Marie Turčičová

# Úvod do testování hypotéz a jednovýběrový t-test

Předpokládejme, že jsme zvážili sedm kojenců (chlapců) a dostali jsme tato data:

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | |
|---|---|---|---|---|---|---|---|
| 8.0 | 8.1 | 8.7 | 8.3 | 7.9 | 6.5 | 9.6 | [kg] |

Chceme otestovat pravdivost tvrzení, že střední hmotnost chlapců v celé populaci (označme si ji $\mu$) je 8.2 kg. Test se nám zdá rozumné provést tak, že porovnáme výběrový průměr z našich dat (označíme ho tradičně $\bar{X}$) s hodnotou 8.2 a bude-li od této hodnoty „moc daleko“, tvrzení označíme za nepravdivé. Přitom bychom chtěli, aby pravděpodobnost omylu byla malá.

Abychom byli schopni kvantifikovat ono „moc daleko“, musíme zavést další předpoklady, které se budou týkat pravděpodobnostního rozdělení sledovaného znaku v populaci. Testů o střední hodnotě existuje celá řada a liší se od sebe právě těmito dodatečnými předpoklady.

Pro náš první statistický test budeme předpokládat, že náš náhodný výběr $X_1, X_2, \ldots, X_n$ má normální rozdělení $\mathrm{N}(\mu,\sigma^2)$, tj. že hmotnost chlapců je v populaci normálně rozdělená a má střední hodnotu $\mu$ a rozptyl $\sigma^2$. Dále budeme předpokládat, že naše veličiny (hmotnosti jednotlivých chlapců) jsou nezávislé (tj. že hmotnosti chlapců se vzájemně nijak neovlivňují), což bývá obvykle triviálně splněno. My chceme otestovat pravdivost tvrzení o střední hmotnosti chlapců v populaci, tedy budeme testovat hypotézu o parametru $\mu$. Matematicky toto zapíšeme jako:

$$\mathrm{H}_0 : \mu = 8.2 \text{ kg}$$

Toto tvrzení se nazývá **nulová hypotéza**. Testovaná hodnota (v našem případě 8.2 kg) se obecně značí jako nějaké $\mu_0$, tedy obecný tvar nulové hypotézy je

$$\mathrm{H}_0 : \mu = \mu_0$$

V případě, že hypotézu $\mathrm{H}_0$ zamítnu, potřebuji se uchýlit k nějakému alternativnímu tvrzení. Tomuto tvrzení se říká **alternativní hypotéza** (zkráceně alternativa), značí se $\mathrm{H}_1$ a může mít jeden z těchto tvarů:

$$\begin{aligned} & \mathrm{H}_1 : \mu \neq 8.2 \text{ kg} \\ \text{nebo} \quad & \mathrm{H}_1 : \mu < 8.2 \text{ kg} \\ \text{nebo} \quad & \mathrm{H}_1 : \mu > 8.2 \text{ kg}. \end{aligned}$$

Vždy je tedy jakýmsi doplňkem nulové hypotézy. Při našem rozhodování ohledně platnosti $\mathrm{H}_0$ můžou nastat následující situace:

| | | skutečnost: | |
|---|---|---|---|
| | | $\mathrm{H}_0$ platí | $\mathrm{H}_0$ neplatí |
| rozhodnutí: | $\mathrm{H}_0$ zamítneme | chyba 1. druhu | OK |
| | $\mathrm{H}_0$ nezamítneme | OK | chyba 2. druhu |

Ráda bych, aby pravděpodobnosti obou druhů chyb byly malé. Avšak minimalizovat obě současně nelze. Udělá se to tedy tak, že velikost pravděpodobnosti chyby 1. druhu se stanoví jako nějaké malé číslo (např. 0.05) a pravděpodobnost chyby 2. druhu už pak prostě „nějak vyjde“, tu už neovlivním. Stanovená pravděpodobnost chyby 1. druhu se tradičně značí $\alpha$ a nazývá se **hladina významnosti testu**, zkráceně *hladina testu* nebo jen *hladina*. Volí se jako nějaké malé číslo, kromě zmíněných 0.05 se ještě používá 0.1 nebo 0.01. Hladina testu má tedy význam pravděpodobnosti, že zamítneme nulovou hypotézu, která ale ve skutečnosti platí (jenom my jsme to bohužel nedokázali poznat, protože jsme měli smůlu na data).

Předpokládejme, že rozptyl $\sigma^2$ je neznámý, což je v praxi nejčastější případ. Pak za platnosti nulové hypotézy (tedy pokud skutečné $\mu$ je opravdu oněch 8.2) platí:

$$\frac{\bar{X} - 8.2}{\frac{S}{\sqrt{n}}} \sim t_{n-1} \qquad \left( \text{obecně: } \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1} \right), \tag{1}$$

kde $S$ značí odhad směrodatné odchylky $\sigma$, tedy výběrovou směrodatnou odchylku. Pro úplnost připomeňme její vzorec

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

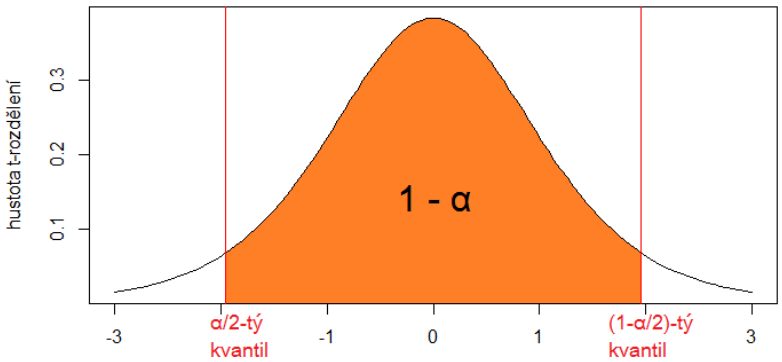
Vztah (1) už byl zmíněn v textu o intervalech spolehlivosti a využijeme ho i zde. Výraz na levé straně (1) si označíme $T$, tedy

$$T = \frac{\bar{X} - 8.2}{\frac{S}{\sqrt{n}}} \qquad \left( \text{obecně: } T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \right)$$

a bude naší **testovou statistikou**, tj. výrazem, pomocí nějž budeme o platnosti $H_0$ rozhodovat. Tato statistika má totiž pro naše účely rozumný tvar - porovnává výběrový průměr s testovanou hodnotou a celé to dělí směrodatnou chybou průměru, tedy zohledňuje jeho přesnost coby odhadu skutečného $\mu$.

Pokusme se nyní konečně odvodit rozhodovací kritérium. Hledejme hodnotu, se kterou bychom hodnotu statistiky $T$ porovnali, a pokud bude $T$ o hodně větší, zamítli bychom nulovou hypotézu. Přitom chceme, aby pravděpodobnost, že se zmýlíme (tj. vypočtená $T$ bude velká, ačkoli $H_0$ ve skutečnosti platí) byla $\alpha$.

Pokud $H_0$ platí, měla by statistika $T$ mít t-rozdělení o $n-1$ stupních volnosti (tak praví výraz (1)). Tedy její konkrétní hodnota (vypočtená z našich dat) by se měla s pravděpodobností (1-$\alpha$) nacházet mezi $\frac{\alpha}{2}$-tým a $\left(1 - \frac{\alpha}{2}\right)$-tým kvantilem tohoto rozdělení (viz obrázek).



Zbylá pravděpodobnost $\alpha$ pak bude odpovídat situaci, kdy $\mathrm{H}_0$ platí, ale my jsme z našich dat dostali hodnotu $T$, která byla příliš velká. Udělali jsme tak chybu 1. druhu, jejíž pravděpodobnost má být $\alpha$, což přesně souhlasí.

Hypotézu $\mathrm{H}_0$ tedy zamítneme, bude-li

$$T \geq \mathrm{qt}_{n-1}\left(1-\frac{\alpha}{2}\right) \quad \text{nebo} \quad T \leq \mathrm{qt}_{n-1}\left(\frac{\alpha}{2}\right),$$

kde $\mathrm{qt}_{n-1}\left(1-\frac{\alpha}{2}\right)$ značí $\left(1-\frac{\alpha}{2}\right)$-tý kvantil t-rozdělení s $n-1$ stupni volnosti. Častěji se používá označení $t_{n-1}\left(1-\frac{\alpha}{2}\right)$. Analogicky pro $\mathrm{qt}_{n-1}\left(\frac{\alpha}{2}\right)$.

t-rozdělení je symetrické kolem nuly, tedy $qt_{n-1}\left(\frac{\alpha}{2}\right) = -\mathrm{qt}_{n-1}\left(1-\frac{\alpha}{2}\right)$. Souhrnně lze tedy naše rozhodovací kritérium zapsat jako:

$$|T| \geq \mathrm{qt}_{n-1}\left(1-\frac{\alpha}{2}\right) \quad \Rightarrow \quad \text{zamítáme } \mathrm{H}_0.$$

Toto rozhodovací kritérium odpovídá situaci, kdy je alternativní hypotéza takzvaně oboustranná, tj. $\mathrm{H}_1$: $\mu \neq \mu_0$.

Pro případ $\mathrm{H}_1$: $\mu < \mu_0$ je rozhodovací kritérium takovéto:

$$T \leq -\mathrm{qt}_{n-1}(1-\alpha) \quad \Rightarrow \quad \text{zamítáme } \mathrm{H}_0$$

a pro případ $\mathrm{H}_1$: $\mu > \mu_0$ zase takovéto

$$T \geq \mathrm{qt}_{n-1}(1-\alpha) \quad \Rightarrow \quad \text{zamítáme } \mathrm{H}_0,$$

přičemž tyto alternativní hypotézy se nazývají jednostranné.

Kvantily t-rozdělení z rozhodovacích kritérií se v kontextu testování hypotéz nazývají **kritické hodnoty** (neboť na nich se láme, zda budeme nulovou hypotézu zamítat či nikoli).

Oblasti hodnot $T$, které způsobí zamítnutí nulové hypotézy, tvoří tzv. **kritický obor**. Pro oboustrannou alternativu je to tedy sjednocení intervalů $(-\infty, \mathrm{qt}_{n-1}\left(\frac{\alpha}{2}\right))$ a $(\mathrm{qt}_{n-1}\left(1-\frac{\alpha}{2}\right), \infty)$. Pro alternativu $\mathrm{H}_1$: $\mu < \mu_0$ je kritickým oborem interval $(-\infty, \mathrm{qt}_{n-1}(\alpha))$, pro alternativu $\mathrm{H}_1$: $\mu > \mu_0$ je to interval $(\mathrm{qt}_{n-1}(1-\alpha), \infty)$.

Tento test se nazývá **jednovýběrový t-test**, protože pracujeme jen s jedním výběrem a protože testová statistika $T$ má t-rozdělení.

V praxi je situace, kdy neznáme rozptyl $\sigma^2$, asi nejčastější. Nicméně, kdybychom ho přeci jen znali, dal by se pro test hypotézy $\mathrm{H}_0 : \mu = \mu_0$ odvodit zcela analogický test. Místo (1) bychom jen vyšly z toho, že za platnosti $\mathrm{H}_0$ platí:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

Testová statistika na levé straně se většinou značí $Z$ (neboť je to vlastně z-skór pro průměr) a výsledný test se označuje jako „z-test“. Konkrétní rozhodovací kritéria jsou uvedena v tabulce se shrnutím.

## Shrnutí

Mějme náhodný výběr $X_1, X_2, \ldots, X_n$ z $\mathrm{N}(\mu, \sigma^2)$ a testujme nulovou hypotézu $\mathrm{H}_0$: $\mu = \mu_0$ na hladině $\alpha$. Pak rozhodovací kritérium má tvar:

| | alternativa | $\mathrm{H}_0$ zamítáme pokud: |
|---|---|---|
| $\sigma^2$ neznámé (t-test) | $\mathrm{H}_1$: $\mu \neq \mu_0$ | $\lvert T\rvert \geq \mathrm{qt}_{n-1}\left(1-\frac{\alpha}{2}\right)$ |
| | $\mathrm{H}_1$: $\mu > \mu_0$ | $T \geq \mathrm{qt}_{n-1}\left(1-\alpha\right)$ |
| | $\mathrm{H}_1$: $\mu < \mu_0$ | $T \leq -\mathrm{qt}_{n-1}\left(1-\alpha\right)$ |
| $\sigma^2$ známé (z-test) | $\mathrm{H}_1$: $\mu \neq \mu_0$ | $\lvert Z\rvert \geq \mathrm{qnorm}\left(1-\frac{\alpha}{2}\right)$ |
| | $\mathrm{H}_1$: $\mu > \mu_0$ | $Z \geq \mathrm{qnorm}\left(1-\alpha\right)$ |
| | $\mathrm{H}_1$: $\mu < \mu_0$ | $Z \leq -\mathrm{qnorm}\left(1-\alpha\right)$ |

„qnorm“ značí opět kvantily normovaného normálního rozdělení, vzorečky pro statistiky $T$ a $Z$ jsou uvedeny v textu výše.

## Důležitá upozornění

### Předpoklady

Před použitím t-testu (popř. z-testu) nezapomeňte nikdy ověřit jeho předpoklady! To jest, nezávislost jednotlivých hodnot ve výběru (tady nelze udělat víc, než se v duchu ujistit, že je tento předpoklad reálný) a normalitu rozdělení sledovaného znaku (použije se QQ-plot nebo Shapiro-Wilkův test)! Za těchto předpokladů byl test odvozen a bez nich neplatí. Testů pro střední hodnotu je ale naštěstí celá řada, takže pokud by například předpoklad o normálním rozdělení nebyl naplněn, lze se obrátit na jiný test (uvidíme později v semestru).

### nezamítnout $\neq$ přijmout

Uvědomte si, že postavení nulové a alternativní hypotézy není symetrické. To, že jsme nulovou hypotézu nezamítli, neznamená, že platí. Možná, že skutečné $\mu$ není 8.2 kg, ale 8.1 kg a my jsme to jen na základě našich dat nedokázali poznat. Pokud tedy v nějaké situaci nezamítnete $\mathrm{H}_0$, musíte to okomentovat opatrnou formulací typu „na základě našich dat nezamítáme nulovou hypotézu, že...“. Ovšem pokud $\mathrm{H}_0$ zamítnete, můžete použít odvážnější tvrzení „na hladině 5 % jsme prokázali, že... (tvrzení alternativní hypotézy)“.

## Jak zvolit alternativní hypotézu?

Z toho, že postavení hypotéz není symetrické, a tedy že zamítnutí nulové hypotézy je silnějším výsledkem než její nezamítnutí, plyne, že v reálném životě se nám více hodí nulové hypotézy zamítat. Proto do nulové hypotézy volíme vždy rovnost a do alternativní hypotézy vždy to, co nás ve skutečnosti zajímá a co bychom rádi prokázali. Pokud jsme tedy vyvinuli nové hnojivo, které vede k vyšším středním výnosům než jaké udává konkurence, zvolíme $\mathrm{H}_0$: $\mu =$ *výnos konkurence* a $\mathrm{H}_1$: $\mu >$ *výnos konkurence*.

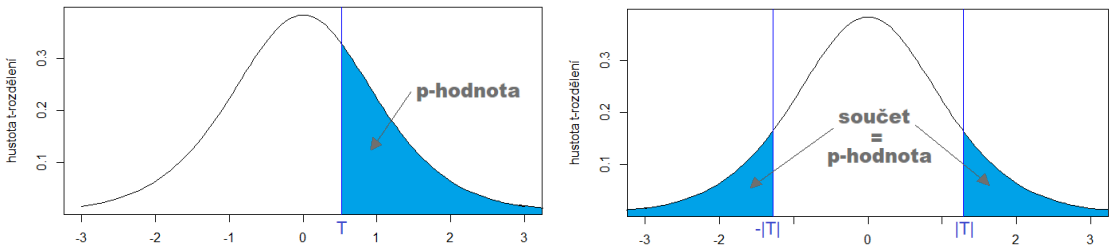
## p-hodnota (angl. p-value)

Protože volba hladiny je věcí každého výzkumníka, bylo nutno vymyslet způsob, jak ve výstupu statistických programů říct, pro které hladiny by došlo k zamítnutí nulové hypotézy a pro které nikoli. K tomuto účelu slouží číslo nazývané „p-hodnota“, které má

význam nejmenší hladiny, na které se $H_0$ zamítá. Nulovou hypotézu tedy zamítáme na každé hladině, která je vyšší nebo rovna p-value. Schématicky:

$$\text{p-value} \leq \alpha \quad \Rightarrow \quad \text{zamítáme } H_0.$$

Jiný způsob jak p-value interpretovat je, že je to pravděpodobnost, že za platnosti $H_0$ dostaneme (při nějakém dalším pokusu) data, která budou ještě „horší“ než ta, co máme. „Horší“ je míněno jako „ještě více svědčící proti $H_0$“. Je to tedy pravděpodobnost, že naše testová statistika nabude (při dalším pokusu) hodnoty stejné nebo vyšší než ta, co máme teď, ačkoli $H_0$ ve skutečnosti platí. Například pro jednovýběrový t-test lze p-hodnotu graficky znázornit jako plochu pod hustotou t-rozdělení, která odpovídá vyšším (tedy „horším“) hodnotám testové statistiky $T$. Na obrázcích je písmenem $T$ označena hodnota $T$ statistiky vypočtená z našich dat. Vlevo je obrázek odpovídající jednostranné alternativě $H_1$: $\mu < \mu_0$, vpravo je obrázek pro oboustrannou alternativu $H_1$: $\mu \neq \mu_0$.




## Síla testu

Doposavad jsme ignorovali chybu 2. druhu, ale i ta má svůj význam. Pravděpodobnost chyby 1. druhu se značí $\alpha$ a nazývá se hladina, pravděpodobnost chyby 2. druhu se značí $\beta$, ale přímo svůj název nemá. Avšak její doplněk do 1, tedy 1-$\beta$, se nazývá **síla testu**. Síla testu je tedy „1−{pravděpodobnost chyby 2. druhu}“, slovy řečeno je to pravděpodobnost, že zamítneme $H_0$, která opravdu neplatí. Jistě bychom byli rádi, kdyby náš test měl velkou sílu, tedy aby dokázal s velkou pravděpodobností zamítat neplatné hypotézy. Sílu si lze představit jako takovou „rozlišovací schopnost“ testu, tedy jak dobře je náš test schopen rozlišovat platné a neplatné nulové hypotézy. Nikoho nepřekvapí, že tato „rozlišovací schopnost“ závisí na rozdílu mezi testovanou hodnotou $\mu_0$ a tou skutečnou (pravdivou) střední hodnotou $\mu$. Samozřejmě, čím větší tento rozdíl je, tím snazší je rozpoznat že $\mu \neq \mu_0$ (tedy že $H_0$ neplatí) a tím větší sílu test v dané situaci má. Zkusme si to znázornit graficky pro náš jednovýběrový t-test.

Mějme tedy výběr z $N(\mu, \sigma^2)$ a testujme hypotézu $H_0$: $\mu = \mu_0$ proti alternativě $H_1$: $\mu > \mu_0$. Naše rozhodovací kritérium má tvar:

$$\text{statistika } T \geq \text{kvantil} \quad \Rightarrow \text{zamítáme } H_0.$$

Doplníme-li tam konkrétní tvar statistiky $T$, dostáváme

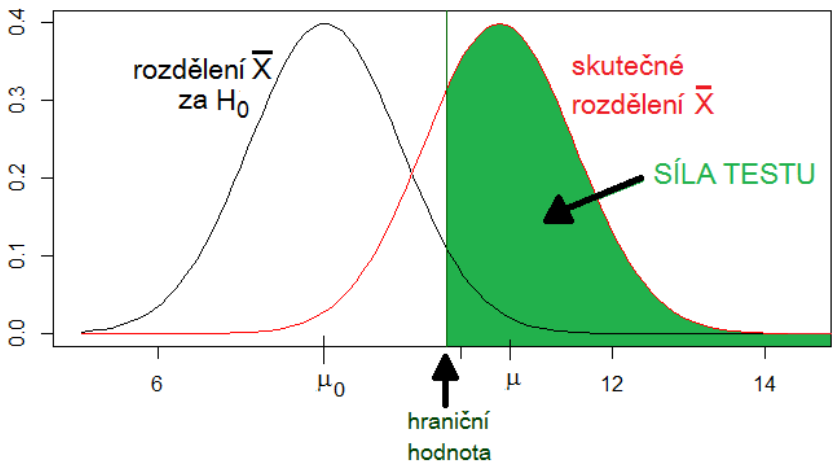
$$\frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}} \geq \text{kvantil} \quad \Rightarrow \text{zamítáme } H_0,$$

což lze dále upravit jako

$$\bar{X} \geq \mu_0 + \frac{S}{\sqrt{n}} \cdot \text{kvantil} \quad \Rightarrow \text{zamítáme } H_0.$$

Hodnota „$\mu_0 + \frac{S}{\sqrt{n}} \cdot$ kvantil“ je jakousi hraniční hodnotou pro průměr, u které se láme rozhodování o tom, zda $H_0$ zamítnout či nikoli.

Na obrázku dole je černě znázorněno rozdělení, jaké by měl mít výběrový průměr $\bar{X}$, pokud nulová hypotéza platí, a červeně je znázorněno jeho skutečné rozdělení (se střední hodnotou $\mu$). V grafu je dále označena zmíněná hraniční hodnota pro průměr. Pravděpodobnost, že zamítnu nulovou hypotézu, když skutečné rozdělení je to červené, je pak rovna ploše pod červenou křivkou vpravo od hraniční hodnoty (viz obrázek níže). Čím vzdálenější je červená křivka od černé (čím větší je rozdíl mezi $\mu$ a $\mu_0$), tím větší ta plocha je a tím větší je tedy síla testu. Hraniční hodnota se samozřejmě odvíjí od zvolené hladiny (skrze kvantil). Odtud je vidět, že čím vyšší hladina, tím více je hraniční hodnota posunuta vlevo, a tím větší je síla. Z toho je patrné, proč nelze současně minimalizovat chybu 1. a 2. druhu.




Vrátíme-li se k představě síly jakožto „rozlišovací schopnosti“, bude nám asi jasné, že při větší směrodatné odchylce $\sigma$ bude těžší od sebe blízké $\mu$ a $\mu_0$ odlišit, a tedy síla testu bude nižší. Naopak s rostoucím počtem pozorování se síla (naše rozlišovací schopnost) zlepšuje. Počet pozorování je také jediný možný způsob, jak lze v praxi sílu testu ovlivnit.

## Tři způsoby, jak rozhodnout o platnosti nulové hypotézy

Na základě výstupu z R (popř. jiného statistického softwaru) lze o platnosti $H_0$ rozhodnout třemi různými způsoby:

1. podívat se na hodnotu testové statistiky $T$ a porovnat ji s příslušným kvantilem t-rozdělení (viz naše rozhodovací kritéria)

2. podívat se na p-hodnotu a porovnat ji se zvolenou hladinou (p-value $\leq \alpha \Rightarrow$ zamítáme $H_0$)

3. spočítat interval spolehlivosti pro $\mu$ (pokud testovaná hodnota $\mu_0$ v tomto intervalu neleží, pak s velkou pravděpodobností nebude tou skutečnou)