Jaroslav Janko

K teorii representativní metody

*Časopis pro pěstování matematiky a fysiky*, Vol. 57 (1928), No. 3-4, 286--290

Persistent URL: http://dml.cz/dmlcz/121372

# K teorii representativní metody.

Dr. *Jaroslav Janko.*

Potřeby a tím úkoly kladené na statistickou službu v nynější době jsou stále rozsáhlejší a podrobnější; rozšiřují se jednak na nové obory, jednak jdou ve všech oborech daleko více do hloubky, než tomu bývalo. Naproti tomu jsou prostředky k provádění statistických šetření věnované značně omezeny. Směřují tudíž snahy statistiků k ekonomisaci, která se projevuje dosud hlavně v racionalisaci zpracování materiálu, ale bude se musiti jeviti také v racionalisaci metod šetření. Proto musí obraceti stále intensivněji statistikové zřetel k parciálním šetřením statistickým, při nichž se nevyšetřuje celý soubor, o němž chceme získati vědomosti, nýbrž jen některá část. Aby pak bylo možno zevšeobecniti výsledky získané částečným šetřením, je nutno opatřiti si representativní soubor, což se děje buď metodou náhodného výběru, při níž se vyjme nějaký počet jednotek pro šetření podle principu mechanického nebo jiného, který není nijak v souvislosti s předmětem a účelem šetření, a výběr je tak prováděn, že každá jednotka celkového souboru má stejnou pravděpodobnost býti do něho pojata, nebo metodou účelového výběru, při níž se vybere nějaký počet skupin jednotek tak, že dávají co možná nejpřibližněji tytéž středy nebo poměry jako celý soubor vzhledem k těm charakteristikám, které jsou již známy. Representativní metoda je jedním z nejobyčejnějších prostředků chápání nejen ve vědeckém badání, ale i v životě obecném. Prakticky tvoříme si celý svůj obraz světa pomocí řady částečných pozorování a jejich generalisováním dospíváme k výsledkům, které považujeme v určitých mezích za všeobecně platné. Správnost těchto výsledků záleží velmi podstatně na náležitém hodnocení representativního charakteru pozorování. Z denního života byla nepozorovaně pak přenesena aplikace této metody do mnohých oborů vědeckého badání.

První otázka, kterou tu statistikové řešili, směřovala k tomu, jak dalece bude náhodný výběr representovati daný původní soubor. Praktická otázka však zní obráceně, totiž co můžeme souditi o původním souboru z daného náhodného výběru. Při jejím řešení se postupovalo tak, že byla hledána nejprve směrodatná odchylka rozdílu mezi pravou hodnotou a pozorovanou, pak

stanovena pravděpodobnost, že může vzniknouti určitá úchylka a konečně aplikován princip inversní pravděpodobnosti.*) Je-li $X'$ daná funkce $n$ náhodných výběrů a neznámá jí odpovídající funkce v původním souboru je $X$, pak můžeme psáti

$$X = X' + x.$$

Můžeme-li dokázati, že pravděpodobnost hodnoty $X'$, když hodnota původního souboru je $X$, je tvaru

$$P_x = P_0 e^{-\frac{1}{2}\frac{x^2}{\sigma^2}},$$

kde $P_0$ je maximální pravděpodobnost, kterou dostáváme je-li $X = X'$, pak můžeme tvrditi, že $X'$ daná výběrem je nejpravděpodobnější hodnotou hledané funkce a že pravděpodobnost odchylky od $X'$ je dána Gaussovou funkcí o směrodatné odchylce $\sigma$. V obecnějším případě proces inverse není tak přímý. Jsou-li všechny hodnoty $X$ a priori stejně pravděpodobné, pak pravděpodobnost, že náhodný výběr je z původního souboru, byla-li hodnota $X$ v mezích $X' \pm x$, jest

$$2\int_0^x P_x\,dx,$$

je-li $x$ malé; a inversní pravděpodobností dostáváme pravděpodobnost, že hodnota původního souboru byla v těchto právě uvedených mezích

$$\frac{2\int_0^x P_x\,dx}{\int_{-\infty}^{+\infty} P_x\,dx} = \frac{2}{\sigma\sqrt{2\pi}}\int_0^x e^{-\frac{1}{2}\frac{x^2}{\sigma^2}}\,dx. \qquad (1)$$

Tímto postupem budeme nyní řešiti problém, do jaké míry lze souditi, že původní soubory dvou náhodných výběrů jevících stejné pravděpodobnosti příznivého případu mají tytéž vlastnosti.

Jsou-li tedy dány dva náhodné výběry, jest třeba nejprve se přesvědčiti, zda jsou na sobě nezávislé a zda byly podmínky, podle nichž se řídí dostavení se pozorované vlastnosti, tytéž pro každý výběr pozorovaných případů, čili zda jsou splněny podmínky normální stability. Tyto jsou, jak známo, splněny tehdy, rovná-li se divergenční koeficient Lexis-Bortkiewiczův jedničce. Budiž rozsah výběrů $n_1$ resp. $n_2$ jednotek. Počet příznivých případů pak je $r_1$ resp. $r_2$. Máme tedy pro pravděpodobnosti příznivého výsledku v těchto náhodných výběrech rovnice

$$n_1 p_1 = r_1, \quad n_2 p_2 = r_2.$$

Příslušné pravděpodobnosti příznivého výsledku v původních souborech buďtež $p_0$ resp. $p'_0$; označme

*) Bowley: Elements of Statistics, str. 409 a násl.

$$\left.\begin{aligned} |\, p_0 - p_1 \,| &= \Delta p_0 \\ |\, p'_0 - p_2 \,| &= \Delta p'_0 \end{aligned}\right\} \qquad (2)$$

Podle rovnice (1) je pak pravděpodobnost, že původní soubor prvního náhodného výběru dává příznivý výsledek s pravděpodobností $p_0$, dána výrazem

$$\frac{2}{\sigma_1\sqrt{2\pi}} \int_0^{\Delta p_0} e^{-\frac{1}{2}\left(\frac{\Delta p_0}{\sigma_1}\right)^2} dp_0.$$

Při tom je

$$\sigma_1^2 = \frac{p_1(1-p_1)}{n_1} + \frac{2p_1(1-p_1)}{n_1^2}(r_{12} + r_{13} + \ldots . r_{23} + \ldots .),$$

kde $r_{12}$, $r_{13}$, ... jsou koeficienty korelace mezi výsledky jednotlivých událostí ve výběru, v němž nejsou události na sobě nezávislé.[2]) Kdyby byly všechny události náhodného výběru na sobě nezávislé, rovnaly by se všechny koeficienty korelace nule a směrodatná odchylka by byla dána prvním členem.

Obdobně platí pro druhý náhodný výběr

$$\frac{2}{\sigma_2\sqrt{2\pi}} \int_0^{\Delta p'_0} e^{-\frac{1}{2}\left(\frac{\Delta p'_0}{\sigma_2}\right)^2} dp'_0,$$

kde

$$\sigma_2^2 = \frac{p_2(1-p_2)}{n_2} + \frac{2p_2(1-p_2)}{n_2^2}(r'_{12} + r'_{13} + \ldots + r'_{23} + \ldots).$$

Případy, v nichž $r_{i,k}$ nebo $r'_{i,k}$ je kladné, zahrnují, jak známo[3]) odchylky od pravidel prostého náhodného výběru vzniklé tím, že je tu podstatný rozdíl mezi místy nebo okolnostmi, v nichž nebo za nichž se výběry pozorování konají nebo že nastala nějaká podstatná změna během periody, po kterou výběr pozorování se dál. Bereme-li totiž výběry postupně z různých souborů, zavádí se tím ihned kladná korelace, i kdyby výsledky událostí při každém pokusu byly na sobě úplně nezávislé. V případech pak, kdy není pravděpodobnost pro každou událost při každém pokusu táž, nebo že nejsou stejné okolnosti, jež upravují dostavení se pozorovaného znaku pro každý jednotlivý případ nebo pro každou dílčí skupinu případů v každém souboru, z něhož výběry jsou vzaty, jest $r_{m,n}$ resp. $r'_{m,n}$ záporné, neboť je-li pravděpodobnost příznivého výsledku pro některý případ nad průměrnou hodnotou, musí býti průměrná pravděpodobnost příznivého výsledku pro zbytek pod touto hodnotou.

Pravděpodobnost, že naše dva náhodné výběry jsou ze souborů o pravděpodobnostech příznivého výsledku $p_0$ a $p'_0$, pak jest

$$\frac{2}{\pi\sigma_1\sigma_2} \int_0^{\Delta p_0} \int_0^{\Delta p'_0} e^{-\frac{1}{2}\left(\left[\frac{\Delta p_0}{\sigma_1}\right]^2 + \left[\frac{\Delta p'_0}{\sigma_2}\right]^2\right)} dp_0\, dp'_0. \qquad (3)$$

[2]) G. U. Yule: Úvod do theorie statistiky, český překlad str. 298.
[3]) Viz na př. Yule: Úvod do theorie statistiky, český překlad str. 298.

Poněvadž cílem našich úvah je posouditi, že tyto soubory jsou stejné nebo prakticky skoro stejné, čili jak se sobě blíží $p_0$ a $p'_0$, budeme uvažovati jejich rozdíl

$$u = p_0 - p'_0. \tag{4}$$

Budeme tedy hledati nejprve pravděpodobnost, že rozdíl

$$\Delta p_0 - \Delta p'_0 - \delta \tag{5}$$

je v intervalu $\delta$ a $\delta + d\delta$. Můžeme dáti první úchylce $\Delta p_0$ určitou hodnotu; pak bude $\Delta p'_0$ obsaženo v mezích $\Delta p_0 - \delta$ a $\Delta p_0 - - \delta \pm d\delta$ s pravděpodobností

$$\frac{2}{\sigma_2\sqrt{2\pi}}\, e^{-\frac{1}{2}\frac{(\Delta p_0 - \delta)^2}{\sigma_2^2}}$$

a příslušná pravděpodobnost jest

$$\frac{2}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{1}{2}\frac{(\Delta p_0)^2}{\sigma_1^2}}.$$

Aplikujeme-li nyní princip složené pravděpodobnosti pro $\Delta p_0$ v intervalu od $-\infty$ do $+\infty$, dostáváme

$$\frac{2}{\sigma_1\sigma_2\pi}\int_{\infty}^{+\infty} e^{-\frac{1}{2}\frac{(\Delta p_0 - \delta)^2}{\sigma_2^2} - \frac{1}{2}\frac{(\Delta p_0)^2}{\sigma_1^2}}\, d(\Delta p_0).$$

Tento integrál je tvaru

$$\int_{-\infty}^{+\infty} e^{-(ax^2 + 2bx + c)}\, dx,$$

kde

$$a = \frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}$$

$$b = -\frac{\delta}{2\sigma_2^2}$$

$$c = \frac{\delta^2}{2\sigma_2^2}.$$

Dosadíme-li

$$x + \frac{b}{a} = \frac{t}{\sqrt{a}},$$

pak exponent jest

$$t^2 + \frac{ac - b^2}{a},$$

takže

$$e^{-\frac{ac - b^2}{a}} \int_{-\infty}^{+\infty} \frac{e^{-t^2}}{\sqrt{a}}\, dt = \sqrt{\frac{\pi}{a}}\; e^{-\frac{ac - b^2}{a}}$$

a dosadíme-li za $a$, $b$, $c$ jejich hodnoty, dostaneme snadno

$$\frac{2}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}}\, e^{-\frac{1}{2}\frac{\delta^2}{\sigma_1^2 + \sigma_2^2}}$$

Píšeme-li $\sigma_1^2 + \sigma_2^2 = \sigma^2$, dostáváme konečně, vzhledem k rovnicím (2), (4), (5)[1]

$$\frac{2}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\frac{(u-p_1+p_2)^2}{\sigma^2}}$$

Můžeme tedy užíti přímo tabulky Bertrandovy pro funkci

$$\Theta(\lambda) = \frac{2}{\sqrt{\pi}} \int_0^{\lambda} e^{-\lambda^2} d\lambda.$$

Prakticky pak můžeme zůstati při hodnotě $\Theta(\lambda) = 1 - 10^{-3}$, jíž odpovídá $\lambda = 2{\cdot}3$. Tu vidíme, že dávají-li dva náhodné výběry stejnou pravděpodobnost příznivého případu, pak máme prakticky jistotu, že rozdíl mezi soubory, z nichž byly tyto náhodné výběry vzaty, je menší než $2{\cdot}3\,\sigma$. Čtverec této směrodatné odchylky se rovná součtu čtverců směrodatných odchylek obou náhodných výběrů. Vrátíme-li se k jejich případné závislosti na korelaci mezi výsledky jednotlivých událostí a zvolíme-li si $r$ jako aritmetický průměr těchto korelačních měr, pak

$$\sigma_1^2 = \frac{p_1(1-p_1)}{n_1}[1 + r(n_1 + 1)]$$

a obdobně pro $\sigma_2^2$, takže vliv průměrné korelace může býti značný, ježto $\sigma_1$ může klesnouti na nulu nebo vzrůsti na $\sqrt{p_1(1-p_1)}$ a podobně $\sigma_2$, z čehož tedy je patrný interval, v němž se může $\sigma$ pohybovati.

•

## Sur la théorie de la méthode représentative.

(Extrait de l'article précédent.)

Puisqu'il faut appliquer de plus en plus la méthode représentative à la statistique, beaucoup de problèmes se présentent dont il faut donner la solution théorique. Dans le présent article, nous essayons de résoudre une de ces questions: à savoir jusqu'à quel point on peut conclure que des ensembles, d'où on a fait deux choix fortuits présentant des probabilités égales pour un événement favorable, possèdent les mêmes propriétés. C'est par l'application de la probabilité inverse qu'on obtient le résultat suivant: on peut juger que la différence entre les ensembles dont on a fait des choix fortuits de même probabilité pour l'événement favorable, est inférieure à $2{\cdot}3\,\sigma$. En même temps, on fait une évaluation de l'influence exercée sur l'écart quadratique $\sigma$ dans les cas où les écarts quadratiques des choix individuels dépendent de la corrélation qui existe entre les résultats des événements respectifs.