ON THE DISTINCTION BETWEEN ENUMERATIVE AND ANALYTIC SURVEYS

By
W. Edwards Deming

Reprinted from the JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
Vol. 48, 1953, pp. 244 - 255

ON THE DISTINCTION BETWEEN ENUMERATIVE AND ANALYTIC SURVEYS

By

W. Edwards Deming

Reprinted from the JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
Vol. 48, 1953, pp. 244 - 255

ON THE DISTINCTION BETWEEN ENUMERATIVE AND ANALYTIC SURVEYS*

W. EDWARDS DEMING
New York University

I. DEFINITION OF THE ENUMERATIVE AND ANALYTIC USES OF DATA

Purpose of the paper. Statistical data are supposedly collected to provide a rational basis for action. The action may call for the enumerative interpretation of the data, or it may call for the analytic interpretation.

The aim here is to exhibit some of the consequences of failing to distinguish between the enumerative and the analytic uses of data. This distinction is necessary in the statement of the aims of a survey, census, or experiment, in order that the plans for the collection of the data and for the tabulations may most economically meet the needs of the consumer, and it is equally important in the interpretation of data.

Thus, to draw on a result from a later paragraph, information obtained in a complete census concerning every person in an area (e.g., on occupation, income, or education) still possesses *for analytic purposes* a sampling error that is actually about a quarter as great as the sampling error of a 6 per cent sample. The consequences are far-reaching. In using a census-table for analytic purposes, even though the figures come from a perfect complete count, it is therefore necessary to bear in mind that small numbers in a cell are unreliable in the sense that they have a standard error, just as if they had arisen in sampling, as indeed they did. Moreover, in the planning of a complete census, it is therefore imperative to use sampling for every bit of information that is not necessary as an aid to complete coverage, or required to give detail for small areas (such as the block statistics). Name, relationship to the head, age, sex, marital status, color are probably all necessary for the sake of completeness of coverage. These things, plus a few questions on rent, tenure, year built, will provide the information required for the block statistics.

To draw on another result, we shall see that it is often impossible to design a survey that will supply economically information for both enumerative and analytic purposes. For example, in a marketing survey, the best design for an estimate of the number of people who prefer to use ground coffee at home, rather than soluble coffee, requires, for

* Delivered at a conference on sampling conducted by the Institute of Statistics, University of North Carolina at the Blue Ridge Assembly, 21 July 1952.

greatest economy, one type of sample design; whereas a study of the *reasons*, or even of the difference in the two proportions, requires another design. One must be prepared to make some sacrifices in precision, as it may not be economical to satisfy both aims simultaneously.

The distinction between the enumerative and analytic uses of data.[1] Briefly, the enumerative question is how many? The analytic question is *why?* is there any difference between the two classes, and if so, how big are the differences?

In the enumerative problem, some action is to be taken because the frequency of some particular characteristic of the universe is found to exceed some critical value. The crop of wheat, according to a sample survey, turns out to be large or small. As a consequence, the market goes down or up, and production of meat, cereal products, and of substitutes shifts one way or another because of this information. The Census, or perhaps a sample study of birth registrations, shows that in a particular city the number of children in the primary schools will be much greater in 4 years than now. Bonds are issued; work commences on a new school building. Inspection of a sample of wool or of cotton may determine its disposition, the price to be paid for it, and what kind of cloth and of garments to make of it. Inspection of a lot of industrial product determines whether it will be accepted or subjected to screening or to a lower rating, or outright rejection.

Such problems are enumerative because they depend purely on a determination of the number of people in an area, or the inventory of grain, or the production of grain, or the quality of a product. They do not involve the analytic question of *why* all these people are there or why the crop this year is what it is; or *why* the wool or cotton or product is so good or so bad.

When certain cities in America were swelled with in-migrants because of war production in the spring of 1944, special censuses were taken with the aim of arriving at equitable allocations of food, gasoline, repair parts for buses and trolley cars, and other necessaries of living. Equitable distribution of supplies to these cities was impossible because no one knew just how many people were in them: assertions of editors and chambers of commerce did not provide a basis for action. The problem was enumerative because the action (viz., allocation

---

[1] These are the terms that I invented for Chapter 7 in my book *Some Theory of Sampling* (John Wiley, 1950). The terms are not important; the concepts are. The concepts are old, but plain statements of what they are and of the consequences of failing to keep them in mind in design and analysis are not easy to find. Similar but not exactly parallel concepts occur in the analysis of variance, under the terms Model I and Model II, a lucid explanation of which occurs in the paper by Churchill Eisenhart, "The assumptions underlying the analysis of variance," *Biometrics*, 3 (1947), 1–21.

of food and materials) depended on how many people were there and not why they were there.

By law, the Social Security program is partly an enumerative problem because federal reimbursement to a state depends on the number of inhabitants 65 and over within the state. Public health programs, agricultural adjustments, and other allotments depend on population and acreage, and are examples of enumerative uses of data. Administrative problems concerned with the long-range aims of these programs, however, are analytic.

In the analytic problem, the action is to be directed at the underlying causes that have made the frequencies of the various classes of the population what they are, in order to govern the frequencies of these classes in time to come. Familiar examples of analytic studies are found in intelligent city-planning. More familiar studies are the differential effects of varieties and of treatments in agriculture and entomology. The particular crops that are measured are of interest only because they aid decisions on what varieties and treatments to use for the best results in crops *yet to be planted*. We may run an experiment with a group of test animals or with patients in a hospital, but when we generalize from these tests we are thinking of the production process: what will it produce in the future? The present tests are important only because they help us to prescribe or to modify the treatment for future use. The control-chart is a splendid example, the purpose being to control the production process and the quality of lots yet to be made. Other examples are medical and social studies wherein interest centers in the causes that produce differences in health, fertility, or death-rate in different segments of a population of people. Current population surveys in the United States and Canada aid studies of employment, unemployment, farm and industrial labor, school attendance, etc. The monthly sample of deaths by causes, published by the National Office of Vital Statistics in the United States, aids in the control of epidemics and the spread of disease. Its use is both enumerative and analytic.

Special reference to the statistical control of quality. Both the enumerative and analytic problems present themselves hourly in the statistical control of quality. A batch of product has been produced, let us suppose, and the machine is already producing another batch. Two questions arise: B (analytic). Shall we leave the machine alone, or shall we adjust it? Shall we make it run slower or faster, or shall we change the type of chemical bath? A (enumerative). What shall we do with the batch of product just made? Shall we send it on to the next operation (which might be into the consumer's market, or into the next bay of the same factory for further work)? Or is the product so defective that we must re-work it, sell it as second-class, or scrap it?

A chemical engineer whose specialty is the production process may have a special interest in Problem B, and little in Problem A, which he leaves to someone else. On the other hand, if we are the purchaser of a batch of product, such as a single automobile, or some paint for our home, or a carpet, we certainly have a special interest in Problem A. We wish to know the quality of this particular batch of product. It is little comfort to know that the process by which it was made was a good one, and was in a fine state of statistical control, if the product that we ourselves purchase turns out to be defective and unsuited to our purpose. A manufacturer, on the other hand, must purchase raw materials and assemblies in quantities, week in and week out. In order to cut the costs of these materials and to improve their quality, he must concern himself not only with Problem A, the inspection of these materials upon receipt; he must in addition take a lively interest in Problem B, the control of the production processes in the plants of his suppliers.

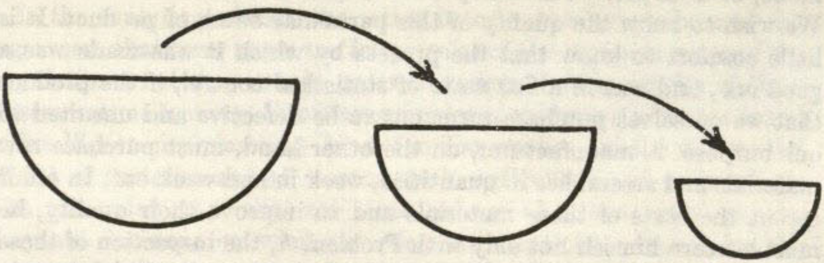
The methods of the Shewhart control chart are essentially analytic, as they tell when to take action on the process. In contrast, the methods of acceptance sampling are primarily enumerative, dealing with the disposal of a lot, although they react secondarily on the process by forcing better control where needed.

II. THE SAMPLING VARIANCES FOR THE TWO TYPES OF DISTRIBUTION

The two uses re-stated in terms of sampling distributions. Re-stated in terms of a mechanism for carrying out the sampling, we may distinguish between the two uses (enumerative and analytic) by consideration of the two distinct types of repetition of the operations that lead to two distinct sampling distributions. In the enumerative case, we take repeated random samples from the same lot, and seek the sampling distribution of the mean or of other statistical measures of these samples. In the analytic case, we take repeated random lots from a supply or cause system, and we select a random sample from each lot; then seek the sampling distribution of the mean or of other statistical measures of these samples.

The use to which the data will be put determines which of the two types of repetition is applicable in any one problem. Unfortunately, sometimes we require data from the same survey to serve both purposes.

It is helpful to look at a diagram. The figure shows three bowls with poker chips, all physically similar, some red and some white. By stirring the contents of any bowl thoroughly, and reaching in blind-folded, it is possible to satisfy satisfactorily the conditions for a random sample. Another way is to give serial numbers to the chips and to draw



Cause System or Production Process	Lot	Sample
Initially	*Initially*	
Mp red	NP red	r red
Mq white	NQ white	$n-r$ white
M total	N total	n total

them with random numbers. The bowl on the left represents the process or cause system. It is a supply of chips. The bowl in the middle represents the lot. It is the people in an area today, or a batch of product, or a crop. The lot has come, we suppose here, as a random sample from the process. This assumption is over-simplified; nevertheless it is a first step to an understanding of analytic problems. The small bowl at the right represents a sample drawn from the lot.

The four possible different variances. Now we are able to state the two problems in terms of estimation. In the enumerative problem the sample is used for an estimate of the contents of the lot, which is described by the proportions P and Q. In the analytic problem the sample is used for an estimate of the contents of the supply, which is described by the proportions p and q. The same sample serves both purposes, but not equally well, for the two estimates have different variances. Hence the proper size of sample, and how to interpret a sample, will depend on whether the aim is enumerative or analytic.

The variances of r/n as estimates of p and of P are shown in the accompanying table. There are four cases, A, B, C, D, depending on how the lot and the sample are drawn.[2] There are two interpretations (enumerative and analytic) in each case.

Table of the variances of $\widehat{p}$ and of $\widehat{P}$
In all cases, $E\widehat{p} = p$ and $E\widehat{P} = P$

The N balls in the lot-container are drawn from the supply	The sample of n is drawn from the lot-container: With replacement	Without replacement
	Case A	Case B
Analytic	$\operatorname{Var} \widehat{p} = \frac{pq}{n}\left\{1 + \frac{n-1}{N}\right\}$	$\operatorname{Var} \widehat{p} = \frac{pq}{n}$
WITH REPLACEMENT		
Enumerative	$\operatorname{Var} \widehat{P} = \frac{PQ}{n}$	$\operatorname{Var} \widehat{P} = \frac{N-n}{N-1}\,\frac{PQ}{n}$
	Case C	Case D
Analytic	$\operatorname{Var} \widehat{p} = \frac{pq}{n}\left\{1 + \frac{n-1}{N}\,\frac{M-N}{M-1}\right\}$	$\operatorname{Var} \widehat{p} = \frac{M-n}{M-1}\,\frac{pq}{n}$
OUT WITH REPLACEMENT		
Enumerative	$\operatorname{Var} \widehat{P} = \frac{PQ}{n}$	$\operatorname{Var} \widehat{P} = \frac{N-n}{N-1}\,\frac{PQ}{n}$

APPLICATIONS

Tabulation plans. The two variances, Var $\widehat{p}$ and Var $\widehat{P}$, are different. A sample therefore contains different amounts of information for the two purposes. How do these observations help us in the design of samples? They tell us that if there is a definite enumerative aim in finding out how many people there are with a given characteristic however rare, then the tabulation and printing of small cells may be justified, provided the universe will still retain enough of its characteristics by the time the sample is tabulated.

On the other hand, if the aim is analytic, there is the sampling error $\sigma_{\widehat{p}}$ even if the figure comes from a complete census, and this error may become troublesome in small cells. It will then be well to economize by

[2] For the derivations of the variances see the author's *Some Theory of Sampling* (John Wiley, 1950), p. 254.

using sampling in the collection of such data, and to determine in advance what consolidations may be made in the tabulating and in the printing. Much space and money are wasted annually on the tabulation of cells that are too small for analytic use, and which have no enumerative use. Too often the excuse for tabulating small cells in a complete census is that they came from a complete census and must therefore be correct. When the use of such tables is analytic only, such arguments do not hold: a reconsideration is due.

Case B is one that often corresponds approximately to many problems in real life. For enumerative use we take the proportion r/n in the sample to estimate the proportion P in the lot. The variance of the estimate

$$\hat{P} = \frac{r}{n} \tag{1}$$

is seen in the table to be

$$\text{Var } \hat{P} = \frac{N-n}{N-1} \frac{PQ}{n}, \tag{2}$$

which reduces to 0 if the sample n is increased to a complete census of the lot, when $n = N$.

In contrast, for analytic purposes in Case B we use the proportion r/n in the sample to estimate the proportion p in the supply. The variance of the estimate

$$\hat{p} = \frac{r}{n} \tag{3}$$

is seen in the table to be

$$\text{Var } \hat{P} = \frac{pq}{n}. \tag{4}$$

The size N of the lots, although they furnished the samples, does not enter into this variance at all. The size N only limits the size of the sample: it cannot be bigger than the lot. To reach greater precision than pq/N (the variance of a complete census) we must draw another lot and sample it also; then combine the two estimates of p.

Effectiveness of a medical treatment. We may see from the following example the two possible ways of interpreting data. Three hundred ninety-eight patients in a hospital went under treatment for a certain disease. After two months a random sample of 250 patients drawn from the list of the original 398 showed no sign of recurrence. How do we figure the odds that there are 0, 1, 2, etc., recurrences in the remaining 148 patients? Stated another way, what is the highest number of recurrences in the original 398 that would permit such a result as often as once in 20 repetitions of the sample? This question is framed as an enumerative one.

This sample was drawn without replacement, and we may see from either Case B or D that the probabilities to apply are the terms of the hypergeometric series.

Let K be the number of patients in the original 398 who actually have a recurrence, none of which showed up in the sample of 250. Then the probability of the result obtained is given by the hypergeometric term

$$P_e = \frac{\binom{398-K}{250}\binom{K}{0}}{\binom{398}{250}}$$

$$= \frac{398-K}{398}\,\frac{398-K-1}{397}\,\frac{398-K-2}{396} \text{ etc. to 250 fractions.} \quad (5)$$

A few results are tabled here, whence we may conclude with odds of about 19:1 that there are not more than 3 recurrences in the lot.

$K = NP$	P	P_e
0	0	1
3	.008	0.051
4	.010	.019
5	.012	.007
6	.015	.002
7	.018	.001

Now suppose that the problem is to predict the proportion that would be cured in a *succession of lots* of patients. This is an analytic question, and the theory to use here then is the binomial series. If p is the proportion of recurrences in the general population (from which by assumption now the 398 patients is a random sample), then the probability of the observed result is

$$P_a = q^{250}p^0 = q^{250}. \tag{6}$$

It is interesting to note that the size 398 of the lot does not come into this probability at all.

A few results are in the table here, whence by interpolation we may conclude with odds of about 19:1 that the proportion p of recurrences in the general population would not be more than 12 in 1000.

p	P_a
0	1
.005	.286
.010	.081
.015	.023
.020	.006
.025	.0002

This example, though oversimplified, may help to guide the design and interpretation of the results of samples. It shows specifically how the interpretation changes when we change our aims from the enumerative to the analytic use.

Allocation of sample. In the symbolism just introduced, the analytic aim is to measure the difference between the two proportions p_1 and p_2 which exist in two cause systems, or to find out if there is any significant difference between p_1 and p_2. We cannot examine the cause systems directly; we can only study two groups of farms, plots, patients, or pupils that the two cause systems have produced. That is, we can only study two *lots*, one from one cause system, and one from another. We shall assume that Case B fits the actual events.

The lots may be of different sizes, N_1 and N_2. At any rate, we take samples therefrom of sizes n_1 and n_2, and we ask what should their sizes be to minimize $Var\ (\hat{p}_2 - \hat{p}_1)$, an analytic purpose. The optimum allocation for this purpose requires that

$$n_1 = k\sigma_1 = k\sqrt{p_1q_1} \tag{7}$$

$$n_2 = k\sigma_2 = k\sqrt{p_2q_2} \tag{8}$$

where

$$k = n/(\sigma_1 + \sigma_2). \tag{9}$$

Now usually, if not always, such problems require the aid of statistical techniques only if p_1 and p_2 are not far apart; if the difference between them were wide, we could observe it without help. Hence we may properly take

$$n_1 = n_2$$

as optimum. Thus, the usual practice of taking equal sizes of sample for clinical or laboratory tests is correct for minimizing the variance of the estimated difference, regardless of the sizes of the populations or of the acreages of the crops whence the samples were drawn.

In contrast, for the enumerative purpose of estimating the over-all average $\bar{x}$ or total X of some characteristic (average rent, total number of unemployed, total acreage in wheat) in the two lots of size N_1 and N_2, the aim is to minimize the variance of $\bar{x}$ or of X, an enumerative purpose. To this end, the optimum size of sample will be

$$\left.\begin{aligned} n_1 &= nN_1/N \\ n_2 &= nN_2/N \end{aligned}\right\} \text{[for proportionate allocation]} \tag{10}$$

or, if one prefers,

$$\left.\begin{aligned} n_1 &= kN_1\sigma_1 \\ n_2 &= kN_2\sigma_2 \end{aligned}\right\} \text{[for disproportionate allocation]} \tag{11}$$

where

$$k = n/(N_1\sigma_1 + N_2\sigma_2).$$

Obviously, the optimum allocations for the two purposes will be different except when the two lots N_1 and N_2 are nearly equal in size.

Unfortunately, the purposes of a survey are often both analytic and enumerative. In a survey to assist the marketing of a certain brand of frozen orange juice, we need to know not only how many people of various income levels buy frozen orange juice of a particular brand, but of all brands, and tinned unfrozen juices as well, and probably fresh fruit besides. These are *enumerative* counts. Then also, probably more important, the survey must discover *why* people of various groups buy or do not buy frozen orange juice and the products that compete with it. This kind of question is *analytic*. The design that is economical for one type may not be economical for the other.

In another study, a research worker wishes to study the variation in the behavior of people, classified by age, education, marital status; perhaps also by religion, urban and rural residence, and occupation. Frequencies are important; but so also are the contributing causes. Thus, this research presents also both enumerative and analytic problems.

One way out is to conduct two surveys—one for enumerative purposes to ascertain the frequencies of certain behavior by class; another, with a more intensive questionnaire, to study the causes.

Another solution is to make some sort of compromise in the design, sacrificing economy and precision in (e.g.) the enumerative results, in order to gain something for the analytic uses. How much to sacrifice, how far to lean, and which way, can be settled only with consideration of the risks and of the losses of making a wrong decision on the basis of information not sufficiently precise, and on consideration of the additional cost of getting more precise information.

A note on acceptance sampling. The probabilities that one encounters in the analytic problem in Case B justify the customary 3-sigma control limits in the form

$$\bar{x} \pm \frac{3\bar{R}}{d_2\sqrt{n}} \quad (\bar{R} \text{ is the average range over a series of samples, and } \bar{R}/d_2 \text{ is an estimate of } \sigma)$$

for the $\bar{x}$-chart, or

$$\bar{p} \pm 3\sqrt{\frac{\bar{p}\bar{q}}{n}} \quad (\bar{p} \text{ is the average fraction defective over a series of samples})$$

for the p-chart, as an aid for detecting uncontrolled variability. It will be observed that the size N of the lot does not appear in these equations even if the sample-size n is 100 per cent of N. This form of computation is now seen to be correct; it is not an approximation. The justification is the absence of N and of any finite multiplier in the analytic Case B.

On the other hand, many writers in dealing with the *producer's risk* in acceptance sampling (the probability that a lot of acceptable quality will be rejected) have recommended hypergeometric terms (like that in Eq. 5), or rather have reluctantly used binomial terms as approximations to the hypergeometric terms. Actually, however, the producer is concerned with the problem of keeping his process in control and at a desired level p. The quality P of the lots that he produces will vary from lot to lot, yet the risk (probability) that a lot will be rejected on a single-sampling plan turns out to be a sum of binomial terms typified by

$$\binom{n}{r} q^{n-r} p^r,$$

into which the size N of the lot does not enter at all. This problem behaves like the *analytic* Case B, for which the binomial terms are correct; they are not approximations.

The *consumer's risk* is another story. The consumer is concerned with the particular lots of product that he is purchasing, and he has a sample from each lot on the basis of which to decide whether to accept or to reject the lot. He may aim to guard against accepting lots with too high a value of P, regardless of p and of the state of control. To compute on this basis the correct probabilities for the consumer's risk, one requires hypergeometric terms. The finite multiplier $(N-n)/(N-1)$ then appears in the variance of $\widehat{P}$, because this is the enumerative Case B.

There are thus, strictly, two operating-characteristic (O.C.) curves, one for the producer, another for the consumer. In practice, however, except when the sample is 20% or more of the lot, the two curves coincide, almost, fortunately, and one curve suffices.

The distinction made here between the different probabilities for the producer's and consumer's risks is not new. It forms the basis for the Dodge-Romig tables, as is clear from their text. An extremely lucid exposition appears in the book *Sampling Inspection* by the Statistical Research Group, Columbia University (McGraw-Hill, 1947), pages 183 and 184. I quote:

> There are two alternative ways of interpreting "percentage of defective items in submitted product," and these lead to somewhat different O.C. curves for small inspection lots. (a) The percentage of defective items can be considered as applying to each inspection lot separately. . . . (b) The percentage of defective items can be considered as applying to the process. . . . If interpretation (b) is adopted, the resultant O.C. curve does not depend on the size of the inspection lot.