Výběrové charakteristiky

Martina Litschmannová

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY | KATEDRA APLIKOVANÉ MATEMATIKY

Obsah (1)

- Opakování: Statistika – základní pojmy
- Princip statistické indukce
- Populační parametry vs. výběrové charakteristiky
- Výběrové charakteristiky a jejich modelování
  - Výběrový průměr
    - Slabý zákon velkých čísel
    - Centrální limitní věta
    - Modelování průměrů výběrů velkých rozsahů
    - Modelování průměrů výběrů malých rozsahů (Studentovo rozdělení)
  - Výběrový úhrn
  - Výběrový rozptyl, popř. výběrová směrodatná odchylka (Chí-kvadrát rozdělení)
  - Výběrový podíl

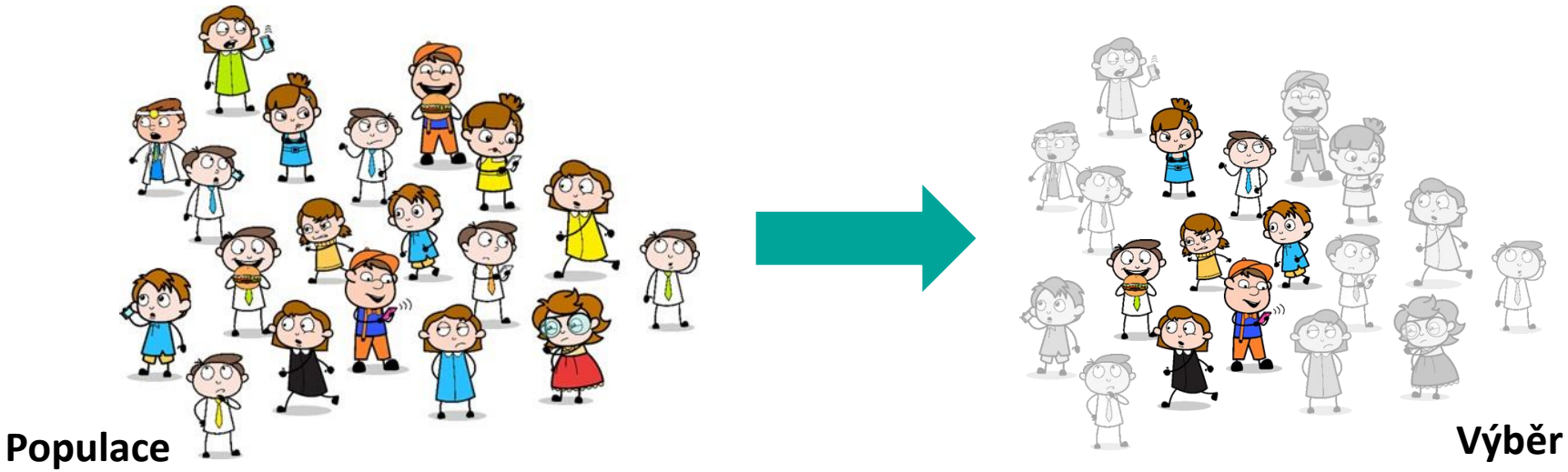
Obsah (2)

- Rozdíl (podíl) výběrových charakteristik a jejich modelování
  - Rozdíl výběrových průměrů
    - známé rozptyly σ_1^2, σ_2^2 nebo $n_1 > 30$, $n_2 > 30$
    - neznámé rozptyly σ_1^2, σ_2^2, ale víme, že $\sigma_1^2 = \sigma_2^2$
    - neznámé rozptyly σ_1^2, σ_2^2, ale víme, že $\sigma_1^2 \neq \sigma_2^2$
  - Podíl rozptylů (Fischerovo – Snedecorovo rozdělení)
  - Rozdíl výběrových podílů

Opakování: Statistika – základní pojmy

- **Populace** (základní soubor) je soubor nějakých prvků, o kterém chceme statistickými metodami něco vypovídat. Definuje se výčtem nebo pomocí zvolené vlastnosti. O každém prvku umíme rozhodnout, zda do populace patří či nikoliv.

- **Výběr** je část dané populace, která má sloužit k odvození závěrů platných pro celou populaci.
  (Pozor na reprezentativnost výběru!)

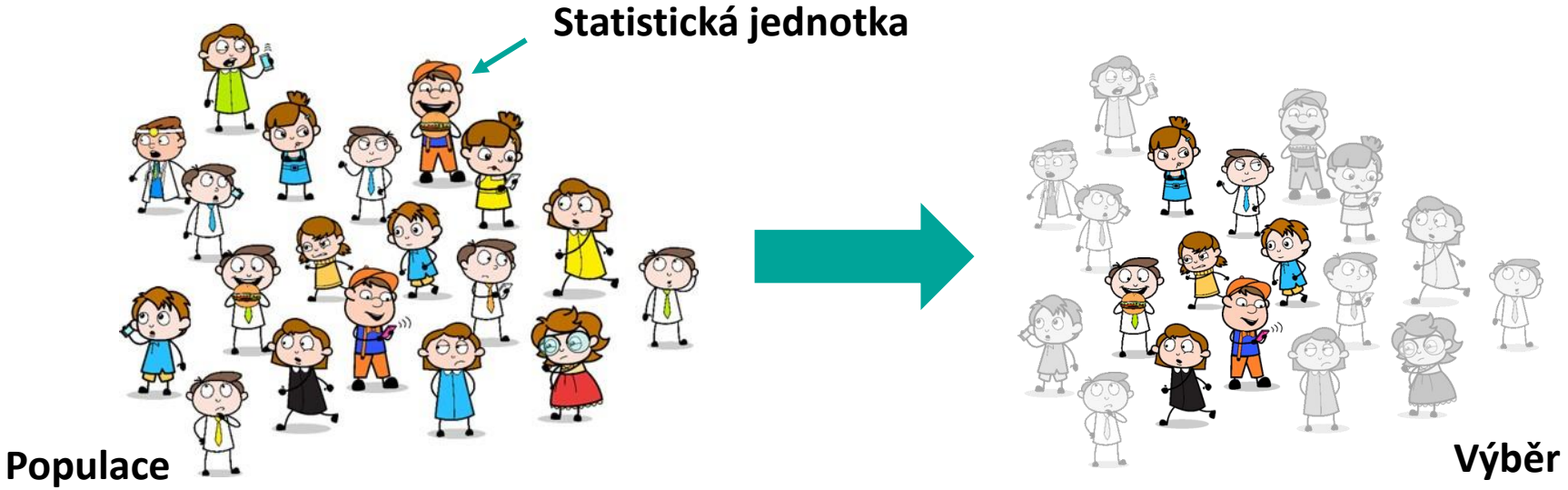


***Zdroj**: https://bazant.wordpress.com/2019/07/23/zaklady-statistiky-cast-1/*

Opakování: Statistika – základní pojmy

- **Statistická jednotka** je prvek populace.

- **Statistický znak (proměnná)** je nějaká měřitelná (zjistitelná) charakteristika statistické jednotky (hmotnost, pohlaví, …).

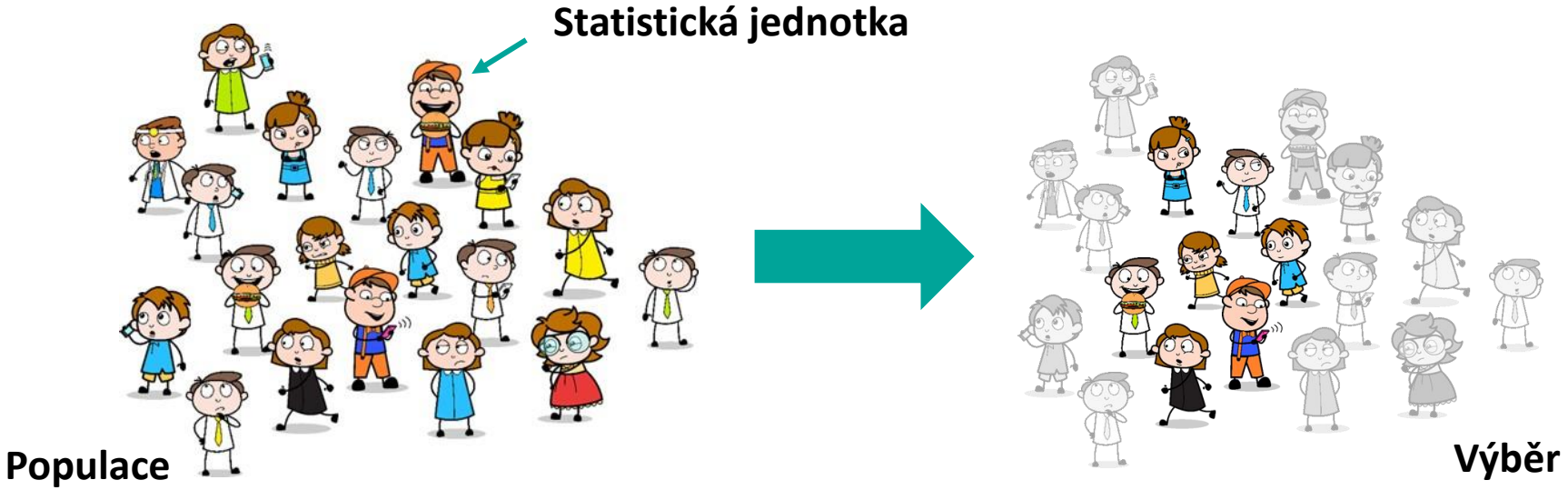


***Zdroj**: https://bazant.wordpress.com/2019/07/23/zaklady-statistiky-cast-1/*

Opakování: Statistika – základní pojmy

Proč provádíme výběrová šetření?

- Omezené zdroje (lidské, finanční, časové)
- Detruktivní zkoušky
- Důvěryhodnost získaných dat (sběr dat u menšího výběru lze provést lépe proškolenými lidmi…)

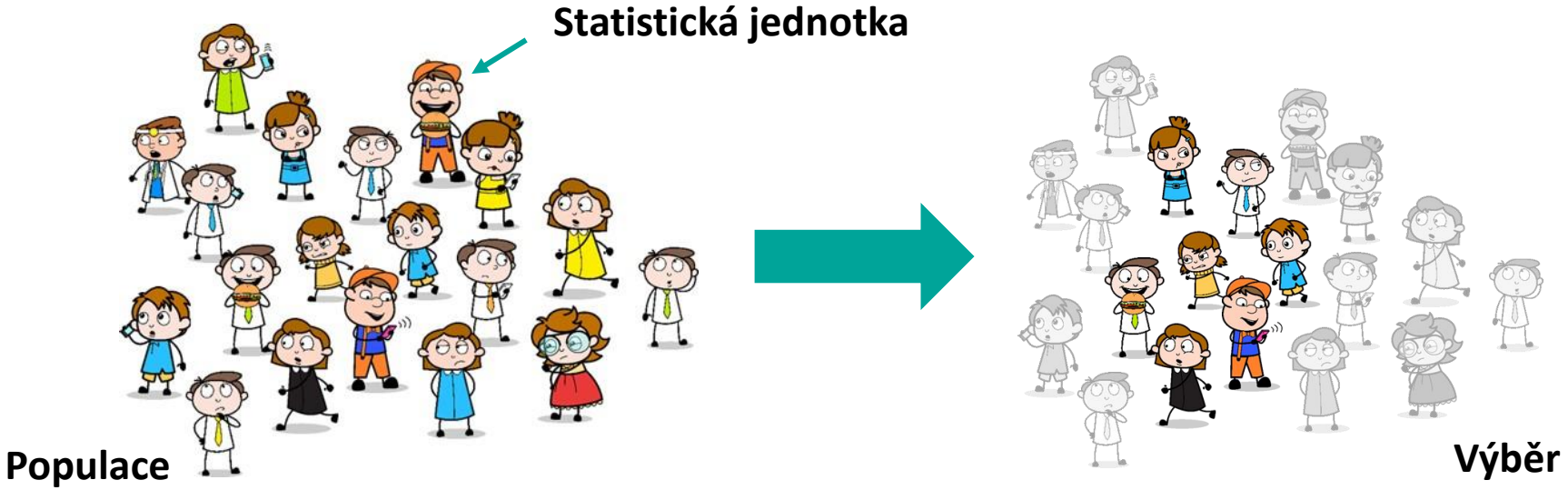


***Zdroj**: https://bazant.wordpress.com/2019/07/23/zaklady-statistiky-cast-1/*

Opakování: Statistika – základní pojmy

- **Explorační analýza** - zjišťuje a sumarizuje informace, zpracovává je ve formě grafů a tabulek

- **Statistická indukce** (inferenční statistika) – na základě informací zjištěných z výběrových šetření predikuje (odhaduje) závěry platné pro celou populaci.

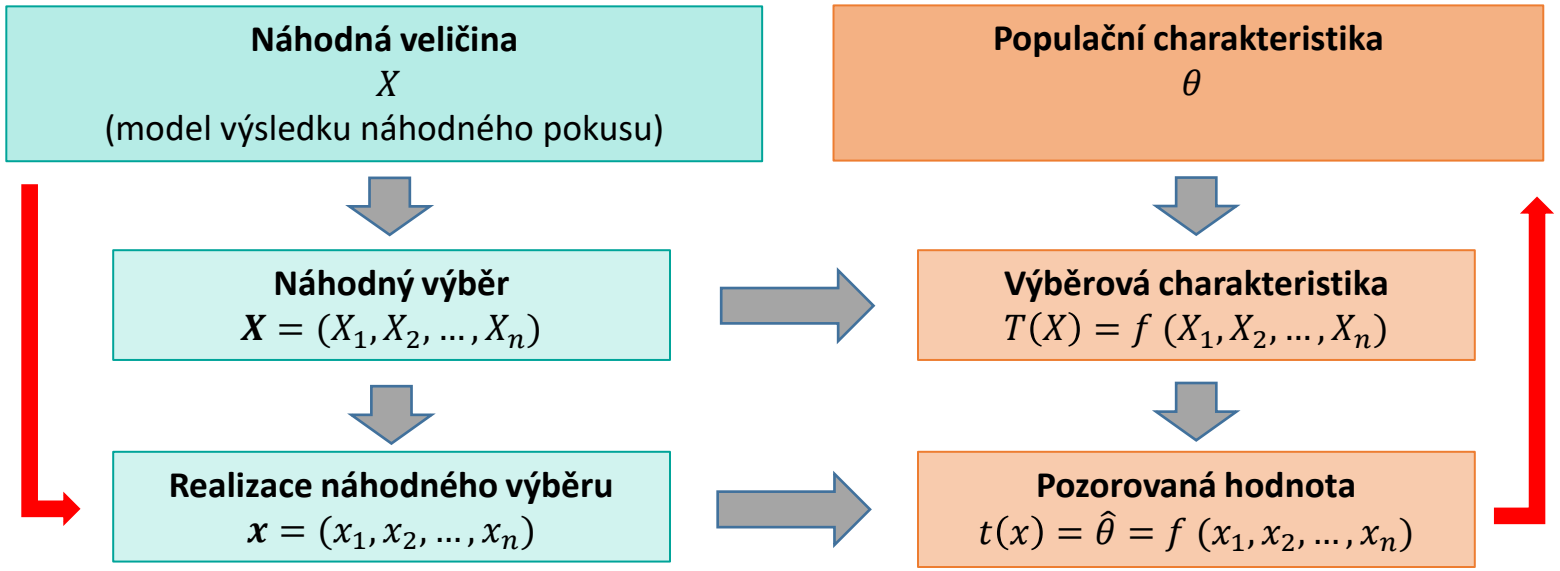


***Zdroj**: https://bazant.wordpress.com/2019/07/23/zaklady-statistiky-cast-1/*

Princip statistické indukce

- **Náhodný výběr** $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ je vektor náhodných veličin, které jsou nezávislé a mají stejné rozdělení.

- **Realizace náhodného výběru** $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ - u každé jednotky, která je vybrána do výběrového souboru, zjistíme hodnotu zkoumaného znaku x_i $(i = 1,2, \dots, n)$. Tuto hodnotu můžeme chápat jako jednu z možných hodnot náhodné veličiny X_1. Často označujeme také jako pozorování nebo vstupní data.

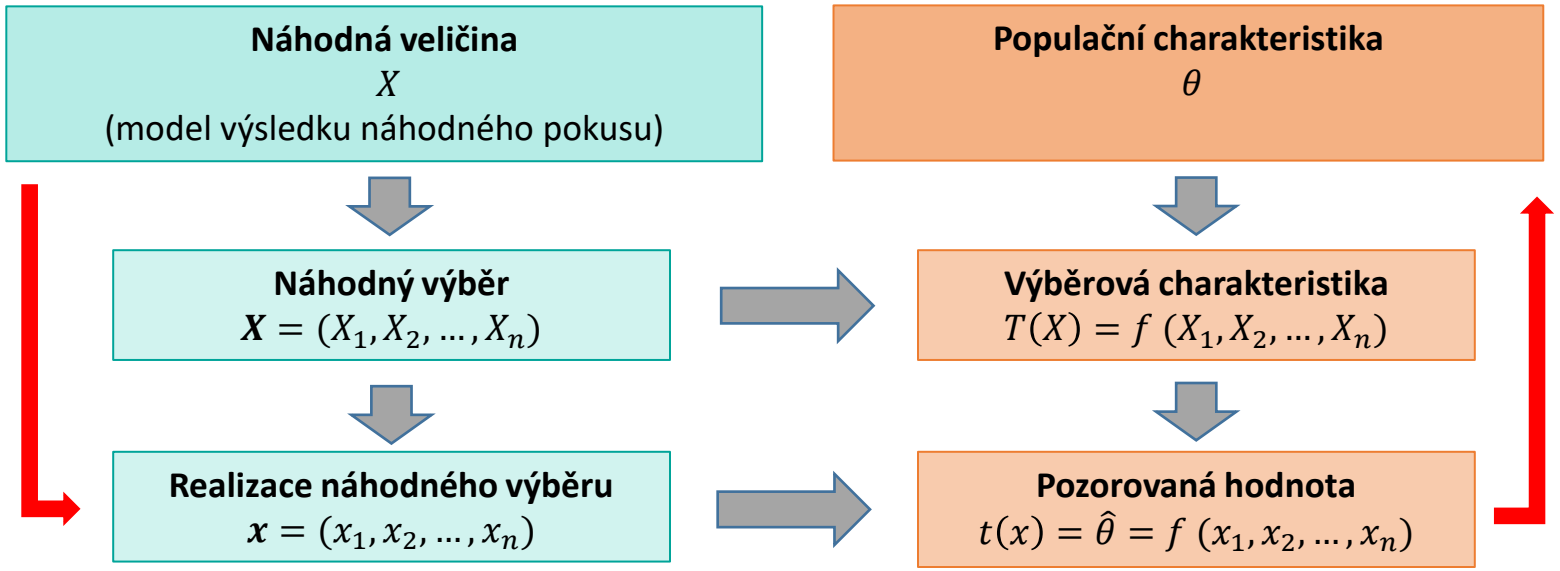


Princip statistické indukce

- **Náhodný výběr** $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ je vektor náhodných veličin, které jsou nezávislé a mají stejné rozdělení.

- **Výběrová charakteristika** $T(\boldsymbol{X})$ je funkce náhodných veličin $X_1, X_2, \dots, X_n$, tj. je také náhodnou veličinou.

- **Pozorovaná hodnota** $t(\boldsymbol{x})$ je konkrétní realizací výběrové charakteristiky $T(\boldsymbol{X})$**.** (reálné číslo)



Populační parametry vs. výběrové charakteristiky

- **Populační charakteristiky** (obecně značíme θ, konkrétně většinou řeckými písmeny) – jsou konstanty, ale většinou nedokážeme určit jejich přesnou hodnotu

- **Výběrové charakteristiky** (obecně značíme $T(X)$, konkrétně latinkou, velkými písmeny) jsou náhodnými veličinami a na základě realizace náhodného výběru dokážeme určit jejich **pozorované hodnoty** (značíme latinkou, malými písmeny)

Populační parametry	**stř. hodnota** μ nebo $E(X)$	**medián** $x_{0,5}$	**rozptyl** σ^2 nebo $D(X)$	**směr. odchylka** σ nebo $\sigma(X)$	**pravděpodobnost** π
Výběrové charakteristiky	**(výběrový) průměr** $\bar{X}$	**výběrový medián** $\tilde{X}_{0,5}$	**výběrový rozptyl** S^2	**výběrová směr. odchylka** S	**rel. četnost** P

Populační parametry vs. výběrové charakteristiky

- **Populační charakteristiky** (obecně značíme θ, konkrétně většinou řeckými písmeny) – jsou konstanty, ale většinou nedokážeme určit jejich přesnou hodnotu

- **Výběrové charakteristiky** (obecně značíme $T(X)$, konkrétně latinkou, velkými písmeny) jsou náhodnými veličinami a na základě realizace náhodného výběru dokážeme určit jejich **pozorované hodnoty** (značíme latinkou, malými písmeny)

Populační parametry	**stř. hodnota** μ nebo $E(X)$	**medián** $x_{0,5}$	**rozptyl** σ^2 nebo $D(X)$	**směr. odchylka** σ nebo $\sigma(X)$	**pravděpodobnost** π
Výběrové charakteristiky	**(výběrový) průměr** $\bar{X}$	**výběrový medián** $\tilde{X}_{0,5}$	**výběrový rozptyl** S^2	**výběrová směr. odchylka** S	**rel. četnost** P
Pozorované hodnoty výběrových charakteristik	$\bar{x}$ nebo $\hat{\mu}$	$\hat{x}_{0,5}$	s^2 nebo $\hat{\sigma}^2$	s nebo $\hat{\sigma}$	p nebo $\hat{\pi}$

(Výběrový) průměr

- **(Výběrový) průměr** (náhodná veličina)

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

- **Pozorovaná hodnota (výběrového) průměru** (reálné číslo určené z konkrétní realizace náh. výběru)

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

- Chceme-li zdůraznit, že se jedná o výb. charakteristiku (pozorovanou hodnotu výb. charakteristiky) výběru o rozsahu n, můžeme pro to využít dolní index: $\bar{X}_n$ ($\bar{x}_n$).
- Opakováním realizací náhodného výběru o rozsahu n můžeme získat libovolný počet pozorovaných hodnot náhodné veličiny $\bar{X}$.
- Vyneseme-li tyto pozorované hodnoty např. do histogramu, můžeme získat představu o rozdělení této náhodné veličiny…

Příklad 1 (motivační)

Jak můžeme empiricky modelovat výběrové charakteristiky?

Navrhněte, jak bychom mohli zjišťovat, jaké rozdělení má (výběrový) průměr hladin cholesterolu 5 osob.

- Označme si hledaný průměr $\bar{X}_5$ a uvědomme si, že se jedná o náhodnou veličinu.
- Jakých hodnot může tato náhodná veličina nabývat? Pokud nemáme žádnou představu, ale můžeme provádět testy na lidech (představme si, že ano ☺), můžeme to zjistit experimentálně (empiricky):
  - ✓ Náhodně vybereme z populace 5 lidí, změříme jim hladinu cholesterolu a určíme průměr z těchto hodnot. Tím získáme $\bar{x}_1$. (Nechť $\bar{x}_1 = 4{,}75\ mmol/l$)
  - ✓ Chceme-li získat lepší představu, pokus opakujeme, tj. znovu vybereme z populace náhodně 5 lidí, změříme jim hladinu cholesterolu a určíme průměr z těchto hodnot. (Nechť $\bar{x}_2 = 5{,}15\ mmol/l$)
  - ✓ Je zřejmé, že čím vícekrát pokus opakujeme, tím lepší představu o chování $\bar{X}_5$ získáme…
  - ✓ Pro lepší představu o rozdělení $\bar{X}_5$ si vynesme získané pozorované hodnoty do krabicového grafu, popř. do histogramu…

Příklad 1 (motivační)

- Mějme fiktivní populaci 10 000 osob se známou průměrnou hladinou cholesterolu v krvi (5 mmol/l). Naším úkolem je zjistit, jaké rozdělení má (výběrový) průměr hladin cholesterolu 5 osob.

- V praxi samozřejmě populaci nemáme k dispozici a proto neznáme ani střední hodnotu zkoumané náhodné veličiny. Hlavním cílem by pak bylo tuto střední hodnotu co nejlépe odhadnout.

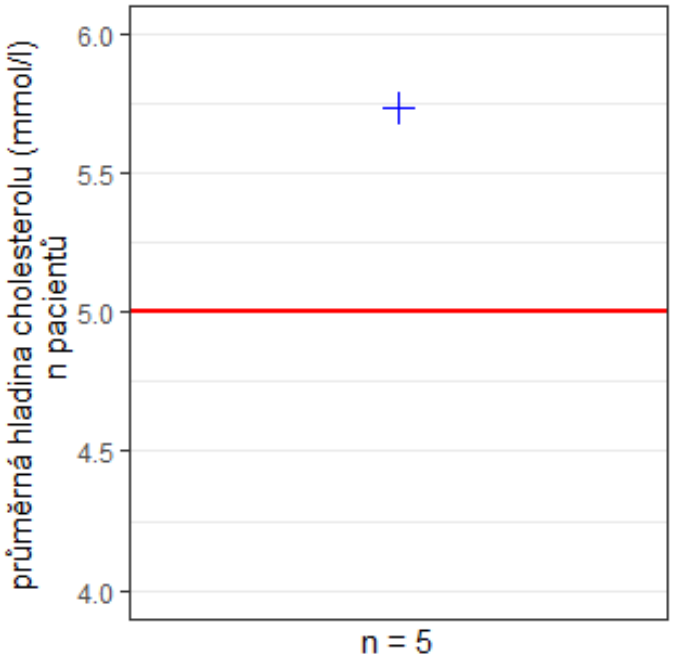
- Chceme-li si ukázat, jak „to funguje“ a přitom se vyhnout testům prováděných na lidech (odběry krve pro zjištění hladiny cholesterolu), musíme si
  - ✓ nasimulovat fiktivní populaci (datový soubor obsahující 10 000 fiktivních hodnot hladin cholesterolu) a
  - ✓ z nich opakovaně provádět náhodný výběr o daném rozsahu, nyní o rozsahu 5 a
  - ✓ vždy vypočíst příslušnou pozorovanou hodnotu průměru…

Příklad 1 (motivační)

- **Fiktivní populace 10 000 osob:**

  X … hladina cholesterolu v krvi (mmol/l)

  $E(X) = 5$



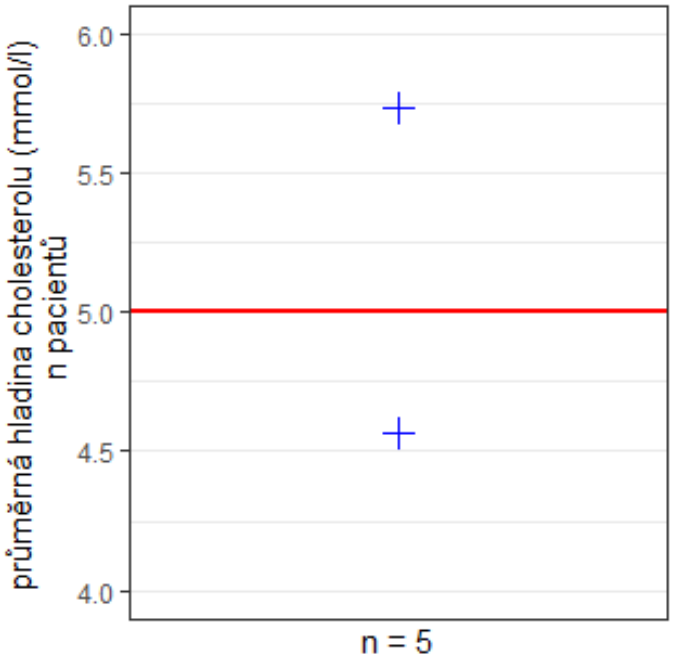
- $\bar{x}_1$

Příklad 1 (motivační)

- **Fiktivní populace 10 000 osob:**

  X … hladina cholesterolu v krvi (mmol/l)

  $E(X) = 5$



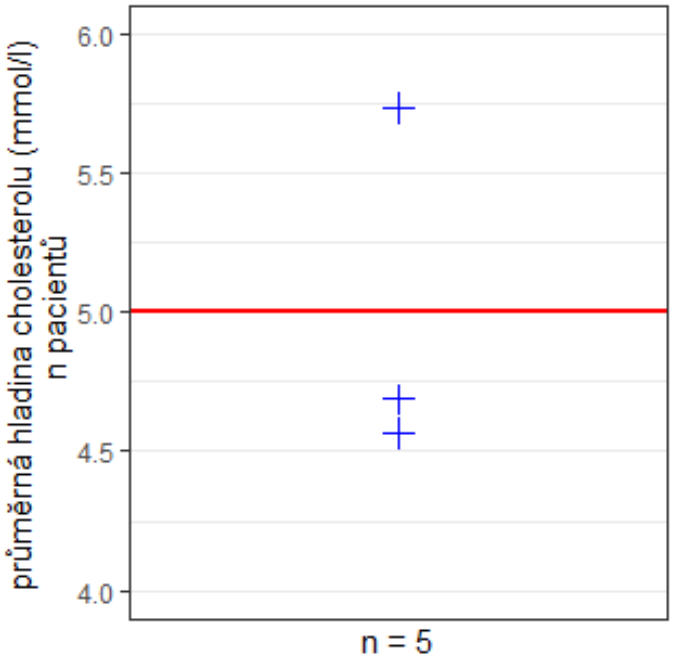
- $\bar{x}_1, \bar{x}_2$

Příklad 1 (motivační)

- **Fiktivní populace 10 000 osob:**

  X … hladina cholesterolu v krvi (mmol/l)

  $E(X) = 5$



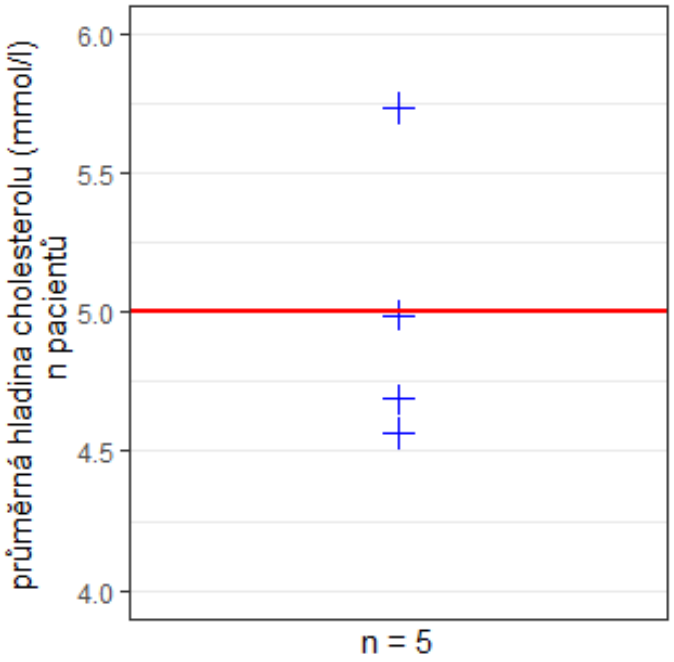
- $\bar{x}_1, \bar{x}_2, \bar{x}_3$

Příklad 1 (motivační)

- **Fiktivní populace 10 000 osob:**

  X … hladina cholesterolu v krvi (mmol/l)

  $E(X) = 5$



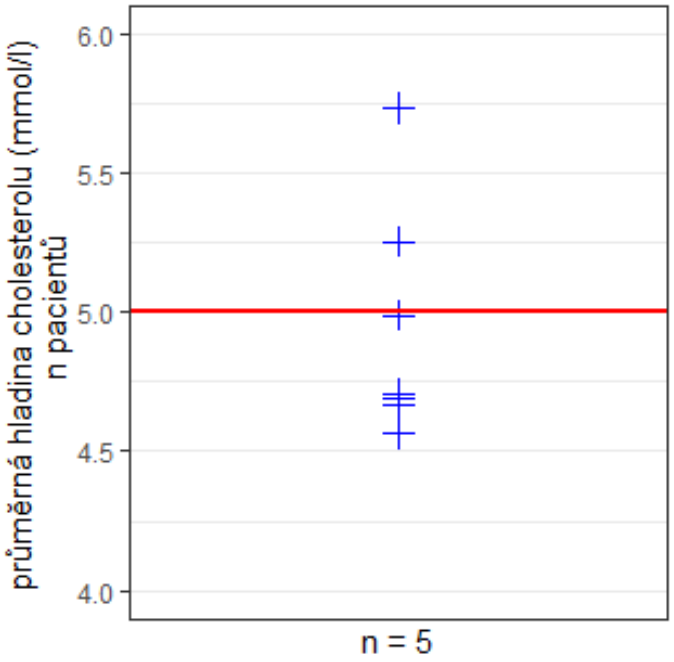
- $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4$

Příklad 1 (motivační)

- **Fiktivní populace 10 000 osob:**

  X … hladina cholesterolu v krvi (mmol/l)

  $E(X) = 5$



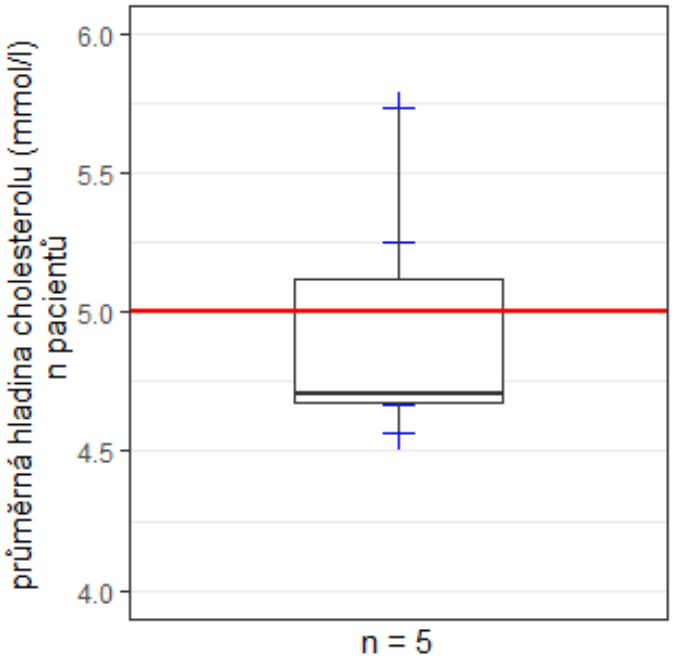
- $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \ldots, \bar{x}_{10}$

Příklad 1 (motivační)

- **Fiktivní populace 10 000 osob:**

  X … hladina cholesterolu v krvi (mmol/l)

  $E(X) = 5$



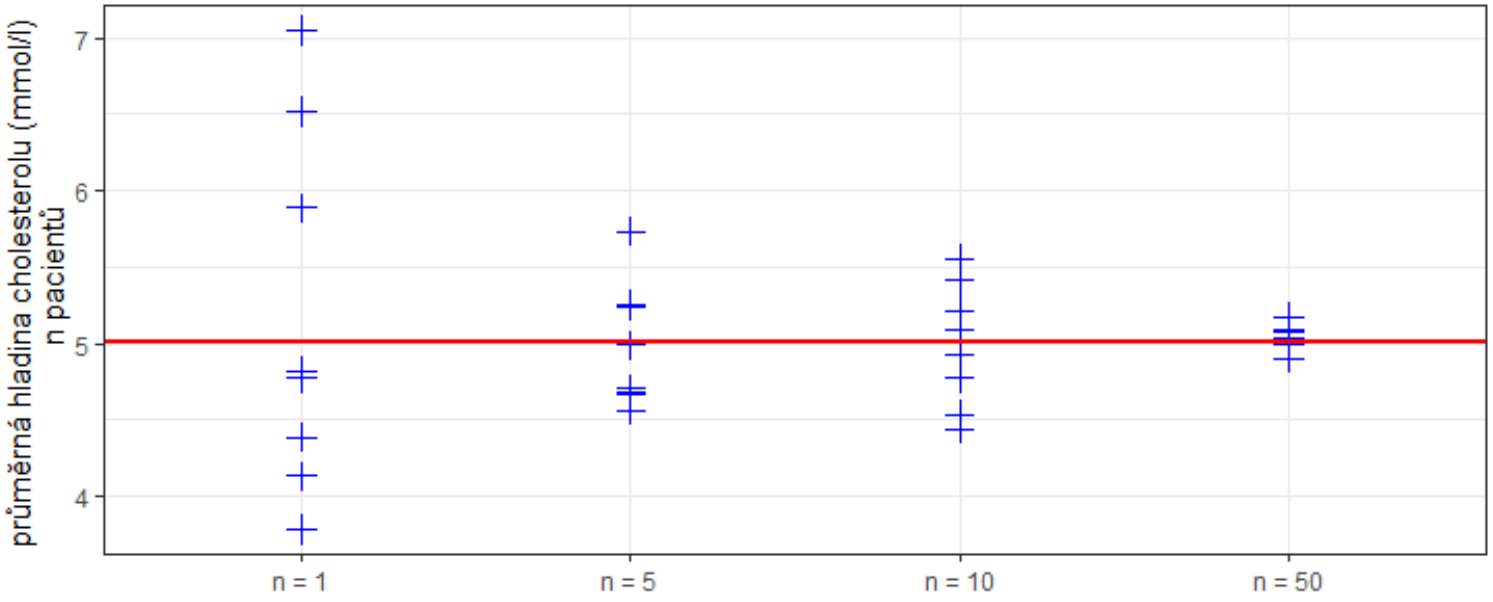
- $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \ldots, \bar{x}_{10}$

Příklad 1 (motivační)

- **Fiktivní populace 10 000 osob:**

  X … hladina cholesterolu v krvi (mmol/l)

  $E(X) = 5$



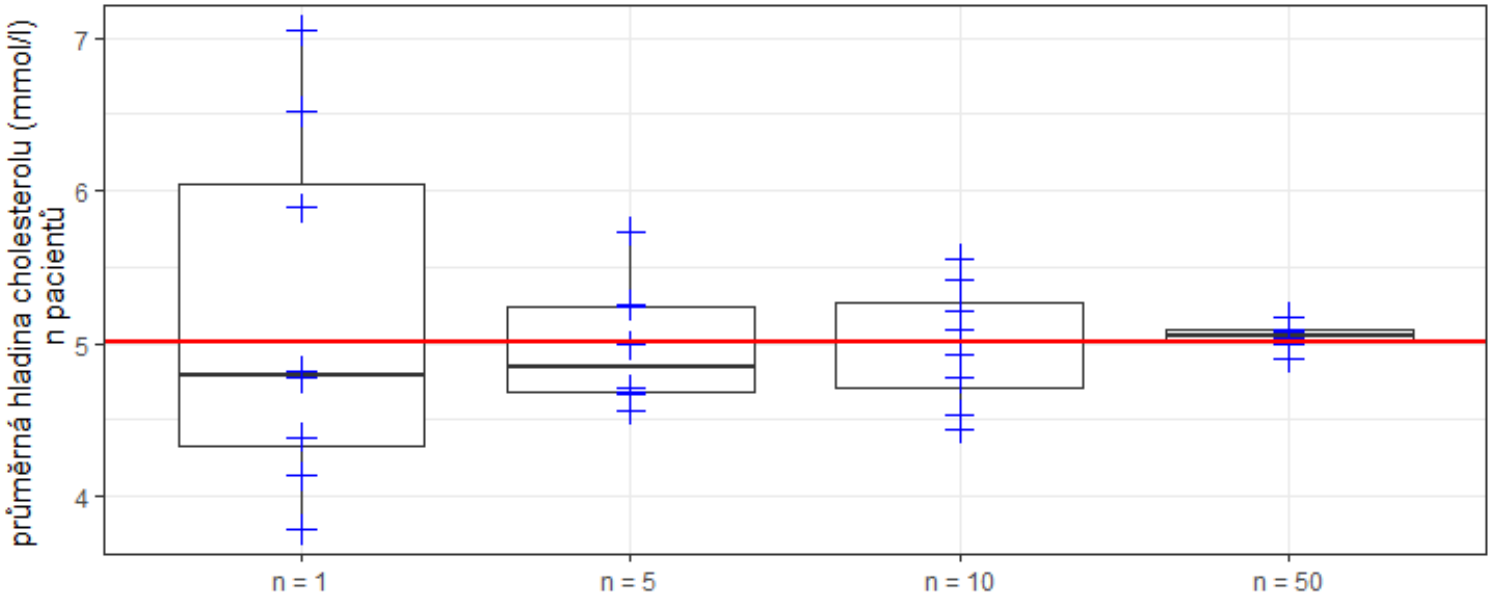
- Srovnejme výsledky simulací pro opakované výběry 1 lidí, 5 (10, popř. 50) lidí (vždy 10 opakování výběru).

Příklad 1 (motivační)

- **Fiktivní populace 10 000 osob:**

  X … hladina cholesterolu v krvi (mmol/l)

  $E(X) = 5$



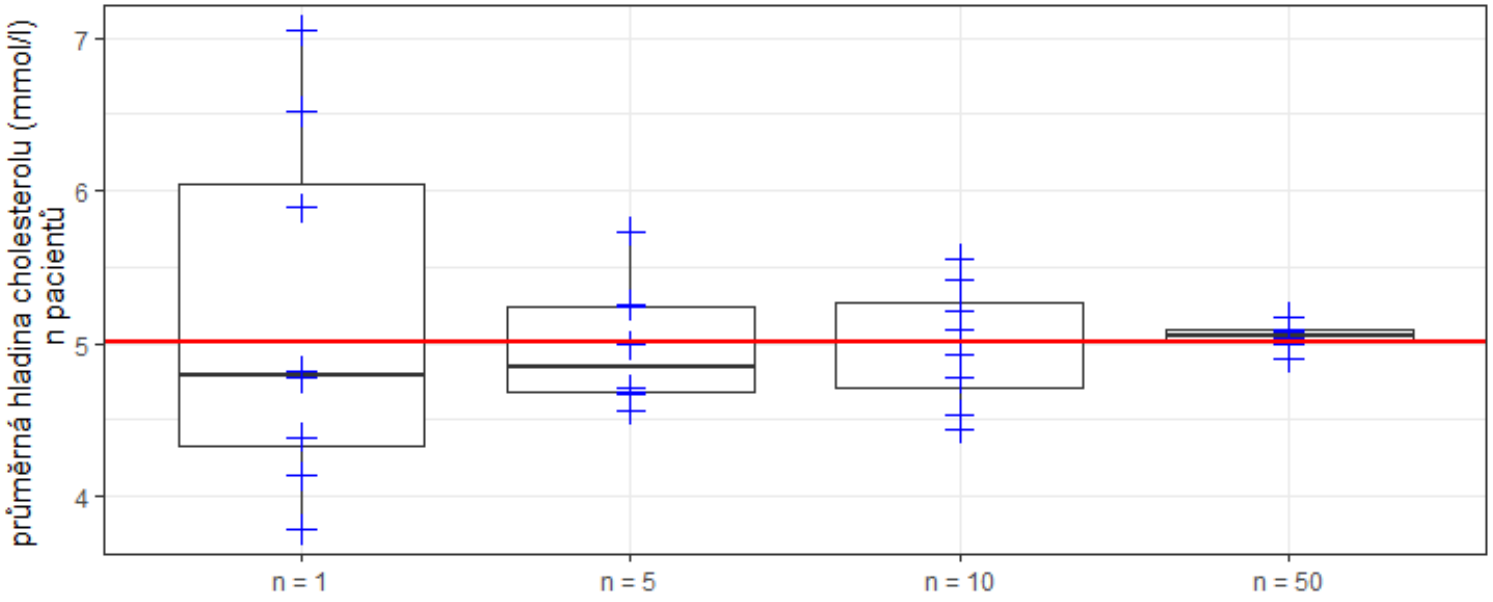
- Srovnejme výsledky simulací pro opakované výběry 1 lidí, 5 (10, popř. 50) lidí (vždy 10 opakování výběru).

Příklad 1 (motivační)

- **Fiktivní populace 10 000 osob:**

  X … hladina cholesterolu v krvi (mmol/l)

  $E(X) = 5$

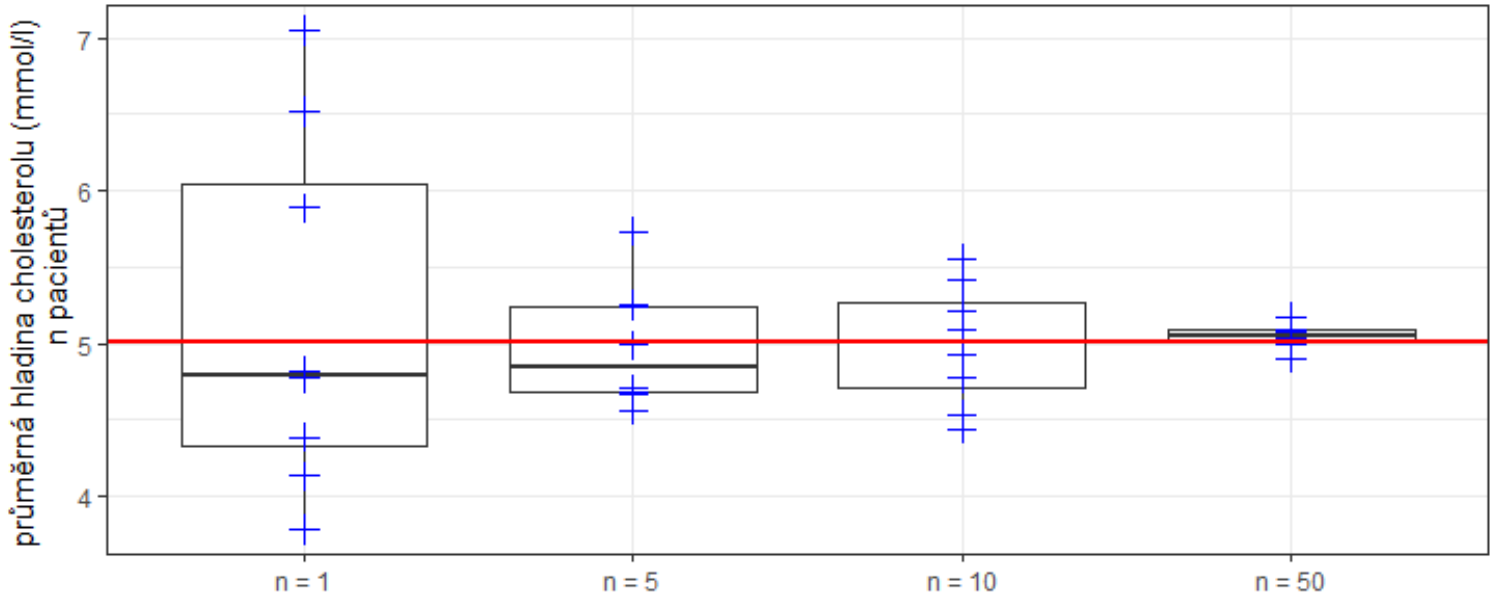


- **Závěr**: Čím větší rozsah výběru, tím lépe se v každém pokusu dařilo odhadnout skutečnou střední hodnotu.

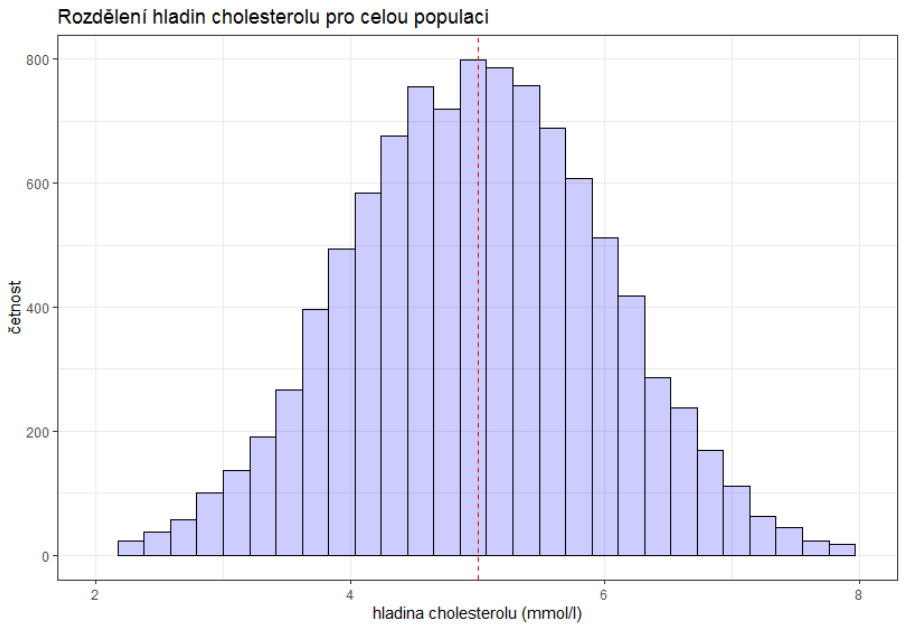
Slabý zákon velkých čísel

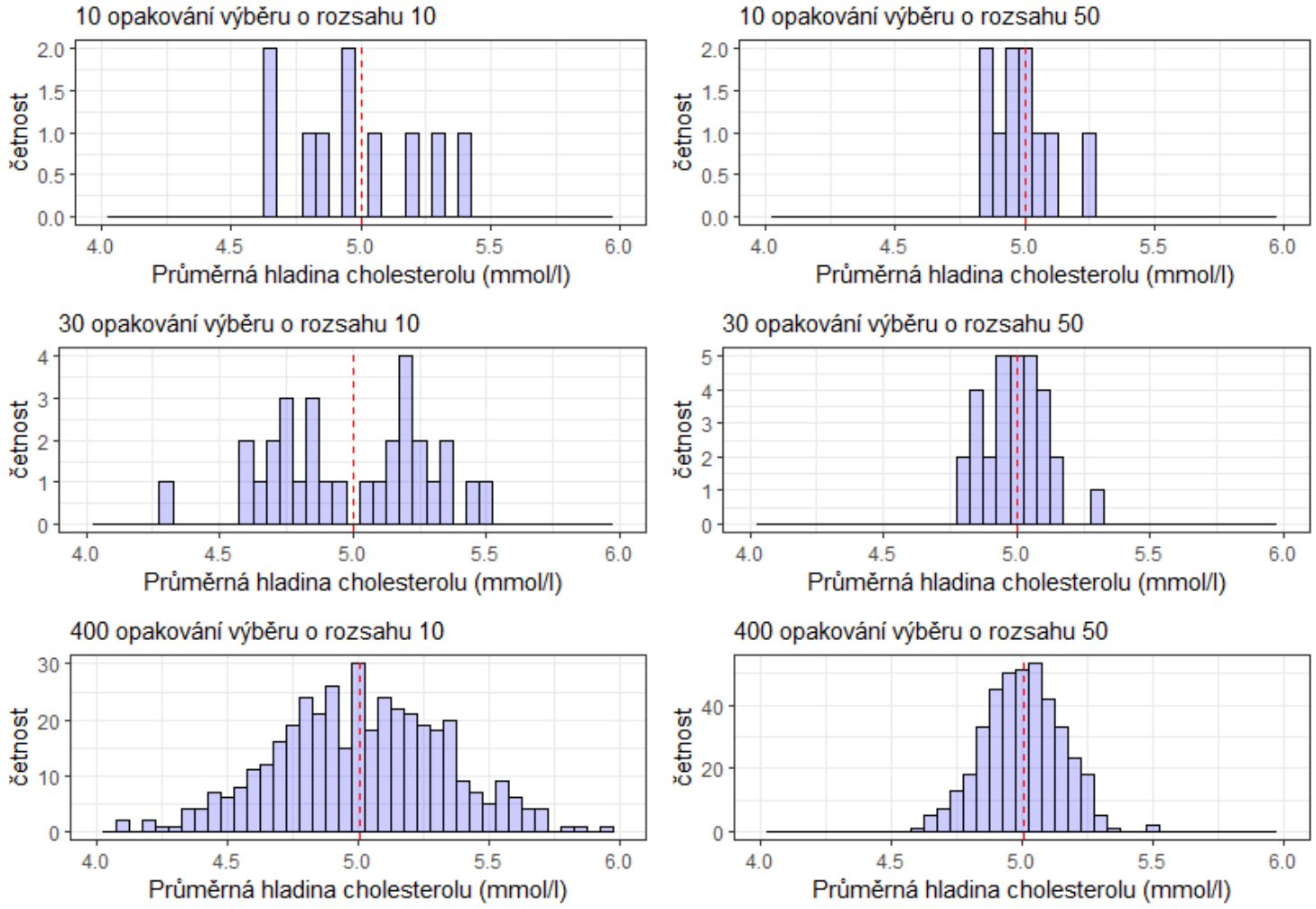
Mějme nekonečný náhodný výběr $X_1, X_2, \ldots$ z rozdělení se střední hodnotou μ_X a konečným rozptylem, kde $X_1, X_2, \ldots$ jsou nekorelované náhodné veličiny. Potom platí, že výběrový průměr $\bar{X}_n$ vypočítaný z prvních n pozorování se pro $n \to \infty$ blíží ke střední hodnotě μ_X, což zapisujeme

$$\lim_{n\to\infty}\left[P\left(\left|\bar{X}_n - \mu_X\right| > \varepsilon\right)\right] = 0 \text{ pro každé } \varepsilon > 0.$$



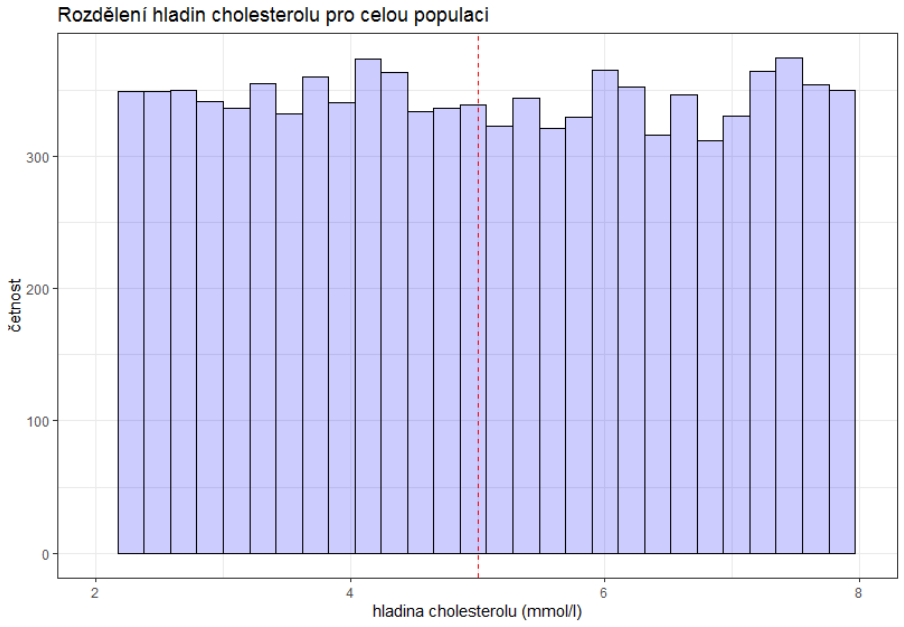
Příklad 1 (motivační)

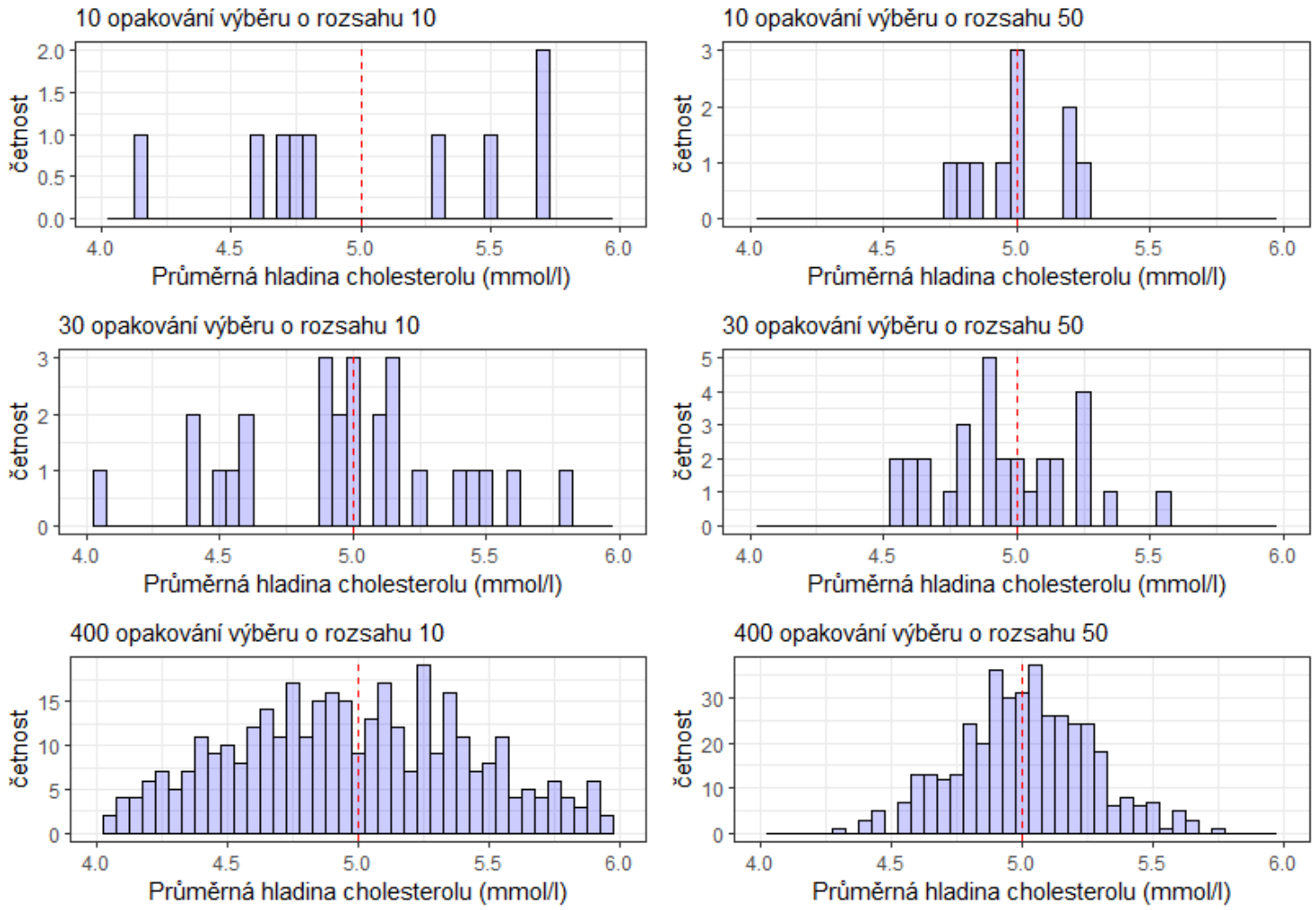




Jak dopadla simulace,
když jsme použili fiktivní populaci,
v níž hladiny cholesterolu měly cca
normální rozdělení?

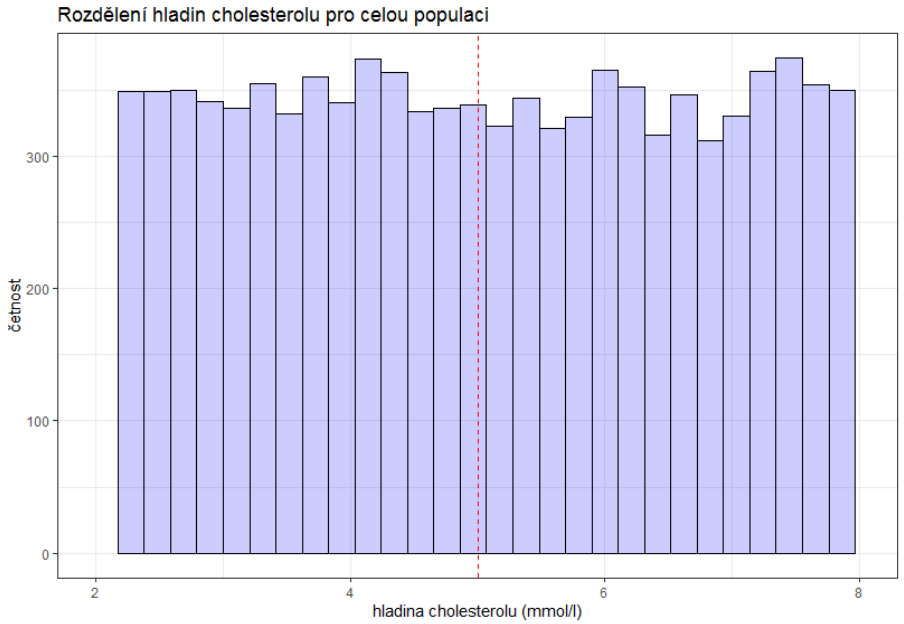
Příklad 1 (motivační)

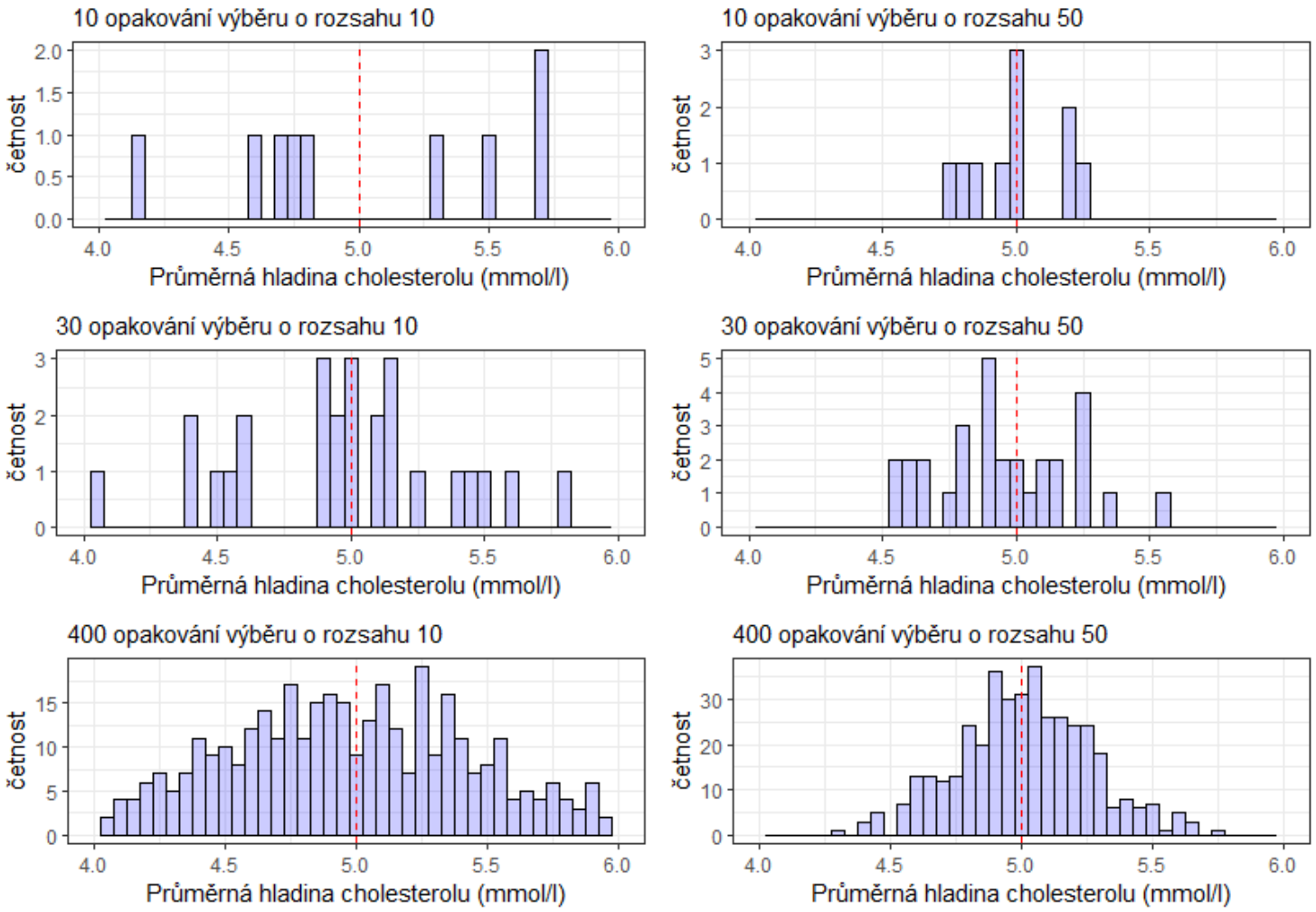




Jak dopadla simulace,
když jsme použili fiktivní populaci,
v níž hladiny cholesterolu měly cca
rovnoměrné rozdělení?

Příklad 1 (motivační)





Závěr:
Zdá se, že průměr má normální rozdělení bez ohledu na rozdělení populace.

Centrální limitní věta (CLV)

Jsou-li X_i $(i = 1, 2, \dots, n)$ nezávislé náhodné veličiny
se stejnou střední hodnotou μ_X a se stejným konečným rozptylem σ_X^2,
pak výběrový průměr $\bar{X}_n$ má **při dostatečně velkém počtu pozorování** přibližně normální rozdělení,
ať už X_i pocházejí z libovolného rozdělení.

Centrální limitní větu zapisujeme

$$\bar{X}_n \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right) \text{ nebo } \frac{\bar{X}_n - \mu_X}{\sigma_X}\sqrt{n} \sim N(0,1).$$

Předpoklady CLV:

- X_i nezávislé náhodné veličiny,
- $E(X_1) = E(X_2) = \cdots = E(X_n) = \mu_X$,
- $D(X_1) = D(X_2) = \cdots = D(X_n) = \sigma_X^2$; $\sigma_X^2 < \infty$,
- $n \to \infty$ (v praxi: $n > 30$, výběr neobsahuje odlehlé pozorování).

Centrální limitní věta

$$\bar{X}_n \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right) \text{ nebo } \frac{\bar{X}_n - \mu_X}{\sigma_X}\sqrt{n} \sim N(0,1)$$

Předpoklady CLV:

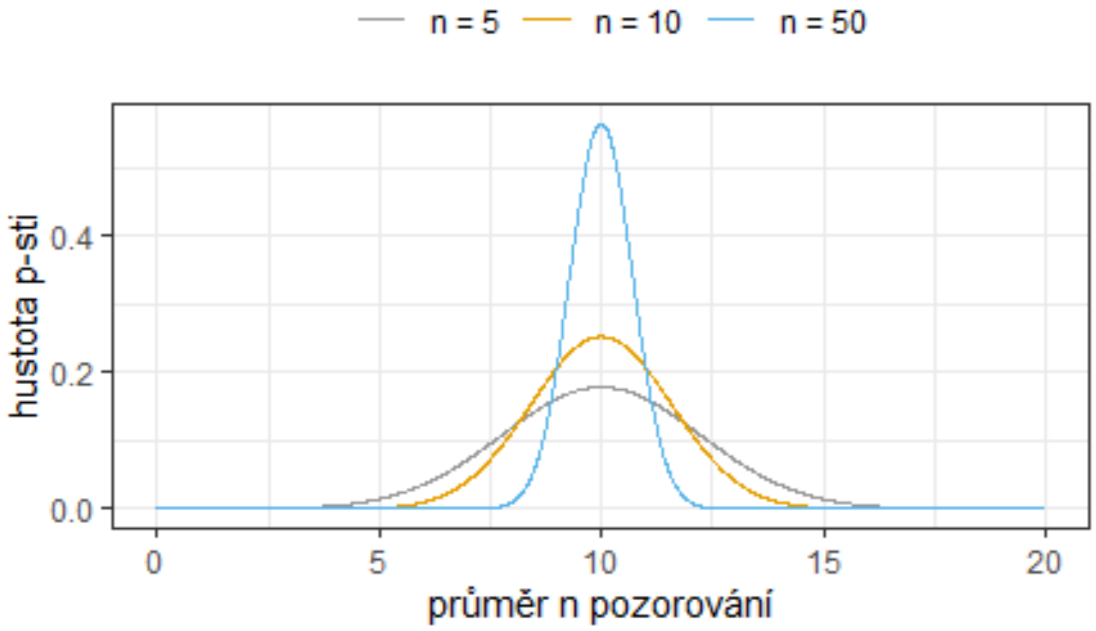
- X_i nezávislé náhodné veličiny,
- $E(X_1) = E(X_2) = \cdots = E(X_n) = \mu_X$,
- $D(X_1) = D(X_2) = \cdots = D(X_n) = \sigma_X^2; \sigma_X^2 < \infty$,
- $n \to \infty$ (v praxi: $n > 30$, výběr neobsahuje odlehlé pozorování).

Proč platí CLV?

- **Věta 1**: Jsou-li X_i nezávislé náh. veličiny, pak má náhodná veličina $X_1 + X_2 + \cdots + X_n = \sum_{i=1}^n X_i$ normální rozdělení.
- **Věta 2**: Má-li náhodná veličina X normální rozdělení, pak náhodná veličina $Y = aX + b$, kde $a,\ b \in \mathbb{R}$, má normální rozdělení.
- $E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots E(X_n) = n\mu_X$.
- Jsou-li X_i nezávislé náh. veličiny, pak $D(X_1 + X_2 + \cdots + X_n) = D(X_1) + D(X_2) + \cdots D(X_n) = n\sigma_X^2$.
- $\sum_{i=1}^n X_i \sim \mathrm{N}(n\mu_X, n\sigma_X^2)$
- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \Rightarrow E(\bar{X}_n) = \frac{1}{n}E(\sum_{i=1}^n X_i) = \mu_X; D(\bar{X}_n) = \left(\frac{1}{n}\right)^2 D(\sum_{i=1}^n X_i) = \frac{\sigma_X^2}{n}$ + dle Věta 2: $\bar{X}_n \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$.
- $X \sim N(\mu, \sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$, tj. $\bar{X}_n \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right) \Rightarrow \frac{\bar{X}_n - \mu_X}{\sigma_X}\sqrt{n} \sim N(0,1)$.

Centrální limitní věta

$$\bar{X}_n \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$



Vliv rozsahu výběru na graf hustoty pravděpodobnosti výběrového průměru

Jak modelovat průměr u výběrů „velkého“ rozsahu?

Za výběry dostatečně velkého rozsahu považujeme takové výběry, kde $n > 30$.

- Známe-li střední hodnotu a rozptyl populace, lze použít CLV, tj. znalosti, že $\bar{X}_n \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$.
- Má-li populace normální rozdělení a neznáme jeho rozptyl a rozsah výběru je velký $(n > 30)$, výběrová směrodatná odchylka je dobrým odhadem populační směr. odchylky, tj. lze předpokládat, že $\sigma_X^2 \cong s_X^2$. Dále lze využít znalosti, že $\bar{X}_n \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$.

Příklad 2 (modelování průměru pro velké rozsahy výběru)

Doba přežití jistého typu pacientů má exponenciální rozdělení se střední hodnotou 2 roky. Určete p-st, že

a) doba přežití pacienta bude vyšší než 27 měsíců,

Řešení:

X … doba přežití pacienta (měs.)

$X \sim Exp\left(\lambda = \frac{1}{24}\right)$

$P(X > 27) = 1 - F(27) = e^{-\frac{27}{24}} \cong \mathbf{0,325}$ (*1-pexp(27,1/24)*)

Příklad 2 (modelování průměru pro velké rozsahy výběru)

Doba přežití jistého typu pacientů má exponenciální rozdělení se střední hodnotou 2 roky. Určete p-st, že

b) průměrná doba přežití 150 pacientů bude vyšší než 27 měsíců.

Řešení:

$\bar{X}_{150}$ … průměrná doba přežití 150 pacientů (měs.) – dostatečně velký rozsah výběru $(n > 30)$, lze použít CLV

X_i … doba přežití pacienta (měs.),

$$X_i \sim \mathrm{Exp}(\lambda = \tfrac{1}{24}) \Rightarrow E(X_i) = \frac{1}{\lambda} = 24,\ D(X_i) = \frac{1}{\lambda^2} = 576$$

$$\bar{X}_{150} \sim N\left(\mu = E(X_i), \sigma^2 = \frac{D(X_i)}{150}\right) \quad \Rightarrow \quad \bar{X}_{150} \sim N\left(\mu = 24, \sigma^2 = 3{,}84\right)$$

$$P(\bar{X}_{150} > 27) = 1 - F(27) = 1 - \Phi\left(\frac{27-24}{\sqrt{3{,}84}}\right) = 1 - \Phi(1{,}53) = \mathbf{0{,}063}$$ (*1-pnorm(27,24,sqrt(576/150))*)

Příklad 3 (modelování průměru pro velké rozsahy výběru)

Doba přežití jistého typu pacientů má střední hodnotu 2 roky a směr. odchylku 2 roky. Určete p-st, že

a) doba přežití pacienta bude vyšší než 27 měsíců,

Řešení:

X … doba přežití pacienta (měs.)

$X \sim ?$

$P(X > 27)$ nelze určit, protože neznáme rozdělení doby přežití pacienta.

Příklad 3 (modelování průměru pro velké rozsahy výběru)

Doba přežití jistého typu pacientů má střední hodnotu 2 roky a směr. odchylku 2 roky. Určete p-st, že

b) průměrná doba přežití 150 pacientů bude vyšší než 27 měsíců.

Řešení:

$\bar{X}_{150}$ … průměrná doba přežití 150 pacientů (měs.) – dostatečně velký rozsah výběru $(n > 30)$, lze použít CLV

X_i … doba přežití i-tého pacienta (měs.),

$E(X_i) = 24, \qquad D(X_i) = 24^2 = 576$

$$\bar{X}_{150} \sim N\left(\mu = E(X_i), \sigma^2 = \frac{D(X_i)}{150}\right) \quad \Rightarrow \quad \bar{X}_{150} \sim N\left(\mu = 24, \sigma^2 = 3{,}84\right)$$

$$P(\bar{X}_{150} > 27) = 1 - F(27) = 1 - \Phi\left(\frac{27-24}{\sqrt{3{,}84}}\right) = 1 - \Phi(1{,}53) = \mathbf{0{,}063}$$ (*1-pnorm(27,24,sqrt(576/150))*)

Jak modelovat průměr u výběrů malého rozsahu?

Pokud rozsah výběru je malý (v praxi menší než 30),
nelze pro modelování průměru použít centrální limitní větu!

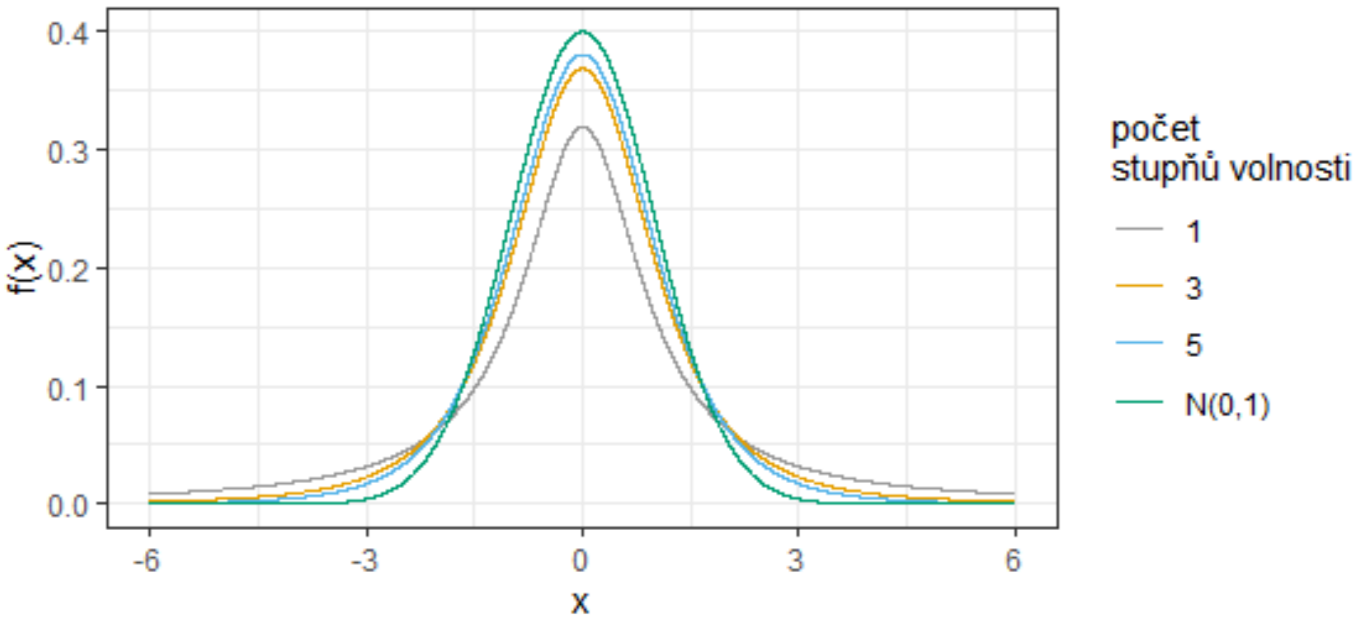
- Má-li populace normální rozdělení a známe jeho rozptyl lze, bez dalších omezujících podmínek, pro modelování průměru použít znalost, že $\bar{X}_n \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$.

- Má-li populace normální rozdělení a neznáme jeho rozptyl a rozsah výběru je malý, výběrová směrodatná odchylka nemusí být dobrým odhadem populační směr. odchylky, proto nelze použít výše uvedenou znalost. Pro modelování průměru lze použít tzv. **Studentovo rozdělení ν stupni volnosti**.

Studentovo rozdělení s ν stupni volnosti (t_ν) - vlastnosti

- Pro $\nu \to \infty$: $X \sim t_\nu \Rightarrow X \approx N(0;1)$
- Pokud náhodné veličiny $X_1, X_2, \dots, X_n$ mají norm. rozdělení $N(\mu, \sigma^2)$ a jsou navzájem nezávislé, lze ukázat, že

$$\boxed{\frac{\bar{X}-\mu}{S}\sqrt{n} \sim t_{n-1}.}$$



Příklad 4 (modelování průměru pro malé rozsahy výběru)

Předpokládejme, že IQ má normální rozdělení se střední hodnotou 100 bodů. Z populace náhodně vybereme 20 lidí. Směrodatná odchylka IQ těchto 20 lidí je 15 bodů. S jakou pravděpodobností průměrné IQ v daném výběru nepřesáhne 110 bodů?

Řešení:

$\bar{X}_{20}$ … průměrné IQ 20 lidí (bod) – malý rozsah výběru $(n < 30)$, **nelze** použít CLV

X_i … IQ i-tého člověka (bod),

$X_i \sim \mathrm{N}\left(\mu = 100, \sigma^2 = ?\right) \Rightarrow$ Platí, že $\frac{\bar{X}-\mu}{S}\sqrt{n} \sim t_{n-1}$.

Nechť $X = \frac{\bar{X}_{20}-\mu}{S}\sqrt{n}$, pak $X \sim t_{19}$.

Obyčejná úprava nerovnice: Nejdříve odečteme μ, poté násobíme $\frac{\sqrt{n}}{S}$.

$$P(\bar{X}_{20} \leq 110) = P\left(\frac{\bar{X}_{20}-\mu}{S}\sqrt{n} \leq \frac{110-\mu}{S}\sqrt{n}\right) = P\left(X \leq \frac{110-100}{15}\sqrt{20}\right) = P(X \leq 2{,}981) = \mathbf{0{,}996}$$

(*pt(2.981,19)*)

Výběrový úhrn

- **Výběrový úhrn** (náhodná veličina)

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

- **Pozorovaná hodnota (výběrového) úhrnu** (reálné číslo určené z konkrétní realizace náh. výběru)

$$\sum_{i=1}^{n} x_i = n \cdot \bar{x}$$

Důsledek centrální limitní věty (viz slide 27):

Nechť:

- ✓ X_i nezávislé náhodné veličiny ,
- ✓ $E(X_1) = E(X_2) = \cdots = E(X_n) = \mu_X$,
- ✓ $D(X_1) = D(X_2) = \cdots = D(X_n) = \sigma_X^2$,
- ✓ $n \to \infty$ (v praxi: $n > 30$, výběr neobsahuje odlehlé pozorování).

pak

$$\boxed{\sum_{i=1}^{n} X_i \sim N\left(n\mu_X, n\sigma_X^2\right) \text{ nebo } \frac{\sum_{i=1}^{n} X_i - n\mu_X}{\sigma_X}\sqrt{n} \sim N(0;1)}$$

Příklad 5 (modelování úhrnu pro velké rozsahy výběru)

Předpokládejme, že průměrná spotřeba elektrické energie na jednoho obyvatele v České republice činí 6,2 MWh (6 200 kWh)/rok na obyvatele, směrodatná odchylka 2,2 MWh.

a) Jakou lze očekávat celkovou roční spotřebu elektrické energie v obci Krmelín, která má 2 365 obyvatel (k 1. 1. 2019)?

b) S jakou p-stí bude celková roční spotřeba el. energie v obci Krmelín menší než 14,5 GWh?

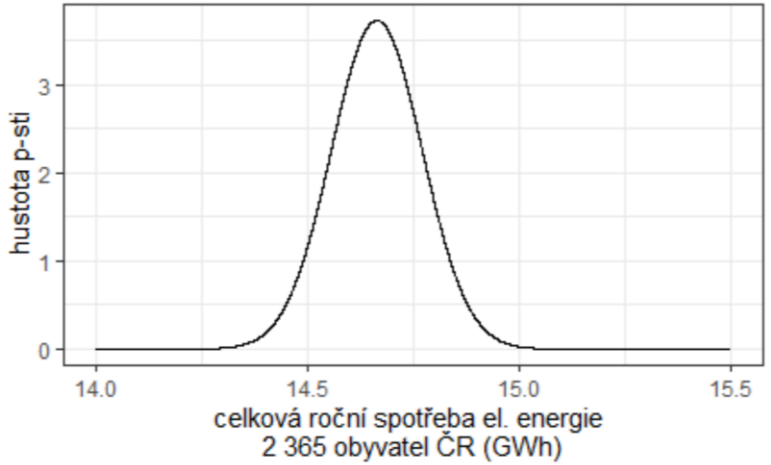
Řešení:

X_i … roční spotřeba el. energie i-tého obyvatele ČR (MWh), $E(X_i) = 6{,}2,\ D(X_i) = 2{,}2^2$

$\sum_{i=1}^{2365} X_i$ … celková roční spotřeba 2 365 obyvatel ČR (MWh) – dostatečně velký rozsah výběru ($n > 30$), lze použít CLV

$$\sum_{i=1}^{2365} X_i \sim N(\mu = 2\,365 \cdot 6{,}2; \sigma^2 = 2\,365 \cdot 2{,}2^2)$$

$$\sum_{i=1}^{2365} X_i \sim N(\mu = 14\,663; \sigma^2 \cong 107^2)$$



ada) Očekávaná celková roční spotřeba 2 365 obyvatel ČR: cca (14,3 – 14,9) GWh.

Příklad 5 (modelování úhrnu pro velké rozsahy výběru)

Předpokládejme, že průměrná spotřeba elektrické energie na jednoho obyvatele v České republice činí 6,2 MWh (6 200 kWh)/rok na obyvatele, směrodatná odchylka 2,2 MWh.

a) Jakou lze očekávat celkovou roční spotřebu elektrické energie v obci Krmelín, která má 2 365 obyvatel (k 1. 1. 2019)?

b) S jakou p-stí bude celková roční spotřeba el. energie v obci Krmelín menší než 14,5 GWh?

Řešení:

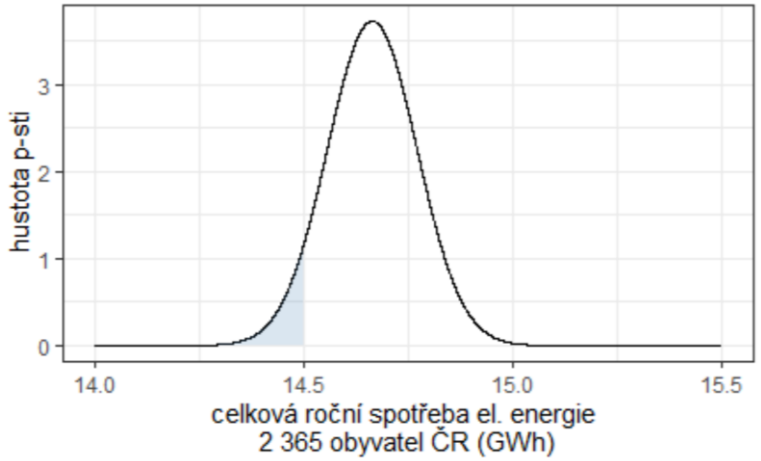
X_i … roční spotřeba el. energie i-tého obyvatele ČR (MWh), $E(X_i) = 6{,}2,\ D(X_i) = 2{,}2^2$

$\sum_{i=1}^{2365} X_i$ … celková roční spotřeba 2 365 obyvatel ČR (MWh) – dostatečně velký rozsah výběru ($n > 30$), lze použít CLV

$\sum_{i=1}^{2365} X_i \sim N(\mu = 2\,365 \cdot 6{,}2; \sigma^2 = 2\,365 \cdot 2{,}2^2)$

$\sum_{i=1}^{2365} X_i \sim N(\mu = 14\,663; \sigma^2 \cong 107^2)$

adb) $P(X < 14500) \cong \mathbf{0{,}064}$ (*pnorm(14500,14663,107)*)



Výběrový rozptyl

- **Výběrový rozptyl** (náhodná veličina)

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

- **Pozorovaná hodnota (výběrového) průměru** (reálné číslo určené z konkrétní realizace náh. výběru)

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

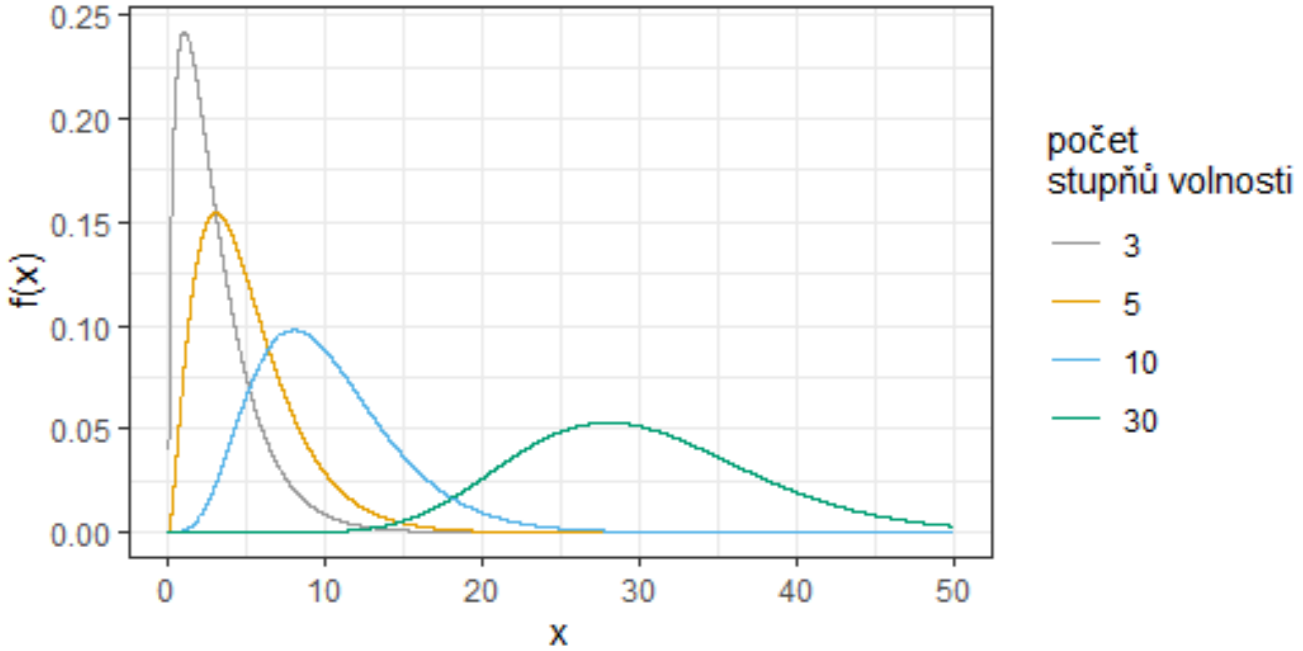
- **POZOR!** Rozdělení výběrového rozptylu neznáme!!!

  Pro úsudky o rozdělení výběrového rozptylu musíme využít tzv. **chí-kvadrát rozdělení s ν stupni volnosti**.

Chí kvadrát rozdělení s ν stupni volnosti

Mějme nezávislé náhodné veličiny $Z_1, Z_2, \dots, Z_\nu$, z nichž každá má normované normální rozdělení. Součet čtverců těchto náhodných veličin, tj. náhodná veličina X má rozdělení χ^2 (čteme „chí-kvadrát“) s ν stupni volnosti, což značíme χ^2_ν.

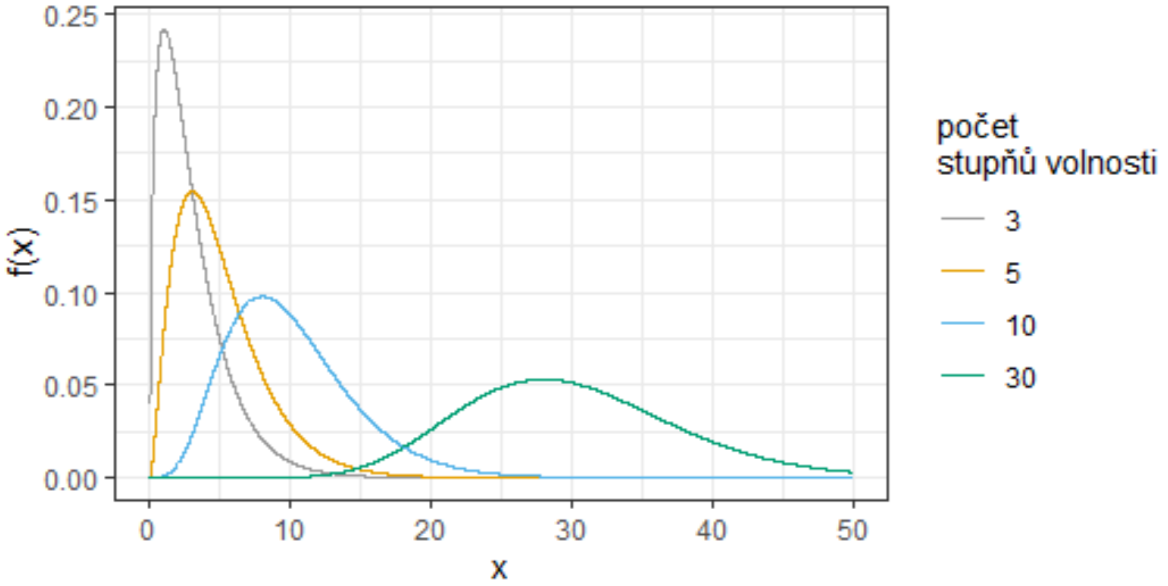
$$\forall i = 1, \dots, n: \;\; Z_i \sim N(0; 1), \text{ pak } \; X = \sum_{i=1}^{\nu} Z_i^2 \sim \chi^2_\nu$$



Chí kvadrát rozdělení s ν stupni volnosti - vlastnosti

- $E(X) = \nu;\ D(X) = 2\nu$. Pro $\nu \to \infty$: $X \sim N(\nu; 2\nu)$
- Mějme náhodný výběr o rozsahu n z populace mající **normální rozdělení** s rozptylem σ^2. Pro uvedený výběr určíme výběrový rozptyl s^2. Lze ukázat, že :

$$\boxed{\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.}$$



Příklad 6 (modelování výběrové sm. odchylky)

Firma Edison vyrábí žárovky Ed. Životnost těchto žárovek je průměrně 5 let se směrodatnou odchylkou 6 měsíců. Pro ověřování kvality výroby bude testováno 20 žárovek. Jaká je pravděpodobnost, že při tomto testu bude zjištěna směrodatná odchylka životnosti vyšší než 7 měsíců? (Předpokládejme, že životnost žárovek má normální rozdělení.)

Řešení:

S … výběrová směrodatná odchylka životnosti 20 žárovek (měs.)

$P(S > 7) = ?$

Neznáme rozdělení S, ALE…!

Nechť $X = \frac{(n-1)S^2}{\sigma^2} \quad \Rightarrow \quad X \sim \chi^2_{n-1}$ (**Pozor!** Lze použít pouze díky předpokladu o normalitě životnosti.)

V našem případě: $X \sim \chi^2_{19}$

Obyčejná úprava nerovnice: Nejdříve umocníme na druhou, poté násobíme $\frac{(n-1)}{\sigma^2}$.

$$P(S > 7) = P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)7^2}{\sigma^2}\right) = P\left(X > \frac{19}{36} \cdot 7^2\right) = P(X > 25{,}86) = \mathbf{0{,}134}$$ (*1-pchisq(25.86,19)*)

Relativní četnost (výběrový podíl)

- **Relativní četnost** (náhodná veličina)

$$\text{nechť } X_i \sim \mathrm{A}(\pi), \text{ pak } P = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **Pozorovaná hodnota (výběrového) průměru** (reálné číslo určené z konkrétní realizace náh. výběru)

$$\text{nechť } X_i \sim \mathrm{A}(\pi), \text{ pak } p = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Důsledek centrální limitní věty:

Nechť:

- ✓ $X_i \sim \mathrm{A}(\pi)$
- ✓ X_i nezávislé náhodné veličiny ,
- ✓ $E(X_1) = E(X_2) = \cdots = E(X_n) = \mu_X = \pi$,
- ✓ $D(X_1) = D(X_2) = \cdots = D(X_n) = \sigma_X^2 = \pi(1-\pi)$,
- ✓ $n \to \infty$ (v praxi: $n > \frac{9}{p(1-p)}$),

pak

$$\boxed{P \sim N\left(\pi; \frac{\pi(1-\pi)}{n}\right) \quad \text{nebo} \quad \frac{P-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n} \sim N(0;1)}$$

Příklad 7 (modelování výběrového podílu)

Pravděpodobnost, že výrobek nebude dosahovat požadované kvality je 4 %.

a) Jaký podíl výrobků nedostatečné kvality lze očekávat při kontrole 300 výrobků?

b) S jakou pravděpodobností bude mezi 300 kontrolovanými výrobky více než 9 (3 %) nekvalitních?

Řešení:

X_i … počet kvalitních výrobků při kontrole 1 výrobků (tj. alternativní (0-1) náh. veličina)

$X_i \sim A(\pi = 0{,}04) \quad \Rightarrow \quad E(X_i) = \pi = 0{,}04,\ D(X_i) = \pi(1-\pi) = 0{,}0384$

P … podíl nekvalitních výrobků mezi 300 kontrolovanými

$P = \frac{1}{300}\sum_{i=1}^{300} X_i$ (tj. průměr X_i)

$P \sim N\left(\mu = 0{,}04, \sigma^2 = \frac{0{,}0384}{300}\right)$ (dle CLV, lze použít protože očekáváme $p \cong 0{,}04 \Rightarrow n > \frac{9}{p(1-p)} (\cong 234)$)

Příklad 7 (modelování výběrového podílu)

Pravděpodobnost, že výrobek nebude dosahovat požadované kvality je 4 %.

a) Jaký podíl výrobků nedostatečné kvality lze očekávat při kontrole 300 výrobků?

b) S jakou pravděpodobností bude mezi 300 kontrolovanými výrobky více než 9 (3 %) nekvalitních?

Řešení:

X_i … počet kvalitních výrobků při kontrole 1 výrobků (tj. alternativní (0-1) náh. veličina)

$X_i \sim A(\pi = 0{,}04) \quad \Rightarrow \quad E(X_i) = \pi = 0{,}04,\ D(X_i) = \pi(1-\pi) = 0{,}0384$

P … podíl nekvalitních výrobků mezi 300 kontrolovanými

$P = \frac{1}{300}\sum_{i=1}^{300} X_i$ (tj. průměr X_i)

$P \sim N\left(\mu = 0{,}04, \sigma^2 \cong 0{,}0001\right)$ (dle CLV, lze použít protože očekáváme $p \cong 0{,}04 \Rightarrow n > \frac{9}{p(1-p)} (\cong 234)$)

Příklad 7 (modelování výběrového podílu)

Pravděpodobnost, že výrobek nebude dosahovat požadované kvality je 4 %.

a) Jaký podíl výrobků nedostatečné kvality lze očekávat při kontrole 300 výrobků?

b) S jakou pravděpodobností bude mezi 300 kontrolovanými výrobky více než 9 (3 %) nekvalitních?

Řešení:

X_i … počet kvalitních výrobků při kontrole 1 výrobků (tj. alternativní (0-1) náh. veličina)

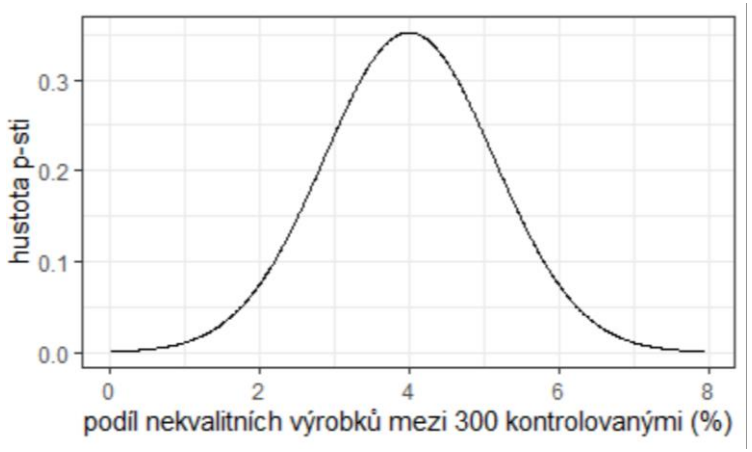
$X_i \sim A(\pi = 0{,}04) \quad \Rightarrow \quad E(X_i) = \pi = 0{,}04,\ D(X_i) = \pi(1-\pi) = 0{,}0384$

P … podíl nekvalitních výrobků mezi 300 kontrolovanými

$P = \frac{1}{300}\sum_{i=1}^{300} X_i$ (tj. průměr X_i)

$P \sim N\left(\mu = 0{,}04, \sigma^2 \cong 0{,}01^2\right)$ (dle CLV)

ada) Očekávaný podíl nekvalitních výrobků mezi 300 kontrolovanými: cca (1 – 7) %.



Příklad 7 (modelování výběrového podílu)

Pravděpodobnost, že výrobek nebude dosahovat požadované kvality je 4 %.

a) Jaký podíl výrobků nedostatečné kvality lze očekávat při kontrole 300 výrobků?

b) S jakou pravděpodobností bude mezi 300 kontrolovanými výrobky více než 9 (3 %) nekvalitních?

Řešení:

X_i … počet kvalitních výrobků při kontrole 1 výrobků (tj. alternativní (0-1) náh. veličina)

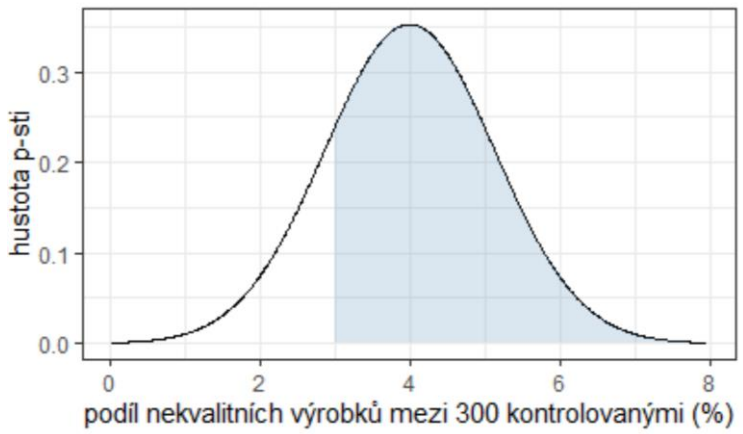
$X_i \sim A(\pi = 0{,}04) \quad \Rightarrow \quad E(X_i) = \pi = 0{,}04,\ D(X_i) = \pi(1-\pi) = 0{,}0384$

P … podíl nekvalitních výrobků mezi 300 kontrolovanými

$P = \frac{1}{300}\sum_{i=1}^{300} X_i$ (tj. průměr X_i)

$P \sim N\left(\mu = 0{,}04, \sigma^2 \cong 0{,}01^2\right)$ (dle CLV)

adb) $P(X > 0{,}03) \cong \mathbf{0{,}841}$ (*1-pnorm(0.03,0.04,0.01)*)



Výběrové charakteristiky - shrnutí

Mějme náh. výběr $\boldsymbol{X}$ ze spojitého rozdělení, tj. $\boldsymbol{X} = (X_1, \dots, X_n)$, $\forall i = 1, \dots, n$: $E(X_i) = \mu$, $D(X_i) = \sigma^2$ a předpokládejme, že rozsah výběru nepřesahuje 5 % velikosti populace ($n \leq 0{,}05N$, neboli $N \geq 20n$).				
Populační parametr	**Výběrová charakteristika**	**Podmínky pro použití modelu**	**Jak modelovat „přímo“?**	**Jak modelovat s využitím „pomocné statistiky“?**
střední hodnota μ	průměr $\bar{X}$	normalita populace nebo $n > 30$, známý rozptyl σ^2	$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$	$\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$
		normalita populace	---	$\frac{\bar{X}-\mu}{S}\sqrt{n} \sim t_{n-1}$
úhrn	výběrový úhrn $\sum_{i=1}^{n} X_i$	normalita populace nebo $n > 30$, známý rozptyl σ^2	$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$	$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma}} \sim N(0,1)$
rozptyl σ^2, směr. odchylka σ	výběrový rozptyl S^2, výb. směr. odchylka S	normalita populace	---	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

Výběrové charakteristiky - shrnutí

Mějme náhodný výběr $\boldsymbol{X}$ z alternativního rozdělení, tj. $\boldsymbol{X} = (X_1, \dots, X_n)$, $\forall i = 1, \dots, n$: $X_i \sim A(\pi)$ a předpokládejme, že rozsah výběru nepřesahuje 5 % velikosti populace ($n \leq 0{,}05N$, neboli $N \geq 20n$).				
Populační parametr	**Výběrová charakteristika**	**Podmínky pro použití modelu**	**Jak modelovat „přímo“?**	**Jak modelovat s využitím „pomocné statistiky“?**
pravděpodobnost π	výběrový podíl (rel. četnost) P	$n > \frac{9}{p(1-p)}$	$P \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$	$\frac{P-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n} \sim N(0,1)$

Tabulky ve zjednodušené podobě najdete i v dokumentu Vzorce a tabulky (str. 4-5), který můžete používat u zkoušky…

Rozdíl průměrů (výběry z populací se známými rozptyly)

Mějme náhodný výběr $X_{11}, \dots, X_{1n_1}$ z rozdělení se střední hodnotou μ_1 a rozptylem σ_1^2

a náhodný výběr $X_{21}, \dots, X_{2n_2}$ z rozdělení se střední hodnotou μ_2 a rozptylem σ_2^2.

Dále nechť jsou splněny následující předpoklady:

- Rozsah každé z populací je dostatečně velký vzhledem k rozsahu příslušného výběru $(N_1 > 20n_1,\ N_2 > 20n_2)$.
- Platí předpoklady CLV, zejména to, že každý z výběrů pochází z normálního rozdělení nebo je dostatečně velký $(n_i > 30)$, pak

$$\boxed{(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \quad \text{nebo} \quad \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).}$$

Rozdíl průměrů (výběry z populací se známými rozptyly)

$$(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \quad \text{nebo} \quad \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Proč?

- Platí-li CLV, pak $\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$, $\bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$.
- $E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$,
- $D(\bar{X}_1 - \bar{X}_2) = D(\bar{X}_1 + (-\bar{X}_2)) = D(\bar{X}_1) + D(-\bar{X}_2) = D(\bar{X}_1) + (-1)^2 D(\bar{X}_2) = D(\bar{X}_1) + D(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
- **Věta 1**: Jsou-li X_i nezávislé náh. veličiny, pak má náhodná veličina $X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$ normální rozdělení.

Rozdíl průměrů
(výběry z populací s normálním rozdělením s neznámými, ale shodnými rozptyly)

Mějme náhodný výběr $X_{11}, \dots, X_{1n_1}$ z normálního rozdělení se střední hodnotou μ_1 a neznámým rozptylem σ_1^2 a náhodný výběr $X_{21}, \dots, X_{2n_2}$ z normálního rozdělení se střední hodnotou μ_2 a neznámým rozptylem σ_2^2.

Dále nechť jsou splněny následující předpoklady:

- Rozsah každé z populací je dostatečně velký vzhledem k rozsahu příslušného výběru ($N_1 > 20n_1,\ N_2 > 20n_2$).
- $\sigma_1^2 = \sigma_2^2$.

Pak

$$\boxed{\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{S_1^2(n_1-1)+S_2^2(n_2-1)}}\sqrt{\frac{n_1n_2(n_1+n_2-2)}{n_1+n_2}} \sim t_{n_1+n_2-2}.}$$

Rozdíl průměrů
(výběry z populací s normálním rozdělením s neznámými neshodnými rozptyly)

Mějme náhodný výběr $X_{11}, \dots, X_{1n_1}$ z normálního rozdělení se střední hodnotou μ_1 a neznámým rozptylem σ_1^2 a náhodný výběr $X_{21}, \dots, X_{2n_2}$ z normálního rozdělení se střední hodnotou μ_2 a neznámým rozptylem σ_2^2.

Dále nechť jsou splněny následující předpoklady:

- Rozsah každé z populací je dostatečně velký vzhledem k rozsahu příslušného výběru $(N_1 > 20n_1,\ N_2 > 20n_2)$.
- $\sigma_1^2 \neq \sigma_2^2$.

Pak

$$\boxed{\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t_\nu, \text{ kde } \nu \cong \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \frac{1}{n_1+1}+\left(\frac{s_2^2}{n_2}\right)^2 \frac{1}{n_2+1}} - 2.}$$

Poměr výběrových rozptylů (výběry z populací s normálním rozdělením)

Mějme náhodný výběr $X_{11}, \dots, X_{1n_1}$ z normálního rozdělení s rozptylem σ_1^2

a náhodný výběr $X_{21}, \dots, X_{2n_2}$ z normálního rozdělení s rozptylem σ_2^2.

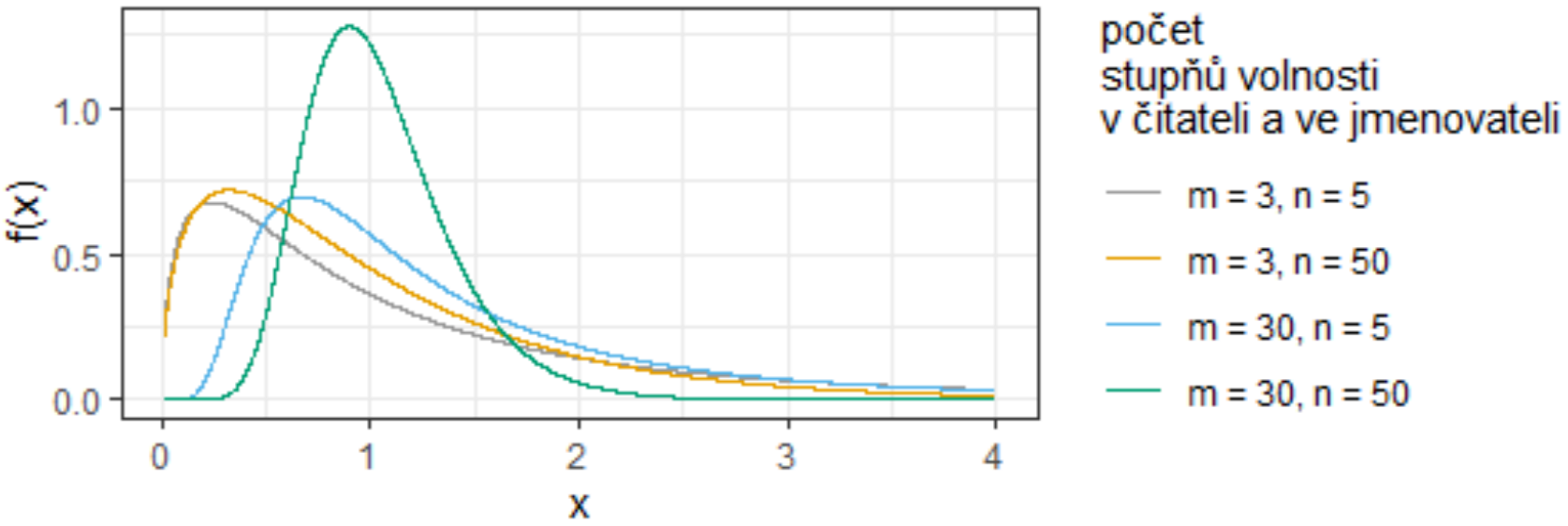
- Pro modelování poměru výběrových rozptylů S_1^2/S_2^2 používáme tzv. **Fischerovo – Snedecorovo rozdělení s m stupni volnosti v čitateli a n stupni volnosti ve jmenovateli**.

Fischerovo - Snedecorovo rozdělení
s m stupni volnosti v čitateli a n stupni volnosti ve jmenovateli

Mějme dvě nezávislé náhodné veličiny V a W s rozdělením χ^2. První z nich má počet stupňů volnosti m, druhá má počet stupňů volnosti n (obecně mají různý počet stupňů volnosti). Pak má náhodná veličina

$$F = \frac{\frac{V}{m}}{\frac{W}{n}}$$

Fisherovo-Snedecorovo rozdělení o m a n stupních volnosti, což značíme $F \sim F_{m,n}$.



Poměr výběrových rozptylů (výběry z populací s normálním rozdělením)

Mějme náhodný výběr $X_{11}, \dots, X_{1n_1}$ z normálního rozdělení s rozptylem σ_1^2

a náhodný výběr $X_{21}, \dots, X_{2n_2}$ z normálního rozdělení s rozptylem σ_2^2.

Označme S_1^2 a S_2^2 příslušné výběrové rozptyly.

Dále nechť je splněn následující předpoklad:

- Rozsah každé z populací je dostatečně velký vzhledem k rozsahu příslušného výběru $(N_1 > 20n_1,\ N_2 > 20n_2)$, pak

$$\boxed{\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F_{n_1-1,n_2-1}.}$$

Příklad 8 (modelování poměru výběrových rozptylů)

Firma Edison vyrábí žárovky Ed. Životnost těchto žárovek je průměrně 5 let se směrodatnou odchylkou 6 měsíců. Uvedené informace specifikujeme: Žárovky jsou vyráběny na dvou linkách. Předpokládejme, že obě linky mají srovnatelné parametry, tj. že průměrná životnost a variabilita životnosti žárovek Ed vyrobených ve firmě Edison nezávisí na tom, na jaké lince byly vyrobeny. Pro ověření kvality výroby bude testována životnost 20 žárovek z linky 1 a 30 žárovek z linky 2. (Předpokládejme, že životnost žárovek má normální rozdělení.)

Jaká je pravděpodobnost, že u vzorku z linky 1 bude zjištěn více než dvojnásobný rozptyl oproti rozptylu zjištěnému u vzorku z linky 2?

Řešení:

S_1^2 ... výběrový rozptyl na lince 1, S_2^2 ... výběrový rozptyl na lince 2

$P\left(\frac{S_1^2}{S_2^2} > 2\right) = ?$

Neznáme rozdělení $\frac{S_1^2}{S_2^2}$, ALE...!

Nechť $X = \left(S_1^2/\sigma_1^2\right)/\left(S_2^2/\sigma_2^2\right) = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \Rightarrow \quad X \sim F_{n_1-1,n_2-1}$ (**Pozor!** Lze použít díky předpokladu o normalitě životnosti.)

V našem případě: $X \sim F_{19,29}$

Dle zadání: $\sigma_1^2 = \sigma_2^2$

$P\left(\frac{S_1^2}{S_2^2} > 2\right) = P\left(\frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} > 2 \cdot \frac{\sigma_2^2}{\sigma_1^2}\right) = P\left(X > 2 \cdot \frac{\sigma_2^2}{\sigma_1^2}\right) = P(X > 2) = \mathbf{0,045}$ (*1-pf(2,19,29)*)

Rozdíl výběrových podílů

Mějme náhodný výběr $X_{11}, \dots, X_{1n_1}$ z alternativního rozdělení $A(\pi_1)$, kde $P_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}$

a náhodný výběr $X_{21}, \dots, X_{2n_2}$ z alternativního rozdělení $A(\pi_2)$, kde $P_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$.

Dále nechť je splněn následující předpoklad:

- Rozsah každé z populací je dostatečně velký vzhledem k rozsahu příslušného výběru $(N_1 > 20n_1, N_2 > 20n_2)$. Pak

Důsledek centrální limitní věty:

Nechť:

- ✓ $X_{1i} \sim A(\pi_1), X_{2j} \sim A(\pi_2)$
- ✓ X_{1i} nezávislé náhodné veličiny, X_{2i} nezávislé náhodné veličiny,
- ✓ $E(X_{1i}) = \pi_1, E(X_{2i}) = \pi_2$,
- ✓ $D(X_{1i}) = \pi_1(1-\pi_1), D(X_{2i}) = \pi_2(1-\pi_2)$,
- ✓ $n_1 \to \infty$ (v praxi: $n_1 > \frac{9}{p_1(1-p_1)}$), $n_2 \to \infty$ (v praxi: $n_2 > \frac{9}{p_2(1-p_2)}$)

pak

$$\boxed{(P_1 - P_2) \sim N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right) \quad \text{nebo} \quad \frac{(P_1-P_2)-(\pi_1-\pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1).}$$

Rozdíl výběrových podílů

$$(P_1 - P_2) \sim N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right) \quad \text{nebo} \quad \frac{(P_1-P_2)-(\pi_1-\pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1)$$

Proč?

- Platí-li CLV, pak $P_1 \sim N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1}\right)$, $P_2 \sim N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2}\right)$.
- $E(P_1 - P_2) = E(P_1) - E(P_2) = \pi_1 - \pi_2$,
- $D(P_1 - P_2) = D(P_1 + (-P_2)) = D(P_1) + D(-P_2) = D(P_1) + (-1)^2 D(P_2) = D(P_1) + D(P_2) =$

  $= \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$
- **Věta 1**: Jsou-li X_i nezávislé náh. veličiny, pak má náhodná veličina $X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$ normální rozdělení.

Rozdíl (podíl) výběrových charakteristik - shrnutí

Mějme **dva nezávislé výběry z normálního rozdělení**.
$\forall i = 1, 2, \dots, n_1$, kde n_1 je rozsah prvního výběru: $X_{1i} \to N(\mu_1; \sigma_1^2)$,
$\forall j = 1, 2, \dots, n_2$, kde n_2 je rozsah druhého výběru: $X_{2j} \to N(\mu_2; \sigma_2^2)$
a předpokládejme, že rozsahy výběrů nepřesahuje 5 % velikosti populace ($n_i \leq 0{,}05 N_i$, neboli $N_i \geq 20 n_i$ pro $i \in \{1,2\}$).

Rozdíl (podíl) populačních parametrů	**Rozdíl (podíl) výběrových charakteristik**	**Další podmínky pro použití modelu**	**Jak modelovat „přímo“?**	**Jak modelovat s využitím „pomocné statistiky“?**
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	známe σ_1^2, σ_2^2 nebo $n_1 > 30$, $n_2 > 30$	$(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$	$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$
		neznáme σ_1^2, σ_2^2, ale víme, že $\sigma_1^2 = \sigma_2^2$	---	$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \sim t_{n_1 + n_2 - 2}$
		neznáme σ_1^2, σ_2^2, ale víme, že $\sigma_1^2 \neq \sigma_2^2$	---	$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\nu$, kde $\nu \cong \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 \frac{1}{n_1 + 1} + \left(\frac{S_2^2}{n_2}\right)^2 \frac{1}{n_2 + 1}} - 2$

Rozdíl (podíl) výběrových charakteristik - shrnutí

Mějme **dva nezávislé výběry z normálního rozdělení**.
$\forall i = 1, 2, \dots, n_1$, kde n_1 je rozsah prvního výběru: $X_{1i} \sim N(\mu_1; \sigma_1^2)$,
$\forall j = 1, 2, \dots, n_2$, kde n_2 je rozsah druhého výběru: $X_{2j} \sim N(\mu_2; \sigma_2^2)$
a předpokládejme, že rozsahy výběrů nepřesahuje 5 % velikosti populace ($n_i \leq 0{,}05 N_i$, neboli $N_i \geq 20 n_i$ pro $i \in \{1,2\}$).

Rozdíl (podíl) populačních parametrů	**Rozdíl (podíl) výběrových charakteristik**	**Další podmínky pro použití modelu**	**Jak modelovat „přímo“?**	**Jak modelovat s využitím „pomocné statistiky“?**
σ_1^2/σ_2^2	S_1^2/S_2^2	---	---	$(S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2) \sim F_{n_1-1,n_2-1}$

Rozdíl (podíl) výběrových charakteristik - shrnutí

Mějme **dva nezávislé výběry z alternativního rozdělení**.
$\forall i = 1, 2, \dots, n_1$, kde n_1 je rozsah prvního výběru: $X_{1i} \sim A(\pi_1)$,
$\forall j = 1, 2, \dots, n_2$, kde n_2 je rozsah druhého výběru: $X_{2j} \sim A(\pi_2)$
a předpokládejme, že rozsahy výběrů splňují podmínku $\left(n_i > \frac{9}{p_i(1-p_i)} \text{ pro } i \in \{1,2\}\right)$.

Rozdíl (podíl) populačních parametrů	**Rozdíl (podíl) výběrových charakteristik**	**Další podmínky pro použití modelu**	**Jak modelovat „přímo“?**	**Jak modelovat s využitím „pomocné statistiky“?**
$\pi_1 - \pi_2$	$P_1 - P_2$	---	$(P_1 - P_2) \sim N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)$	$\frac{(P_1-P_2)-(\pi_1-\pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1)$

Tabulky ve zjednodušené podobě najdete i v dokumentu Vzorce a tabulky (str. 5), který můžete používat u zkoušky…

Děkuji za pozornost!

martina.litschmannova@vsb.cz

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY | KATEDRA APLIKOVANÉ MATEMATIKY