Jaroslav Hájek

O rozdělení některých statistik za přítomnosti vnitrotřídní korelace

Časopis pro pěstování matematiky, Vol. 83 (1958), No. 3, 327–329

Persistent URL: `http://dml.cz/dmlcz/108296`

Časopis pro pěstování matematiky, roč. 83 (1958), Praha

O ROZDĚLENÍ NĚKTERÝCH STATISTIK ZA PŘÍTOMNOSTI VNITROTŘÍDNÍ KORELACE

JAROSLAV HÁJEK, Praha

(Došlo dne 6. září 1957)

DT: 519.271
519.272.2

Velmi jednoduchým způsobem je dokázáno, že přítomnost vnitrotřídní korelace nemá vliv na rozdělení některých důležitých statistik. Dosažený výsledek v sobě zahrnuje výsledek, ke kterému dospěl M. Halperin v práci [1].

Věta. *Mějme normální náhodné veličiny $y_1, y_2, \ldots, y_n$ s libovolnými středními hodnotami. Tvrdíme, že následující dvě hypotesy* H_1 a H_2 *o rozptylu a korelačních koeficientech*

$$H_1: \quad Dy_i = \sigma^2\,, \quad i, j = 1, \ldots, n\,; \qquad \varrho_{ij} = \varrho\,, \quad i \neq j\,,\ -\frac{1}{n-1} < \varrho < 1\,; \tag{1}$$

$$H_2: \quad Dy_i = (1 - \varrho)\,\sigma^2\,, \quad i, j = 1, \ldots, n\,; \qquad \varrho_{ij} = 0\,, \quad i \neq j\,, \tag{2}$$

vedou k témuž rozdělení u vektoru $(y_1 - \bar{y}, \ldots, y_n - \bar{y})$, kde $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, a tudíž i u všech statistik tvaru $s = s(y_1 - \bar{y}, \ldots, y_n - \bar{y})$; kromě toho, při obou hypotesách je výběrový průměr $\bar{y}$ nezávislý na vektoru $(y_1 - \bar{y}, \ldots, y_n - \bar{y})$.

Důkaz. Nechť normální náhodné veličiny $z_1, \ldots, z_n$ se chovají podle hypotesy H_2, tj. jsou nezávislé na rozptylu $(1 - \varrho)\,\sigma^2$. Přímým výpočtem se snadno dokáže, že potom náhodné veličiny

$$y_i = z_i - \frac{1}{n}\left(1 - \sqrt{\frac{1 + (n-1)\,\varrho}{1 - \varrho}}\right) \sum_{i=1}^{n} (z_i - Mz_i) \tag{3}$$

se chovají podle hypotesy H_1, tj. mají rozptyl σ^2 a korelační koeficienty $\varrho_{ij} = \varrho$; kromě toho střední hodnoty jsou stejné, $My_i = Mz_i$. Z (3) plyne, že $y_i - \bar{y} = z_i - \bar{z}$, takže $(y_1 - \bar{y}, \ldots, y_n - \bar{y}) = (z_1 - \bar{z}, \ldots, z_n - \bar{z})$, odkud plyne první tvrzení věty. Dále z (3) a z $My_i = Mz_i$ vyplývá, že

$$\bar{y} - M\bar{y} = \sqrt{\frac{1 + (n-1)\,\varrho}{1 - \varrho}}\,(\bar{z} - M\bar{z})\,, \tag{4}$$

takže nezávislost $\bar{z}$ a $(z_i - \bar{z}, \ldots, z_n - \bar{z})$, jež je dobře známa, implikuje nezávislost $\bar{y}$ a $(y_1 - \bar{y}, \ldots, y_n - \bar{y})$. Tím je důkaz proveden.

Věta umožňuje převést výsledky vypracované pro obvyklou hypotesu H_2, znamenající nezávislost, na poněkud obecnější případ popsaný hypotesou H_1, znamenající přítomnost vnitrotřídní korelace, tj. korelace konstantní pro všechny páry náhodných veličin.

Příklad 1. Příkladem statistik tvaru $s = s(y_1 - \bar{y}, \ldots, y_n - \bar{y})$ jsou nejlepší lineární odhady b_j regresních koeficientů β_j ve vztahu

$$My_i = \mu + \sum_{j=1}^{k} \beta_j(x_j - \bar{x}) , \tag{5}$$

dále residuální rozptyl $\sum_{i=1}^{n}(y_i - \bar{y} - \Sigma b_j(x_j - \bar{x}))^2$, dále vektor, jehož komponentami jsou právě jmenované statistiky apod. Přítomnost vnitrotřídní korelace má na tyto a jim podobné statistiky stejný účinek jako změna rozptylu z σ^2 na $(1 - \varrho)\, \sigma^2$ při ponechání předpokladu nezávislosti.

Příklad 2. Při náhodném výběru n elementů z N elementů bez opakování je $\varrho = - \dfrac{1}{N - 1}$. Za předpokladu, že rozdělení pozorování je zhruba normální, z naší věty vyplývá, že Studentův poměr

$$t = \frac{\bar{y} - \bar{\bar{y}}}{\sqrt{\dfrac{\left(1 - \dfrac{n}{N}\right)}{n(n - 1)} \sum_{i=1}^{n} (y_i - \bar{y})^2}} ,$$

kde $\bar{y}$ resp. $\bar{\bar{y}}$ je průměr hodnot na vybraných resp. na všech elementech, má zhruba Studentovo rozdělení s $n - 1$ stupni a nikoliv s $\dfrac{n - 1}{1 - \dfrac{n}{N}}$, jak je doporučováno v Cramérově knize [2] na str. 523. K témuž závěru dospějemo, předpokládáme-li, že hodnoty $y_1, \ldots, y_N$ tvoří náhodný výběr z normálního rozdělení a hodnoty $y_1, \ldots, y_n$ část tohoto výběru.

Poznámka. Representace (3), která je poněkud jednodušší než ta, kterou použil J. E. Walsh v [3], nám dovoluje ihned najít hustotu pozorování $(y_1, \ldots, y_n)$ při platnosti hypotesy H_1. Skutečně hustota vektoru $(z_1, \ldots, z_n)$ je rovna

$$K_1 e^{-\frac{1}{2} \sum_{i=1}^{n} \frac{(z_i - Mz_i)^2}{(1-\varrho)\sigma^2}} = K_1 e^{-\frac{1}{2} \left[\sum_{i=1}^{n} \frac{(z_i - Mz_i - \bar{z} + M\bar{z})^2}{(1-\varrho)\sigma^2} + \frac{n(\bar{z} - M\bar{z})^2}{(1-\varrho)\sigma^2} \right]} .$$

Lineární transformací (3), z které plyne $z_i - Mz_i - \bar{z} + M\bar{z} = y_i - My_i - \bar{y} + M\bar{y}$ a (4), dostáváme hustotu $(y_1, \ldots, y_n)$ ve tvaru

$$K_2 e^{-\frac{1}{2} \left[\sum_{i=1}^{n} \frac{(y_i - y - My_i + M\bar{y})^2}{(1-\varrho)\sigma^2} + \frac{n}{\sigma^2[1 + (n-1)\varrho]} (\bar{y} - M\bar{y})^2 \right]} . \tag{6}$$

Determinant kovariační matice při hypotese H_1 je roven

$$\sigma^{2n}[1 + (n - 1)\,\varrho]\,(1 - \varrho)^{n-1}\,,$$

takže

$$K_2 = (2\pi)^{-\frac{n}{2}}\sigma^{-n}(1 - \varrho)^{-\frac{n-1}{2}}\,[1 + (n - 1)\,\varrho]^{-\frac{1}{2}}\,.$$

Z hustoty (6) vyplývá, že maximálně věrohodné odhady regresních koeficientů v (5) jsou tytéž při obou hypotesách H_1 i H_2. Nemá tedy Halperin pravdu, když v [1] tvrdí, že nalezení maximálně věrohodných odhadů při hypotese H_1 je neúměrně složité. Ve skutečnosti odhady (3.3) uvedené v jeho práci [1] nejsou ničím jiným než maximálně věrohodnými odhady.

LITERATURA

[1] *M. Halperin:* „Normal regression theory in the presence of intra-class correlation", Ann. Math. Statist. *22* (1951), 573—580.

[2] *H. Cramér:* „Mathematical methods of statistics", Princeton University Press, 1946.

[3] *J. E. Walsh:* „Concerning the effect of intraclass correlation on certain significance tests". Ann. Math. Statist. *18* (1947), 88—96.

Резюме

О РАСПРЕДЕЛЕНИИ НЕКОТОРЫХ СТАТИСТИК ПРИ НАЛИЧИИ МЕЖДУКЛАССОВОЙ КОРРЕЛЯЦИИ

ЯРОСЛАВ ГАЕК (Jaroslav Hájek), Прага
(Поступило в редакцию 6/IX 1957 г.)

В статье доказывается, что n-мерному нормальному распределению с постоянной дисперсией σ^2 и с коэффициентом корреляции ϱ соответсвует плотность вероятности (6). Из вида последней следует, что максимально правдоподобные оценки коэффициентов регрессии β_j в соотношении $My_i = a + \sum\beta_j(x_{ji} - \bar{x}_j)$ не зависят от ϱ. Одновременно выводятся некоторые результаты, содержащиеся в [1] и [3].

Summary

ON THE DISTRIBUTION OF SOME STATISTICS IN THE PRESENCE OF INTRACLASS CORRELATION

JAROSLAV HÁJEK, Praha
(Received September 6, 1957)

The n-dimensional normal distribution with constant variance σ^2 and correlation coefficient ϱ is proved to possess the probability density (6). From its form it is easily seen that the maximal-likelihood estimates of the regression coefficients β_j in the relation $My_i = a + \Sigma\beta_j(x_{ji} - \bar{x}_j)$ are independent of ϱ. Some results contained in [1] a [3] are derived in addition.