Aplikace matematiky

Zbyněk Šidák; Jiří Vondráček

Jednoduchý neparametrický test rozdílnosti polohy dvou populací

Aplikace matematiky, Vol. 2 (1957), No. 3, 215–221

Persistent URL: http://dml.cz/dmlcz/102568

JEDNODUCHÝ NEPARAMETRICKÝ TEST ROZDÍLNOSTI POLOHY DVOU POPULACÍ

ZBYNĚK ŠIDÁK, JIŘÍ VONDRÁČEK

(Došlo dne 31. srpna 1956.) DT: 519.27

V článku je zdokonalen Rosenbaumův [1] test nulové hypothesy totožnosti dvou populací proti alternativní hypothese jejich vzájemného posunutí. Výhodou testu je jednoduché provedení. Pro praktické používání jsou uvedeny tabulky.

1. Testová charakteristika a její rozložení

Budiž $x_1 < x_2 < \ldots < x_n$ uspořádaný náhodný výběr rozsahu n z rozložení $F(x)$ a $y_1 < y_2 < \ldots < y_m$ uspořádaný náhodný výběr rozsahu m z rozložení $G(y)$. Distribuční funkce $F(x)$, $G(y)$ nechť jsou spojité. (Tímto předpokladem jsou tedy vyloučeny rovnosti mezi výběrovými hodnotami.) Úkolem je testovat hypothesu $F(x) \equiv G(x)$ proti alternativní hypothese, že $G(y)$ je posunuta směrem k větším hodnotám vzhledem k $F(x)$.

Jako testové charakteristiky užijeme $Q = R + S$, kde R jest počet prvků z výběru $\{x_i\}$, které jsou menší než y_1, a S jest počet prvků z výběru $\{y_j\}$, které jsou větší než x_n. Předpokládejme totožnost distribucí a odvoďme rozložení testové charakteristiky Q.

Označme $P^*(r, s)$ pravděpodobnost, že mezi x_r a x_n bude $m - s$ prvků z výběru $\{y_j\}$ a nad x_n bude s prvků z téhož výběru. Podle Wilksova [2] obecného vzorce dostaneme

$$P^*(r, s) = \frac{\binom{m + n - s - r - 1}{m - s}}{\binom{m + n}{n}} \qquad \begin{matrix} \text{pro } r = 1, 2, \ldots, n - 1 , \\ s = 0, 1, 2, \ldots, m . \end{matrix}$$

Pravděpodobnost $P(r, s)$, že nad x_n bude s prvků z výběru $\{y_j\}$ a prvek y_1 bude mezi x_r a x_{r+1}, je zřejmě

$$P(r, s) = P^*(r, s) - P^*(r + 1, s) = \frac{\binom{m + n - s - r - 2}{m - s - 1}}{\binom{m + n}{n}} \tag{1}$$

$$\text{pro } r = 1, 2, \ldots, n - 2 , \; s = 0, 1, 2, \ldots, m - 1 .$$

Lze snadno dokázat, že vzorec (1) platí celkem pro $r = 0, 1, 2, \ldots, n - 1$ a $s = 0, 1, 2, \ldots, m - 1$ a jest

$$P(n, m) = \frac{1}{\binom{m+n}{n}}. \tag{2}$$

Případ $r < n$, $s = m$ nebo $r = n$, $s < m$ nemůže nastat, tedy pro tyto případy $P(r, s) = 0$.

$P(r, s)$ je tedy pravděpodobnost, že $R = r$, $S = s$, a pravděpodobnost, že $Q = q$, je rovna

$$\sum_{r+s=q} P(r, s) \quad \text{pro} \quad q = 0, 1, \ldots, m + n - 2, \quad m + n. \tag{3}$$

Je-li $0 < \alpha < 1$, označme k_α nejmenší celé číslo takové, že $\sum_{r+s=k_\alpha}^{m+n} P(r, s) \leqq \alpha$. Když testová charakteristika $Q \geqq k_\alpha$, zamítneme nulovou hypothesu rovnosti obou distribucí na hranici významnosti menší nebo rovné α. Postup výpočtu kritických hodnot k_α je uveden v části 2.

Poznámka. Pravděpodobnost $P(r, s)$ lze snadno odvoditi kombinatoricky. Za předpokladu nulové hypothesy můžeme výběry $\{x_i\}$ a $\{y_j\}$ sloučit a vytvořit tak jediný výběr $z_1 < z_2 < \ldots < z_{m+n}$. Každé uspořádání prvků výběrů $\{x_i\}$ a $\{y_j\}$ ve výběru $\{z_k\}$ odpovídá tomu, že n prvků výběru $\{z_k\}$ označíme znakem x a zbylých m prvků znakem y, t. j. vybíráme kombinace n-té třídy z $n + m$ prvků. Každá tato kombinace má stejnou pravděpodobnost $\frac{1}{\binom{m+n}{n}}$. Má-li být nad x_n právě s prvků y a pod y_1 právě r prvků x, musíme od konce výběru $\{z_k\}$ označit s prvků znakem y a další prvek znakem x a právě tak od začátku r prvků označit x a další prvek y; ostatní prvky můžeme označit libovolně. Tedy ze zbylých $n + m - s - r - 2$ prvků máme ještě $n - r - 1$ prvků označit znakem x, což dává $\binom{n+m-s-r-2}{n-r-1}$ příznivých možností. Odtud vyplynou opět vzorce (1) a (2).

2. Výpočet tabulek kritických hodnot a asymptotické kritické hodnoty

Pro pevné rozsahy výběrů n, m označme $P_{n,m}(k)$ pravděpodobnost, že $Q \geqq k$. Když dodefinujeme $\binom{a}{b} = 0$ pro $a < b$ nebo $b < 0$, $\binom{0}{0} = 1$, potom pro $n \geqq 2$, $m \geqq 2$, $0 \leqq k \leqq n + m - 2$ platí

$$P_{n,m}(k) = \sum_{r+s=k}^{m+n} P(r, s) = \frac{1}{\binom{m+n}{n}} \left\{ \sum_{r=k}^{n-1} \binom{m+n-r-2}{m-1} + \right.$$

$$+ \sum_{r=k-1}^{n-1} \binom{m+n-r-3}{m-2} + \ldots + \left. \sum_{r=1}^{n-1} \binom{m+n-r-k-1}{m-k} \right) +$$

$$+\sum_{r=0}^{n-1}\binom{m+n-r-k-2}{m-k-1}+\sum_{r=0}^{n-1}\binom{m+n-r-k-3}{m-k-2}+$$

$$+\ldots+\sum_{r=0}^{n-1}\binom{n-r-1}{0}+1\Big\}=\frac{1}{\binom{m+n}{n}}\left\{\binom{m+n-k}{n}+\binom{m+n-k-1}{m}+\right.$$

$$\left.+\binom{m+n-k-1}{m-1}+\ldots+\binom{m+n-k-1}{m-k+1}\right\}. \tag{4}$$

Píšeme-li $P_{n,m}(k)=\frac{1}{\binom{m+n}{n}}\cdot A_{n,m}(k)$, potom pro $0\leqq k\leqq n+m-2$ jest $A_{n,m}(k)$ rovno výrazu ve svorkové závorce v (4). Pro krajní hodnoty platí $A_{n,m}(n+m)=A_{n,m}(n+m-1)=1$.

Pomocí známých kombinatorických vzorců dostáváme pro $1\leqq k\leqq n+m-2$

$$2A_{n,m}(k)=2\binom{m+n-k}{n}+\binom{m+n-k-1}{m}+\binom{m+n-k-1}{m-k+1}+$$

$$+\binom{m+n-k}{m}+\binom{m+n-k}{m-1}+\ldots+\binom{m+n-k}{m-k+2}=2\binom{m+n-k}{n}+$$

$$+\binom{m+n-k-1}{m}+\binom{m+n-k-1}{m-k+1}+A_{n,m}(k-1)-\binom{m+n-k+1}{n}=$$

$$=A_{n,m}(k-1)+\binom{m+n-k-1}{n}+\binom{m+n-k-1}{m}. \tag{5}$$

Jelikož $A_{n,m}(0)=A_{m,n}(0)=\binom{m+n}{n}$, ze vzorce (5) indukcí plyne, že $A_{n,m}(k)=A_{m,n}(k)$, tedy také $P_{n,m}(k)=P_{m,n}(k)$ a tabulka kritických hodnot bude zřejmě symetrická podle diagonály $n=m$.

Kritické hodnoty v tabulkách byly vypočteny pro hranice významnosti $\alpha=0{,}01$ a $\alpha=0{,}05$ a pro rozsahy výběrů $n, m=1, 2, 3, \ldots, 26$ tím, že bylo nalezeno nejmenší k takové, že

$$A_{n,m}(k)\leqq\binom{m+n}{n}\cdot\alpha.$$

Ke kontrole výpočtu tabulek bylo užito vzorce (5). Další kontrolu tabulek poskytují následující nerovnosti.

$$\left.\begin{array}{l}\textit{Když } n \textit{ je pevné, } m \textit{ roste, potom:}\\ \textit{je-li } m<n-1, \textit{ při přechodu od } m \textit{ k } m+1 \textit{ kritická hodnota nevzroste,}\\ \textit{je-li } m=n-1, \textit{ při přechodu od } m \textit{ k } m+1 \textit{ kritická hodnota je táž,}\\ \textit{je-li } m>n-1, \textit{ při přechodu od } m \textit{ k } m+1 \textit{ kritická hodnota neklesne.}\end{array}\right\} \tag{6}$$

V důsledku symetrie platí obdobné vztahy pro m pevné, n rostoucí.

Tvrzení (6) dokážeme ve dvou krocích.

Pomocné tvrzení.[1]) Pro $m \lessgtr n - 1$ platí

$$(m+1)\left\{\binom{m+n-k}{n} + \binom{m+n-k}{m+1}\right\} \lessgtr (m+n+1)\left\{\binom{m+n-k-1}{n} + \binom{m+n-k-1}{m}\right\}.$$

Důkaz. Úpravou a pomocí kombinatorických vzorců dostaneme ekvivalentní nerovnost $\binom{m}{k} \lessgtr \binom{n-1}{k}$, která jest podle předpokladu správná.

Tvrzení. Pro $m \lessgtr n - 1$ platí $P_{n,m+1}(k) \lessgtr P_{n,m}(k)$.

Důkaz. Uvedená nerovnost je ekvivalentní s nerovností

$$(m+1)\, A_{n,m+1}(k) \lessgtr (m+n+1)\, A_{n,m}(k)\,. \tag{7}$$

Pro $k = 0$ nastává ve vzorci (7) rovnost, tedy platí $\leqq$ a současně $\geqq$. Předpokládejme, že (7) je dokázáno pro $k - 1$. Dosazením vzorce (5) a přičtením nerovnosti z pomocného tvrzení plyne platnost vzorce (7) pro k.

Tímto tvrzením je zřejmě také dokázáno (6).

Rovněž je možno dokázat, že pro stejné rozsahy výběrů $m = n$ při přechodu po diagonále k $m + 1 = n + 1$ kritické hodnoty neklesnou.

Poněvadž nám nebyly dosažitelné tabulky binomických koeficientů pro větší hodnoty m a n, nemohli jsme rozšířit tabulky pro tyto větší rozsahy. Avšak s pomocí hořejších nerovností jsme zjistili, že pro stejné rozsahy výběrů $m = n = 27, \ldots, 50$ je kritická hodnota na 1% hranici významnosti 9 a na 5% hranici je kritická hodnota 7.

Když rozsahy výběrů n, m rostou nade všechny meze tak, že $\frac{n}{m} \to 1$, pak snadným výpočtem dostaneme

$$\lim_{n,m\to\infty} P(r,s) = \lim_{n,m\to\infty} \frac{(m+n-s-r-2)!\, n!\, m!}{(m-s-1)!(n-r-1)!(m+n)!} =$$

$$= \lim_{n,m\to\infty} \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{r}{n}\right)\left(1-\frac{1}{m}\right)\left(1-\frac{2}{m}\right)\ldots\left(1-\frac{s}{m}\right)}{B_{n,m}} = 2^{-(r+s+2)},$$

$$B_{n,m} = \left(1+\frac{n}{m}\right)\left(1+\frac{n}{m}-\frac{1}{m}\right)\left(1+\frac{n}{m}-\frac{2}{m}\right)\ldots\left(1+\frac{n}{m}-\frac{s}{m}\right)\left(1+\frac{m}{n}-\right.$$

$$\left.-\frac{s+1}{n}\right)\ldots\left(1+\frac{m}{n}-\frac{s+r+1}{n}\right).$$

Jest

$$\sum_{r+s=k}^{\infty} 2^{-(r+s+2)} = \frac{k+2}{2^{k+1}}\,;$$

[1]) Symbolu $\lessgtr$ v následujících výpočtech jest rozuměti tak, že platí buď ve všech nerovnostech $\leqq$ nebo ve všech nerovnostech $\geqq$.

odtud plyne, že

pro $\alpha = 0{,}01$ jest asymptotická kritická hodnota[2]) $k_\alpha = 10$,
pro $\alpha = 0{,}05$ jest asymptotická kritická hodnota $k_\alpha = 7$.

3. Závěr

Podnětem k této práci byl ROSENBAUMŮV článek [1], v kterém jest jako testové charakteristiky užito pouze S (podle našeho značení). Je nasnadě myšlenka, že test se zlepší, užijeme-li místo Rosenbaumovy „jednostranné" charakteristiky S „oboustranné" charakteristiky $Q = R + S$, jak jsme to učinili v tomto článku. Je vidět, že Rosenbaumův test je velmi citlivý na různost rozptylů. Uvedený test tuto citlivost zmenšuje. Jsou známy některé neparametrické testy tohoto druhu, které jsou mohutnější, ovšem za cenu zkomplikování výpočtů. Test navržený v tomto článku má proti nim výhodu značné jednoduchosti.

Tabulka kritických hodnot k pro 1% hranici významnosti

n \ m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	—																									
2	—	—																								
3	—	—	—																							
4	—	—	—	—																						
5	—	—	—	8	8																					
6	—	—	—	8	8	8																				
7	—	—	9	9	9	8	8																			
8	—	—	10	9	9	9	8	8																		
9	—	—	10	10	9	9	9	9	9																	
10	—	—	11	10	10	9	9	9	9	9																
11	—	—	12	11	10	10	9	9	9	9	9															
12	—	—	12	11	11	10	9	9	9	9	9	9														
13	—	14	13	12	11	10	10	9	9	9	9	9	9													
14	—	15	14	13	12	11	10	10	9	9	9	9	9	9												
15	—	16	14	13	12	11	11	10	10	9	9	9	9	9	9											
16	—	17	15	14	13	12	11	10	10	10	9	9	9	9	9	9										
17	—	18	16	14	13	12	11	11	10	10	10	9	9	9	9	9	9									
18	—	19	17	15	14	13	12	11	11	10	10	10	9	9	9	9	9	9								
19	—	19	17	16	14	13	12	12	11	10	10	10	10	9	9	9	9	9	9							
20	—	20	18	16	15	14	13	12	11	11	10	10	10	10	9	9	9	9	9	9						
21	—	21	19	17	15	14	13	12	12	11	11	10	10	10	10	9	9	9	9	9	9					
22	—	22	20	18	16	15	14	13	12	11	11	11	10	10	10	9	9	9	9	9	9	9				
23	—	23	20	18	17	15	14	13	12	12	11	11	10	10	10	10	9	9	9	9	9	9	9			
24	—	24	21	19	17	16	15	14	13	12	12	11	11	10	10	10	10	9	9	9	9	9	9	9		
25	—	25	22	20	18	16	15	14	13	12	12	11	11	11	10	10	10	10	9	9	9	9	9	9	9	
26	—	26	23	20	18	17	15	14	13	13	12	12	11	11	10	10	10	10	10	9	9	9	9	9	9	9

[2]) Pro $k = 9$ jest $\frac{k+2}{2^{k+1}} = 0{,}010742$, pro $k = 10$ jest $\frac{k+2}{2^{k+1}} = 0{,}00586$. Při $m, n \to \infty$ pro $k = 9$ uvedený součet tedy jen o málo převýší 1% hranici; tím se vysvětluje, že i pro dosti velké rozsahy výběrů $m = n = 50$ je kritická hodnota stále jen 9, ačkoli asymptoticky jest rovna 10.

Tabulky kritických hodnot k pro 5% hranici významnosti

n \ m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	—																									
2	—	—																								
3	—	—	5																							
4	—	—	6	6																						
5	—	6	6	6	6																					
6	—	7	6	6	6	6																				
7	—	8	7	6	6	6	6																			
8	—	8	7	7	6	6	6	6																		
9	—	9	8	7	7	6	6	6	6																	
10	—	10	9	8	7	7	6	6	6	6																
11	—	11	9	8	8	7	7	7	6	6	6															
12	—	11	10	9	8	7	7	7	7	6	6	6														
13	—	12	10	9	8	8	7	7	7	7	6	6	6													
14	—	13	11	10	9	8	8	7	7	7	7	6	6	6												
15	—	14	12	10	9	8	8	7	7	7	7	7	6	6	6											
16	—	14	12	11	10	9	8	8	7	7	7	7	7	7	6	6										
17	—	15	13	11	10	9	8	8	7	7	7	7	7	7	7	7	7									
18	—	16	13	12	10	9	9	8	8	7	7	7	7	7	7	7	7	7								
19	19	17	14	12	11	10	9	8	8	8	7	7	7	7	7	7	7	7	7							
20	20	17	15	13	11	10	9	9	8	8	7	7	7	7	7	7	7	7	7	7						
21	21	18	15	13	12	11	10	9	8	8	8	7	7	7	7	7	7	7	7	7	7					
22	22	19	16	14	12	11	10	9	9	8	8	8	7	7	7	7	7	7	7	7	7	7				
23	23	20	16	14	13	11	10	10	9	8	8	8	7	7	7	7	7	7	7	7	7	7	7			
24	24	20	17	15	13	12	11	10	9	9	8	8	8	7	7	7	7	7	7	7	7	7	7	7		
25	25	21	18	15	13	12	11	10	9	9	8	8	8	8	7	7	7	7	7	7	7	7	7	7	7	
26	26	22	18	16	14	12	11	10	10	9	9	8	8	8	7	7	7	7	7	7	7	7	7	7	7	7

LITERATURA

[1] *S. Rosenbaum:* Tables for a nonparametric test of location; Ann. Mat. Stat. 25 (1954) No. 1, str. 146—150.

[2] *S. S. Wilks:* Statistical prediction with special reference to the problem of tolerance limits; Ann. Mat. Stat. 13 (1942), No. 4, str. 400—409.

Резюме

ПРОСТОЙ НЕПАРАМЕТРИЧЕСКИЙ КРИТЕРИЙ РАЗНОСТИ ПОЛОЖЕНИЯ ДВУХ РАСПРЕДЕЛЕНИЙ

ЗБЫНЕК ШИДАК, ИРЖИ ВОНДРАЧЕК (Zbyněk Šidák, Jiří Vondráček)

(Поступило в редакцию 31/VIII 1956 г.)

Эта статья является двусторонним улучшением критерия Розенбаума [1].

Пусть $x_1 < \ldots < x_n$, $y_1 < \ldots < y_m$ — две упорядоченные независимые случайные выборки из непрерывных распределений $F(x)$, соответственно $G(y)$.

Нашей задачей будет проверить нулевую гипотезу о равенстве распределений $F(x)$, $G(y)$ против альтернативной гипотезы, что $G(y)$ подвинута к большим значениям.

Критерием служит $Q = R + S$, где R означает число наблюдений из первой выборки $\{x_i\}$, которые меньше, чем y_1, и S означает число наблюдений из выборки $\{y_j\}$, которые больше, чем x_n.

Выведена функция распределения этой статистики, хвост этого распределения дан в формуле (4).

Таблицы 1% и 5% критических пределов k даны для $m, n = 1, 2, \ldots, 26$. Для $m = n = 27, \ldots, 50$, 1% критический предел $k = 9$, для 5% $k = 7$. Для $m \approx n \to \infty$ асимптотическими критическими пределами являются $k = 10$ для 1% и $k = 7$ для 5% уровня значимости.

Если $Q \geqq k$, нулевая гипотеза отвергается.

Summary

A SIMPLE NONPARAMETRIC TEST OF THE DIFFERENCE OF LOCATION OF TWO POPULATIONS

ZBYNĚK ŠIDÁK, JIŘÍ VONDRÁČEK

(Received August 31, 1956.)

This paper appears to be a twosided improvement of Rosenbaum's test [1].

Let $x_1 < \ldots < x_n$; $y_1 < \ldots < y_m$ be two ordered independent random samples from populations with continuous distribution functions $F(x)$, $G(y)$ respectively. The problem is to test the null hypothesis whether these two samples come from the same population against the alternative hypothesis that $G(y)$ has slipped towards higher values in comparison with $F(x)$.

The test characteristic is $Q = R + S$, where R denotes the number of points from the sample $\{x_i\}$ which are less than y_1, and S denotes the number of points from the sample $\{y_j\}$ which are greater than x_n. The distribution of this characteristic was determined and the tail of this distribution is given by the formula (4).

Tables of 1% and 5% critical values k are given for $m, n = 1, 2, \ldots, 26$. For $m = n = 27 \ldots, 50$ the 1% critical value is $k = 9$ and the 5% one is $k = 7$, while for $m \approx n \to \infty$ the asymptotical critical values are $k = 10$ for 1% level of significance and $k = 7$ for the 5% one.

The null hypothesis is rejected if $Q \geqq k$.