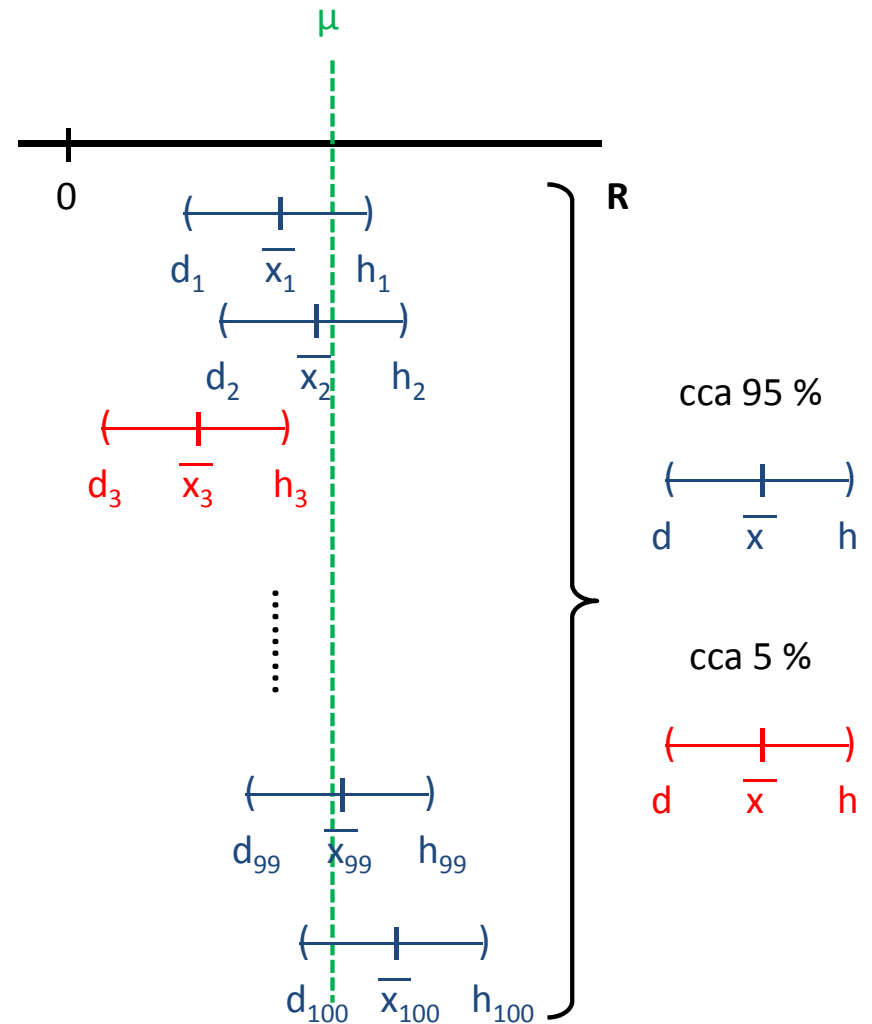


Opakování – vlastnosti výběrového průměru

- ➡ Čím lze vysvětlit následující vlastnosti výběrového průměru?
- ➡ **Rozdělení pravděpodobnosti výběrového průměru je tím méně variabilní čím více pozorování je v průměru zahrnuto.**
- ➡ **Rozdělení pravděpodobnosti výběrového průměru se s rostoucím n přestává podobat rozdělení původních dat a začíná se podobat rozdělení normálnímu.**

Opakování – interpretace intervalu spolehlivosti

➡ Jak lze interpretovat např. 95% interval spolehlivosti pro odhad střední hodnoty?



1. Motivace

Odhady a testy

➡ Zatím jsme se bavili hlavně o odhadech – pomocí odhadů popisujeme charakteristiky cílové populace.

➡ **Chceme jít v rozhodování dál:**

1. Chceme srovnávat.

➡ 1 náhodný výběr s předpokládanou hodnotou

➡ 2 náhodné výběry mezi sebou

➡ Více náhodných výběrů mezi sebou

2. Chceme hodnotit změnu náhodné veličiny vzhledem k nějaké intervenci.

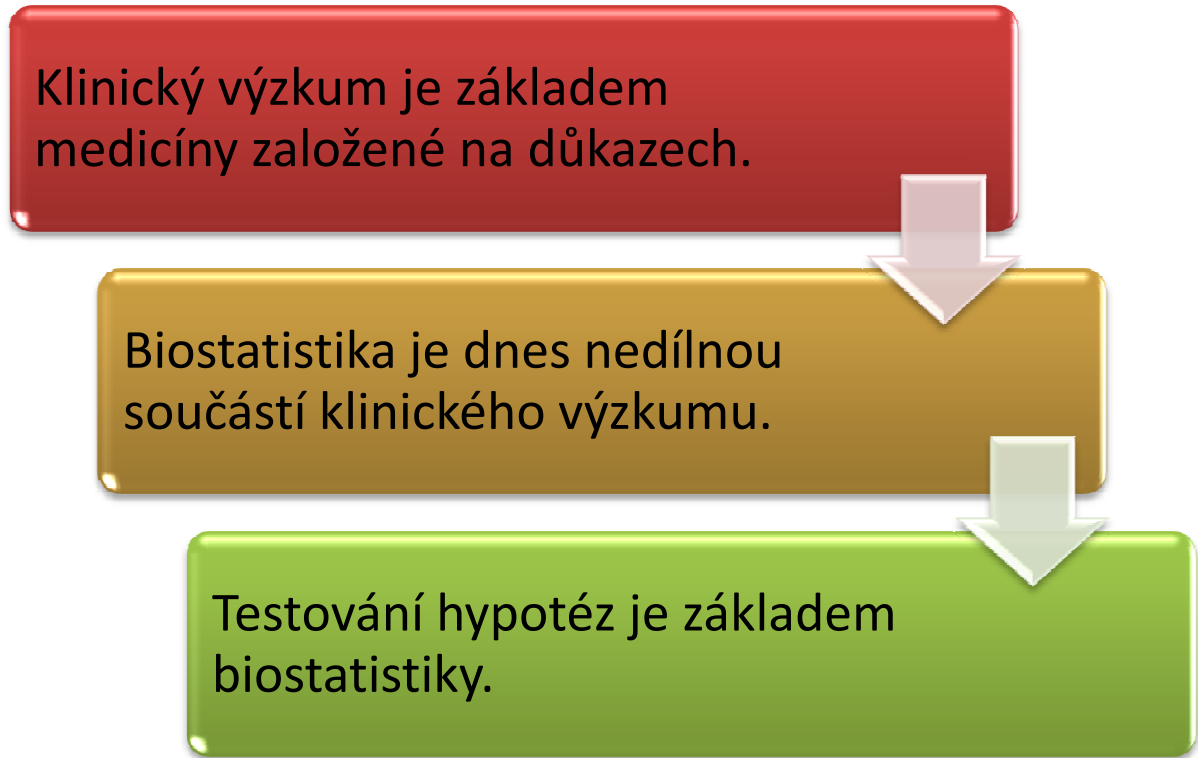
3. Chceme rozhodovat o nezávislosti dvou náhodných veličin.

4. Chceme rozhodovat o charakteru rozdělení náhodné veličiny.

Medicína založená na důkazech

- ➔ Úkolem zdravotního systému je zajistit dostupnými prostředky nejlepší možný zdravotní a psychický stav národa.
- ➔ K naplňování tohoto úkolu by měl pomoci princip nazvaný **medicína založená na důkazech** („evidence based medicine“).
- ➔ **Medicína založená na důkazech je proces zabývající se systematickým hledáním, hodnocením a hlavně využitím současných výsledků klinického výzkumu při poskytování péče jednotlivým pacientům.**
 - ➔ Poskytování důkazů pomocí klinického výzkumu a vědecké literatury.
 - ➔ Vytváření klinických doporučení (založených na důkazech) a jejich distribuce.
 - ➔ Implementace účinných a efektivních postupů pomocí výuky a řízení kvality.
 - ➔ Hodnocení dodržování doporučených postupů pomocí klinických auditů, indikátorů kvality a výsledků léčebné péče.

Role testování hypotéz



2. Principy a pojmy testování hypotéz

Hypotézy

➔ **Nulová hypotéza** („null hypothesis“) – tvrzení o neznámých vlastnostech rozdělení pravděpodobnosti sledované náhodné veličiny (na cílové populaci). Může být tvrzením o parametrech rozdělení nebo tvaru rozdělení pravděpodobnosti.

➔ **Nulová hypotéza má tvar:** $H_0 : \theta = \theta_0$

➔ **Alternativní hypotéza** – tvrzení o neznámých vlastnostech rozdělení pravděpodobnosti sledované náhodné veličiny, které popírá platnost nulové hypotézy. Vymezuje, jaká situace nastává, když nulová hypotéza neplatí.

➔ **Alternativní hypotéza má tvar:**

$$H_1 : \theta \neq \theta_0$$

$$H_1 : \theta < \theta_0$$

$$H_1 : \theta > \theta_0$$



Testování hypotéz

- ➡ Testování hypotéz se zabývá rozhodováním o platnosti stanovených hypotéz na základě pozorovaných dat.
- ➡ Platnost hypotéz ověřujeme pomocí **statistického testu** – rozhodovacího pravidla, které každému náhodnému výběru přiřadí právě jedno ze dvou možných rozhodnutí – H_0 nezamítáme nebo H_0 zamítáme.
- ➡ Základy moderního testování hypotéz položili J. Neyman a E. S. Pearson.



Příklady – řešené problémy

1. Urychluje použití antibiotika ve srovnání s použitím běžné dezinfekce hojení rány?
2. Je průměrný objem prostaty mužů nad 70 let stejný jako průměrný objem prostaty celé mužské populace?
3. Je efekt snížení systolického tlaku novým antihypertenzivem stejný u hypertoniků, kteří kouří, jako u hypertoniků, kteří nekouří?
4. Liší se AML, ALL, CML a CLL v aktivitě vybraných genů?

Příklady – hypotézy

1. Urychluje použití antibiotika ve srovnání s použitím běžné dezinfekce hojení rány?

Střední doba hojení s antibiotiky: θ_1 $H_0 : \theta_1 = \theta_2$

Střední doba hojení bez antibiotik: θ_2 $H_1 : \theta_1 \neq \theta_2$

2. Je průměrný objem prostaty mužů nad 70 let stejný jako průměrný objem prostaty celé mužské populace?

Střední objem prostaty mužů nad 70 let : θ_1 $H_0 : \theta_1 = \theta_0$

Populační hodnota (konstanta): θ_0 $H_1 : \theta_1 > \theta_0$

Příklady – hypotézy

3. Je efekt snížení systolického tlaku novým antihypertenzivem stejný u hypertoniků, kteří kouří, jako u hypertoniků, kteří nekouří?

Střední hodnota efektu u kuřáků: θ_1 $H_0 : \theta_1 = \theta_2$

Střední hodnota efektu u nekuřáků: θ_2 $H_1 : \theta_1 < \theta_2$

4. Liší se AML, ALL, CML a CLL v aktivitě vybraných genů?

Střední hodnota exprese genu g u AML, ALL, CML, CLL: $\theta_{AML}^g, \theta_{ALL}^g, \theta_{CML}^g, \theta_{CLL}^g$

$$H_0 : \theta_{AML}^g = \theta_{ALL}^g = \theta_{CML}^g = \theta_{CLL}^g$$

H_1 : nejméně jedno θ^g je odlišné od ostatních

Proč nulová hypotéza vyjadřuje nepřítomnost efektu?

Proč nulová hypotéza vyjadřuje nepřítomnost efektu?

- ➡ Nulová hypotéza odráží fakt, že se něco nestalo nebo neprojevilo → je stanovena obvykle jako opak toho, co chceme experimentem prokázat.
- ➡ **Nulová hypotéza je postavena tak, abychom ji mohli pomocí pozorovaných hodnot vyvrátit.**
- ➡ Pro zamítnutí platnosti nulové hypotézy nám totiž stačí najít jeden příklad, kdy nulová hypotéza neplatí – tím příkladem má být náš náhodný výběr (naše pozorovaná data).
- ➡ Zamítnout nulovou hypotézu je jednodušší než nulovou hypotézu potvrdit.

Co se při rozhodování může stát

➔ Vzhledem k nulové hypotéz máme čtyři možnosti výsledku rozhodovacího procesu:

Rozhodnutí	Skutečnost	
	H_0 platí	H_0 neplatí
H_0 nezamítneme	správné přijetí platné nulové hypotézy	chyba II. druhu
H_0 zamítneme	chyba I. druhu	správné zamítnutí neplatné nulové hypotézy

➔ **Při rozhodování se můžeme mýlit**, můžeme se dopustit dvou chybných úsudků.



Analogie se soudním procesem

- ➔ Ctíme presumpci nevinny = předpokládáme, že nulová hypotéza platí.
- ➔ Požadujeme důkaz pro prokázání viny = na základě dat chceme ukázat, že nulová hypotéza neplatí.
- ➔ Když nám bude stačit málo důkazů, zvýší i procento odsouzených, kteří jsou skutečně vinni = **správné zamítnutí neplatné nulové hypotézy**, ale zároveň se zvýší se procento odsouzených nevinných = **chyba I. druhu**.
- ➔ Když budeme požadovat hodně důkazů, zvýší se procento nevinných, kteří budou osvobozeni = **správné přijetí platné nulové hypotézy**, ale zároveň se zvýší i procento vinných, kteří budou osvobozeni = **chyba II. druhu**.

Pravděpodobnost výsledků rozhodovacího procesu

Rozhodnutí	Skutečnost	
	H_0 platí	H_0 neplatí
H_0 nezamítneme	správné rozhodnutí $P = 1 - \alpha$	chyba II. druhu $P = \beta$
H_0 zamítneme	chyba I. druhu $P = \alpha$	správné rozhodnutí $P = 1 - \beta$

➡ Jak je vidět z analogie se soudním procesem, nelze zároveň minimalizovat α i β . V praxi je nutné více hlídat α → předem stanovíme maximální hranici pro α (**hladina významnosti testu**, „level of significance“) a za této podmínky minimalizujeme β .



Proč hlídat spíše α než β ?

➡ Benjamin Franklin:

„It is better that 100 guilty persons should escape than that one innocent person should suffer.“

Statistický test

- ➔ Testování hypotéz probíhá na základě dat (ve frekventistické statistice výhradně na základě dat).
- ➔ **Testované hypotéze odpovídá statistický test**, respektive testová statistika, která umožní ověřit platnost nulové hypotézy.
- ➔ **Testová statistika** je transformací pozorovaných dat s rozdělením pravděpodobnosti, sama tedy **má také rozdělení pravděpodobnosti**. Rozdělení pravděpodobnosti testové statistiky za platnosti H_0 se označuje jako „null distribution“.
- ➔ **Na základě dat vypočítáme hodnotu testové statistiky, kterou srovnáme s kvantilem (kritickou hodnotou) jejího rozdělení odpovídajícím zvolené hladině významnosti testu α .**

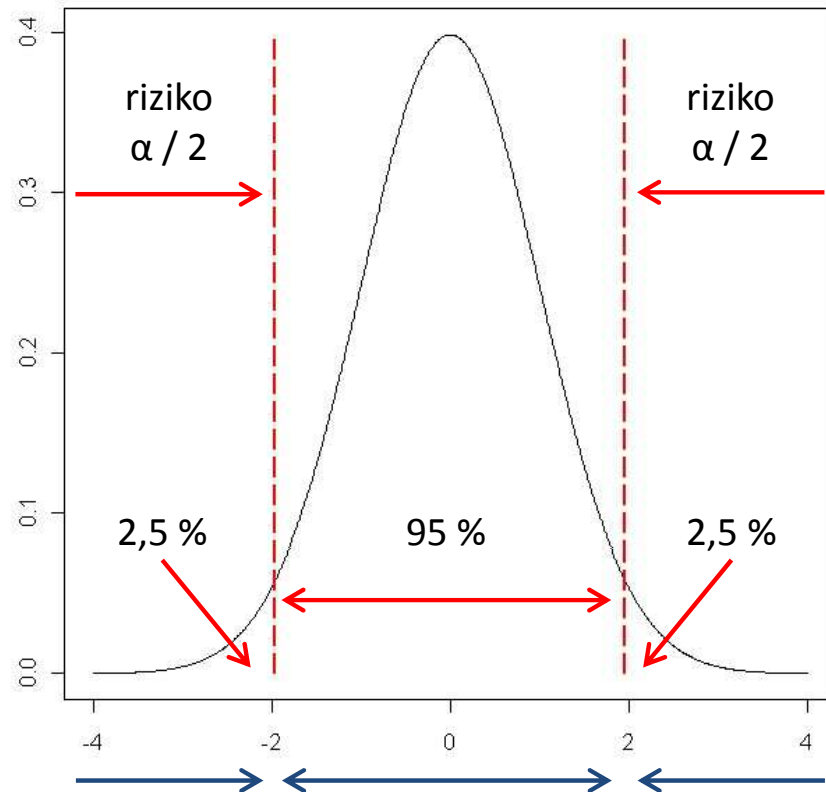
Zamítnutí / nezamítnutí nulové hypotézy

- ➡ Hodnotu testové statistiky srovnáme s kvantilem (**kritickou hodnotou**) jejího rozdělení odpovídajícím zvolené hladině významnosti testu α .
- ➡ Představuje-li pozorovaná hodnota testové statistiky extrémnější (méně pravděpodobnou) hodnotu v rámci rozdělení odpovídajícího nulové hypotéze než je kritická hodnota (kvantil) odpovídající zvolenému riziku α , pak nulovou hypotézu zamítáme.
- ➡ Riziko špatného rozhodnutí, které podstupujeme, buď rovnoměrně rozdělujeme na obě extrémní varianty výsledku (**oboustranný test**) nebo uvažujeme pouze jednu extrémní variantu výsledku (**jednostranný test**).

Zamítnutí / nezamítnutí nulové hypotézy

Oboustranný test při $\alpha = 0,05$

$$H_0 : \theta_1 = \theta_2 \quad H_1 : \theta_1 \neq \theta_2$$



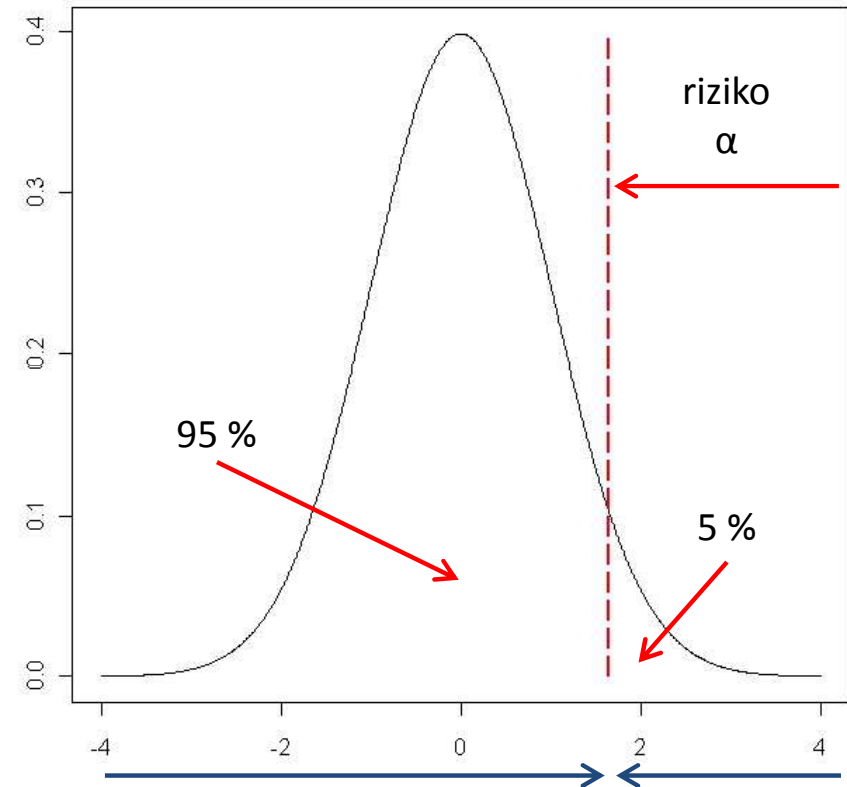
Padne-li testová statistika sem – **zamítáme H_0**

Padne-li testová statistika sem – **nezamítáme H_0**

Padne-li testová statistika sem – **zamítáme H_0**

Jednostranný test při $\alpha = 0,05$

$$H_0 : \theta_1 = \theta_0 \quad H_1 : \theta_1 > \theta_0$$



Padne-li testová statistika sem – **nezamítáme H_0**

Padne-li testová statistika sem – **zamítáme H_0**

Co znamená „padnutí testové statistiky“

➡ Je-li hodnota testové statistiky větší než kvantil příslušný riziku α , pak mohly nastat dvě situace:

1. **buď H_0 platí a my jsme pozorovali málo pravděpodobný jev**
2. **H_0 neplatí**

➡ My pracujeme s rizikem α , tedy málo pravděpodobné jevy jsou součástí našeho rizika, proto v tomto případě volíme možnost 2 a zamítáme H_0 .

Příklad – z-test pro jeden výběr

- ➔ Při populačním epidemiologickém průzkumu se zjistilo, že průměrný objem prostaty u mužů je 32,73 ml ($SD = 18,12$ ml). Na hladině významnosti testu $\alpha = 0,05$ chceme ověřit, jestli se muži nad 70 let liší od celé populace. Máme náhodný výběr o velikosti $n = 100$ a výběrový průměr 36,60 ml.
- ➔ Chceme ověřit platnost $H_0 : \mu = 32,73$ proti $H_1 : \mu \neq 32,73$
- ➔ Platí-li H_0 , pak $\bar{X} \sim N(\mu = 32,73, \sigma/\sqrt{n} = 1,812)$ (předpokládáme, že známe σ)
- ➔ **Jak na to?**

Příklad – z-test pro jeden výběr

- ➔ Při populačním epidemiologickém průzkumu se zjistilo, že průměrný objem prostaty u mužů je 32,73 ml ($SD = 18,12$ ml). Na hladině významnosti testu $\alpha = 0,05$ chceme ověřit, jestli se muži nad 70 let liší od celé populace. Máme náhodný výběr o velikosti $n = 100$ a výběrový průměr 36,60 ml.
- ➔ Chceme ověřit platnost $H_0 : \mu = 32,73$ proti $H_1 : \mu \neq 32,73$
- ➔ Platí-li H_0 , pak $\bar{X} \sim N(\mu = 32,73, \sigma/\sqrt{n} = 1,812)$ (předpokládáme, že známe σ)
- ➔ Z CLV víme, že by mělo platit: $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- ➔ Pokud tedy výběrový průměr patří do rozdělení $N(\mu = 32,73, \sigma/\sqrt{n} = 1,812)$ neměla by jeho hodnota být vzhledem k tomuto rozdělení nijak extrémní.

Příklad – z-test pro jeden výběr

- Chceme ověřit platnost $H_0 : \mu = 32,73$ proti $H_1 : \mu \neq 32,73$
- Platí-li H_0 , pak $\bar{X} \sim N(\mu = 32,73, \sigma/\sqrt{n} = 1,812)$ (předpokládáme, že známe σ)

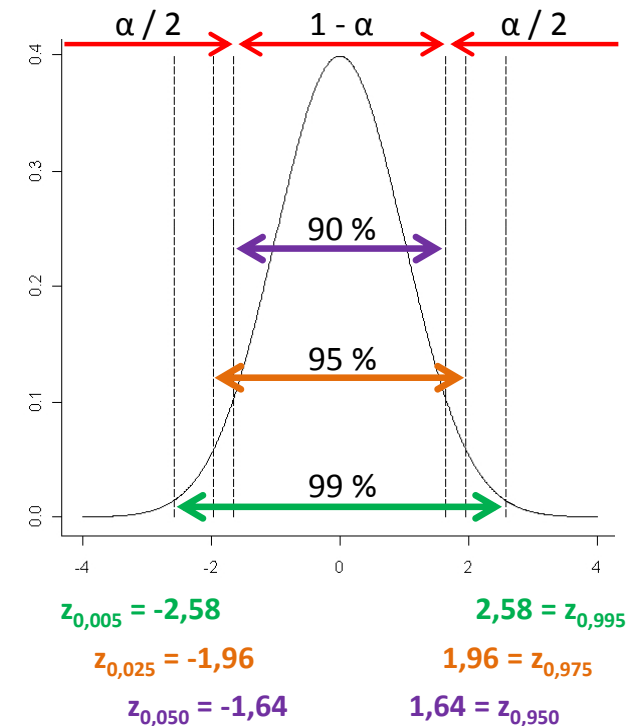
→ Z CLV víme, že by mělo platit: $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

→ Vypočteme hodnotu testové statistiky:

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

$$z = \frac{36,60 - 32,73}{18,12/\sqrt{100}} = \frac{3,87}{1,812} = 2,14$$

- **Můžeme zamítnout nulovou hypotézu na hladině významnosti testu $\alpha = 0,05$ nebo ne?**



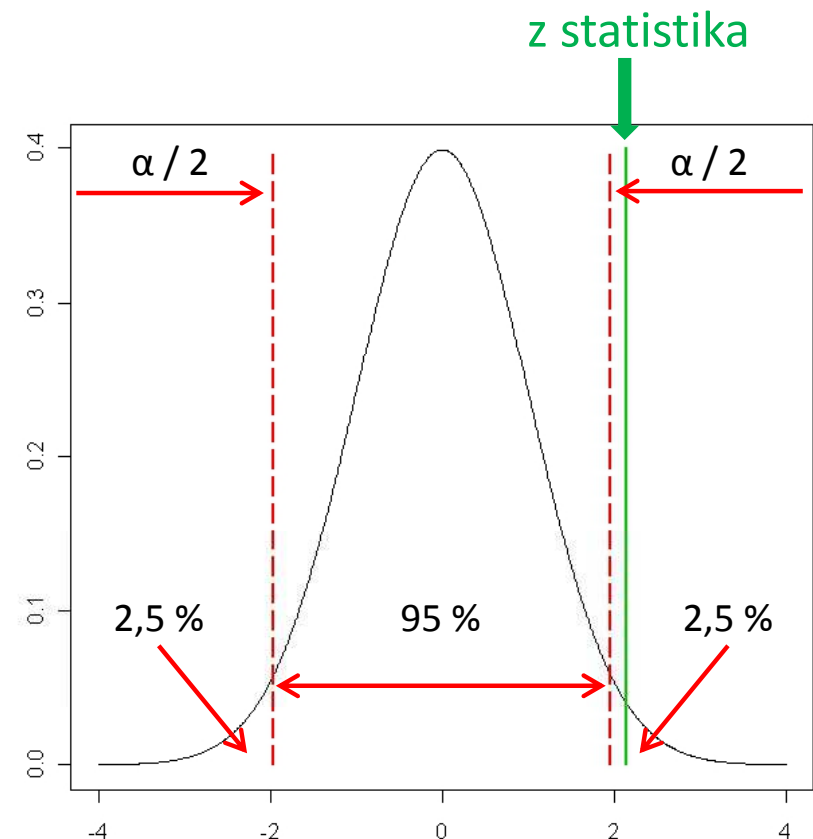
Příklad – z-test pro jeden výběr

➔ Hodnota testové statistiky: $z = \frac{36,60 - 32,73}{18,12 / \sqrt{100}} = \frac{3,87}{1,812} = 2,14$

➔ Můžeme zamítnout nulovou hypotézu na hladině významnosti testu $\alpha = 0,05$ nebo ne?

$$z = 2,14 > 1,96 = z_{0,975} = z_{1-\alpha/2}$$

➔ Nulovou hypotézu o rovnosti objemu prostaty u mužů nad 70 let populační hodnotě 32,73 ml zamítáme na hladině významnosti $\alpha = 0,05$, protože výsledná hodnota z statistiky je větší než kritická hodnota (příslušný kvantil) rozdělení $N(0,1)$.



3. P-hodnota a její interpretace

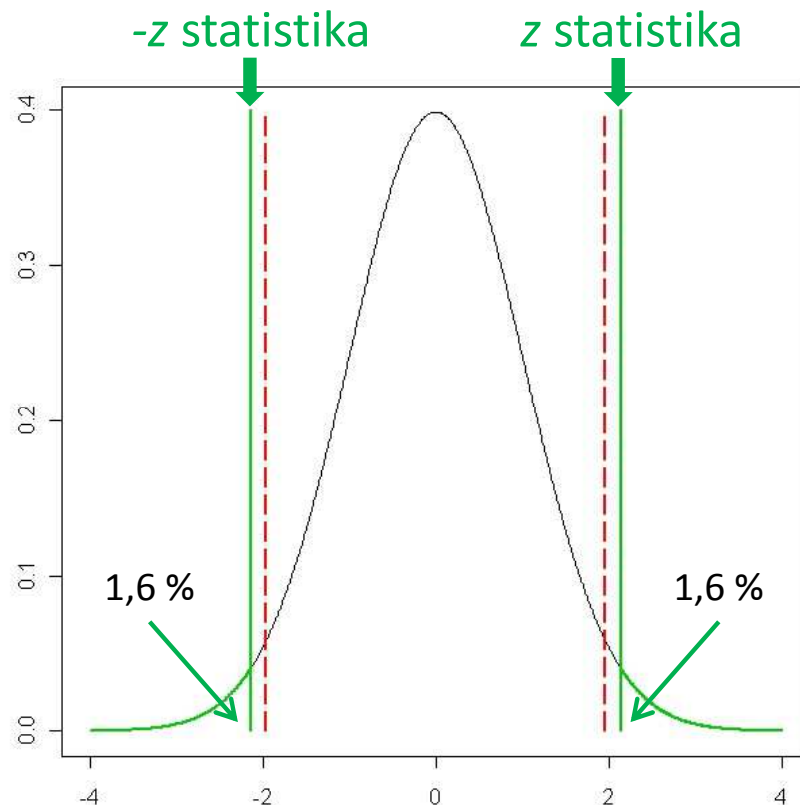
P-hodnota

- ➔ P-hodnota vyjadřuje pravděpodobnost za platnosti H_0 , s níž bychom získali stejnou nebo extrémnější hodnotu testové statistiky (samozřejmě vzhledem k jednostrannosti nebo oboustrannosti testu).
- ➔ Číselně (ale ne filozoficky) ekvivalentní je tzv. dosažená hladina významnosti testu („attained significance level“), což je nejmenší hladina významnosti α , při které bychom ještě zamítnuli H_0 .
- ➔ V praxi se často obě hodnoty zaměňují!



Příklad – p -hodnota

- ➡ Vraťme se k příkladu s objemem mužské prostaty – hodnota testové statistiky $z = 2,14$. Jaká jí odpovídá p -hodnota?
- ➡ **Důležité je uvědomit si, že máme oboustrannou alternativní hypotézu!**



p -hodnota pro oboustrannou alternativu:

$$p = 2 * (1 - P(Z \leq z_{test}))$$

p -hodnota pro z-test z příkladu:

$$p = 2 * (1 - P(Z \leq 2,14))$$

$$p = 2 * (1 - 0,984) = 0,032$$

Velikost vzorku a významnost výsledku

➔ Vraťme se k příkladu s objemem mužské prostaty – ALE s $n = 10$!

➔ Chceme ověřit platnost $H_0 : \mu = 32,73$ proti $H_1 : \mu \neq 32,73$

➔ Platí-li H_0 , pak $\bar{X} \sim N(\mu = 32,73, \sigma/\sqrt{n} = 5,73)$ (předpokládáme, že známe σ)

➔ Pak

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

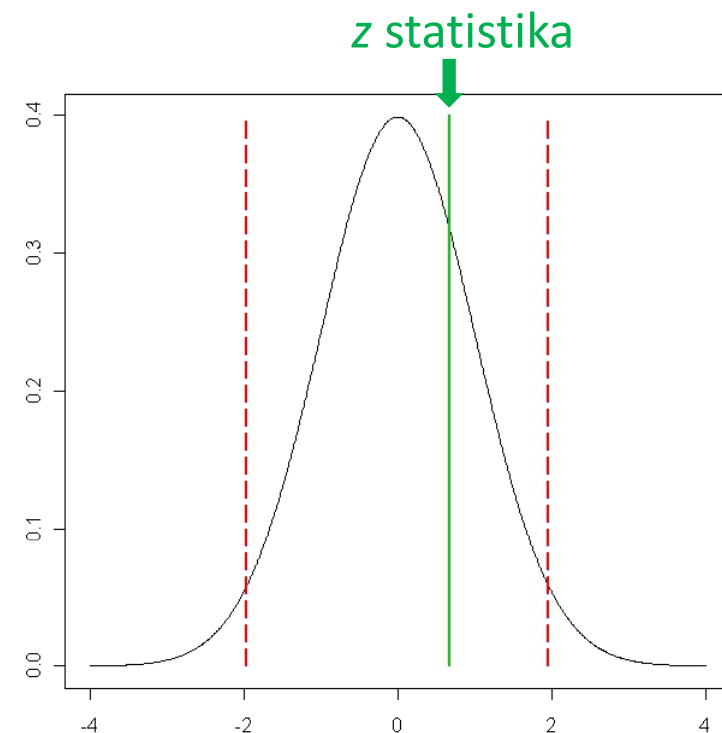
$$z = \frac{36,60 - 32,73}{18,12/\sqrt{10}} = \frac{3,87}{5,73} = 0,68$$

$$z = 0,68 < 1,96 = z_{0,975} = z_{1-\alpha/2}$$

➔ **H_0 nyní nezamítáme!**

➔ **Rozdíl se nezměnil, pouze je menší n !**

➔ **Máme méně informace.**



Důležité poznámky k testování hypotéz

- ➔ **Nezamítnutí nulové hypotézy neznamena automaticky její přijetí!** Může se jednat o situaci, kdy pro zamítnutí nulové hypotézy nemáme dostatečné množství informace.
- ➔ **Dosažená hladina významnosti testu (ať už 0,05, 0,01 nebo 0,10) nesmí být slepě brána jako hranice pro existenci/neexistenci testovaného efektu.** Neexistuje jasná hranice pro významnost či nevýznamnost – často je velmi malý rozdíl mezi p -hodnotou 0,04 a p -hodnotou 0,06.
- ➔ **Malá p -hodnota nemusí znamenat velký efekt.** Hodnota testové statistiky a odpovídající p -hodnota může být ovlivněna velkou velikostí vzorku a malou variabilitou pozorovaných dat.
- ➔ **Výsledky testování musí být nahlíženy kriticky** – jedná se o závěr založený „pouze“ na jednom výběrovém souboru.

Spojitosť s intervaly spolehlivosti

- ➔ Opět se vraťme příkladu s objemem mužské prostaty (pro $n = 100$).
- ➔ Chceme ověřit platnost $H_0 : \mu = 32,73$ proti $H_1 : \mu \neq 32,73$
- ➔ Na základě $n = 100$ jsme nulovou hypotézu zamítli.
- ➔ **Zkusme vypočítat 95% interval spolehlivosti pro μ (tedy IS s $\alpha = 0,05$):**

$$Z = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \quad \bar{X} = 36,60; \quad \frac{\sigma}{\sqrt{n}} = \frac{18,12}{\sqrt{100}} = 1,812; \quad z_{0,975} = 1,96$$

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

$$95\% \text{ IS} = (36,60 - 1,812 * 1,96; 36,60 + 1,812 * 1,96) = (33,05; 40,15)$$

- ➔ **95% IS neobsahuje H_0 . Co nám to říká?**

Spojitost s intervaly spolehlivosti

- ➔ Můžeme zamítnout H_0 , protože 95% IS neobsahuje předpokládanou hodnotu neznámého parametru (neobsahuje H_0). Opět podstupujeme riziko $\alpha = 0,05$, že se mýlíme (tedy, že jsme naším 95% IS nepokryli hodnotu μ).
- ➔ Testování hypotéz a intervaly spolehlivosti jsou velmi často ekvivalentní, ale popisují trochu něco jiného: **Interval spolehlivosti charakterizuje přesnost bodového odhadu, zatímco test nulové hypotézy se zaměřuje na pravděpodobnostní model v pozadí.**
- ➔ Vždy by měl být vedle výsledku testu publikována i velikost dosaženého efektu s příslušným intervalem spolehlivosti. Ze samotné p -hodnoty testu nebo rozhodnutí zamítáme H_0 /nezamítáme H_0 není zřejmé, v jakých mezích se velikost účinku (rozdílu) pohybuje.

4. Síla testu

Síla testu

- ➡ Pravděpodobnost chyby II. druhu značíme β .
- ➡ $1 - \beta$ se nazývá **síla testu** a vyjadřuje pravděpodobnost, že zamítneme H_0 ve chvíli, kdy H_0 opravdu neplatí.
- ➡ Snažíme se sílu testu optimalizovat při zachování hladiny významnosti testu α
→ princip výpočtu velikosti experimentálního vzorku před provedením studie
(budeme se tomu věnovat někdy příště).
- ➡ Optimalizovat sílu testu a velikost vzorku předem není triviální, můžeme narazit na spoustu problémů – biologické limity, etické limity, finanční limity.

Faktory ovlivňující sílu testu

- ➔ **Velikost vzorku:** čím více pozorování (informace o platnosti nulové hypotézy), tím větší má test sílu. Stejně jako u intervalů spolehlivosti, síla testu roste s odmocninou z n .
- ➔ **Velikost efektu (účinku):** velikost rozdílu v neznámých parametrech také ovlivňuje sílu testu. Vždy je jednodušší identifikovat jako významný velký efekt, např. velký rozdíl ve středních hodnotách objemu prostaty dvou populací. Naopak je těžší prokázat jako významný menší efekt (menší rozdíl).
- ➔ **Variabilita dat:** variabilita dat zvyšuje variabilitu odhadů a ztěžuje tak rozhodnutí o H_0 . Čím více jsou pozorované hodnoty variabilní, tím více dat bude potřeba pro přesný odhad velikosti účinku (rozdílu).
- ➔ **Hladina významnosti:** snížíme-li hladinu významnosti testu (např. zvolíme 0,01 místo 0,05), bude obtížnější H_0 zamítnout → sníží se síla testu.

5. Statistická versus klinická/biologická významnost

Klíčové principy – významnost

Praktická významnost

		Praktická významnost	
		ANO	NE
Statistická významnost	ANO	OK, praktická i statistická významnost jsou ve shodě.	Významný výsledek je statistický artefakt, prakticky nevyužitelný.
	NE	Výsledek může být pouhá náhoda, neprůkazný výsledek.	OK, praktická i statistická významnost jsou ve shodě.



Statisticky nevýznamný výsledek neznamená, že pozorovaný rozdíl ve skutečnosti neexistuje! Může to být způsobeno nedostatečnou informací v pozorovaných datech!



Problém násobného testování hypotéz

- ➔ V klinickém výzkumu se často potřebujeme testovat více hypotéz zároveň – např. při hodnocení stejného primárního parametru v rámci různých podskupin souboru pacientů (A , B , C a D) → Je zajímavé podívat se na rozdíl i mezi podskupinami, tedy podívat se, jak se liší skupiny A a B , B a C , apod.
- ➔ **Tento fenomén v praxi vede k tzv. problému násobného testování hypotéz.**
- ➔ **S narůstajícím počtem testovaných hypotéz nám roste také pravděpodobnost získání falešně pozitivního výsledku, tedy pravděpodobnost toho, že se při našem testování zmýlíme a ukážeme na statisticky významný rozdíl tam, kde ve skutečnosti žádný neexistuje (chyba I. druhu).**
- ➔ Použití korekčních procedur: Bonferroniho procedura, metoda Steela a Dwasse.

Příklad

- ➔ **Modelová situace:** provedeme zároveň 60 testů (v době srovnávání biochemických a genetických parametrů to není zase tolik). Použijeme-li klasickou hladinu významnosti 0,05 (resp. 5 %), máme pro každý test 5% riziko získání falešně pozitivního výsledku. **Vynásobíme-li 60 a 0,05, vyjde nám, že zhruba u 3 testů bychom měli dospět k falešně statisticky významnému závěru.**
- ➔ V případě genomických analýz, kde jsou často různé testy pouze formou exploratorní analýzy, nemusí být přítomnost falešně pozitivních výsledků fatální, v klinické praxi to však může vést k zavádějícím výsledkům a mylným interpretacím.