Základy biostatistiky

(MD710P09)
ak. rok 2007/2008

Karel Zvára

karel.zvara@mff.cuni.cz
http://www.karlin.mff.cuni.cz/∼zvara
katedra pravděpodobnosti a matematické statistiky MFF UK

(naposledy upraveno 19. března 2008)

populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku) $\Rightarrow$ rozdělení náhodné veličiny
- **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- **reprezentativní** výběr obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- **statistika** funkce náhodného výběru (pozorování)
- **odhad** statistika použitá k odhadu parametru

populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku) $\Rightarrow$ rozdělení náhodné veličiny
- **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- **reprezentativní** výběr obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- **statistika** funkce náhodného výběru (pozorování)
- **odhad** statistika použitá k odhadu parametru

populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku) $\Rightarrow$ rozdělení náhodné veličiny
- **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- **reprezentativní** výběr obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- **statistika** funkce náhodného výběru (pozorování)
- **odhad** statistika použitá k odhadu parametru

populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku) $\Rightarrow$ rozdělení náhodné veličiny
- **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- **reprezentativní** výběr obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- **statistika** funkce náhodného výběru (pozorování)
- **odhad** statistika použitá k odhadu parametru

populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku) $\Rightarrow$ rozdělení náhodné veličiny
- **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- **reprezentativní** výběr obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- **statistika** funkce náhodného výběru (pozorování)
- **odhad** statistika použitá k odhadu parametru

populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku) $\Rightarrow$ rozdělení náhodné veličiny
- **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- **reprezentativní** výběr obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- **statistika** funkce náhodného výběru (pozorování)
- **odhad** statistika použitá k odhadu parametru

populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku) $\Rightarrow$ rozdělení náhodné veličiny
- **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- **reprezentativní** výběr obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- **statistika** funkce náhodného výběru (pozorování)
- **odhad** statistika použitá k odhadu parametru

průměr z náhodného výběru

- $X_1, \ldots, X_n$ **nezávislé**, stejné rozdělení — **náhodný výběr**
  $\mu_{X_i} = \mathrm{E}\, X_i = \mu$ (stejná střední hodnota) — populační průměr
  $\sigma^2_{X_i} = \mathrm{var}\, X_i = \sigma^2$ (stejný rozptyl) — populační rozptyl
- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ — výběrový průměr
- $\mu_{\bar{X}} = \mathrm{E}\, \bar{X} = \mu$
  - výběrový průměr $\bar{X}$ je opět náhodná veličina
  - je **nestranným** odhadem [unbiased estimator] parametru μ
  - nestranným odhadem populačního průměru (střední hodnoty)
  - když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

průměr z náhodného výběru

- $X_1, \ldots, X_n$ **nezávislé**, stejné rozdělení — **náhodný výběr**
  $\mu_{X_i} = \mathrm{E}\, X_i = \mu$ (stejná střední hodnota) — populační průměr
  $\sigma^2_{X_i} = \mathrm{var}\, X_i = \sigma^2$ (stejný rozptyl) — populační rozptyl
- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ — výběrový průměr
- $\mu_{\bar{X}} = \mathrm{E}\, \bar{X} = \mu$
  - výběrový průměr $\bar{X}$ je opět náhodná veličina
  - je **nestranným** odhadem [unbiased estimator] parametru μ
  - nestranným odhadem populačního průměru (střední hodnoty)
  - když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

průměr z náhodného výběru

- $X_1, \ldots, X_n$ **nezávislé**, stejné rozdělení — **náhodný výběr**
  $\mu_{X_i} = \mathrm{E}\, X_i = \mu$ (stejná střední hodnota) — populační průměr
  $\sigma^2_{X_i} = \mathrm{var}\, X_i = \sigma^2$ (stejný rozptyl) — populační rozptyl
- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ — výběrový průměr
- $\mu_{\bar{X}} = \mathrm{E}\, \bar{X} = \mu$
  - výběrový průměr $\bar{X}$ je opět náhodná veličina
  - je **nestranným** odhadem [unbiased estimator] parametru μ
  - nestranným odhadem populačního průměru (střední hodnoty)
  - když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

průměr z náhodného výběru

- $X_1, \ldots, X_n$ **nezávislé**, stejné rozdělení — **náhodný výběr**
  $\mu_{X_i} = \mathrm{E}\, X_i = \mu$ (stejná střední hodnota) — populační průměr
  $\sigma^2_{X_i} = \mathrm{var}\, X_i = \sigma^2$ (stejný rozptyl) — populační rozptyl
- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ — výběrový průměr
- $\mu_{\bar{X}} = \mathrm{E}\, \bar{X} = \mu$
  - výběrový průměr $\bar{X}$ je opět náhodná veličina
  - je **nestranným** odhadem [unbiased estimator] parametru μ
  - nestranným odhadem populačního průměru (střední hodnoty)
  - když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

průměr z náhodného výběru

- $X_1, \ldots, X_n$ **nezávislé**, stejné rozdělení — **náhodný výběr**
  $\mu_{X_i} = \mathrm{E}\, X_i = \mu$ (stejná střední hodnota) — populační průměr
  $\sigma^2_{X_i} = \mathrm{var}\, X_i = \sigma^2$ (stejný rozptyl) — populační rozptyl
- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ — výběrový průměr
- $\mu_{\bar{X}} = \mathrm{E}\, \bar{X} = \mu$
  - výběrový průměr $\bar{X}$ je opět náhodná veličina
  - je **nestranným** odhadem [unbiased estimator] parametru μ
  - nestranným odhadem populačního průměru (střední hodnoty)
  - když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

průměr z náhodného výběru

- $X_1, \ldots, X_n$ **nezávislé**, stejné rozdělení — **náhodný výběr**
  $\mu_{X_i} = \mathrm{E}\, X_i = \mu$ (stejná střední hodnota) — populační průměr
  $\sigma^2_{X_i} = \mathrm{var}\, X_i = \sigma^2$ (stejný rozptyl) — populační rozptyl
- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ — výběrový průměr
- $\mu_{\bar{X}} = \mathrm{E}\, \bar{X} = \mu$
  - výběrový průměr $\bar{X}$ je opět náhodná veličina
  - je **nestranným** odhadem [unbiased estimator] parametru μ
  - nestranným odhadem populačního průměru (střední hodnoty)
  - když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

průměr z náhodného výběru

- $X_1, \ldots, X_n$ **nezávislé**, stejné rozdělení — **náhodný výběr**
  $\mu_{X_i} = \mathrm{E}\, X_i = \mu$ (stejná střední hodnota) — populační průměr
  $\sigma^2_{X_i} = \mathrm{var}\, X_i = \sigma^2$ (stejný rozptyl) — populační rozptyl
- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ — výběrový průměr
- $\mu_{\bar{X}} = \mathrm{E}\, \bar{X} = \mu$
  - výběrový průměr $\bar{X}$ je opět náhodná veličina
  - je **nestranným** odhadem [unbiased estimator] parametru μ
  - nestranným odhadem populačního průměru (střední hodnoty)
  - když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

rozptyl průměru z náhodného výběru

$$\sigma^2_{\bar{X}} = \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{var} X_i = \frac{\sigma^2}{n} = \left(\frac{\sigma}{\sqrt{n}}\right)^2 = \left(\text{S.E.}(\bar{X})\right)^2$$

- $\text{S.E.}(\bar{X})$ – **střední chyba průměru** [standard error of mean]
- variabilita průměrů (měřená rozptylem) z výběrů rozsahu n je n-krát menší, než variabilita jednotlivých pozorování σ^2
- střední chyba průměru je $\sqrt{n}$-krát menší než σ
- čím jsou rozsahy výběru větší, tím méně výběrové průměry kolísají (kolem populačního průměru)
- speciálně pro normální rozdělení $X_i \sim \text{N}(\mu, \sigma^2)$:

$$\bar{X} \sim \text{N}(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim \text{N}(0, 1)$$

(všimněte si závislosti na n)

rozptyl průměru z náhodného výběru

$$\sigma^2_{\bar{X}} = \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{var} X_i = \frac{\sigma^2}{n} = \left(\frac{\sigma}{\sqrt{n}}\right)^2 = \left(\text{S.E.}(\bar{X})\right)^2$$

- $\text{S.E.}(\bar{X})$ – **střední chyba průměru** [standard error of mean]
- variabilita průměrů (měřená rozptylem) z výběrů rozsahu n je n-krát menší, než variabilita jednotlivých pozorování σ^2
- střední chyba průměru je $\sqrt{n}$-krát menší než σ
- čím jsou rozsahy výběru větší, tím méně výběrové průměry kolísají (kolem populačního průměru)
- speciálně pro normální rozdělení $X_i \sim \text{N}(\mu, \sigma^2)$:

$$\bar{X} \sim \text{N}(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim \text{N}(0, 1)$$

(všimněte si závislosti na n)

rozptyl průměru z náhodného výběru

$$\sigma^2_{\bar{X}} = \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{var} X_i = \frac{\sigma^2}{n} = \left(\frac{\sigma}{\sqrt{n}}\right)^2 = \left(\text{S.E.}(\bar{X})\right)^2$$

- $\text{S.E.}(\bar{X})$ – **střední chyba průměru** [standard error of mean]
- variabilita průměrů (měřená rozptylem) z výběrů rozsahu n je n-krát menší, než variabilita jednotlivých pozorování σ^2
- střední chyba průměru je $\sqrt{n}$-krát menší než σ
- čím jsou rozsahy výběru větší, tím méně výběrové průměry kolísají (kolem populačního průměru)
- speciálně pro normální rozdělení $X_i \sim \text{N}(\mu, \sigma^2)$:

$$\bar{X} \sim \text{N}(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim \text{N}(0, 1)$$

(všimněte si závislosti na n)

rozptyl průměru z náhodného výběru

$$\sigma^2_{\bar{X}} = \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{var} X_i = \frac{\sigma^2}{n} = \left(\frac{\sigma}{\sqrt{n}}\right)^2 = \left(\text{S.E.}(\bar{X})\right)^2$$

- $\text{S.E.}(\bar{X})$ – **střední chyba průměru** [standard error of mean]
- variabilita průměrů (měřená rozptylem) z výběrů rozsahu n je n-krát menší, než variabilita jednotlivých pozorování σ^2
- střední chyba průměru je $\sqrt{n}$-krát menší než σ
- čím jsou rozsahy výběru větší, tím méně výběrové průměry kolísají (kolem populačního průměru)
- speciálně pro normální rozdělení $X_i \sim \text{N}(\mu, \sigma^2)$:

$$\bar{X} \sim \text{N}(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim \text{N}(0, 1)$$

(všimněte si závislosti na n)

rozptyl průměru z náhodného výběru

$$\sigma^2_{\bar{X}} = \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{var} X_i = \frac{\sigma^2}{n} = \left(\frac{\sigma}{\sqrt{n}}\right)^2 = \left(\text{S.E.}(\bar{X})\right)^2$$

- $\text{S.E.}(\bar{X})$ – **střední chyba průměru** [standard error of mean]
- variabilita průměrů (měřená rozptylem) z výběrů rozsahu n je n-krát menší, než variabilita jednotlivých pozorování σ^2
- střední chyba průměru je $\sqrt{n}$-krát menší než σ
- čím jsou rozsahy výběru větší, tím méně výběrové průměry kolísají (kolem populačního průměru)
- speciálně pro normální rozdělení $X_i \sim \text{N}(\mu, \sigma^2)$:

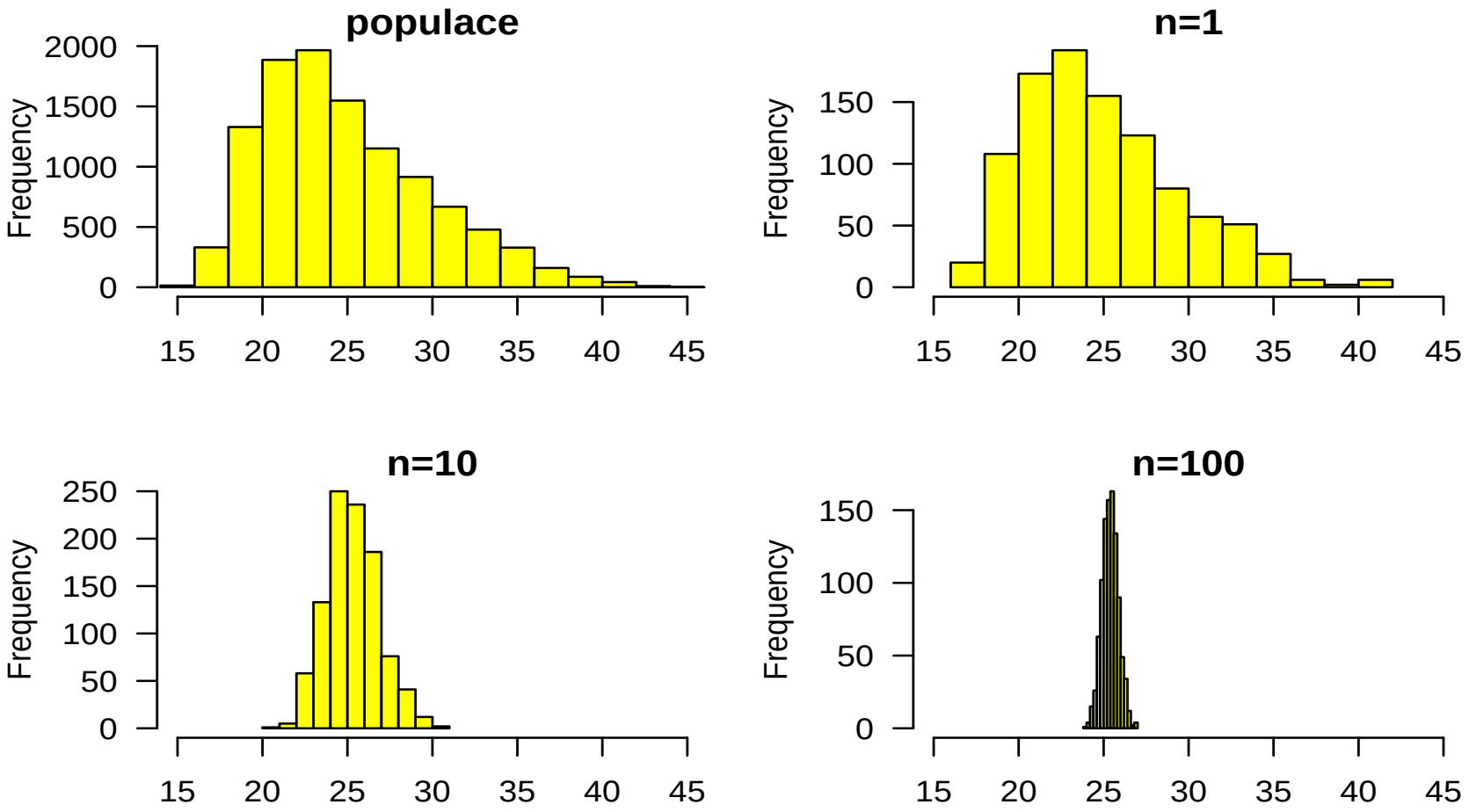
$$\bar{X} \sim \text{N}(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim \text{N}(0, 1)$$

(všimněte si závislosti na n)

příklad: věk matek

populace - 10 916 matek, opakované výběry rozsahu $n = 1, 10, 100$

je patrná variabilita klesající s rostoucím n



příklad: věk matek – shrnutí

- velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- náhodně vybráno 100 matek (vlastně průměry výběrů rozsahu $n = 1$), nakreslen histogram
- 100 krát náhodně vybráno vždy $n = 10$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- 100 krát náhodně vybráno vždy $n = 100$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- skutečné rozptyly (odhady ze 100 realizací): 23,5; 2,20; 0,21

příklad: věk matek – shrnutí

- velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- náhodně vybráno 100 matek (vlastně průměry výběrů rozsahu $n = 1$), nakreslen histogram
- 100 krát náhodně vybráno vždy $n = 10$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- 100 krát náhodně vybráno vždy $n = 100$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- skutečné rozptyly (odhady ze 100 realizací): 23,5; 2,20; 0,21

příklad: věk matek – shrnutí

- velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- náhodně vybráno 100 matek (vlastně průměry výběrů rozsahu $n = 1$), nakreslen histogram
- 100 krát náhodně vybráno vždy $n = 10$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- 100 krát náhodně vybráno vždy $n = 100$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- skutečné rozptyly (odhady ze 100 realizací): 23,5; 2,20; 0,21

příklad: věk matek – shrnutí

- velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- náhodně vybráno 100 matek (vlastně průměry výběrů rozsahu $n = 1$), nakreslen histogram
- 100 krát náhodně vybráno vždy $n = 10$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- 100 krát náhodně vybráno vždy $n = 100$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- skutečné rozptyly (odhady ze 100 realizací): 23,5; 2,20; 0,21

příklad: věk matek – shrnutí

- velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- náhodně vybráno 100 matek (vlastně průměry výběrů rozsahu $n = 1$), nakreslen histogram
- 100 krát náhodně vybráno vždy $n = 10$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- 100 krát náhodně vybráno vždy $n = 100$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- skutečné rozptyly (odhady ze 100 realizací): 23,5; 2,20; 0,21

příklad: věk matek – shrnutí

- velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- náhodně vybráno 100 matek (vlastně průměry výběrů rozsahu $n = 1$), nakreslen histogram
- 100 krát náhodně vybráno vždy $n = 10$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- 100 krát náhodně vybráno vždy $n = 100$ matek, vždy spočítán průměr, nakreslen histogram průměrů
- podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- skutečné rozptyly (odhady ze 100 realizací): 23,5; 2,20; 0,21

centrální limitní věta (CLT)
[Central Limit Theorem]

- Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$ (nemusí pocházet z normálního rozdělení).
  Potom **pro velké** n má průměr $\bar{X}$ přibližně rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, součet $X_1 + \ldots + X_n$ pak rozdělení $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- prakticky: **průměr** má pro dost velká n **normální rozdělení** s rozptylem n-krát menším než jednotlivá pozorování, a to bez ohledu na výchozí rozdělení jednotlivých pozorování
- CLT je často důvodem předpokladu o normálním rozdělení, výsledná hodnota je ovlivněna součtem velikého počtu nahodilých malých vlivů
- příklad: průměrný věk matek z velkých výběrů má už (téměř) normální rozdělení

centrální limitní věta (CLT)
[Central Limit Theorem]

- Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$ (nemusí pocházet z normálního rozdělení).
  Potom **pro velké** n má průměr $\bar{X}$ přibližně rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, součet $X_1 + \ldots + X_n$ pak rozdělení $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- prakticky: **průměr** má pro dost velká n **normální rozdělení** s rozptylem n-krát menším než jednotlivá pozorování, a to bez ohledu na výchozí rozdělení jednotlivých pozorování
- CLT je často důvodem předpokladu o normálním rozdělení, výsledná hodnota je ovlivněna součtem velikého počtu nahodilých malých vlivů
- příklad: průměrný věk matek z velkých výběrů má už (téměř) normální rozdělení

centrální limitní věta (CLT)
[Central Limit Theorem]

- Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$ (nemusí pocházet z normálního rozdělení).
  Potom **pro velké** n má průměr $\bar{X}$ přibližně rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, součet $X_1 + \ldots + X_n$ pak rozdělení $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- prakticky: **průměr** má pro dost velká n **normální rozdělení** s rozptylem n-krát menším než jednotlivá pozorování, a to bez ohledu na výchozí rozdělení jednotlivých pozorování
- CLT je často důvodem předpokladu o normálním rozdělení, výsledná hodnota je ovlivněna součtem velikého počtu nahodilých malých vlivů
- příklad: průměrný věk matek z velkých výběrů má už (téměř) normální rozdělení

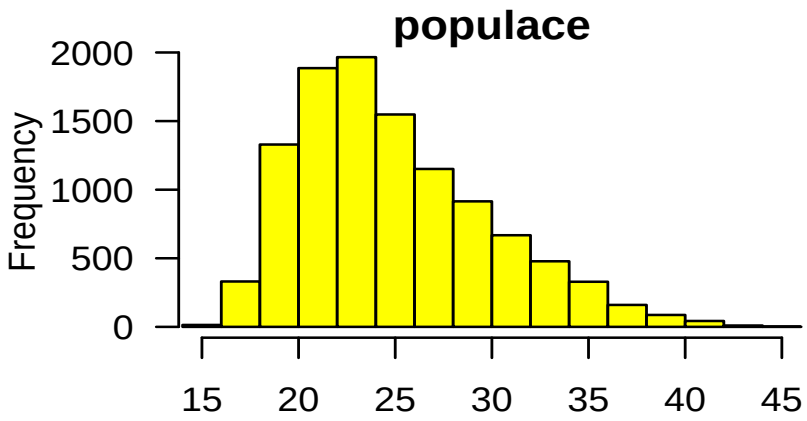
centrální limitní věta (CLT)
[Central Limit Theorem]

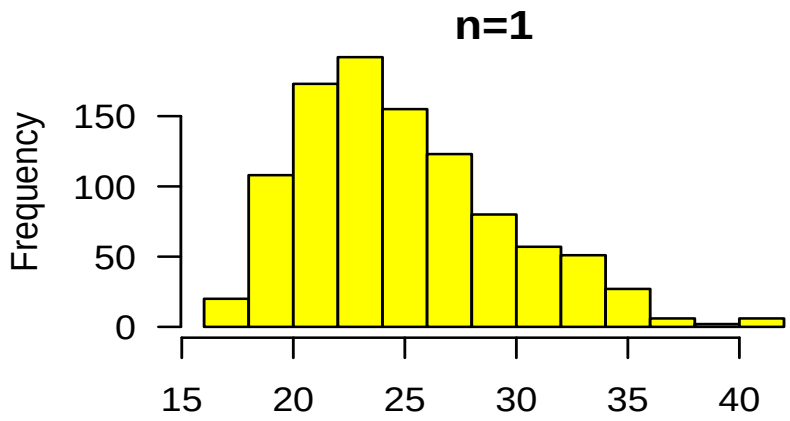
- Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$ (nemusí pocházet z normálního rozdělení).
  Potom **pro velké** n má průměr $\bar{X}$ přibližně rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, součet $X_1 + \ldots + X_n$ pak rozdělení $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- prakticky: **průměr** má pro dost velká n **normální rozdělení** s rozptylem n-krát menším než jednotlivá pozorování, a to bez ohledu na výchozí rozdělení jednotlivých pozorování
- CLT je často důvodem předpokladu o normálním rozdělení, výsledná hodnota je ovlivněna součtem velikého počtu nahodilých malých vlivů
- příklad: průměrný věk matek z velkých výběrů má už (téměř) normální rozdělení

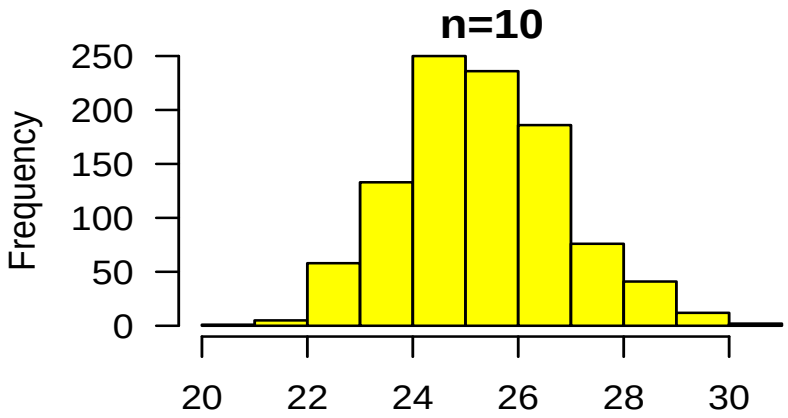
příklad: věk matek

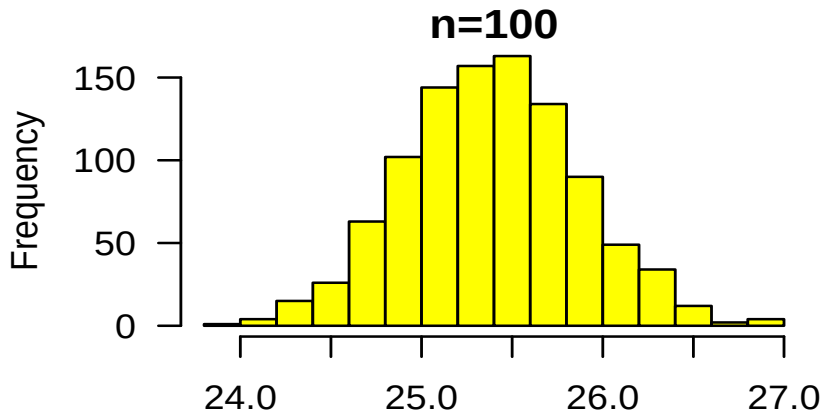
populace - 10 916 matek, opakované výběry rozsahu $n = 1, 10, 100$

je patrno, že s rostoucím n se histogram blíží histogramu norm. rozdělení









příklad: věk matek

průměrný věk matek v opakovaných výběrech

rozsah výběru n	průměr průměrů	směr. odch. průměrů	šikmost průměrů	špičatost průměrů
1	24,74	4,848	0,682	-0,040
10	25,14	1,482	0,743	-0,199
100	25,40	0,455	0,087	-0,076
populace	μ =25,41	σ =4,932	γ_1 =0,771	γ_2 =0,189

interval spolehlivosti pro μ (výběr z $\mathrm{N}(\mu, \sigma^2)$)

[confidence interval]

- víme, že $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$, tedy $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1)$

$$\mathrm{P}(|Z| < 1{,}96) = \mathrm{P}\left(\frac{|\bar{X}-\mu|}{\sigma}\sqrt{n} < 1{,}96\right) = 0{,}95$$

- což je totéž, jako (μ se od $\bar{X}$ liší nejvýše . . .)

$$\mathrm{P}\left(|\bar{X}-\mu| < 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

- tedy (všimněte si zkracování intervalu s rostoucím n)

$$\mathrm{P}\left(\bar{X} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

- dostali jsme 95% **interval spolehlivosti**

interval spolehlivosti pro μ (výběr z $\mathrm{N}(\mu, \sigma^2)$)

[confidence interval]

- víme, že $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$, tedy $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1)$

$$\mathrm{P}(|Z| < 1{,}96) = \mathrm{P}\left(\frac{|\bar{X}-\mu|}{\sigma}\sqrt{n} < 1{,}96\right) = 0{,}95$$

- což je totéž, jako (μ se od $\bar{X}$ liší nejvýše ...)

$$\mathrm{P}\left(|\bar{X}-\mu| < 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

- tedy (všimněte si zkracování intervalu s rostoucím n)

$$\mathrm{P}\left(\bar{X} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

- dostali jsme 95% **interval spolehlivosti**

interval spolehlivosti pro μ (výběr z $\mathrm{N}(\mu, \sigma^2)$)

[confidence interval]

- víme, že $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$, tedy $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1)$

$$\mathrm{P}(|Z| < 1{,}96) = \mathrm{P}\left(\frac{|\bar{X}-\mu|}{\sigma}\sqrt{n} < 1{,}96\right) = 0{,}95$$

- což je totéž, jako (μ se od $\bar{X}$ liší nejvýše ...)

$$\mathrm{P}\left(|\bar{X}-\mu| < 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

- tedy (všimněte si zkracování intervalu s rostoucím n)

$$\mathrm{P}\left(\bar{X} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

- dostali jsme 95% **interval spolehlivosti**

interval spolehlivosti pro μ (výběr z $\mathrm{N}(\mu, \sigma^2)$)

[confidence interval]

- víme, že $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$, tedy $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1)$

$$\mathrm{P}(|Z| < 1{,}96) = \mathrm{P}\left(\frac{|\bar{X}-\mu|}{\sigma}\sqrt{n} < 1{,}96\right) = 0{,}95$$

- což je totéž, jako (μ se od $\bar{X}$ liší nejvýše ...)

$$\mathrm{P}\left(|\bar{X}-\mu| < 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

- tedy (všimněte si zkracování intervalu s rostoucím n)

$$\mathrm{P}\left(\bar{X} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

- dostali jsme 95% **interval spolehlivosti**

interpretace intervalu spolehlivosti

- je to **intervalový** odhad hodnoty μ
- $\bar{X}$ je **bodový** odhad
- **základní vlastnost**: 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé** μ (**odhadovaný parametr**)
- kdybychom postup prováděli opakovaně, pak asi v 95 % případů interval překryje skutečnou hodnotu μ, ve zbylých asi 5 % zůstane skutečné μ mimo interval spolehlivosti
- pro obecné α (spolehlivost $1 - \alpha$):

$$\mathrm{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2)\right) = 1 - \alpha$$

interpretace intervalu spolehlivosti

- je to **intervalový** odhad hodnoty μ
- $\bar{X}$ je **bodový** odhad
- **základní vlastnost**: 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé** μ (**odhadovaný parametr**)
- kdybychom postup prováděli opakovaně, pak asi v 95 % případů interval překryje skutečnou hodnotu μ, ve zbylých asi 5 % zůstane skutečné μ mimo interval spolehlivosti
- pro obecné α (spolehlivost $1 - \alpha$):

$$\mathrm{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2)\right) = 1 - \alpha$$

interpretace intervalu spolehlivosti

- je to **intervalový** odhad hodnoty μ
- $\bar{X}$ je **bodový** odhad
- **základní vlastnost**: 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé** μ (**odhadovaný parametr**)
- kdybychom postup prováděli opakovaně, pak asi v 95 % případů interval překryje skutečnou hodnotu μ, ve zbylých asi 5 % zůstane skutečné μ mimo interval spolehlivosti
- pro obecné α (spolehlivost $1 - \alpha$):

$$\mathrm{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2)\right) = 1 - \alpha$$

interpretace intervalu spolehlivosti

- je to **intervalový** odhad hodnoty μ
- $\bar{X}$ je **bodový** odhad
- **základní vlastnost**: 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé** μ (**odhadovaný parametr**)
- kdybychom postup prováděli opakovaně, pak asi v 95 % případů interval překryje skutečnou hodnotu μ, ve zbylých asi 5 % zůstane skutečné μ mimo interval spolehlivosti
- pro obecné α (spolehlivost $1-\alpha$):

$$\mathrm{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2)\right) = 1 - \alpha$$

interpretace intervalu spolehlivosti

- je to **intervalový** odhad hodnoty μ
- $\bar{X}$ je **bodový** odhad
- **základní vlastnost**: 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé** μ (**odhadovaný parametr**)
- kdybychom postup prováděli opakovaně, pak asi v 95 % případů interval překryje skutečnou hodnotu μ, ve zbylých asi 5 % zůstane skutečné μ mimo interval spolehlivosti
- pro obecné α (spolehlivost $1-\alpha$):

$$\mathrm{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2)\right) = 1 - \alpha$$

interval spolehlivosti pro neznámé σ

- pro malé n (asi do 50) a pro X_i s normálním rozdělením je třeba použít kritické hodnoty Studentova t-rozdělení (pozor na **jinak značené** kritické hodnoty Studentova t-rozdělení)

$$\mathrm{P}\left(\bar{X} - \frac{S_x}{\sqrt{n}} t_{n-1}(\alpha) < \mu < \bar{X} + \frac{S_x}{\sqrt{n}} t_{n-1}(\alpha)\right) = 1 - \alpha$$

- jako odhad σ^2 se použije výběrový rozptyl

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

- interval spolehlivosti se počítá i při odhadu jiných parametrů
- je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**

interval spolehlivosti pro neznámé σ

- pro malé n (asi do 50) a pro X_i s normálním rozdělením je třeba použít kritické hodnoty Studentova t-rozdělení (pozor na **jinak značené** kritické hodnoty Studentova t-rozdělení)

$$\mathrm{P}\left(\bar{X} - \frac{S_x}{\sqrt{n}} t_{n-1}(\alpha) < \mu < \bar{X} + \frac{S_x}{\sqrt{n}} t_{n-1}(\alpha)\right) = 1 - \alpha$$

- jako odhad σ^2 se použije výběrový rozptyl

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

- interval spolehlivosti se počítá i při odhadu jiných parametrů
- je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**

interval spolehlivosti pro neznámé σ

- pro malé n (asi do 50) a pro X_i s normálním rozdělením je třeba použít kritické hodnoty Studentova t-rozdělení (pozor na **jinak značené** kritické hodnoty Studentova t-rozdělení)

$$\mathrm{P}\left(\bar{X} - \frac{S_x}{\sqrt{n}} t_{n-1}(\alpha) < \mu < \bar{X} + \frac{S_x}{\sqrt{n}} t_{n-1}(\alpha)\right) = 1 - \alpha$$

- jako odhad σ^2 se použije výběrový rozptyl

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

- interval spolehlivosti se počítá i při odhadu jiných parametrů
- je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**

interval spolehlivosti pro neznámé σ

- pro malé n (asi do 50) a pro X_i s normálním rozdělením je třeba použít kritické hodnoty Studentova t-rozdělení (pozor na **jinak značené** kritické hodnoty Studentova t-rozdělení)

$$\mathrm{P}\left(\bar{X} - \frac{S_x}{\sqrt{n}} t_{n-1}(\alpha) < \mu < \bar{X} + \frac{S_x}{\sqrt{n}} t_{n-1}(\alpha)\right) = 1 - \alpha$$

- jako odhad σ^2 se použije výběrový rozptyl

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

- interval spolehlivosti se počítá i při odhadu jiných parametrů
- je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**

příklad: věk matek

normální rozdělení dáno CLT a velkým n

- 95% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left(25{,}7 - 1{,}98 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 1{,}98 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (24{,}9; 26{,}5)$$

[confint(lm(vek.m∼1,data=Kojeni))]

- 99% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek (bude užší nebo širší?)
- větší jistota způsobí delší interval spolehlivosti (méně vypovídající tvrzení)

$$\left(25{,}7 - 2{,}63 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 2{,}63 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (24{,}6; 26{,}8)$$

[confint(lm(vek.m∼1,data=Kojeni),level=0.99)]

příklad: věk matek

normální rozdělení dáno CLT a velkým n

- 95% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left(25{,}7 - 1{,}98 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 1{,}98 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (24{,}9; 26{,}5)$$

[confint(lm(vek.m∼1,data=Kojeni))]

- 99% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek (bude užší nebo širší?)
- větší jistota způsobí delší interval spolehlivosti (méně vypovídající tvrzení)

$$\left(25{,}7 - 2{,}63 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 2{,}63 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (24{,}6; 26{,}8)$$

[confint(lm(vek.m∼1,data=Kojeni),level=0.99)]

příklad: věk matek

normální rozdělení dáno CLT a velkým n

- 95% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left(25{,}7 - 1{,}98 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 1{,}98 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (24{,}9; 26{,}5)$$

[confint(lm(vek.m∼1,data=Kojeni))]

- 99% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek (bude užší nebo širší?)
- větší jistota způsobí delší interval spolehlivosti (méně vypovídající tvrzení)

$$\left(25{,}7 - 2{,}63 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 2{,}63 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (24{,}6; 26{,}8)$$

[confint(lm(vek.m∼1,data=Kojeni),level=0.99)]

příklad: věk matek II

normální rozdělení dáno CLT a velkým n

- 90% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left(25{,}7 - 1{,}66 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 1{,}66 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (25{,}0; 26{,}4)$$

[confint(lm(vek.m∼1,data=Kojeni),level=0.9)]

- příklady **nesprávné interpretace** 90% intervalu spolehlivosti:
  - *90 % žen má věk v intervalu (25,0; 26,4)*
    např. mezi našimi 99 matkami je jen 12 ve věku 25 a 10 věku 26 roků, navíc, s rostoucím n se interval zužuje
  - *výběrový průměr věku matek je s pravděpodobností 90 % v intervalu (25,0; 26,4)*
    výběrový průměr je **vždy uvnitř** intervalu

příklad: věk matek II

normální rozdělení dáno CLT a velkým n

- 90% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left(25{,}7 - 1{,}66 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 1{,}66 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (25{,}0; 26{,}4)$$

[confint(lm(vek.m∼1,data=Kojeni),level=0.9)]

- příklady **nesprávné interpretace** 90% intervalu spolehlivosti:
  - *90 % žen má věk v intervalu (25,0; 26,4)*
    např. mezi našimi 99 matkami je jen 12 ve věku 25 a 10 věku 26 roků, navíc, s rostoucím n se interval zužuje
  - *výběrový průměr věku matek je s pravděpodobností 90 % v intervalu (25,0; 26,4)*
    výběrový průměr je **vždy uvnitř** intervalu

příklad: věk matek II

normální rozdělení dáno CLT a velkým n

- 90% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left(25{,}7 - 1{,}66 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 1{,}66 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (25{,}0; 26{,}4)$$

[confint(lm(vek.m∼1,data=Kojeni),level=0.9)]

- příklady **nesprávné interpretace** 90% intervalu spolehlivosti:
  - *90 % žen má věk v intervalu (25,0; 26,4)*
    např. mezi našimi 99 matkami je jen 12 ve věku 25 a 10 věku 26 roků, navíc, s rostoucím n se interval zužuje
  - *výběrový průměr věku matek je s pravděpodobností 90 % v intervalu (25,0; 26,4)*
    výběrový průměr je **vždy uvnitř** intervalu

příklad: věk matek II

normální rozdělení dáno CLT a velkým n

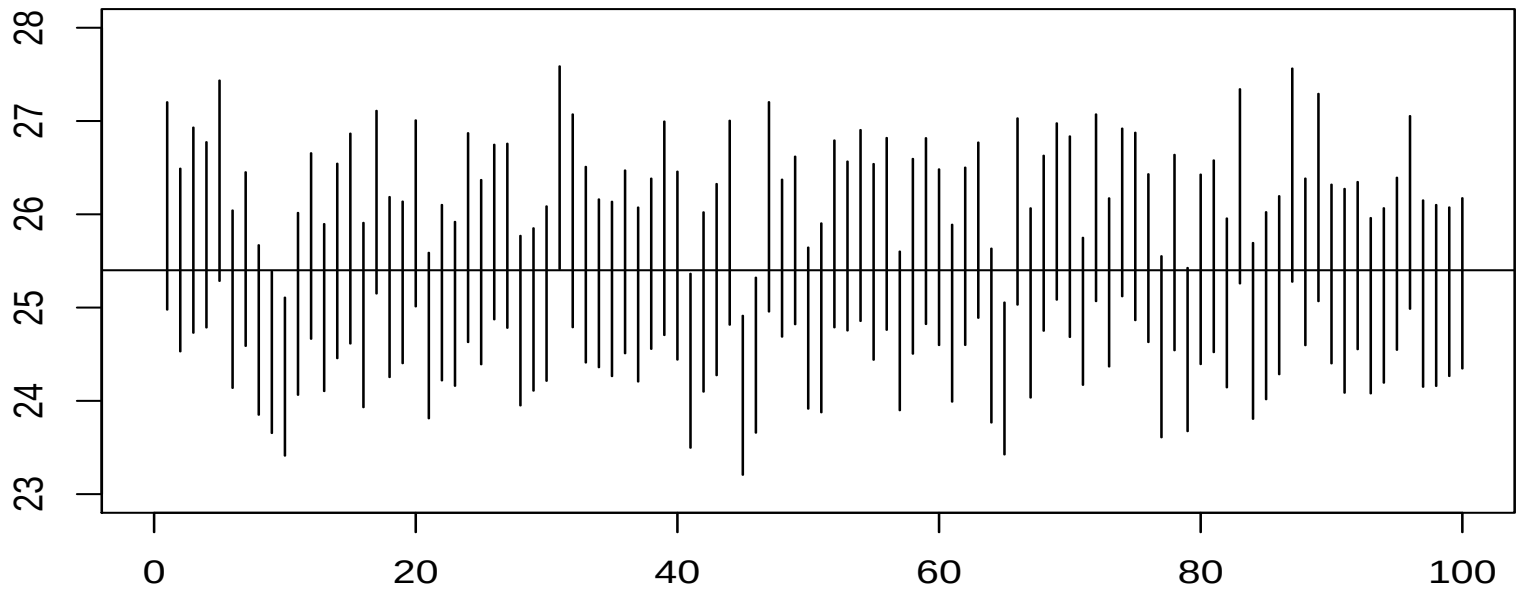
- 90% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left(25{,}7 - 1{,}66 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 1{,}66 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (25{,}0; 26{,}4)$$

[confint(lm(vek.m∼1,data=Kojeni),level=0.9)]

- příklady **nesprávné interpretace** 90% intervalu spolehlivosti:
  - *90 % žen má věk v intervalu (25,0; 26,4)*
    např. mezi našimi 99 matkami je jen 12 ve věku 25 a 10 věku 26 roků, navíc, s rostoucím n se interval zužuje
  - *výběrový průměr věku matek je s pravděpodobností 90 % v intervalu (25,0; 26,4)*
    výběrový průměr je **vždy uvnitř** intervalu

simulované výběry pro $n = 100$ (věk matek)



znázorněno celkem 100 95% intervalů spolehlivosti pro μ
ve skutečnosti mimořádně víme, že $\mu = 25{,}4$
v 7 případech je μ nepřekryto

centrální limitní věta pro četnosti

- (CLT obecně:) Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$. Potom pro velké n má průměr z nich rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, jejich součet rozdělení $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- $Y \sim \mathrm{bi}(n, \pi)$: Y je absolutní četnost výskytu jevu s pstí π v n nezáv. pokusech
- $Y = \sum_{i=1}^{n} X_i$ je součet nezávislých náhodných veličin X_i s alternativním rozdělením, $X_i \sim \mathrm{alt}(\pi), \mathrm{var}\, X_i = \pi(1-\pi)$
- podle CLT proto přibližně $Y \sim \mathrm{N}(n\pi, n\pi(1-\pi))$
- relativní četnost $Y/n = \bar{X}$ je průměr veličin s alternativním rozdělením, označme $\hat{\pi} = Y/n$
- podle CLT je přibližně $\hat{\pi} \sim \mathrm{N}(\pi, \pi(1-\pi)/n)$
- $\hat{\pi}$ je **nestranný** odhad π

centrální limitní věta pro četnosti

- (CLT obecně:) Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$. Potom pro velké n má průměr z nich rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, jejich součet rozdělení $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- $Y \sim \mathrm{bi}(n, \pi)$: Y je absolutní četnost výskytu jevu s pstí π v n nezáv. pokusech
- $Y = \sum_{i=1}^{n} X_i$ je součet nezávislých náhodných veličin X_i s alternativním rozdělením, $X_i \sim \mathrm{alt}(\pi), \mathrm{var}\, X_i = \pi(1-\pi)$
- podle CLT proto přibližně $Y \sim \mathrm{N}(n\pi, n\pi(1-\pi))$
- relativní četnost $Y/n = \bar{X}$ je průměr veličin s alternativním rozdělením, označme $\hat{\pi} = Y/n$
- podle CLT je přibližně $\hat{\pi} \sim \mathrm{N}(\pi, \pi(1-\pi)/n)$
- $\hat{\pi}$ je **nestranný** odhad π

centrální limitní věta pro četnosti

- (CLT obecně:) Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$. Potom pro velké n má průměr z nich rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, jejich součet rozdělení $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- $Y \sim \mathrm{bi}(n, \pi)$: Y je absolutní četnost výskytu jevu s pstí π v n nezáv. pokusech
- $Y = \sum_{i=1}^{n} X_i$ je součet nezávislých náhodných veličin X_i s alternativním rozdělením, $X_i \sim \mathrm{alt}(\pi), \mathrm{var}\, X_i = \pi(1-\pi)$
- podle CLT proto přibližně $Y \sim \mathrm{N}(n\pi, n\pi(1-\pi))$
- relativní četnost $Y/n = \bar{X}$ je průměr veličin s alternativním rozdělením, označme $\hat{\pi} = Y/n$
- podle CLT je přibližně $\hat{\pi} \sim \mathrm{N}(\pi, \pi(1-\pi)/n)$
- $\hat{\pi}$ je **nestranný** odhad π

centrální limitní věta pro četnosti

- (CLT obecně:) Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$. Potom pro velké n má průměr z nich rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, jejich součet rozdělení $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- $Y \sim \mathrm{bi}(n, \pi)$: Y je absolutní četnost výskytu jevu s pstí π v n nezáv. pokusech
- $Y = \sum_{i=1}^{n} X_i$ je součet nezávislých náhodných veličin X_i s alternativním rozdělením, $X_i \sim \mathrm{alt}(\pi), \mathrm{var}\, X_i = \pi(1-\pi)$
- podle CLT proto přibližně $Y \sim \mathrm{N}(n\pi, n\pi(1-\pi))$
- relativní četnost $Y/n = \bar{X}$ je průměr veličin s alternativním rozdělením, označme $\hat{\pi} = Y/n$
- podle CLT je přibližně $\hat{\pi} \sim \mathrm{N}(\pi, \pi(1-\pi)/n)$
- $\hat{\pi}$ je **nestranný** odhad π

centrální limitní věta pro četnosti

- (CLT obecně:) Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$. Potom pro velké n má průměr z nich rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, jejich součet rozdělení $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- $Y \sim \mathrm{bi}(n, \pi)$: Y je absolutní četnost výskytu jevu s pstí π v n nezáv. pokusech
- $Y = \sum_{i=1}^{n} X_i$ je součet nezávislých náhodných veličin X_i s alternativním rozdělením, $X_i \sim \mathrm{alt}(\pi), \mathrm{var}\, X_i = \pi(1-\pi)$
- podle CLT proto přibližně $Y \sim \mathrm{N}(n\pi, n\pi(1-\pi))$
- relativní četnost $Y/n = \bar{X}$ je průměr veličin s alternativním rozdělením, označme $\hat{\pi} = Y/n$
- podle CLT je přibližně $\hat{\pi} \sim \mathrm{N}(\pi, \pi(1-\pi)/n)$
- $\hat{\pi}$ je **nestranný** odhad π

centrální limitní věta pro četnosti

- (CLT obecně:) Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$. Potom pro velké n má průměr z nich rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, jejich součet rozdělení $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- $Y \sim \mathrm{bi}(n, \pi)$: Y je absolutní četnost výskytu jevu s pstí π v n nezáv. pokusech
- $Y = \sum_{i=1}^{n} X_i$ je součet nezávislých náhodných veličin X_i s alternativním rozdělením, $X_i \sim \mathrm{alt}(\pi), \mathrm{var}\, X_i = \pi(1-\pi)$
- podle CLT proto přibližně $Y \sim \mathrm{N}(n\pi, n\pi(1-\pi))$
- relativní četnost $Y/n = \bar{X}$ je průměr veličin s alternativním rozdělením, označme $\hat{\pi} = Y/n$
- podle CLT je přibližně $\hat{\pi} \sim \mathrm{N}(\pi, \pi(1-\pi)/n)$
- $\hat{\pi}$ je **nestranný** odhad π

centrální limitní věta pro četnosti

- (CLT obecně:) Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$. Potom pro velké n má průměr z nich rozdělení $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, jejich součet rozdělení $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- $Y \sim \mathrm{bi}(n, \pi)$: Y je absolutní četnost výskytu jevu s pstí π v n nezáv. pokusech
- $Y = \sum_{i=1}^{n} X_i$ je součet nezávislých náhodných veličin X_i s alternativním rozdělením, $X_i \sim \mathrm{alt}(\pi), \mathrm{var}\, X_i = \pi(1-\pi)$
- podle CLT proto přibližně $Y \sim \mathrm{N}(n\pi, n\pi(1-\pi))$
- relativní četnost $Y/n = \bar{X}$ je průměr veličin s alternativním rozdělením, označme $\hat{\pi} = Y/n$
- podle CLT je přibližně $\hat{\pi} \sim \mathrm{N}(\pi, \pi(1-\pi)/n)$
- $\hat{\pi}$ je **nestranný** odhad π

interval spolehlivosti pro pravděpodobnost π

- odmocnina z rozptylu odhadu $\hat{\pi}$ je $\sqrt{\frac{\pi(1-\pi)}{n}}$
- střední chyba relativní četnosti = směrodatná odchylka relativní četnosti
- pravděpodobnost π neznáme, odhadneme ji pomocí relativní četnosti $\hat{\pi} = Y/n$
- odtud je $100(1-\alpha)\%$ přibližný interval spolehlivosti pro π

$$\left(\hat{\pi} - z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}; \hat{\pi} + z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)$$

- existují přesnější (pracnější) postupy
  [prop.test(y,n,correct=FALSE)]
  [binom.test(y,n)]

interval spolehlivosti pro pravděpodobnost π

- odmocnina z rozptylu odhadu $\hat{\pi}$ je $\sqrt{\frac{\pi(1-\pi)}{n}}$
- střední chyba relativní četnosti = směrodatná odchylka relativní četnosti
- pravděpodobnost π neznáme, odhadneme ji pomocí relativní četnosti $\hat{\pi} = Y/n$
- odtud je $100(1-\alpha)\%$ přibližný interval spolehlivosti pro π

$$\left(\hat{\pi} - z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}; \hat{\pi} + z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)$$

- existují přesnější (pracnější) postupy
  [prop.test(y,n,correct=FALSE)]
  [binom.test(y,n)]

interval spolehlivosti pro pravděpodobnost π

- odmocnina z rozptylu odhadu $\hat{\pi}$ je $\sqrt{\frac{\pi(1-\pi)}{n}}$
- střední chyba relativní četnosti = směrodatná odchylka relativní četnosti
- pravděpodobnost π neznáme, odhadneme ji pomocí relativní četnosti $\hat{\pi} = Y/n$
- odtud je $100(1-\alpha)\%$ přibližný interval spolehlivosti pro π

$$\left(\hat{\pi} - z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}; \hat{\pi} + z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)$$

- existují přesnější (pracnější) postupy
  [prop.test(y,n,correct=FALSE)]
  [binom.test(y,n)]

interval spolehlivosti pro pravděpodobnost π

- odmocnina z rozptylu odhadu $\hat{\pi}$ je $\sqrt{\frac{\pi(1-\pi)}{n}}$
- střední chyba relativní četnosti = směrodatná odchylka relativní četnosti
- pravděpodobnost π neznáme, odhadneme ji pomocí relativní četnosti $\hat{\pi} = Y/n$
- odtud je $100(1-\alpha)\%$ přibližný interval spolehlivosti pro π

$$\left(\hat{\pi} - z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}; \hat{\pi} + z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)$$

- existují přesnější (pracnější) postupy
  [prop.test(y,n,correct=FALSE)]
  [binom.test(y,n)]

interval spolehlivosti pro pravděpodobnost π

- odmocnina z rozptylu odhadu $\hat{\pi}$ je $\sqrt{\frac{\pi(1-\pi)}{n}}$
- střední chyba relativní četnosti = směrodatná odchylka relativní četnosti
- pravděpodobnost π neznáme, odhadneme ji pomocí relativní četnosti $\hat{\pi} = Y/n$
- odtud je $100(1-\alpha)\%$ přibližný interval spolehlivosti pro π

$$\left(\hat{\pi} - z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}; \hat{\pi} + z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)$$

- existují přesnější (pracnější) postupy
  [prop.test(y,n,correct=FALSE)]
  [binom.test(y,n)]

příklad: hody s hrací kostkou

- odhadujeme pravděpodobnost šestky, $\alpha = 0{,}05$
- kostka A: $n = 100, y = 17, \hat{\pi}_A = 0{,}17$

$$\left(0{,}17 - 1{,}96 \cdot \sqrt{\frac{0{,}17 \cdot 0{,}83}{100}}; 0{,}17 + 1{,}96 \cdot \sqrt{\frac{0{,}17 \cdot 0{,}83}{100}}\right) = (0{,}10; 0{,}24)$$

- kostka B: $n = 100, y = 41, \hat{\pi}_B = 0{,}41$

$$\left(0{,}41 - 1{,}96 \cdot \sqrt{\frac{0{,}41 \cdot 0{,}59}{100}}; 0{,}41 + 1{,}96 \cdot \sqrt{\frac{0{,}41 \cdot 0{,}59}{100}}\right) = (0{,}31; 0{,}51)$$

- důležitý rozdíl: u kostky A patří $1/6 = 0{,}167$ do 95% intervalu spolehlivosti; u kostky B nikoliv

příklad: hody s hrací kostkou

- odhadujeme pravděpodobnost šestky, $\alpha = 0{,}05$
- kostka A: $n = 100, y = 17, \hat{\pi}_A = 0{,}17$

$$\left(0{,}17 - 1{,}96 \cdot \sqrt{\frac{0{,}17 \cdot 0{,}83}{100}}; 0{,}17 + 1{,}96 \cdot \sqrt{\frac{0{,}17 \cdot 0{,}83}{100}}\right) = (0{,}10; 0{,}24)$$

- kostka B: $n = 100, y = 41, \hat{\pi}_B = 0{,}41$

$$\left(0{,}41 - 1{,}96 \cdot \sqrt{\frac{0{,}41 \cdot 0{,}59}{100}}; 0{,}41 + 1{,}96 \cdot \sqrt{\frac{0{,}41 \cdot 0{,}59}{100}}\right) = (0{,}31; 0{,}51)$$

- důležitý rozdíl: u kostky A patří $1/6 = 0{,}167$ do 95% intervalu spolehlivosti; u kostky B nikoliv

příklad: hody s hrací kostkou

- odhadujeme pravděpodobnost šestky, $\alpha = 0{,}05$
- kostka A: $n = 100, y = 17, \hat{\pi}_A = 0{,}17$

$$\left(0{,}17 - 1{,}96 \cdot \sqrt{\frac{0{,}17 \cdot 0{,}83}{100}}; 0{,}17 + 1{,}96 \cdot \sqrt{\frac{0{,}17 \cdot 0{,}83}{100}}\right) = (0{,}10; 0{,}24)$$

- kostka B: $n = 100, y = 41, \hat{\pi}_B = 0{,}41$

$$\left(0{,}41 - 1{,}96 \cdot \sqrt{\frac{0{,}41 \cdot 0{,}59}{100}}; 0{,}41 + 1{,}96 \cdot \sqrt{\frac{0{,}41 \cdot 0{,}59}{100}}\right) = (0{,}31; 0{,}51)$$

- důležitý rozdíl: u kostky A patří $1/6 = 0{,}167$ do 95% intervalu spolehlivosti; u kostky B nikoliv

příklad: hody s hrací kostkou

- odhadujeme pravděpodobnost šestky, $\alpha = 0{,}05$
- kostka A: $n = 100, y = 17, \hat{\pi}_A = 0{,}17$

$$\left(0{,}17 - 1{,}96 \cdot \sqrt{\frac{0{,}17 \cdot 0{,}83}{100}}; 0{,}17 + 1{,}96 \cdot \sqrt{\frac{0{,}17 \cdot 0{,}83}{100}}\right) = (0{,}10; 0{,}24)$$

- kostka B: $n = 100, y = 41, \hat{\pi}_B = 0{,}41$

$$\left(0{,}41 - 1{,}96 \cdot \sqrt{\frac{0{,}41 \cdot 0{,}59}{100}}; 0{,}41 + 1{,}96 \cdot \sqrt{\frac{0{,}41 \cdot 0{,}59}{100}}\right) = (0{,}31; 0{,}51)$$

- důležitý rozdíl: u kostky A patří $1/6 = 0{,}167$ do 95% intervalu spolehlivosti; u kostky B nikoliv