DISSERTATION

\- - -

SOME TWO-STEP SAMPLING PROCEDURES

Submitted by

Terrence Lee Connell

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

February 1966

COLORADO STATE UNIVERSITY

February 1 1966

IT IS RECOMMENDED THAT THE DISSERTATION PREPARED BY

Terrence Lee Connell

ENTITLED Some Two-Step Sampling Procedures

be accepted as fulfilling this part of the requirement for the degree of Doctor of Philosophy.

Committee on Graduate Work

F. A. Graybill — Major Professor
Paul W. Mielke Jr.
J M Sutherland
E E Remmenga

Examination Satisfactory

Committee on Final Examination

Paul W. Mielke Jr.
J M Sutherland
E E Remmenga
F. A. Graybill — Chairman

Permission to publish this dissertation or any part of it must be obtained from the Dean of the Graduate School.

ACKNOWLEDGMENTS

Most publications require the efforts of more than one person. This dissertation is no exception. The author wishes to express his gratitude first of all to his wife Susan for the help and encouragement she gave him, and to his major professor, Dr. Franklin A. Graybill, who suggested the research problems, some possible attacks on them, and gave helpful suggestions along the way. Next he wishes to thank Mrs. Wayne Deason who spent many hours typing the contents and doing other tasks for him. He also wishes to thank his graduate committee for reviewing this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

I. THE SAMPLE SIZE PROBLEM

One of the most important statistical problems is estimating the value of an unknown parameter in a given frequency function. If a point estimate is desired and the sample size is not fixed in advance then the experimenter must decide how large a sample should be taken. For most problems the cost of an experiment increases with the sample size. This increase in cost may be in financial terms or perhaps in terms of time or effort. On the other hand, a decrease in sample size may increase the variance of the estimate (loss of precision) or decrease the "closeness" of the estimate to the true value of the parameter, i.e., the probability that the estimate is within a given distance of the true value decreases. Thus the problem is to devise some procedure for determining the smallest sample size which still allows the experimenter to obtain an estimate of the parameter with certain restrictions on precision, closeness, or some other criterion.

An experimenter may prefer to obtain an interval estimate of the unknown parameter. The desirability of small sample size is the same as in the point estimation case. For interval estimation a reduction in sample size will in general increase the width of the interval or decrease the confidence coefficient. The problem is to obtain the smallest sample size possible with certain restrictions, determined by the experimenter, on the width or confidence coefficient of the interval estimate.

When the experimenter has limited financial resources or time, it is possible that he might be forced to relax desired restrictions to reduce sample size. In this case the experimenter must evaluate relative losses between increased sample size and decreased usefulness of the results. This indicates that the quality of results in experiments is often related to financial resources and time restrictions.

Dantzig [12] in 1940 was the first to show that not all sample size problems can be solved by a one-step procedure. In particular, Dantzig's results showed that a one-step procedure cannot be devised to test "student's" hypothesis such that its power function is independent of the variance.

For the class of all distributions for which the mean exists, Bahadur and Savage [3] have shown that a purely sequential sampling scheme is not sufficient to provide a universal procedure for interval estimation of a mean with specified width and confidence coefficient. Also it has been shown by Farrell [14] that for estimation of the median, within the class of distributions possessing a unique median, a purely sequential scheme is both necessary and sufficient.

Sequential tests of statistical hypotheses as discussed by Wald [32], are described as "any statistical test procedure which gives a specific rule, at any stage of the experiment (at the nth trial for each integral value of n), for making one of the following three decisions: (1) accept the hypothesis being tested (null hypothesis), (2) to reject the null hypothesis, (3) to continue the experiment by making an additional observation." Thus, after each trial in a succession of trials one of three decisions is made. Either decision

(1) or (2) is made, terminating the procedure, or (3) is made and another trial is performed. Some of the work in this field is contained in the set of references [2], [7], [11], [22], [23], [24], [26], [31], and [32].

Even though a sequential procedure may result in the smallest sample size possible for a given problem, the sequential method does present certain difficulties. It is not always possible or practical to sample from a population at an indefinite number of different times as must be done with this method. The same population may not be available to take more than a limited number of samples or it may be too costly to take samples at different times.

For certain problems it is possible to find a one-step procedure. The advantage of taking only one sample is obvious. Greenwood and Sandomire [20] have solved one such problem which is amenable to this type of sampling procedure. Their procedure gives the sample size required such that a confidence interval can be placed on the standard deviation of a normal population within 100p percent of the true value. Given p and α, their method determines the sample size n such that

$$P[\,|\hat{\sigma}_n - \sigma| < p\sigma\,] \geq 1 - \alpha$$

where

$$\hat{\sigma}_n = [\Sigma(X_i - \overline{X})^2 / (n - 1)]^{1/2}$$

and the X_i are n independent observations from a normal density with mean μ and variance σ^2. Graybill and Connell [17] give a one-step procedure to obtain sample sizes such that the ratio of variances from two independent normal populations can be estimated within $100p$ percent of the true ratio with specified confidence coefficient. Similarly Epstein [13], using a one-step procedure, has estimated the mean in the exponential distribution within 100δ percent of its true value with specified confidence coefficient. In many common densities it is possible to construct a one-step procedure to estimate a parameter within a given percent of its true value.

A two-step sampling procedure for estimation of an unknown parameter can be defined as a procedure for computing an estimate, under certain desired restrictions, based on a sample of size n, where n is determined by a first or preliminary sample. Some writers use the terms two-stage sampling, two-sampling, or double sampling to mean the same as two-step sampling.

Blum and Rosenblatt [9] give sufficient conditions for the existence of a two-step procedure for constructing confidence intervals of prescribed widths and confidence coefficients. Let G be a family of distribution functions and let $\theta(.)$ be a real-valued functional defined on G. It is desired to make an interval estimate of $\theta(F)$ based on a sample from $F \varepsilon G$. For each positive integer k let F_k denote the product distribution function on Euclidean k-space induced by F. The corresponding probability measure is denoted by P_{F_n}. Let $G(m, \gamma, \delta) = [F \varepsilon G : m(F, \gamma, \delta) \geq m]$ where $m(F, \gamma, \delta)$ is the smallest positive integer such that for all $n \geq m(F, \gamma, \delta)$ we have

$$P_{F_n}[\,|\,\theta_n(X_i,\ldots,X_n) - \theta(F)\,| \leq \delta\,] \geq 1 - \alpha.$$

Using this notation, Blum and Rosenblatt give the following theorem:

Suppose there exists a decreasing sequence $[N_j]$ of Borel subsets of R^k such that

(i) $N_o = R^k$ and $\lim_{j\to\infty} P_{F_k}[N_j] = 0$ for every $F \varepsilon G$,

and

(ii) There exists $\gamma \varepsilon (o, \alpha)$ and for each integer j

a positive integer n_j such that

$$\inf_{F \varepsilon G(n_j,\gamma,\delta)} P_{F_k}[N_j] > (1 - \alpha)/(1 - \gamma).$$

Then there exists a two-stage procedure with k observations in the first stage for constructing a confidence interval for $\theta(F)$, of length 2δ and confidence $1 - \alpha$.

A generalization of this theorem is also given for sufficient conditions which require an n-step procedure. Sufficient conditions for the existence of a two-step procedure for constructing confidence intervals of preassigned widths and confidence coefficients using only one observation in the first step are given by Abbott and Rosenblatt [1].

The first two-step procedure was given in 1945 by Stein [30] for estimating the mean μ from a normal population. In this procedure d, m, and α are specified and the sample size n is determined from a preliminary sample of size m such that

$$P[\,|\,\mu - \hat{\mu}\,| \leq d] \geq 1 - \alpha \tag{1.1}$$

where $\hat{\mu}$ is the mean of the combined sample. He also generalizes his method to confidence regions for means of several normal populations with equal but unknown variance.

Four years after the publication of Stein's article Ruben [28], working in ignorance of Stein's [30] results and using a different type of argument, rediscovered them. Ruben's results were achieved more simply and directly and have the advantage that the method generalizes quite naturally to deal with the more difficult problem of sampling from normal populations with unequal and unknown variances.

Procedures to determine the preliminary sample size in Stein's procedure have been given by Seelbinder [29] and Moshman [25]. These methods require some estimate of the range for the variance σ^2 of the population and are concerned with minimum expected sample size.

In Stein's two-step procedure to estimate the mean of a normal population the experimenter specifies d and $1 - \alpha$ in (1.1) in advance and the total number of observations is a random variable. In this case the cost of the experiment is not predetermined and may extend beyond the experimenter's resources. To exercise some control over cost Wormleighton [36] generalizes Stein's procedure so that a first sample can be taken to give an estimate of the variance after which the experimenter can decide on the total number of observations and the number of units the estimate of the mean may differ from the true value with a given confidence coefficient. Using Wormleighton's procedure the experimenter is still able to use all his data in making the estimate. Stein's results are extended by Chapman [10] to test hypotheses concerning the ratio of means of two normal populations with power independent of the unknown variances. To do this, Stein's procedure is used for each population, under certain restrictions dependent upon the hypothesized ratio, and a test involving the difference of two student's t-variables is given. Also Chapman uses Stein's technique to test the hypothesis

$$H : b = b_o$$

in the regression problem where the Y_i are independent random variables with σ_{Y_i} unknown and

$$E(Y_i) = a + bx_i .$$

In this case the power is shown to be a function of $(b'-b_o)(x_1-x_2)/\sqrt{z}$, where b' is the true value of b, x_1 and x_2 are the x values at the ends of the range which are used in the first step of Stein's procedure, and z can be chosen to obtain any prescribed power given b'. This power function is independent of σ_{Y_i} as desired.

Healy [21] also extends Stein's results and gives two-step procedures to construct simultaneous confidence intervals of prescribed widths and confidence coefficients for the following: (1) all normalized linear functions of means, (2) all differences between means, and (3) means of k independent normal populations with common unknown variances. A partial generalization of (2) has been given by Ghurye and Robbins [15]. The method estimates the difference between means from normal populations with different variances. The second sample size, restricted by a cost constraint, is determined on the basis of the size of the preliminary sample and estimate of variance. The variance of the estimate is given and is shown to be asymptotic to the minimum variance which would be obtained if the variances were known.

Bechhofer, Dunnett, and Sobel [5] give a two-step procedure to rank several normal populations according to their means when these populations have equal but unknown variance. This method is similar to Stein's. They also extend their solution to the more general case where the variances are unequal but the ratios are known. For the case of known variances Bechhofer [4] gives a single-sample multiple decision procedure.

Weiss [33] gives a two-step procedure for obtaining a confidence interval of preassigned width and confidence coefficient for quantiles of a continuous distribution. The only assumption is that the density function is unimodal.

Graybill [16] gives sufficient conditions for two-step estimation in certain parametric cases. The sample size is determined such that the probability is β^2 that the width of a confidence interval, with prescribed confidence coefficient $1 - \alpha$, will be less than some preassigned value d. Graybill's theorem is as follows:

> Let the chance variable X be the width of a confidence interval on a parameter μ based on a sample of size n. Suppose that X depends on n and on an unknown parameter θ(θ may be the parameter μ). Suppose also that there exists a function of X, θ, and n, say $g(X; \theta, n)$, such that if $Y = g(X; \theta, n)$, then the distribution of Y does not depend on any unknown parameters except n. Let $f(n)$ be a function of n such that

$$P[Y < f(n)] = \beta \qquad \text{for any} \qquad 0 < \beta < 1 .$$

Let the solution of the equation $g(x; \theta, n), = f(n)$ for x be $x = h(\theta, n)$, and suppose the following are true for $x > 0$:

(a) $g(x; \theta, n)$ is monotonic increasing in x for every n and θ.

(b) $h(\theta, n)$ is monotonic increasing for every n.

(c) $h(\theta, n)$ is monotonic decreasing in n for every θ.

(d) z is random variable which is available from step one of the procedure such that $P[t(z) > \theta] = \beta$ for $0 < \beta < 1$, where $t(z)$ is a function of z which does not depend on any unknown parameters or on n.

Let d and β be specified in advance. Then if n is such that the equation

$$h[t(z), n] \leq d$$

is satisfied $[t(z)$ is known$]$ then the following inequality is true:

$$P(X \leq d) \geq \beta^2$$

This is a two-step procedure which is applicable for many common distributions. In particular, this procedure is applied to the variance of a normal distribution by Graybill and Morrison [19].

Birnbaum and Healy [8] attacked the sample size problem by giving rules for sampling in two steps so as to obtain an unbiased estimator of a given parameter, having variance equal to, or not exceeding, a prescribed bound. They assume conditions satisfied by many distributions. It is applied to the means of the binomial, Poisson, and hypergeometric distributions, scale parameters in general and of the gamma distribution in particular, the variance of a normal, and a component of variance. This procedure can be applied to problems of interval estimation in two-steps by the use of Tchebycheff's inequality.

Graybill and Connell [18] give a two-step procedure to estimate the parameter in the uniform density,

$$f(u) = 1/\theta \quad ; \quad o \leq u \leq \theta ,$$

within d units of the true value with specified confidence coefficient. This procedure is shown to give smaller sample sizes than that possible with Birnbaum and Healy's method using Tchebycheff's inequality.

Another type of sample size problem has been solved graphically by Birnbaum and Zuckerman [6]. To determine the smallest sample size for which the minimum and the maximum of a sample are the $100\ \beta$ % distribution-free tolerance limits at the probability level α, one has to solve the equation

$$N\beta^{N-1} - (N-1)\beta^{N} = 1 - \alpha$$

given by Wilks [34]. The graph presented makes it possible to solve this equation with sufficient accuracy for almost all useful values of β and α.

The preceding is a resume of the type of work that has been done in the sample size problems. In this dissertation two estimation problems will be solved with two-step procedures. In Chapter III a two-step procedure will be given to estimate the variance of a normal distribution within d units with a specified confidence coefficient using an inequality derived in Chapter II. With minor modifications, the results of Chapter III can be extended to estimate the mean of the gamma distribution. A two-step procedure derived by a different type of argument will be given in Chapter IV to estimate the mean of a Poisson distribution within d units with a specified confidence coefficient. The results in both Chapters III and IV are compared with Birnbaum and Healy's [8] method using Tchebycheff's inequality. It is anticipated that the techniques used in this dissertation can be applied to other similar types of problems. Chapter V presents some of the sample size problems which have not yet been solved and discusses some of the problems which are associated with the solutions in Chapters III and IV.

II. A TCHEBYCHEFF TYPE INEQUALITY FOR GAMMA

2.1 Introduction

A Tchebycheff type inequality is useful in many situations in statistics, but for certain densities it may be improved upon. For example, in Chapter III it's desirable to sharpen the inequality somewhat for the gamma density. The purpose of this chapter is to find an inequality that is an improvement of Tchebycheff's inequality for a random variable that is distributed as gamma with parameters r and λ.

$$\text{Let} \quad f(x) = \frac{\lambda(x\lambda)^{r-1}}{\Gamma(r)} e^{-\lambda x}, \quad x > o .$$

$$= o \qquad , \quad x \leq o .$$

The problem is to prove that

$$P(\mid x - r/\lambda \mid \; < a\, r/\lambda) > 1 - e^{-a\sqrt{2r-1}/\sqrt{\pi}}$$

for all $a > o$, $r \geq 1/2$, and $\lambda > o$. Let $v = \lambda x/r$, $r = n/2$. Then the problem is equivalent to showing that

$$\int_{1-a}^{1+a} f_1(v)\, dv > 1-e^{-a\sqrt{n-1}/\sqrt{\pi}} \tag{2.1}$$

for all $a > o$ and $n \geq 1$, where $f_1(.)$ is the density of a chi-square divided by n, its degrees of freedom. Also, the inequality in (2.1) will be compared with

$$\int_{1-a}^{1+a} f_1(v)\, dv \geq 1 - 2/a^2 n$$

which is Tchebycheff's inequality for this problem. We shall assume n given. In the proof we shall use $y \sim z$ to mean $yz > o$.

2.2 Solution

By definition

$$f_1(v) = \frac{(n/2)^{n/2}}{\Gamma(n/2)} v^{(n/2)-1} e^{-(n/2)v}, \quad o < v < \infty$$

$$= o \qquad , -\infty < v \leq o.$$

Let

$$f_2(v) = \frac{\sqrt{n-1}}{2\sqrt{\pi}} \exp\left[- |v-1| \sqrt{n-1}/\sqrt{\pi} \right], -\infty < v < \infty .$$

Define $h(a)$ by

$$h(a) = \int_{1-a}^{1+a} [\, f_1(v)-f_2(v)\,]\, dv, \quad a \geq o, \quad n \geq 1.$$

Thus (2.1) is true if

$$h(a) > o \text{ for all } o < a < \infty \text{ and } n \geq 1. \tag{2.2}$$

From Wilton [35] we obtain

$$\Gamma(m+1) < \sqrt{2\pi}\,(m+1/2)^{m+1/2}\, e^{-(m+1/2)} \quad , \; m \geq -1/2 \; .$$

If we let $m = n/2 - 1$ we find that

$$\frac{(n/2)^{n/2}}{\Gamma(n/2)}\, e^{-n/2} > \frac{\sqrt{n-1}}{2\sqrt{\pi}} \quad , \quad n \geq 1 \; . \tag{2.3}$$

Let

$$g_1(v) = [f_1(1-v) + f_1(1+v)]\,/2f_2(1+v), \quad v \geq o \; .$$

From (2.3) we obtain

$$g_1(o) > 1 \quad , \quad n \geq 1 \; .$$

Equation (2.2) is true, which implies (2.1) is true, if there exists a v_1 such that

$$\frac{d}{dv}\, g_1(v) \begin{cases} > o, & o \leq v < v_1 \\ \leq o, & v_1 \leq v < \infty \end{cases} \tag{2.4}$$

since

$$\frac{d}{da}\, h(a) \sim g_1(a) - 1 \quad , \quad o < a < \infty$$

and

$$h(\infty) = o \; .$$

We shall show that a v_1 exists such that (2.4) is true. By definition

$$g_1(v) = \begin{cases} r[(1-v)^{p-1}\, e^{(q+p)v} + (1+v)^{p-1}\, e^{(q-p)v}\,], & o \leq v < 1 \\ r(1+v)^{p-1}\, e^{(q-p)v} & , \ 1 \leq v < \infty \end{cases}$$

where $p = n/2$, $q = \sqrt{n-1}/\sqrt{\pi}$, $r = p^p\, e^{-p}/q\, \Gamma\,(p)$.

Thus

$$\frac{d}{dv}\, g_1(v) \sim \frac{e^{-qv}}{r} \quad \frac{d}{dv}\, g_1(v)$$

$$= \begin{cases} [(q+1) - (p+q)\, v\,]\,(1-v)^{p-2}\, e^{pv} & \\ \qquad + [(q-1) - (p-q)v\,]\,(1+v)^{p-2}\, e^{-pv} & , \ o \leq v < 1 \\ [(q-1) - (p-q)\, v\,]\,(1+v)^{p-2}\, e^{-pv} & , \ 1 \leq v < \infty\,. \end{cases}$$

By definition $p > q^2$ which implies

$$(q+1)\,/\,(p+q) > (q-1)\,/\,(p-q) \quad , \ n \geq 1\,.$$

Therefore

$$\frac{d}{dv}\, g_1(v) \leq o \quad , \ v \geq (q+1)\,/\,(p+q) \tag{2.5}$$

To show the existence of such a v_1 in (2.4) we have proved (2.5) that v_1 must be less than or equal to $(q+1)\,/\,(p+q)$. To find v_1 we shall discuss three cases.

Case (i) : $4 \leq n < \infty$

Let

$$g_2(v) = [(1-v) / (1+v)]^{p-2}$$

$$g_3(v) = e^{-2pv} [(q-1) - (p-q) v] / [(q+1) - (p+q) v] .$$

Hence

$$\frac{d}{dv} g_1(v) \sim g_2(v) + g_3(v) , \quad o \leq v < (q+1) / (p+q) \tag{2.6}$$

Differentiating we obtain

$$\frac{d}{dv} g_2(v) = (p-2) [(1-v) / (1+v)]^{p-3} [-2/(1+v)^2] \leq o ,$$

and

$$\frac{d}{dv} g_3(v) = -2p\, g_3(v) - e^{-2pv} (p-q) / [(q+1) - (p+q) v]$$

$$+ e^{-2pv}(p+q)[(q-1) - (p-q)v] / [(q+1) - (p+q)v]^2$$

$$\sim -p[(q-1) - (p-q)v] [(q+1) - (p+q)v] - p + q^2$$

$$= -p(p^2 - q^2)v^2 + 2pq(p-1)v - q^2(p-1)$$

$$\leq -p(p^2 - q^2) [q(p-1) / (p^2 - q^2)]^2$$

$$+ 2pq(p-1) [q(p-1) / (p^2 - q^2)] - q^2(p-1)$$

$$\sim -p + q^2 < o .$$

Using (2.5), (2.6), and the knowledge that

$$g_2(o) + g_3(o) > o$$

and

$$\frac{d}{dv} [g_2(v) + g_3(v)] < o, \quad o \leq v < (q+1) / (p+q) ,$$

we see that there exists a $v_1 < (q+1) / (p+q)$ such that (2.4) is true, implying (2.1) is true for this case.

Case (ii) : $1 \leq n \leq 2$

In this case it can be shown that (2.4) is true for $v_1 = 1$, implying (2.1) is true.

Case (iii) : $2 < n < 4$

By similar, though tedious, manipulations it can be shown that (2.4) is true for some v_1 in the interval $(q/p, (q+1) / (p+q))$.

Thus (2.1) is true for all $n \geq 1$.

2.3 Comparison With Tchebycheff's Inequality

Let the density of v be $f_1(\cdot)$, a chi-square divided by n, its degrees of freedom. By Tchebycheff's inequality

$$P [| v-1 | < a] \geq 1 - 2/a^2 n . \tag{2.7}$$

In this chapter, by (2.1), the corresponding inequality is

$$P [| v-1 | < a] \geq 1 - e^{-a \sqrt{n-1}/ \sqrt{\pi}} . \tag{2.8}$$

Consider the ratio of $e^{-a\sqrt{n-1}}/\sqrt{\pi}$ and $2/a^2n$, i.e.

$$k(n, a) = .5a^2ne^{-a\sqrt{n-1}}/\sqrt{\pi}\ , \quad o < a < \infty\ , \quad 1 \leq n < \infty\ .$$

If $k(n, a) < 1$, then (2.8) provides a better (larger) lower bound than (2.7). We shall show the values of a and n where $k(n, a) < 1$ and hence where the method described in this chapter is better than Tchebycheff's inequality for the gamma density. Figure 2.1 shows that $k(n, a) > 1$ only in a limited region. For instance $k(n, a) < 1$ for $n \geq 7$ and $o < a < \infty$. Also, $k(n, a) < 1$ for $o < a < 1.36$ and $1 \leq n < \infty$.

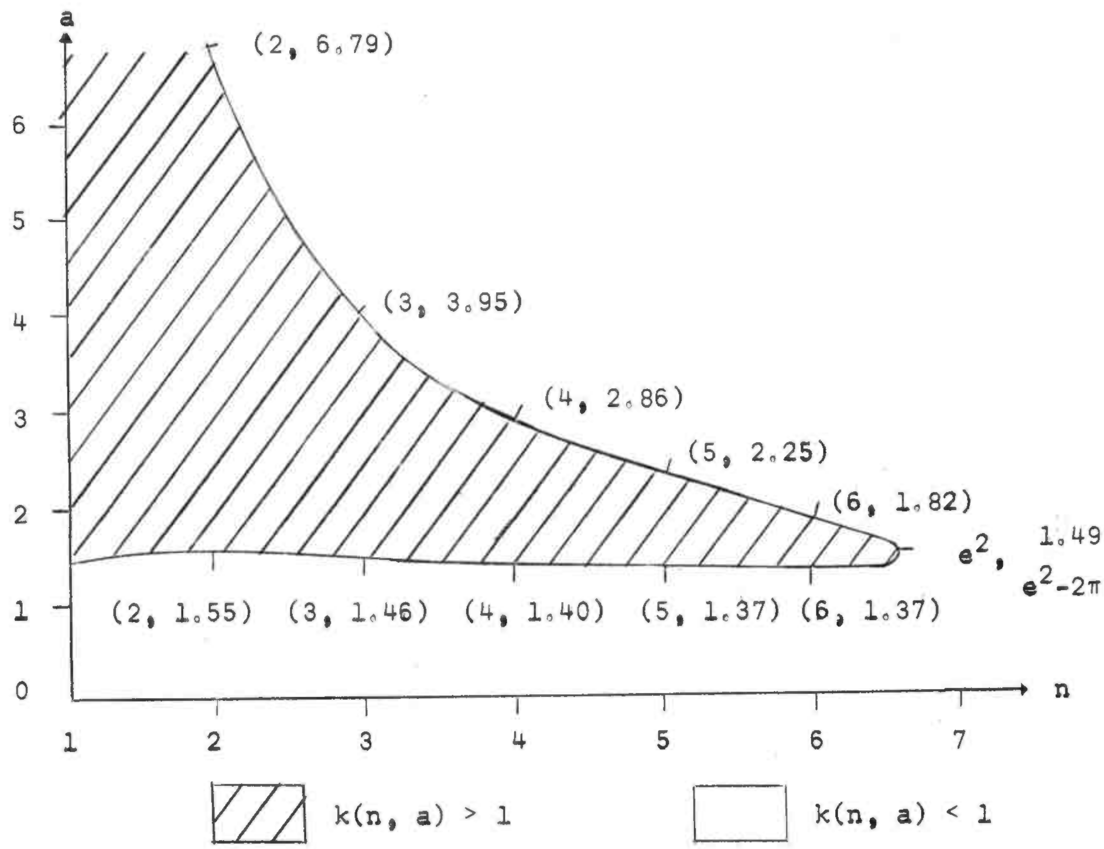


Figure 2.1

III. SAMPLE SIZE REQUIRED FOR ESTIMATING THE VARIANCE WITHIN d UNITS OF THE TRUE VALUE

3.1 Introduction

The problem of estimating the variance (σ^2) of a normal population arises in many experimental situations. J. A. Greenwood and M. M. Sandomire [20] have presented a means of obtaining the sample size required to estimate the variance of a normal population within a given per cent of its true value. An investigator may prefer to estimate the variance within a given number of units. This chapter will provide the sample size required to solve that problem.

Assume a preliminary sample of size m; $z_1, z_2, \ldots, z_m$, is taken from a normal density with variance σ^2. The unbiased estimator of the variance, s_m^2, is computed by the formula $s_m^2 = (m-1)^{-1}\Sigma(z_i-\bar{z})^2$, and d and $1-\alpha$ are specified in advance. It is desired to determine n, on the basis of the preliminary sample, such that

$$P\left[\,\left|\, s_{n+1}^2 - \sigma^2 \,\right| < d\,\right] > 1-\alpha \tag{3.1}$$

where s_{n+1}^2 is equal to $(1/n) \sum_{i=1}^{n+1} (y_i - \bar{y})^2$ and where

$y_1, y_2, \ldots, y_{n+1}$ is a random sample of size $n+1$, from a normal density with variance σ^2.

The tables in section 3.3 provide the sample size $n+1$, such that (3.1) is true, for

$$1-\alpha = .90,\ .95,\ .99$$

$$m = 5(5)20(10)50(25)150(50)300(100)500(250)1000.$$

$$\frac{s_m^2}{d} = .33,\ .5,\ .67,\ 1(1)5(5)20,\ 30.$$

The only other known method for solving this problem is given in [8] which requires the use of Tchebycheff's inequality. It can be shown that the method presented in this chapter provides a significantly smaller sample size than does [8]. For some comparisons with [8], see Table 3.4.

3.2 Solution

Equation (3.1) may be written as

$$P[\ |\ s_{n+1}^2 - \sigma^2\ |\ < d] = E_n\ \{\ P[(1-a) < v < (1+a)\ |\ n]\}$$

$$= \int_1^\infty g(n) \int_{(1-a)}^{(1+a)} f_1(v)\ dv\ dn$$

where E_n is expectation with respect to n; $a = \frac{d}{\sigma^2}$; $v = \frac{s_{n+1}^2}{\sigma^2}$; $g(.)$ is the density of n, and $f_1(.)$ is the density of a chi-square variable divided by n, its degrees of freedom. We shall restrict n such that $n \geq 1$. By definition

$$f_1(v) = \frac{\left(\frac{n}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\ v^{(n/2\ -\ 1)}\ e^{-(n/2)v}\quad ,\quad o < v < \infty$$

$$= 0 \quad , -\infty < v \leq 0 \, .$$

In Chapter II it was shown that

$$\int_{1-a}^{1+a} f_1(v)\, dv > \int_{1-a}^{1+a} f_2(v)\, dv \quad , \quad \text{for all } \; a > 0, \; n \geq 1,$$

where

$$f_2(v) = \frac{\sqrt{n-1}}{2\sqrt{\pi}} e^{-\frac{\sqrt{n-1}}{\sqrt{\pi}} | v-1 |} \qquad , -\infty < v < \infty$$

and

$$\int_{1-a}^{1+a} f_2(v)\, dv = 1 - e^{-\frac{\sqrt{n-1}}{\sqrt{\pi}} a} \quad .$$

If a were known, we might let n be equal to

$$\frac{\pi \log^2 \alpha}{a^2} + 1,$$

since in that case we would have

$$P[\; | s_{n+1}^2 - \sigma^2 | < d] > E_n \int_{1-a}^{1+a} f_2(v)\, dv$$

$$= E_n(1-\alpha)$$

$$= 1-\alpha \, .$$

Because a is assumed unknown let

$$n = \frac{\pi \log^2 \alpha}{d^2} k^2 s_m^4 + 1 \tag{3.2}$$

$$= \frac{R^2}{a^2} u^2 + 1$$

where k is some constant, independent of a, such that

$$E_n \left\{ \int_{1-a}^{1+a} f_2(v)\, dv \right\} = 1 - \alpha,$$

and where

$$R = \frac{\sqrt{\pi}\ |\log \alpha|}{m-1} k$$

$$= \frac{\sqrt{\pi}\ \log(1/\alpha)}{m-1} k$$

and

$$u = \frac{(m-1)\ s_m^2}{\sigma^2} \quad .$$

The density of u is chi-square with $m-1$ degrees of freedom; that is

$$f_3(u) = \frac{1}{2^{\frac{m-1}{2}}\ \Gamma\left(\frac{m-1}{2}\right)} u^{\frac{(m-1)}{2} - 1} e^{-(1/2)u} \quad , \quad u \geq o$$

But

$$u = \frac{a(n-1)^{1/2}}{R} \quad .$$

Thus

$$g(n) = \frac{a^{(m-1)/2}}{2^{\frac{m+1}{2}}\ R^{\frac{m-1}{2}}\ \Gamma\left(\frac{m-1}{2}\right)} (n-1)^{(m-1)/4 - 1} e^{-(a/2R)(n-1)^{1/2}} , \ n \geq 1.$$

Therefore

$$E_n\left\{\int_{1-a}^{1+a} f_2(v)\,dv\right\}$$

$$= E_n\left\{1-e^{-\frac{\sqrt{n-1}}{\sqrt{\pi}}a}\right\}$$

$$= 1 - \frac{a^{(m-1)/2}}{2^{\frac{m+1}{2}} R^{\frac{m-1}{2}} \Gamma\left(\frac{m-1}{2}\right)} \int_1^\infty (n-1)^{(m-1)/4 - 1} \cdot e^{-\left[\frac{a(n-1)^{1/2}}{2R} + \frac{a(n-1)^{1/2}}{\sqrt{\pi}}\right]} dn$$

$$= 1 - \frac{1}{2^{\frac{m-1}{2}} R^{\frac{m-1}{2}} \Gamma\left(\frac{m-1}{2}\right)} \int_0^\infty w^{(m-1)/2 - 1} e^{-(1/2R+1/\sqrt{\pi})w}\, dw$$

$$= 1 - \frac{1}{2^{\frac{m-1}{2}} R^{\frac{m-1}{2}} \left(\frac{\sqrt{\pi}+2R}{2R\sqrt{\pi}}\right)^{\frac{m-1}{2}}}$$

$$= 1 - \left[1 + \frac{2}{(m-1)}\log(1/\alpha)k\right]^{-\frac{(m-1)}{2}}$$

If we set

$$k = \frac{(m-1)}{2}\left[\frac{(1/\alpha)^{\frac{2}{(m-1)}} - 1}{\log(1/\alpha)}\right] \tag{3.3}$$

we have

$$E_n\left\{\int_{1-a}^{1+a} f_2(v)\,dv\right\} = 1 - \alpha.$$

Thus if we substitute k in Equation (3.3) into Equation (3.2) we get

$$n+1 = \frac{\pi}{4}\left[(1/\alpha)^{\frac{2}{(m-1)}} - 1\right]^2 (m-1)^2 \frac{s_m^4}{d^2} + 2 \qquad (3.4)$$

We have proved that if the sample size $n+1$, given in Equation (3.4) is used for the second step sample, the following inequality is satisfied:

$$P[\,|\,s_{n+1}^2 - \sigma^2\,| < d] > 1-\alpha.$$

The expected sample size in Equation (3.4) is

$$E_n(n+1) = \frac{\pi}{4}\left[(1/\alpha)^{\frac{2}{(m-1)}} - 1\right]^2 (m-1)(m+1)\frac{\sigma^4}{d^2} + 2 \qquad (3.5)$$

3.3 Sample Size Tables

The sample size $n+1$ as given in Equation (3.4) insures that (3.1) is true. To find the sample size, compute s_m^2/d, where s_m^2 is available from the preliminary sample of the procedure and d is the desired allowable deviation from the true variance, and use Table 3.1, 3.2, or 3.3 depending on the appropriate $1-\alpha$ level (m is the sample size on which s_m^2 is computed in the preliminary sample).

To find $n+1$ for values of s_m^2/d other than those in Tables 3.1, 3.2, and 3.3 use Table 3.4 as follows. Compute s_m^4/d^2, multiply by the entry in Table 3.4 which corresponds to the appropriate $1-\alpha$ level and m, and add 2.

Table 3.5 shows some comparisons between the sample size given in (3.4) and the sample size obtained in [8]. The quantities tabled are

$$h(m,\alpha) = \frac{n-1}{n'-1} = \frac{\pi}{8}\alpha(m-3)(m-5)[(1/\alpha)^{2/(m-1)} - 1]^2 \; ; \quad m \geq 6$$

where $n+1$ is given in (3.4) and n' is the sample size given in [8]. It is noted that

$$h(m,\alpha) = \frac{E(n-1)}{E(n'-1)} \; .$$

It can be demonstrated that

$$\begin{aligned} h(m,\alpha) &< h(m,\alpha_o) \\ &< \lim_{m\to\infty} h(m,\alpha_o) \\ &= 2\pi e^{-2} \\ &\cong .85 \end{aligned}$$

where $\alpha_o = \left(\frac{m-5}{m-1}\right)^{(m-1)/2}$. This shows that the sample size using (3.4) is never more than 85% of the sample size obtained by using the method in [8].

TABLE 3.1

Sample Size n+1 such that $P[\,|\, s^2_{n+1} - \sigma^2 \,|\, < d] > 1-\alpha$

$1-\alpha = .90$

m \ $\frac{s_m^2}{d}$	.33	.5	.67	1.	1.33	2	3
5	8.40	16.68	28.37	60.75	105.92	237.01	530.78
10	5.09	9.09	14.74	30.39	52.22	115.58	257.56
15	4.54	7.83	12.48	25.35	43.30	95.41	212.18
20	4.32	7.33	11.57	23.32	39.72	87.31	193.96
30	4.13	6.89	10.78	21.56	36.60	80.25	178.07
40	4.04	6.69	10.42	20.76	35.19	77.06	170.89
50	3.99	6.57	10.21	20.31	34.49	75.24	166.80
75	3.93	6.43	9.95	19.73	33.36	72.92	161.58
100	3.90	6.36	9.83	19.45	32.87	71.81	159.07
125	3.88	6.32	9.76	19.28	32.58	71.15	157.59
150		6.29	9.71	19.18	32.39	70.72	156.62
200		6.26	9.65	19.04	32.15	70.18	155.42
250		6.24	9.61	18.96	32.01	69.87	154.70
300		6.22	9.59	18.91	31.92	69.66	154.23
400		6.21	9.56	18.84	31.80	69.39	153.64
500		6.20	9.54	18.81	31.73	69.24	153.29
750		6.18	9.52	18.75	31.64	69.03	152.83
1000		6.18	9.51	18.73	31.59	68.93	152.59

m \ $\frac{s_m^2}{d}$	4	5	10	15	20	30
5	942	1471	5877	13222	23503	52880
10	456	712	2842	6391	11360	25559
15	376	586	2337	5257	9343	21020
20	343	535	2135	4801	8534	19198
30	315	491	1958	4404	7828	17610
40	302	471	1879	4224	7508	16891
50	295	460	1833	4122	7326	16482
75	286	445	1775	3992	7095	15960
100	281	438	1747	3929	6983	15709
125	279	434	1731	3892	6917	15562
150	277	432	1720	3868	6874	15465
200	275	428	1707	3838	6821	15344
250	273	426	1699	3820	6789	15273
300	273	425	1694	3808	6768	15226
400	272	423	1687	3793	6742	15167
500	271	422	1683	3784	6726	15132
750	270	421	1678	3773	6706	15085
1000	270	420	1675	3767	6695	15062

TABLE 3.2

Sample Size $n+1$ such that $P[\ |s_{n+1}^2 - \sigma^2| < d] > 1-\alpha$

$1-\alpha = .95$

m \ $\frac{s_m^2}{d}$	.33	.5	.67	1.	1.33	2	3
5	18.49	39.87	70.00	153.49	269.98	607.98	1365.47
10	8.19	16.22	27.55	58.91	102.68	229.67	514.26
15	6.78	12.97	21.71	45.91	79.68	177.66	397.25
20	6.24	11.74	19.49	40.96	70.92	157.87	352.70
30	5.78	10.69	17.61	36.78	63.53	141.15	315.09
40	5.58	10.23	16.78	34.94	60.26	133.76	298.46
50	5.47	9.97	16.32	33.90	58.42	129.60	289.10
75	5.33	9.64	15.73	32.58	56.10	124.35	277.29
100	5.26	9.49	15.45	31.96	55.00	121.84	271.65
125	5.22	9.39	15.28	31.59	54.35	120.38	268.35
150	5.19	9.33	15.17	31.35	53.92	119.41	266.19
200	5.16	9.26	15.04	31.05	53.40	118.23	263.52
250	5.14	9.22	14.96	30.88	53.08	117.52	261.93
300	5.13	9.19	14.91	30.76	52.88	117.06	260.89
400	5.11	9.15	14.84	30.62	52.62	116.48	259.58
500	5.10	9.13	14.80	30.53	52.47	116.13	258.81
750	5.09	9.10	14.75	30.42	52.27	115.68	257.78
1000	5.08	9.09	14.73	30.36	52.17	115.45	257.27

m \ $\frac{s_m^2}{d}$	4	5	10	15	20
5	2426	3789	15152	34089	60601
10	913	1425	5694	12809	22769
15	705	1100	4394	9883	17569
20	625	976	3899	8770	15589
30	559	872	3481	7829	13917
40	529	826	3296	7414	13178
50	512	800	3192	7180	12762
75	491	767	3061	6884	12237
100	481	751	2998	6743	11987
125	476	742	2962	6661	11840
150	472	736	2937	6607	11744
200	467	728	2908	6540	11625
250	464	724	2890	6500	11555
300	462	721	2879	6474	11508
400	460	718	2864	6442	11450
500	459	715	2855	6422	11416
750	457	713	2844	6397	11370
1000	456	711	2838	6384	11348

TABLE 3.3

Sample Size $n+1$ such that $P[|s_{n+1}^2 - \sigma^2| < d] > 1-\alpha$

$1-\alpha = .99$

m \ $\frac{s_m^2}{d}$	.33	.5	.67	1	1.33	2
5	112.84	256.46	458.92	1019.87	1802.52	4073.51
10	24.01	52.53	92.74	204.14	359.57	810.58
15	16.52	35.33	61.85	135.34	237.86	535.36
20	14.01	29.58	51.52	112.32	197.14	443.28
30	12.05	25.07	43.43	94.30	165.27	371.21
40	11.23	23.19	40.05	86.76	151.94	341.06
50	10.78	22.16	38.19	82.64	144.64	324.56
75	10.22	20.88	35.91	77.55	135.64	304.21
100	9.96	20.29	34.84	75.17	131.43	294.69
125	9.81	19.94	34.22	73.79	129.00	289.18
150	9.72	19.72	33.82	72.89	127.40	285.58
200	9.60	19.44	33.33	71.79	125.45	281.17
250	9.52	19.28	33.03	71.14	124.30	278.57
300	9.48	19.17	32.84	70.71	123.55	276.86
400	9.42	19.04	32.60	70.18	122.61	274.73
500	9.39	18.96	32.46	69.86	122.05	273.47
750	9.34	18.86	32.27	69.45	121.31	271.80
1000	9.32	18.81	32.18	69.24	120.94	270.97

m \ $\frac{s_m^2}{d}$	3	4	5	10	15
5	9162.90	16288	25449	101790	229025
10	1821.30	3236	5056	20217	45485
15	1202.07	2135	3336	13336	30004
20	994.88	1767	2760	11034	24824
30	832.73	1479	2310	9232	20770
40	764.89	1358	2121	8479	19074
50	727.76	1292	2018	8066	18146
75	681.98	1211	1891	7557	17002
100	660.57	1173	1831	7319	16466
125	648.16	1151	1797	7182	16156
150	640.07	1136	1774	7092	15954
200	630.14	1119	1747	6981	15706
250	624.29	1108	1731	6916	15559
300	620.44	1101	1720	6874	15463
400	615.65	1093	1707	6820	15343
500	612.82	1088	1699	6789	15273
750	609.06	1081	1688	6747	15179
1000	607.19	1078	1683	6726	15132

TABLE 3.4

Entries are $(\pi/4)[(1/\alpha)^{2/(m-1)}-1]^2(m-1)^2$

$1-\alpha$ \ m	5	10	15	20	50	100	200	500	1000
.90	58.75	28.40	23.35	21.33	18.31	17.45	17.05	16.81	16.73
.95	151.50	56.92	43.92	38.97	31.80	29.96	29.06	28.53	28.36
.99	1017.88	202.14	133.34	110.32	80.64	73.17	69.79	67.87	67.24

TABLE 3.5

Comparison of sample size: n+1 given in (2.3), n' given in [1]

$$h(m,\alpha) = \frac{n-1}{n'-1} = \frac{E(n-1)}{E(n'-1)}$$

m \ α	.01	.05	.10
10	.437	.615	.613
100	.344	.704	.820
1000	.334	.705	.832

IV. SAMPLE SIZE REQUIRED TO ESTIMATE THE PARAMETER IN THE POISSON DISTRIBUTION

4.1 Introduction

In this chapter some two-step procedures will be presented to estimate the Poisson Parameter within d units with a specified confidence coefficient. The Poisson density is

$$P(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!} \quad , \quad x = 0,1,2,\ldots \tag{4.1}$$

Let m, d, and $1-\varepsilon$ be specified in advance and let $x_1, x_2, \ldots, x_m$ be a preliminary sample of size m from $P(\cdot;\lambda)$. The problem is to determine n, the size of a second sample $y_1, y_2, \ldots, y_n$ from $P(\cdot;\lambda)$, based on the values of the first sample, as m, d, and $1-\varepsilon$, such that

$$\Pr[|\hat{\lambda}_n - \lambda| < d] \geq 1 - \varepsilon \tag{4.2}$$

where $\hat{\lambda}_n$ is some function of the second sample. If n were fixed the maximum likelihood estimator of λ is $\bar{y}_n$, the mean of the second sample. Since n is a random variable the maximum likelihood estimator of λ depends on the density of n. In this chapter n is not defined in explicit terms and the actual maximum likelihood estimator cannot easily be found. Theorem 4.1 shows that

$$\hat{\lambda}_n = \bar{y}_n \tag{4.3}$$

is an unbiased estimator of λ. Equation (4.3) will be used as the definition of $\hat{\lambda}_n$ throughout the remainder of this chapter. Theorem 4.1 assumes n is a proper random variable whose range consists of all the positive integers. Later in this chapter n will be a continuous random variable, but by letting the second sample size be the next largest integer for fractional values of n, a new random variable is defined replacing n. In this case, equation (4.2) will be true if it was true for n.

Although $\bar{y}_n$ is computed on the basis of a random sample of size n, the unconditional distribution of the random variable $\bar{y}_n$ is independent of n. For convenience the subscript n has been added.

Theorem 4.1: Let $\hat{\lambda}_n = \bar{y}_n$, the mean of the second sample $y_1, y_2, \ldots, y_n$ from $P(\cdot;\lambda)$. Then

$$E(\hat{\lambda}_n) = \lambda$$

Proof:

By definition

$$E(\hat{\lambda}_n) = E(\bar{y}_n) .$$

Thus

$$\begin{aligned} E(\hat{\lambda}_n) &= E_n[E(\bar{y}_n \mid n)] \\ &= E_n[\lambda] \\ &= \lambda . \end{aligned}$$

It is further noted that $\bar{y}_n$ given the value of n is the traditional estimator for λ. It is the minimum variance unbiased estimator and also is a squared error consistent estimator of λ.

Let

$$t_1(n;\lambda,d) = \Pr[\,|\hat{\lambda}_n - \lambda\,| < d \mid n]$$

$$= \sum_{v=[n\lambda-nd]+1}^{[n\lambda+nd]} \frac{e^{-n\lambda}(n\lambda)^v}{v!} \tag{4.4}$$

where $[k]$ means the integral value of k. The density of $n\bar{y}_n$ given n is $P(\cdot;n\lambda)$, i.e., $n\bar{y}_n$ given n is Poisson with parameter $n\lambda$. Thus

$$\Pr[\,|\hat{\lambda}_n - \lambda\,| < d] = E_n\{\Pr[\,|\hat{\lambda}_n - \lambda\,| < d \mid n]\}$$

$$= E_n\{\Pr[\,|n\bar{y} - n\lambda\,| < nd \mid n]\}$$

$$= E_n\{t_1(n;\lambda,d)\}. \tag{4.5}$$

In Section 2 we shall prove certain monotonic properties of t_1 and a generalization of t_1 which will lead to a determination of n, the size of the second sample.

Sections 3, 4, 5 and 6 present various methods for determining n. The procedure in Section 3 is the easiest to use and demonstrates the basic technique employed in this chapter to determine n. Section 4 gives a second solution which leads to a smaller sample size n but the confidence coefficient is not predetermined. This difficulty is removed by a similar solution in Section 5. This solution allows for a preassigned confidence coefficient and reduces sample size compared with the solution in Section 3 but is more difficult to use. The basic solution is further generalized in Section 6 but the confidence coefficient is not preassigned.

In Section 8 some comparisons of the results of Sections 3 and 5 are made with the solution to this problem which can be obtained by Birnbaum and Healy's [8] method using Tchebycheff's inequality. These comparisons show a significant reduction in sample size.

To apply the results of this chapter some examples are given in Section 9 along with sample size graphs for procedures developed in Sections 3 and 5. Also Section 7 shows how to use this chapter's methods to estimate the mean in a Poisson stochastic process.

In the remainder of this chapter the following definitions for z, $\bar{x}_m$, and H will be used:

$$z = m\bar{x}_m ,$$

where $\bar{x}_m$ is the mean of the preliminary sample, and

$$H(c;z) = \sum_{v=o}^{z} \frac{e^{-cz}(cz)^v}{v!} \tag{4.6}$$

Also,

$$a \sim b$$

will mean that a and b have the same sign.

4.2 Monotonic Properties

Equation (4.4) can be rewritten as

$$t_1(n;\lambda,d) = \sum_{v=[n\lambda-nd]+1}^{[n\lambda+nd]} P(v;n\lambda) \tag{4.7}$$

where $P(v;n\lambda)$ is defined by (4.1). Consider

$$h_1(v) = P(v+1;n\lambda)/P(v;n\lambda)$$

$$= n\lambda/(v+1)$$

$$\begin{cases} > 1, & v < n\lambda - 1 \\ < 1, & v > n\lambda - 1. \end{cases}$$

Thus, for integral values of v, the function $P(v;n\lambda)$ is monotonic increasing in v for $v>[n\lambda]$ and is monotonic decreasing for $v>[n\lambda]$. Thus the mode of $P(.;n\lambda)$ is $[n\lambda]$. If $n\lambda$ is an integer, then $P(.;n\lambda)$ is bimodal with modes at $n\lambda-1$ and $n\lambda$ since

$$P(n\lambda-1;n\lambda) = \frac{e^{-n\lambda}(n\lambda)^{n\lambda-1}}{(n\lambda-1)!}$$

$$= \frac{e^{-n\lambda}(n\lambda)^{n\lambda}}{(n\lambda)!}$$

$$= P(n\lambda;n\lambda) .$$

Let

$$t(n;\lambda,d) = \int_{n\lambda-nd+.5}^{n\lambda+nd-.5} f(v;n\lambda)dv \qquad (4.8)$$

where

$$f(v;\lambda) = (1-v+[v])P([v];\lambda)+(v-[v])P([v]+1;\lambda)$$

and it is assumed that nd is greater than or equal to 2. Otherwise let t be zero. The following two theorems show the relationship of

$t(n;\lambda,d)$ to $t_1(n;\lambda,d)$ and some monotonic properties of t essential to the determination of n.

Theorem 4.2: Let t_1 and t be defined by (4.7) and (4.8) respectively. Then

$$E_n\{t(n;\lambda,d)\} \geq 1 - \varepsilon$$

implies equation (4.2) is true, i.e.,

$$\Pr[|\hat{\lambda}_n - \lambda| < d] \geq 1 - \varepsilon.$$

Proof: From (4.8) we observe that $f(v;n\lambda)$ consists of straight lines joining the values of $P(v;n\lambda)$ at adjacent integral values of v. Therefore

$$\int_k^{k+1} f(v;n\lambda)dv = .5\{P(k;n\lambda)+P(k+1;n\lambda)\}$$

for integral values of k. Also,

$$\int_{n\lambda-nd+.5}^{[n\lambda-nd+1.5]} f(v;n\lambda)dv = .5\{[n\lambda-nd+1.5]-(n\lambda-nd+.5)\}\{f(n\lambda-nd+.5;n\lambda) + P([n\lambda-nd+1.5];n\lambda)\}.$$

Thus, if $[n\lambda-nd+1.5] = [n\lambda-nd+1]$, then

$$\begin{aligned}\int_{n\lambda-nd+.5}^{[n\lambda-nd+1.5]} f(v;n\lambda)dv &< .5(.5)\{f(n\lambda-nd+.5;n\lambda)+P([n\lambda-nd+1];n\lambda)\} \\ &\leq .25\{f([n\lambda-nd+1];n\lambda)+P([n\lambda-nd+1];n\lambda)\} \\ &= .5\, P([n\lambda-nd+1];n\lambda),\end{aligned}$$

since $f(v;n\lambda)$ is monotonic increasing for v less than $[n\lambda]$. Otherwise $[n\lambda-nd+1.5] = [n\lambda-nd+2]$, which implies

$$\int_{n\lambda-nd+.5}^{[n\lambda-nd+1.5]} f(v;n\lambda)dv \leq \int_{[n\lambda-nd+1]}^{[n\lambda-nd+2]} f(v;n\lambda)dv$$

$$< .5\{P([n\lambda-nd+2];n\lambda)+P([n\lambda-nd+1];n\lambda)\}$$

Similarly, if $[n\lambda+nd-.5] = [n\lambda+nd]$, then

$$\int_{[n\lambda+nd-.5]}^{n\lambda+nd-.5} f(v;n\lambda)dv = .5\{(n\lambda+nd-.5)-[n\lambda+nd-.5]\}\{P([n\lambda+nd-.5];n\lambda)$$

$$+f(n\lambda+nd-.5;n\lambda)\}$$

$$< .5(.5)\{P([n\lambda+nd];n\lambda)+f(n\lambda+nd-.5;n\lambda)\}$$

$$\leq .25\{P([n\lambda+nd];n\lambda)+f([n\lambda+nd];n\lambda)\}$$

$$= .5\ P([n\lambda+nd];n\lambda),$$

since $f(v;n\lambda)$ is monotonic decreasing for v greater than $[n\lambda]$. In the case where $[n\lambda+nd-.5] = [n\lambda+nd-1]$, then

$$\int_{[n\lambda+nd-.5]}^{n\lambda+nd-.5} f(v;n\lambda)dv \leq \int_{[n\lambda+nd-1]}^{[n\lambda-nd]} f(v;n\lambda)dv$$

$$< .5\{P([n\lambda+nd-1];n\lambda)+P([n\lambda+nd];n\lambda)\}.$$

Therefore

$$t(n;\lambda,d) = \int_{n\lambda-nd+.5}^{n\lambda+nd-.5} f(v;n\lambda)dv$$

$$< \sum_{v=[n\lambda-nd]+1}^{[n\lambda+nd]} P(v;n\lambda)$$

$$= t_1(n;\lambda,d).$$

Hence

$$\Pr[|\hat{\lambda}_n - \lambda| < d] = E_n\{t_1(n;\lambda,d)\}$$

$$> E_n\{t(n;\lambda,d)\}.$$

If nd is less than 2, then the inequality is still true which completes the proof.

Theorem 4.3: Let $t(n;\lambda,d)$ be defined by (4.8). Then assuming $nd \geq 2$, we have

(a) $\frac{\partial t}{\partial d} > 0$

(b) $\frac{\partial t}{\partial \lambda} < 0$

(c) $\frac{\partial t}{\partial n} > 0$.

Proof:

(a) Differentiating t with respect to d we obtain

$$\frac{\partial t}{\partial d} = nf(n\lambda+nd-.5;n\lambda)+nf(n\lambda-nd+.5;n\lambda)$$

$$> 0 .$$

(b) Also

$$\frac{\partial t}{\partial \lambda} = nf(n\lambda+nd-.5;n\lambda)-nf(n\lambda-nd+.5;n\lambda)+\int_{n\lambda-nd+.5}^{n\lambda+nd-.5} \frac{\partial t}{\partial \lambda}(v;n\lambda)dv,$$

where

$$\frac{\partial f}{\partial \lambda}(v;n\lambda) = (1-v+[v])\frac{\partial P}{\partial \lambda}([v];n\lambda)+(v-[v])\frac{\partial}{\partial \lambda}P([v]+1;n\lambda).$$

But

$$\frac{\partial P}{\partial \lambda}(v;n\lambda) = (\frac{v}{\lambda} -n)P(v;n\lambda)$$

$$= nP(v-1;n\lambda)-nP(v;n\lambda).$$

Thus

$$\begin{aligned}\frac{\partial f}{\partial \lambda}(v;n\lambda)&=n(1-v+[v])P([v]-1;n\lambda)+n\{-(1-v+[v])+(v-[v])\}P([v];n\lambda)\\ &\qquad -n(v-[v])P([v]+1;n\lambda)\\ &=n\{(1-v+[v])P([v]-1;n\lambda)+(v-[v])P([v];n\lambda)\}-n\{(1-v+[v])P([v];n\lambda)\\ &\qquad +(v-[v])P([v]+1;n\lambda)\}\\ &=nf(v-1;n\lambda)-nf(v;n\lambda).\end{aligned}$$

Therefore

$$\frac{\partial t}{\partial \lambda} = nf(n\lambda+nd-.5;n\lambda)-nf(n\lambda-nd+.5;n\lambda)+n\int_{n\lambda-nd+.5}^{n\lambda+nd-.5}\{f(v-1;n\lambda)-f(v;n\lambda)\}dv$$

$$= -nf(n\lambda-nd+.5;n\lambda)+n\int_{n\lambda-nd-.5}^{n\lambda-nd+.5} f(v;n\lambda)dv - n\int_{n\lambda+nd-1.5}^{n\lambda+nd-.5} f(v;n\lambda)dv+nf(n\lambda+nd-.5;n\lambda)$$

$$< 0 ,$$

since $f(v;n\lambda)$ is monotonic increasing for v in the interval

$(n\lambda-nd-.5,\ n\lambda-nd+.5)$ and is monotonic decreasing in the interval $(n\lambda+nd-1.5,\ n\lambda+nd-.5)$.

(c) Finally, differentiating with respect to n we have

$$\frac{\partial t}{\partial n} = (\lambda+d)f(n\lambda+nd-.5;n\lambda)-(\lambda-d)f(n\lambda-nd+.5;n\lambda)+\int_{n\lambda-nd+.5}^{n\lambda+nd-.5}\frac{\partial}{\partial n}f(v;n\lambda)dv,$$

where

$$\frac{\partial f}{\partial n}(v;n\lambda) = \lambda f(v-1;n\lambda)-\lambda f(v;n\lambda)$$

since n appears everywhere λ does in $f(v;n\lambda)$ and from part (b)

$$\frac{\partial f}{\partial \lambda}(v;n\lambda) = nf(v-1;n\lambda)-nf(v;n\lambda).$$

Thus

$$\frac{\partial t}{\partial n} = (\lambda+d)f(n\lambda+nd-.5;n\lambda)-(\lambda-d)f(n\lambda-nd+.5;n\lambda)+\lambda\int_{n\lambda-nd-.5}^{n\lambda-nd+.5}f(v;n\lambda)dv-\lambda\int_{n\lambda+nd-1.5}^{n\lambda+nd-.5}f(v;n\lambda)dv.$$

Let

$L = n\lambda$,

$D = nd$,

and

$$S_{\pm} = \pm\ (L\pm D)f(L\pm D\mp.5;L)\mp\ L\int_{L\pm D-1\mp.5}^{L\pm D\mp.5}f(v;L)dv\ .$$

Therefore

$$\frac{\partial t}{\partial n} \sim n\frac{\partial t}{\partial n}$$

$$= S_{+} + S_{-}\ .$$

Let $s_{\pm} = L \pm D \mp .5 - [L \pm D \mp .5]$.

For convenience the subscript on s will be deleted below. Hence

$$S_{\pm} = \pm(L\pm D)\{(1-s)P([L\pm D\mp.5];L)+sP([L\pm D\mp.5]+1;L)\}$$

$$\mp L\{.5(1-s)f(L\pm D-1\mp.5;L)+.5(1-s)P([L\pm D\mp.5];L)$$

$$+.5sP([L\pm D\mp.5];L)+.5sf(L\pm D\mp.5;L)\}$$

$$= \mp.5L(1-s)^2P([L\pm D\mp.5]-1;L)$$

$$\pm\{(L\pm D)(1-s)-.5L(1-s)s-.5L-.5Ls(1-s)\}P([L\pm D\mp.5];L)$$

$$\pm\{(L\pm D)s-.5Ls^2\}P([L\pm D\mp.5]+1;L).$$

But

$$P([L\pm D\mp.5]-1;L) = \frac{[L\pm D\mp.5]}{L}P([L\pm D\mp.5];L)$$

$$= \frac{(L\pm D\mp.5-s)}{L}P([L\pm D\mp.5];L).$$

Hence

$$S_{\pm} \sim 2S_{\pm}$$

$$= \pm\{-(L\pm D\mp.5-s)(1-s)^2+2(L\pm D)(1-s)-2Ls(1-s)-L\}P([L\pm D\mp.5];L)$$

$$\pm\{2(L\pm D)s-Ls^2\}P([L\pm D\mp.5]+1;L)$$

$$\sim (2S_{\pm})L/P([L\pm D\mp.5]+1;L)$$

$$= \pm\{(-2s+s^2)L\pm(1-s^2)D+(s\pm.5)(1-s)^2\}(L\pm D\mp.5+1-s)$$

$$\pm\{(2s-s^2)L\pm 2sD\}L$$

$$= LD+(.5\mp s\pm s^2)L\pm(1-s^2)D^2$$

$$+\{1\mp s-(3\mp 1)s^2+2s^3\}D\pm(s\pm .5)(1-s)^2(1-s\mp .5)$$

$$= LD+\{.5\mp s(1-s)\}L\pm(1-s)(1+s)D^2+s(1-s)D$$

$$+(1-s)(1\mp s-2s^2)D\pm(1-s)^2(s\pm .5)\{1-(s\pm .5)\}$$

$$> LD+.25L-D^2-.25$$

$$= (L-D)D+.25(L-1)$$

$$> 0$$

Therefore

$$\frac{\partial t}{\partial n} \sim S_+ + S_-$$

$$> 0$$

which completes the proof.

4.3 One Point Solution

Let m, d, α and β be specified in advance and observe a preliminary sample $x_1, x_2, \ldots x_m$. Define n_1 such that

$$t(n_1;\lambda,d) = 1 - \alpha \tag{4.9}$$

(note Figure 4.1).

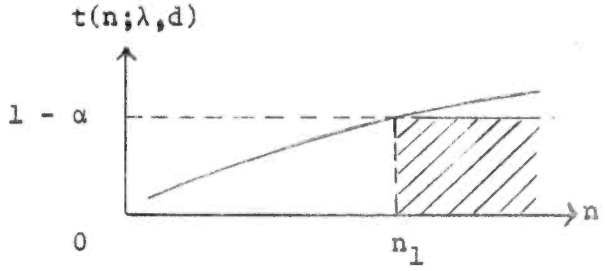


Figure 4.1

Determine the value n of the random variable, the size of the second sample, such that

$$t(n;c\bar{x}_m;d) = 1 - \alpha \tag{4.10}$$

where t is defined by (4.8), H by (4.6), and C is defined by

$$H(c;z) = \beta \ . \tag{4.11}$$

From (4.9) and (4.10) we have

$$t(n_1;\lambda,d) = t(n;c\bar{x}_m,d). \tag{4.12}$$

Thus, using Theorem 4.3 (c) and (b), we obtain

$$\begin{aligned} Pr(n > n_1) &= Pr[t(n;\lambda,d) > t(n_1;\lambda,d)] \\ &= Pr[t(n;\lambda,d) > t(n;c\bar{x}_m,d)] \\ &= Pr(\lambda < c\bar{x}_m) \ . \end{aligned} \tag{4.13}$$

<u>Theorem 4.4</u>: Let $H(c;z)$ be defined as in (4.6). Then

$$Pr(\lambda < c\bar{x}_m) = 1 - H(c;z).$$

Proof: The random variable $m\bar{x}_m$ is distributed as Poisson with mean m. Therefore

$$\Pr(\bar{x}_m > \lambda_{1-\beta}) = 1 - \beta ,$$

where $\lambda_{1-\beta}$ is defined such that

$$\sum_{v=0}^{m\lambda_{1-\beta}} \frac{e^{-m\lambda}(m\lambda)^v}{v!} = \beta .$$

This implies that

$$\Pr(\lambda < \lambda') = 1 - \beta$$

where λ' is defined by

$$\sum_{v=0}^{m\bar{x}_m} \frac{e^{-m\lambda'}(m\lambda')^v}{v!} = \beta .$$

Hence

$$\Pr(\lambda < c\bar{x}_m) = 1 - \sum_{v=0}^{m\bar{x}_m} \frac{e^{-mc\bar{x}_m}(mc\bar{x}_m)^v}{v!}$$

$$= 1 - H(c;z),$$

since λ' and β are in a one to one correspondence with each other, thus completing the proof.

Therefore, from (4.11), (4.13) and Theorem 4.4, we obtain

$$\Pr(n > n_1) = 1 - H(c;z)$$

$$= 1 - \beta \qquad (4.14)$$

From Theorem 4.3 (c) we obtain

$$t(n;\lambda,d) \geq \begin{cases} 0 & , \quad 0 < n < n_1 \\ t(n_1;\lambda,d), & n \geq n_1 \end{cases}$$

which is depicted by the shaded area in Figure 4.1. Thus, by choosing n as defined in (4.10), we obtain

$$E_n[t(n;\lambda,d)] > 0 \cdot Pr(0 < n < n_1) + t(n_1;\lambda,d)Pr(n \geq n_1)$$

$$= (1 - \alpha)(1 - \beta),$$

by (4.9) and (4.14). From Theorem 4.2 we conclude that

$$Pr[\,|\hat{\lambda}_n - \lambda\,| < d] > 1 - \varepsilon \,,$$

where

$$1 - \varepsilon = (1 - \alpha)(1 - \beta).$$

Graphs are given in Section 9 to find n for various values of m, d, $\bar{x}_m$ and $1 - \varepsilon$.

4.4 Two Point Solution:

Let m, d, α, β, and δ be preassigned and let $x_1, x_2, \ldots, x_m$ be a preliminary sample. Define n_1, n_2 such that

$$\begin{aligned} &\text{(a)} \quad t(n_1;\lambda,d) = \gamma(1 - \alpha) \\ &\text{(b)} \quad t(n_2;\lambda,d) = 1 - \alpha \end{aligned} \qquad (4.16)$$

where γ is calculated by (4.19) (note Figure 4.2).

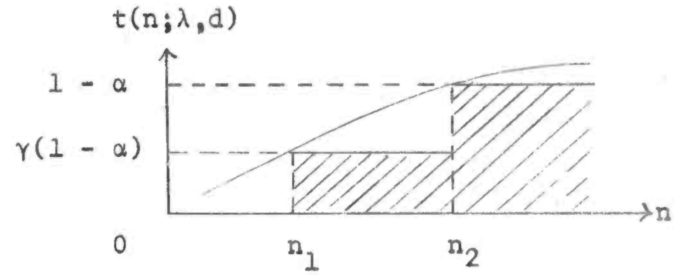


Figure 4.2

Define $c_1; c_2$ such that

$$\begin{aligned} &\text{(a)} \quad H(c_1;z) = (1 - \delta)\beta \\ &\text{(b)} \quad H(c_2;z) = \beta. \end{aligned} \tag{4.17}$$

Determine the value n of the random variable, the size of the second sample, such that

$$t(n;c_2\bar{x}_m,d) = 1 - \alpha \tag{4.18}$$

and calculate γ using

$$t(n;c_1\bar{x}_m,d) = \gamma(1 - \alpha). \tag{4.19}$$

Proceeding as in Section 3 and using (4.16), (4.18) and (4.19) we have

$$t(n_i;\lambda,d) = t(n;c_i\bar{x}_m,d), \quad i=1,2.$$

Thus, by Theorem 4.3(c) and (b), we obtain

$$\begin{aligned} \Pr(n > n_i) &= \Pr[t(n;\lambda,d) > t(n_i;\lambda,d)] \\ &= \Pr[t(n;\lambda,d) > t(n;c_i\bar{x}_m,d)] \\ &= \Pr(\lambda < c_i\bar{x}_m), \quad i=1,2. \end{aligned}$$

Hence, by (4.17) and Theorem 4.4, we have

$$\Pr(n > n_1) = 1 - (1 - \delta)\beta \tag{4.20}$$

and

$$\Pr(n > n_2) = 1 - \beta$$

which implies that

$$\begin{aligned} \Pr(n_1 < n < n_2) &= 1 - (1 - \delta)\beta - (1 - \beta) \\ &= \delta\beta \end{aligned} \tag{4.21}$$

From Theorem 4.3 (c) we obtain

$$t(n;\lambda,d) \geq \begin{cases} 0 & 0 < n \leq n_1 \\ t(n_1;\lambda,d), & n_1 < n \leq n_2 \\ t(n_2;\lambda,d), & n > n_2 \end{cases}$$

This is represented by the shaded portion in Figure 4.2. Thus by (4.16), (4.20) and (4.21) we conclude that

$$\begin{aligned} E_n[t(n;\lambda,d)] &\geq 0 \cdot \Pr(0 < n \leq n_1) + t(n_1;\lambda,d)\Pr(n_1 < n \leq n_2) \\ &\qquad + t(n_2;\lambda,d)\Pr(n > n_2) \\ &= \gamma(1 - \alpha)\delta\beta + (1 - \alpha)(1 - \beta) \\ &= (1 - \alpha)(1 - \beta + \gamma\delta\beta) . \end{aligned}$$

By Theorem 4.2 this implies

$$\Pr[|\hat{\lambda}_n - \lambda| < d] > 1 - \varepsilon$$

where

$$1 - \varepsilon = (1 - \alpha)(1 - \beta + \gamma\delta\beta) .$$

The value of n is determined by (4.18) and it is observed that this would result in the same value as determined in the one point solution assuming α and β are the same. The value of $1 - \varepsilon$, the confidence coefficient, is increased however. The amount of increase depends upon γ which is calculated after the value of n is determined. Thus the confidence coefficient is not preassigned as would normally be desired. Section 5 shows one way to avoid this difficulty.

4.5 Two Point Solution with Preassigned Confidence Coefficient

Specify m, d, and α, β, γ, δ such that $(1 - \alpha)(1 - \beta + \gamma\delta\beta)$ equals some prescribed value, say $1 - \varepsilon$. Let $x_1, x_2, \ldots, x_m$ be a preliminary sample. Define n_1, n_2 as in (4.16) and c_1, c_2 as in (4.17). Determine the values n', n'' of two random variables such that

$$\begin{aligned} &\text{(a)} \quad t(n';c_1\bar{x}_m,d) = \gamma(1 - \alpha) \\ &\text{(b)} \quad t(n'';c_2\bar{x}_m,d) = (1 - \alpha). \end{aligned} \tag{4.22}$$

From (4.16) and (4.22) we have

$$\begin{aligned} &\text{(a)} \quad t(n_1;\lambda,d) = t(n';c_1\bar{x}_m,d) \\ &\text{(b)} \quad t(n_2;\lambda,d) = t(n'';c_2\bar{x}_m,d) \end{aligned} \tag{4.23}$$

Thus, by Theorem 4.3(c) and (b) we obtain

$$\begin{aligned} \text{(a)} \quad Pr(n' > n_1) &= Pr[t(n';\lambda,d) > t(n_1;\lambda,d)] \\ &= Pr[t(n';\lambda,d) > t(n';c_1\bar{x}_m,d)] \end{aligned}$$

$$= \Pr(\lambda < c_1\bar{x}_m)$$

(b) $\Pr(n'' > n_2) = \Pr[t(n'';\lambda,d) > t(n_2;\lambda,d)]$

$$= \Pr[t(n'';\lambda,d) > t(n'';c_2\bar{x}_m,d)]$$

$$= \Pr(\lambda < c_2\bar{x}_m) .$$

Hence, by (4.17) and Theorem 4.4, we have

$$\begin{aligned} &(a) \quad \Pr(n' > n_1) = 1 - (1 - \delta)\beta \\ &(b) \quad \Pr(n'' > n_2) = 1 - \beta. \end{aligned} \tag{4.24}$$

Now choose the value n of the random variable, the size of the second sample, such that

$$n = \max(n', n''). \tag{4.25}$$

Therefore, by (4.24), we have

$$\begin{aligned} (a) \quad \Pr(n > n_1) &\geq \Pr(n' > n_1) \\ &= 1 - (1 - \delta)\beta \\ (b) \quad \Pr(n > n_2) &\geq \Pr(n'' > n_2) \\ &= 1 - \beta. \end{aligned} \tag{4.26}$$

Define n_1' and n_2' such that

$$\begin{aligned} &(a) \quad \Pr(n > n_1') = 1 - (1 - \delta)\beta \\ &(b) \quad \Pr(n > n_2') = 1 - \beta, \end{aligned} \tag{4.27}$$

which implies that

$$\Pr(n_1' < n \leq n_2') = \delta\beta \tag{4.28}$$

comparing (4.26) and (4.27) it is seen that

(a) $n_1' \geq n_1$

(b) $n_2' \geq n_2$,

which, by (4.16) and Theorem 4.3(c), implies

$$\begin{aligned}
&\text{(a)} \quad t(n_1';\lambda,d) \geq t(n_1;\lambda,d) \\
&\qquad\qquad = \gamma(1 - \alpha) \\
&\text{(b)} \quad t(n_2';\lambda,d) \geq t(n_2;\lambda,d) \\
&\qquad\qquad = 1 - \alpha.
\end{aligned} \tag{4.29}$$

From Theorem 4.3(c) we have

$$t(n;\lambda,d) \geq \begin{cases} 0 \quad , & 0 < n \leq n_1' \\ t(n_1';\lambda,d), & n_1' < n \leq n_2' \\ t(n_2';\lambda,d), & n > n_2' \ . \end{cases}$$

Thus, by (4.27b), (4.28), and (4.29), we obtain

$$\begin{aligned}
E_n[t(n;\lambda,d)] &\geq 0 \cdot \Pr(0 < n \leq n_1') + t(n_1';\lambda,d)\Pr(n_1' < n \leq n_2') \\
&\qquad + t(n_2';\lambda,d)\Pr(n > n_2') \\
&\geq \gamma(1 - \alpha)\delta\beta + (1 - \alpha)(1 - \beta)
\end{aligned}$$

$$= (1 - \alpha)(1 - \beta + \gamma\delta\beta)$$

$$= 1 - \varepsilon.$$

Thus, by Theorem 4.2, if n is determined by (4.25) we conclude that

$$\Pr[|\hat{\lambda}_n - \lambda| < d] > 1 - \varepsilon,$$

a predetermined confidence coefficient.

This solution is more difficult to work with but yields smaller second sample sizes. In Section 9 graphs are given for various values of m, d, $\overline{x}_m$ and $1 - \varepsilon$. Thus, in using the graphs an experimenter does not have to specify α, β, γ and δ.

4.6 The k Point Solution

Let m, d, k, β_i $(i=1,2,\ldots,k)$, and γ_1 be specified in advance and define $c_i (i=1,2,\ldots,k)$ such that

$$H(c_i;z) = \beta_i, \quad i=1,2,\ldots,k. \tag{4.30}$$

Determine the value n of the random variable, the size of the second sample, by (a), and $\gamma_i (i=2,3,\ldots,k)$ such that

$$\begin{aligned} &\text{(a)} \quad t(n;c_1\overline{x}_m,d) = \gamma_1 \\ &\text{(b)} \quad t(n;c_i\overline{x}_m,d) = \gamma_i, \quad i=2,3,\ldots,k. \end{aligned} \tag{4.31}$$

Further define $n_i (i=1,2,\ldots,k)$ such that

$$t(n_i;\lambda,d) = \gamma_i, \quad i=1,2,\ldots,k \tag{4.32}$$

(note Figure 4.3).

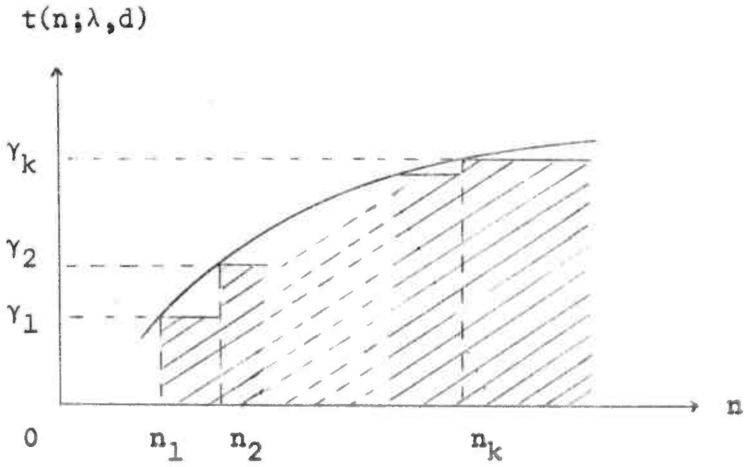


Figure 4.3

From (4.31) and (4.32) it is seen that

$$t(n_i;\lambda,d) = t(n;c_i\overline{x}_m,d), \quad i=1,2,\ldots,k$$

Thus, by Theorem 4.3(c) and (b), we obtain

$$\Pr(n > n_i) = \Pr[t(n;\lambda,d) > t(n_i;\lambda,d)]$$

$$= \Pr[t(n;\lambda,d) > t(n;c_i\overline{x}_m,d)]$$

$$= \Pr(\lambda < c_i\overline{x}_m), \quad i=1,2,\ldots,k$$

Hence, by (4.30) and Theorem 4.4; we have

$$\Pr(n > n_i) = 1 - \beta_i, \quad i=1,2,\ldots,k$$

which implies that

$$\Pr(n_i < n \leq n_{i+1}) = \beta_{i+1} - \beta_i, \quad i=1,2,\ldots,k \qquad (4.33)$$

where we define

$$n_{k+1} = \infty$$

and

$$\beta_{k+1} = 1 .$$

From Theorem 4.3(c) it is seen that

$$t(n;\lambda,d) \geq \begin{cases} 0 , & 0 < n \leq n_2 \\ t(n_1;\lambda,d) , & n_1 < n \leq n_2 \\ t(n_2;\lambda,d) , & n_2 < n \leq n_3 \\ \quad - - - & \\ t(n_k;\lambda,d) , & n > n_k \end{cases}$$

(note the shaded portion in Figure 4.3).

Thus, by (4.32) and (4.33) we have

$$E_n[t(n;\lambda,d)] \geq \sum_{i=1}^{k} t(n_i;\lambda,d)Pr(n_i < n \leq n_{i+1})$$

$$= \sum_{i=1}^{k} \gamma_i(\beta_{i+1} - \beta_i) .$$

By Theorem 4.2 this implies that

$$Pr[|\hat{\lambda}_n - \lambda| < d] > 1 - \varepsilon$$

where

$$1 - \varepsilon = \sum_{i=1}^{k} \gamma_i(\beta_{i+1} - \beta_i) .$$

Since the $\gamma_i (i=2,3,\ldots,k)$ are not determined in advance, the confidence coefficient is not predetermined in this extension of the two point solution in Section 4.

4.7 The Poisson Stochastic Process

In a Poisson stochastic process we have a counting process $\{N(t),\ t \geq 0\}$ such that $\{N(t),\ t \geq 0\}$ has stationary independent increments and

$$\Pr[N(t) - N(s) = k] = \frac{e^{-\lambda(t-s)}[\lambda(t-s)]^k}{k!},\quad k=0,1,2,\ldots.$$

To apply the method of this chapter to estimate λ, pick some constant $\tau > 0$. Then let the random variables X_i, Y_j, $i=1,2,\ldots$, $j=1,2,\ldots$ be defined such that

$$X_i = N(t_i+\tau)-N(t_i),\quad Y_j = N(s_j+\tau)-N(s_j)$$

where the intervals $(t_i,\ t_i+\tau)$ and $(s_j,\ s_j+\tau)$ for $i=1,2,\ldots$, $j=1,2,\ldots$ are disjoint. Therefore the X_i and Y_i are independent and identically distributed as Poisson with parameter $\tau\lambda$. Thus, to estimate λ within d_1 units use the method presented here with $d=\tau d_1$, with x_i equal to the value of the random variable X_i, and y_i equal to the value of the random variable Y_i. Therefore we obtain

$$\Pr[|\tau\lambda - \overline{y}_n| < \tau d_1] > 1 - \varepsilon$$

which implies that

$$\Pr[|\lambda - \overline{y}_n/\tau| < d_1] > 1 - \varepsilon$$

the desired result.

4.8 Birnbaum and Healy's Solution

Birnbaum and Healy [8] give a two-step procedure to estimate λ with the unbiased estimator $\bar{y}_n$ having variance not exceeding a prescribed bound. By using Tchebycheff's inequality their method gives

$$P[|\hat{\lambda}_n - \lambda| < d] > 1 - \varepsilon$$

if

$$n = (m\bar{x}_m + 1)/m\varepsilon d^2, \tag{4.34}$$

where $\bar{x}_m$ is the mean in a preliminary sample of size m.

Table 4.1 gives some comparisons of the second sample size when determined according to Birnbaum and Healy's solution, the one point solution from Section 3, and the two point solution from Section 5 for various values of $1 - \varepsilon$, m, d, and $\bar{x}_m$. In these examples sample sizes were reduced by varying amounts from 50 to 90%.

TABLE 4.1

Comparisons of Second Sample Sizes

Column I gives n for Birnbaum and Healy's solution (equation 4.34).

Column II gives n for the one point solution (equation 4.10).

Column III gives n for the two point solution (equation 4.25).

$1-\varepsilon$	m	$\bar{x}_m$	d	I	II	n'	n"	III
.90	5	100	2	205.5	78	72	73	73
.95	"	"	"	501	112	98	104	104
.99	"	"	"	2505	196	180	178	180
.95	1	100	2	505	130	121	112	121
"	2	"	"	502.5	120	109	108	109
"	5	"	"	501	112	98	104	104
"	10	"	"	500.5	108	94	102	102
"	25	"	"	500.2	104	91	100	100
"	∞	"	"	500	97	84	97	97
.95	5	20	2	101	26	23.5	22	23.5
"	"	50	"	251	58	53	53	53
"	"	100	"	501	130	98	104	104
"	"	250	"	1251	265	233	252	252
.95	5	100	1	2004	444	400	416	416
"	"	"	2	501	112	98	104	104
"	"	"	10	20.04	4.4	4.0	4.2	4.2
.95	1	10	1	220	91	95	53	95
"	5	"	"	204	60	55	47	55
"	10	"	"	202	52	47	46	47
"	25	"	"	200.8	47	42	44	44
"	∞	"	"	200	40	34	40	40
.95	5	2	.1	4400	1800	1900	1230	1900
"	10	"	"	4200	1450	1440	1060	1440
"	25	"	"	4080	1160	1130	940	1130
"	∞	"	"	4000	800	680	800	800
.95	56	2.23	.2	1125	285	250	245	250

4.9 Second Sample Size Graphs

To determine the second sample size, first specify m, d, and $1-\varepsilon$ in advance. Then take a preliminary sample of size m and compute $m\bar{x}_m$, the sum of the m observations. If $m\bar{x}_m$ is less than 500 use Figure 4.4 to find c for a one point solution or to find c_1 and c_2 if a two point solution is desired. For a one point solution the lower curve labeled c is to be used for $1-\varepsilon = .90$ or $.95$. Similarly the upper curve labeled c is to be used for $1-\varepsilon = .99$. If a two point solution is desired the value of $1-\varepsilon$ is immaterial. For $m\bar{x}_m$ greater than 500 the same procedure applies to Figure 4.5. Next compute $c\bar{x}_m$ or $c_1\bar{x}_m$ and $c_2\bar{x}_m$. Consider the one point solution. If the ratio of $c\bar{x}_m$ to d is less than 30:1 for $1-\varepsilon = .99$, less than 50:1 for $1-\varepsilon = .95$, or less than 60:1 for $1-\varepsilon = .90$ use Figure 4.6. For larger ratios it is necessary to use Figure 4.7. Plot a point of the form $(kd, kc\bar{x}_m)$, for some value of k, on the graph. The value of k chosen is immaterial but the larger the value used the more accurate will be the result. With a straight edge connect the plotted point with the origin. At the point of intersection between this line and the appropriate curve, depending upon the size of $1-\varepsilon$, record the value on the horizontal axis. This value is nd and thus the second sample size, n, is found by dividing this by d. The method to compute n'', needed for the two point solution, is identical with the above except c_2 is used instead of c. To determine n', Figure 4.8 is used for ratios of $c_1\bar{x}_m$ to d less than 33:1 for $1-\varepsilon = .99$, less than 54:1 for $1-\varepsilon = .95$, and less than 63:1 for $1-\varepsilon = .90$. Larger ratios require Figure 4.9. Again the procedure is the same as that for finding n'' except c_2 is replaced by c_1. Finally, to determine the second sample size in the two point solution take the larger of n' and n".

As an example, consider an actual experiment in which flowers were exposed to low level irradiation and the number of discolored sectors per petal were counted for 56 petals.

TABLE 4.2

Frequency Count	0	1	2	3	4	5	6	7
Observed	7	11	16	11	8	2	0	1

In this experiment $m=56$, $\bar{x}_m = 2.23$ and $s^2 = 2.22$. Assuming this follows the Poisson distribution the true value of λ can be estimated within .2 of a unit with confidence coefficient $1-\varepsilon = .95$ as follows. We compute $m\bar{x}_m = 125$. From Figure 4.4 we see that $c = 1.287$ which implies that $c\bar{x}_m = 2.87$. The ratio of $c\bar{x}_m$ to d is 14.35:1 enabling us to use Figure 4.6. The largest value of k possible to use is 1000, which corresponds to the point (200, 2870). The intersection of the straight line joining this point with the origin and the $1-\varepsilon = .95$ curve has a value of $nd = 57$ on the horizontal axis. Thus the second sample size for the one point solution is $n=57/d=285$. For the two point solution c_1 and c_2 are read from Figure 4.4 as $c_1 = 1.377$ and $c_2 = 1.122$ implying $c_1\bar{x}_m = 3.07$ and $c_2\bar{x}_m = 2.50$. Again setting $k = 1000$ we plot the point (200, 3070) in Figure 4.8, connect it to the origin, and the intersection with the $1-\varepsilon = .95$ curve is at $n'd = 49$ on the horizontal axis. Thus $n' = 49/d = 245$. Similarly, plotting the point (200,2500) in Figure 4.6 yields $n''d=50$ implying $n''=50/d=250$. Therefore the two point second sample size is $n=250$.

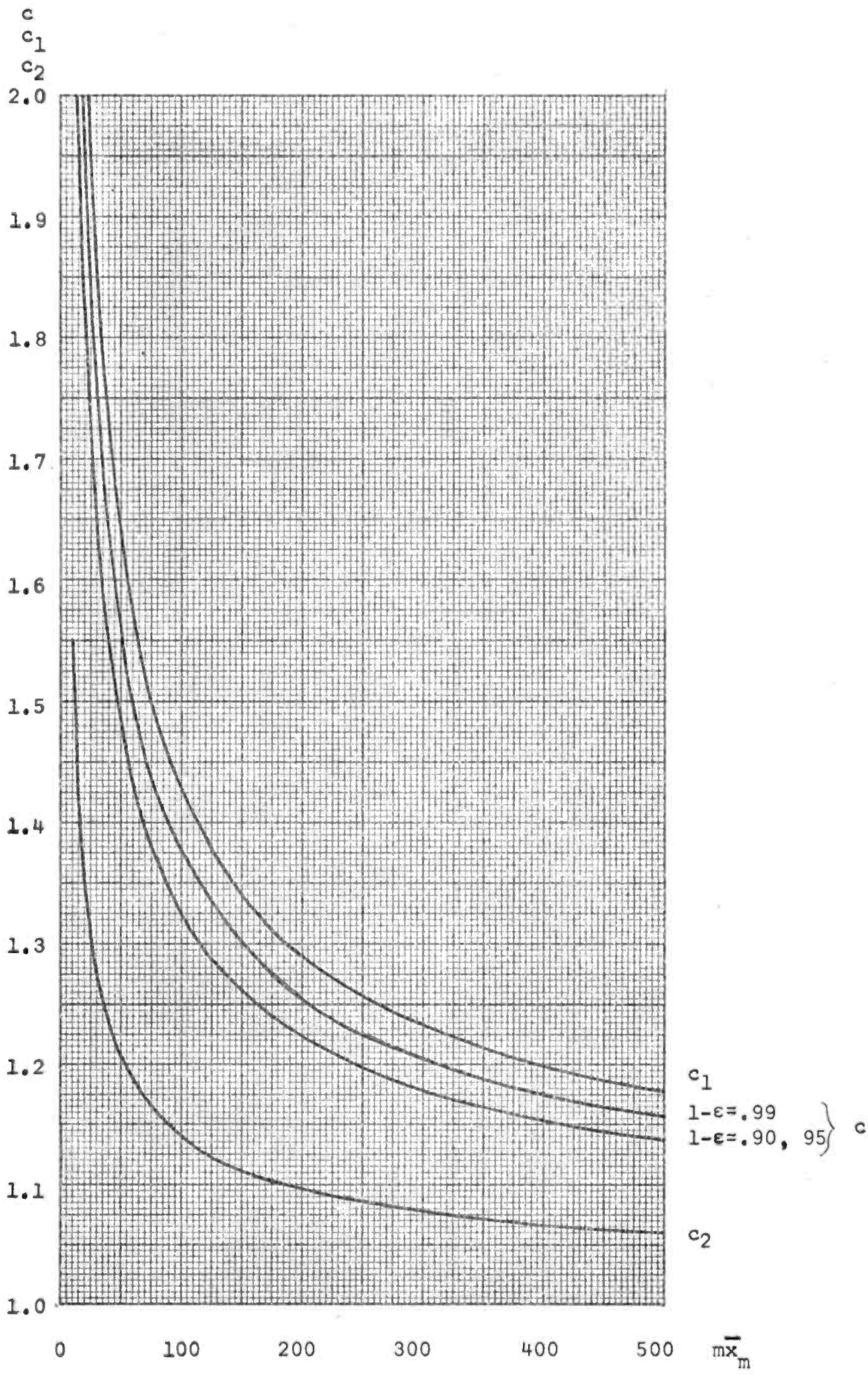

c
c1
c2
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0
100
200
300
400
500
mx̄m
c1
1-ε=.99
1-ε=.90, 95
c
c2

Figure 4.4

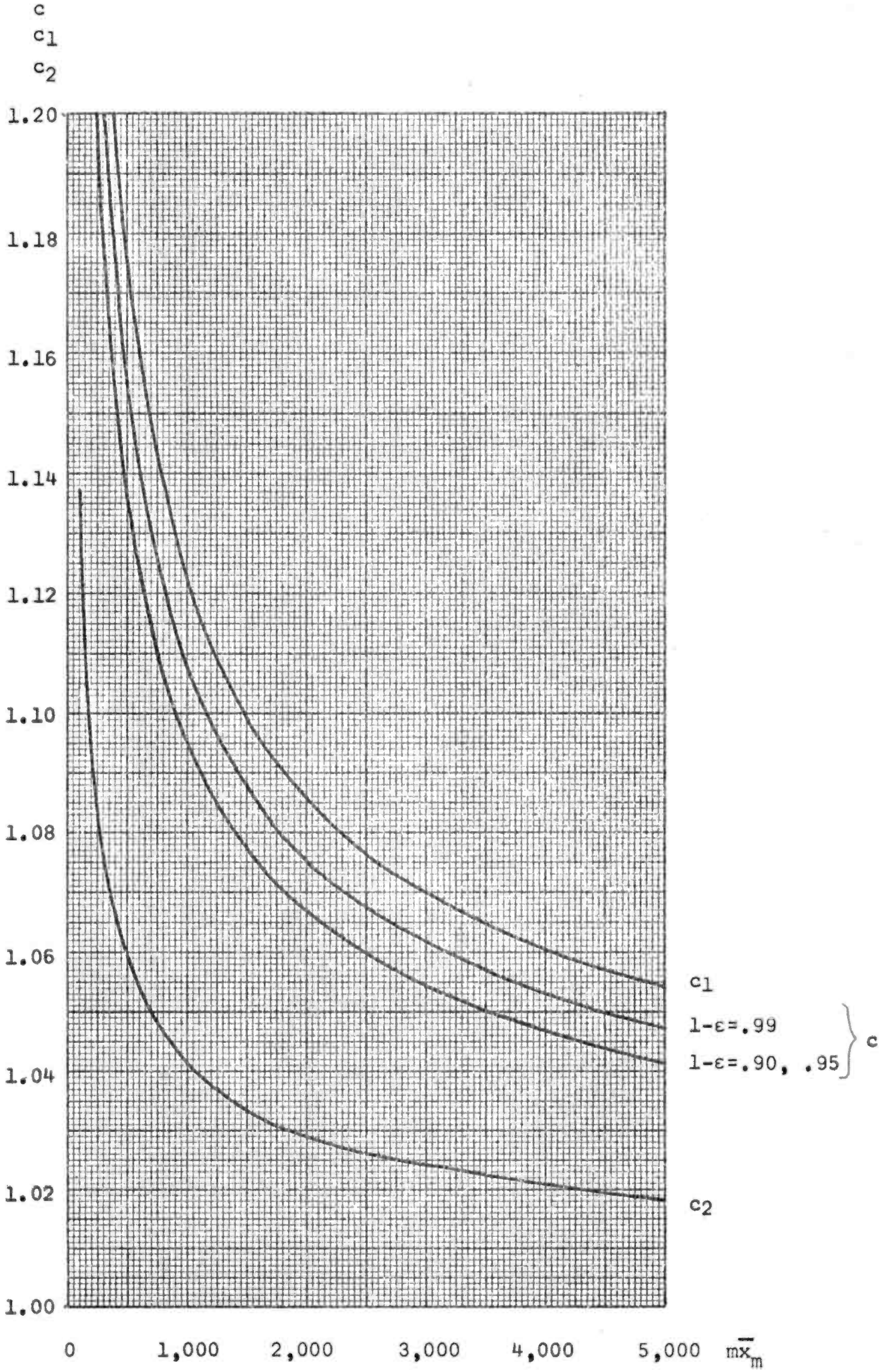

c
c1
c2
1.20
1.18
1.16
1.14
1.12
1.10
1.08
1.06
1.04
1.02
1.00
0
1,000
2,000
3,000
4,000
5,000
mx̄m
c1
1-ε=.99
1-ε=.90, .95
c
c2

Figure 4.5

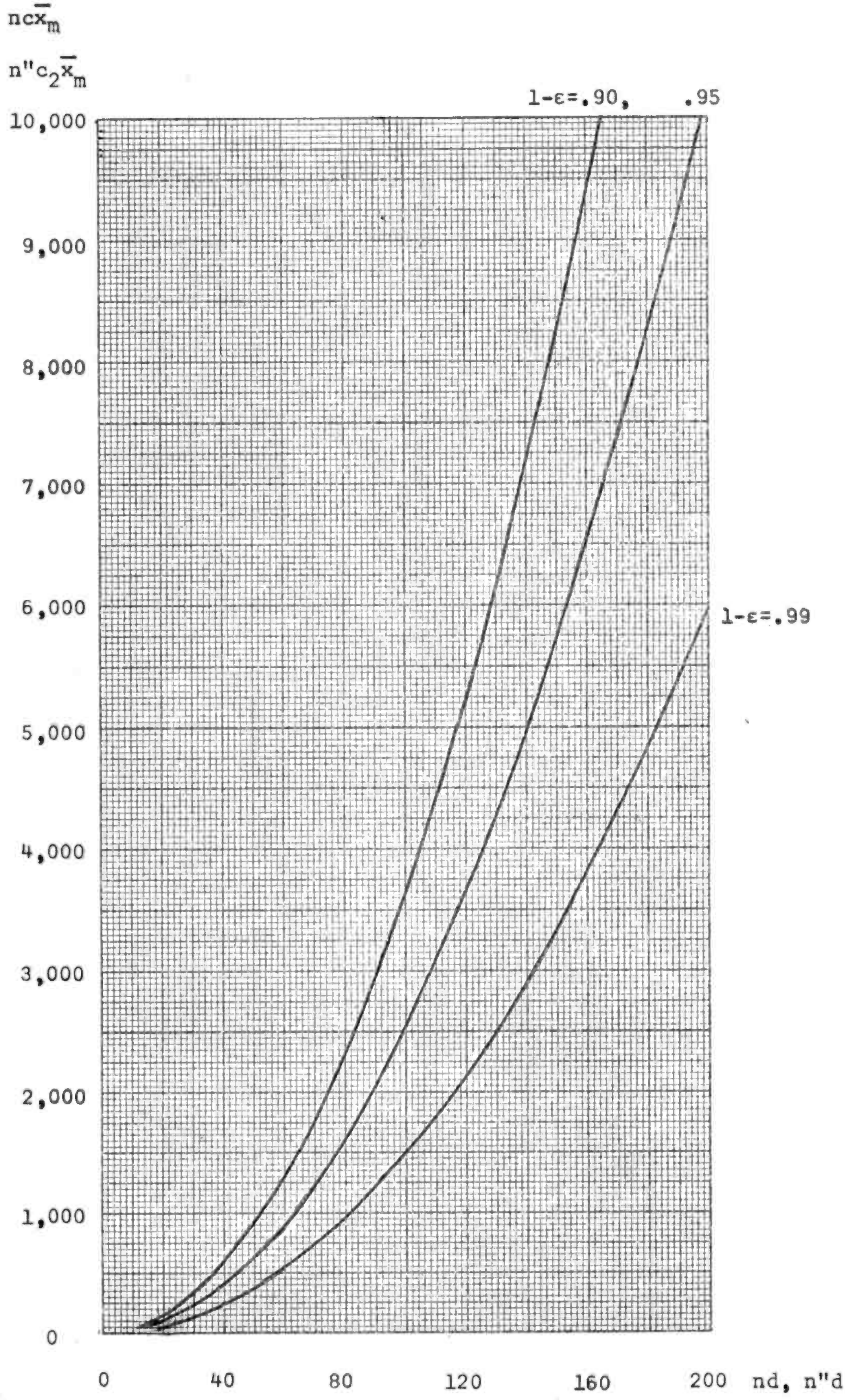
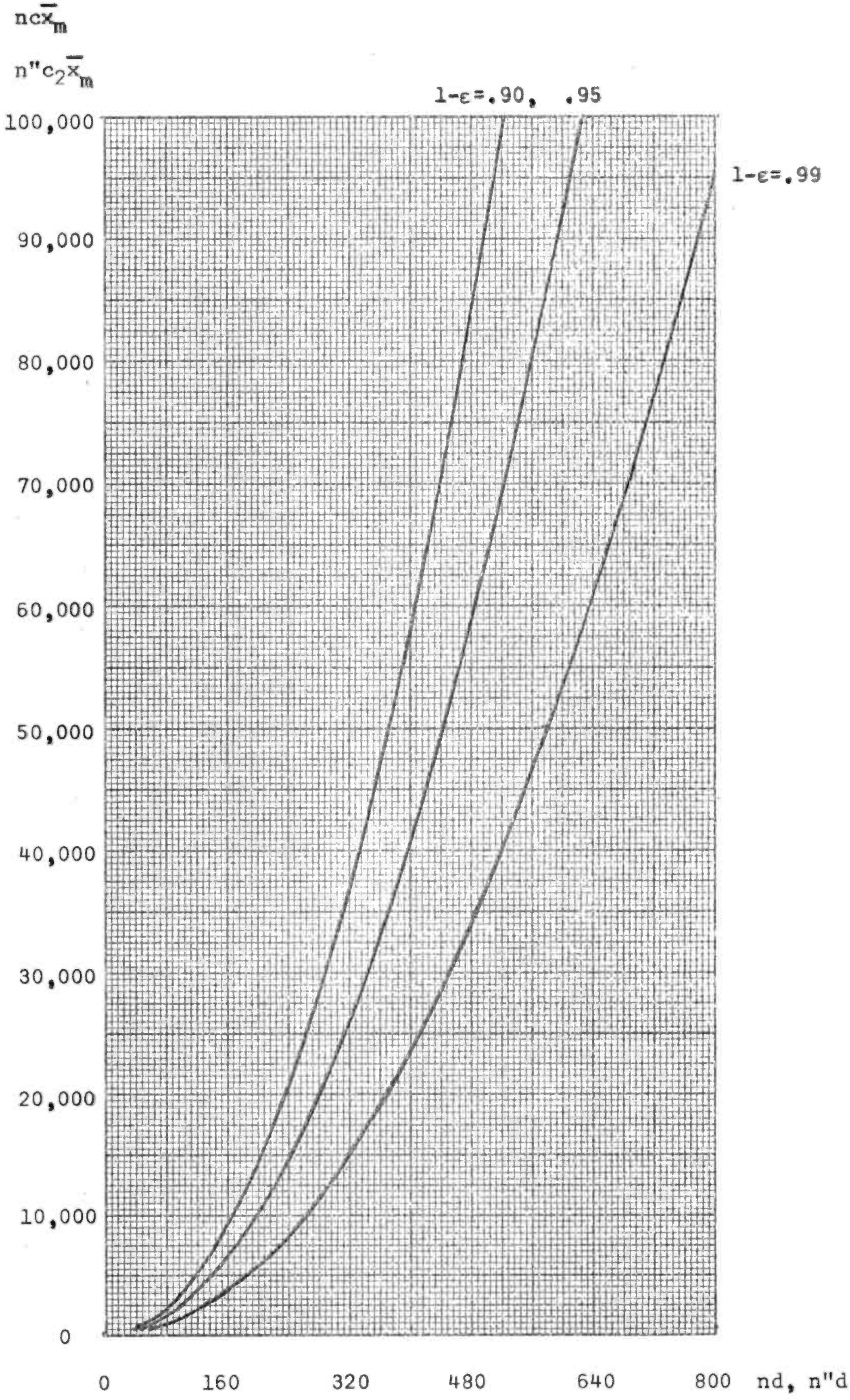
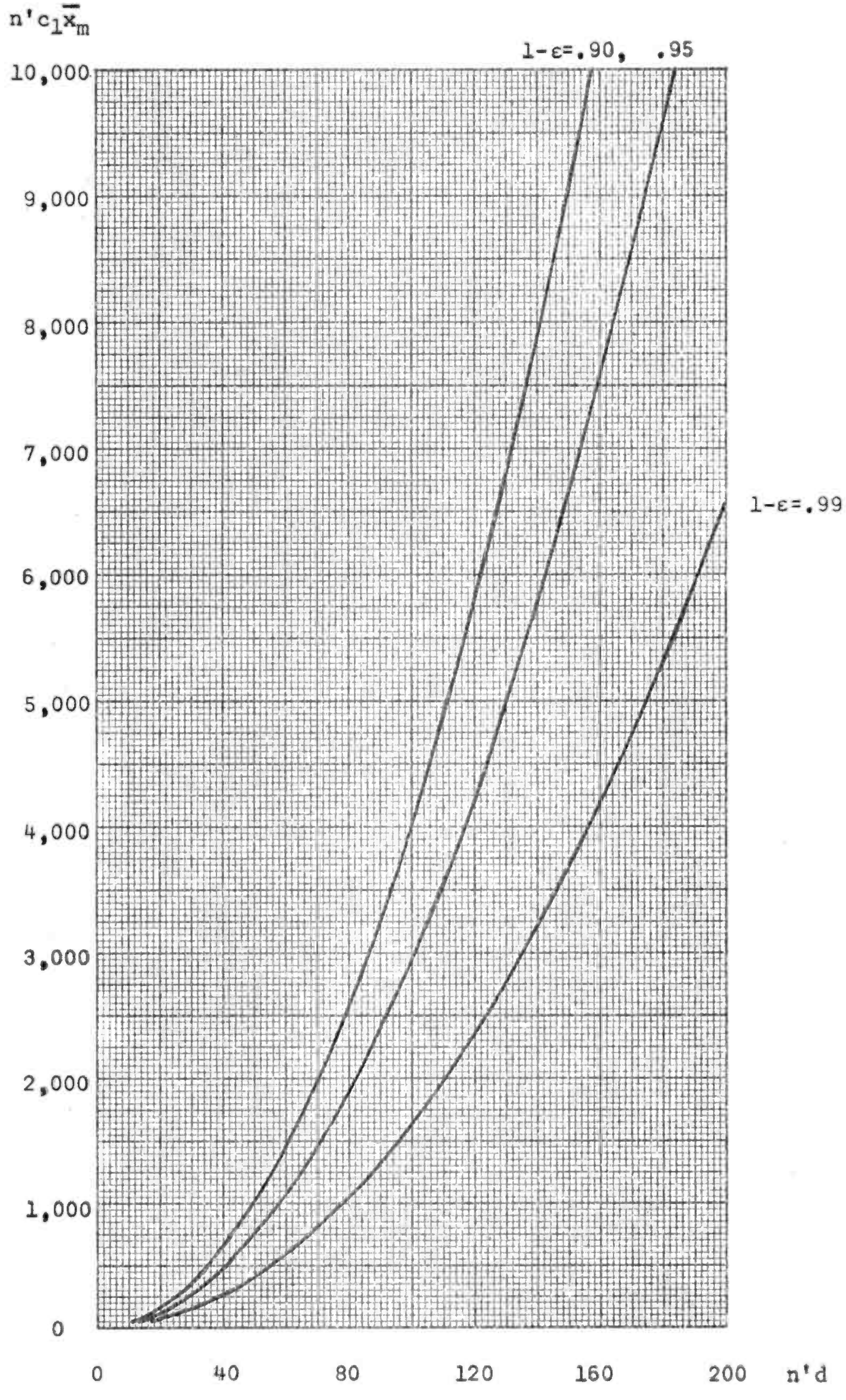

$nc\overline{x}_m$
$n''c_2\overline{x}_m$
1-ε=.90,
.95
1-ε=.99
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
0
40
80
120
160
200
nd, n"d

Figure 4.6



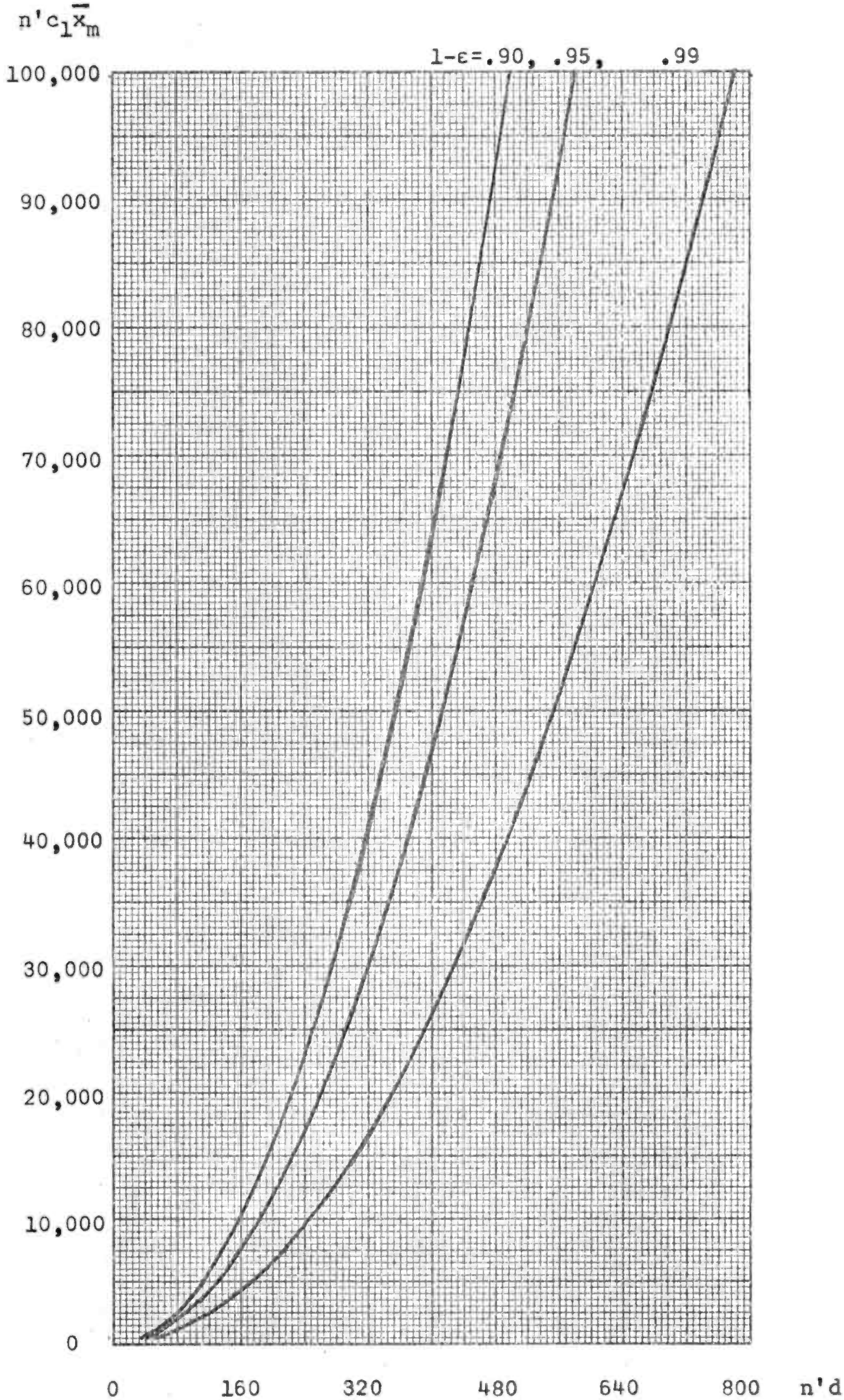
ncx̄m
n"c2x̄m
1-ε=.90, .95
1-ε=.99
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
0
160
320
480
640
800
nd, n"d

Figure 4.7



$n'c_1\overline{x}_m$
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
$1-\varepsilon=.90$, .95
$1-\varepsilon=.99$
0
40
80
120
160
200
$n'd$

Figure 4.8



$n'c_1\bar{x}_m$
$1-\varepsilon=.90,\ .95,\ \ .99$
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
0
160
320
480
640
800
$n'd$

Figure 4.9

To derive the graphs for the one point solution various values of β and α were tried for the particular case where $m=5$, $\bar{x}_m=100$, and $d=2$ such that $(1-\beta)(1-\alpha)=1-\varepsilon$. The optional values creating the smallest sample size for $1-\varepsilon=.99$ were $\beta=.0005$, $1-\alpha=.9905$. For $1-\varepsilon=.90$ and $1-\varepsilon=.95$ the optimum values for β were so close that $\beta=.002$ worked for both with negligible loss in sample size but a gain in simplicity. The $1-\alpha$ values were therefore .902 and .952 respectively. These values were used in Figures 4.4, 4.5, 4.6, and 4.7.

In the two point solution it was also necessary to optimize γ and δ (see Section 5). Again the case where $m=5$, $\bar{x}_m=100$, and $d=2$ was chosen to optimize. In particular, for $1-\varepsilon=.95$ the optimum was at $1-\varepsilon=.9520$, $\beta=.1379$, $\gamma=.9865$, and $\delta=.999$. This allowed the use of the same graph as in the one point solution with $1-\varepsilon=.95$ to compute n''. Therefore to optimize in the $1-\varepsilon=.90$ case I set $1-\alpha=.902$ and the other values were optimized at $\beta=.1003$, $\gamma=.9800$, and $\delta=.998$. Finally, to optimize for $1-\varepsilon=.99$ I set $1-\alpha=.9905$ and $(1-\delta)\beta=.0001$ (in the $1-\varepsilon=.90$ case $(1-\delta)\beta$ optimized at .0002 and for $1-\varepsilon=.95$ at .00014). This led to the optimum values $\beta=.0961$, $\gamma=.9958$, and $\delta=.999$. Because the sample size is fairly insensitive to small changes in β, a common value of $\beta=.10$ was chosen. Similarly, the same value of $\delta=.999$ was used with negligible loss. Thus $(1-\delta)\beta=.0001$ for all three cases. This consolidation then forced new values of γ such that $1-\varepsilon=(1-\alpha)(1-\beta+\beta\gamma\delta)$ as desired. The new values of γ were $\gamma=.981$, .981, and .996 for $1-\varepsilon=.90$, 95 and .99 respectively. These changes greatly reduced the number of graphs involved but did not materially change the sample size.

Figures 4.4 and 4.5 graph c versus z for given values of $H(c;z)$ where $H(c;z)$ is defined by (4.6). In particular, the curves from top to bottom represent $H=.0001$, $.0005$, $.002$, and $.10$. Figures 4.6 and 4.7 graph $nc\bar{x}_m$ versus nd for various values of $t(n;c\bar{x}_m,d)$ which is defined by (4.8). This type of graph is practical because $t(n;c\bar{x}_m,d) = t(1;nc\bar{x}_m,nd)$. From left to right the curves correspond to $t=.902$, $.952$, and $.9905$. Figures 4.8 and 4.9 are similar to Figures 4.6 and 4.7 with the curves corresponding to $t=\gamma(1-\alpha)=.8849$, $.9339$, and $.9865$ from left to right.

The computations in Figures 4.4 and 4.5 were made from existing tables for values of $z \leq 50$. Table 4.3 displays these. For larger values of z the normal approximation to the Poisson was used. No loss of accuracy was evident in several examples which were checked. The values for Figures 4.6 - 4.9 were obtained by use of the IBM 1620 computer, using a program which computed the actual values by summation of the appropriate Poisson distribution.

Table 4.3

	$m\bar{x}_m$	10	20	30	40	50
	c_1	2.785	2.120	1.870	1.730	1.638
c	$1-\varepsilon=.99$	2.525	1.968	1.753	1.636	1.556
c	$1-\varepsilon=.90,\ .95$	2.299	1.830	1.650	1.548	1.480
	c_2	1.550	1.354	1.277	1.234	1.206

V. UNSOLVED PROBLEMS

The two procedures given in Chapters III and IV clearly are an improvement over existing methods. These solutions, however, raise additional problems to be investigated.

Both procedures supply a lower bound but the upper bound should also be given consideration. There is some information on an upper bound in the Poisson problem. For instance, in the one point solution

$$\Pr[|\lambda-\hat{\lambda}_n| < d] < 1 \cdot \Pr[n \geq n_1]+(1-\alpha)\Pr[n < n_1]$$

$$= 1 \cdot (1-\beta)+(1-\alpha)\beta = 1-\alpha\beta \;.$$

For $1-\varepsilon=.90$, $.95$, $.99$ this gives upper bounds of .999804, .999904, and .99999525 respectively. These are not very useful. Since they were obtained using the optimal values for α and β there exist other values of α and β with $(1-\alpha)(1-\beta)=1-\varepsilon$, that will yield larger sample sizes and at the same time decrease the upper bound; $1-\alpha=1-\beta=\sqrt{1-\varepsilon}$ for example. An upper bound for a large sample size would also be an upper bound for a small sample, thus the upper bound could be lowered. Other techniques should be investigated to put better limits on the confidence coefficient, since a confidence coefficient much larger than the level desired increases sample size and wastes resources.

Another problem left unanswered is whether or not more information can be used. For instance, both procedures described here neglect the first sample once it has been used to determine the size of the second sample. Could the preliminary sample be used somehow with the second sample in the estimator? This seems to be a complex problem. As a starting difficulty, what estimator should be used?

An important problem to consider is that of the size of the first sample. To determine this some knowledge is needed of the approximate size of the parameter. A three-step procedure may be required, where the first step would be needed to find a first, rough approximation to the size of the parameter.

In Chapter IV, Section 7, it was shown how to estimate the parameter in a Poisson process with a two-step procedure. It was necessary to specify a constant τ to do this. Some investigation should be made to devise a scheme for picking the best value for τ. Perhaps a three-step solution would be necessary. The first step would be used to find τ.

The graphs in Section 9 of Chapter IV were derived by finding the combination of $\alpha, \beta, \gamma, \delta$ which minimized the second sample size for the particular values $m=5$, $d=2$, and $\bar{x}_m=100$. This is fully described in Chapter IV, Section 9. If the second sample size was minimized for each individual set of values of m, d, and $\bar{x}_m$, reductions would be made in sample size. The feasibility of doing this should be investigated. For one indication of how good the particular optimization procedure used is, for the two point solution, examine the difference between n' and n''. Table 4.1 shows that $n'=95$ and $n''=53$ for $1-\varepsilon=.95$, $m=1$, $\bar{x}_m=10$, and $d=1$, indicating the optimization was not good for these values. In fact n is greater than the one point solution sample size which is 91. Similar results occur for $1-\varepsilon=.95$, $m=5$, $\bar{x}_m=2$, and $d=.1$. It is noted though that larger values of m reduce this difference in both cases and also decrease the two point solution second sample size below that of the one point solution.

An investigation could be made into procedures for solving the problems presented in Chapters III and IV by minimizing expected sample size. This would require a different approach from those in Chapters III and IV.

Also of interest would be a Bayes' type of solution in which various loss functions could be considered.

The solution for the variance of a normal problem in Chapter III can be generalized to estimate the mean in the gamma distribution with little modification. To solve other problems with this technique it would be necessary to first derive improvements on Tchebycheff's inequality for the distribution involved. The solution for the Poisson mean problem in Chapter V seems to be very useful and the same technique could be used in many problems requiring a two-step procedure. In fact, it may result in lower sample sizes in estimating the variance of the normal. This problem and many others should be investigated using the technique in Chapter IV.

BIBLIOGRAPHY

1. Abbott, J. H. and Rosenblatt, Judah, Two stage estimation with one observation on first stage, Annals of the Institute of Statistical Mathematics, 1963, 14:229-235.

2. Anderson, T. W., A modification of the sequential probability ratio test to reduce sample size, Annals of Mathematical Statistics, 1960, 31:165-197.

3. Bahadur, R. R. and Savage, L. J., The nonexistence of certain statistical procedures in nonparametric problems, Annals of Mathematical Statistics, 1956, 27:1115-1122.

4. Bechhofer, R. E., A single-sample multiple decision procedure for ranking means of normal populations with known variances, Annals of Mathematical Statistics, 1954, 25:16-39.

5. Bechhofer, R. E., Dunnett, C. W. and Sobel, M., A two-sample multiple decision procedure for ranking means of normal distributions with a common unknown variance, Biometrika, 1954, 41:170-176.

6. Birnbaum, Z. W. and Zuckerman, H. S., A graphical determination of sample size for Wilks' tolerance limits, Annals of Mathematical Statistics, 1949, 20:313-316.

7. Birnbaum, Allan, Sequential tests for variance ratios and components of variance, Annals of Mathematical Statistics, 1958, 29:504-514.

8. Birnbaum, A., and Healy, W. C., Jr., Estimates with prescribed variance based on two-stage sampling, Annals of Mathematical Statistics, 1960, 31:662-676.

9. Blum, J. R. and Rosenblatt, Judah, On multistage estimation, Annals of Mathematical Statistics, 1963, 34:1452-1458.

10. Chapman, Douglas G., Some two-sample tests, Annals of Mathematical Statistics, 1950, 21:601-606.

11. Chernoff, H., Sequential design of experiments, Annals of Mathematical Statistics, 1959, 30:755-770.

12. Dantzig, G. B., On the nonexistence of tests of "Student's" hypothesis having power functions independent of σ, Annals of Mathematical Statistics, 1940, 11:186-192.

13. Epstein, B., Estimates of bounded relative error for the mean life of an exponential distribution, Technometrics, 1961, 3:107-109.

14. Farrell, Roger, Sequentially determined bounded length confidence intervals, Unpublished doctoral thesis, University of Illinois.

15. Ghurye, S. G. and Robbins, Herbert, Two stage procedures for estimating the difference between means, Biometrika, 1954, 41:146-152.

16. Graybill, F. A., Determining sample size for a specified width confidence interval, Annals of Mathematical Statistics, 1958, 29:282-287.

17. Graybill, Franklin A. and Connell, Terrence L., Sample size required to estimate the ratio of variances with bounded relative error, Journal of the American Statistical Association, 1963, 58: 1044-1047.

18. Graybill, Franklin A. and Connell, Terrence L., Sample size required to estimate the parameter in the uniform density within d units of the true value, Journal of the American Statistical Association, 1964, 59:550-556.

19. Graybill, F. A. and Morrison, R. D., Sample size for a specified width confidence interval on the variance of a normal distribution, Biometrics, 1960, 16:636-641.

20. Greenwood, J. A. and Sandomire, Marion M., Sample size required for estimating the standard deviation as a percent of its true value, Journal of the American Statistical Association, 1950, 45:257-260.

21. Healy, W. D., Jr., Two-sample procedures in simultaneous estimation, Annals of Mathematical Statistics, 1956, 27:687-702.

22. Hoeffding, W., Lower bounds for the expected sample size and the average risk of a sequential procedure, Annals of Mathematical Statistics, 1960, 31:352-368.

23. Lehmann, E. L. and Stein, Charles, Completeness in the sequential case, Annals of Mathematical Snatistics, 1950, 21:376-285.

24. Lindley, D. V., Binomial sampling schemes and the concept of information, Biometrika, 1957, 44:179-186.

25. Moshman, Jack, A method for selecting the size of the initial sample in Stein's two-sample procedure, Annals of Mathematical Statistics, 1958, 29:1271-1275.

26. Nordin, J. A., Determining sample size, Journal of the American Statistical Association, 1944, 39:497-506.

27. Putter, Joseph, Sur une methode de double echantillonnage pour estimer la moyenne d'une population la placienne stratifee, Review of the International Statistical Institute, 1951, Part 3, 19:1-8.

28. Ruben, H., Studentisation of two-stage sample means from normal populations with unknown common variance, Sankhya, Series A, 1961, 23:231-250.

29. Seelbinder, B. M., On Stein's two-stage sampling scheme, Annals of Mathematical Statistics, 1953, 24:640-649.

30. Stein, Charles, A two-sample test for a linear hypothesis whose power is independent of the variance, Annals of Mathematical Statistics, 1945, 16:245-258.

31. Wald, A., Sequential tests of statistical hypotheses, Annals of Mathematical Statistics, 1945, 16:117-186.

32. Wald, A., Sequential method of sampling for deciding between two courses of action, Journal of the American Statistical Association, 1945, 40:277-306.

33. Weiss, L., Confidence intervals of preassigned length for quantiles of unimodal populations, Naval Research Logistics Quarterly, 1960, 7:251-256.

34. Wilks, S. S., Mathematical Statistics, Princeton University Press, 1943, 284 pp.

35. Wilton, T. R., A proof of Burnside's formula for $\log(x + 1)$ and certain allied properties of Riemann's ζ-function, Messenger of Mathematics, 1922, 52:90-93.

36. Wormleighton, E., A useful generalization of the Stein two-sample procedure, Annals of Mathematical Statistics, 1960, 31:217-221.

ABSTRACT OF DISSERTATION

SOME TWO-STEP SAMPLING PROCEDURES

Two-step sampling procedures are presented to estimate the variance of a normal distribution and the mean of a Poisson distribution within d units with a specified confidence coefficient.

The procedure to estimate the variance of a normal is based on a Tchebycheff type inequality derived especially for the gamma distribution. A different type of argument, which could be applied to many other distributions, was used to solve the problem for the Poisson distribution.

Sampling sizes are presented in tables and graphs to implement the two solutions. Also, favorable comparisons are made with existing methods.

Terrence Lee Connell

Department of Mathematics and Statistics

Colorado State University

February 1966