Marie Turčičová

Dvouvýběrové testy

Představme si jednoduchou situaci, kdy máme dva výběry (ze dvou populací), např. se může jednat o hmotnosti kojenců v kilogramech:

1.výběr ($n_x = 5$)				
x_1	x_2	x_3	x_4	x_5
8.0	8.2	8.4	8.9	9.6

2.výběr ($n_y = 6$)					
y_1	y_2	y_3	y_4	y_5	y_6
7.9	7.2	9.1	8.7	7.8	8.1

(Jak jste si asi všimli, výběry nemusejí mít stejný počet prvků, tj. nevadí, když $n_x \neq n_y$.)

Naším cílem bude prozkoumat rovnost středních hodnot hmotností v obou populacích. Pojďme ty dvě populace nazývat *populace* $\mathcal{X}$ a *populace* $\mathcal{Y}$ a příslušné střední hodnoty hmotností označme μ_x a μ_y. Chceme testovat hypotézu o rovnosti těchto populačních průměrů[1]:

$$H_0 : \mu_x = \mu_y$$
$$H_1 : \mu_x \neq \mu_y \qquad (\text{nebo } H_1 : \mu_x < \mu_y, \text{ nebo } H_1 : \mu_x > \mu_y).$$

Kdybychom nevěděli nic o testování hypotéz, jak bychom si intuitivně udělali názor na platnost H_0? Asi bychom porovnali výběrové průměry $\bar{x} = 8.62$ kg a $\bar{y} = 8.13$ kg, a pokud by byly od sebe nějak hodně daleko, shodu populačních průměrů v obou populacích bychom zamítli. To je taky přesně myšlenka dvouvýběrového t-testu :o)

Dvouvýběrový t-test

Zapišme si ještě jednou formálně celou situaci:

- máme dva výběry $X_1, X_2, \ldots, X_{n_x}$ (výběr z populace $\mathcal{X}$) a $Y_1, Y_2, \ldots, Y_{n_y}$ (výběr z populace $\mathcal{Y}$) a předpokládáme, že jsou nezávislé (tj. žádný prvek z $X_1, \ldots, X_{n_x}$ nemá nic společného s žádným prvkem $Y_1, \ldots, Y_{n_y}$)
- dále předpokládáme, že oba výběry pocházejí z normálního rozdělení[2], tj. $X_i \sim N(\mu_x, \sigma^2)$, pro všechna $i = 1, \ldots, n_x$ a $Y_j \sim N(\mu_y, \sigma^2)$ pro všechna $j = 1, \ldots, n_y$ (střední hodnoty obou rozdělení jsou samozřejmě ty populační střední hmotnosti)

Sami jsme uhodli, že testovou statistiku je dobré založit na rozdílu výběrových průměrů $\bar{X}$ a $\bar{Y}$, ale bude potřeba zohlednit i směrodatnou chybu (značeno *s.e.* = *standard error*) tohoto rozdílu. Testová statistika dvouvýběrového t-testu má tedy tvar:

$$T = \frac{\bar{X} - \bar{Y}}{s.e.(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y}}{s} \sqrt{\frac{n_x n_y}{n_x + n_y}}, \tag{1}$$

kde $s = \sqrt{s^2}$, přičemž $s^2 = \frac{n_x - 1}{n_x + n_y - 2} s_x^2 + \frac{n_y - 1}{n_x + n_y - 2} s_y^2$ je jakýsi společný odhad rozptylu (je to vážený průměr výběrových rozptylů z obou výběrů). Pokud je H_0 pravdivá, tj. skutečně jsou střední hodnoty shodné ($\mu_x = \mu_y$), tak rozdíl $\bar{X} - \bar{Y}$ bude malý a statistika T ze vzorce (1) bude také někde blízko u nuly. Formálně je toto pozorování vyjádřené tím, že statistika T má t-rozdělení (které, jak víme, je silně koncentrované kolem nuly). Konkrétně jde o t-rozdělení s $n_x + n_y - 2$ stupni volnosti, tj.

$$T = \frac{\bar{X} - \bar{Y}}{s} \sqrt{\frac{n_x n_y}{n_x + n_y}} \overset{\text{za } H_0}{\sim} t_{n_x + n_y - 2}. \tag{2}$$

[1] Připomínám, že *střední hodnota = populační průměr*.

[2] Abychom měli test na čem založit...

Kritické obory (tj. hodnoty T při kterých H_0 zamítneme) se samozřejmě odvíjejí od tvaru alternativní hypotézy. Rozhodovací kritéria mají tvar:

- Pro $H_1 : \mu_x \neq \mu_y$:

$$\boxed{\text{Je-li } |T| \geq qt_{n_x+n_y-2}(1-\tfrac{\alpha}{2}), \text{ pak zamítáme } H_0.}$$

- Pro $H_1 : \mu_x < \mu_y$:

$$\boxed{\text{Je-li } T \leq -qt_{n_x+n_y-2}(1-\alpha), \text{ pak zamítáme } H_0.}$$

- Pro $H_1 : \mu_x > \mu_y$:

$$\boxed{\text{Je-li } T \geq qt_{n_x+n_y-2}(1-\alpha), \text{ pak zamítáme } H_0.}$$

Poznámka 1 *Pokud bychom nechtěli testovat rovnost středních hodnot, ale to, že se liší o nějakou konstantu c, to jest*

$$H_0 : \mu_x - \mu_y = c$$
$$H_1 : \mu_x - \mu_y \neq c,$$

tak testová statistika má tvar

$$T = \frac{\bar{X} - \bar{Y} - c}{s} \sqrt{\frac{n_x n_y}{n_x + n_y}} \tag{3}$$

a rozhodovací kritérium opět zní:

$$\boxed{\textit{Je-li } |T| \geq qt_{n_x+n_y-2}(1-\tfrac{\alpha}{2}), \textit{ pak zamítáme } H_0.}$$

Pro jednostranné alternativní hypotézy je to analogicky.

Ověřování předpokladů dvouvýběrového t-testu

Jak jste si asi všimli z modrého textu v úvodu, t-test je založen na několika předpokladech. Pojďme se na ně nyní detailněji podívat a ukažme si možnosti jejich ověření.

1. Nezávislost

Potřebujeme, aby hodnoty v prvním výběru byly vzájemně nezávislé, stejně tak hodnoty ve 2. výběru a navíc ještě oba výběry navzájem. Ověřovat to nelze, to musíme mít zajištěno z designu experimentu.

2. Normalita

Potřebujeme, aby oba výběry pocházely z normálního rozdělení (tj. předpokládáme, že rozdělení zkoumaného znaku v obou populacích je normální). K ověřování normality již máme z dřívějška dva nástroje:

- normální diagram (QQ plot)
- Shapirův-Wilkův test

Zvolenou metodu pak aplikujeme na oba výběry zvlášť! (Testovat to najednou nelze, protože normální rozdělení má v obou případech odlišné parametry).

3. Shoda rozptylů

Všimněte si, že se předpokládá $X_1, \ldots, X_{n_x} \sim N(\mu_x, \sigma^2)$ a $Y_1, \ldots, Y_{n_y} \sim N(\mu_y, \sigma^2)$, tj. rozptyl σ^2 je v obou případech stejný! To je zcela nový předpoklad, se kterým jsme se zatím nesetkali. Pro jeho ověřování se nabízejí dva nástroje:

- boxplot - Podívat se na krabicový diagram u obou výběrů a porovnat, zda jsou krabičky stejně vysoké (tj. zda jsou podobná mezikvartilová rozpětí). Je to opět ale pouze „okometrická“ metoda, jejíž závěr je subjektivní.
- F**-test shody rozptylů**: to je rigorózní způsob, jak shodu rozptylů ověřit. Příslušné hypotézy mají tvar:

$$H_0 : \frac{\sigma_x^2}{\sigma_y^2} = 1 \qquad\qquad H_1 : \frac{\sigma_x^2}{\sigma_y^2} \neq 1,$$

kde σ_x^2 značí rozptyl zkoumaného znaku v populaci $\mathcal{X}$ a σ_y^2 v populaci $\mathcal{Y}$. Testová statistika má (překvapivě ☺) tvar podílu výběrových rozptylů. Je-li nulová hypotéza správná, má tato statistika F-rozdělení[3] (odtud název testu). Formálně to lze zapsat jako

$$F = \frac{s_x^2}{s_y^2} \overset{\text{za } H_0}{\sim} F_{n_x-1, n_y-1}. \tag{4}$$

Nulovou hypotézu o rovnosti populačních rozptylů zamítnu, pokud bude testová statistika F příliš vzdálená od jedničky, tj. příliš malá, nebo příliš velká. Rozhodovací kritérium zní:

$$\boxed{H_0 \text{ zamítnu, pokud } F \geq qF_{n_x-1,n_y-1}(1-\tfrac{\alpha}{2}), \text{ anebo } F \leq qF_{n_x-1,n_y-1}(\tfrac{\alpha}{2}) = \frac{1}{qF_{n_x-1,n_y-1}(1-\frac{\alpha}{2})},}$$

kde jako qF označuji kvantil F-rozdělení. Nedojde-li k zamítnutí nulové hypotézy, tak můžeme říct, že shodu rozptylů lze předpokládat (není to tedy tak, že bychom ji prokázali, pouze se nám ji nepodařilo zamítnout).

Co dělat, když nejsou splněny některé předpoklady

1. Nezávislost

S tím se nedá dělat nic ☹. Je to důsledek špatného designu experimentu.

2. Shoda rozptylů

Použije se modifikace dvouvýběrového t-testu pro nestejné rozptyly, tzv. **Welchův test**. Detailněji si ho zde ale ukazovat nebudeme. Dalo by se říci, že předpoklad shody rozptylů je tak trochu zbytečný, protože v případě, kdy shoda rozptylů platí, má Welchův test jen nepatrně menší sílu než dvouvýběrový t-test popsaný výše (který shodu rozptylů předpokládá), takže ho lze používat automaticky všude a shodou rozptylů vůbec nezkoumat.

3. Normalita

V případě, kdy není splněn předpoklad normálního rozdělení (tj. alespoň u jednoho z výběrů zamítneme normalitu), musíme se uchýlit k nějakému z *pořadových testů*. Pořadové se tyto testy jmenují proto, že místo původních hodnot pracují pouze s jejich pořadími. Takovou „pořadovou“ alternativou k dvouvýběrovému t-testu je **dvouvýběrový Wilcoxonův test**, popř. jeho modifikace zvaná Mannův-Whitneyův.

[3] Celým názvem je to Fisherovo-Snedecorovo rozdělení. Je to další ze spojitých rozdělení, na tvar jeho hustoty se můžete podívat na internetu (`https://en.wikipedia.org/wiki/F-distribution`). Toto rozdělení má dva parametry, oba se nazývají stupně volnosti a píší se do dolního indexu, např. $F_{n,m}$ je F-rozdělení se stupni volnosti n a m.

Dvouvýběrový Wilcoxonův test

Tento test předpokládá, že máme dva výběry $X_1, \dots, X_{n_x}$ a $Y_1, \dots, Y_{n_y}$ z nějakého neznámého rozdělení. Testují se zde proti sobě hypotézy, které vypadají malinko jinak než u dvouvýběrového t-testu:

H_0 : rozdělení znaku v populaci $\mathcal{X}$ je shodné s rozdělením v populaci $\mathcal{Y}$ (a tudíž i populační mediány musejí být stejné)

H_1 : rozdělení znaku v populaci $\mathcal{X}$ se liší od rozdělení v populaci $\mathcal{Y}$

To je pro pořadové testy typické. Místo o střední hodnotě se ve formulaci hypotéz většinou mluví o celém rozdělení daného znaku, popř. o jeho mediánu. Místo rovnosti populačních průměrů tak vlastně zkoumáme shodu celých rozdělení, která nám pak implikuje shodu populačních mediánů (ty jakoby suplují populační průměry).

Jak již bylo řečeno, pořadový test pracuje s pořadími hodnot ve výběru. Všem hodnotám tedy přiřadíme pořadí v rámci sdruženého výběru (= výběr, který vznikne spojením obou výběrů do jednoho). Pro náš úvodní příklad s hmotnostmi by to vypadalo následovně:

	x_1	x_2	x_3	x_4	x_5	y_1	y_2	y_3	y_4	y_5	y_6
hodnoty	8.0	8.2	8.4	8.9	9.6	7.9	7.2	9.1	8.7	7.8	8.1
pořadí:	4	6	7	9	11	3	1	10	8	2	5

Pokud nulová hypotéza platí, tj. daný znak má v obou populacích stejné rozdělení, měly by být výběry v rámci toho sdruženého výběru dobře promíchány. Tj. u obou výběrů by se v tabulce měla objevit jak nízká, tak vysoká pořadí.

Klíčovým prvkem testové statistiky je součet pořadí od výběru $Y_1, \dots, Y_{n_y}$[4]. Tento součet pořadí si označme $W_y = \sum_{i=1}^{n_y} R(Y_i)$, kde $R(Y_i)$ značí pořadí Y_i. V příkladu s hmotnostmi by vyšlo $W_y = 29$. Jsou-li výběry „dobře promíchány“, nemělo by být číslo W_y příliš velké, ani příliš malé. Přesné rozdělení veličiny W_y, na němž by bylo možné založit test, není známo, proto se pro ni používá tzv. normální aproximace: od hodnoty W_y se odečte její střední hodnota $\mathrm{E}\,W_y$ a rozdíl se podělí směrodatnou odchylkou $\sqrt{\mathrm{var}(W_y)}$ (což jsou obě hodnoty, které lze určit z rozsahů výběrů n_x a n_y). Výsledná testová statistika pak má asymptoticky[5] (tj. pro velká n_x a n_y) normované normální rozdělení $N(0,1)$. Formálně to zapíšeme jako

$$Z = \frac{W_y - \mathrm{E}\,W_y}{\sqrt{\mathrm{var}\,W_y}} = \frac{W_y - \frac{1}{2}n_y(n_x + n_y + 1)}{\sqrt{n_x n_y (n_x + n_y + 1)\frac{1}{12}}} \overset{\text{za } H_0}{\underset{as.}{\sim}} N(0,1). \tag{5}$$

Jak jsme již vypozorovali, nulovou hypotézu o shodě rozdělení zamítneme, bude-li W_y (a tudíž i celá statistika Z) příliš malá nebo příliš velká, což lze kvantifikovat pomocí příslušných kvantilů normálního rozdělení.

$$\boxed{H_0 \text{ tedy zamítneme, pokud } |Z| \geq qnorm\left(1 - \tfrac{\alpha}{2}\right).}$$

Poznámka 2 *S Wilcoxonovým testem je ekvivalentní Mannův-Whitneyův test, který místo W_y používá veličinu*

$$U_y = W_y - \frac{n_y(n_y + 1)}{2}, \tag{6}$$

která vyjadřuje počet dvojic (X_i, Y_i) pro něž je $X_i < Y_i$. Test je pak opět založen na normální aproximaci pro U_y (přičemž $\mathrm{E}\,U_y = \frac{1}{2}n_x n_y$ a $\mathrm{var}\,U_y = \mathrm{var}\,W_y$), která vede na stejnou statistiku Z, jako je ve vzorci (5).

Poznámka 3 *Tento test je citlivý (má dobrou schopnost správně rozhodnout) zejména v situaci, liší-li se rozdělení znaku v populaci $\mathcal{X}$ a $\mathcal{Y}$ pouze o posunutí. V řeči distribučních funkcí bychom to zapsali jako $F_x(z) = F_y(z - a)$, pro nějaké $a \neq 0$. Méně citlivý je v případě, kdy se rozdělení liší spíše variabilitou.*

Poznámka 4 *O platnosti předpokladů a volbě statistického postupu by se správně mělo rozhodovat ještě před tím, než jsou známa data. Předpoklady bychom měli formulovat např. na základě historických dat a podobně.*

[4] Stejně tak by se dalo vyjít ze součtu pořadí od výběru $X_1, \dots, X_{n_x}$, je to jedno.

[5] Zkratka *as.*