Reprinted from

CONTRIBUTIONS TO STATISTICS

PRESENTED TO

PROFESSOR P. C. MAHALANOBIS

ON THE OCCASION OF HIS
70TH BIRTHDAY

ON SOME OF THE CONTRIBUTIONS OF INTERPENETRATING NETWORKS OF SAMPLES

By

W. EDWARDS DEMING

Paper received : March, 1963

PERGAMON PRESS
OXFORD · LONDON · NEW YORK · PARIS · FRANKFURT
STATISTICAL PUBLISHING SOCIETY
CALCUTTA

JUNE, 1964

Printed in India
at the Eka Press, Calcutta-35

ON SOME OF THE CONTRIBUTIONS OF INTERPENETRATING NETWORKS OF SAMPLES

By W. EDWARDS DEMING

Consultant in Statistical Surveys, Washington

Breadth of application

It is a special pleasure and privilege for me to present a paper in honour of Professor Mahalanobis, as for 14 years I have used only interpenetrating networks of samples (IPNS), initiated by him, as everyone knows, about 1936. Applications in my own work cover many types of samples of human populations in several countries for studies in consumer research, labour force, morbidity, evaluation of inventory in process, measurement of physical deterioration of plant and buildings belonging in public utilities, studies in the fertility of schizophrenics, psychological problems of the deaf, studies of mental retardation, estimates of the costs of operations (e.g., switching railway cars), cost of installing certain types of equipment, usage of telephone circuits, studies of the characteristics of rail and motor freight, air passenger-traffic, census information, early tabulation of complete censuses, agricultural production, studies of housing conditions, and otherwise.

Advantages of IPNS

The main feature of IPNS is simplicity in the calculation of the standard error of an estimate. It also enables one to estimate rapidly the mathematical bias, if any, in the formula of estimation (*vide infra*). It helps to detect gross blunders in selection, recording, and processing. It permits evaluation of variances between investigators, coders, and other workers in the various statistical stages of processing. A section further on deals with the detection of gross blunders.

In a consumer-survey in the US, housewives answered questions (*inter alia*) with respect to purchases of aluminium foil. The survey showed that Brand X, during the past year, had lost a good share of the market to another brand, which I will call Brand Y. On the other hand, shipments of Brand X from the factory had increased, and the manufacturer doubted the results of the survey. This was for Brand X a critical matter in plans for the coming year in respect to output, distribution, and advertising. It turned out later that the survey was correct : heavy shipments from the factory had represented increases in wholesale inventory, not demand by the consumer. Without standard errors, one could not be so sure that the results for Brand X indicated a real decrease for Brand X, and a real increase for Brand Y, and could not have given the manufacturer results for planning at the time when he needed them.

Standard errors also serve as a guide to day-to-day improvement of survey-procedures. For example, I plotted standard errors of 50 characteristics computed from the results of a national consumer-survey by the format shown further on. Forty-two of the standard errors fell between $\sqrt{(pq/n)}$ and $\sqrt{(cpq/n)}$, c being the size of segment (4 dwelling units to the segment in this case). Strangely, 8 standard errors were well above $\sqrt{(cpq/n)}$. A standard error as high as $\sqrt{(cpq/n)}$ is possible if the c dwelling units within every segment gave solid answers, all yes or all no. Variance greater than $\sqrt{(cpq/n)}$ indicates that the interviewers themselves were supplying some answers. A glance at the questionnaire showed that these 8 characteristics were questions about personal products, concerning which interviewers do not enquire: they guess, and put down answers. The standard errors thus indicated that the results for these 8 characteristics were largely fictitious. Two courses of action were open: (a) retrain the interviewers, or (b) revise the questionnaire. Without standard errors at hand, these observations and improvements would not have taken place.

One may lay out the sample so that pairs of subsamples are worked in the field, in the coding, and even in the punching and tabulating, by separate groups of workers. The pairs of subsamples are competitive. Wide disagreement between the pairs of subsamples, compared with the variance within pairs, may indicate operational errors or gross misunderstanding of definitions or procedures, to be found and corrected. Use of subsamples in this way does not replace other statistical controls, but it has been very helpful in my own work, especially in the detection of gross blunders.

Conversely, one may interpret a small standard error, where the design measures the total variance, including the variance between investigators, as indication of fairly uniform performance.

A further advantage is important, when one uses IPNS in the form of the Tukey plan (*vide infra*), and that is easier communication of the procedure, and easier interpretation of results, to the nonstatistician.

Detection of persistent errors and of gross blunders

Various designs are possible for measurement of the variance between investigators. Some designs give rise to low-power measurement of this variance, at little additional cost, as where a pair of sampling units in different subsamples are allotted at random to a pair of investigators. Each pair of sampling units yields one degree of freedom. Other designs give more degrees of freedom per sampling unit, but require more travel, and cost more. No extended account is needed here, though a reference to Chapter 12 in my book *Sample Design in Business Research* (Wiley, 1960) might not be out of place.

As an example of detection of a blunder in selection, I may cite a continuing study of motor freight in the US, the purpose of which is to estimate the distributions of shipments by type of rate, by weight-bracket, by mileage-bracket, type of handling,

circuity, and other characteristics that affect costs and profits. As a regular part of the sampling procedure, the sample for each motor carrier is summarized for a month, and estimates prepared for certain grand totals, viz., number of shipments, pounds carried, and revenue, and compared with figures that come from the accounting department of the company. The purpose of these comparisons is to detect gross misunderstandings in procedure and in definition.

This sample is laid out in 2 subsamples, wherefore the estimates from the 2 subsamples should under good reporting bracket the figure from the accounting records half the time; both subsamples should be below the figure from the accounting records a quarter of the time, and both above it a quarter of the time, in a random manner.

Such comparisons have disclosed a number of times the fact that some of the accounting departments have been reporting incorrectly, through misunderstanding of definitions (for example, confusion of line-haul with storage, actual weight for billed weight). The extra care that is possible with the sample detects errors in classification and definition that had caused biases month in and month out in the regular 100% reporting.

A run of successive months in which both subsamples fall short of the figure from the accounting records for the month indicates a persistent error, either in the sample, or in the accounting records. To save space, I will not show actual figures for an example, but will only describe a nonparametric comparison whereby plus (+) indicates that the estimate from a subsample fell above the figure from the accounting records, and minus (−) that it fell below.

A run of negative signs in the TL-category (TL for truckload) was observed for one carrier. Although it is not the function of a statistical test to identify a cause, but only to detect the fact that a cause exists, if one exists, it is often true, as every statistician knows, that statistical tests do often point the finger pretty definitely at the source of trouble, as happened in this case. This motor carrier hauls a special commodity (steel) by whole truckload, under a special type of rate. The sampling unit is the freight bill for a shipment, a piece of paper about 5 by 7 inches with a serial number. The freight bills for steel are printed in a different colour from the others, with a different run of serial numbers, and they are filed separately, in filing cabinets of a different colour. The people that carried out the selection somehow assumed, in spite of careful admonition to study all traffic, that TL-shipments of steel were not subject to the sampling plan. The accounting department, however, reports revenue from all sources including steel: hence the persistent discrepancy. Agreement improved when TL-shipments of steel were sampled.

The nonparametric test just described is simple but helpful. Other tests, some parametric, are also possible because of the replication provided by IPNS.

Brief description of the method

One employs any type of sample-design, then replicates it k times with samples of size $1/k$ times the original size intended, with the use of paper zones. Each replicate is called a subsample. This sounds simple, and it is, but it is necessary to observe that the k subsamples must be independent. For example, for validity of the simple formulas shown further on, the system of selection must permit a primary unit to contribute to more than one subsample: the fact that one subsample fell into a primary unit must not preclude a 2-*nd* one from falling into it. However, the ultimate unit of investigation (for example, a household or group of households, or a segment of area) may be restricted to one subsample. That is, we may draw the ultimate unit without replacement and apply the finite correction factor $1-n/N$ to the estimate of variance.

The requirement that 2 random numbers be permitted to fall into one primary unit in order to validate the simple estimates of variance shown further along was first pointed out to me by my friend and colleague Professor John W. Tukey in 1949. I thereupon gave the name *Tukey plan* to the procedure of IPNS in its simplified form, with paper zones, now to be described.

For a brief description of the method, we may use for illustration a national sample of the population, as for consumer research, or for census data.

The primary frame will be a list of areas with definite boundaries, such as counties or metropolitan districts. These areas will preferably be areas for which we have recent Census-figures, although one must sometimes draw up plans without Census-figures, or with obsolete figures. If there are no figures at all, one may resort to the use of equal sizes, or sizes modified by judgment.

In any case, we show for each area the best available measure of size. We then decide on the size of the sampling unit, and cumulate sampling units in the primary frame. The next step is to decide the zoning interval for the paper zones; then to construct a sampling table, which will show the serial numbers of the sampling units selected for the sample within every paper zone.

Every random number in the sampling table will fall into a primary area that is definitely identifiable. The next step is the same as in multi-stage sampling, viz., to break up the primary unit into portions, and to draw one or more portions, break them up, and continue nesting until local frames provide lists of the ultimate small sampling units (e.g., segments of area containing a number of dwelling units), random drawings of which will provide the workloads for interviewing.

Paper zones contain a fixed number of sampling units. This number is called the zoning interval, Z. $1/Z$ is the probability that any specified sampling unit will fall into any designated subsample, and (when we draw without replacement) k/Z is the overall probability of selection.

In a national sample, k would almost always be 2, to get maximum benefit of geographic stratification. For smaller areas, such as a city or a few counties that are fairly homogeneous, I use 10 subsamples, for simplicity.

Details of the procedure are not necessary in this paper, as I have described it with numerous examples in my book *Sample Design in Business Research* (Wiley, 1960).

One will of course stratify in the use of IPNS, just as he would in any other design. One may use systematic subsamples, with k random starts, for k subsamples, if systematic sampling appears to be advantageous, or one may draw k fresh random numbers in every zone where there is danger of cycles (an ever-present hazard, in my experience). The theory for the optimum size of segment and for the choice of sampling unit are the same as they are for any other sampling procedure.

One may vary the size of the sampling unit, and may vary the size of segment within a sampling unit. For example, in areas where there are no recent maps, or where travel is expensive, or where carving is for any other reason especially difficult, one may double the size of segment, or may avoid wide dispersion of segments. This will cut down on the number of segments that come into the sample, and will save cost of carving segments, and cost of travel. One may also double the size of the sampling unit in a primary area, and thus halve the number of sampling units therein. What this does is to decrease the probability that this county will contribute to the sample, and decrease the cost of carving segments and the cost of travel in case a random number does fall into this primary unit.

These alterations do not change the overall probability of selection, nor the procedure of tabulation.

Example of format for calculation

A convenient format for calculation of the standard error with 2 subsamples may be of interest, as local modifications thereof will fit a variety of uses (see Table).

This example comes from a national survey carried out with a replicated design. Let y (with appropriate subscript) denote the number of dwelling units in a segment, and x denote the number of women that answer yes. Use of dwelling units as the base permits use of figures from the Census for ratio-estimates of the total number of females that would have answered yes to the question tabulated had the sample been 100 per cent.

The size of the survey in this example was small, being only about 1200 interviews. The number of dwelling units in the survey ($y = y_1 + y_2 = 1961$ in the table) is not the number of interviews, but the number of dwelling units in the 400 segments of area selected for the sample. The universe was female homemakers, and the question was this: "Do you make your own cake frosting at home?" Some dwellings had no female homemakers whom the questions would apply to: moreover, there is always some nonresponse; hence the number y will be bigger than x.

The 400 segments of area were not drawn as one sample, but as 2 independent samples, each of about 200 segments of area. The subscripts denote the two subsamples

1 and 2. The symbols NE, NC, etc., in the heading denote Census regions—Northeast, North Central, South, and West. The sampling plan, including the formulas, are in Chapter 11 of my book, cited earlier.

The standard error calculated here includes the variable performance of the interviewers, as there was no attempt, in this instance, to randomize interviewers and subsamples in an orthogonal design for separate evaluation of the variance between interviewers.

characteristic	metropolitan				non-metropolitan				US
	NE	NC	S	W	NE	NC	S	W	
$i =$	1	2	3	4	5	6	7	8	1–8
x_1	98	100	40	38	35	33	47	33	424
x_2	79	69	38	45	17	38	52	29	367
$x = x_1 + x_2$	177	169	78	83	52	71	99	62	791
$x_1 - x_2$	19	31	2	−7	18	−5	−5	4	57
y_1	213	203	123	87	64	94	174	62	1020
y_2	176	162	110	89	55	128	165	56	941
$y = y_1 + y_2$	389	365	233	176	119	222	339	118	1961
$y_1 - y_2$	37	41	13	−7	9	−34	9	6	79
$p = x/y$	.455	.463	.355	.472	.437	.320	.292	.525	.40
$h = (x_1 - x_2) - p(y_1 - y_2)$	2.16	12.01	−2.35	−6.05	14.06	5.88	−7.62	.850	25.16
$S^2 = h^2$									482.68
S									20.7
$\hat{\sigma}_p = S/y$									.01

Final estimate for the US, $p = .40$; standard error .01. Upper and lower 1% fiducial limits, .43 and .37 respectively, based on 7 degrees of freedom.

The results show $p = .40$ for an estimate of the proportion of dwelling units with women that prefer to make their own cake frosting at home, and $\hat{\sigma}_p = .01$ for the standard error of this estimate (rather, for the standard error of the sampling procedure that gave p). The number of degrees of freedom in this estimate of the standard error is something less than 8 because the variances contributed by the different geographic regions are unequal.[1]

The upper 1% fiducial limit, for the estimated proportion, based on 7 degrees of freedom, is .40 plus 3 standard errors, or .43. The fiducial 1% lower limit would be .40 less than 3 standard errors. The upper and lower 1% fiducial limits of the estimated proportion are thus .43 and .37.

[1]One may estimate the number of degrees of freedom by a formula given by F. E. Satterthwaite, (1946): An approximate distribution of estimates of variance components. *Biometrics*, **2**, 110–114. Satterthwaite's formula appears in the appendix to Chapter 11 of my book cited earlier.

The standard error of any other characteristic that the survey measured may be calculated in like manner. Usually, one calculates only the standard errors of chief importance, such as the total dwelling units in the US (for comparison), the total number of purchasers of a product, one's share of the market, a chief competitor's share, and possibly a few other characteristics. In this connexion I may add the fact that the estimate of the total number of dwelling units in the US, made from this survey, came out 2% short of the Census figure; ; or about 1 standard error short. One may possibly interpret this difference and its standard error as evidence that persistent undercoverage of assigned segments of area, if there be any, must have been of small magnitude.

Estimates of variance and of bias

Estimates of variance. The k replications (or k subsamples) give k estimates x_i of Ex, k estimates y_i of Ey, and k estimates f_i of $f(Ex, Ey)$. where f_i is some function of x_i and y_i. Let $\tilde{x}$ be the estimate of Ex calculated from the whole sample, and define $\tilde{y}$, likewise. Let f denote $f(\tilde{x}, \tilde{y})$.

Under the rules of selection stated above, we obtain an estimate of the variance of $\tilde{x}$ by the simple form

$$\hat{\text{var}}\ \tilde{x} = \frac{1}{k(k-1)} \Sigma (x_i - \bar{x})^2 \qquad \ldots \quad (1)$$

[The sums run overall k subsamples]

and an approximation to the variance of $f = \tilde{x}/\tilde{y}$ by the formula

$$\hat{\text{var}}\ f = \frac{k}{(k-1)\tilde{y}^2} \Sigma (x_i - f y_i)^2. \qquad \ldots \quad (2)$$

In practice, we may simplify our estimate of the variance of any function f even further by merely calculating

$$\hat{\text{var}}\ f = \frac{1}{k(k-1)} \Sigma (f_i - f)^2 \qquad \ldots \quad (3)$$

It is easy to increase the number of degrees of freedom, if k—1 is not enough (cf. my book cited earlier, pages 198 and 199).

An improved estimate of Ex/Ey, together with its variance, appears further along, useful under the rare condition of troublesome bias in the estimate $f(\tilde{x}, \tilde{y})$.

Correction of bias. I here adapt some theory of Quenouille[2] and of Durbin,[3] along with suggestions privately communicated to me from Dr. Tukey, to illustrate a

[2] M. H. Quenouille (1941): Approximate tests of and correlation in time-series. *J. Roy. Stat. Soc.*, Series B, **11**, 68–84; p. 70 in particular. M. H. Quenouille (1956): Notes on bias in estimation. *Biometrika*, **43**, 353–360. See also H. O. Hartley and A. Ross (1954): Unbiased ratio estimators. *Nature*, **174**, 7 August, 170; H. O. Hartley and L. A. Goodman (1958): The precision of unbiased ratiotype estimators. *J. Amer. Stat. Ass.*, **53**, 491–508. John W. Tukey (1958): Bias and confidence in not-quite large samples. *Ann. Math. Stat.*, **29**, 614.

[3] J. Durbin (1959): A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, **46**, 477–480.

graphical evaluation of the variance of a function, and an improvement in the estimate of the variance of a function of x, y.

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, furnishes the estimates x_i and y_i of Ex and Ey ($i = 1, 2, \ldots, k$). Each subsample, we suppose, is a valid sample of the whole frame. All k subsamples 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 introduce variation between subsamples.

As an example, $\bar{n}$ might be the number of sampling units 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 xy for the area of a rectangle, x and y being the measured sides.

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

$$x = Ex + \Delta x \qquad \ldots \quad (4)$$

where Δx is the sampling error in x, and

$$E\,\Delta x = 0 \qquad \ldots \quad (5)$$

$$E\,\Delta x^2 = \sigma_x^2 \qquad \ldots \quad (6)$$

$$E\,\Delta x^3 = \mu_{3x} \qquad \ldots \quad (7)$$

$$E\,\Delta x^4 = \mu_{4x} = \beta_2 \sigma_x^4 \qquad \ldots \quad (8)$$

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

$$f(x, y) = f(Ex, Ey) + f_x \Delta x + f_y \Delta y + f_{xx} \Delta x^2 + 2 f_{xy} \Delta x\, \Delta y + f_{yy} \Delta y^2 + \ldots \qquad \ldots \quad (9)$$

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

$$E\,f(x, y) = f(Ex, Ey) + \frac{A}{n} + \frac{B}{n^2} + \frac{C}{n^3} + \ldots \qquad \ldots \quad (10)$$

where

$$\left.\begin{aligned} A &= E\,\Delta x^2 + E\,\Delta y^2 + 2E\,\Delta x\,\Delta y = \sigma_x^2 + \sigma_y^2 + 2E\,\Delta x\,\Delta y \\ B &= E(\Delta x^3 + 3\Delta x^2\,\Delta y + 3\Delta x\,\Delta y^2 + \Delta y^3) \\ C &= E(\Delta x^4 + \text{etc.}) \end{aligned}\right\} \qquad \ldots \quad (11)$$

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 pre-assigned 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 $E f(x, y)$ 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), we may obtain a specified degree of approximation by writing

$$E f(x, y) = f(Ex, Ey) + A/n. \qquad \ldots \quad (12)$$

Let there be k subsamples, $\bar{n}$ sampling units per subsample, $n = k\bar{n}$ sampling units in all k subsamples combined. We need now the symbols $x_{(i)}, y_{(i)}, f_{(i)}$ to denote estimates calculated from all subsamples excluding Subsample i.

Let
$$f_{\cdot} = \frac{1}{k}\sum_{1}^{k} f_{(i)}. \qquad \ldots \quad (13)$$

Then Equation (12) with the full sample leads to the approximation

$$f = \tilde{f} + \frac{A}{k\bar{n}} \qquad \ldots \quad (14)$$

while with the estimate $f_{\cdot}$ it leads to the approximation

$$f_{\cdot} = \tilde{f} + \frac{A}{(k-1)\bar{n}}. \qquad \ldots \quad (15)$$

We may solve these 2 equations for A and $\tilde{f}$, finding that

$$A = \bar{n}(k-1)(f_{\cdot} - \tilde{f}) \qquad \ldots \quad (16)$$

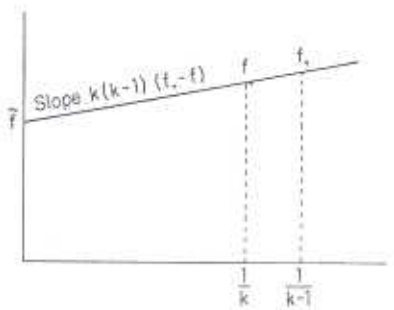
and
$$\tilde{f} = kf - (k-1)f_{\cdot} \qquad \ldots \quad (17)$$

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

$$\tilde{f} = 2f - \tfrac{1}{2}(f_1 + f_2) \qquad \ldots \quad (18)$$

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 $f_{\cdot}$. The



line drawn through the 2 points $(1/k, f)$ and $(1/[k-1], f_\bullet)$ 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 Equation (17) and is an estimate of $f(Ex, Ey)$. The slope of the line is $k(k-1)(f_\bullet - \tilde{f})$, which would be 0 if A were 0—that is, if there were no bias. The variance of $\tilde{f}$ may be estimated as

$$\hat{\text{var}}\ \tilde{f} = \frac{k-1}{k} \Sigma [f_{(i)} - f_\bullet]^2 \qquad \ldots \quad (19)$$

which is equivalent to

$$\hat{\text{var}}\ \tilde{f} = \frac{1}{k(k-1)} \Sigma [\tilde{f}_{(i)} - \tilde{f}]^2 \qquad \ldots \quad (20)$$

wherein

$$\tilde{f}_{(i)} = kf - (k-1) f_{(i)}. \qquad \ldots \quad (21)$$