A WILCOXON-STATISZTIKA KÉT MÓDOSÍTÁSÁRÓL

CSÁKI ENDRE

Bevezetés

Két minta összehasonlítására jól ismert próba a Wilcoxon-próba, mely a következő statisztikán alapul: legyen $\xi_1, \xi_2, \ldots, \xi_n$, ill. $\eta_1, \eta_2, \ldots, \eta_m$ két egymástól független statisztikai sokaságból vett minta. Tekintsük a (ξ_i, η_j) párok közül azokat, melyekre $\eta_j < \xi_i$ $(i = 1,2,\ldots,n\,;\ j = 1,2,\ldots,m)$. Ezek száma legyen U. Jelöljük a ξ változók közös eloszlásfüggvényét $F(x)$-el, az η változók közös eloszlásfüggvényét $G(x)$-el. Legyenek ezek folytonosak. A $H_0 : F(x) \equiv G(x)$ hipotézist visszautasítjuk, ha U túl kicsi, vagy túl nagy, ellenkező esetben elfogadjuk.

A Wilcoxon-próba azonban, mint már MANN és WHITNEY [4] valamint D. VAN DANTZIG [5] kimutatta, csak olyan alternatív ellenhipotézissel szemben konzisztens, melyre $p = \int G(x)\, dF(x) \neq \frac{1}{2}$, vagyis a próba nem mindig azt dönti el, hogy $F(x)$ különbözik-e $G(x)$-től, vagy nem, hanem csak azt, hogy $p = \frac{1}{2}$, vagy nem. Ezért természetes törekvés úgy módosítani a Wilcoxon-próbát, hogy az minden $F(x) \not\equiv G(x)$ hipotézissel szemben konzisztens legyen. Ilyen módosításokat LEHMANN [6] és RÉNYI [7] adott meg. Az alábbi 1. §-ban kimutatjuk, hogy a két módosítás lényegében azonos. Ez egy LEHMANN—SCHEFFÉ lemma közvetlen következménye, de közvetlenül is belátható. A 2. §-ban meghatározzuk a statisztika szórását, melyre vonatkozólag a következő egyszerű formulához jutunk:

$$\mathbf{D}^2(V) = \frac{(n+m+1)\,(n+m-2)}{45\binom{n}{2}\binom{m}{2}},$$

ahol V a LEHMANN-statisztika.

1. §. A V és W statisztika

LEHMANN a következő módosítást javasolja [6]: a (ξ_i, η_j) párok helyett tekintsük a $(\xi_i, \xi_k\,;\ \eta_j, \eta_l)$ számnégyeseket, ahol $i \neq k$, $j \neq l$. Legyen V_1 azon számnégyesek száma, melyekre $\max(\xi_i, \xi_k) < \min(\eta_j, \eta_l)$, V_2 pedig azoké, melyekre $\min(\xi_i, \xi_k) > \max(\eta_j, \eta_l)$. A továbbiakban a $\max(x_1, x_2) <$

8 A Matematikai Kutató Intézet Közleményei IV/3—4.

$< \min(y_1, y_2)$ eseményt $\{x_1, x_2 < y_1, y_2\}$-vel jelöljük. A statisztika a következő:

$$V = \frac{V_1 + V_2}{\binom{n}{2}\binom{m}{2}}. \tag{1}$$

Rényi módosítása a következő [7]: Tekintsük a (ξ_i, η_j, η_l) számhármasok közül (ahol $j \neq l$) azokat, melyekre $\eta_j < \xi_i$, $\eta_l < \xi_i$. Legyen ezeknek a száma W_1. Tekintsük a (ξ_i, ξ_k, η_j) számhármasok közül $(i \neq k)$ azokat, melyekre $\xi_i < \eta_j$, $\xi_k < \eta_j$. Ezek száma legyen W_2. A W statisztika a következő:

$$W = \frac{W_1}{n\binom{m}{2}} + \frac{W_2}{m\binom{n}{2}}. \tag{2}$$

A V és W statisztika között a következő egyszerű lineáris kapcsolat van:

$$V = 2W - 1. \tag{3}$$

Számítsuk ki először a V, ill. W valószínűségi változó várható értékét.

$$\mathbf{M}(V) = \mathbf{P}(\{\xi, \xi' < \eta, \eta'\} + \{\eta, \eta' < \xi, \xi'\}) =$$

$$= \int (1-F)^2\, dG^2 + \int (1-G)^2\, dF^2 = 2 + \int (F^2\, dG^2 + G^2\, dF^2) - 4\int FG\, d(F+G) =$$

$$= 2 + \int d(F^2 G^2) - 4\int FG\, d(F+G) =$$

$$= 3 - 2\int [(F+G)^2 - (F-G)^2]\, d\left(\frac{F+G}{2}\right) = \frac{1}{3} + 2\int (F-G)^2\, d\left(\frac{F+G}{2}\right).$$

Míg

$$\mathbf{M}(W) = \mathbf{P}(\eta, \eta' < \xi) + \mathbf{P}(\xi, \xi' < \eta) = \int F^2\, dG + \int G^2\, dF =$$

$$= \int (F+G)^2\, d(F+G) - \int F^2\, dF - \int G^2\, dG - 2\int FG\, d(F+G) =$$

$$= \frac{8}{3} - \frac{1}{3} - \frac{1}{3} - \int [(F+G)^2 - (F-G)^2]\, d\left(\frac{F+G}{2}\right) = \frac{2}{3} + \int (F-G)^2\, d\left(\frac{F+G}{2}\right).$$

Látható, hogy a várható értékekre fennáll az

$$\mathbf{M}(V) = 2\,\mathbf{M}(W) - 1$$

összefüggés. Megjegyezzük, hogy a (3) összefüggés már ebből is következik, ha felhasználjuk a következő, Lehmann—Scheffé lemmát (lásd [6]).

Legyen $f(F, G)$ valós funkcionál értelmezve minden folytonos $F(x)$, $G(x)$ eloszlásfüggvényre. Akkor legfeljebb egy olyan $t_{n,m}$ függvény létezik, hogy $t_{n,m}(\xi_1, \xi_2, \ldots, \xi_n; \eta_1, \eta_2, \ldots \eta_m)$ szimmetrikus az első n és utolsó m változójára és torzítatlan becslése $f(F, G)$-nek minden folytonos $F(x)$, $G(x)$-re. Ha egy ilyen $t_{n,m}$ függvény létezik és véges szórása van, akkor ennek $f(F, G)$ minden torzítatlan becslése között legkisebb a szórása.

E lemma első része azt mondja ki, hogy egy $f(F, G)$-nek lényegében csak egy szimmetrikus torzítatlan becslése van. Így $\mathbf{M}(V)$-nek és $\mathbf{M}(2W - 1)$-nek szimmetrikus torzítatlan becslései azonosak, azaz fennáll a (3) összefüggés.

A (3) relációt azonban közvetlenül is be lehet látni a LEHMANN—SCHEFFÉ lemma felhasználása nélkül. Azt fogjuk bizonyítani, hogy

$$V_1 + V_2 = (n-1)\,W_1 + (m-1)\,W_2 - \binom{n}{2}\binom{m}{2}. \tag{4}$$

Legyen

$$\beta(x_1, x_2; y_1, y_2) = \begin{cases} 1, & \text{ha } \{x_1, x_2 < y_1, y_2\}, \text{ vagy } \{y_1, y_2 < x_1, x_2\} \\ 0 & \text{különben}, \end{cases}$$

továbbá

$$\gamma(x\,;\, y_1, y_2) = \begin{cases} 1, & \text{ha } y_1 < x \text{ és } y_2 < x, \\ 0 & \text{különben.} \end{cases}$$

Ekkor nyilvánvalóan

$$V_1 + V_2 = \sum_{(i,k)} \sum_{(j,l)} \beta(\xi_i, \xi_k\,;\, \eta_j, \eta_l),$$

valamint

$$W_1 = \sum_{(j,l)} \sum_{i=l}^{n} \gamma(\xi_i\,;\, \eta_j, \eta_l),$$

$$W_2 = \sum_{(i,k)} \sum_{j=1}^{m} \gamma(\eta_j, \xi_i, \xi_k),$$

ahol $\sum_{(i,k)}'$, ill. $\sum_{(j,l)}'$ azt jelenti, hogy az összegezés kiterjesztendő az $1,2,\ldots,n$, ill. az $1,2,\ldots,m$ elemekből alkotott összes (i,k), ill. (j,l) elempárra, ahol $i<k$, ill. $j<l$.

Így

$$(n-1)\,W_1 + (m-1)\,W_2 = \sum_{(j,l)} (n-1) \sum_{i=1}^{n} \gamma(\xi_i;\, \eta_j, \eta_l) + \sum_{(i,k)} (m-1) \sum_{j=1}^{m} \gamma(\eta_j;\, \xi_i, \xi_k).$$

Ez a nyilvánvaló

$$(n-1) \sum_{i=1}^{n} a_i = \sum_{(i,k)}' (a_i + a_k)$$

azonosság felhasználásával a következőképpen alakítható át:

$$\sum_{(j,l)} (n-1) \sum_{i=1}^{n} \gamma(\xi_i;\, \eta_j, \eta_l) + \sum_{(i,k)} (m-1) \sum_{j=1}^{m} \gamma(\eta_j;\, \xi_i, \xi_k) =$$

$$= \sum_{(j,l)} \sum_{(i,k)} [\gamma(\xi_i;\, \eta_j, \eta_l) + \gamma(\xi_k;\, \eta_j, \eta_l)] + \sum_{(i,k)} \sum_{(j,l)} [\gamma(\eta_j;\, \xi_i, \xi_k) + \gamma(\eta_l;\, \xi_i, \xi_k)] =$$

$$= \sum_{(i,k)} \sum_{(j,l)} [\gamma(\xi_i;\, \eta_j, \eta_l) + \gamma(\xi_k;\, \eta_j, \eta_l) + \gamma(\eta_j;\, \xi_i, \xi_k) + \gamma(\eta_l;\, \xi_i, \xi_k)].$$

Fennáll a következő azonosság:

$$-1 + \gamma(\xi_i;\, \eta_j, \eta_l) + \gamma(\xi_k;\, \eta_j, \eta_l) + \gamma(\eta_j;\, \xi_i, \xi_k) + \gamma(\eta_l;\, \xi_i, \xi_k) = \beta(\xi_i, \xi_k\,;\, \eta_j, \eta_l),$$

így

$$(n-1)\,W_1 + (m-1)\,W_2 = \sum_{(i,k)} \sum_{(j,l)} [\beta(\xi_i, \xi_k\,;\, \eta_j, \eta_l) + 1] = V_1 + V_2 + \binom{n}{2}\binom{m}{2}$$

ahonnan adódik a (4) reláció.

Azt, hogy a V-statisztikán alapuló próba minden $F \not\equiv G$ ellenhipotézissel szemben konzisztens, már LEHMANN kimutatta [6] és lényegében abból következik, hogy $\mathbf{D}^2(V) \to 0$, midőn $\min(n, m) \to \infty$, valamint $\mathbf{M}(V) = \frac{1}{3}$ azaz $\int (F - G)^2 \, d\left(\frac{F+G}{2}\right) = 0$, akkor és csak akkor áll fenn, ha $F \equiv G$.

Megjegyezzük még, hogy a V, ill. a W statisztika a következőképpen is kifejezhető:

Legyen $r_1, r_2, \ldots, r_n$, ill. $s_1, s_2, \ldots, s_m$ azon helyek sorszáma, ahol a $\xi_1, \xi_2, \ldots, \xi_n, \eta_1, \eta_2, \ldots, \eta_m$ mintaelemek nagyság szerint rendezett egyesített sorozatában a $\xi_1^*, \xi_2^*, \ldots, \xi_n^*$, ill. az $\eta_1^*, \eta_2^*, \ldots, \eta_m^*$ elemek állnak.

Ekkor

$$V = \frac{\sum_{i=1}^{n} (n-i)\binom{r_i - i}{2} + \sum_{j=1}^{m} (m-j)\binom{s_j - j}{2}}{\binom{n}{2}\binom{m}{2}}, \tag{5}$$

míg

$$W = \frac{\sum_{i=1}^{n} \binom{r_i - i}{2}}{n\binom{m}{2}} + \frac{\sum_{j=1}^{m} \binom{s_j - j}{2}}{m\binom{n}{2}}. \tag{6}$$

Ebből is igazolható a (3) összefüggés, csak azt kell belátni, hogy

$$\sum_{i=1}^{n} (i-1)\binom{r_i - i}{2} + \sum_{j=1}^{m} (j-1)\binom{s_j - j}{2} = \binom{n}{2}\binom{m}{2}.$$

Ez az összefüggés pedig majdnem nyilvánvaló. Azt kell belátni ugyanis, hogy a baloldalon az összes $(\xi_i, \xi_k; \eta_j, \eta_l)$ számnégyeseknek a száma áll.

2. §. A V-statisztika szórása

Míg a Wilcoxon-statisztika eloszlása lényegében ismeretes, addig a V-statisztika eloszlásáról igen keveset tudunk. LEHMANN egy tételéből ugyan következik, hogy V aszimptotikus eloszlása normális (lásd [6]), midőn $n \to \infty$ és $m \to \infty$, úgy hogy $\frac{n}{m} = C = \text{konst.}$ függetlenül attól, hogy igaz-e a nullhipotézis, vagy nem, azonban V-nek még csak a szórása sem volt eddig ismeretes. Ebben a §-ban kiszámítjuk a V statisztika szórását, midőn $F(x) \equiv G(x)$. Megjegyezzük, hogy ezen eljárás $F(x) \not\equiv G(x)$ esetén is alkalmazható, azonban túl bonyolult formulához jutunk.

Kiindulunk abból, hogy

$$V_1 + V_2 = \sum_{(i,k)} \sum_{(j,l)} \beta(\xi_i, \xi_k; \eta_j, \eta_l),$$

így

$$\mathbf{M}[(V_1 + V_2)^2] = \mathbf{M}\Big[\Big(\sum_{(i,k)} \sum_{(j,l)} \beta(\xi_i, \xi_k; \eta_j, \eta_l)\Big)^2\Big] =$$

$$= \binom{n}{2}\binom{m}{2}\Big\{\binom{n-2}{2}\binom{m-2}{2}\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi'', \xi'''; \eta'', \eta''')] +$$

$$+ \binom{n-2}{2} 2(m-2)\,\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi'', \xi'''; \eta, \eta''')] +$$

$$+ 2(n-2)\binom{m-2}{2}\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi, \xi'''; \eta'', \eta''')] +$$

(7) $$+ 2(n-2)\,2(m-2)\,\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi, \xi'''; \eta, \eta''')] +$$

$$+ \binom{n-2}{2}\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi'', \xi'''; \eta, \eta')] +$$

$$+ \binom{m-2}{2}\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi, \xi'; \eta'', \eta''')] +$$

$$+ 2(n-2)\,\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi, \xi'''; \eta, \eta')] +$$

$$+ 2(m-2)\,\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi, \xi'; \eta, \eta''')] +$$

$$+ \mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi, \xi'; \eta, \eta')]\Big\},$$

— ahol a $\xi, \xi', \xi'', \xi''', \eta, \eta', \eta'', \eta'''$ valszínűségi változók egymástól függetlenek és azonos eloszlásúak. Egyszerűség kedvéért a továbbiakban tegyük fel, hogy a fenti változók egyenletes eloszlásúak a (0,1) intervallumban.

Könnyen látható, hogy

$$\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi'', \xi'''; \eta'', \eta''')] = \frac{1}{3}\cdot\frac{1}{3} = \frac{1}{9},$$

$$\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi'', \xi'''; \eta, \eta''')] = \mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi, \xi'''; \eta'', \eta''')] =$$

$$= \mathbf{P}(\xi, \xi' < \eta, \eta'; \xi'', \xi''' < \eta, \eta''') + 2\,\mathbf{P}(\xi, \xi' < \eta, \eta'; \xi'', \xi''' > \eta, \eta''') +$$

$$+ \mathbf{P}(\xi, \xi' > \eta, \eta'; \xi'', \xi''' > \eta, \eta''') =$$

$$= \int_0^1\int_0^1 [1 - \max(x, y)]\,(1-x)\,(1-y)\,dx^2\,dy^2 +$$

$$+ 2\int_0^1\int_0^y (y-x)\,(1-x)\,y\,dx^2\,d\,[1-(1-y)^2] +$$

$$+ \int_0^1\int_0^1 \min(x, y)\,xy\,d\,[1-(1-x)^2]\,d\,[1-(1-y)^2] = \frac{1}{9}.$$

$$\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi, \xi'''; \eta, \eta''')] = 2\,\mathbf{P}(\xi, \xi' < \eta, \eta'; \xi, \xi''' < \eta, \eta''') =$$

$$= 2\int_0^1\int_0^1\int_0^1 \min(x, y, z)\,\min(x, y)\,\min(x, z)\,dx\,dy\,dz = \frac{11}{90}.$$

$$\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi'', \xi'''; \eta, \eta')] = \mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi, \xi'; \eta'', \eta''')] =$$

$$= \mathbf{P}(\xi, \xi' < \eta, \eta'; \xi'', \xi''' < \eta, \eta') + 2\,\mathbf{P}(\xi, \xi' < \eta, \eta'; \xi'', \xi''' > \eta, \eta') +$$

$$+ \mathbf{P}(\xi, \xi' > \eta, \eta'; \xi'', \xi''' > \eta, \eta') = \int_0^1 x^4\,d\,[1 - (1-x)^2] +$$

$$+ 2\int_0^1\int_0^y (y-x)^2\,dx^2\,d\,[1-(1-y)^2] + \int_0^1 x^2\,d\,[1-(1-x)^4] = \frac{7}{45},$$

$$\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi, \xi'''; \eta, \eta')] = \mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi, \xi'; \eta, \eta''')] =$$

$$= \mathbf{P}(\xi, \xi' < \eta, \eta'; \xi, \xi''' < \eta, \eta') + \mathbf{P}(\xi, \xi' > \eta, \eta'; \xi, \xi''' > \eta, \eta') =$$

$$= \int_0^1 x^3\,d\,[1-(1-x)^2] + \int_0^1 x^2\,d\,[1-(1-x)^3] = \frac{1}{5},$$

$$\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi, \xi'; \eta, \eta')] = \mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')] = \frac{1}{3}.$$

Igy tehát

$$\mathbf{D}^2(V_1 + V_2) = \mathbf{M}[(V_1 + V_2)^2] - [\mathbf{M}(V_1 + V_2)]^2 = \binom{n}{2}\binom{m}{2}\frac{(n+m+1)(n+m-2)}{45},$$

tehát

$$\mathbf{D}^2(V) = \frac{(n+m+1)(n+m-2)}{45\binom{n}{2}\binom{m}{2}}. \tag{8}$$

Megjegyezzük még, hogy a (7) formulában szereplő várható értékek egyszerű kombinatorikus meggondolással is kiszámíthatók. A fenti módszert azonban előnyben kell részesíteni a kombinatorikussal szemben, mert ez csak abban az esetben alkalmazható, midőn $F(x) \equiv G(x)$, míg az előbbi módszerrel tetszőleges $F(x)$, $G(x)$ esetén célhoz jutunk, csak bonyolultabbak lesznek a formulák.

Határozzuk meg például kombinatorikus úton a következő várható értéket:

$$\mathbf{M}\,[\beta(\xi, \xi'; \eta, \eta')\,\beta(\xi'', \xi'''; \eta, \eta')].$$

Ez a következő (diszjunkt) események összegének valószínűsége:

$$\{\xi, \xi', \xi'', \xi''' < \eta, \eta'\};\ \{\xi, \xi' < \eta, \eta' < \xi'', \xi'''\};\ \{\xi'', \xi''' < \eta, \eta' < \xi, \xi'\};$$

$$\{\eta, \eta' < \xi, \xi', \xi'', \xi'''\}.$$

A $\xi, \xi', \xi'', \xi''', \eta, \eta'$ változók összes lehetséges sorrendjének száma 6! Az első és negyedik esemény lehetőségeinek száma 4 ! 2 ! míg a másodiké és harmadiké 2 ! 2 ! 2 !

Igy

$$\mathbf{M}[\beta(\xi, \xi'; \eta, \eta')\, \beta(\xi'', \xi'''; \eta, \eta')] = \frac{2 \cdot 4!\, 2! + 2 \cdot 2!\, 2!\, 2!}{6!} = \frac{7}{45}.$$

(Beérkezett: 1959. május 28.)

IRODALOM

[1] WILCOXON, F.: „Individual comparisons by ranking methods". *Biometrics Bulletin* **1** (1945) 80—83.
[2] WILCOXON, F.: „Individual comparisons of grouped data by ranking methods". *Journ. Econ. Entomology* **39** (1946) 269.
[3] WILCOXON, F.: „Probability tables for individual comparisons by ranking methods". *Biometrics Bulletin* **3** (1947) 119—122.
[4] MANN, H. B.—WHITNEY, D. R.: „On a test whether one of two random variables is stochastically larger than the other". *The Annals of Mathematical Statistics* **18** (1947) 50—60.
[5] VAN DANTZIG, D.: „On the consistency and the power of Wilcoxon's two sample test". *Indagationes Mathematicae* **13** (1951) 1—8.
[6] LEHMANN, E. L.: „Consistency and unbiasedness of certain nonparametric tests". *The Annals of Mathematical Statistics* **22** (1951) 165—180.
[7] RÉNYI A.: „Újabb kritériumok két minta összehasonlítására". *A Magyar Tudományos Akadémia Alkalmazott Matematikai Intézetének Közleményei* **2** (1953) 243—257.

О ДВУХ ВИДОИЗМЕНЕНИЯХ СТАТИСТИКИ WILCOXON-A

E. CSÁKI

Резюме

LEHMANN [6] и RÉNYI [7] видоизменили статистику WILCOXON-a относящуюся к сравнению двух выборок, так, что она уже конзистентна относительно всякой альтернативной контргипотезы. В настоящей работе доказывается, что между двумя изменёнными статистиками существует простая линейная связь. Вычисляется дисперсия статистики, что делает возможным применение пробы к большой выборке приближением с помощью нормального распределения.

ON TWO MODIFICATIONS OF THE WILCOXON-STATISTIC

E. CSÁKI

Summary

LEHMANN [6] and RÉNYI [7] have modified the Wilcoxon-statistic concerning the comparison of two samples, in a form which is consistent against every alternative hypothesis. This paper shows, that between the two modified statistics a simple linear relation exists. The dispersion of the statistic is determined, which gives in case of large sample the possibility to use this test by approximation with normal distribution.