Statistické metody

Základní metody:

- odhad parametrů
- testování hypotéz

Pokročilejší metody:

- shluková (klastrová) analýza
- faktorová analýza
- analýza hlavních komponent (PCA)
- …

Statistické metody

Základní soubor (populace) je příliš velký a nemůžeme ho celý „proměřit“.

Proto dělám reprezentativní výběr, ten změřím, tedy náhodným procesem získávám konkrétní hodnoty náhodných veličin.

Spočítám výběrové charakteristiky souboru.

Tyto výběrové charakteristiky chci vztáhnout na celý základní soubor. Musím nějak kvantifikovat jistotu či nejistotu, že moje odhady se potkávají s neznámou skutečností.

Připomínka značení:

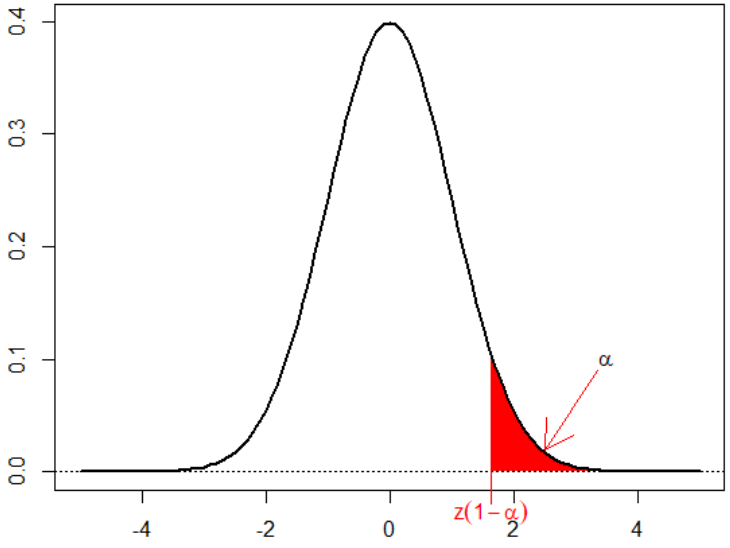
μ vs. $\bar{X}$ … skutečný neznámý parametr vs. náš odhad

σ^2 vs. S^2

$z(1-\alpha)$ … (1 - α)% kvantil rozdělení prstí, pro který platí

$t_{df}(1-\alpha)$

$$P\left(X > z(1-\alpha)\right) = \alpha$$



Bodový odhad parametru [point estimate of the parameter]

Základní předpoklad dalšího odvozování:
mám výběr ***n*** hodnot $(X_1, X_2, X_3, \dots, X_n)$, které jsou ***iid.***, tedy vzájemně nezávislé a všechny pocházejí ze stejného rozdělení prstí.

K odhadu typické hodnoty (charakteristika polohy) nejčastěji používáme

výběrový průměr $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ [sample mean]

Protože výběrový průměr je náhodná veličina, má smysl se ptát:

a) jaká je jeho střední hodnota [expected value of the estimate]
b) jaký je jeho rozptyl [variance of the estimate]
c) jaká je jeho směrodatná odchylka [standard error of the estimate]

Tady shrnutí, odvození dále:

$$E\bar{X} = \mu \qquad var\,\bar{X} = \frac{\sigma^2}{n} \qquad sd\,\bar{X} = \frac{\sigma}{\sqrt{n}}$$

Odvození pro výběrový průměr:

(a) Střední hodnota výběrového průměru:

$$E\bar{X} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}\cdot n \cdot \mu = \mu$$

vlastnost střední hodnoty: $\mathbf{E}(X + Y) = EX + EY$

- tento odhad je nestranný, protože $E\bar{X} = \mu$

Odvození pro výběrový průměr:

(b) Rozptyl výběrového průměru:

$$var\bar{X} = var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\, var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n} varX_i\right) = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{\sigma^2}{n}$$

$var(\beta \cdot X) = \beta^2 \cdot varX$

(1) všechna X_i jsou ***iid.***, proto $cov(X_i, X_j) = 0$ pro $\forall i, j$
(2) $var(X + Y) = varX + varY + 2cov(X, Y)$

$$var\bar{X} = \frac{\sigma^2}{n}$$

- $n = 1 \rightarrow var\,\overline{X_1} = \sigma^2$
- větší $n \rightarrow$ menší rozptyl $\bar{X}$
- problém: σ^2 většinou neznáme

Odvození pro výběrový průměr:

(c) Směrodatná odchylka výběrového průměru:

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

- říkáme jí **střední chyba průměru** [standard error of mean, SEM]
- často se uvádí ve výsledcích článků
- charakterizuje „přesnost“ odhadu (pozor: přesnost odhadu ve smyslu střední kvadratické chyby (viz dále) zahrnuje i vychýlení odhadu)
- platí: čím větší výběr (n), tím přesnější odhad
- *SEM* závisí na parametru σ, který většinou neznáme a nahrazujeme ho vhodným odhadem, např. výběrovým rozptylem (za chvilku). Slovní označení „střední chyba“ se používá i tehdy, když místo σ použiji odhad.

Bodový odhad variance – výběrový rozptyl

K odhadu variability hodnot v populaci nejčastěji používáme

výběrový rozptyl $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ [sample variance]

- střední hodnota výběrového rozptylu:
  $ES^2 = \sigma^2$
- rozptyl výběrového rozptylu běžně nepotřebujeme, proto neuvádím

- výběrový momentový rozptyl $S_n{}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ většinou nepoužíváme, protože o $\frac{1}{n}$ podhodnocuje skutečný parametr σ^2 (dále)

Vsuvka – jiný tvar výběrového rozptylu:

užitečný tvar pro „ruční“ výpočet, používá se v algoritmech (je rychlejší):

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

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

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}{X_i}^2 - 2\bar{X}\cdot n\frac{\sum X_i}{n} + n\bar{X}^2\right) =$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}{X_i}^2 - 2\bar{X}\cdot n\bar{X} + n\bar{X}^2\right) = \boldsymbol{\frac{1}{n-1}\left(\sum_{i=1}^{n}{X_i}^2 - n\bar{X}^2\right)}$$

Odvození výpočtu střední hodnoty výběrového rozptylu

$$\boldsymbol{ES^2} = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = E\left(\frac{1}{n-1}\left(\sum_{i=1}^{n} {X_i}^2 - n\bar{X}^2\right)\right) =$$

$$E(\beta \cdot X) = \beta \cdot EX$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} E\,{X_i}^2 - n \cdot E\,\bar{X}^2\right) =$$

$$= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n \cdot \left(\frac{\sigma^2}{n} + \mu^2\right)\right) = \frac{1}{n-1}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2) =$$

$$= \frac{\mathbf{1}}{\boldsymbol{n-1}} \cdot \boldsymbol{(n-1)\sigma^2} = \boldsymbol{\sigma^2}$$

→ $var\,X_i = E(X_i - EX_i)^2 = E\left({X_i}^2 - 2X_i\,EX_i + (EX_i)^2\right) = E\,{X_i}^2 - 2 \cdot EX_i \cdot EX_i + (EX_i)^2 = E\,{X_i}^2 - (EX_i)^2$

odtud: $\boldsymbol{E\,{X_i}^2 = varX_i + (EX_i)^2 = \sigma^2 + \mu^2}$

→ podobně: $var\bar{X} = E(\bar{X} - E\bar{X})^2 = \cdots = E(\bar{X})^2 - (E\bar{X})^2$

odtud: $\boldsymbol{E\bar{X}^2 = var\bar{X} + (E\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2}$

Bodový odhad populační SD – výběrová směrodatná odchylka

$$S = \sqrt{S^2}$$

- tento odhad je vychýlený, skutečnou směr. odchylku v průměru podhodnocuje, protože platí $ES < \sigma$.

Vlastnosti bodového odhadu

Nestranný odhad (nevychýlený, nezkreslený) [unbiased estimation]

- když střední hodnota odhadu = teoretickému parametru
- právě jsme měli: $\boldsymbol{E\overline{X} = \mu}$ a $\boldsymbol{ES^2 = \sigma^2}$
- nestranný odhad systematicky nenadhodnocuje ani nepodhodnocuje odhadovaný parametr

- příklad vychýleného odhadu – výběrový momentový rozptyl:

$$\boldsymbol{ES_n{}^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \cdots = \frac{n-1}{n}\sigma^2}$$

vychýlení značíme $B\left(\sigma^2, S_n{}^2\right) = \boldsymbol{ES_n{}^2 - \sigma^2} = \frac{n-1}{n}\sigma^2 - \frac{n}{n}\sigma^2 = -\frac{1}{n}\boldsymbol{\sigma^2}$

$\boldsymbol{S_n{}^2}$ podhodnocuje skutečný parametr $\boldsymbol{\sigma^2}$.

Vlastnosti bodového odhadu

Asymptoticky nestranný odhad

- když odhad je sice vychýlený, ale se zvyšujícím se rozsahem výběru n se vychýlení zmenšuje až k nule
- to je případ výběrového momentového rozptylu:

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

$$ES_n^2 = \frac{n-1}{n}\sigma^2$$

vychýlení $ES_n^2 - \sigma^2 = -\frac{1}{n}\sigma^2$

$$\lim_{n\to\infty} -\frac{1}{n}\sigma^2 \to -\frac{1}{\infty}\sigma^2 = 0$$

Vlastnosti bodového odhadu

Konzistentní odhad [consistent estimation]

- pokud se s rostoucím rozsahem výběru ***n*** odhad zpřesňuje
- $E(odhadu) = parametr$
- a zároveň $\lim_{n\to\infty}\left(var(odhadu)\right) = 0$

- platí např. pro výběrový průměr:

$$\boldsymbol{E\bar{X} = \mu}$$

$$\boldsymbol{var\bar{X} = \frac{\sigma^2}{n} \quad \xrightarrow[n\to\infty]{} \frac{\sigma^2}{\infty} = 0}$$

Vlastnosti bodového odhadu

Vydatný, eficientní, nejlepší nestranný odhad [efficient estimation]

- má nejmenší rozptyl mezi všemi nestrannými odhady téhož parametru

Přesnost, kvalita odhadu [quality of the estimation]

- měříme pomocí střední kvadratické chyby odhadu
- výběrová chyba odhadu: $odhad - parametr$
- zkratka MSE(odhadu) [mean squared error]
- Kromě variability zahrnuje i vychýlení odhadu. Pro nestranné odhady (vychýlení = 0) je to totéž jako var(odhadu) a potažmo S.E.(odhadu)
- $MSE(odhadu) = E(odhad - parametr)^2 = var(odhadu) + B^2(odhadu) = \\ = E(odhad - E(odhadu))^2 + (E(odhadu) - parametr)^2$
- příklad: $MSE\left(S_n{}^2\right) = E\left(S_n{}^2 - \sigma^2\right)^2 = \cdots$

Ze statistického slovníku:

Robustní = odolný

přibližně řečeno je to schopnost spočítat „spolehlivý“ výsledek, přestože jsou narušeny předpoklady testu, odhadu apod.

Konečnostní násobitel

Většinou zahrnuje náš výběr méně než 5 % jedinců z celé populace, proto můžeme takovou populaci považovat za nekonečnou.

Pokud ovšem vybíráme z menší konečné populace a rozsah výběru je větší než 5 % všech jedinců, potom výběrový průměr $\bar{X}$ zůstává nestranným odhadem populačního průměru, ale rozptyl $\bar{X}$ se poněkud zmenší. Aby byly odhadované vlastnosti $\bar{X}$ správné, je třeba rozptyl vynásobit konečnostním násobitelem $\frac{N-n}{N-1}$.

Tedy:

$$E\bar{X} = \mu \qquad \text{... to je stejné}$$

$$var\,\bar{X} = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$$

(Citace: Zvára, Karel: Biostatistika. Karolinum, Praha 2008.)

Intervalový odhad parametru [confidence interval of the parameter]

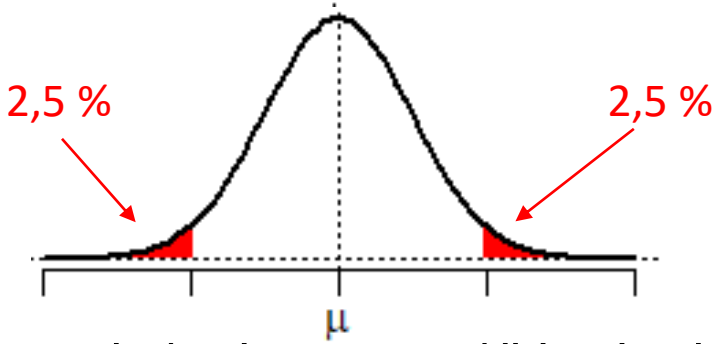
také konfidenční interval či interval spolehlivosti.

Konstrukci intervalu provedeme na příkladu výběrového průměru, teorie však platí pro odhady všech parametrů.

- Výběrový průměr $\overline{X}$ je náhodná veličina, má tedy i své rozdělení pravděpodobností.
- Víme, že $E\overline{X} = \mu$ a $var\ \overline{X} = \frac{\sigma^2}{n}$ (skutečné, ale neznámé parametry).
- Pokud výběr pochází z normálního rozdělení $N(\mu, \sigma^2)$, potom také náh. veličina $\overline{X}$ má normální rozdělení s parametry $N\left(\mu, \frac{\sigma^2}{n}\right)$.
- Když výběr nepochází z normálního rozdělení (histogram je šikmý nebo hrbatý), potom záleží na velikosti výběru n funguje centrální limitní věta (dále) a podle té má $\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$ i když původní data nejsou z normálního rozdělení.

Intervalový odhad parametru

Teoreticky: hodnoty, kterých může nabývat průměr $\bar{X}$ jsou popsány normálním rozdělením $N\left(\mu, \frac{\sigma^2}{n}\right)$:



Chceme sestrojit interval takový, aby pokrýval „rozumné" hodnoty $\bar{X}$ a abychom znali pravděpodobnost chybného tvrzení o tomto intervalu.

Zvolíme velikost možné chyby $\alpha = 0,05$, tj. 5 % (například).

Pomůžeme si normovaným tvarem $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ se známými kvantily:

$$P\left(-z(1-{}^{\alpha}/_{2}) < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z(1-{}^{\alpha}/_{2})\right) = 0,95$$

N(0, 1) je souměrné, proto $z(1-{}^{\alpha}/_{2}) = -z({}^{\alpha}/_{2})$.

$$\approx P\left(\bar{X} - z(1-{}^{\alpha}/_{2}) \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z(1-{}^{\alpha}/_{2}) \cdot \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

Intervalový odhad parametru

$$\rightarrow P\left(\bar{X} - z(1-{}^{\alpha}/_{2}) \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z(1-{}^{\alpha}/_{2}) \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha = 0,95$$

Tedy jsem zpět v $\sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Jiný tvar: $\mathbf{P}\left(\mu \in \left\{\bar{X} - z(1-{}^{\alpha}/_{2}) \cdot \frac{\sigma}{\sqrt{n}} \; ; \; \bar{X} + z(1-{}^{\alpha}/_{2}) \cdot \frac{\sigma}{\sqrt{n}}\right\}\right) = 1 - \alpha$

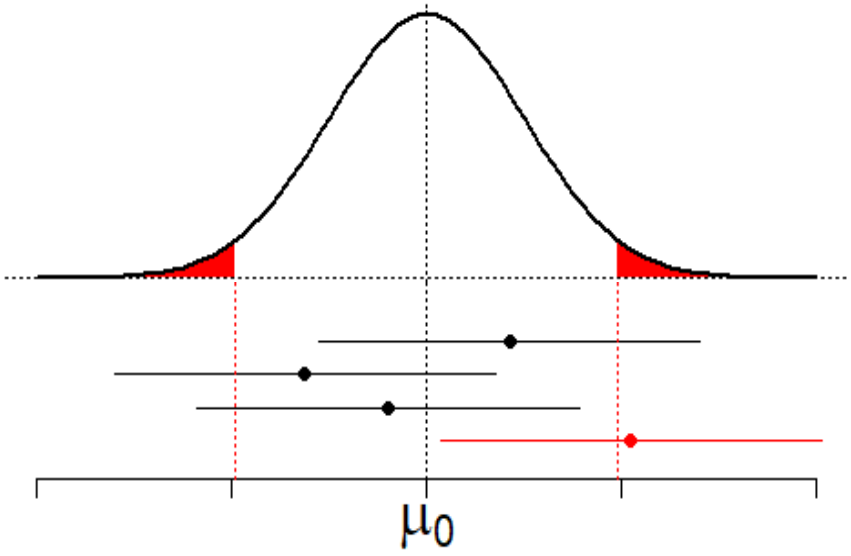
konfidenční interval odhadu parametru μ na hladině **α** = 0.05.

Další způsob zápisu: $\bar{X} \pm z(1-{}^{\alpha}/_{2}) \cdot \frac{\sigma}{\sqrt{n}}$

Výsledek 15.3 ± 3.65 čteme např. takto: *střední hodnotu odhadujeme hodnotou 15.3, přičemž skutečná hodnota střední hodnoty leží na 95 % v rozmezí 15.3 – 3.65 a 15.3 + 3.65.* Je třeba uvádět také pravděp. nebo α.

Intervalový odhad parametru – graficky

$$\mathbf{P}\left(\boldsymbol{\mu} \in \left\{\bar{X} - z(1 - {}^{\alpha}/_{2}) \cdot \frac{\sigma}{\sqrt{n}} \; ; \; \bar{X} + z(1 - {}^{\alpha}/_{2}) \cdot \frac{\sigma}{\sqrt{n}}\right\}\right) = 1 - \alpha = 0,95$$



Červený interval je to „chybné tvrzení o intervalu spolehlivosti“. Červený interval nezahrnuje (nepokrývá) skutečnou hodnotu μ_0.
Pravděpodobnost této chyby je α (5 %).

Intervalový odhad parametru – nahrazení neznámého σ^2

Většinou neznáme σ^2 a nahrazujeme ho odhadem rozptylu $S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$.

Potom místo $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ dostáváme $\boldsymbol{T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}}$

a mění se i konfidenční interval:

$$\mathbf{P}\left(\mu \in \left\{\bar{X} - t_{(n-1)}\left(1 - \alpha/2\right) \cdot \frac{S}{\sqrt{n}} \; ; \; \bar{X} + t_{(n-1)}\left(1 - \alpha/2\right) \cdot \frac{S}{\sqrt{n}}\right\}\right) = 1 - \alpha$$

- Rozdělení ***t*** má ***(n – 1)*** stupňů volnosti! (viz odvození dále)
- Interval spolehlivosti spočítaný z t-rozdělení je širší, protože $t_{(n-1)}(1 - \alpha/2) > z(1 - \alpha/2)$. Odpovídá to nejistotě přidané použitím odhadu S^2.
- Odvození T statistiky dále:

Intervalový odhad parametru – odvození T statistiky

$$T = \frac{Z}{\sqrt{\frac{W}{k}}} \overset{??}{\longrightarrow} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}, \qquad \text{kde } Z \sim N(0,1) \text{ a } W = \sum_{i=1}^{k} Z_i^{\,2}, Z_i \sim N(0,1)$$

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}, \qquad \text{protože } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{... Normování } \bar{X}$$

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} \longrightarrow (n-1) \cdot S^2 = \sum (X_i - \bar{X})^2$$

$$\frac{(n-1) \cdot S^2}{\sigma^2} = \frac{\sum (X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^{\,2} = W$$

Normovaná X_i vypadá takto: $\frac{X_i - \mu}{\sigma}$

Chci tam $\bar{X}$ místo μ, ale ztrácím tím jeden stupeň volnosti.

Přidat σ je snadné, ale musím ji přidat na obě strany rovnice!

$$T = \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1) \cdot S^2}{\sigma^2}}{n-1}}} = \frac{(\bar{X} - \mu)\frac{\sqrt{n}}{\sigma}}{\sqrt{\frac{S^2 \cdot (n-1)}{\sigma^2 \cdot (n-1)}}} = \frac{(\bar{X} - \mu)\frac{\sqrt{n}}{\sigma}}{\frac{S}{\sigma}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$$

Centrální limitní věta (CLV) [central limit theorem]

Vysvětluje, proč můžeme použít na odhad střední hodnoty datového souboru s nenormálním rozdělením
aproximaci normálním rozdělením, pokud je výběr „dostatečně velký“.

Máme-li posloupnost $X_1, X_2, X_3, \dots, X_n$ nezávislých, stejně rozdělených náh. veličin, které mají nějaké rozdělení (nemusí být normální)
s $\boldsymbol{EX = \mu}$ a $\boldsymbol{varX = \sigma^2}$, a počet $\boldsymbol{n}$ jde do ∞,
potom za velmi obecných předpokladů konverguje normovaný součet $\boldsymbol{X_i}$ pro $\boldsymbol{n \to \infty}$ k rozdělení $\boldsymbol{N(0,1)}$:

≠ součet normovaných X_i

součet $\boldsymbol{X_i}$: $\sum_{i=1}^{n} X_i$ → normovaný součet: $\boldsymbol{Z_n = \frac{\sum X_i - E(\sum X_i)}{\sqrt{var(\sum X_i)}}}$

$$\boldsymbol{Z_n = \frac{\sum X_i - E(\sum X_i)}{\sqrt{var(\sum X_i)}} = \frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}} \xrightarrow[n \to \infty]{} \sim N(0,1)}$$

Toto lze aplikovat např. na průměr, relativní četnost či součet pořadí, také na testy o střední hodnotě nějakého rozdělení.

Použití CLV na aproximaci binomického rozdělení

$\boldsymbol{Y \sim Bi(n,p)}$, kde $\boldsymbol{Y = \sum_{i=1}^{n} X_i}$ a $\boldsymbol{X_i \sim Alt(p)}$

víme, že $\boldsymbol{\mathrm{E}X_i = p}$ a $\boldsymbol{varX_i = p(1-p)}$

tedy $\boldsymbol{EY = n \cdot p}$ a $\boldsymbol{varY = n \cdot p \cdot (1-p)}$

Podle CLV má náh. vel. $\boldsymbol{Z = \frac{Y - np}{\sqrt{np(1-p)}} \sim N(0,1)}$ pro velká $\boldsymbol{n}$.

Proto $\boldsymbol{Y \sim Bi(n,p)}$ může být pro velká n aproximována $\boldsymbol{\sim N(np, np(1-p))}$.

Zkušenosti starších říkají, že aproximace je dobře použitelná pro

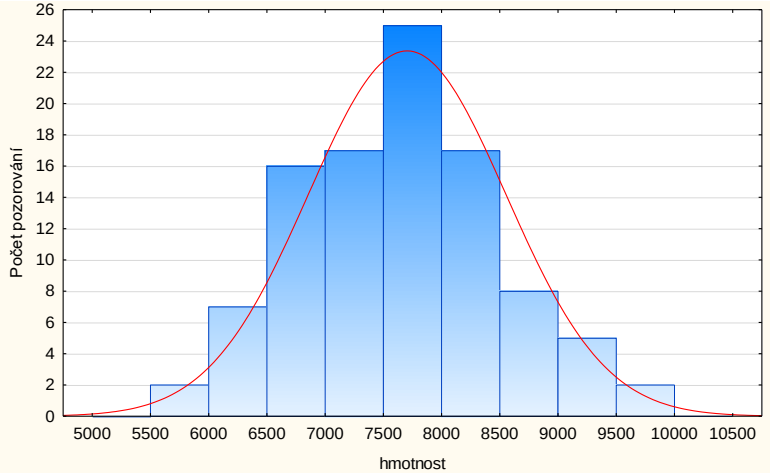
$\boldsymbol{np(1-p) > 9}$ nebo

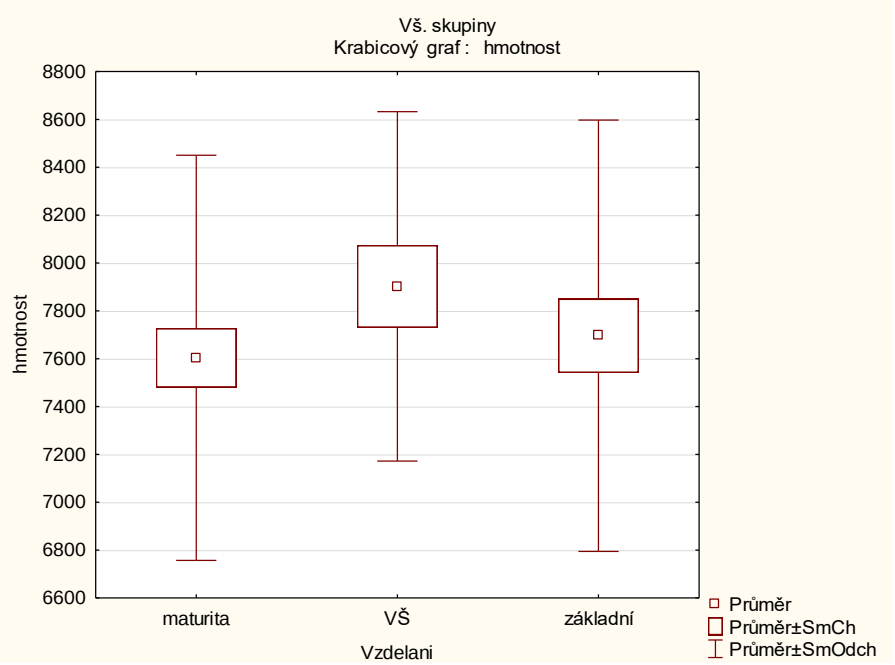
p	→	n
0.5		≥ 30
0.4 a ***0.6***		≥ 50
0.3 a ***0.7***		≥ 80
0.2 a ***0.8***		≥ 200
0.1 a ***0.9***		≥ 600

Prezentace odhadu a jeho vlastností - graficky a tabulkou

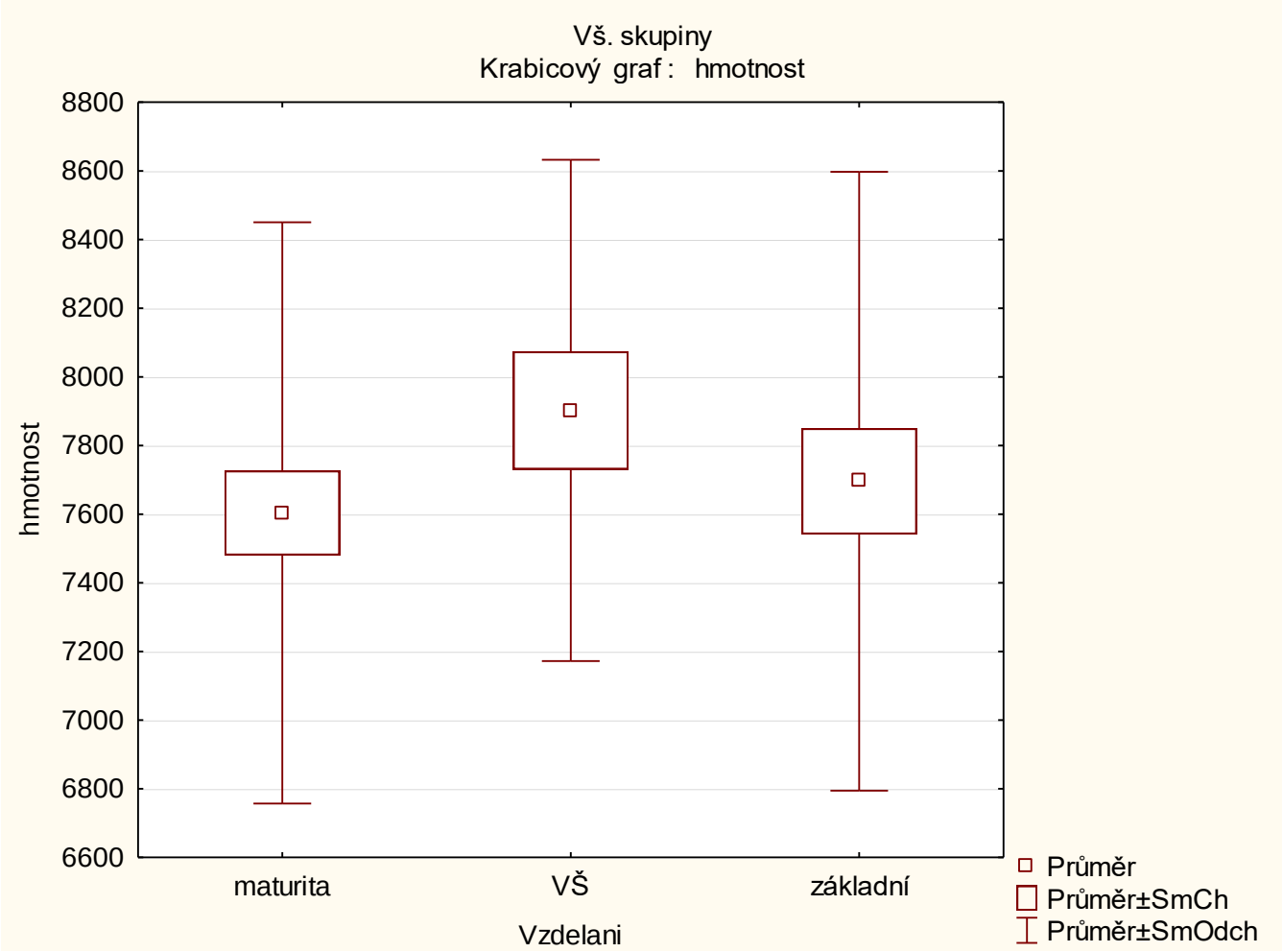
Hmotnosti miminek ve 24. týdnu podle vzdělání matky

Vzdělání matky	n	průměr	medián	SD	SE
Základní	34	7695,9	7775,0	901,2	154,6
Maturita	47	7603,5	7600,0	846,6	123,5
VŠ	18	7902,2	8000,0	730,0	172,1





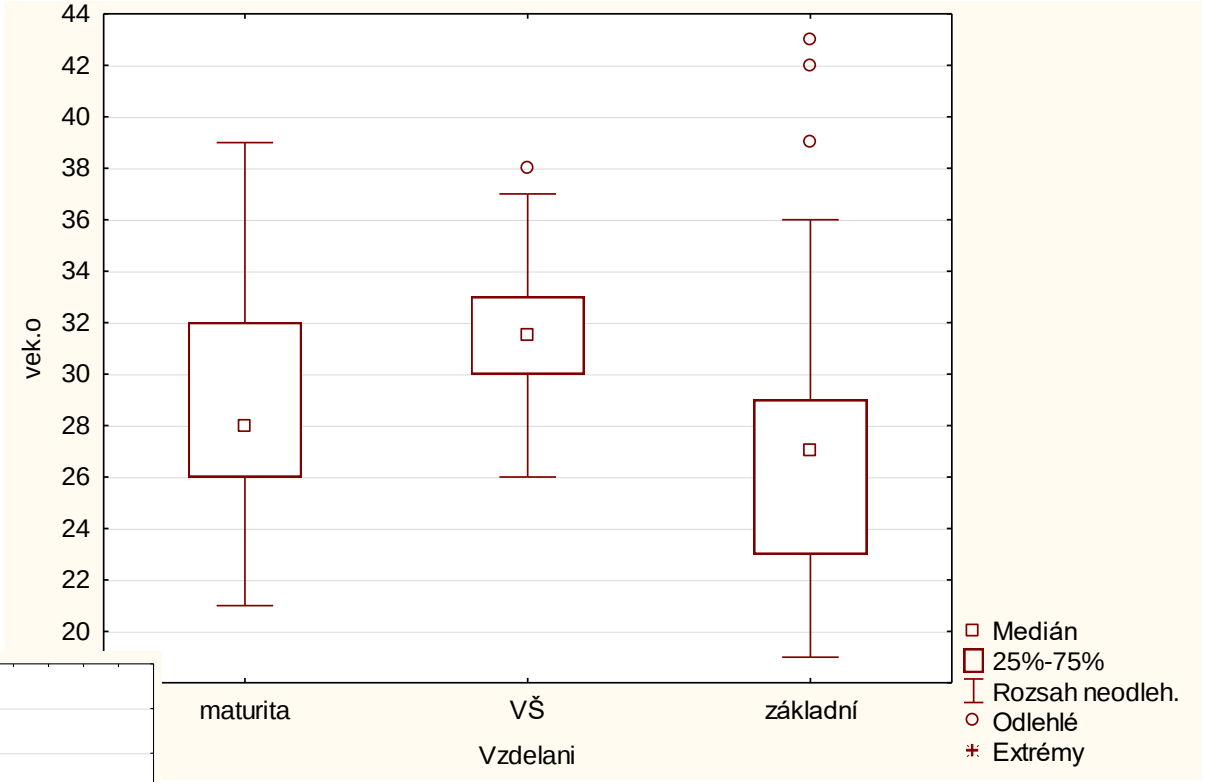
Prezentace odhadu a jeho vlastností - graficky a tabulkou

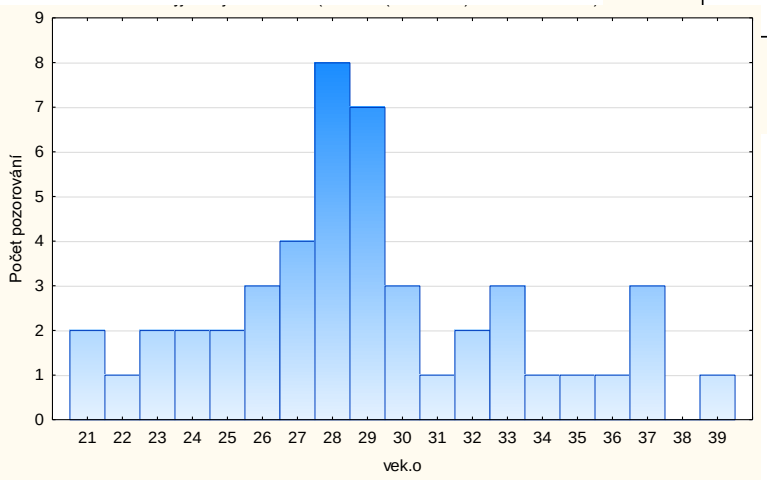


Jsou-li data symetrická, zobrazíme PRŮMĚR a jeho STŘEDNÍ CHYBU (SE)

Prezentace odhadu a jeho vlastností - graficky a tabulkou

Pro zešikmená data volím raději medián a kvartily…





… data nejsou ani normální ani šikmá…

Testování hypotéz [hypotheses testing]

Příklady:

- Z histogramu vidím, že data mají zhruba normální rozdělení. Ale tvrzení, že výběr pochází z normálního rozdělení, musím podepřít testem.
- Mám data o hmotnosti samců a samiček nějakého druhu a z grafické prezentace je vidět, že samečci jsou těžší. Statistický test řekne, zda je rozdíl mezi pohlavími „systematický“ nebo zda bylo věcí náhody, že někteří samečci byli těžší a posunuli průměr napravo.

Základní poučka metodologie vědy: shoda dat s hypotézou ještě neznamená, že hypotéza je pravdivá; na druhou stranu data odporující hypotéze ukazují, že hypotéza pravdivá není.
Proto hypotézu nelze na základě dat dokázat,
ale hypotézu lze na základě dat vyvrátit.

Ad příklad 2) chci vyvrátit tvrzení, že samci i samičky mají stejnou hmotnost.

Formulujeme **nulovou hypotézu H_0** [null hypothesis] a její negaci , tzv. **alternativní hypotézu H_1**, příp. **H_A** [alternative hypothesis].

Příklad. H_0: dva datové soubory mají stejnou střední hodnotu, $\mu_1 = \mu_2$;
H_1: střední hodnoty se liší, $\mu_1 \neq \mu_2$.

H_0: výběr pochází z normálního rozdělení;
H_1: výběr nepochází z normálního rozdělení

Máme 2 možná rozhodnutí: H_0 zamítáme nebo H_0 nezamítáme.
Následují 4 možné situace:

	SKUTEČNOST	
NAŠE ROZHODNUTÍ	H_0 platí	H_0 neplatí (platí H_1)
H_0 zamítáme	**Chyba 1. druhu: α** Prst. chyby ≤ α	SPRÁVNÉ ROZHODNUTÍ $P = 1 - \beta$ **síla testu**
H_0 nezamítáme	SPRÁVNÉ ROZHODNUTÍ $(P \geq 1 - \alpha)$	**Chyba 2. druhu: β** β většinou neznáme

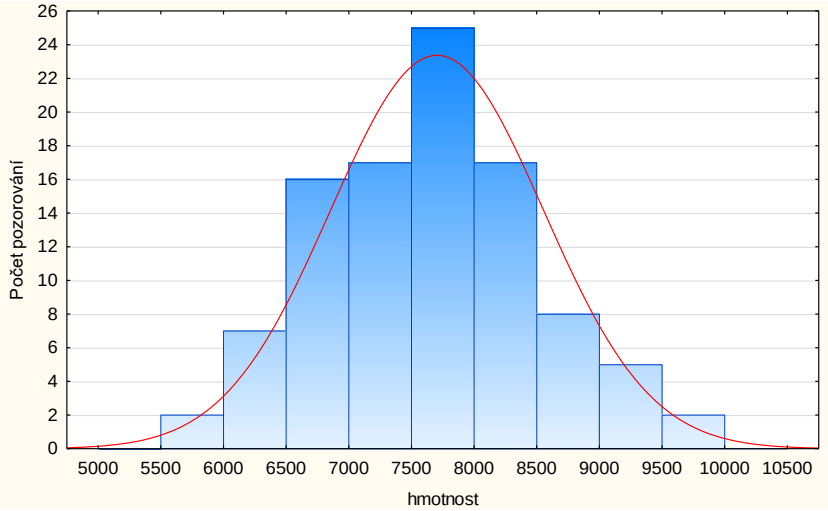
Testování hypotéz

Nulová hypotéza souvisí s nějakým předem daným uspořádáním dat.
Toto uspořádání je popsáno nějakým teoretickým rozdělením prstí nějaké náhodné veličiny.
Naše výběrová data tedy porovnáváme s určeným teoretickým rozdělením pomocí odhadu určené náhodné veličiny.

Nulovou hypotézu zamítáme tehdy, když naše uspořádání výběrového souboru je za předpokladu platnosti H_0 velmi nepravděpodobné.

Příklad: Test hypotézy o střední hodnotě normálního rozdělení.

Data „kojeni“: váha miminek ve 24. týdnu
Tabulková váha = 7600 g
Průměrná váha VŠ = 7900 g
$H_0: \mu_{vs} = 7600$; $H_1: \mu_{vs} \neq 7600$



Test hypotézy o střední hodnotě normálního rozdělení

Data „kojeni“: váha miminek ve 24. týdnu

- Mám výběr $X_1, X_2, \dots, X_n$ — *Jen matky s VŠ: n = 18*
- Předpokládám, že $X_i \sim N(\mu_X, {\sigma_X}^2)$ a jsou iid.
- Testuji, zda $\mu_X = \mu_0$ … μ_0 nějaké číslo, často 0
- Hypotéza H_0: $\mu_X = \mu_0$, H_1: $\mu_X \neq \mu_0$ — H_0: $\mu_X = 7600$; H_1: $\mu_X \neq 7600$
- Určím přípustnou chybu $\boldsymbol{\alpha}$ – **hladinu testu** — $\alpha = 0.05$
- μ_X odhadnu pomocí $\bar{X}$, protože vím, $E\bar{X} = \mu$ — $\bar{X} = 7900$
- Rozhodovací pravidlo: $|\bar{X} - \mu_0|$ … bude-li velký rozdíl, H_0 zamítnu
- Využijeme znalostí o $Z = \frac{Y - EY}{sd\ Y} \sim N(0,1)$
- V tomto příkladu $Y \to \bar{X}$, $EY \to E\bar{X} = \mu_X$, $sd\ Y \to sd\ \bar{X} = \frac{\sigma_X}{\sqrt{n}}$
- Zapracujeme předpoklad H_0: $\mu_X = \mu_0$ $\to$ $Z = \frac{\bar{X} - \mu_0}{\frac{\sigma_x}{\sqrt{n}}} = \frac{\bar{X} - \mu_0}{\sigma_x}\sqrt{n}$

Test hypotézy o střední hodnotě normálního rozdělení

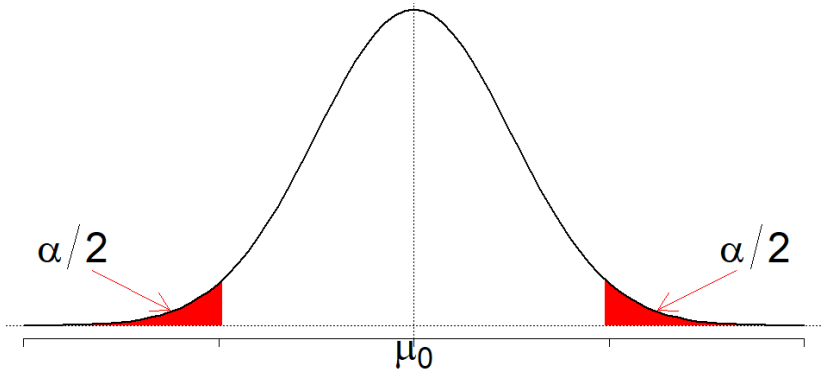
Odvodili jsme testovou statistiku Z, která má – za platnosti H_0 – rozdělení N(0, 1):

$$Z = \frac{\bar{X} - \mu_0}{\sigma_x}\sqrt{n} \sim N(0,1) \qquad \dots \ \mu_0 = \text{známé číslo}$$

V tuto chvíli otazník jen u σ_x
- a) známe
- b) neznáme

a) σ_x známe: rozhod. pravidlo bude $|Z| \geq z(1 - {}^{\alpha}/_{2})$, protože H_1: $\mu_X \neq \mu_0$

oboustranná alternativa
[two-tailed test]



b) σ_x neznáme: nahradíme ho odhadem $\sqrt{S_X{}^2} = S_X$

test. statistika $T = \frac{\bar{X} - \mu_0}{S_x}\sqrt{n} \sim t_{n-1}$ a rozhod. pravidlo $|T| \geq t_{n-1}\left(1 - \frac{\alpha}{2}\right)$.

Test hypotézy o střední hodnotě normálního rozdělení v číslech:

Příklad: Data „kojeni“: váha miminek ve 24. týdnu, kde matky mají VŠ

H_0: $\mu_{vs} = 7600$ g; H_1: $\mu_{vs} \neq 7600$ g

$\overline{X} = 7902$ g , σ_x neznáme → odhad S = 730

Testová statistika: $T = \frac{7902-7600}{730}\sqrt{18} = 1.76$

Kvantil $t_{(17)}(1 - 0{,}025) = 2.11$

Rozhodnutí: $|1.76| < 2.11$, proto nezamítám H_0, že skutečná $\mu_{vs} = 7600$ g.

P-hodnota provedeného testu p = 0.097, tj. 9.7 %

Test průměrů vůči referenční konstantě (hodnotě) (data_kojeni)
Zhrnout podmínku: Vzdelani="VŠ"

Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
hmotnost	7902,222	729,9602	18	172,0533	7600,000	1,756562	17	0,096989

Dosažená hladina významnosti testu

Také **p-hodnota** [p-value]
Je to pravděpodobnost, které odpovídá testová statistika coby kvantil.
Dnes je toto číslo velmi cennou informací v publikacích, proto je častou součástí výsledků.

Co nastává: Zvolili jsme $\alpha = 0.05$ (5 %) a …

- p-hodnota vyjde 0.0023, tj. 0.23 %. Výsledek je tedy hluboko za kritickou hranicí, výsledek (rozdíl) je evidentně průkazný. Hurá!
- $p = 0.049$, tedy zamítám H_0, ale jen velmi těsně.
- $p = 0.052$, tedy nezamítám H_0, ale také velmi těsně.
- $p = 0.43$, tedy H_0 nezamítám a je zřejmé, že se výsledek hranici 5 % ani zdaleka neblíží.

Formulace nulové hypotézy

a) Vidím, že samci a samičky mají skoro stejnou charakteristiku a chci je spojit do jedné skupiny. Potřebuji testem ukázat, že v datech není rozpor se „sjednocením“.
→ hledaný výsledek: „*nezamítám H_0*“, „*rozdíl mezi samci a samičkami je neprůkazný*“, apod. Tvrzení podporuje velká p-hodnota, např. 0.3 a větší.

b) Chci ukázat, že dvě skupiny se v nějaké charakteristice liší. Potom H_0 formuluji tak, abych ji na základě svých dat mohla zamítnout.
→ Hledaný výsledek: „*zamítám H_0 o tom, že mezi charakteristikami první a druhé skupiny není rozdíl*“. Tvrzení musí mít p-hodnotu ≤ α.

- Nezamítnutí H_0 znamená spíše nedostatek důkazů pro zamítnutí, než potvrzení platnosti H_0.
- **Pouze zamítnutím H_0 něco vědecky dokazujeme.**
- Odpověď při neúspěchu: „*Na základě dat nemůžeme zamítnout H_0.*“
- Nelze napsat: „*dokázali jsme nulovou hypotézu…*“ CHYBA!!

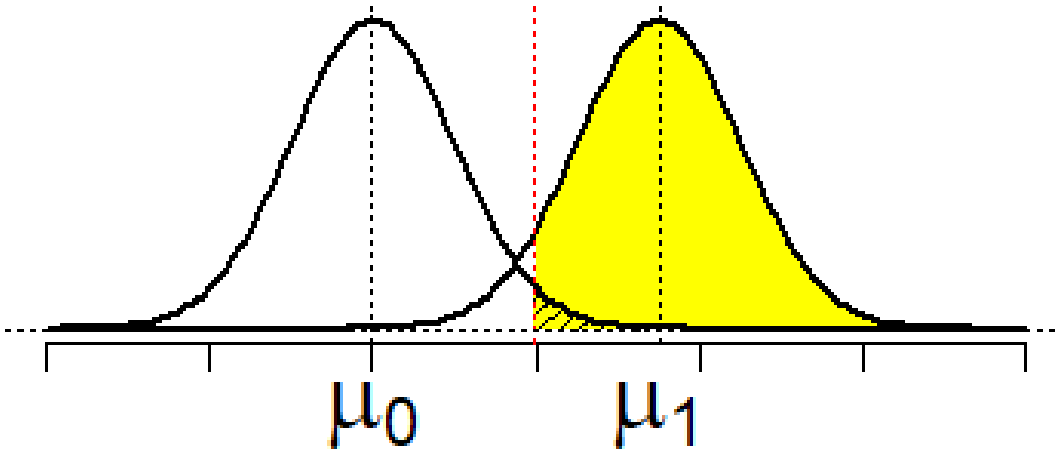
Neprůkaznost rozdílu, který jsme očekávali, je nejčastěji důsledkem toho, že buď rozdíly neexistují, nebo máme málo dat.

Poznámky k postupu

- Statistik má nejdříve formulovat hypotézu, zvolit rozhodovací pravidlo, určit hladinu testu, podle toho spočítat minimální rozsah výběru, a pak teprve sbírat data.
- Biolog nasbírá data, polovinu jich vyřadí a pak se ptá, co z toho lze otestovat ☺
- Přesto máme pokusy, kdy je třeba o rozsahu výběru i o hladině testu uvažovat předem -> plánování experimentů, výpočet potřebného rozsahu výběru tak, aby bylo možné dosáhnout potřebné hladiny testu α.

Síla testu (1 – β) [power of the test]

= pravděpodobnost, že nulovou hypotézu zamítneme, když ona neplatí
= pravděpodobnost, s jakou odhalíme neplatnost hypotézy → ta žlutá prst.



- Sílu testu většinou neznáme. Závisí na skutečném rozdělení výběrového souboru.
- Víme ale, že síla testu roste s odchylkou od nulové hypotézy a také s počtem pozorování (rozsahem výběru).
- Také platí, že čím menší je α, tím větší bude β.

Síla testu (1 – β)

Různé typy testů mají také různé síly, tím se zabývá teorie.
Nás pak zajímají praktické poznámky typu

- „test B je silnější než běžně používaný test A“
- „test C je silný, ale je citlivý na porušení předpokladů o normalitě dat“ (tzn. mám pěkná data z normálního rozd. => beru test C)
- „test D je spíše slabý, ale je robustní k narušení předpokladů“ (tzn. použiju ho tam, kde data nejsou zrovna příkladně gaussovská).

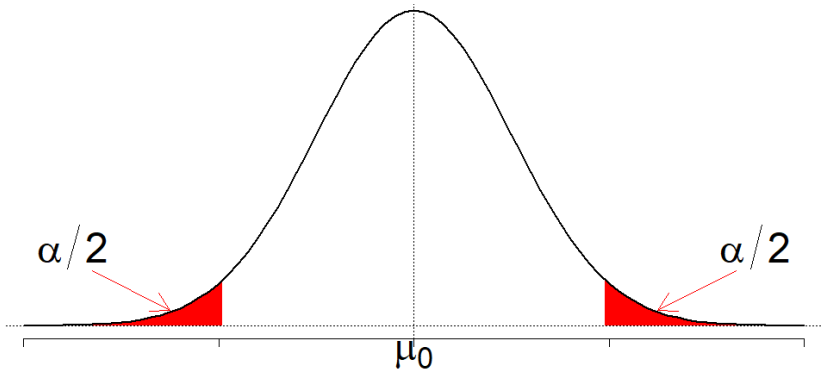
Oboustranná vs. jednostranná alternativa

[two-tailed vs. one-tailed alternative]

Oboustranná alternativa

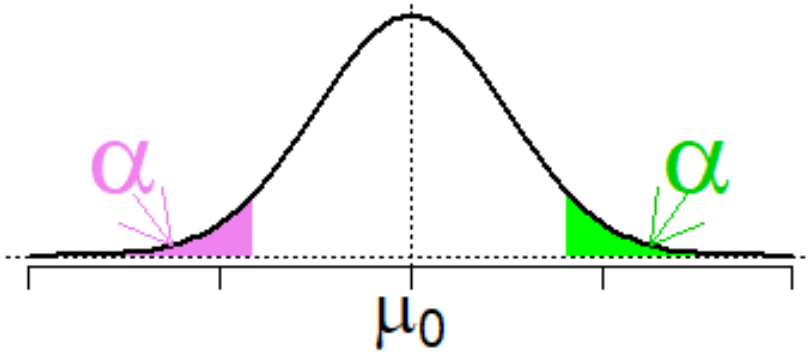
$H_0: \mu_X = \mu_0$, $\mathbf{H_1: \mu_X \neq \mu_0}$

… tedy $\boldsymbol{\mu_x}$ může být větší. Teorie případu nenapovídá nic o tom, na kterou stranu se rozdělení dat může posunout (přestože nám to napovídají čísla!)



Jednostranná alternativa

Pokud z povahy případu vyplývá, že např. střední hodnota může být jedině menší (větší) než testovaná hodnota μ_0, zapracuju tento fakt do H_1:



$H_0: \mu_X = \mu_0$, $\mathbf{H_1: \mu_X < \mu_0}$ nebo $\mathbf{H_1: \mu_X > \mu_0}$

Rozhodovací pravidlo: $T < t_{n-1}(\alpha)$ nebo $T > t_{n-1}(1-\alpha)$

Testování hypotéz - slovníček

Chyba 1. druhu – Type I error
Chyba 2. druhu – Type II error
Síla testu – power of the test
Hladina testu – signifikance level
Zamítnout hypotézu – to reject hypothesis
Oboustranný test – two-tailed test
Jednostranný test – one-tailed test, left/right-tailed test

Kritický obor – takové výsledky testové statistiky, kdy H_0 zamítáme
Obor přijetí – takové výsledky testové statistiky, kdy H_0 nezamítáme

Potřebný rozsah výběru

… v další přednášce.