Josef Machek
Test rovnosti kvantilů dvou normálních rozdělení

Aplikace matematiky, Vol. 5 (1960), No. 3, 210–215

Persistent URL: http://dml.cz/dmlcz/102707

TEST ROVNOSTI KVANTILŮ DVOU NORMÁLNÍCH ROZDĚLENÍ

JOSEF MACHEK

(Došlo dne 24. července 1959)

Je navržen přibližný test hypothesy, že dvě normální rozdělení s neznámými průměry a rozptyly mají stejný $100p\%$ a $100q\%$ kvantil (p může být různé od q). Zároveň je uveden přibližný výraz pro kvantil necentrálního t-rozdělení. Navržený test je založen na „necentrální analogii" WELCHOVA řešení Fisher-Behrensova problému.

1. ÚVOD

V některých aplikacích není pro srovnání dvou normálních rozdělení rozhodující ani pouhá vzájemná poloha jejich středních hodnot, ani pouhý vzájemný poměr jejich rozptylů, nýbrž vzájemná poloha jejich určitých kvantilů. Tak např. může mít význam otázka, zda 10% kvantil jednoho rozdělení není menší než 10% kvantil druhého rozdělení. Někdy dokonce může jít u jednoho rozdělení o jiný kvantil než u druhého, např. je třeba srovnat 10% kvantil s 5%, apod. Zejména pro rozdělení životnosti (trvanlivosti) nebo pro rozdělení odolnosti vůči určitému podnětu jsou nejdůležitější charakteristikou kvantily a nikoliv samotné střední hodnoty nebo rozptyly.

Jestliže rozptyly obou rozdělení jsou známy, pak se hypothesa o jejich kvantilech redukuje na hypothesu o jejich středních hodnotách. Jestliže rozptyly jsou stejné a srovnávají se stejné kvantily, pak jde rovněž o test hypothesy o rozdílu mezi průměry dvou normálních rozdělení a úloha je řešena Studentovým testem. Jde-li o dva různé kvantily a rozptyly jsou stejné, pak — jak ukázal WALSH v [1] — je úloha testu rovnosti dvou kvantilů a úloha intervalového odhadu rozdílu mezi kvantily přesně řešena pomocí necentrálního t-rozdělení. Nelze-li předpokládat, že rozptyly obou rozdělení jsou stejné, a je-li jejich poměr neznám, je úloha podstatně složitější a bude zde jen přibližně vyřešena.

V oddíle 3 bude popsán test rovnosti dvou kvantilů pro případ, kdy poměr rozptylů je znám, v oddíle 4 budou uvedeny dva přibližné testy pro případ, kdy poměr rozptylů je neznám.

2. FORMULACE ÚLOHY A ZNAČENÍ

Jsou dány dva soubory s normálním rozdělením, P_1 se střední hodnotou μ_1 a s rozptylem σ_1^2 a P_2 se střední hodnotou μ_2 a s rozptylem σ_2^2. Dále nechť ξ značí $100p\%$ kvantil souboru P_1 a η $100q\%$ kvantil souboru P_2, kde p, q jsou daná čísla z intervalu $(0,1)$. Z P_1 se provádí náhodný výběr rozsahu n_1, $X_1, X_2, \ldots, X_{n1}$, z P_2 náhodný výběr $Y_1, Y_2, \ldots, Y_{n2}$ rozsahu n_2. Na základě těchto výběrů jest ověřiti hypothesu

$$\xi = \eta$$

proti alternativě

$$\xi > \eta .$$

Odhad střední hodnoty μ_1 označíme $\bar{x}$, odhad střední hodnoty μ_2 pak $\bar{y}$. Nestranné odhady rozptylů σ_1^2 a σ_2^2 označíme s_1^2 a s_2^2, tedy

$$\bar{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i , \quad \bar{y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i ,$$

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{x})^2 , \quad s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{y})^2 .$$

Hypothesa $\xi = \eta$ je zřejmě ekvivalentní hypothese

$$\mu_1 - \mu_2 = u_q \sigma_2 - u_p \sigma_1 ,$$

kde u_p a u_q jsou kvantily standardisovaného normálního rozdělení,

$$\int_{-\infty}^{u_p} (2\pi)^{-\frac{1}{2}} \mathrm{e}^{-\frac{t^2}{2}} \mathrm{d}t = p , \quad \int_{-\infty}^{u_q} (2\pi)^{-\frac{1}{2}} \mathrm{e}^{-\frac{t^2}{2}} \mathrm{d}t = q ,$$

a alternativa $\xi > \eta$ je ekvivalentní alternativě

$$\mu_1 - \mu_2 > u_q \sigma_2 - u_p \sigma_1 .$$

Hypothesu $\xi = \eta$ zamítneme ve prospěch alternativy $\xi > \eta$, jestliže odhad rozdílu $\mu_1 - \mu_2$ nabude „příliš vysoké hodnoty", tj. když nastane jev

$$\bar{x} - \bar{y} > A(s_1, s_2) , \tag{1}$$

kde $A(s_1, s_2)$ je funkce odhadů s_1, s_2 směrodatných odchylek σ_1, σ_2, stanovená tak, aby

$$\mathbf{P}\{\bar{x} - \bar{y} > A(s_1, s_2) \mid \xi = \eta\} = \alpha , \tag{2}$$

tj., aby pravděpodobnost splnění nerovnosti (1), když hypothesa $\xi = \eta$ je správná, byla rovna zvolené hladině významnosti α.

Hlavní problém spočívá v nalezení funkce $A(s_1, s_2)$.

3. ŘEŠENÍ PRO PŘÍPAD ZNÁMÉHO POMĚRU ROZPTYLŮ

Označme poměr rozptylů σ_2^2 a σ_1^2 symbolem λ^2,

$$\frac{\sigma_2^2}{\sigma_1^2} = \lambda^2 .$$

Za platnosti testované hypothesy má rozdíl $\bar{x} - \bar{y}$ zřejmě normální rozdělení se střední hodnotou

$$\sigma_2 u_q - \sigma_1 u_p = \sigma_1(\lambda u_q - u_p)$$

a s rozptylem

$$\sigma_1^2 \left\{\frac{1}{n_1} + \frac{\lambda^2}{n_2}\right\}.$$

Rozptyl σ_1^2 odhadneme ze spojených výběrů jako

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2/\lambda^2}{n_1 + n_2 - 2}.$$

Veličina

$$\frac{\bar{x} - \bar{y}}{s\left[\dfrac{1}{n_1} + \dfrac{\lambda^2}{n_2}\right]^{\frac{1}{2}}}$$

má za platnosti testované hypothesy necentrální t-rozdělení s parametrem necentrality

$$\delta = (\lambda u_q - u_p)\left[\frac{1}{n_1} + \frac{\lambda^2}{n_2}\right]^{-\frac{1}{2}}$$

a s $f = n_1 + n_2 - 2$ stupni volnosti.

Funkce $A(s_1, s_2)$ je tedy pro případ, kdy rozptyly σ_2^2, σ_1^2 jsou ve známém poměru λ^2, dána výrazem

$$A(s_1, s_2) = s\left[\frac{1}{n_1} + \frac{\lambda^2}{n_2}\right]^{\frac{1}{2}} .\, t\left(n_1 + n_2 - 2\,,\ \frac{\lambda u_q - u_p}{\left[\dfrac{1}{n_1} + \dfrac{\lambda^2}{n_2}\right]^{\frac{1}{2}}}\,,\ \alpha\right),$$

kde $t(f, \delta, \alpha)$ značí $100\alpha\%$ kritickou hodnotu (tj. $100(1 - \alpha)\%$ kvantil) necentrálního t-rozdělení s f stupni volnosti a s parametrem necentrality δ. Tuto kritickou hodnotu lze stanovit z tabulek necentrálního t-rozdělení, připravených Johnsonem a Welchem (viz [3] a [4]), nebo vypočítat z následujícího rozvoje. Rozvoj $100\gamma\%$ kvantilu necentrálního t-rozdělení a f stupni volnosti a s parametrem necentrality δ v mocninách převrácené hodnoty počtu stupňů volnosti až do členů druhého řádu jest

$$t(f, \delta, 1 - \gamma) = A_\gamma\left\{1 + \frac{1 + A_\gamma u_\gamma}{4f} + \frac{1 + A_\gamma u_\gamma}{96f^2} .\, (8A_\gamma u_\gamma - 3A_\gamma^2 + 4) + \frac{7A_\gamma^2 - 1}{96f^2}\right\}, \tag{3}$$

kde A_γ značí $\delta + u_\gamma$.

4. ŘEŠENÍ PRO PŘÍPAD NEZNÁMÉHO POMĚRU ROZPTYLŮ

Pro případ, kdy poměr rozptylů není znám, je nutna jakási „necentrální analogie“ řešení Fisher-Behrensova problému, jaké podal Welch v [2], po případě vhodné rozšíření tabulek A. Aspinové [5], které z tohoto řešení vycházejí. Stanovení funkce $A(s_1, s_2)$ odhadů s_1 a s_2, splňující podmínku (2), je zde však mnohem složitější než v případě srovnání průměrů. Stejným postupem, jakého užil Welch v [2], byl hledán rozvoj funkce $A(s_1, s_2)$ v řadu v mocninách $\frac{1}{f_1}, \frac{1}{f_2}$, kde f_i je počet stupňů volnosti pro odhad s_i, $i = 1, 2$. Složitost členů roste však mnohem rychleji než v „centrálním případě“, takže již členy druhého řádu v $\frac{1}{f_i}$ by neměly pro praktický výpočet význam.

Pro zjednodušení zápisu položme

$$-u_p = a_1\,, \quad u_q = a_2\,, \quad \frac{1}{n_1} = b_1\,, \quad \frac{1}{n_2} = b_2\,.$$

V tomto značení jsou parametry normálního rozdělení charakteristiky $\bar{x} - \bar{y}$ za platnosti testované hypothesy rovny:

střední hodnota

$$\mathbf{E}(\bar{x} - \bar{y}) = \sum_{i=1}^{2} a_i \sigma_i$$

a rozptyl

$$\mathbf{D}(\bar{x} - \bar{y}) = \sum_{i=1}^{2} b_i \sigma_i^2\,.$$

Do členů prvního řádu v $\frac{1}{f_i}$ je funkce $A(s_1, s_2)$ z (1) rovna

$$A(s_1, s_2) = \sum_1^2 a_i s_i + u_{1-\alpha}\left[\sum_1^2 b_i s_i^2\right]^{\frac{1}{2}} + \sum_{i=1}^2 \frac{a_i s_i}{4f_i} + \tag{4}$$

$$+ u_{1-\alpha}\left[\sum_1^2 b_i s_i^2\right]^{-\frac{3}{2}} \sum_1^2 \left\{\left(a_i s_i\left[\sum b_i s_i^2\right]^{\frac{1}{2}} + u_{1-\alpha} b_i s_i^2\right)^2 + b_i^2 s_i^4\right\} \Big/ (4 f_i)\,.$$

Položíme-li zde $a_i = 0$, $i = 1, 2$ (což odpovídá testu rovnosti 50% kvantilů), dostaneme výraz, shodný s prvními členy rozvoje kritické hodnoty pro test hypothesy $\mu_1 = \mu_2$ při různých rozptylech, jak je uveden v [2].

Jiná možnost přibližného stanovení funkce $A(s_1, s_2)$ plyne z nahrazení lineární kombinace náhodných veličin s χ^2 rozdělením, $\Sigma\beta_i\chi_i^2(f_i)$, náhodnou veličinou tvaru $c\,\chi^2(f)$, kde c a f je určeno tak, aby první dva momenty veličiny $c\,\chi^2(f)$ se shodovaly s prvními dvěma momenty veličiny $\Sigma\beta_i\,\chi_i^2(f_i)$, čili

$$c = (\sum\beta_i f_i)(\sum\beta_i^2 f_i)^{-1}\,, \quad f = (\sum\beta_i f_i)^2(\sum\beta_i^2 f_i)^{-1}\,.$$

Odhad $\sum_1^2 b_i s_i^2$ má rozdělení jako $\sum_1^2 \beta_i \chi_i^2(f_i)$ s $\beta_i = b_i \sigma_i^2 / f_i$, $f_i = n_i - 1$. Nahradíme-li toto rozdělení rozdělením $c\,\chi^2(f)$, pak kriterium

$$(5) \qquad (\bar{x} - \bar{y})\left(\sum_1^2 b_i s_i^2\right)^{-\frac{1}{2}}$$

má rozdělení přibližně jako

$$(\bar{x} - \bar{y})(c\,\chi^2(f))^{-\frac{1}{2}} = \frac{(\bar{x} - \bar{y})(\Sigma \beta_i f_i)^{-\frac{1}{2}}}{[\chi^2(f)]^{\frac{1}{2}}[(\Sigma \beta_i f_i)^2/(\Sigma \beta_i^2 f_i)]^{\frac{1}{2}}} = \frac{Y}{[\chi^2(f)/f]^{\frac{1}{2}}},$$

kde Y má normální rozdělení se střední hodnotou $\Sigma a_i \sigma_i / \sqrt{\Sigma b_i \sigma_i^2}$ a s rozptylem 1. Charakteristika (5) má pak přibližně necentrální t-rozdělení s parametrem necentrality $\Sigma a_i \sigma_i / \sqrt{\Sigma b_i \sigma_i^2}$ a s

$$f = \frac{(\Sigma \beta_i f_i)^2}{\Sigma \beta_i^2 f_i} = \frac{(\Sigma b_i \sigma_i^2)^2}{\Sigma b_i^2 \sigma_i^4}$$

stupni volnosti. Jako přibližné hodnoty pro $A(s_1, s_2)$ by tedy bylo možno užít

$$(6) \qquad A_1(s_1, s_2) = \sqrt{\Sigma b_i s_i^2}\,.\,t\left(\frac{(\Sigma b_i s_i^2)^2}{\Sigma b_i^2 s_i^4}, \frac{\Sigma a_i s_i}{\sqrt{\Sigma b_i s_i^2}}, \alpha\right),$$

kde $t(f, \delta, \alpha)$ je $100\alpha\,\%$ kritická hodnota necentrálního t-rozdělení. Do výrazů pro parametr necentrality a pro počet stupňů volnosti byly za neznámé parametry σ_i dosazeny odhady s_i, $i = 1, 2$.

Pro srovnání obou aproximací byl vypočten rozdíl mezi výrazy (5) a (6), kde v (6) byl kvantil

$$t\left(\frac{(\Sigma b_i s_i^2)^2}{\Sigma b_i^2 s_i^4}, \frac{\Sigma a_i s_i}{\sqrt{\Sigma b_i s_i^2}}, \alpha\right)$$

rozvinut podle (3). Aproximace se liší již ve členech řádu prvního, zatímco přibližné kritické hodnoty pro test rozdílu mezi průměry se odchylují od rozvoje přesného řešení až ve členech druhého řádu.

Literatura

[1] *J. E. Walsh:* On the use of the non-central t-distribution for comparing percentage points of normal populations; Annals of Mathematical Statistics 19 (1948), 93.

[2] *B. L. Welch:* The generalization of Student's problem when several different population variances are involved; Biometrika 34 (1947), 28.

[3] *N. L. Johnson, B. L. Welch:* Applications of the non-central t-distribution; Biometrika 31 (1939), 362.

[4] *J. Janko:* Statistické tabulky; Praha 1958.

Резюме

ПРОВЕРКА РАВЕНСТВА КВАНТИЛЕЙ ДВУХ НОРМАЛЬНЫХ РАСПРЕДЕЛЕНИЙ

ИОСЕФ МАХЕК (Josef Machek)

В настоящей заметке приведен приближенный критерий для проверки гипотезы, что две нормальных совокупности имеют равные друг другу $100p\%$ и $100q\%$ квантили (здесь p может быть разным от q), если взаимное отношение дисперсий неизвестно. В то же время дано приближенное выражение (3) для квантили нецентрального t—распределения. Предложенный критерий является каким-то ,,нецентральным аналогом" критерия Вельча (см. [2]) для проверки равенства средних значений двух нормальных совокупностей, если их дисперсии вполне неизвестны, не равны друг другу, и неизвестно даже их взаимное отношение.

Summary

ON A COMPARISON OF TWO NORMAL POPULATIONS

JOSEF MACHEK

In this note an approximate test is given for the hypothesis that two normal populations with completely unknown means and variances, have equal $100p\%$ and $100q\%$ quantiles (p and q are given numbers, p possibly different from q). Also an approximation to the quantile of the non-central t-distribution is given. The proposed significance test might be said to be a "non-central analogue" of Welch's solution of the Behrens — Fisher problem.