NATIONAL BUREAU OF STANDARDS REPORT

1621

EMPIRICAL SAMPLING DISTRIBUTIONS II

by

D. Teichroew

Institute of Statistics

University of North Carolina

and

National Bureau of Standards

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

U. S. DEPARTMENT OF COMMERCE
Charles Sawyer, Secretary

NATIONAL BUREAU OF STANDARDS
A. V. Astin, Acting Director

THE NATIONAL BUREAU OF STANDARDS

The scope of activities of the National Bureau of Standards is suggested in the following listing of the divisions and sections engaged in technical work. In general, each section is engaged in specialized research, development, and engineering in the field indicated by its title. A brief description of the activities, and of the resultant reports and publications, appears on the inside of the back cover of this report.

1. ELECTRICITY. Resistance Measurements. Inductance and Capacitance. Electrical Instruments. Magnetic Measurements. Electrochemistry.

2. OPTICS AND METROLOGY. Photometry and Colorimetry. Optical Instruments. Photographic Technology. Length. Gage.

3. HEAT AND POWER. Temperature Measurements. Thermodynamics. Cryogenics. Engines and Lubrication. Engine Fuels.

4. ATOMIC AND RADIATION PHYSICS. Spectroscopy. Radiometry. Mass Spectrometry. Physical Electronics. Electron Physics. Atomic Physics. Neutron Measurements. Nuclear Physics. Radioactivity. X-Rays. Betatron. Nucleonic Instrumentation. Radiological Equipment. Atomic Energy Commission Instruments Branch.

5. CHEMISTRY. Organic Coatings. Surface Chemistry. Organic Chemistry. Analytical Chemistry. Inorganic Chemistry. Electrodeposition. Gas Chemistry. Physical Chemistry. Thermochemistry. Spectrochemistry. Pure Substances.

6. MECHANICS. Sound. Mechanical Instruments. Aerodynamics. Engineering Mechanics. Hydraulics. Mass. Capacity, Density, and Fluid Meters.

7. ORGANIC AND FIBROUS MATERIALS. Rubber. Textiles. Paper. Leather. Testing and Specifications. Organic Plastics. Dental Research.

8. METALLURGY. Thermal Metallurgy. Chemical Metallurgy. Mechanical Metallurgy. Corrosion.

9. MINERAL PRODUCTS. Porcelain and Pottery. Glass. Refractories. Enameled Metals. Building Stone. Concreting Materials. Constitution and Microstructure. Chemistry of Mineral Products.

10. BUILDING TECHNOLOGY. Structural Engineering. Fire Protection. Heating and Air Conditioning. Exterior and Interior Coverings. Codes and Specifications.

11. APPLIED MATHEMATICS. Numerical Analysis. Computation. Statistical Engineering. Machine Development.

12. ELECTRONICS. Engineering Electronics. Electron Tubes. Electronic Computers. Electronic Instrumentation.

13. ORDNANCE DEVELOPMENT. Mechanical Research and Development. Electromechanical Fuzes. Technical Services. Missile Fuzing Research. Missile Fuzing Development. Projectile Fuzes. Ordnance Components. Ordnance Tests. Ordnance Research.

14. RADIO PROPAGATION. Upper Atmosphere Research. Ionospheric Research. Regular Propagation Services. Frequency Utilization Research. Tropospheric Propagation Research. High Frequency Standards. Microwave Standards.

15. MISSILE DEVELOPMENT. Missile Engineering. Missile Dynamics. Missile Intelligence. Missile Instrumentation. Technical Services. Combustion.

NBS PROJECT
3011-60-0002

4-4-52

NBS REPORT
1621

EMPIRICAL SAMPLING DISTRIBUTIONS II[1,2]

by

D. Teichroew

Institute of Statistics, University of North Carolina

and

National Bureau of Standards

PREPRINT

[1]Address delivered at the meeting of the Institute at Boston, Massachusetts on December 26, 1951.

[2]This report was completed under a National Bureau of Standards contract with the University of North Carolina.

NBS
4-4-52
PREPRINT

Empirical Sampling Distributions II[1,2]

by

D. Teichroew

Institute of Statistics, University of North Carolina

and

National Bureau of Standards

[1]Address delivered at the meeting of the Institute at Boston, Massachusetts on December 26, 1951.

[2]This report was completed under a National Bureau of Standards' contract with the University of North Carolina.

1. INTRODUCTION

The calculation of distribution functions is one of the important tasks in mathematical statistics. Of particular importance are the sampling distributions of statistics used in testing hypotheses. These require the determination of the distribution of a function of, say n, random variables, each of which has the same basic distribution. The sampling distribution can be expressed in the form of a definite integral and, if this integral can be evaluated analytically, it is only necessary to prepare a table of the resulting function, if not, the table may be obtained by numerical integration.

An empirical method of determining the sampling distribution suggests itself immediately from the formulation of the problem. This method consists of taking, say N, samples, each consisting of n observations on the random variable and calculating the value of the statistic for each sample. The "empirical" distribution can then be obtained by arranging the values of the statistic in order of size and counting the proportion that are less than, or equal to a given value. (In practice, it is usually desirable to apply a smoothing process to the resulting distribution).

The empirical method has been used since it was introduced by Student [6], without the indication that it could be considered as a way of getting approximate values for a definite integral. In recent years, much attention has been directed towards the numerical solution of difficult mathematical problems. It has

been found that probability interpretations may be used, not only to evaluate definite integrals, but also to invert matrices and to solve differential equations. The general class of methods has been given the code name of Monte Carlo and now has a fairly extensive literature, (see for example, Curtiss [2], Kahn [3], and The Monte Carlo Method [8]).

An outstanding contribution of Monte Carlo theory is the use of so-called "importance sampling". The concept will be developed in Section 3, the word "importance", however, will be avoided since it is not particularly appropriate. As far as we know, no one has yet utilized this concept in calculating sampling distributions. In this paper, some of the theoretical aspects of such a procedure will be examined, with particular attention to the practical aspects of using the method on high-speed electronic calculators.

It might be mentioned here that empirical sampling, because of probability considerations, cannot, with a practical number of samples, give results which are accurate to many significant figures. Even with the fastest electronic calculators now being built, statisticans, by virtue of economic necessity, may have to be satisfied with thousands of samples instead of millions. Nevertheless, there are many practical problems which could be solved by results obtainable most economically by empirical sampling. For example, one such problem is to determine which of a number of non-parametric tests is most powerful against a

specified class of alternatives. A problem of this type is being prepared for computation by the SWAC (National Bureau of Standards Western Automatic Computer).

2. STATEMENT OF PROBLEM AND NOTATION

The basic model consists of:

(i) A random variable x which has a density function f(x)

(ii) A sample size n

(iii) A statistic s which is a function of n independent observations on x.

In any practical problem, various values of n and various functional forms of s may arise, and this may affect the method of solution but for the present only one value of n and one form of s will be used.

The problem is to calculate (and tabulate) the sampling distribution of s, i.e. the probability that the statistic will be less than or equal to some given value. Let H(s) denote this function of s, then

$$H(s) = \text{Prob.}\left\{ s(X_1, X_2, \cdots X_n) \leq s \right\}$$
$$= \int \cdots \int f(x_1, x_2, \cdots x_n)\, dx_1\, dx_2 \cdots dx_n \qquad (2.1)$$

where the integration is taken over the region in the n dimensional sample space in which

$$\left\{ s(x_1, x_2, \cdots x_n) \leq s \right\}$$

In accordance with standard usage, capital letters will indicate observed values, a dash beneath a letter denotes a vector, and the n dimensional space of possible sample values will be

denoted by R_n. In addition $f(\underline{x})$ will be written for $f(x_1, x_2 \cdots x_n)$; $d\underline{x}$ for $dx_1 \cdots dx_n$, and a single integral sign will be used for n-fold integration. The range of summation will be indicated by the subscript, summation over individuals in a sample will be for j from 1 to n and summation over samples will be for i from 1 to N. In addition it will be convenient to define:

$$A(s) = \left\{R_n : s(\underline{x}) \leq s\right\} \tag{2.2}$$

= set in R_n in which $s(\underline{x}) \leq s$.

$$t(\underline{x}) = 1 \text{ if } s(\underline{x}) \leq s \qquad \text{i.e if } \underline{x} \subset A(s) \tag{2.3}$$

$$= 0 \text{ otherwise} \qquad \text{i.e if } \underline{x} \subset R_n - A(s)$$

$$g(\underline{x}) = \text{any n dimensional density function} \tag{2.4}$$

i.e $g(\underline{x}) \geq 0$ for all $\underline{x}$ in R_n; $\int_{R_n} g(\underline{x})\, d\underline{x} = 1$.

$$v(\underline{x}) = \frac{t(\underline{x})\, f(\underline{x})}{g(\underline{x})} \tag{2.5}$$

$$w(\underline{x}) = \frac{f(\underline{x})}{g(\underline{x})} \tag{2.6}$$

With this notation (2.1) becomes

$$H(s) = \int_{A(s)} f(\underline{x})\, d\underline{x} = \int_{R_n} t(\underline{x})\, f(\underline{x})\, d\underline{x} \tag{2.7}$$

3. SAMPLING MODELS

Equation (2.7) can be regarded either as giving the value of H(s) as a definite integral or as stating that if a random variable $\underline{x}$ has a density function $f(\underline{x})$ then:

$$E\left\{t(\underline{x})\right\} = H(s)$$

since the expectation operator is defined by just such an equation (Cramer [1], p. 170). This means that the average value of $t(\underline{X})$ in N observations on $\underline{x}$ will tend to H(s) as N increases, i.e. the variance or the dispersion of these average values around H(s) is a decreasing

function of N. Therefore one way of estimating H(s) is to take many samples (according to the basic model) and calculate the average value of $t(\underline{X})$.

For this case the behaviour of the estimator is well-known (and is furthermore independent of $f(\underline{x})$) since $t(\underline{X})$ is a binomial variate; i.e.

$$t(\underline{X}) = 1 \text{ with probability } H(s)$$
$$= 0 \text{ with probability } 1 - H(s)$$

and therefore it follows that the estimator

$$p_1(s) = \frac{1}{N} \Sigma \; t(\underline{X}_i) \tag{3.1}$$

has

$$E\left\{p_1(s)\right\} = H(s) \tag{3.2}$$

$$N \text{ Var } p_1(s) = H(s) - [H(s)]^2 \tag{3.3}$$

This sampling model has consisted of two steps:

1. The space R_n is divided into two mutually exclusive sets of points, A(s) and $R_n - A(s)$.

2. Points are selected from R_n, the probability of a point coming from A(s) is H(s) and the probability of it coming from $R_n - A(s)$ is $1 - H(s)$. An estimate of H(s) therefore is the proportion of points coming from A(s). The important point here is that the first part is a matter of geometry only and is independent of probability considerations.

From a consideration of step 2, it is clear that the probability of a point coming from A(s) does not have to be H(s) and it might be possible to reduce the variance of the estimate of H(s) by changing

it. The mechanics of the procedure are apparent if the integrand in (2.7) is multiplied and divided by a density function $g(\underline{x})$. The equation becomes

$$H(s) = \int_{R_n} t(\underline{x}) \frac{f(\underline{x})}{g(\underline{x})} \; g(\underline{x}) \; d\underline{x} = \int_{R_n} v(\underline{x}) \; g(\underline{x}) \; d\underline{x} \qquad (3.4)$$

(where $v(\underline{x})$ is defined by (2.5)) and is interpreted as stating that

$$E\left\{v(\underline{x})\right\} = H(s).$$

where $\underline{x}$ has a density function $g(\underline{x})$. Now, in step 2, the probability that a point comes from A(s) is

$$\int_{A(s)} g(\underline{x}) \; d\underline{x}$$

and an estimate of H(s) is the average of $v(\underline{X}_i)$. The basis of "importance" sampling is to choose a $g(\underline{x})$, such that regions in R_n which have a large contribution to the variance of the estimate are sampled more heavily than other regions (since the variance is a decreasing function of N) and therefore the overall variance of the estimate of H(s) will be reduced.

The word "importance" will not be used in this paper, instead the two sampling models will be referred to as Model I and Model II. Both of these models will give an empirical answer to the problem outlined in section 2 and differ in the distribution of the random variable which is sampled. These distributions are as follows:

Model I: The random variable sampled has density $f(\underline{x})$.

Model II: The random variable sampled has density $g(\underline{x}) \neq f(\underline{x})$.

IIa: $g(\underline{x}) = g(x_1) \; g(x_2) \cdots g(x_n)$

IIb: $g(\underline{x})$ is not factorable as in IIa.

Any statement made about Model II without referring to a or b, applies to both.

It may be noted that Model II sampling bears a certain analogy to the stratified sampling used in surveys. A survey, in its simplest form, consists of selecting a sample of m individuals from a population of M individuals and observing the value of a characteristic on each of the selected individuals. The average value obtained in this way is an estimate of the average of that characteristic for the population. If the population can be divided into k strata, in each of which the characteristic has a smaller variance than in the whole population, an estimate of the mean with smaller variance can sometimes be obtained by allocating different numbers of samples to the different strata. Formulas can be determined for the optimum sample sizes for each strata on the basis of strata variances and number of individuals in each strata.

In Model II sampling the population consists of the points $\underline{x}$ of the n dimensional sample space and the characteristic is $s(\underline{x})$. A strata consists of points which have their values of s in a certain interval. The allocation of samples to the different strata is accomplished by choosing a $g(\underline{x})$.

4. ESTIMATES OF H(s)

The problem, under Model II sampling might be looked at as a selection of a $g(\underline{x})$, such that the variance of some estimator of H(s) is a minimum. The minimum turns out to be zero, as has been shown by Kahn [3]. This trivial solution arises because the problem is, incorrectly, being considered as one in statistical inference. In statistical inference a sample is taken from a distribution which is a function of an unknown parameter and it is desired to estimate

this unknown value from information obtained from the sample. In the present problem the distribution from which a sample is taken is completely specified and from this sample an "estimate" of H(s) is desired. However, H(s) is not unknown because its value is given by the integral (2.1). Quite naturally, therefore, the trivial solution states that an estimate of H(s) with zero variance can be obtained by evaluating the integral. The word "estimate" will be used in this paper in the special sense that the unknown quantity is not a parameter of the distribution which is being sampled.

For sampling under Model I, the estimator (3.1) is the only reasonable one to use. Its variance (3.3) and, in fact, its sampling distribution are known and do not depend on the functional form of $f(\underline{x})$. Under Model II the problem is to estimate the expected value of $v(\underline{x})$. This is a much more complex problem because the form of the distribution of $v(\underline{x})$ may depend, not only on $g(x)$, but also on $f(\underline{x})$. Hence, the average value may not be a desirable estimator of $E\left\{v(\underline{x})\right\}$ and furthermore, this distribution may be such that the variance may not be a suitable criterion of desirability. Therefore the choice of estimator will depend on the particular situation. For the general discussion, the variance will be used and two classes of estimators having general applicability will be considered, first those depending on the sample values and second, those depending on the ordered observations only.

Of the first class, the most natural one is the sample average, which has already been mentioned in the last section. Let

$$p_2(s) = \frac{1}{N} \sum v(\underline{X}_i) = \bar{v} \quad . \qquad (4.1)$$

This estimator has the unfortunate property that $p_2(s^*)$ where s^* is the largest observed value of s, is in general not equal to

unity, i.e. $p_2(s)$ is not a distribution function in the usual sense. The estimator may be normalized to make it run from zero to one by dividing by $p_2(s^*)$. This leads us to consider the "ratio" estimator,

$$p_3(s) = \frac{\Sigma\, v(\underline{X}_i)}{\Sigma\, w(\underline{X}_i)} = \frac{\bar{v}}{\bar{w}} \qquad (4.2)$$

where $w(\underline{x})$ is defined by (2.6).

Under the class of estimators depending on the ordered values of $v(\underline{X}_i)$, estimators such as the median, etc., might be considered. The problem of forcing the empirical sampling distribution to cover the interval from zero to one would arise here too. Investigation of this class is of secondary importance, because the time and memory space required to rank the $s(\underline{X}_i)$ increases too rapidly as N increases to make the use of these order statistics attractive for high speed computing machines.

5. COMPARISON OF VARIANCE OF ESTIMATORS

$p_1(s)$ obtained under Model I has expectation and variance as given by (3.2) and (3.3) respectively. For any particular value of s, confidence limits for H(s) may be calculated by exact binomial theory if N is small or the asymptotic normality of $p_1(s)$ if N is large. Confidence limits for the entire function H(s) of the form

$$p_1(s) - \lambda N^{-\frac{1}{2}} < H(s) < p_2(s) + \lambda N^{-\frac{1}{2}}$$

may be obtained by the method given by Kolmogorov [4] if N is large or by tables given by Massey [5] if N is small. Wald and Wolfowitz [7] have given a method for calculating confidence limits of other forms.

$p_2(s)$ is defined by (4.1). Since

$$E\left\{v(\underline{x})\right\} = H(s)$$

and $$\text{var } v(\underline{x}) = \int_A [f(\underline{x})]^2 [g(\underline{x})]^{-1} d\underline{x} - [H(s)]^2, \tag{5.1}$$

it follows that

$$E\left\{p_2(s)\right\} = H(s) \tag{5.2}$$

$$N \text{ var } p_2(s) = \int_A [f(\underline{x})]^2 [g(\underline{x})]^{-1} d\underline{x} - [H(s)]^2. \tag{5.3}$$

$\therefore$ Var, $p_2(s) = 0$ if

$$g(\underline{x}) = f(\underline{x}) [H(s)]^{-1} \quad \text{if } \underline{x} \subset A(s)$$
$$= 0 \qquad\qquad\qquad \text{otherwise} \tag{5.4}$$

This is the zero variance solution for $g(\underline{x})$ mentioned in section 4. It is not of much use in actual problems because it requires a knowledge of $H(s)$; applies to only one value of s and furthermore is defined in terms of the set $A(s)$. If Model IIa were used, where $g(\underline{x}) = \pi g(x_j)$, $g(\underline{x})$ could not be defined, in general, in terms of $A(s)$, unless the statistic were of a similar form, i.e. $s(\underline{x}) = \pi s(x_j)$.

Next consider the more modest requirement of choosing $g(\underline{x})$ so that var $p_1(s) >$ var $p_2(s)$, i.e.

$$\int_A f(\underline{x})\, d\underline{x} > \int_A [f(\underline{x})]^2 [g(\underline{x})]^{-1} d\underline{x} \ . \tag{5.5}$$

A sufficient condition for this inequality is

$$g(\underline{x}) > f(\underline{x}) \text{ for all } \underline{x} \in A(s) \ . \tag{5.6}$$

This however is not a necessary condition.

Since $p_3(s)$ is a ratio of two random variables $\bar{v}(\underline{x})$ and $\bar{w}(\underline{x})$ its distribution, or even its expected value and variance cannot, in general, be calculated explicitly. The best we can do is to follow the procedure usually used in a situation of this kind and consider the expansion of a function of two variables, $\bar{v}$, and $\bar{w}$ in a Taylor series about the point $E(\bar{v})$, $E(\bar{w})$. Taking expected values of both sides of the equation gives the expected value of the ratio. A similar series can be obtained for the variance, etc. If an upper bound can be found for the sum of the series beginning with the (m+1)th term the first m terms can be used as an approximation with a known maximum error. This is difficult to do in the general case but in any particular situation it may be possible.

In the general case, if the terms are decreasing functions of N, as N becomes large, all terms except the first in the series will become negligible. These statements are made precise in the theorem given by Cramer [1] p. 355 and 366. Here we need only the expected value and variance of which the first terms are

$$E\, p_3(s) \sim H \tag{5.7}$$

$$\text{var}\, p_3(s) \sim N^{-1} \left\{ [H(s)]^2 \text{ var } w + \text{var } v - 2\, H(s) \text{ cov } (v,w) \right\}. \tag{5.8}$$

It will be noticed that N^{-1} var v is the var $p_2(s)$ therefore $p_3(s)$ can have a lower variance than $p_2(s)$ only if cov $(v(\underline{x}), w(\underline{x}))$ is positive. Since

$$\text{cov } (v(\underline{x}), w(\underline{x})) = \int_A [f(\underline{x})]^2 [g(\underline{x})]^{-1} d\underline{x} - H(s) \tag{5.9}$$

var $p_3(s) <$ var $p_2(s)$ only if

$$\int_A [f(\underline{x})]^2 [g(\underline{x})]^{-1} d\underline{x} > \int_A f(\underline{x})\, d\underline{x} \quad . \tag{5.10}$$

This is the opposite inequality to the one given in (5.5) for var $p_2(s) <$ var $p_1(s)$. The two methods of estimation are to some extent complementary.

6. COMPUTATION

Before considering some of the applications of Model II sampling it may be worthwhile to outline the changes in computational procedure that would be necessary. The computations carried out by a high speed machine in computing an empirical sampling distribution under Model I may be divided into three steps:

1. The generation and testing of random numbers and the transformation of these random numbers to observations from $f(\underline{x})$.
2. The calculation of the statistic from n of these observations.
3. The preparation of an empirical distribution of N values of the statistic. This could take the form of a printed list of the N values (in order of size, if possible) or a printed frequency distribution, each frequency being a count of the number of values that fell in a certain interval (the intervals having been determined in advance).

Under Model II, step 1 would contain the transformation of random numbers to observations from $g(\underline{x})$. Step 2 would be unchanged. However, in step 3, it is now necessary to record not only the value of the statistic but also the value of $v(\underline{X})$ for each sample. The empirical distribution could take the form of a printed list of N values (in order of size, if possible) together with the value of $v(\underline{X})$ for that sample or a printed distribution, each frequency this time being the sum of the $v(\underline{X})$'s for all samples having the value of the statistic fall in that interval.

The major change arises in the computation of the variance of the estimates. Under Model I, the var $p_1(s)$, may be estimated from $p_1(s)\ (1 - p_1(s))$ for which nothing is required except p_1 itself. Var $p_2(s)$ however must be estimated from the variance of the $v(\underline{x})$'s and if this cannot be done by evaluating the integral (5.1) theoretically, it will be necessary to store the sum of $[v(\underline{X})]^2$ for all samples having values falling in each interval in the frequency distribution. $p_3(s)$ presents an even more difficult problem since, in general, no exact formula for its variance exists. One estimate of var $p_3(s)$ can be obtained by substituting estimates from the sample in (5.8). Another estimate can be obtained by dividing the N samples into k groups of N samples each, estimating p_3 from each group and calculating the variance of the k estimates.

If the computations are carried out on IBM machines, the calculation of $v(\underline{X})$ may be time-consuming. Prof. J. W. Tukey has suggested that the work could be simplified by dividing the x axis into a finite number of intervals and in each one choosing $g(x)$ so that $[f(\underline{x})]\ [g(\underline{x})]^{-1} = a^i$ where a is some constant and $i = 0, \pm 1, \pm 2, \pm 3$, etc. Then $v(\underline{X})$ can be calculated for each sample merely by summing the i's that appear in that sample and raising a to that power. This procedure greatly simplifies the calculations but leads to estimates with large variance unless a is close to unity.

7. APPLICATIONS

It will be advantageous to use Model II sampling only if it reduces the total cost; where the cost may be determined either by trying to estimate H(s) with minimum cost for a certain amount of precision or with maximum precision for a given cost. Since a problem of this type is relatively easy to program for a high speed computer, the major expense in an empirical sampling study will be the cost of time of the computer. The amount of computing time may be decreased either by decreasing the number of samples required or by decreasing the number of computations required for each sample.

If Model I is used the number of samples required can be determined by fixing the value of H(s) and var $p_1(s)$. There may exist a $g(\underline{x})$ such that if it were used with Model II var $p_2(s)$ or var $p_3(s)$ would be substantially lower and hence the number of samples required would be sufficiently smaller to more than offset the increased amount of computing. Another case in which Model II could reduce the number the samples is where the sampling is being carried out mainly to determine the "tails" of the distribution of s. 1000 samples under Model I will not give much information about the value of s_o for which $H(s_o) = .9999$. Under Model II a $g(\underline{x})$ could be selected such that, say, an average of 10 out of every 1000 samples would have $s \geq s_o$ and hence provide an estimate of $H(s_o)$.

Model II sampling could also be used to decrease the computing time required for each sample. The first step in the computation

consists of transforming uniform variates to observations from $f(\underline{x})$. In some cases this involves considerable computation and there may exist a $g(\underline{x})$ such that it is quicker to transform the uniform variates to observations from $g(\underline{x})$ and compute $v(\underline{x})$. Sometimes the distribution of a statistic is required for a set of functions $f_1(\underline{x})$, $f_2(\underline{x})$, ... Instead of using each of these in Model I sampling it may be advisable to select one of them as $g(\underline{x})$ and use Model II, particularly if it is desirable and/or permissable to use the same set of uniform variates for each of the functions. Under this procedure, the value of the statistic would have to be calculated only once for each sample instead of once for each $f_i(\underline{x})$ and this could be a great saving, if, for example the calculation of the statistic involved the inversion of a matrix.

In using Model II sampling, it is therefore necessary to determine $g(\underline{x})$ in a way that will achieve the desired effect. Unfortunately this is a difficult thing to do. Equation (5.5) and (5.10) give conditions under which estimates under Model II sampling may have a lower variance than estimates under Model I. As methods of determining in advance which $g(\underline{x})$ should be used they are not much help since the integrals involved will generally be as difficult to evaluate as the original one. Therefore the answer must come largely from intuition and empirical work and in difficult cases mainly from the latter.

8. NUMERICAL EXAMPLE

The following example has been chosen because the expressions for the variances of $p_2(s)$ and $p_3(s)$ can be evaluated analytically.

Suppose the sampling distribution of the sum of square of n standardized normal deviates is required and that Model II sampling will be used with g(x) a normal distribution with mean zero and variance σ^2. Table I and II give the variance of p_2 and the asymptotic variance of p_3 for different σ^2 and H(s); for n = 1 and 9 respectively. (The numbers given are actually N times the variance). These values were calculated as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$$

$$\frac{[f(x)]^2}{g(x)} = (\sigma\sigma') \frac{1}{\sigma'\sqrt{2\pi}} e^{-x^2/2\sigma'^2} \quad \text{where } \sigma'^2 = \frac{\sigma^2}{2\sigma^2-1}$$

$$\frac{[f(\underline{x})]^2}{g(\underline{x})} = (\sigma\sigma')^n \left(\frac{1}{\sigma'\sqrt{2\pi}}\right)^n e^{-\Sigma x_j^2/2\sigma'^2}$$

$$s(\underline{x}) = \Sigma x_j^2$$

$$A(s) = \left\{R_n : s < s_H\right\}$$

$$\text{where } H = H(s) = \text{Prob.} \left\{s < s_H\right\} = \int_{A(s)} f(\underline{x})\, d\underline{x} = \int_0^{s_H} K_n(s)ds$$

where $K_n(s)$ is the Chi-Square distribution with n degrees of freedom. Therefore

$$\int_{A(s)} [f(x)]^2 [g(\underline{x})]^{-1} d\underline{x} = (\sigma\sigma')^n \int_0^{\frac{s_H}{\sigma'^2}} K_n(s)\, ds$$

and this integral can be evaluated directly from table of the Chi-Square distribution.

The variance under Model I is given by $\sigma^2 = 1$. It may be noticed that if $\sigma^2 < 1$, $p_2(s)$ will sometimes give a smaller variance, if $\sigma^2 > 1$, $p_3(s)$ will sometimes give a smaller variance. For $n = 9$ and $H = .80$ all the variances are given to indicate how rapidly the variances can increase if an inappropriate estimator is used.

April 15, 1952

TABLE I

Comparison of Variances of Estimators
under Models I and II, for n = 1

	H = .80		H = .90		H = .95	
σ^2	Var P_2	Var* P_3	Var P_2	Var* P_3	Var P_2	Var* P_3
0.55	.048		.047		.0871	
0.60	.085		.072		.0929	
0.70	.107		.060		.0419	
0.80	.117		.064		.0378	
0.90	.138		.075		.0385	
1.00	.160	.160	.090	.090	.0475	.0475
1.25		.139		.068		.0318
1.50		.131		.060		.0252
2.00		.127		.052		.0197
3.00		.133		.050		.0171
4.00				.051		.0167
5.00				.053		.0169

*Asymptotic Variance

TABLE II

Comparison of Variances of Estimators
under Models I and II, for n = 9

	H = .80		H = .90		H = .95	
σ^2	Var P_2	Var* P_3	Var P_2	Var* P_3	Var P_2	Var* P_3
0.60	.684	8.222	1.306		2.218	
0.75	.177	.597	.265		.371	
0.80	.135	.391	.159		.191	
0.90	.118	.222	.057		.059	
1.00	.160	.160	.090	.090	.048	.048
1.25	.610	.136		.059		.024
1.50	.947	.119		.046		.017
2.00	2.896	.214		.062		.018

*Asymptotic Variance

REFERENCES

1. H. Cramer, (1946) Mathematical Methods of Statistics, Princeton University Press.

2. J. H. Curtiss, (1949) "Sampling methods applied to differential and difference equations." Proc. of Sem. on Scientific Computation, IBM, New York, p.87-109.

3. H. Kahn, (1949) "Modification of the Monte Carlo method.", Proc. of Sem. on Scientific Computing, IBM, New York, p. 20-27.

4. A. Kolmogorov, (1941) "Confidence limits for an unknown distribution function." Annals of Math. Stat., V. 12, p. 461-3

5. F. J. Massey, (1950) "A note on the estimation of a distribution function by confidence limits." Annals of Math. Stat. V. 21, p. 116-9.

6. Student (1908) "The probable error of a mean", Biometricka V.6, p. 1

7. A Wald and J. Wolfowitz (1939) "Confidence limits for continuous distribution functions", Annals of Math. Stat. V. 10, p. 105-18.

8. Monte Carlo Method (1951) National Bureau of Standards. Applied Mathematics Series No. 12.

THE NATIONAL BUREAU OF STANDARDS

Functions and Activities

The National Bureau of Standards is the principal agency of the Federal Government for fundamental and applied research in physics, mathematics, chemistry, and engineering. Its activities range from the determination of physical constants and properties of materials, the development and maintenance of the national standards of measurement in the physical sciences, and the development of methods and instruments of measurement, to the development of special devices for the military and civilian agencies of the Government. The work includes basic and applied research, development, engineering, instrumentation, testing, evaluation, calibration services, and various scientific and technical advisory services. A major portion of the NBS work is performed for other government agencies, particularly the Department of Defense and the Atomic Energy Commission. The functions of the National Bureau of Standards are set forth in the Act of Congress, March 3, 1901, as amended by Congress in Public Law 619, 1950. The scope of activities is suggested in the listing of divisions and sections on the inside of the front cover.

Reports and Publications

The results of the Bureau's work take the form of either actual equipment and devices or published papers and reports. Reports are issued to the sponsoring agency of a particular project or program. Published papers appear either in the Bureau's own series of publications or in the journals of professional and scientific societies. The Bureau itself publishes three monthly periodicals, available from the Government Printing Office: the Journal of Research, which presents complete papers reporting technical investigations; the Technical News Bulletin, which presents summary and preliminary reports on work in progress; and Basic Radio Propagation Predictions, which provides data for determining the best frequencies to use for radio communications throughout the world. There are also five series of nonperiodical publications: the Applied Mathematics Series, Circulars, Handbooks, Building Materials and Structures Reports, and Miscellaneous Publications.

Information on the Bureau's publications can be found in NBS Circular 460, Publications of the National Bureau of Standards ($1.00). Information on calibration services and fees can be found in NBS Circular 483, Testing by the National Bureau of Standards (25 cents). Both are available from the Government Printing Office. Inquiries regarding the Bureau's reports and publications should be addressed to the Office of Scientific Publications, National Bureau of Standards, Washington 25, D. C.


NBS