ANNALES
UNIVERSITATIS MARIAE CURIE-SKŁODOWSKA
LUBLIN—POLONIA
VOL. IV, 10 SECTIO A 1950

Z Zakładu Statystyki Matematycznej Wydziału Matematyczno-Przyrodniczego U.M.C.S.
Kierownik: z. prof. dr M. Olekiewicz

M. OLEKIEWICZ

Determining number of independent observations n', equivalent to n observations that are not independently obtained

Wyznaczenie liczby spostrzeżeń niezależnych równoważnej n spostrzeżeniom otrzymanym w sposób zależny

As long as all observations on a random variable are independently obtained, that is, are drawn by simple sampling from the same population, the amount of information contained in samples is proportional to their sizes. When, however, the observations are not independently obtained, as for instance, when there are b individuals chosen at random, each measured k times on trait x, so that altogether there are $n=kb$ observations in the sample, then in order to be able to compare such a sample with other samples consisting of individuals measured but once, it seems indicated to determine the number of independent observations n', that could be considered equivalent to n dependently obtained observations on x.

To determine n' we shall write the expression for sampling variance of mean of x:

$$\sigma_{\bar{x}}^2 = \frac{\sigma_m^2}{b} \qquad [1]$$

where σ_m^2 is population variance of individual means, m_i' s, given by

$$m_i = \bar{x}_i = \frac{1}{k} \sum_j^k x_{ij}, \qquad i = 1, 2, \ldots b \qquad [2]$$

The sampling variance of $\bar{x}$ based on independent observations would be

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n'} \qquad [3]$$

where σ_x^2 is population variance of x.

Now, if we equate [1] and [3], we shall be able to determine n'. Thus from

$$\frac{\sigma_m^2}{b} = \frac{\sigma_x^2}{n'} \qquad [4]$$

we obtain

$$n' = \frac{b\sigma_x^2}{\sigma_m^2} \qquad [5]$$

To evaluate this expression we note that

$$\sigma_x^2 = \sigma_\infty^2 + \sigma_e^2 \ , \qquad [6]$$

where σ_∞^2 is population variance „between individuals", while σ_e^2 is population variance „within individuals".

On the other hand we have

$$\sigma_m^2 = \sigma_\infty^2 + \frac{\sigma_e^2}{k} \ . \qquad [7]$$

Solving for σ_∞^2 and substituting in [6] and then in [5], we obtain

$$n' = b + \frac{b(k-1)\,\sigma_e^2}{k\,\sigma_m^2} = b + \frac{b(n-b)\,\sigma_e^2}{n\,\sigma_m^2} \qquad [8]$$

which expresses n' in terms of readily estimable parameters:

$$\left.\begin{aligned} \hat{\sigma}_e^2 &= \frac{ns_e^2}{n-b} \\ \hat{\sigma}_m^2 &= \frac{bs_m^2}{b-1} \end{aligned}\right| \qquad [9]$$

where $s_e^2 = \frac{1}{n}\sum_i^b \sum_j^k (x_{ij} - \bar{x}_i)^2$, and $s_m^2 = \frac{1}{b}\sum_i^b (\bar{x}_i - \bar{\bar{x}}^2)$.

Now n' can be estimated by n'^*

$$n'^* = b + \frac{(b-1)\,ns_e^2}{ns_m^2} \ . \qquad [10]$$

With the aid of identity

$$ns^2 = ns_m^2 + ns_e^2 \ , \qquad [11]$$

where $s^2 = \frac{1}{n}\sum_{i,j}^{n}(x_{ij} - \bar{x})^2$,

the formula [10] can be put into a more convenient form for calculation:

$$n'^{*} = \frac{(b-1)\, ns^2}{ns_m^2} + 1 \,. \qquad [12]$$

In cases when different individuals have (not too excessively) varying numbers of measurements, n_i's, the formula can be used as an approximation, with s_m^2 defined as

$$s_m^2 = \frac{1}{n} \sum_i^b n_i (\bar{x}_i - \bar{x})^2 \,. \qquad [13]$$

Since dependently obtained observations may turn out to be statistically independent, it is desirable to determine limit of significance for n'^{*} on hypothesis of independent observations.

By simple transformation [10] can be written in the following form:

$$n'^{*} = b + \frac{n-b}{F_m^0} \qquad [14]$$

where

$$F_m^0 = \frac{ns_m^2}{b-1} \div \frac{ns_e^2}{n-b} \qquad [15]$$

It can be seen that F_m^0 is the well known „variance ratio", F, defined on null hypothesis and used in tests of significance in connection with the analysis of variance. The critical point for n'^{*} will be obtained by putting in [14]

$$F_m^0 = F_p \,, \; \nu_1 = b-1 \,, \; \nu_2 = n-b \qquad [16]$$

where F_p is $100\,\mathrm{P}\%$ point to be read off from F table with the indicated d.f. (by the nature of our assumptions the critical region will consist of one tail only). If, when calculated by formula [12], n'^{*} is less than its critical point, the observations can be considered significantly dependent, and, for appraising amount of information contained in the sample, n'^{*} is to be used rather than n. If n'^{*} exceeds its critical point, n may be used, provided that error of the 2nd kind with the admitted tolerance be smaller than error of the 1st kind.

The substitution of n'^{*} for n serves only as a correction for gross error. As a random variable, n'^{*} is a linear function of F, on the assumption of independency:

$$n'^{*} = b + (n-b)\, F_e^0 \qquad [17]$$

where $F_e^0 = \dfrac{1}{F_m^0}$

is distributed as F with $\nu_1 = n-b \quad \nu_2 = b-1$.

The confidence limits for n' can be determined by defining F_m without assumption of independency.

$$F_m = \frac{\dfrac{b\, s_m^2}{\sigma_m^2} \Big/ (b-1)}{\dfrac{n s_e^2}{\sigma_e^2} \Big/ (n-b)} = \frac{\sigma_e^2}{\sigma_m^2} \cdot \frac{b\,(n-b)\, s_m^2}{n\,(b-1)\, s_e^2} \qquad [18]$$

Since by [8] $\dfrac{\sigma_e^2}{\sigma_m^2} = \dfrac{n(n'-b)}{b(n-b)}$, we shall have the following confidence interval for n':

$$b + F_1\,(n'^{*} - b) \leqslant n' \leqslant b + F_2\,(n'^{*} - b) \qquad [19]$$

where F_1 and F_2 are critical points taken from F table at a chosen level of confidence. The upper limit, of course, cannot exceed n. The interval will be the shorter, the closer to b is n'^{*}.

The use of the above transformation is not limited to comparing amounts of information. It is applicable, e. g., in testing significance of correlation by means of Student's ratio, when both variables are equally reliable:

$$|t| = \frac{|r|}{\sqrt{1-r^2}} \cdot \sqrt{n-2} \ , \quad \nu = n-2 \qquad [20]$$

where n is to be substituted by n'^{*} in cases of dependent observations.

When it is wished to estimate standard error of regression coefficient, b_{yx}, the formula

$$\sigma^{*}_{b_{yx}} = \frac{s_{y\cdot x}}{s_x \sqrt{n-2}} \ , \qquad [21]$$

where

$$s_{y\cdot x} = s_y \sqrt{1-r_{xy}^2} = \sqrt{s_y^2 - b_{yx}^2 s_x^2} \qquad [22]$$

s_y and s_x being standard deviations of y and x, and r_{xy} coefficient of correlation, can be valid only when n observations of $y.x$ are independent.

When n observations ar not independently obtained, what we need is to estimate number of independent observations on $y \cdot x$, i.e on y with fixed values of x, n', to be substituted for n.

In this case, to determine n' we shall equate two expressions for sampling variance of mean of $y \cdot x$:

$$\sigma^2_{\bar{y}\cdot x} = \frac{\sigma^2_{m\cdot x}}{b} = \frac{\sigma^2_{y\cdot x}}{n'} \qquad [23]$$

where

$$\left.\begin{aligned} \sigma^2_{y\cdot x} &= \sigma^2_y\,(1-\varrho^2_{xy}) \\ \sigma^2_{m\cdot x} &= \sigma^2_m\,(1-\varrho^2_{xm}) \\ m_i &= \bar{y}_i\ ,\quad i = 1,2,\ldots b, \end{aligned}\right\} \qquad [24]$$

the coefficient of correlation between m and x being given by

$$\varrho_{xm} = \frac{\varrho_{xy}\,\sigma_y}{\sigma_m}\ . \qquad [25]$$

Substituting [25] and [24] into [23], and solving for n' we obtain

$$n' = \frac{b\,(\sigma^2_y - \sigma^2_y\,\varrho^2_{xy})}{\sigma^2_m - \sigma^2_y\,\varrho^2_{xy}} \qquad [26]$$

Now, if regression is linear, i. e., if deviations of means of y's for given values of x from regression line are not significant, the population variance „between columns" (say, σ^2_a), can be considered equal to the variance of regression values ($\sigma^2_y\,\varrho^2_{yx}$).

On this assumption we can substitute σ^2_a for $\sigma^2_y\,\varrho^2_{xy}$ in [26]:

$$n' = \frac{b\,(\sigma^2_y - \sigma^2_a)}{\sigma^2_m - \sigma^2_a} \qquad [27]$$

Applying [6] and [7], we have

$$n' = \frac{n\,(\sigma^2_\infty - \sigma^2_a + \sigma^2_e)}{k\,(\sigma^2_\infty - \sigma^2_a) + \sigma^2_e} \qquad [28]$$

where σ^2_∞ and σ^2_e refer to variable y.

But $\sigma^2_\infty - \sigma^2_a$ is variance „between individuals freed from the effect of variation „between columns". Denoting this by σ^2_b, we obtain

$$n' = \frac{n\,(\sigma^2_b + \sigma^2_e)}{k\,\sigma^2_b + \sigma^2_e} \qquad [29]$$

which expresses n' in terms of readily estimable parameters.

The parameter σ^2_b can be estimated by

$$\hat{\sigma}_b^2 = \frac{b}{b-a}\left(\frac{ns_b^2 - (b-a)\,\hat{\sigma}_e^2}{n}\right), \qquad [30]$$

where $s_b^2 = s_m^2 - s_a^2$, $s_a^2 = \frac{1}{n}\sum_h^a n_h(\bar{y}_h - \bar{y})^2$, a being number of columns in regression table, and $\hat{\sigma}_e^2$ given in [9], but referring now to variable y. Substituting these estimates in [29], we obtain

$$n'^* = b + \frac{(b-a)\,ns_e^2}{ns_b^2} \qquad [31]$$

Making use of identity

$$ns^2 = ns_a^2 + ns_b^2 + ns_e^2, \qquad [32]$$

a more convenient formula can be written:

$$n'^* = a + \frac{(b-a)\,(ns^2 - ns_m^2)}{ns_m^2 - ns_a^2}. \qquad [33]$$

To determine limit of significance for n'^* on hypothesis of independent observations of $y \cdot x$ we shall write [31] in the form:

$$n'^* = b + \frac{n-b}{F_b^0}, \qquad [34]$$

where

$$F_b^0 = \frac{ns_b^2}{b-a} \div \frac{ns_e^2}{n-b} \qquad [35]$$

is hypothetical F with d. f. $\nu_1 = b-a$, $\nu_2 = n-b$.

The limit of significance for n'^* will be found thus by putting in [34]

$$F_b^0 = F_P, \quad \nu_1 = b-a, \quad \nu_2 = n-b \qquad [36]$$

where F_P is $100\,\mathrm{P}\%$ point to be read off from F table with the indicated d. f.

To determine confidence interval for n' we shall define F_b without assumption of independency:

$$F_b = \frac{b\,s_b^2}{(b-a)\left(\sigma_b^2 + \frac{\sigma_e^2}{k}\right)} \div \frac{ns_e^2}{(n-b)\,\sigma_e^2} = \frac{\sigma_e^2}{k\,\sigma_b^2 + \sigma_e^2} \cdot \frac{(n-b)\,s_b^2}{(b-a)\,s_e^2} \qquad [37]$$

Since by [29] $\frac{\sigma_e^2}{n\,(k-1)} = \frac{n'k-n}{k\sigma_b^2+\sigma_e^2}$, we find using [35] and [34]

$$b + F_1(n'^* - b) \leqslant n' \leqslant b + F_2(n'^* - b), \quad n' \leqslant n \qquad [38]$$

which is of the same form as [19], and differs only in d. f. of random variable F which are now: $\nu_1 = b - a$, $\nu_2 = n - b$, for confidence interval for n' at a chosen level of confidence.

Streszczenie

Gdy spostrzeżenia zmiennej losowej otrzymuje się w sposób niezależny (to znaczy że losuje się je metodą prostej próby z tej samej populacji), ilość informacji zawartej w próbach pozostaje proporcjonalna do ich wielkości. Gdy jednak spostrzeżenia powstają w sposób zależny, np. gdy mamy b indywiduów wybranych na chybił trafił i dla każdego z nich mamy po k spostrzeżeń dotyczących jakiejś cechy x, tak że ogółem posiadamy $n = k\,b$ spostrzeżeń w próbie, to chcąc porównywać taką próbę ze względu na ilość informacji z innymi próbami opartymi na indywiduach mierzonych tylko 1 raz, należałoby wyznaczyć liczbę spostrzeżeń niezależnych n', równoważną liczbie n spostrzeżeń otrzymanych w sposób zależny.

W celu wyznaczenia tej liczby n', piszemy wyrażenia na zmienność próbową średniej arytmetycznej w dwu postaciach, [1] i [3], których przyrównanie pozwala nam wyznaczyć n' jako funkcję zmienności populacyjnej pojedyńczych spostrzeżeń (σ_x^2) oraz zmienności populacyjnej średnich indywidualnych (σ_m^2) (wzór [5]). Tą ostatnią daje się łatwo oszacować ($\hat{\sigma}_m^2 = \frac{b s_m^2}{b-1}$). W celu oszacowania pierwszej wyrażamy ją jako sumę σ_∞^2 i σ_e^2 gdzie σ_∞^2 jest zmiennością „między indywiduuami", ześ σ_e^2 jest zmiennością „wewnątrz indywiduów" wzór [6]. Następnie wyrażamy σ_m^2 jako funkcję σ_∞^2 i σ_e^2 (wzór [7]). Otrzymane stąd wyrażenie σ_x^2 podstawiamy do [5], otrzymując wzór [8] wyznaczający n' jako funkcję σ_m^2 i σ_e^2. Wzory [9] dają oszacowanie tych parametrów. Na podstawie tych wzorów uzyskujemy wzór na ocenę n', n'^* ([10]), lub, po wykorzystaniu tożsamości [11], dogodniejszy do obliczeń wzór [12].

W wypadku gdy liczby spostrzeżeń odnoszące się do poszczególnych indywiduów różnią się między sobą, ale różnice te nie są zbyt duże, wzór [12] można używać jako przybliżenie, z tym że s_m^2 zostanie obliczone według wzoru [13].

Ponieważ spostrzeżenia otrzymane w sposób zależny mogą okazać się w rzeczywistości statystycznie niezależne, pożądanym jest wyznaczyć punkt krytyczny dla n'^* na założeniu niezależności spostrzeżeń (hipoteza zerowa). W tym celu drogą prostego przekształcenia [10]

otrzymujemy [14], w którym F_m^0 (zdefiniowane w [15]) jest znanym „ilorazem zmienności" obliczonym zgodnie z założeniem o niezależności spostrzeżeń, posiadającym dwuparametrowy rozkład $F=e^{2z}$, gdzie z jest funkcją Fisher'a[1]). Dla wyznaczenia punktu krytycznego dla n'^* należy podstawić w [14] na miejsce F_m^0 wielkość F_P odczytaną z tablicy rozkładu F przy liczbach stopni swobody $\nu_1=b-1$ i $\nu_2=n-b$, przy czym P oznacza tu obszar prawego tylko ogona odpowiedniej krzywej F. Jeżeli n'^* wyznaczone według wzoru [12] okaże się mniejsze od punktu krytycznego, to spostrzeżenia mogą być uznane (z ryzykiem błędu 100 P%) za zależne. Jeżeli natomiast n'^* okaże się większe od tego punktu, to n' może być uznane za równe n, jeżeli tylko błąd 2 rodzaju nie będzie większy od błędu 1 rodzaju.

Jako zmienna losowa n'^* jest funkcją liniową F wyrażoną wzorem [17] z liczbami stopni swobody $\nu_1=n-b$, $\nu_2=b-1$.

Przedział ufności dla n' można wyznaczyć definiując F_m bez założenia o niezależności spostrzeżeń ([18]), skąd otrzymujemy wzór [19] na przedział ufności, gdzie F_1 i F_2 są punktami krytycznymi odpowiadającymi w rozkładzie F obszarom dwu ogonów. Górny kres przedziału ufności nie może oczywiście przekraczać n. Przedział ufności będzie tym krótszy, czym bliżej do b wypadnie n'^*.

Użytek z n'^* nie ogranicza się do porównywania ilości informacji. Wielkość ta jest np. stosowalna przy sprawdzianie istnienia korelacji metodą ilorazu *Student'* a podanego w [20], gdzie n należy zastąpić przez n'^* w wypadku gdy spostrzeżenia są zależne. Oczywiście sprawdzian *Student'* a staje się przez to niedokładny, jednakże (zwłaszcza badając kres dolny i górny przedziału ufności dla n') unika się przez to grubego błędu, który by został popełniony przez pozostawienie n niepoprawionego.

Dla oszacowania błędu standardowego spółczynnika regresji b_{yx} wzór podany w [21] i [22] jest ważny tylko wtedy, gdy n spostrzeżeń zmiennej $y\cdot x$ są niezależne. Gdy spostrzeżenia te otrzymano w sposób zależny, chodzić nam będzie o to, aby wyznaczyć liczbę n' spostrzeżeń niezależnych zmiennej $y\cdot x$ równoważną liczbie n spostrzeżeń otrzymanych w sposób zależny (zmienna $y\cdot x$ jest zmienną y przy ustalonych wartościach x).

[1]) R. A. Fisher „Statistical Methods for Research Workers". Oliver and Boyd, London, 1948.

R. A. Fisher and F. Yates „Statistical Tables of Biological, Agricultural and Medical Research". Oliver and Boyd, London, 1948.

W tym celu przyrównujemy do siebie dwa wyrażenia na zmienność średniej arytmetycznej $y \cdot x$ (wzór [23], [24], i [25]. Podstawiając, otrzymujemy wzór na n' podany w [26].

Jeżeli regresja jest liniowa t. j. jeżeli odchylenia średnich arytmetycznych y' ów przy ustalonych wielkościach x' ów od linii regresji można uznać za losowe, to wtedy zmienność populacyjną tych średnich, t. j. zmienność „między kolumnami" w tablicy regresji, σ_a^2, można uznać za równą zmienności wartości regresyjnych ($\sigma_y^2 \varrho_{xy}^2$). Na tym założeniu piszemy wzór [27], z którego następnie wyprowadzamy, korzystając ze wzorów [6] i [7], wzór [28]. Wielkość $\sigma_\infty^2 - \sigma_a^2$, jest to zmienność „między indywiduuami" pozbawiona wpływu zmienności „między kolumnami". Oznaczając ją przez σ_b^2, piszemy wzór [29]. Ocenę nieobciążoną σ_b^2 podajemy w [30]. Po podstawieniu właściwych ocen parametrów występujących w [29], otrzymujemy ocenę n' wyrażoną wzorem [31], lub, po wykorzystaniu tożsamości [32], ocenę wyrażoną bardziej dogodnym do obliczeń wzorem [33].

Dla wyznaczenia punktu krytycznego dla n'^* przy założeniu niezależności spostrzeżeń $y \cdot x$ piszemy [31] w postaci podanej w [34] gdzie F_b^0 jest zdefinjowane w [35]. Punkt krytyczny wyznaczamy podstawiając w [34] na miejsce F_b^0 wielkość graniczną F_P odczytaną z tablicy F przy $\nu_1 = b - a$, $\nu_2 = n - b$.

Dla wyznaczenia przedziału ufności dla n' definjujemy F_b w [37], nie opierając się już na założeniu niezależności, skąd wyprowadzamy wzór [38] dla $100(1-P)\%$-owego przedziału ufności dla n'.