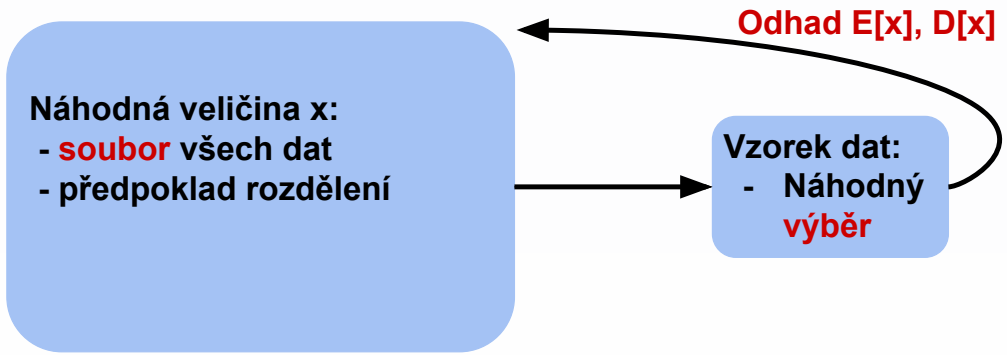
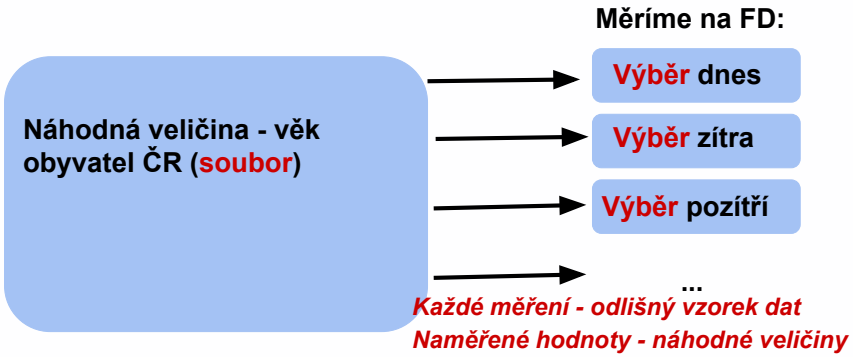
Přednáška 5 – soubor, výběr, bodový odhad, limitní věty

- Jak správně naměřit data?
- Jak z dat odhadnout charakteristiky náhodné veličiny?



Příklad: věk obyvatel ČR

- není možné obejít všechny
- náhodný výběr ⇒ průměrný věk



Výběr

je množina nezávislých a stejně rozdělených náhodných veličin:

$X = [X_1, X_2, \ldots, X_n]$

n – počet hodnot výběru

- Proč stejně rozdělené N.V.? – věk obyvatel ČR neměříme jinde
- Proč nezávislé N.V.?
  - reprezentativní výběr (~~mateřská školka~~, ~~domov důchodců~~)
  - nezávislá skupina obyvatel (MHD, aquapark, atd.)

Statistika

- Cíl – odhad charakteristik (parametrů) souboru pomocí výběru

Předpoklad – normální rozdělení:

věk obyvatel ČR $\sim N(\mu, \sigma^2)$, μ – střední hodnota, σ^2 – rozptyl

- Odhad μ, σ^2 z výběru – statistika

Statistika

je funkce výběru, hodnota které nám dá bodový odhad parametrů

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

θ – parametry μ, σ^2

- výběr X se při každém měření liší
- statistika $T(X)$ – ~~přesná hodnota~~ θ, při každém výpočtu – jiný $\hat{\theta}$
- statistika T je také náhodná veličina se svým rozdělením $f(T)$

Konstrukce statistiky – samostatná úloha

- metoda momentů, metoda maximální věrohodnosti
- hotové statistiky – výběrový průměr, výběrový rozptyl, výb. podíl

Výběrový průměr – funkce výběru (statistika) $T = \bar{X}$

$$\underbrace{\bar{X}}_{N.V.} = \frac{1}{n}\sum_{i=1}^{n}\underbrace{X_i}_{N.V.}$$

Příklad: věk cestujících v metru

[12 36 4 41 15 16 28 35 33 61 ...]

3 výběry:

$X = [12\ \ 4\ \ 28],\ \ \bar{X} = 14.67$

$X = [41\ \ 35\ \ 33],\ \ \bar{X} = 36.33$

$X = [4\ \ 28\ \ 15],\ \ \bar{X} = 15.67$

Výběrový průměr je v každém případě jiný

Výběrový průměr

$\bar{X}$ – náhodná veličina se střední hodnotou $E[\bar{X}]$ a roztpylem $D[\bar{X}]$

- Střední hodnota výběrového průměru se rovná střední hodnotě souboru:

$$E[\bar{X}] = \mu$$

Význam:

- každý výběr z MHD dá odlišný $\bar{X}$,
  ale v průměru – průměrný věk obyvatel ČR μ
- má smysl pracovat s výběrem

Důkaz: dáme $\bar{X}$ podle definice do závorek:

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] =$$

$\frac{1}{n}$ je konstanta, počítání se střední hodnotou je $E[ax] = aE[x]$, tedy platí:

$$= \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] =$$

dále střední hodnota součtu se rovná součtu středních hodnot $E[x + y] = E[x] + E[y]$, proto

$$= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] =$$

$E[X_i] = \mu$ je střední hodnota souboru (věku) podle definice, protože věk $\sim N(\mu, \sigma^2)$. Platí tedy

$$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{E\left[X_i\right]}_{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu.$$

Rozptyl výběrového průměru

se rovná rozptylu souboru vydělenému počtem dat $D[\bar{X}] = \frac{\sigma^2}{n}$

Důkaz: Dáme $\bar{X}$ podle definice do závorek:

$$D[\bar{X}] = D\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] =$$

$\frac{1}{n}$ je konstanta, počítání s rozptylem je $D[ax] = a^2 D[x]$, tj:

$$= \frac{1}{n^2} D\left[\sum_{i=1}^{n} X_i\right] =$$

pro nezávislé N.V. $D[x+y] = D[x] + D[y]$, tj podle definice výběru

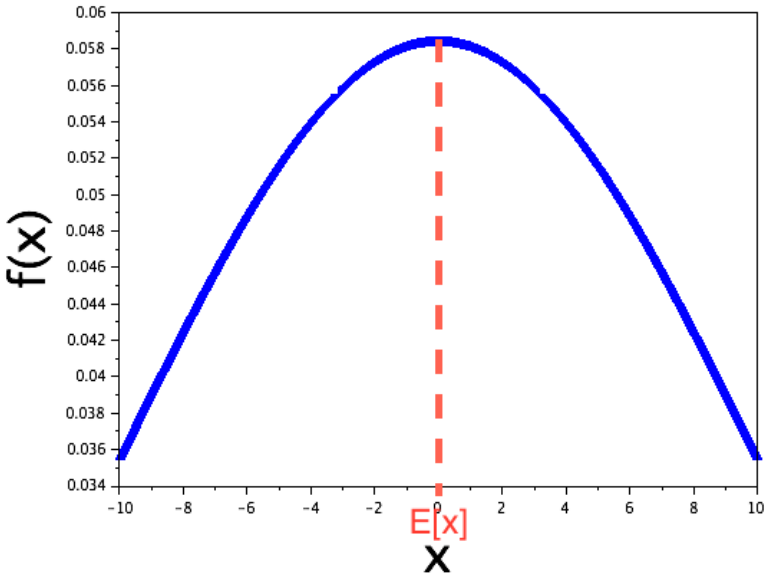
$$= \frac{1}{n^2}\sum_{i=1}^{n} D\left[X_i\right] =$$

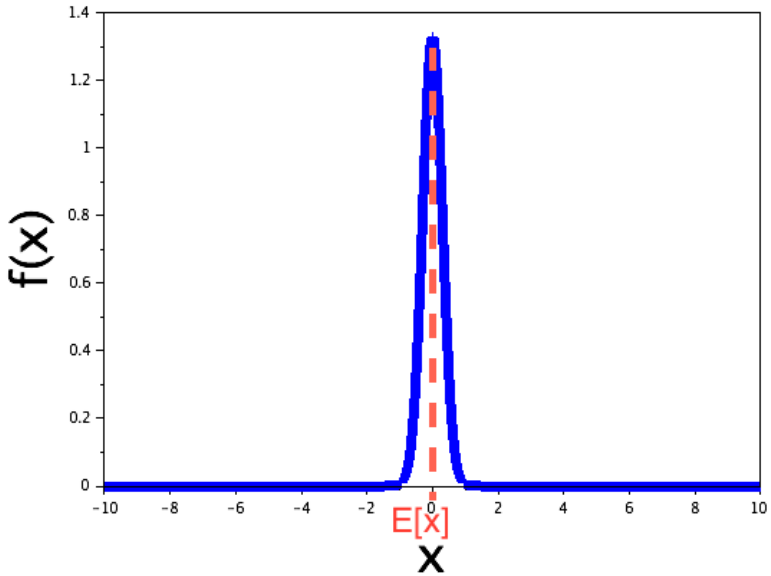
$D[X_i] = \sigma^2$ je rozptyl souboru, protože věk $\sim N(\mu, \sigma^2)$, tedy

$$= \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{D\left[X_i\right]}_{\sigma^2} = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Význam:

- v praxi – použití co největšího výběru
- více dat (větší n) – menší rozptyl výběrového průměru
  - přesnější odhad blíž ke střední hodnotě souboru





Výběrový rozptyl – funkce výběru (statistika) $T = s^2$

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

Výběrový rozptyl s^2 – náhodná veličina

- Střední hodnota výběrového rozptylu se rovná rozptylu souboru

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

Příklad: výběr z metra – průměrná rozptýlenost hodnot výběru je stejná jako pro soubor obyvatel ČR

Poznámka: výběrová směrodatná odchylka, výběrová kovariance, výběrový korelační koeficient

Výběrový podíl – funkce výběru (statistika) $T = p$

- alternativní rozdělení souboru (úspěch / neúspěch)

$$p = \frac{n^+}{n}, \qquad n^+ \text{ je počet úspěchů ve výběru}$$

Příklad: podíl cestujících s kočárkem

Limitní věty – vztah výběru $\bar{X}, s^2$ a souboru μ, σ^2

Zákon velkých čísel

Při velkém rozsahu výběru $n \to \infty$ se hodnoty výběrových charakteristik neomezeně blíží k charakteristikám souboru:

$$\bar{X} \to \mu \qquad s^2 \to \sigma^2$$

Příklad: výběrový průměr $\bar{X}$ z velkého výběru z MHD $\to$ průměrný věk obyvatel ČR (charakteristika souboru μ)

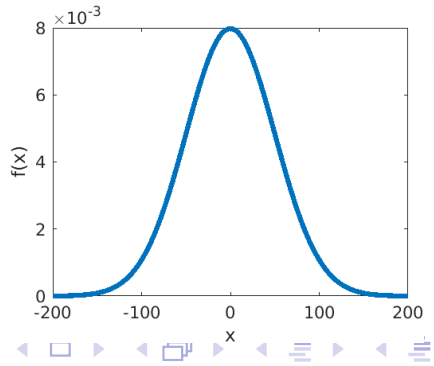
Centrální limitní věta

Při velkém rozsahu výběru $n \to \infty$ má výběrový průměr přibližně normální rozdělení

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Význam:

- $\bar{X}$ z velkého výběru z MHD – μ souboru
- Větší výběr ($\Uparrow n$) – menší rozptyl ($\Downarrow \frac{\sigma^2}{n}$)
- jiné rozdělení výběru – $\bar{X}$ má normální rozdělení



Vlastnosti statistiky

Pro určení bodového odhadu $\hat{\theta} = \{\hat{\mu}, \hat{\sigma}^2\}$ – tři vlastnosti statistiky:

- nestrannost
- konzistence
- vydatnost

Nestrannost

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

$$E[T] = \theta$$

Příklad: výběrový průměr

$\bar{X}$ – nestranná statistika

$$\underbrace{E[\bar{X}]}_{E[T]} = \underbrace{\mu}_{\theta}$$

Příklad: výběrový rozptyl

s^2 – nestranná statistika

$$\underbrace{E[s^2]}_{E[T]} = \underbrace{\sigma^2}_{\theta}$$

Neplatí pro $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ – není nestranná

Význam: bodový odhad je v průměru přesný

Konzistence

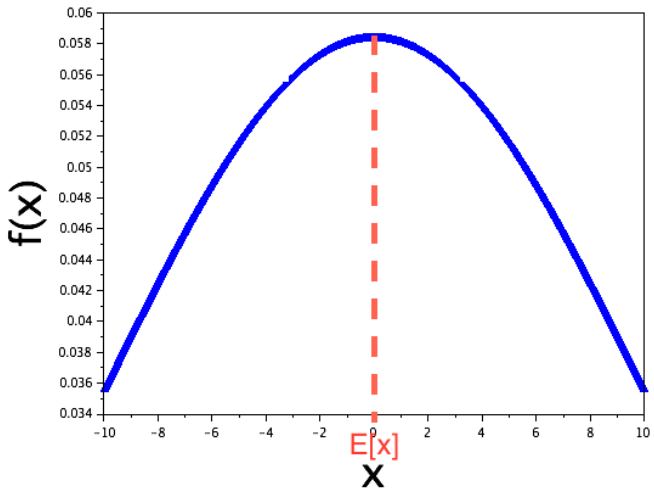
Statistika je konzistentní, pokud s rostoucím rozsahem výběru $n \to \infty$ se rozptyl statistiky blíží k nule $D[T] \to 0$

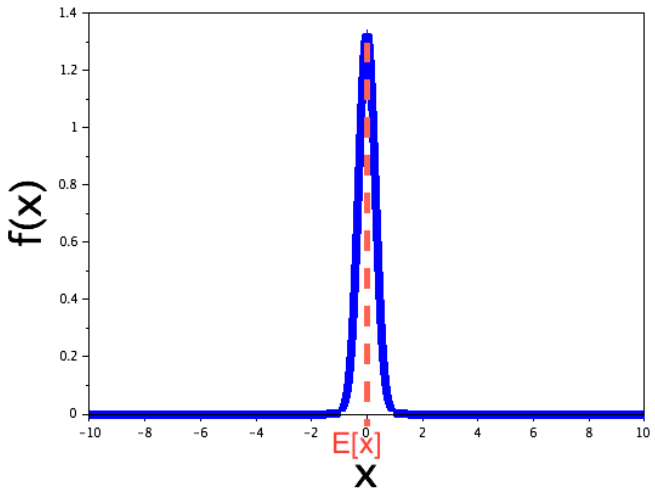
Příklad: výběrový průměr $\bar{X}$ – konzistentní statistika

$$\underbrace{D[\bar{X}]}_{D[T]} = \frac{\sigma^2}{n}$$

Význam:

Větší výběr ($\Uparrow n$) – menší rozptyl ($\Downarrow \frac{\sigma^2}{n}$) – přesnější odhad





Vydatnost

Čím menší je rozptyl statistiky $D[T]$, tím je vydatnější (přesnější):

$$D[T] \Downarrow \quad \text{přesnost odhadu} \Uparrow$$

Příklad: dvě nestranné statistiky T_1 a T_2

$$\text{pokud} \quad D[T_1] < D[T_2], \qquad T_1 - \text{vydatnější}$$

Význam: Nestranná statistika s menším rozptylem je přesnější

Shrnutí přednášky:

- v praxi pracujeme s náhodným výběrem
- pomocí statistik z výběru odhadujeme charakteristiky souboru
- statistiky musí mít tři vlastnosti:
  - nestrannost
  - konzistence
  - vydatnost