Statistika

(D360P03Z, D360P03U)

Karel Zvára

6. přednáška konaná

12. listopadu 2004

chování výběrového průměru

- nechť $X_1, X_2, \ldots X_n$ jsou nezávislé náhodné veličiny s libovolným rozdělením se střední hodnotou μ a rozptylem σ^2 - náhodný výběr z onoho rozdělení
- pro průměr z těchto veličin platí

$$\mathsf{E}\,(\bar{X}) = \mathsf{E}\,\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mu \qquad\qquad \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

- průměr $\bar{X}$ má tedy rozptyl n-krát menší, než jednotlivá pozorování
- **střední chyba** průměru = směrodatná odchylka průměru

$$\mathsf{S.E.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

výběrový průměr z normálního rozdělení

- nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny s rozdělením $\mathsf{N}\left(\mu, \sigma^2\right)$ – **náhodný výběr** z $\mathsf{N}\left(\mu, \sigma^2\right)$
- pro průměr z nich platí

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

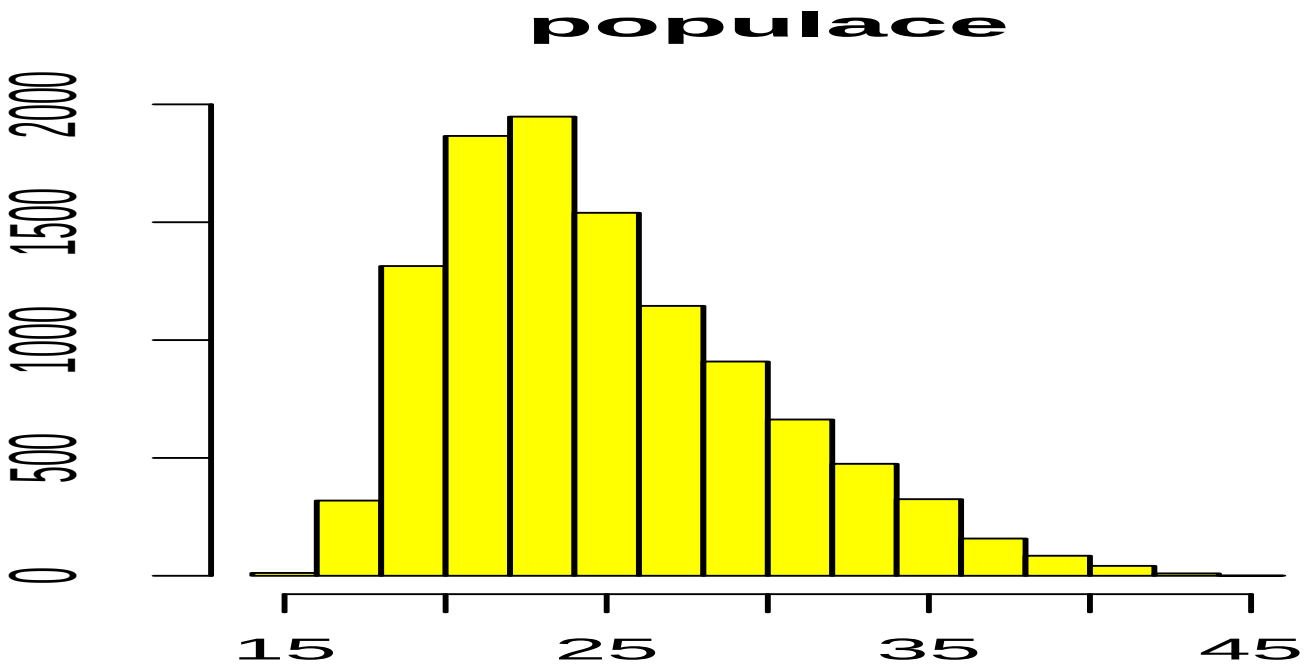
- opět je střední chyba $\bar{X}$ rovna $\frac{\sigma}{\sqrt{n}}$
- proto je

$$Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim \mathsf{N}(0, 1)$$

- chování Z lze popsat pomocí distribuční funkce $\Phi(z)$

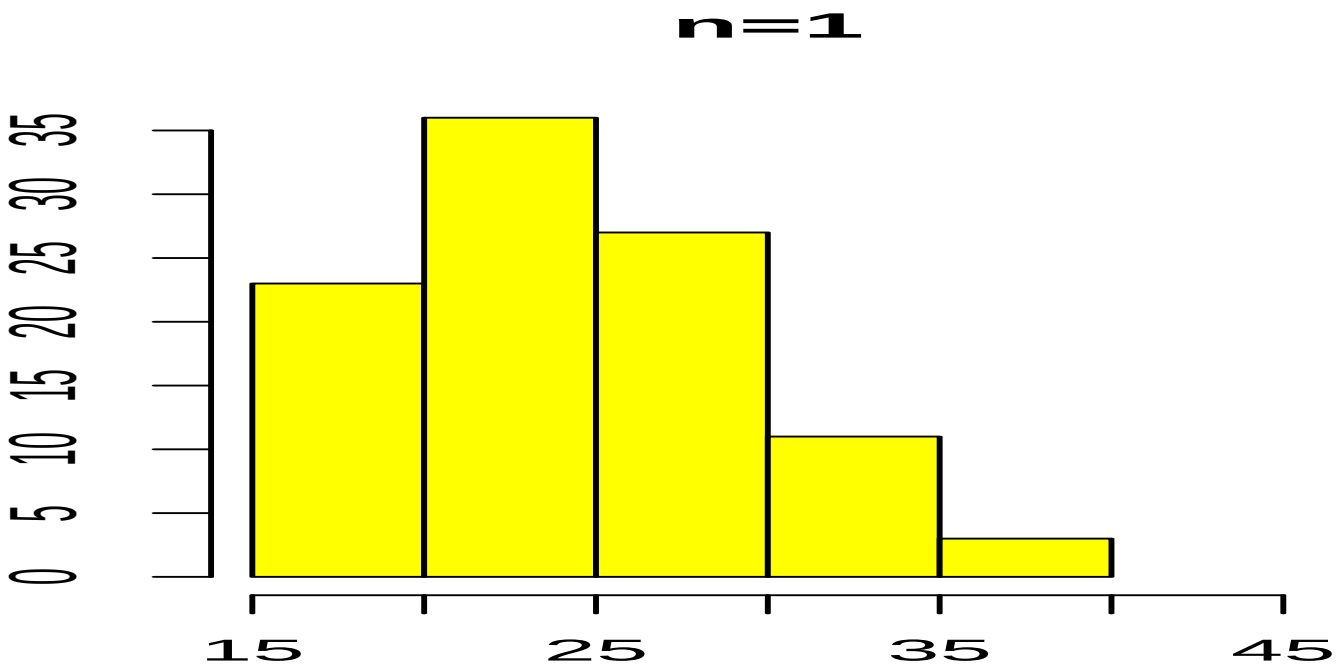
příklad: věk matek

- velká populace rodičů (11 tisíc)



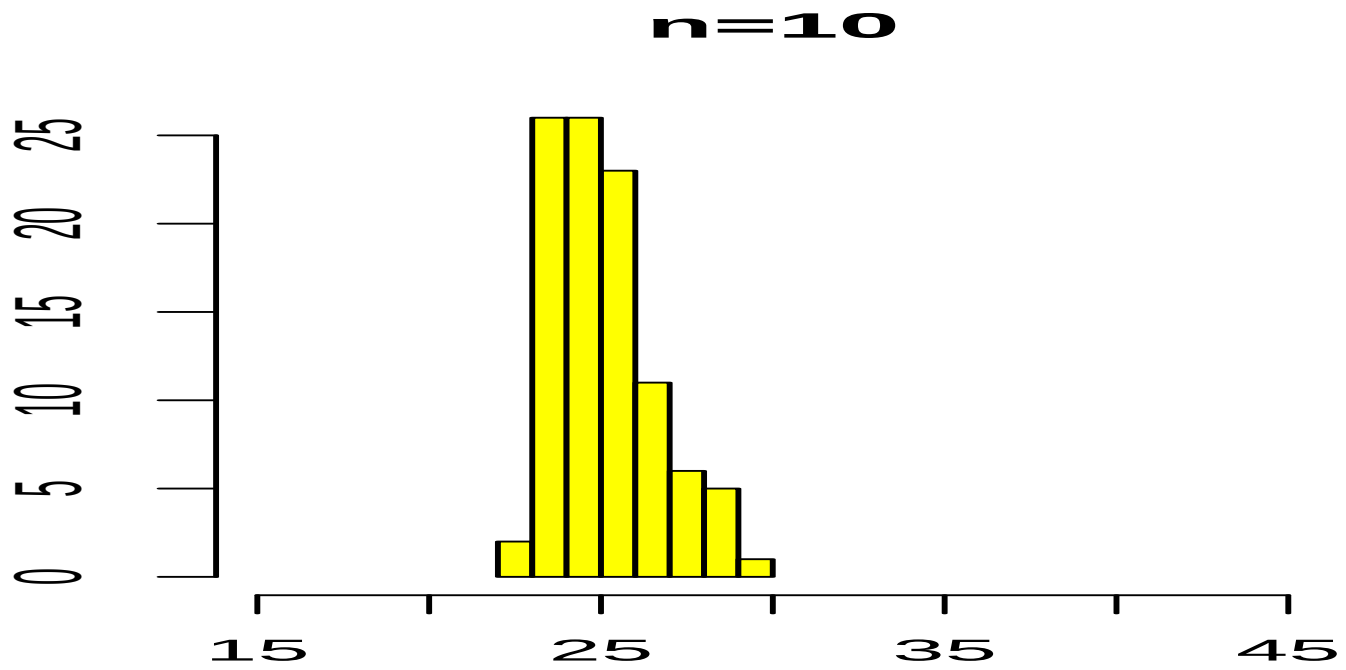
příklad: věk matek

- náhodně vybráno 100 matek



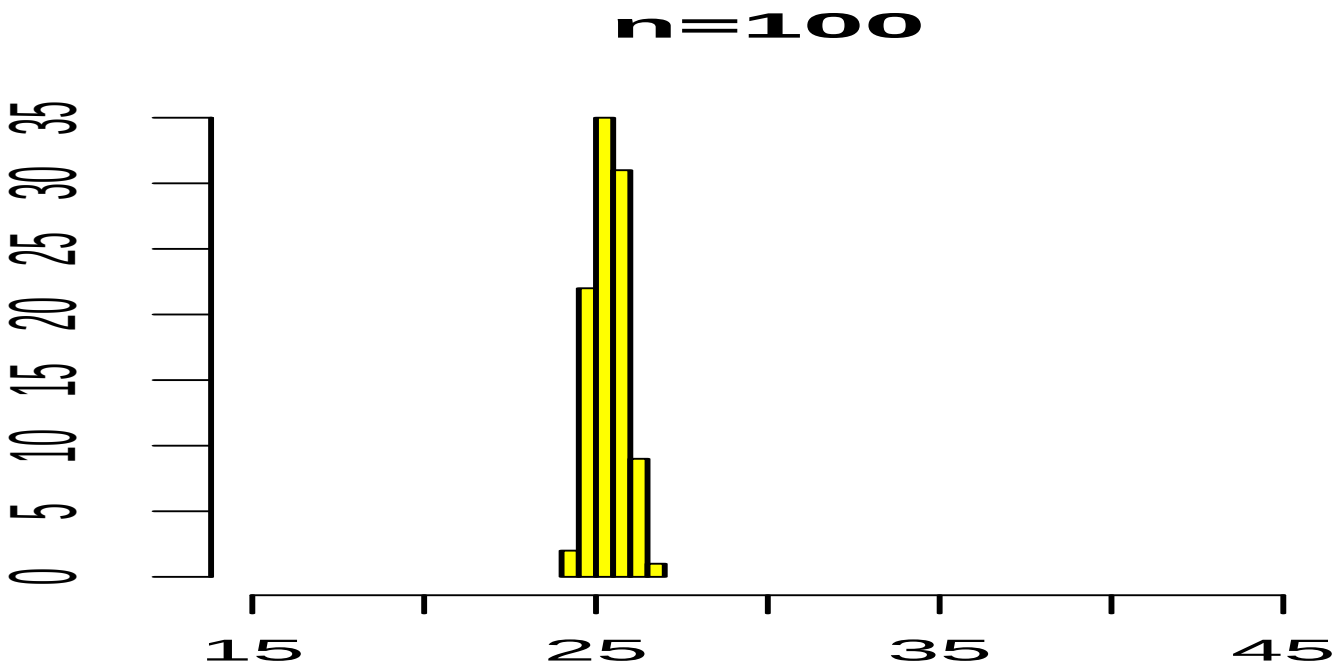
příklad: věk matek

- náhodně vybráno 100 krát po $n = 10$ matkách, průměry:



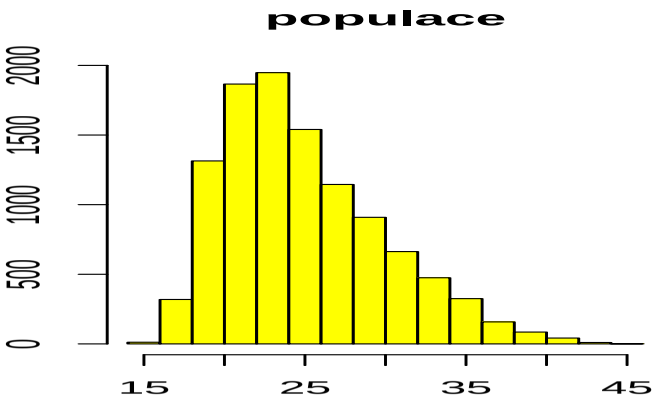
příklad: věk matek

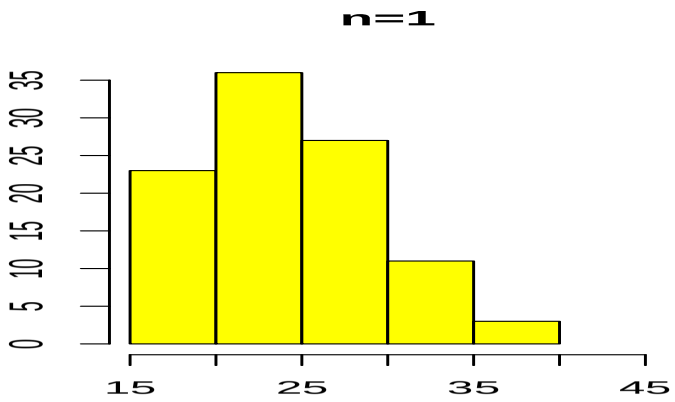
- náhodně vybráno 100 krát po $n = 100$ matkách, průměry:

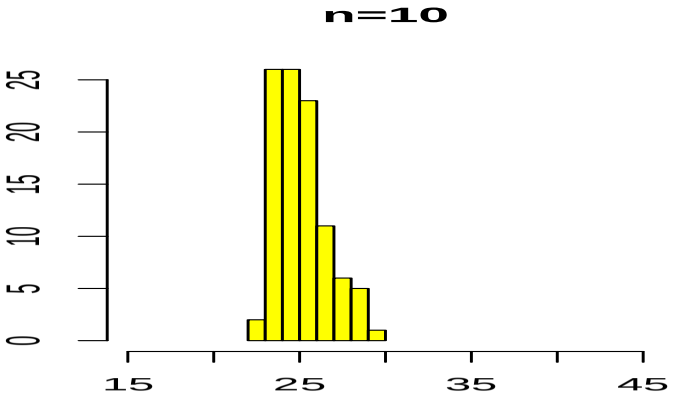


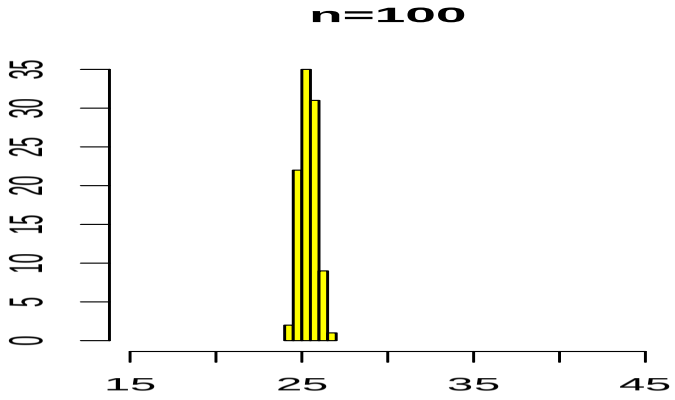
příklad: věk matek

- velká populace rodičů (11 tisíc)
- náhodně vybráno 100 matek (vlastně průměry výběrů rozsahu $n = 1$), nakreslen histogram
- 100 krát náhodně vybráno vždy $n = 10$ matek, spočítán průměr, nakreslen histogram průměrů
- 100 krát náhodně vybráno vždy $n = 100$ matek, spočítán průměr, nakreslen histogram průměrů
- podle teorie by každý další rozptyl ze 100 průměrů měl být 10 krát menší
- skutečnost: 23,5; 2,20; 0,21


populace
0
500
1000
1500
2000
15
25
35
45

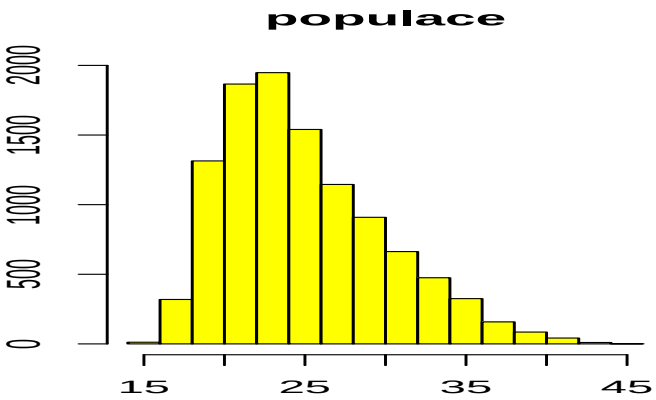

n=1
0
5
10
15
20
25
30
35
15
25
35
45

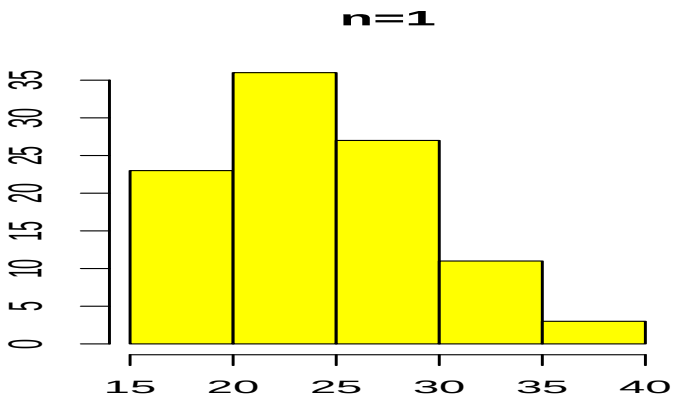

n=10
0
5
10
15
20
25
15
25
35
45

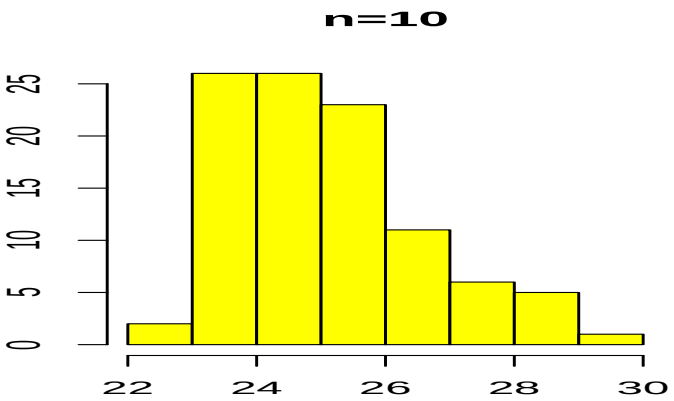

n=100
0
5
10
15
20
25
30
35
15
25
35
45

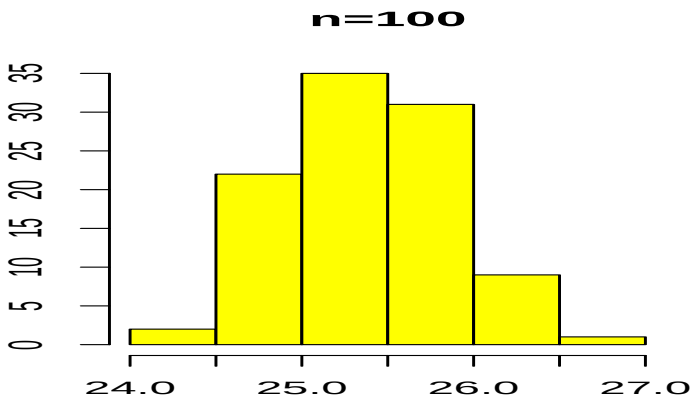
centrální limitní věta

- 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 $\mathrm{N}\left(n\mu, n\sigma^2\right)$.
- prakticky: pro dost velká n má průměr normální rozdělení
- příklad: průměrný věk matek z velkých výběrů už (téměř) normální rozdělení


populace
0
500
1000
1500
2000
15
25
35
45


n=1
0
5
10
15
20
25
30
35
15
20
25
30
35
40


n=10
0
5
10
15
20
25
22
24
26
28
30


n=100
0
5
10
15
20
25
30
35
24.0
25.0
26.0
27.0

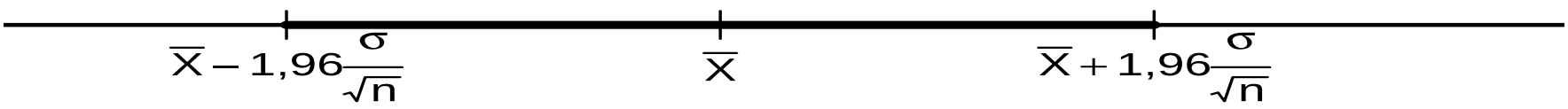
interval spolehlivosti (1)

- protože je $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, platí

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

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

- dostali jsme **95% interval spolehlivosti** pro μ



interval spolehlivosti (2)

- 95% interval spolehlivosti překryje s pravděpodobností 95 % neznámé μ (odhadovaný parametr)
- kdybychom postup prováděli opakovaně, pak asi v 95 % případů překryjeme skutečnou hodnotu μ, ve zbylých asi 5 % zůstane skutečné μ mimo interval spolehlivosti
- pro velké n lze neznámé σ nahradit odhadem s_x
- pro obecné α:

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

interval spolehlivosti (3)

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

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

- interval spolehlivosti lze počítat i pro jiné parametry
- je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**

příklad: věk matek

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

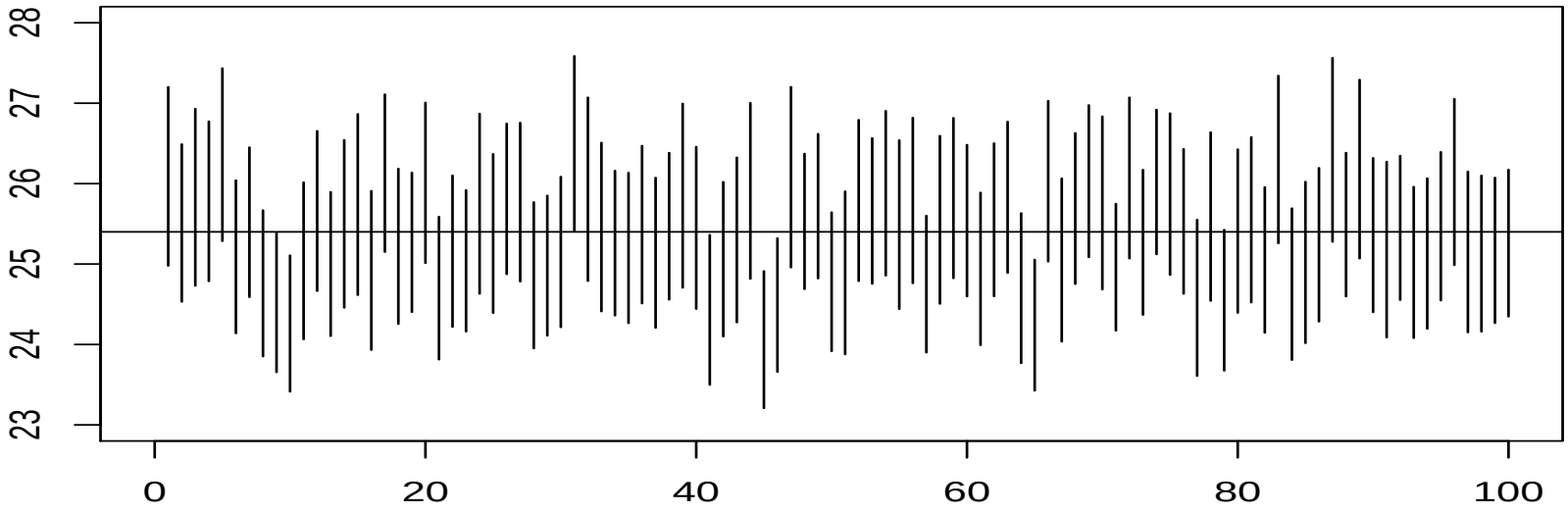
$$\left(25{,}7 - 1{,}98 \cdot \frac{4{,}1}{\sqrt{99}}; 25{,}7 + 1{,}98 \cdot \frac{4{,}1}{\sqrt{99}}\right) = (24{,}9; 26{,}5)$$

- 99% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek (bude užší nebo širší?)

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

- větší jistota $\Leftrightarrow$ větší šířka

příklad: simulované výběry pro $n = 100$



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

centrální limitní věta pro četnosti

- 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 $\mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)$, jejich součet rozdělení $\mathsf{N}\left(n\mu, n\sigma^2\right)$.
- absolutní četnost Y
  - Y – součet veličin s alternativním rozdělením
  - $Y \sim \mathsf{bi}(n, \pi)$, proto přibližně $Y \sim \mathsf{N}(n\pi, n\pi(1-\pi))$
- relativní četnost $f = Y/n$
  - f – průměr veličin s alternativním rozdělením
  - $f \sim \mathsf{N}(\pi, \pi(1-\pi)/n)$

interval spolehlivosti pro podíl (1)

- populace: **podíl** π prvků s danou vlastností
- π – **pravděpodobnost**, že vlastnost má náhodně vybraný prvek
- výběr: **relativní četnost** ve výběru
- relativní četnost je průměr nula-jedničkové veličiny – pro velké n má přibližně normální rozdělení
- nula-jedničková veličina má rozptyl $\pi(1-\pi)$
- relativní četnost (=průměr) má rozptyl $\frac{\pi(1-\pi)}{n}$

interval spolehlivosti pro podíl (2)

- střední chyba relativní četnosti = směrodatná odchylka relativní četnosti = odmocnina z rozptylu je tedy $\sqrt{\frac{\pi(1-\pi)}{n}}$
- pravděpodobnost π neznáme, odhadneme ji pomocí relativní četnosti f
- odtud je 95% interval spolehlivosti pro π

$$\left(f - 1{,}96 \cdot \sqrt{\frac{f(1-f)}{n}}; f + 1{,}96 \cdot \sqrt{\frac{f(1-f)}{n}} \right)$$

- existuje přesnější (pracnější) postup

příklad: hody s hrací kostkou

- odhadujeme pravděpodobnost šestky
- kostka A: $n = 100, n_A = 17, f_A = 0{,}17$

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

- kostka B: $n = 100, n_B = 41, f_B = 0{,}41$

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

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

proč **testování hypotéz**

- nelze bezpečně poznat, že kostka B je falešná nebo že kostka A není falešná
- intervaly spolehlivosti určily rozmezí, kde by skutečná pravděpodobnost šestky měla být, jejich spolehlivost je velká, ale omezená
- znamená něco, když $1/6$ neleží v 95% intervalu spolehlivosti?
- musíme připustit, že jsme mohli mít smůlu, že se v našich pokusech náhodou realizovaly málo pravděpodobné možnosti, přestože k takové smůle dochází jen zřídka