#112

Metrika

Zeitschrift für theoretische und angewandte Statistik

Sonderdruck aus
Band 6
1963

PHYSICA-VERLAG · WÜRZBURG

On the Correction of Mathematical Bias by Use of Replicated Designs

By W. E. DEMING, Washington[1])

Purpose. Use of replicated sampling designs for ease in calculation of standard errors is well known. Not so well known is the fact that a replicated design also enables one to remove most of the mathematical bias in the formula of estimation, if any bias exists. The purpose of this paper is to illustrate the removal of mathematical bias, and an improved calculation of the variance, following procedures described by QUENOUILLE[2]) and by DURBIN[3]).

Replicated designs furnish automatically the random variates of equal expected value and variance that one needs for removal of bias and for estimation of variances. This type of design was described by MAHALANOBIS[4]) in earlier years as an interpenetrating network of samples. In 1949 my friend Professor JOHN W. TUKEY showed me a simplified version of replication, which I have used ever since. It went under the name of the TUKEY plan in my book *Some Theory of Sampling* (WILEY, 1950), with an extended treatment in my later book *Sample Design in Business Research* (WILEY, 1960).

I give here a separate proof of the efficiency of QUENOUILLE's methods, before passing on to an illustration.

Theory. Suppose that we wish to estimate, by a sample-survey, the numerical value of some function $f(Ex, Ey)$. Ex and Ey may both be unknown, but the sample furnishes estimates of either or both, hence also of $f(Ex, Ey)$. A sample replicated in k subsamples,

[1]) W. EDWARDS DEMING, Ph. D., LL. D., Consultant in Statistical Surveys, 4924 Butterworth Place, Washington 16.

[2]) M. H. QUENOUILLE: "Approximate tests of correlation in time-series". J. Royal Statistical Society, Series B, vol. 11, 1949, pages 68—84; page 70 in particular. Biometrika, "Notes on bias in estimation", vol. 43, 1956, pages 353—360. See also H. O. HARTLEY and A. ROSE: "Unbiased ratio estimators", Nature, vol. 174, 7. Aug. 1954, page 170. H. O. HARTLEY and L. A. GOODMAN: "The precision of unbiased ratio-type estimators". J. American Statistical Association, vol. 53, 1958, pages 491—508.

[3]) J. DURBIN: "A note on the application of Quenouille's method of bias reduction to the estimation of ratios". Biometrika, vol. 46, 1959, pages 477—480.

[4]) P. C. MAHALANOBIS: "On large-scale sample-surveys". Phil. Trans. Soc., vol. 231B, 1944, pp. 329—451; "Recent experiments in statistical sampling". J. Royal Stat. Soc., vol. cix, 1946, pp. 325—378.

furnishes the estimates x_i and y_i of Ex and Ey ($i = 1, 2, \cdots, k$). Each subsample, we suppose, is a valid sample of the whole frame. All k subsample, are precisely of the same design. They belong to the same probability-system, and their results differ only because the selections of the sampling units in each came from different random numbers, and because accidental errors of performance also introduce variation between subsamples.

As an example, $\bar{n}$ might be the number of segments of area drawn into each subsample, x_i the number of packages of some product that the y_i families in Subsample i purchased last week. Then $f(x_i, y_i)$ might be x_i/y_i, the average number of packages purchased per family. Or, x_i might be the number of defective items in Subsample i, y_i the number of items tested, in which case $f(x_i, y_i) = x_i/y_i$ would be the so-called fraction defective. The function $f(x, y)$ could of course have any form, such as πx^2 for the area of a circle, x being the measured radius.

Let x be any random variable with expected value Ex. Then

$$x = Ex + \Delta x \tag{1}$$

where Δx is the sampling error in x, and

$$E\Delta x = 0 \tag{2}$$

$$E\Delta x^2 = \sigma_x^2 \tag{3}$$

$$E\Delta x^3 = \mu_{3x} \tag{4}$$

$$E\Delta x^4 = \mu_{4x} = \beta_2 \sigma_x^4 \tag{5}$$

with similar forms for y. Then for any function $f(x, y)$ that possesses derivatives,

$$f(x, y) = f(Ex, Ey) + f_x \Delta x + f_y \Delta y + f_{xx} \Delta x^2 + 2f_{xy} \Delta x \Delta y + f_{yy} \Delta y^2 + \cdots \tag{6}$$

where the subscripts on f denote derivatives evaluated at Ex, Ey.

For a sample of size n sampling units drawn with random numbers with replacement

$$Ef(x, y) = f(Ex, Ey) + \frac{A}{n} + \frac{B}{n^2} + \frac{C}{n^3} + \cdots \tag{7}$$

where

$$\left.\begin{aligned} A &= E\Delta x^2 + E\Delta y^2 + 2E\Delta x \Delta y = \sigma_x^2 + \sigma_y^2 + 2E\Delta x \Delta y \\ B &= E(\Delta x^3 + 3\Delta x^2 \Delta y + 3\Delta x \Delta y^2 + \Delta y^3) \\ C &= E(\Delta x^4 + \text{etc.}) \end{aligned}\right\} \tag{8}$$

There will always be, for any function $f(x, y)$ that possesses derivatives, a sample so big that the remainder after any term will be smaller than any preassigned number ε. Just what this size of sample is depends on the number ε, on the function $f(x, y)$, and on the moment coefficients of the distribution of the sampling units in the frame.

Thus the difference between $Ef(x, y)$ and $f(Ex, Ey)$, commonly called the bias in the sampling procedure, decreases with n. Hence, for samples sufficiently large (with proper assumptions about f),

$$Ef(x, y)=f(Ex, Ey)+A/n \tag{9}$$

Let there be k subsamples, $\bar{n}$ sampling units per subsample, $n = k\bar{n}$ sampling units in all k subsamples combined. Let x_i be the value of x derived from Subsample i, and y_i the corresponding value of y. Let also

$$x=x_1+x_2+\cdots+x_k \tag{10}$$

$$y=y_1+y_2+\cdots+y_k \tag{11}$$

$$x_{(i)}=x-x_i \tag{12}$$

$$y_{(i)}=y-y_i \tag{13}$$

$$f_{(i)}=f(x_{(i)}, y_{(i)}) \tag{14}$$

$$f_{\cdot}=\frac{1}{k}\sum_{1}^{k} f_{(i)} \tag{15}$$

Use $\tilde{f}$ for an estimate of $f(Ex, Ey)$ in Eq. 9. Then the full sample (all k subsamples) leads to the approximation

$$f=\tilde{f}+\frac{A}{k\bar{n}} \tag{16}$$

while the average estimate $f_{\cdot}$ leads to

$$f_{\cdot}=\tilde{f}+\frac{A}{(k-1)\,\bar{n}} \tag{17}$$

The solution of these 2 equations is

$$(k-1)(f_{\cdot}-\tilde{f})=A/\bar{n} \tag{18}$$

$$kf=k\tilde{f}+A/\bar{n}$$

$$=\tilde{f}+(k-1)f_{\cdot} \tag{19}$$

whence

$$\tilde{f}=kf-(k-1)f_{\cdot} \tag{20}$$

We may take $\tilde{f}$ as an estimator of $f(Ex, Ey)$, good to within powers of $1/n$. In the case of 2 subsamples,

$$\tilde{f}=2f-\frac{1}{2}(f_1+f_2) \tag{21}$$

which QUENOUILLE proposed in 1949, f_1 being $f(x_1, y_1)$, and f_2 being $f(x_2, y_2)$.

A simple graph illustrates the solution (see figure). The horizontal coordinates are the reciprocals of the relative sizes of the samples that make up f and $f_.$. The line drawn through the 2 points $(1/k, f)$ and $(1/[k-1], f_.)$ intersects the vertical axis at $1/k = 0$, corresponding to infinite size of sample, where the bias would be 0. The intercept $\tilde{f}$ is thus the solution of Eq. 20 and is an estimate of $f(Ex, Ey)$. The slope of the line is $k(k-1)(f_. - \tilde{f})$, which would be 0 if A were 0 — that is, if there were no bias. The variance of $\tilde{f}$ is

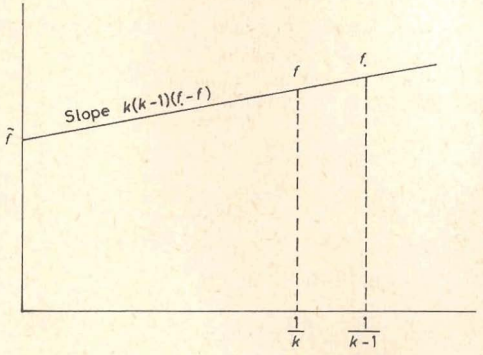
$$\hat{V}ar\,\tilde{f} = \frac{k-1}{k}\sum[f_{(i)} - f_.]^2 \tag{22}$$

which is equivalent to

$$\hat{V}ar\,\tilde{f} = \frac{1}{k(k-1)}\sum[\tilde{f}_{(i)} - \tilde{f}]^2 \tag{23}$$

wherein

$$\tilde{f}_{(i)} = kf - (k-1)f_{(i)} \tag{24}$$



Holes. Use of $f_{(i)}$ offers a valid simple way out of the difficulty that occurs when some rare item fails to appear in 1 or more subsamples (called by TUKEY a *hole*), provided the item appears in at least 2 subsamples. An example is loading coils in manholes or on telephone poles, in a study of the property owned by a telephone company. Loading coils are rare; on the average, only 1 manhole or 1 pole in 20 carries a loading coil. Moreover, the loading coils, when they do appear, often do so in clusters of from 1 to 30 in one manhole or on one pole. They are nevertheless important in the inventory. It often happens in practice that 2 or 3 of 10 subsamples in the inventory contain no loading coil.

Clearly, we get a solution by use of the methods of this paper, provided a rare item appears in at least 2 subsamples.

Example. For a numerical example, I take a study of the aerial property of a telephone company. The aim of the study was to estimate the cost of repairing the average repeater or loading coil, to put it in 1st class condition. The sampling unit was a telephone-pole, and there about 30 telephone-poles in each subsample; 300 poles in the entire sample. At least 1 repeater appeared in each subsample, but loading coils failed to appear in 3 subsamples (see table).

Estimates of maintenance required x_{ji} and y_{ji} are observed. The other figures are calculated

Subsample	Repeaters		Loading coils		$x_i = x_{1i} + y_{2i}$	$y_i = y_{1i} + y_{2i}$	$x_{(i)} = x - x_i$	$y_{(i)} = y - y_i$	$f_{(i)} = x_{(i)}/y_{(i)}$	$\tilde{f}_{(i)} = 10f - 9f_{(i)}$
	x_{1i}	y_{1i}	x_{2i}	y_{2i}						
1	300	4	500	4	800	8	5820	69	84.3478	100.6101
2	425	3	0	0	425	3	6195	74	83.7162	106.2945
3	550	13	0	0	550	13	6070	64	94.8438	6.1461
4	275	3	0	0	275	3	6345	74	85.7432	88.0515
5	575	1	600	10	1175	11	5445	66	82.5000	117.2403
6	425	8	300	2	725	10	5895	67	87.9851	67.8744
7	350	2	170	2	520	4	6100	73	83.5616	107.6859
8	375	7	150	1	525	8	6095	69	88.3333	64.7406
9	550	5	250	3	800	8	5820	69	84.3478	100.6101
10	400	3	425	6	825	9	5795	68	85.2206	92.7549
All 10	4225	49	2395	28	$x=6620$	$y=77$	59,580	693	860.5994	852.0084

It is perfectly permissible to make separate estimates for repeaters and for loading coils. The 7 subsamples that contain loading coils furnish a valid estimate of the cost of repairing loading coils, and for the variance of this estimate[1]).

However, for the sum of the repairs required for repeaters and loading coils combined, we do not add the separate estimates, as there is the possibility of correlation when repeaters and loading coils appear on the same pole. Use of $f_{(i)}$ nevertheless provides a uniform procedure of calculation, in which $x - x_i$ is the cost of repairing the $y - y_i$ repeaters and loading coils combined, in Subsample i.

The table shows x_{ji} in dollars for the cost of repairs for the y_{ji} items of Class j in Subsample i. Numerical calculations give

$$f = x/y = \$\,59.580/693 = \$\,85.9740$$

$$f_{\cdot} = \$\,85.2008$$

$$\tilde{f} = 10f - 9f_{\cdot} = 859.740 - 9 \times 85.2008 = \$\,85.20$$

[1]) HOWARD L. JONES: "Investigating the properties of a subsample mean by employing random subsample means". J. American Statistical Association, vol. 51, 1956, pp. 54—83, p. 78 in particular.

This figure is an estimate of the average cost of repairing a repeater or a loading coil. For the standard error of this estimate we find that

$$\hat{Var}\,\tilde{f} = \frac{1}{10(10-1)} \sum [\tilde{f}_{(i)} - \tilde{f}]^2 = 105.32$$

$$\hat{\sigma}_{\tilde{f}} = \sqrt{105.32} = \$\,10.26$$

The bias $f - \tilde{f}$ in $f = x/y$ is clearly small, being only 85.97 — 85.20 = 0.77, which is less than $\hat{\sigma}_{\tilde{f}}$. I may add that this is my usual experience: bias in the ratio-estimate is almost always insignificant. It is nevertheless satisfaction to have at hand the theory contributed by QUENOUILLE and DURBIN, and straightforward arithmetic procedures for computing $\tilde{f}$ and its variance.

I am indebted to a number of people for the ideas in this paper, especially my friend Professor JOHN W. TUKEY of Princeton and the Bell Telephone Laboratories, who showed me the use of $f_{(i)}$; also to Professor HOWARD L. JONES, formerly with the Illinois Bell Telephone Company, now Professor of Statistics at the University of Chicago. The numerical calculations are the work of my wife, LOLA S. DEMING, M. A. Professor G. S. WATSON assisted ... with references to QUENOUILLE and DURBIN.