Testy pro porovnání vlastností dvou skupin

Petr Pošík

Části dokumentu jsou převzaty (i doslovně)
z *Mirko Navara: Pravděpodobnost a matematická statistika*,
`https://cw.felk.cvut.cz/lib/exe/fetch.php/courses/a6m33ssl/pms_print.pdf`
s laskavým svolením autora.

Porovnání dvou skupin

Víte, jaký je rozdíl mezi socialismem a kapitalismem?

V socialismu jeden člověk využívá druhého.
V kapitalismu je to přesně naopak.

:-)

Recept: Test rozptylů dvou normálních rozdělení

Předpoklad: Máme 2 *nezávislé* výběry

$(X_1, \dots, X_m)$ z rozdělení $\mathrm{N}(\mathrm{E}\,X, \mathrm{D}\,X)$ a
$(Y_1, \dots, Y_n)$ z rozdělení $\mathrm{N}(\mathrm{E}\,Y, \mathrm{D}\,Y)$.

Je-li $\mathrm{D}\,X = \mathrm{D}\,Y$, pak také $S_X^2 \doteq S_Y^2$.
Realizaci testové statistiky

$$t = \frac{s_X^2}{s_Y^2}$$

porovnáme s kvantily *Fisherova rozdělení* $\mathrm{F}(m-1, n-1)$:

H_0	H_A	H_0 zamítáme, když	dosažená významnost P
$\mathrm{D}\,X \le \mathrm{D}\,Y$	$\mathrm{D}\,X > \mathrm{D}\,Y$	$t > q_{\mathrm{F}(m-1,n-1)}(1-\alpha)$	$1 - F_{\mathrm{F}(m-1,n-1)}(t)$
$\mathrm{D}\,X \ge \mathrm{D}\,Y$	$\mathrm{D}\,X < \mathrm{D}\,X$	$t < q_{\mathrm{F}(m-1,n-1)}(\alpha)$	$F_{\mathrm{F}(m-1,n-1)}(t)$
$\mathrm{D}\,X = \mathrm{D}\,Y$	$\mathrm{D}\,X \neq \mathrm{D}\,Y$	$t > q_{\mathrm{F}(m-1,n-1)}(1-\frac{\alpha}{2})$ nebo $t < q_{\mathrm{F}(m-1,n-1)}(\frac{\alpha}{2})$	$2\min(F_{\mathrm{F}(m-1,n-1)}(t),$ $1 - F_{\mathrm{F}(m-1,n-1)}(t))$

F-rozdělení

Fisherovo-Snedecorovo rozdělení $\mathrm{F}(\xi, \eta)$ s ξ a η stupni volnosti je rozdělení náhodné veličiny

$$F = \frac{\frac{U}{\xi}}{\frac{V}{\eta}},$$

kde U a V jsou *nezávislé* náhodné veličiny s rozdělením $\chi^2(\xi)$, resp. $\chi^2(\eta)$.

V našem případě, je-li $\mathrm{D}\,X = \mathrm{D}\,Y = \sigma^2$, pak

$$\begin{aligned}
U &:= \frac{(m-1)S_X^2}{\sigma^2} \text{ má rozdělení } \chi^2(m-1),\\
V &:= \frac{(n-1)S_Y^2}{\sigma^2} \text{ má rozdělení } \chi^2(n-1),\\
\xi &:= m-1,\\
\eta &:= n-1,\\
F &= \frac{\frac{U}{\xi}}{\frac{V}{\eta}} = \frac{\frac{(m-1)S_X^2}{(m-1)\sigma^2}}{\frac{(n-1)S_Y^2}{(n-1)\sigma^2}} = \frac{S_X^2}{S_Y^2} = T,
\end{aligned}$$

kde T je testová statistika testu z předchozího slidu.

Test rozptylů: praktické poznámky

Pro každou hladinu významnosti potřebujeme 2D tabulku kvantilů indexovanou ξ a η. Obvykle je tabelována jen polovina, druhou je třeba dopočítat podle vzorce

$$q_{\mathrm{F}(\xi,\eta)}(\alpha) = \frac{1}{q_{\mathrm{F}(\eta,\xi)}(1-\alpha)}.$$

POZOR na opačné pořadí indexů!

V praxi se často používá alternativní postup:

1. Pro $s_x^2 \geq s_y^2$ testujeme $t = \frac{s_x^2}{s_y^2} \geq 1$ na rozdělení $\mathrm{F}(m-1, n-1)$:

H_0	H_A	H_0 zamítáme, když	dosažená významnost P
$\mathrm{D}\,X \leq \mathrm{D}\,Y$	$\mathrm{D}\,X > \mathrm{D}\,Y$	$t > q_{\mathrm{F}(m-1,n-1)}(1-\alpha)$	$1 - F_{\mathrm{F}(m-1,n-1)}(t)$
$\mathrm{D}\,X \geq \mathrm{D}\,Y$	$\mathrm{D}\,X < \mathrm{D}\,X$	nezamítáme	žádná
$\mathrm{D}\,X = \mathrm{D}\,Y$	$\mathrm{D}\,X \neq \mathrm{D}\,Y$	$t > q_{\mathrm{F}(m-1,n-1)}(1-\frac{\alpha}{2})$	$2(1 - F_{\mathrm{F}(m-1,n-1)}(t))$

2. Pro $s_x^2 \leq s_y^2$ testujeme $t = \frac{s_y^2}{s_x^2} \geq 1$ na rozdělení $\mathrm{F}(n-1, m-1)$:

H_0	H_A	H_0 zamítáme, když	dosažená významnost P
$\mathrm{D}\,X \leq \mathrm{D}\,Y$	$\mathrm{D}\,X > \mathrm{D}\,Y$	nezamítáme	žádná
$\mathrm{D}\,X \geq \mathrm{D}\,Y$	$\mathrm{D}\,X < \mathrm{D}\,X$	$t > q_{\mathrm{F}(n-1,m-1)}(1-\alpha)$	$1 - F_{\mathrm{F}(n-1,m-1)}(t)$
$\mathrm{D}\,X = \mathrm{D}\,Y$	$\mathrm{D}\,X \neq \mathrm{D}\,Y$	$t > q_{\mathrm{F}(n-1,m-1)}(1-\frac{\alpha}{2})$	$2(1 - F_{\mathrm{F}(n-1,m-1)}(t))$

Test stř. hodnot dvou normálních rozdělení se známým rozptylem

Předpoklad: Máme 2 *nezávislé* výběry

$(X_1, \dots, X_m)$ z rozdělení $\mathrm{N}(\mathrm{E}\,X, \sigma^2)$ a
$(Y_1, \dots, Y_n)$ z rozdělení $\mathrm{N}(\mathrm{E}\,Y, \sigma^2)$.

Platí, že

$$\bar{\boldsymbol{X}}_m \text{ má rozdělení } \mathrm{N}\left(\mathrm{E}\,X, \frac{\sigma^2}{m}\right),$$

$$\bar{\boldsymbol{Y}}_n \text{ má rozdělení } \mathrm{N}\left(\mathrm{E}\,Y, \frac{\sigma^2}{n}\right), \text{ takže}$$

$$\bar{\boldsymbol{X}}_m - \bar{\boldsymbol{Y}}_n \text{ má rozdělení } \mathrm{N}\left(\mathrm{E}\,X - \mathrm{E}\,Y, \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)\right).$$

Za předpokladu $\mathrm{E}\,X = \mathrm{E}\,Y$

$$T := \frac{\bar{\boldsymbol{X}}_m - \bar{\boldsymbol{Y}}_n}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \text{ má rozdělení } \mathrm{N}(0,1).$$

Testujeme realizaci t na rozdělení $\mathrm{N}(0,1)$, jako jsme to dělali v testu střední hodnoty $\mathrm{N}(\mu, \sigma^2)$ při známém σ^2.

Test stř. hodnot dvou normálních rozdělení s neznámým rozptylem

Předpoklad: $\mathrm{D}\,X = \mathrm{D}\,Y = \sigma^2$.
Nejprve je nutné tento předpoklad ověřit, provedeme test rovnosti rozptylů.
(*Ve skutečnosti nemůžeme předpoklad ověřit, jedině vyvrátit. Pokusíme se o to, a pokud se nám to nepodaří, pokračujeme. Bez tohoto předpokladu by byl další postup složitější.*)

Máme 2 odhady ($S^2_{\boldsymbol{X}}$ a $S^2_{\boldsymbol{Y}}$) téhož parametru σ^2. Vytvoříme z nich 1 sdružený odhad S^2 parametru σ^2: použijeme jejich průměr vážený rozsahy výběrů (-1 kvůli výpočtu výběrového průměru):

$$S^2 = \frac{(m-1)S^2_{\boldsymbol{X}} + (n-1)S^2_{\boldsymbol{Y}}}{m+n-2}$$

Při výpočtu testové statistiky pak místo skutečné směrodatné odchylky σ použijeme její odhad S. To ale vnáší do výpočtu další zdroj neurčitosti, je proto třeba použít místo normálního rozdělení Studentovo.

Za předpokladu $\mathrm{E}\,X = \mathrm{E}\,Y$

$$T := \frac{\bar{\boldsymbol{X}}_m - \bar{\boldsymbol{Y}}_n}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \text{ má rozdělení } t(m+n-2).$$

Testujeme realizaci t na rozdělení $t(m+n-2)$, jako jsme to dělali v testu střední hodnoty $\mathrm{N}(\mu, \sigma^2)$ při neznámém σ^2.

Test stř. hodnot s neznámým rozptylem: Odvození

Ukažme nejprve, že sdružený odhad rozptylu S^2 je nestranným odhadem σ^2. Víme, že

$$\frac{(m-1)S_X^2}{\sigma^2} \text{ má rozdělení } \chi^2(m-1), \ \frac{(n-1)S_Y^2}{\sigma^2} \text{ má rozdělení } \chi^2(n-1), \text{ takže jejich součet}$$

$$\frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} \text{ má rozdělení } \chi^2(m+n-2) \text{ se střední hodnotou } m+n-2. \text{ Proto}$$

$$\frac{(m-1)S_X^2 + (n-1)S_Y^2}{(m+n-2)\sigma^2} = \frac{S^2}{\sigma^2} \text{ má střední hodnotu 1, takže}$$

$$S^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{(m+n-2)} \text{ je nestranný odhad rozptylu } \sigma^2.$$

Nyní ukažme, že testová statistika T má rozdělení $t(m+n-2)$. Víme, že

$$\frac{\bar{X}_m - \bar{Y}_n}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}} \text{ má rozdělení } \mathrm{N}(0,1) \text{ a že}$$

$$\frac{(m+n-2)S^2}{\sigma^2} = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} \text{ má rozdělení } \chi^2(m+n-2), \text{ takže}$$

$$T := \frac{\bar{X}_m - \bar{Y}_n}{S\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{\frac{\bar{X}_m - \bar{Y}_n}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}}{\sqrt{\frac{S^2}{\sigma^2}}} \text{ má rozdělení } t(m+n-2).$$

Příklad: test středních hodnot při neznámém rozptylu

Zadání: Vliv alkoholismu matky na inteligenci dítěte. Ověřte hypotézu, že alkoholismus matek nijak nesnižuje IQ dětí.
Skupina 1: matky chronické alkoholičky, $m = 6$, $\bar{x} = 78$, $\sum_{i=1}^m (x_i - \bar{x})^2 = 1805$.
Skupina 2: kontrolní, „normální“ matky, $n = 46$, $\bar{y} = 99$, $\sum_{j=1}^n (y_j - \bar{y})^2 = 11520$.

Řešení: Jednostranný test středních hodnot 2 norm. rozdělení s neznámým rozptylem.

Krok 1: Ověření rovnosti rozptylů. Otestujme $H_0 : \mathrm{D}\,X = \mathrm{D}\,Y$. Realizace testové statistiky:

$$t = \frac{s_x^2}{s_y^2} = \frac{\frac{1}{m-1}\sum_{i=1}^m (x_i - \bar{x})^2}{\frac{1}{n-1}\sum_{j=1}^n (y_j - \bar{y})^2} = \frac{\frac{1805}{5}}{\frac{11520}{45}} = 1.41.$$

Dosažená hladina významnosti:

$$p = 2\min(F_{\mathrm{F}(m-1,n-1)}(t), 1 - F_{\mathrm{F}(m-1,n-1)}(t)) =$$
$$= 2\min(F_{\mathrm{F}(5,45)}(1.41), 1 - F_{\mathrm{F}(5,45)}(1.41)) = 2\min(0.76, 0.24) = 0.48.$$

Rozdíl rozptylů mezi skupinami není statisticky významný, nezamítáme H_0 o rovnosti rozptylů, můžeme pokračovat.

Krok 2: Test středních hodnot. Testujeme $H_0 : \mathrm{E}\,X \geq \mathrm{E}\,Y$ proti $H_A : \mathrm{E}\,X < \mathrm{E}\,Y$. Sdružený odhad rozptylu:

$$s^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2} = \frac{\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{j=1}^n (y_j - \bar{y})^2}{m+n-2} =$$
$$= \frac{1805 + 11520}{6 + 46 - 2} = 266.5 \quad \Rightarrow \quad s = 16.3248$$

Realizace testové statistiky:

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{78 - 99}{16.3248\sqrt{\frac{1}{6}+\frac{1}{46}}} = -2.9636$$

Dosažená hladina významnosti: $p = F_{t(m+n-2)}(t) = F_{t(50)}(-2.9636) = 0.0023$.

Závěr: Na hladinách významnosti vyšších než 0.23 % můžeme zamítnout hypotézu, že alkoholismus matky nesnižuje IQ dětí.

Párový pokus

Příklad: Porovnání průměrných teplot na dvou místech. Teploty měříme vždy současně na obou místech.

- Rozdíl teplot (pokud nějaký existuje) může být malý v porovnání s proměnlivostí teplot (v noci 0 st. Celsia, ve dne 20 st. Celsia).
- Standardní test středních hodnot je slabý kvůli velkému rozptylu.
- Rozptyl má ale *společnou příčinu*, která se projevuje v obou výběrech stejně: *výběry nejsou navzájem nezávislé*.

Předpoklad: Prvky náhodných výběrů $\boldsymbol{X}_n$ a $\boldsymbol{Y}_n$, tj. náhodné veličiny $X_j, Y_j, j = 1, \ldots, n$, mají normální rozdělení $\mathrm{N}(\mu_j, \sigma^2)$ s konstantním rozptylem σ^2 a *proměnnými středními hodnotami* $\mu_j = \mathrm{E}\, X_j = \mathrm{E}\, Y_j$.

- Náhodné veličiny $U_j := X_j - \mu_j$ a $V_j := Y_j - \mu_j$, $j = 1, \ldots, n$, *jsou nezávislé* a mají rozdělení $\mathrm{N}(0, \sigma^2)$.
- Náhodné veličiny $\Delta_j := U_j - V_j = X_j - Y_j$, $j = 1, \ldots, n$, *jsou nezávislé* a mají rozdělení $\mathrm{N}(0, 2\sigma^2)$.
- Výběrový průměr $\bar{\boldsymbol{\Delta}}$ má rozdělení $\mathrm{N}\left(0, \frac{2\sigma^2}{n}\right)$.

Testy středních hodnot: párový pokus

1. Pro *známý* rozptyl σ^2:
   - Neznámé parametry sdruženého rozdělení jsou $\mu_1, \ldots, \mu_n$, ale nepotřebujeme je.
   - Testujeme
     $$T := \frac{\bar{\boldsymbol{\Delta}}}{\sigma}\sqrt{\frac{n}{2}} = \frac{\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}}{\sigma}\sqrt{\frac{n}{2}}$$
     na rozdělení $\mathrm{N}(0, 1)$.
2. Pro *neznámý* rozptyl:
   - Neznámé parametry sdruženého rozdělení jsou $\sigma^2, \mu_1, \ldots, \mu_n$, ale potřebujeme z nich pouze $\sigma^2 = \mathrm{D}\, X$.
   - Můžeme pracovat přímo s výběrem $(\Delta_1, \ldots, \Delta_n)$ z normálního rozdělení.
   - Testujeme
     $$T := \frac{\bar{\boldsymbol{\Delta}}}{S_\Delta}\sqrt{n}$$
     na rozdělení $t(n-1)$.

Příklad: Test středních hodnot, párový pokus

Zadání: Vliv hydrochlorothiazidu na krevní tlak. Skupině 11 hypertoniků byl nejprve změřen (systolický) tlak po podání placeba a o měsíc později po podání hydrochlorothiazidu (viz tabulka). Ověřte, že hydrochlorothiazid snižuje krevní tlak.

Placebo (X)	211	210	210	203	196	190	191	177	173	170	163
Hydrochlorothiazid (Y)	181	172	196	191	167	161	178	160	149	119	156
Rozdíl (Δ)	30	38	14	12	29	29	13	17	24	51	7

Řešení: Zavedli jsme náhodnou veličinu $\Delta = X - Y$, z níž máme k dispozici náhodný výběr $\mathbf{\Delta}_n$ rozsahu n, $\Delta_i = X_i - Y_i, i = 1, \ldots, n$. Zkusíme vyvrátit hypotézu $H_0 : \mathrm{E}\, X \leq \mathrm{E}\, Y$, tj. $\mathrm{E}\, \Delta \leq 0$.

- $n = 11$, $\bar{\delta} = 24$, $s_\delta = 13.092$
- Realizace testové statistiky:

$$t = \frac{\bar{\delta}}{s_\delta}\sqrt{n} = \frac{24}{13.092}\sqrt{11} = 6.08$$

- Dosažená hladina významnosti:

$$p = 1 - F_{t(n-1)}(t) = 1 - F_{t(10)}(6.08) = 5.94 \times 10^{-5}.$$

Závěr: Zamítáme H_0 a přijímáme H_A, tj. hydrochlorothiazid snižuje krevní tlak.

Recept: Testy parametrů dvou alternativních rozdělení

Předpoklad: Máme 2 *nezávislé* výběry $\boldsymbol{X}_m = (X_1, \ldots, X_m)$ z rozdělení $Ber(q_1)$ a $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)$ z rozdělení $Ber(q_2)$.

Platí-li $q_1 = q_2 = q$, můžeme pro parametr q použít maximálně věrohodný odhad pomocí obou výběrů:

$$R = \frac{m\bar{\boldsymbol{X}} + n\bar{\boldsymbol{Y}}}{m+n}.$$

Pro dostatečně velké rozsahy výběrů ($m > 100$, $n > 100$) lze *rozdělení výběrových relativních četností* $\bar{\boldsymbol{X}}$ a $\bar{\boldsymbol{Y}}$ aproximovat normálními rozděleními:

$$\bar{\boldsymbol{X}} \text{ má přibližně rozdělení } \mathrm{N}\left(R, \frac{R(1-R)}{m}\right),$$
$$\bar{\boldsymbol{Y}} \text{ má přibližně rozdělení } \mathrm{N}\left(R, \frac{R(1-R)}{n}\right), \text{ takže}$$
$$\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}} \text{ má přibližně rozdělení } \mathrm{N}\left(0, \frac{R(1-R)}{m} + \frac{R(1-R)}{n}\right).$$

Testovou statistiku

$$T := \frac{\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}}{\sqrt{\frac{R(1-R)}{m} + \frac{R(1-R)}{n}}} = \frac{\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}}{\sqrt{R(1-R)\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

testujeme na rozdělení $\mathrm{N}(0, 1)$.

Příklad: Test rovnosti populačních pravděpodobností

Zadání: Přijímací řízení na Berkeley v roce 1973. Z 8300 přihlášených mužů bylo přijato 3700, z 4300 přihlášených žen bylo přijato 1500. Ověřte hypotézu, že pravděpodobnost přijetí mužů a žen je stejná.

Řešení: Oboustranný test hypotézy o rovnosti pravděpodobnosti přijetí pro muže q_m a ženy q_z:

- Počet přihlášených mužů je $m = 8300$ a žen $n = 4300$.
- Realizace relativních výběrových četností (úspěšnost) je pro muže $\bar{x} = \frac{3700}{8300} = 0.4458$ a pro ženy $\bar{y} = \frac{1500}{4300} = 0.3488$.
- Platí-li $q_m = q_z = q$, můžeme q odhadnout pomocí $r = \frac{m\bar{x}+n\bar{y}}{m+n} = \frac{3700+1500}{8300+4300} = 0.4127$.
- Realizace testové statistiky je

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{r(1-r)\left(\frac{1}{m} + \frac{1}{n}\right)}} = \frac{0.4458 - 0.3488}{\sqrt{0.4127(1-0.4127)\left(\frac{1}{8300} + \frac{1}{4300}\right)}} = 10.4802$$

- Dosažená hladina významnosti

$$p = 2(1 - \Phi(t)) \doteq 0.$$

Závěr: Zamítáme H_0, na základě těchto dat existuje jen mizivá šance, že by pravděpodobnosti přijetí muže a ženy mohly být shodné.

Je to důkaz pohlavní diskriminace?

Příklad: Test rovnosti populačních pravděpodobností (pokr.)

Zadání: Přijímací řízení na Berkeley v roce 1973. Tentokrát zohledněme i to, na jaký směr (umění nebo věda) se zájemci hlásili.

Řešení: Kromě celkových výsledků z předchozího slidu, níže uvedená tabulka obsahuje stejné výsledky také pro každý směr zvlášť.

	Muži			Ženy			Celkem	Test	
	Přihl.	Přij.	Poměr	Přihl.	Přij.	Poměr	Poměr	Stat.	Dos. význ.
Směr	m		$\bar{x}$	n		$\bar{y}$	r	t	p
Umění	2300	700	0.3043	3200	900	0.2813	0.2909	1.8604	0.0628
Věda	6000	3000	0.5000	1100	600	0.5455	0.5070	-2.7720	0.0056
Celkem	8300	3700	0.4458	4300	1500	0.3488	0.4127	10.4802	0.0000

Závěry:

- V uměleckých směrech nelze zamítnout hypotézu o stejné pravděpodobnosti přijetí mužů a žen.
- Ve vědeckých směrech tuto hypotézu zamítnout lze, dosažená hladina významnosti je cca 0.5 %.
- Zajímavé ovšem je, že na uměleckých směrech, kam se hlásí více žen, mají vyšší pravděpodobnost přijetí muži, zatímco na vědeckých směrech, kam se hlásí více mužů, mají vyšší pravděpodobnost přijetí ženy. Dochází tedy spíše k *pozitivní diskriminaci*.

Simpsonův paradox: Co platí pro části, nemusí platit pro celek.

- Muži a ženy jsou přijímáni přibližně shodně. Ženy ovšem mají tendenci hlásit se na umělecké směry, kde je výběr přísnější, což vysvětluje jejich celkově nižší úspěšnost v přijímačkách.
- Z celkových čísel nelze správně pochopit efekt pohlaví na přijetí kvůli matoucímu faktoru (směr), který nebyl řízen. Když se zařadil do studie, dostali jsme mnohem přesnější obrázek.