velkého počtu vzájemně nezávislých náhodných veličin X_i se stejným rozdělením, má za velmi obecných podmínek přibližně normální rozdělení.
    - *Příklad: Opakovaný hod kostkou*