# Testování statistických hypotéz

## Statistická hypotéza

Statistická hypotéza je tvrzení, které je možno na základě dat (náhodného výběru) přijmout nebo zamítnout (vyloučit).

Pravidlo, podle kterého hypotézu přijmeme nebo zamítneme se nazývá **test hypotézy**.

**Princip:** Prostřednictvím vhodné výběrové statistiky $T(\vec{X})$ sledujeme odchylku dat od hypotézy a pokud je náhodný vznik tak velké odchylky málo pravděpodobný, hypotézu zamítneme.

### Typické úlohy

- Je měřena hodnota (náhodným výběrem), o které teorie (výrobce, ...) tvrdí, že má hodnotu $x$. Na základě výsledků náhodného výběru lze někdy tuto hypotézu zamítnout, a to pokud by naměřená data byla v případě platnosti hypotézy velmi málo pravděpodobná. Takovou hypotézu nelze nikdy potvrdit. Pokud je hypotéza přijata, znamená to, že data nejsou s hypotézou v rozporu.
- Statistická přejímka zboží - Výrobce garantuje, že v dodávce $N$ výrobků je maximálně $n$ vadných výrobků. Pokud chci dodávku úspěšně reklamovat, musím tuto hypotézu vyloučit. Naopak, pokud by chtěl výrobce prokázat své tvrzení, musel by vyloučit hypotézu "v dodávce je více než $n$ vadných výrobků".
- Jsou 2 metody přípravy vzorků, zajímá nás nějaký parametr $a$. Chtěl bych ukázat, že nová metoda je lepší, tj. vede hodnotě $a_n$ větší než je hodnota $a_s$ pro starou metodu. Hypotézu potvrdíme, pokud dokážeme vyloučit hypotézu $a_n \leq a_s$. Pokud je hypotéza potvrzena, znamená to, rozdíl je **statisticky významný**. Takové tvrzení nic neříká o tom, jak je takový rozdíl velký. Při velkém rozsahu výběru a malém rozptylu dat mohou být i relativně malé změny dat statisticky významné.
- Hypotéza o korelaci dat - data jsou nekorelovaná. Alternativní hypotéza "korelace dat je statisticky významná".

# Testy pro výběry z normálního rozdělení

Tyto testy jsou velmi rozšířené a najdou se ve všech učebnicích a příručkách.

## Hypotézy o střední hodnotě - Studentův t-test

**Hypotéza** - Byl proveden výběr z normálního rozdělení s neznámou disperzí. Hypotéza tvrdí – střední hodnota $EX = \mu$. Máme ji zamítnout nebo přijmout?

Určíme průměr $\overline{x}$, výběrový rozptyl $S^2$ a odtud vypočteme parametr

$$t = \frac{\overline{x} - \mu}{S}\sqrt{n}$$

Pokud hypotéza platí, veličina $t$ má Studentovo rozdělení. Ptám se, jaká je pravděpodobnost, že tak velká nebo větší odchylka mohla vzniknout náhodně, tedy jaká je pravděpodobnost $P(t'; |t'| \geq |t|)$

$$P(t'; |t'| \geq |t|) = \int_{-\infty}^{-|t|} + \int_{|t|}^{\infty} f(t')\ \mathrm{d}t' = 2\frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\Gamma\left(\frac{n-1}{2}\right)} \int_{|t|}^{\infty} \frac{\mathrm{d}t'}{\left(1 + \frac{t^2}{n-1}\right)^{n/2}}$$

Tato pravděpodobnost je tabelována a lze ji vyjádřit pomocí "neúplné $\beta$ funkce"

$$P(t'; |t'| \geq |t|) = I_{\frac{\nu}{\nu+t^2}}\left(\frac{\nu}{2}, \frac{1}{2}\right)$$

kde počet stupňů volnosti je $\nu = n - 1$ a neúplná $\beta$ funkce je

$$I_x(a, b) = \frac{1}{B(a, b)} \int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t$$

kde pro $\forall a, b \;\; I_0(a, b) = 0, \; I_1(a, b) = 1.$

Jestliže $P(t'; |t'| \geq |t|) \leq \alpha$, hypotézu zamítneme na **hladině významnosti** $\alpha$. Hladina významnosti se obvykle volí 5 % nebo 1 %.

**Příklad** Předpokládáme, že výška člověka odpovídá normálnímu rozdělení. Hypotéza tvrdí $EX = \mu = 166$. Náhodně byla změřena výška 10 lidí a získány hodnoty: 160, 160, 167, 170, 173, 176, 178, 178, 181, 181.

*Řešení* Průměr je $\overline{x} = 172.4$, výběrový rozptyl

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{10}(X_i - \overline{x})^2 = 62.93 \qquad \Rightarrow \qquad S = 7.93$$

Výběrová směrodatná odchylka průměru je tedy $S_{\overline{x}} = S/\sqrt{n} = 2.51$. Hodnota $t$ při $\nu = n - 1 = 9$ stupních volnosti je

$$t_9 = \frac{\overline{x} - \mu}{S_{\overline{x}}} = \frac{172.4 - 166}{2.51} = 2.55 \qquad P(|t_9| \geq 2.55) = 0.031$$

Hranice pro 5 % hladinu významnosti je $|t_9| = 2.26$ ($P(|t_9| \geq 2.26) = 0.05$). Zjištěná hodnota $t$ je tedy statisticky významná při 5 % hladině významnosti a hypotézu při 5 % hladině hladině významnosti zamítneme.

Kdybychom zvolili přísnější požadavek na statistickou významnost, např. 1 % hladinu významnosti, hranice významnosti by byla $|t_9| = 3.24$ ($P(|t_9| \geq 3.24) = 0.01$). Zjištěná hodnota $t$ není statisticky významná při této hladině významnosti a hypotézu je při 1 % hladině významnosti nutno přijmout (nelze ji zamítnout).

## Souvislost testování hypotéz s intervalovým odhadem

Jako konfidenční interval hodnot $\mu$ s parametrem spolehlivosti $1 - \alpha$, vezmeme oblast $\mu$ takových, kde hypotézu $EX = \mu$ nelze zamítnout při hladině významnosti $\alpha$.

Aby nedošlo k zamítnutí hypotézy při hladině významnosti $\alpha = 0.05$, musí být

$$-2.26 < t_9 = \frac{\overline{x} - \mu}{S_{\overline{x}}} < 2.26$$

Vyřešíme-li tyto dvě nerovnosti vzhledem k $\mu$, dostáváme 95 % konfidenční interval $\mu \in (166.7, 178.1)$. Často se pak uvádí, že $\mu$ leží s pravděpodobností 0.95 v uvedeném intervalu.

## Srovnání střední hodnoty dvou souborů

Studujeme rozdíly mezi 2 základními soubory $x$ a $y$ reprezentovanými náhodnými výběry $X_i$, $Y_j$. Chceme například prokázat, že soubor $y$ má větší střední hodnotu.
**Alternativní hypotéza** je tvrzení "soubor $Y$ má střední hodnotu menší nebo rovnu střední hodnotě souboru $X$". Pokud alternativní hypotézu zamítneme, prokážeme platnost původní hypotézy.

Předpokládáme nekorelované hodnoty $X_i$, $Y_i$, předpokládáme, že výběrové rozptyly se podstatně neliší a tedy $\sigma_x = \sigma_y$. Pak výběrový rozptyl je

$$S^2 = \frac{\sum_{i=1}^{n_x} (X_i - \overline{x})^2 + \sum_{i=1}^{n_y} (Y_i - \overline{y})^2}{n_x + n_y - 2}$$

Označme $R = \overline{y} - \overline{x}$, pak

$$S_{\overline{x}}^2 = \frac{S^2}{n_x} \qquad S_{\overline{y}}^2 = \frac{S^2}{n_y} \qquad \Rightarrow \qquad S_R^2 = S_{\overline{y}}^2 + S_{\overline{x}}^2 = S^2 \left( \frac{1}{n_x} + \frac{1}{n_y} \right)$$

Pak

$$t_{n_x+n_y-2} = \frac{\overline{y} - \overline{x}}{S_R}$$

Pokud $P(t' \geq t_{n_x+n_y-2}) \leq \alpha$, alternativní hypotézu zamítneme při hladině významnosti $\alpha$.

*Pozn.* Zda se rozptyly $S_X^2$ a $S_Y^2$ podstatně neliší, lze určit testem hypotézy $S_X^2 = S_Y^2$. Orientačně – pokud platí $S_X^2 < 2S_Y^2 \wedge S_Y^2 < 2S_X^2$ pro $n_x, n_y \simeq 10$, pak se rozptyly $x$ a $y$ podstatně neliší. Existuje modifikovaný postup testu, pokud se rozptyly podstatně liší.

*Pozn.* Pokud jsou hodnoty výběrů korelované ($X_i$, $Y_i$ jsou naměřeny na i-tém vzorku), označme rozsah výběru $n$ a $R = \overline{y} - \overline{x}$. Pak

$$S_R^2 = \frac{S_X^2 + S_Y^2 - 2S_{XY}}{n} \qquad\qquad t_{n-1} = \frac{\overline{y} - \overline{x}}{S_R}$$

kde $S_{XY}$ je výběrová kovariance a $(n-1)$ je počet stupňů volnosti.

# Teorie testování hypotéz

2 prvky v rozhodovacím procesu - hypotézu $H$ přijmout, hypotézu $H$ zamítnout

2 typy chyb – **Chyba I.druhu** - zamítnout správnou hypotézu
**Chyba II.druhu** - přijmout nesprávnou hypotézu

**Alternativní hypotéza** $A$ - tvrzení, které platí právě tehdy, pokud $H$ neplatí. Pokud $A$ zamítneme, potvrdíme platnost $H$. To je silnější než tvrzení, že $H$ nelze zamítnout.

**Test hypotézy** – Pravidlo, podle kterého rozhodujeme o přijetí, resp. zamítnutí hypotézy. Je určeno **testovou statistikou** $T(\vec{X})$ a **kritickým oborem** $\mathcal{T}$ – množinou hodnot z $R$ takových, že hypotézu zamítneme právě tehdy, když $T(\vec{X}) \in \mathcal{T}$.

**Hladina významnosti** Pravděpodobnost chyby I. druhu, tj. pravděpodobnost zamítnutí správné hypotézy.

2 druhy hypotéz – **oboustranná** – hypotéza o rovnosti - sčítáme pravděpodobnosti velkých odchylek na obě strany
**jednostranná** – hypotéza o nerovnosti - pravděpodobnost velké odchylky na jednu stranu

2 druhy hypotéz – **parametrické** - známý typ rozdělení s neznámými parametry $\vec{\theta}$, hypotéza se týká některého parametru $\theta_h$ nebo funkce $g(\vec{\theta})$
**neparametrické** - obecnější, nemusí předpokládat typ rozdělení, často nepoužívá přímo sledovaného znaku

Parametrická hypotéza – prostor $\Omega$ parametrů $\vec{\theta}$ se dělí na 2 podprostory
$\omega$ – platí $H$, $(\Omega - \omega)$ – platí $A$
Testy parametrických hypotéz lze konstruovat nejen pro normální rozdělení.

**Silofunkce testu** - Funkce přiřazující $\forall \vec{\theta} \in \Omega$ podmíněnou pravděpodobnost $P[T(\vec{X}) \in \mathcal{T} | \vec{\theta}]$

# Test hypotézy o mediánu - znaménkový test

**Hypotéza** $MeX = x_{0.5}$ - oboustranná neparametrická hypotéza
Ve výběru o rozsahu $n$ je $m$ hodnot $X_i \neq x_{0.5}$, $n_1$ hodnot menších než $x_{0.5}$ a $n_2 = m - n_1$ větších než $x_{0.5}$. Nechť např. $n_2 > m/2$. Pokud hypotéza platí, je pravděpodobnost, že počet hodnot $X_i$ větších než $x_{0.5}$ bude $n_2' \geq n_2$ je

$$P(n_2' \geq n_2) = \sum_{k=n_2}^{m} \binom{m}{k} \left(\frac{1}{2}\right)^m$$

Hypotézu vyloučíme, pokud pravděpodobnost tak velké odchylky bude $\leq \alpha$

$$\begin{aligned} P\left(|n_2' - m/2| \geq n_2 - m/2\right) &= P(n_2' \geq n_2) + P(n_2' \leq m - n_2) = \\ = 2P(n_2' \geq n_2) &= 2\sum_{k=n_2}^{m} \binom{m}{k} \left(\frac{1}{2}\right)^m \leq \alpha \end{aligned}$$

## Další použití znaménkového testu

Nechť je sledována dvojice znaků $(x, y)$. Hypotéza tvrdí, že u většiny prvků základního souboru je $y > x$. Alternativní hypotéza "u většiny prvků je $y \leq x$". Jde o jednostrannou neparametrickou hypotézu, stačí testovat alternativní hypotézu pro případ rovnosti $P(y \leq x) = P(y = x)$. Zamítneme-li rovnost, tím spíše bychom zamítli jakýkoliv zápornou hodnotu rozdílu $y - x$. Z náhodného výběru o rozsahu $n$ je pro $n_+$ prvků $Z_i = Y_i - X_i > 0$. Pokud

$$P(n' \geq n_+) = \sum_{k=n_+}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^n \leq \alpha$$

pak alternativní hypotézu zamítneme a prokážeme tím původní hypotézu.

## Hypotézy o statistickém rozdělení – testy dobré shody

Odpovídá pravděpodobnostní rozdělení základního souboru teorii? Mají dva soubory stejná rozdělení pravděpodobnosti?
(2 typy rozdělení – diskrétní a spojitá)

**Diskrétní rozdělení** – $\chi^2$ test
I-tou hodnotu má $n_i$ prvků výběrového souboru o rozsahu n, podle teorie má být $m_i = n\ p_i$ ($m_i$ nemusí být celé). Počet prvků výběru, které mají i-tou hodnotu, má binomické rozdělení $Bi(n, p_i)$ se střední hodnotou $\mu_i = m_i$ a s rozptylem $\sigma_i^2 = m_i$. Pokud $m_i \geq 5$, lze binomické rozdělení nahradit normálním rozdělením (buňky s $m_i < 5$ sdružíme do skupin). Pak

$$h = \sum_{i=1}^{k} \frac{(n_i - m_i)^2}{m_i}$$

má přibližně rozdělení

$$h \sim \chi^2_\nu \qquad \nu = k - 1 - p$$

kde $k$ je počet hodnot (skupin), $p$ je počet parametrů teoretického rozdělení získaných z dat, počet stupňů volnosti je snížen o 1 vzhledem k podmínce $\sum_{i=1}^{k} n_i = n$. Pravděpodobnost, že náhodou vznikne $h' \geq h$ je

$$P(h' \geq h) = \int_h^\infty \chi^2_\nu(h')\ \mathrm{d}h' = Q\left(\frac{\nu}{2}, \frac{h}{2}\right)$$

Pokud $P(h' \geq h) \leq \alpha$, hypotézu o shodě rozdělení s teorií zamítneme.

**Spojitá rozdělení** – Kolmogorov-Smirnovův test
Nechť hypotéza určuje rozdělení včetně jeho parametrů, nechť $F(x)$ je jeho distribuční funkce, $\tilde{F}_n(x)$ je empirická distribuční funkce výběrového souboru. Pak test je založen na statistice

$$D = \sup_{-\infty < x < \infty} \left|\tilde{F}_n(x) - F(x)\right|$$

Hypotézu zamítneme na hladině významnosti $\alpha$, pokud

$$D \geq D_n(1-\alpha) \simeq \sqrt{-\frac{1}{2n} \ln\left(\frac{\alpha}{2}\right)}$$

kde přibližná rovnost platí pro $n > 100$. Pro shodu 2 empirických rozdělení výraz $n_1 n_2/(n_1 + n_2)$ nahradí $n$.

# Testy korelace a nezávislosti

## Test nekorelovanosti (nezávislosti) pro normální rozdělení

Pro parametrický test nekorelovanosti lze využít lineární korelační koeficient $r_{XY}$, ale častěji se využívá veličina

$$T = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}}$$

Pokud $(X_i, Y_i)$ je výběr z dvourozměrného normálního rozdělení s korelačním koeficientem $\rho_{XY} = 0$, pak veličina $T$ má Studentovo $t$ rozdělení o $(n-2)$ stupních volnosti. Pokud $P(|t'| \geq |T|) \leq \alpha$, hypotéza o nekorelovanosti je zamítnuta na hladině významnosti $\alpha$.

## Neparametrický test nezávislosti

Test založený na pořadí – pro libovolné sdružené rozdělení $(X, Y)$ – založen na Spearmanově korelačním koeficientu $r_{XY}^S$.

Hypotézu o nezávislosti zamítneme na hladině významnosti $\alpha$, pokud

$$|r_{XY}^S| \geq k_n(\alpha)$$

kde kritické hodnoty $k_n(\alpha)$ jsou tabelovány pro malé rozsahy výběru ($n \leq 30$) a pro větší $n$ lze aproximovat $k_n(\alpha) \simeq \Phi^{-1}(1-\alpha/2)/\sqrt{n-1}$, kde $\Phi$ je distribuční funkce rozdělení $N(0,1)$ a její inverzní funkce $\Phi^{-1}$ je kvantilová funkce.

## Test nezávislosti dat v kontingenční tabulce

Prostý náhodný výběr o rozsahu $n$, náhodný vektor $(X, Y)$, $X$ může nabývat hodnot $(A_1, A_2, \ldots, A_K)$, $Y$ hodnot $(B_1, B_2, \ldots, B_L)$. Označme $n_{ij}$ četnost pozorování $(A_i, B_j)$, $n_{i\cdot}$ marginální četnost $A_i$, $n_{\cdot j}$ marginální četnost $B_j$.

Četnosti lze reprezentovat pomocí **kontingenční tabulky**

| třída | $B_1$ | $B_2$ | $\ldots$ | $B_l$ | marg. $X$ |
|---|---|---|---|---|---|
| $A_1$ | $n_{11}$ | $n_{12}$ | $\ldots$ | $n_{1L}$ | $n_{1\cdot}$ |
| $A_2$ | $n_{21}$ | $n_{22}$ | $\ldots$ | $n_{2L}$ | $n_{2\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ldots$ | $\vdots$ | $\vdots$ |
| $A_K$ | $n_{K1}$ | $n_{K2}$ | $\ldots$ | $n_{KL}$ | $n_{K\cdot}$ |
| marg. $Y$ | $n_{\cdot 1}$ | $n_{\cdot 2}$ | $\ldots$ | $n_{\cdot L}$ | $n$ |

Označme $m_{ij} = n_{i\cdot} n_{\cdot j}/n$. Test nezávislosti $X$ a $Y$ je založen na statistice

$$\chi^2_\nu = \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \qquad \nu = (K-1)(L-1)$$

Za předpokladu nezávislosti má tato statistika $\chi^2$ rozdělení. Pokud $P(\tilde{\chi}^2_\nu \geq \chi^2_\nu) \leq \alpha$, hypotézu o nezávislosti zamítneme.

Zamítnutí hypotézy o nezávislosti znamená, že závislost je **statisticky významná** (statisticky prokazatelná), nic však neříká o síle závislosti.

Sílu závislosti charakterizuje Cramerovo $V$

$$V = \sqrt{\frac{\chi^2}{n \min(K-1, L-1)}} \qquad 0 \leq V \leq 1$$

nebo míry asociace založené na entropii.