ANALÝZA DAT V R

4. CENTRÁLNÍ LIMITNÍ VĚTA, TESTOVÁNÍ HYPOTÉZ

Mgr. Markéta Pavlíková
Katedra pravděpodobnosti a matematické statistiky MFF UK

www.biostatisticka.cz

NORMÁLNÍ (GAUSSOVO) ROZDĚLENÍ

- značíme $N(\mu,\sigma^2)$,
- spojité rozdělení
- interval hodnot $(-\infty, \infty)$
- symetrické okolo střední hodnoty μ
- $X \sim \mathrm{N}(\mu, \sigma^2)$, pak $Z = \frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)}$$

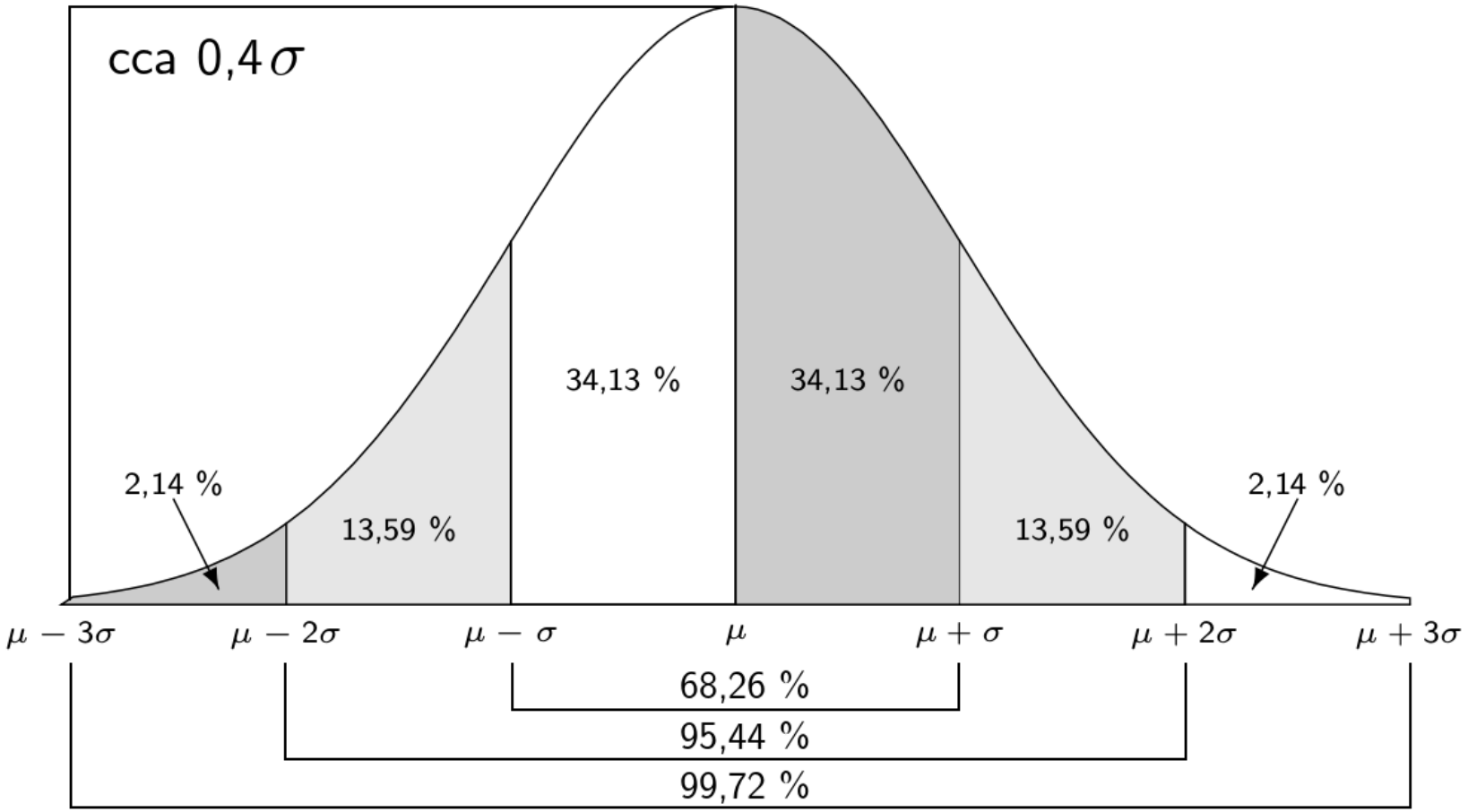
$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$$

- model vzniku: součet velkého počtu nepatrně velkých příspěvků (uvidíme i dále)
- proto tak hezky modeluje fenotypy s multifaktoriální etiologií

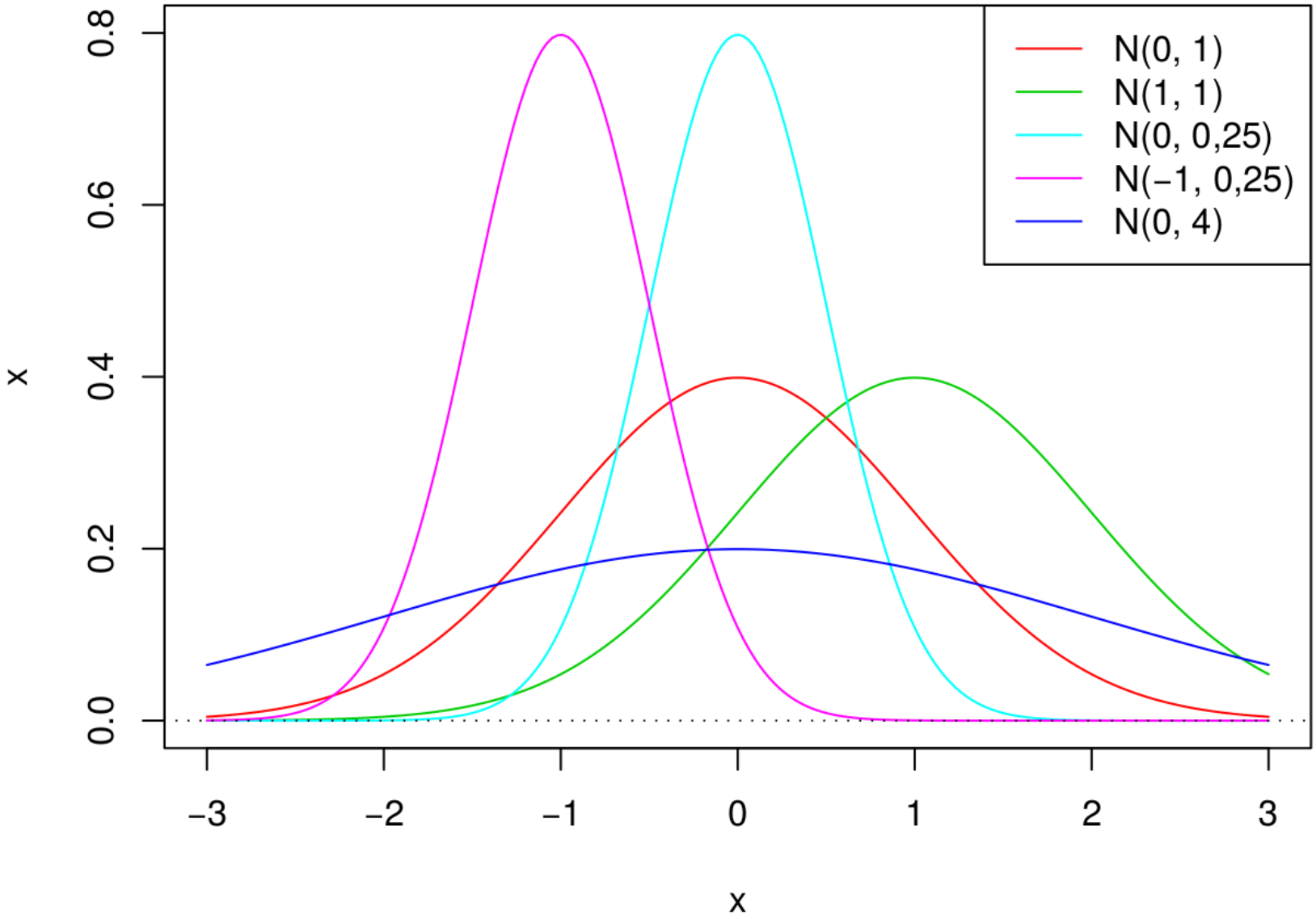
NORMÁLNÍ (GAUSSOVO) ROZDĚLENÍ

graf hustoty $N(\mu,\sigma^2)$

výpočet hustoty: [dnorm(x,mu,sigma)]



NORMÁLNÍ (GAUSSOVO) ROZDĚLENÍ

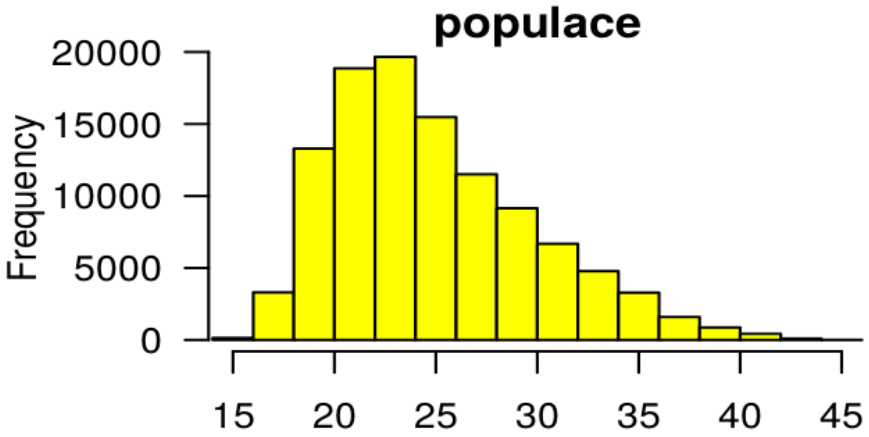


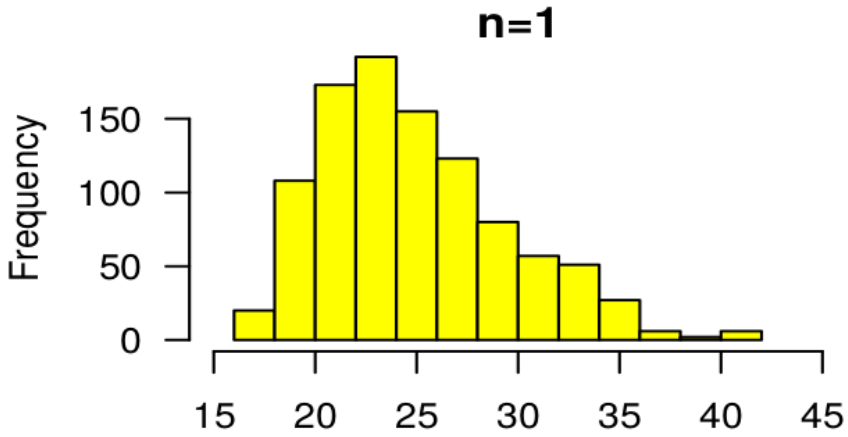
CENTRÁLNÍ LIMITNÍ VĚTA

JAK SE PŘIBLIŽOVAT REALITĚ

- modelový příklad: chceme zjistit průměrný věk rodiček v ČR v roce 2016
- rodičky v roce 2016 jsou populace s populačním průměrem věku μ, rozdělení neznáme (a je to jedno)
- Pokus č. 1: náhodně oslovíme ženu s kojencem na ulici zeptáme se na její věk = realizace náhodné veličiny „věk při porodu“ se střední hodnotou μ
- její věk použiji jako odhad μ
- jak blízko skutečnosti ten odhad je? co se stane, když oslovím jinou?
- oslovím dalších 999 matek, mám 1000 odhadů

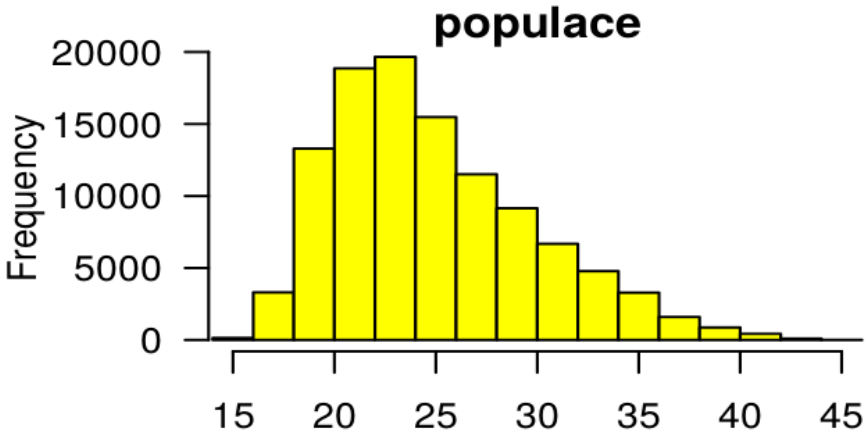
MODELOVÝ PŘÍKLAD – VĚKY RODIČEK

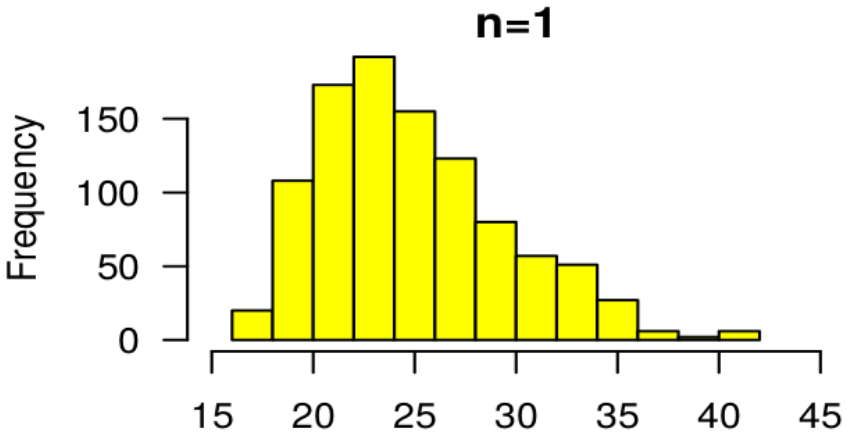


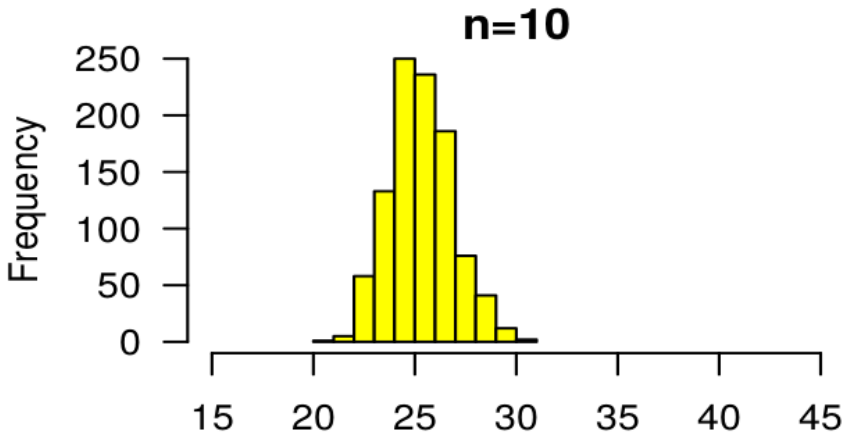


- výběrový rozptyl mých odhadů je 23.5 (± 4.6 roku)
- zeptat se jen jedné ženy zjevně není dost přesné
- zeptám se tedy deseti a udělám průměr
- bude výsledek lepší? jak poznám, zda je lepší?
- zopakuji 999x, uvidím, jak se mi ty průměry od sebe liší

MODELOVÝ PŘÍKLAD – VĚKY RODIČEK

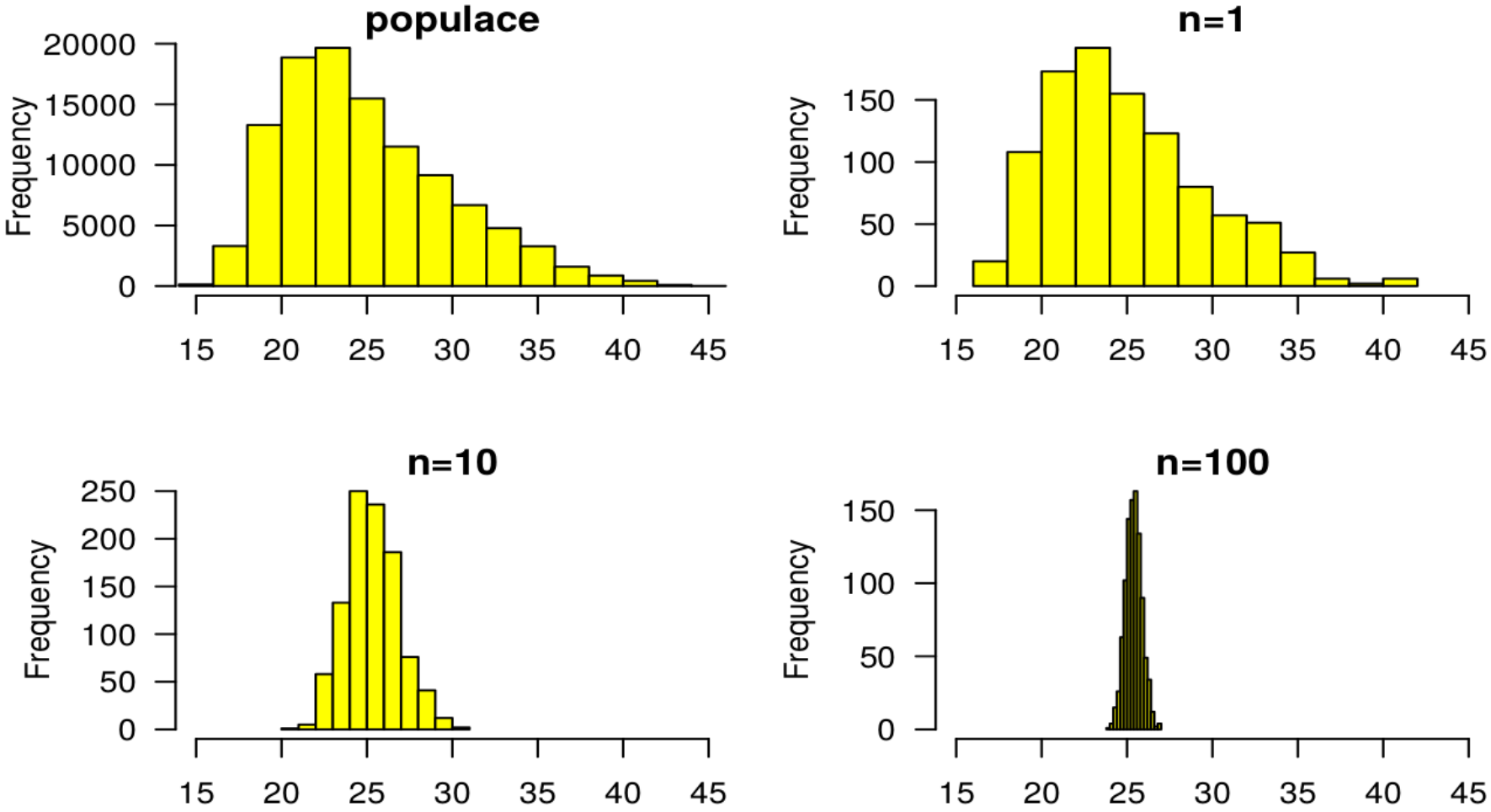






- výběrový rozptyl mých odhadů je 2.38 (± 1.54 roku)
- to je o hodně přesnější!
- pozorování: cca 10x
- můžeme to vylepšit?
- zeptáme se 100 žen!

MODELOVÝ PŘÍKLAD – VĚKY RODIČEK



- výběrový rozptyl mých odhadů je 0.23 (± 0.48 roku)
- 100x menší než při jednom pokusu – náhoda?

MODELOVÝ PŘÍKLAD – VĚKY RODIČEK

- čím více realizací náhodné veličiny (i.e. oslovených žen) jsme použili na výpočet průměru, tím pravděpodobněji jsme se strefili poblíž skutečného μ

rozsah výběru n	průměr průměrů	směr. odch. průměrů	rozptyl průměrů	rozptyl průměrů teoreticky
1	25,42	4,625	21,388	24,428
10	25,35	1,544	2,385	2,443
100	25,39	0,480	0,231	0,244
1000	25,40	0,150	0,022	0,024
populace	μ =25,40	σ =4,932	σ^2 =24,428	

PRŮMĚR Z NÁHODNÉHO VÝBĚRU

- $X_1, \ldots, X_n$ **nezávislé**, mají 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
- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$ — 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
- z příkladu víme, že rozptyl $\bar{X}$ závisí na n

PRŮMĚR Z NÁHODNÉHO VÝBĚRU

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

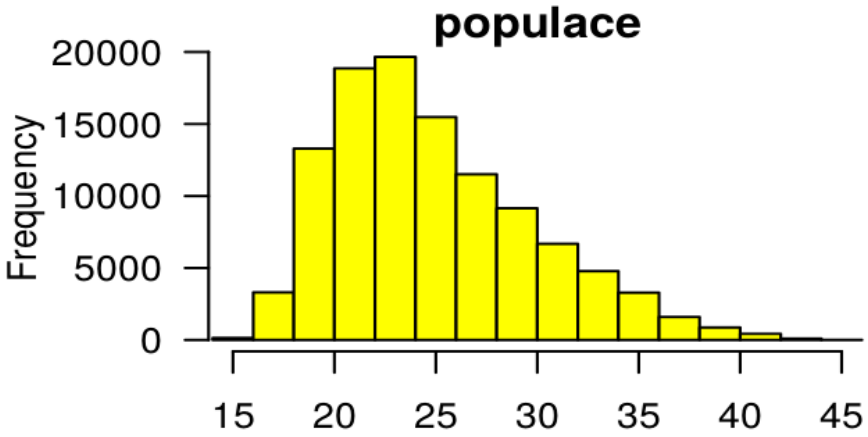
- $\text{S.E.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ – **střední chyba průměru**

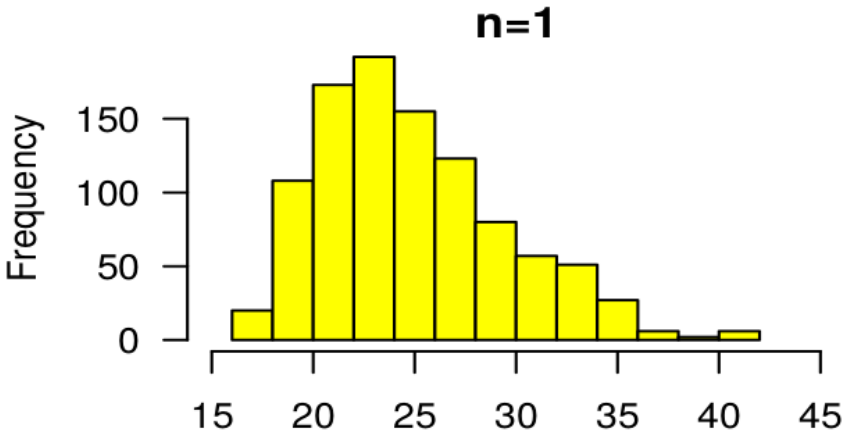
  [standard error of mean]
- variabilita průměrů (měřená rozptylem) z výběrů rozsahu n je n-krát menší, než variabilita jednotlivých pozorování σ^2
- střední chyba průměru je $\sqrt{n}$-krát menší než σ
- čím jsou rozsahy výběru větší, tím méně výběrové průměry kolísají (kolem populačního průměru)

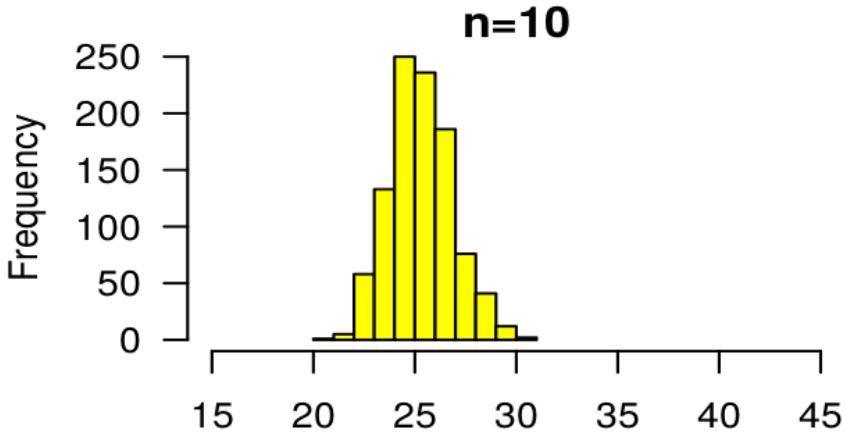
PRŮMĚR Z NÁHODNÉHO VÝBĚRU

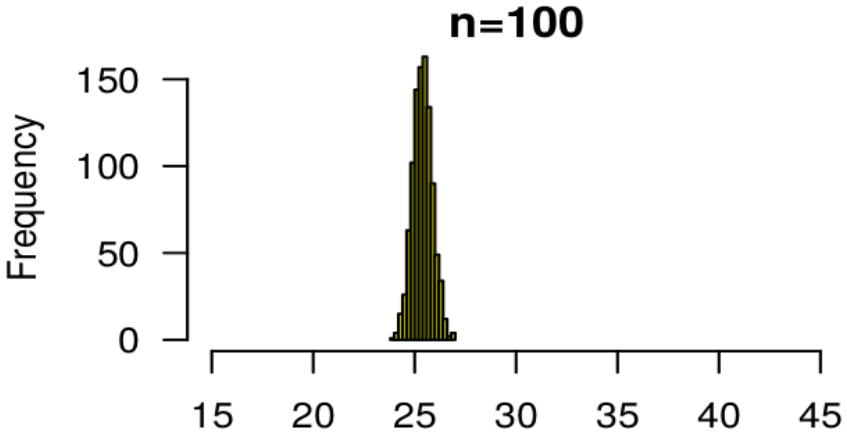
- náhodná veličina, obecné rozdělení, $EX = \mu$, $\text{var } X = \sigma^2$
- víme o průměru:
  - má střední hodnotu μ
  - má rozptyl σ^2/n
  - jaké má rozdělení?

MODELOVÝ PŘÍKLAD – VĚKY RODIČEK

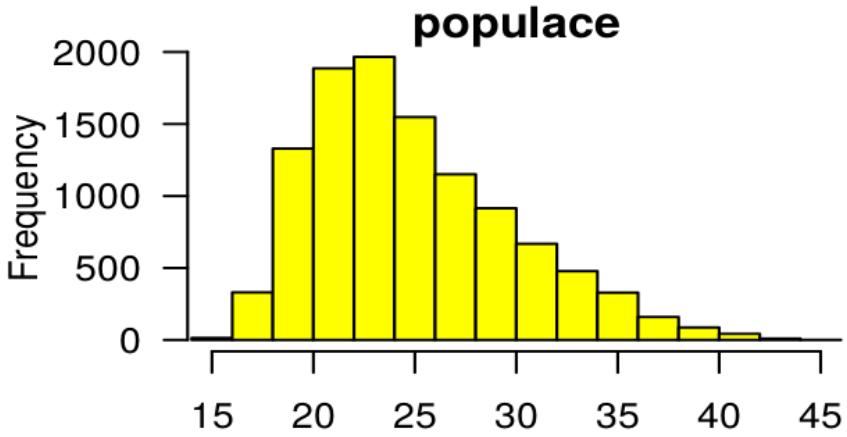


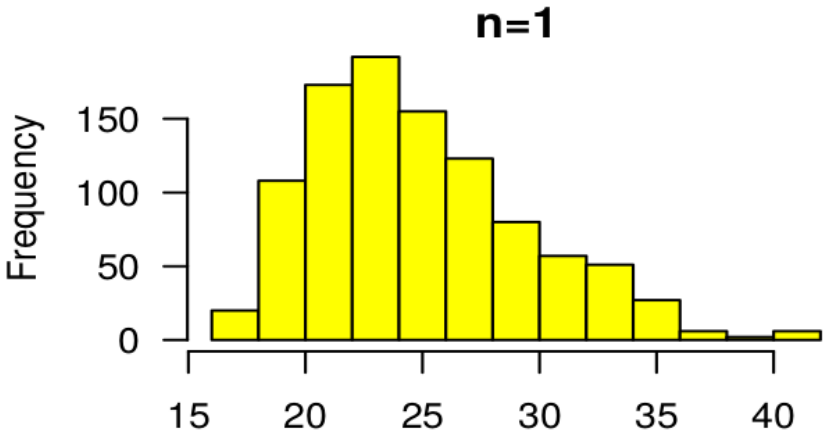


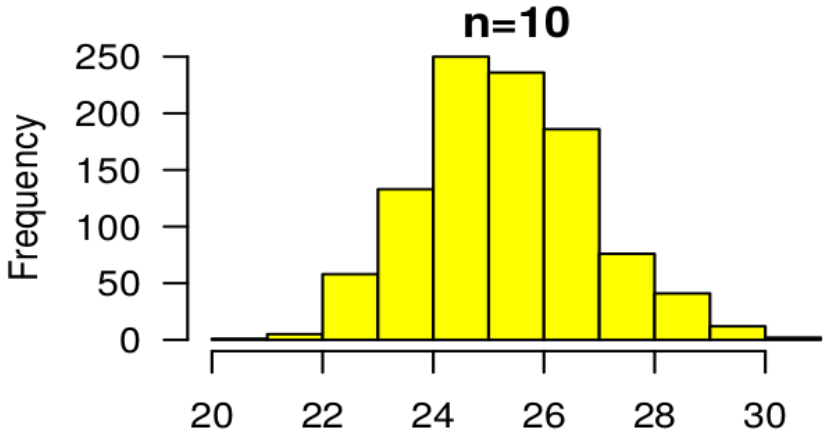


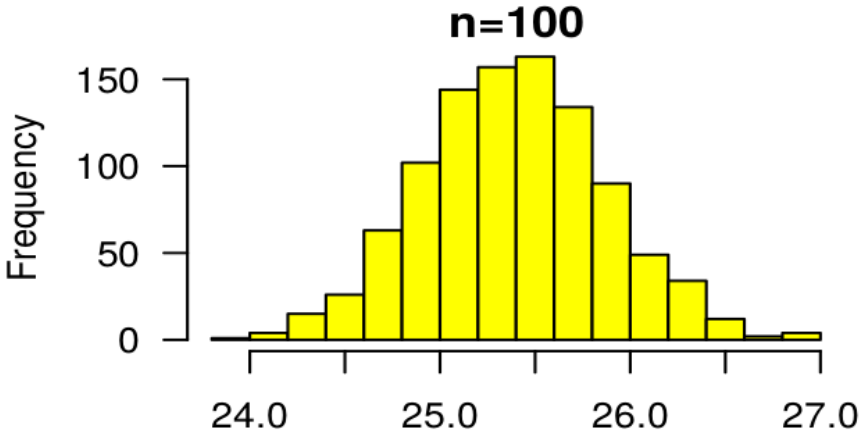


MODELOVÝ PŘÍKLAD – VĚKY RODIČEK









- nějak se to symetričtí :)

CENTRÁLNÍ LIMITNÍ VĚTA

Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou μ a rozptylem $\sigma^2 > 0$ (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: pro dost velká n má průměr normální rozdělení s rozptylem n-krát menším než jednotlivá pozorování, a to bez ohledu na výchozí rozdělení pozorování
- důvod častého předpokladu normality odhadu střední hodnoty

KONFIDENČNÍ INTERVAL PRO μ

- víme, že $\bar{X} \sim \mathrm{N}\left(\mu, \sigma^2/n\right)$, 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 ...)

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

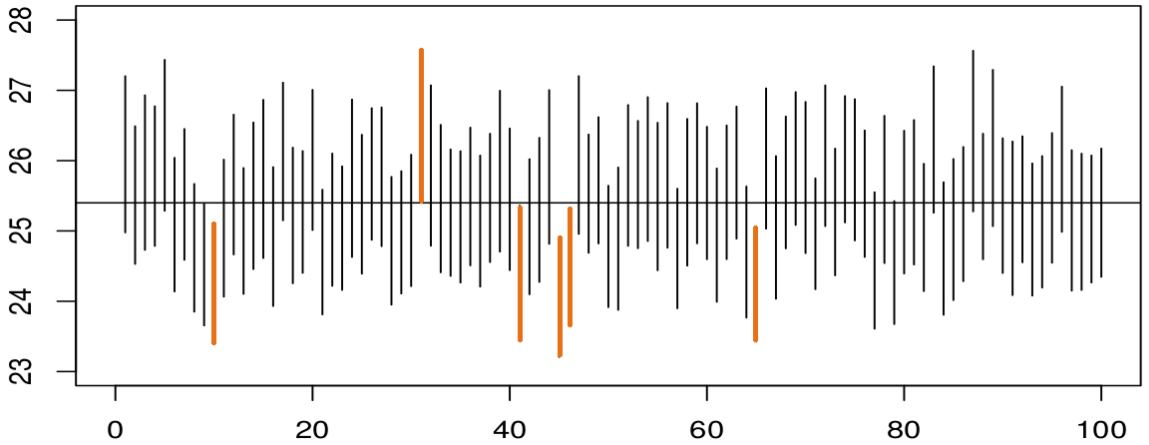
- tedy (všimněte si zkracování intervalu s rostoucím n)

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

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

KONFIDENČNÍ INTERVAL PRO μ

- je to vlastně intervalový odhad neznámého m
- $\bar{X}$ je oproti tomu bodový odhad téhož
- základní vlastnost: 95% interval spolehlivosti překryje s pravděpodobností 95% neznámé μ
- kdybychom prováděli opakovaně, tak v cca 95% případů překryje a v cca 5% bude mimo



KONFIDENČNÍ INTERVAL PRO μ

- pokud je σ neznámé, použijeme jeho odhad = směrodatnou odchylku
- místo kvantilu normálního rozdělení kvantil t-rozdělení
- ale i tady platí CLV a pro dost velké *n* (>50) lze aptoximovat normálním

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

- jako odhad σ se použije **výběrová** směrodatná odchylka

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

TESTOVÁNÍ HYPOTÉZ

PŘIPOMENUTÍ TERMINOLOGIE

[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

MODELOVÝ PŘÍKLAD – DESETILETÍ CHLAPCI

- v roce 1951 bylo provedeno rozsáhlé měření výšky desetiletých hochů, výška byla vyšetřena v populaci desetiletých chlapců: zjištěno $\mu = 136{,}1$ cm, $\sigma = 6{,}4$ cm
- na základě výběru pořízeného v roce 1961 máme rozhodnout, zda se po deseti letech výška populace desetiletých **zvýšila**
- hodnoty zjištěné v roce 1961 [cm]: 130, 140, 136, 141, 139, 133, 149, 151, 139, 136, 138, 142, 127, 139, 147
- $\bar{x} = 139{,}13$ cm, $s^2 = 6{,}56^2$ cm^2
- jiný (další) výběr z roku 1961 by obsahoval jiných 15 hochů, tedy by vedl k jinému výběrovému průměru (náhodná veličina)
- stačí rozdíl $139{,}13 - 136{,}1 = 3{,}03$ (realizace náhodné veličiny, proč?), abychom prokázali, že se **populační průměr** výšek desetiletých chlapců po deseti letech změnil?

TESTOVÁNÍ HYPOTÉZ - TERMINOLOGIE

[hypothesis testing, null hypothesis, alternative hypothesis, critical (rejection) region, Type I (II) error, significance level]

- **nulová hypotéza** H_0: tvrzení o populaci (parametru), o jehož platnosti rozhodujeme (**ne**ní rozdíl, **ne**závisí, **ne**liší se od . . .)
- **alternativní hypotéza** H_1: (alternativa) zbývající možnost (k H_0), často „vědecká hypotéza“, kterou chceme dokázat

- **kritický obor**: možné výsledky pokusu, kdy H_0 zamítáme; zpravidla popsán pomocí statistiky (např. $|Z| \geq z(1-\alpha/2)$)
- **obor přijetí**: možné výsledky pokusu, kdy H_0 nezamítáme
- **chyba prvního druhu**: (náhodný jev) rozhodnutí zamítnout H_0, když platí H_0, tj. falešně prokázat „vědeckou hypotézu“
- **chyba druhého druhu**: (náhodný jev) rozhodnutí nezamítnout H_0, když platí H_1, tj. nepoznat neplatnost H_0

TESTOVÁNÍ HYPOTÉZ - ROZHODOVÁNÍ

[significance level, power, p-value]

- **hladina testu** α (zpravidla $\alpha = 5\ \%$)
  - maximální dovolená pravděpodobnost chyby prvního druhu
  - volí se před pokusem, nezávisle na jeho výsledku
  - **pevná** (nenáhodná) hodnota
- **síla testu** $1 - \beta$
  - pravděpodobnost zamítnutí neplatné H_0
  - pst, s jakou prokážeme platnou „vědeckou hypotézu“
  - závisí na skutečné hodnotě parametru
- p-**hodnota**
  - za platnosti H_0 určená pst, že dostaneme statistiku, která stejně nebo ještě méně podporuje H_0
  - nejmenší hladina α, na které lze ještě H_0 zamítnout
  - „stupeň důvěry“ v platnost nulové hypotézy
  - je to **náhodná veličina**, nikoliv pravděpodobnost H_0
- H_0 se **zamítá**, právě když $\boxed{p \leq \alpha}$ (**zapamatovat**)

MODELOVÝ PŘÍKLAD – DESETILETÍ CHLAPCI

- zvolíme klasickou hladinu $\alpha = 5\ \%$
- v roce 1951 $\mu = \mu_0 = 136{,}1$ cm, $\sigma = 6{,}4$ cm
- v roce 1961 změřeno $n = 15$ náhodně vybraných desetiletých hochů, $\bar{x} = 139{,}13$ cm
- stačí tento vzrůst k důkazu, že nová generace je vyšší?
- vzrostla výška desetiletých ? $H_0 : \mu = \mu_0$ proti $H_1 : \mu > \mu_0$

KONSTRUKCE TESTOVÉ STATISTIKY A TESTU

(σ známé)

- $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ **nezávislé**; $\sigma > 0$ známe
- $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$, tedy $\mathrm{S.E.}(\bar{X}) = \sigma/\sqrt{n}$
- $\mathrm{H}_0 : \mu = \mu_0$ (dané číslo, jiný zápis $\mathrm{H}_0 : \mu - \mu_0 = 0$)
- platí-li H_0, pak $\boxed{Z = \dfrac{\bar{X} - \mu_0}{\mathrm{S.E.}(\bar{X})} = \dfrac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1)}$

KONSTRUKCE TESTOVÉ STATISTIKY A TESTU

(σ známé)

nebo jiné rozdělení a hodně velké n

- $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ **nezávislé**; $\sigma > 0$ známe
- $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$, tedy $\mathrm{S.E.}(\bar{X}) = \sigma/\sqrt{n}$
- $\mathrm{H}_0 : \mu = \mu_0$ (dané číslo, jiný zápis $\mathrm{H}_0 : \mu - \mu_0 = 0$)
- platí-li H_0, pak $\boxed{Z = \dfrac{\bar{X} - \mu_0}{\mathrm{S.E.}(\bar{X})} = \dfrac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1)}$ testová statistika

KONSTRUKCE TESTOVÉ STATISTIKY A TESTU

(σ známé)

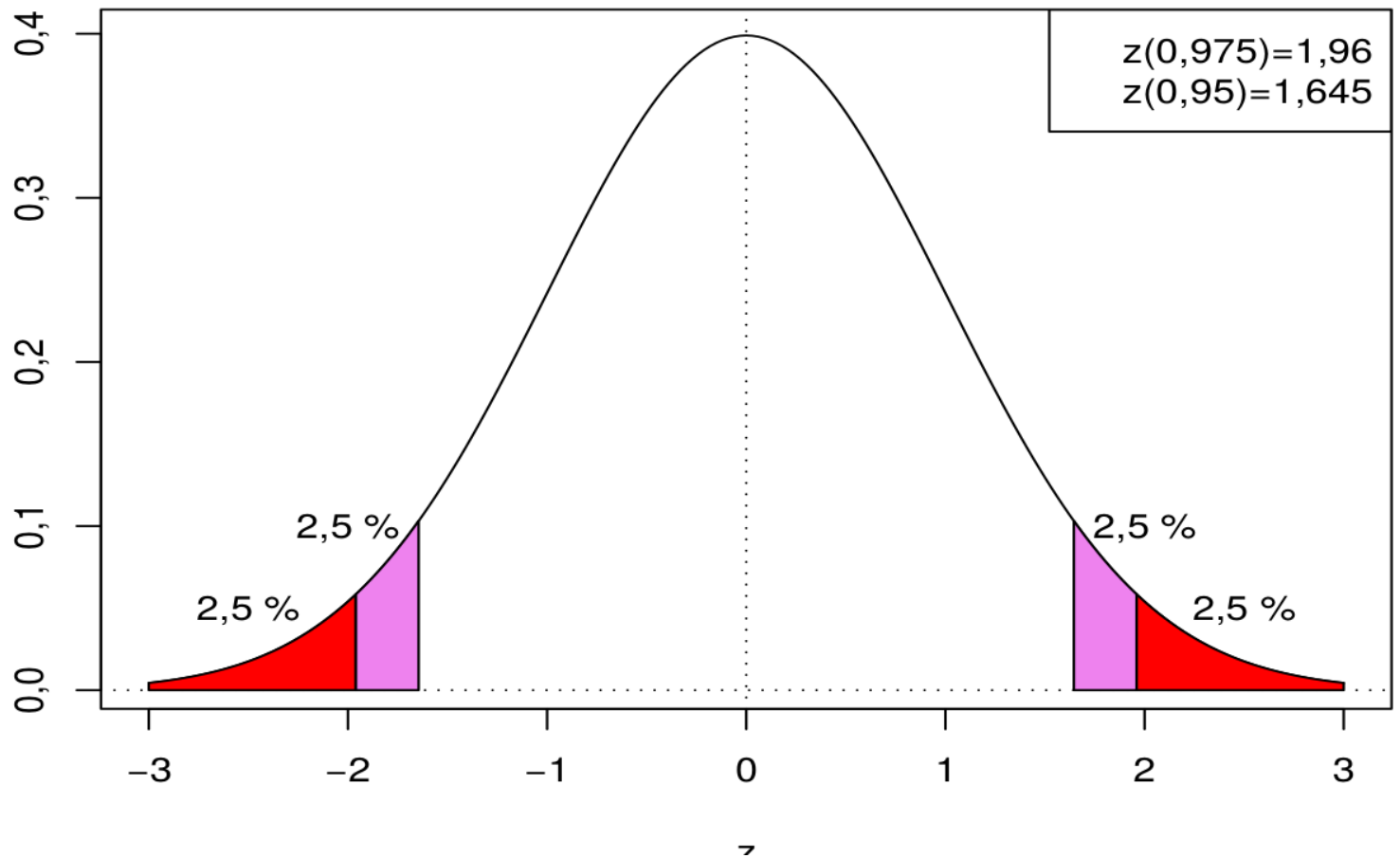
nebo jiné rozdělení a hodně velké *n*

- $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ **nezávislé**; $\sigma > 0$ známe
- $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$, tedy $\mathrm{S.E.}(\bar{X}) = \sigma/\sqrt{n}$
- $\mathrm{H}_0 : \mu = \mu_0$ (dané číslo, jiný zápis $\mathrm{H}_0 : \mu - \mu_0 = 0$)
- platí-li H_0, pak $\boxed{Z = \dfrac{\bar{X} - \mu_0}{\mathrm{S.E.}(\bar{X})} = \dfrac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1)}$ testová statistika
- $\mathrm{H}_1 : \mu \neq \mu_0 \Rightarrow$ kritický obor: $|Z|$ velké, tj. $\boxed{|Z| \geq z(1-\alpha/2)}$
- $\mathrm{H}_1 : \mu > \mu_0$: zamítnout pro $\boxed{Z \geq z(1-\alpha)}$ test
- $\mathrm{H}_1 : \mu < \mu_0$: zamítnout pro $\boxed{Z \leq z(\alpha) = -z(1-\alpha)}$
- volba jednostranné alternativy jen podle zadání úlohy, nikoliv podle výsledku pokusu (nezávisle na datech)

KONSTRUKCE TESTU

kritický obor pro $Z = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}$

červeně na 5% hladině, červeně a fialově na 10% hladině, hustota Z za H_0



MODELOVÝ PŘÍKLAD – DESETILETÍ CHLAPCI

pozor, **jednostranná** alternativa!

- zvolíme klasickou hladinu $\alpha = 5\ \%$
- v roce 1951 $\mu = \mu_0 = 136{,}1$ cm, $\sigma = 6{,}4$ cm
- v roce 1961 změřeno $n = 15$ náhodně vybraných desetiletých hochů, $\bar{x} = 139{,}13$ cm
- stačí tento vzrůst k důkazu, že nová generace je vyšší?
- vzrostla výška desetiletých ? $H_0 : \mu = \mu_0$ proti $H_1 : \mu > \mu_0$

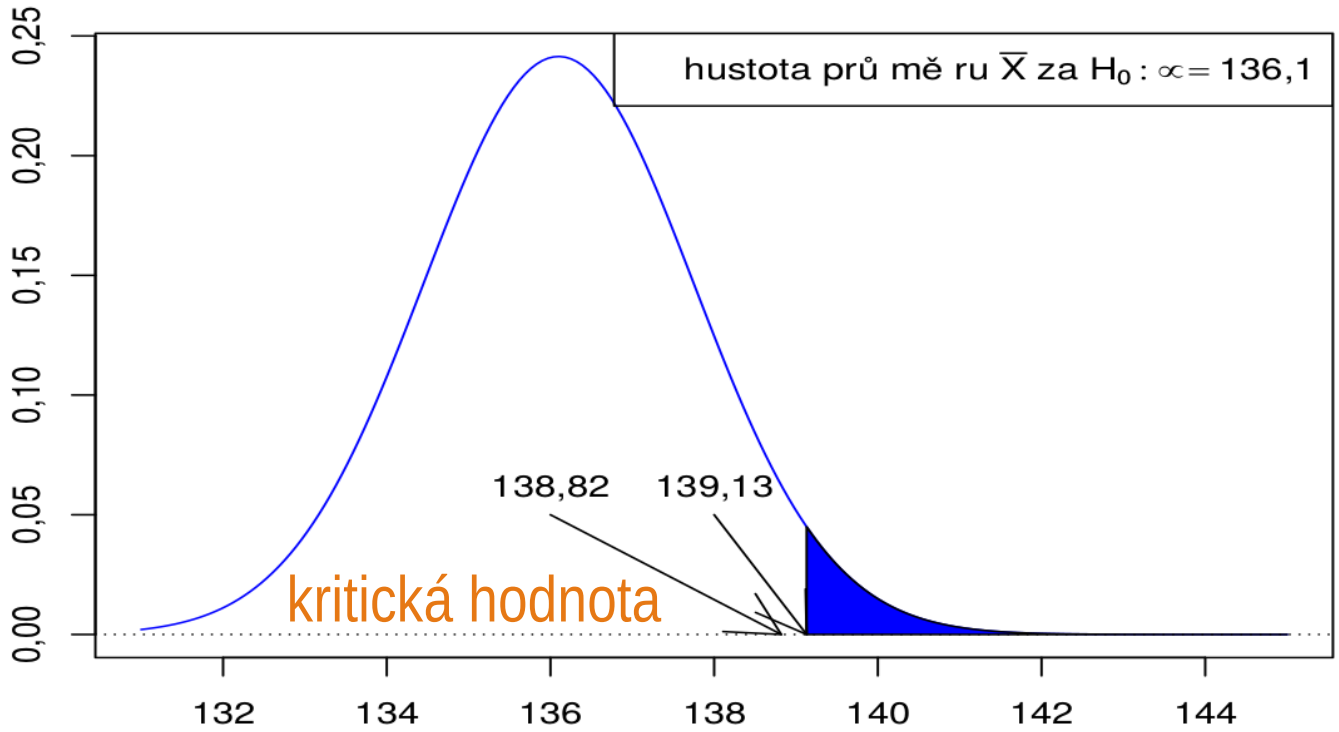
$$z = \frac{139{,}13 - 136{,}1}{6{,}4}\sqrt{15} = 1{,}836$$

- $z(0{,}05) = 1{,}645 < 1{,}836$, tedy H_0 na 5% hladině **zamítáme**
- statisticky **významný** výsledek
- na 5% hladině jsme prokázali, že nová generace je vyšší
- v případě, že nová generace není vyšší, riskovali jsme jen 5% pravděpodobnost, že budeme nesprávně tvrdit, že vyšší je

MODELOVÝ PŘÍKLAD – DESETILETÍ CHLAPCI

hustota $\bar{X}$ za platnosti hypotézy $H_0 : \mu = 136{,}1$, $H_1 : \mu > \mu_0$ při $\sigma = 6{,}4$

- p-hodnota je pst, že za H_0: $Z = (\bar{X} - \mu_0)\sqrt{n}/\sigma > 1{,}836$ tj. $\bar{X} > 136{,}1 + 1{,}836 \cdot 6{,}4/\sqrt{15} = 139{,}13$ [1-pnorm(1.836)]
- p-hodnota: modrá plocha napravo od 139,13, $p = 3{,}3$ %



$$136{,}1 + \frac{6{,}4}{\sqrt{15}} \cdot 1{,}645 = 138{,}82$$

JEDNOVÝBĚROVÝ T-TEST

výběr z $\mathrm{N}(\mu, \sigma^2)$, σ **neznámé**

nebo jiné rozdělení a hodně velké *n*

- n nezávislých pozorování $X_1, \ldots, X_n$ z rozdělení $\mathrm{N}(\mu, \sigma^2)$
- $\mathrm{H}_0 : \mu = \mu_0$ (populační průměr roven dané konstantě)
- nutno odhadnout neznámý rozptyl σ^2

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

- statistika (místo σ použijeme S_x)

$$\boxed{T = \frac{\bar{X} - \mu_0}{\mathrm{S.E.}(\bar{X})} = \frac{\bar{X} - \mu_0}{S_x}\sqrt{n}}$$

testová statistika

- $\mathrm{H}_1 : \mu \neq \mu_0$ zamítat při $|T| \geq t_{n-1}(1 - \alpha/2)$
- $\mathrm{H}_1 : \mu > \mu_0$ zamítat při $T \geq t_{n-1}(1 - \alpha)$
- $\mathrm{H}_1 : \mu < \mu_0$ zamítat při $T \leq t_{n-1}(\alpha) = -t_{n-1}(1 - \alpha)$

Studentovo t-rozdělení o n-1 stupních volnosti

MODELOVÝ PŘÍKLAD – DESETILETÍ CHLAPCI

- $H_0 : \mu = 136{,}1$ proti $H_1 : \mu > 136{,}1$ $(\alpha = 5\ \%)$

$$\bar{x} = 139{,}133 \qquad s_x^2 = 6{,}556^2$$

$$t = \frac{139{,}133 - 136{,}1}{6{,}556}\sqrt{15} = 1{,}792 > 1{,}761 = t_{14}(0{,}95)$$

$$p = \mathrm{P}(T \geq 1{,}792) = 0{,}047 \qquad (\text{ tj. } 4{,}7\ \%)$$

p-hodnota

- na 5% hladině jsme prokázali zvýšení populačního průměru (H_0 se na 5% hladině **zamítá**)
- [t.test(hosi,mu=136.1,alternative="greater")]

T-TEST: INTERVAL SPOLEHLIVOSTI

při oboustranné alternativě

- oboustranný interval spolehlivosti pro μ (viz str. 112)

$$\left(\bar{X} - \frac{S_x}{\sqrt{n}}\, t_{n-1}(1-\alpha/2), \bar{X} + \frac{S_x}{\sqrt{n}}\, t_{n-1}(1-\alpha/2)\right)$$

- μ_0 patří do intervalu spolehlivosti, právě když platí

$$|\bar{X} - \mu_0| < \frac{S_x}{\sqrt{n}} t_{n-1}(1-\alpha/2)$$

- tedy, právě když se nezamítne hypotéza $H_0 : \mu = \mu_0$ při oboustranné alternativě $H_1 : \mu \neq \mu_0$
- interval spolehlivosti obsahuje takové hodnoty μ_0, pro které bychom **nezamítli** hypotézu $H_0 : \mu = \mu_0$
- podobně u jednostranných intervalů spolehlivosti a jednostranných alternativ

TESTOVÁNÍ HYPOTÉZ - ROZHODOVÁNÍ

rozhodnutí	skutečnost	
	H_0 platí	H_0 neplatí
H_0 zamítnout (reject)	**chyba 1. druhu** ($\text{pst} \leq \alpha$)	správné rozhodnutí ($\text{pst} = 1 - \beta$)
H_0 nezamítnout (accept, přijmout)	správné rozhodnutí ($\text{pst} \geq 1 - \alpha$)	chyba 2. druhu ($\text{pst} = \beta$)

- zamítnutí $\Leftrightarrow$ výsledek pokusu v kritickém oboru
- přijetí $\Leftrightarrow$ výsledek pokusu v oboru přijetí
- nikdy spolehlivě nevíme, zda H_0 platí
- chybu 1. druhu nechceme dělat často $\Rightarrow$ α volíme malé

DĚKUJI ZA POZORNOST

www.biostatisticka.cz