DETERMINATION OF SAMPLE SIZE

By

AARON GOLDMAN

Bachelor of Science

Oklahoma Agricultural and Mechanical College

Stillwater, Oklahoma

1954

Submitted to the faculty of the Graduate School of
the Oklahoma State University in partial
fulfillment of the requirements
for the degree of
MASTER OF SCIENCE
May, 1958

OKLAHOMA
STATE UNIVERSITY
LIBRARY

NOV 5 1958

DETERMINATION OF SAMPLE SIZE

Thesis Approved:

Franklin A. Graybill
Thesis Adviser

Earl E. Marshall

Robert MacVean
Dean of the Graduate School

409881

PREFACE

When conducting scientific research, the experimenter is confronted with the problem of taking a minimum sample size and still be confident of the results.

This thesis will help provide an answer to this problem. Four methods are given for determining the sample size necessary from an infinite population. A desirable sample size from a finite population is briefly discussed. An applied problem in textile research is also presented.

Indebtedness is acknowledged to Dr. Franklin Graybill for his valuable guidance and assistance in preparing this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

When we desire the value of a parameter in a given distribution, we are faced with the problem of determining a sample size that will yield a realistic result. Of course, if we had unlimited time and resources there would not be any problem because in that case we could utilize most of the population. We shall consider the case where we are sampling from a normal population of random variables. We want to minimize time, work, and money and at the same time have confidence that a sample will give reliable results.

The following represents a hypothetical situation: An experimenter arbitrarily decides to use a sample of size 100 in an effort to estimate μ, the population mean. Therefore a confidence interval is placed on μ. (We will discuss later how intervals are determined and show the relationship between the sample size and the width of the interval.)

The experimenter finds that the computed interval is too wide. Therefore, in order to decrease the interval width he takes a sample of size 500 feeling that this is a "good" number. Another interval is placed on μ but this time the interval is so small that half as large a sample would have done the job satisfactorily. This is perhaps extreme. However, it is conceivable that there would be many advantages to the experimenter if he were to have an idea of how many samples to take for

a given reliability.

This paper will enumerate methods that will give the experimenter an indication of how many samples are necessary to obtain a specified width confidence interval on μ at a given probability level.

The following notations will be used:

X_i : X_i is a random variable from a normal population with mean μ and variance σ^2.

$\bar{X}$: $\sum_{i=1}^{n} X_i/n$

s^2 : $\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$

$1 - \alpha$: the confidence coefficent

w: width of a confidence interval

d: a predetermined width of a confidence interval

g^2: a predetermined choice of σ^2; it could be (1) an unbiased estimate of σ^2 determined from another experiment, (2) an actually known value of σ^2, or (3) an approximation.

E(K): expected value of K

β: the specified probability that w is less than d

Z_α : Standard normal deviate such that the probability of a larger value is $1 - \alpha/2$

t_α : Student's "t" variate such that the probability of a larger value is $1 - \alpha/2$

$X^2_{(1-\beta)}$: A chi-square variate such that the probability of a smaller value is β.

[Q]: If Q is any number then [Q] is the smallest integer greater than or equal to Q.

Outline of Procedure: The probability that $\bar{X}$ is equal to μ is zero; consequently we use confidence intervals to determine the probability that the interval will include μ. We will explain how to determine n such that certain relationships involving d and w will be established. These will be presented under three specified conditions in the following order:

I. When σ^2 is known (d, α specified)

   A. $P(d = w) = 1.00$
   B. $P(\bar{X} - d/2 < \mu < \bar{X} + d/2) = 1 - \alpha$

II. When σ^2 is unknown and g^2 is used (d, g^2, α, and β specified)

   A. Case A

      1. $P(\bar{X} - w/2) < \mu < \bar{X} + w/2) = 1 - \alpha$
      2. $E(w) = d$

   B. Case B

      1. $P(\bar{X} - w/2 < \mu < \bar{X} + w/2) = 1 - \alpha$
      2. $P(w \leq d) = \beta$

III. When σ^2 is not known; Stein's Two Step Test (d, α specified)

   A. $P(\bar{X} - w/2 < \mu < \bar{X} + w/2) = 1 - \alpha$
   B. $P(w < d) = 100\%$

CHAPTER I

THE POPULATION VARIANCE IS KNOWN

A known variance is a rather trivial case because seldom does the experimenter know the true variance of the population. The experimenter would like to have a sample size such that the width of the confidence interval on μ will be equal to d. At the same time, he wants $1 - \alpha$ to be exact. If the experimenter knows σ^2, he is in a position which is as close to ideal as he could possibly hope to obtain.

We desire an n such that the probability that $|\bar{X} - \mu|$ is greater than $d/2$ equals $1 - \alpha$. This can be written as

$$P(|\bar{X} - \mu| \geq d/2) = 1 - \alpha$$

$\sqrt{n}(\bar{X}-\mu)/\sigma$ is distributed normally with mean zero and variance one.

We obtain

$$P(|\bar{X} - \mu|/\sigma\sqrt{n}) \geq Z_\alpha = 1 - \alpha .$$

This is equivalent to

$$P(|\bar{X} - \mu| \geq Z_\alpha \sigma/\sqrt{n}) = 1 - \alpha .$$

Thus

$$d/2 = Z_\alpha \sigma/\sqrt{n}$$

which gives

$$n = 4Z_\alpha^2 \sigma^2/d^2.$$

Note that Z_α may be found by looking up the tabular value in normal tables. Thus, the desired width is obtained. Hence, if the variance is known, we may find an n such that a specified confidence length will be attained and the confidence interval will have a $1 - \alpha$ probability of including μ.

Example 1:

$$d = 2.0 \quad 1 - \alpha = .95 \quad \sigma = 3.1 \quad Z_\alpha = 1.96$$

$$n = 4\,(1.96)^2\,(3.1)^2 / (2)^2 = 36.917$$

A sample of size 37 is required to produce a confidence interval of width equal 2 regardless of $\bar{X}$ and

$$P(|\bar{X} - \mu| > 1) = .95$$

or

$$P(\bar{X} - 1 < \mu < \bar{X} + 1) = .95.$$

Example 1 illustrates that a value of n equal to 37 under the given conditions will always yield a $1 - \alpha$ confidence interval with width equal to 2.

CHAPTER II

THE VARIANCE IS UNKNOWN

If σ^2 is unknown, we cannot be assured of finding an n such that w will be equal to d. We will use g^2 in place of σ^2 to determine n in the following cases:

(A) $E(w) = d$

(B) $P(w \leq d) = \beta$

The effects of g^2 will be discussed in each case.

CASE (A):

We desire an n such that

$$P(|\bar{X} - \mu| > w/2) = 1 - \alpha$$

and $E(w) = d$.

$\sqrt{n}\,(\bar{X} - \mu)/s$ is distributed as Students' distribution with $n - 1$ degrees of freedom.

Therefore

$$P[(|\bar{X} - \mu|\sqrt{n})/s \geq t_\alpha] = 1 - \alpha$$

and

$$P[|\bar{X} - \mu| \geq t_\alpha\, s/\sqrt{n}] = 1 - \alpha$$

or

$$P[\bar{X} - \frac{t_\alpha s}{\sqrt{n}} < \mu < \bar{X} + \frac{t_\alpha s}{\sqrt{n}}] = 1 - \alpha$$

Hence

$$w = 2t_{\alpha}\, s/\sqrt{n}\ .$$

We want $E(w) = d$. Therefore

$$E(w) = \frac{2t_{\alpha}}{\sqrt{n}}\ \frac{\sqrt{2}}{\sqrt{n-1}}\ \frac{(\frac{n-2}{2})!}{(\frac{n-3}{2})!}\ \sigma.$$

We substitute c_2 and g to obtain

$$d = E(w) = \frac{2t_{\alpha}}{\sqrt{n-1}}\ c_2 g$$

where

$$c_2 = \sqrt{2/n}\ \frac{(\frac{n-2}{2})!}{(\frac{n-3}{2})!}\ .$$

The effects of using g: Because it is not certain that g^2 is equal to σ^2, our results are not exact. If $g > \sigma$ then E(w) is proportionally larger than d, and n is larger than required. If $g < \sigma$ then E(w) is proportionally smaller than d and n will be smaller than required.

The value of g does not affect the confidence coefficient; hence $1 - \alpha$ is exact.

Example 2:

$$d = 2.0;\ \alpha = .05;\ g^2 = 9.0$$

$$2.0 = \frac{(2)\ (2.03)\ (0.978)\ (3)}{\sqrt{n-1}};\ \ n = 36$$

A graph is furnished so that n may be readily determined. To use the graph in Example 2, we procede as follows:

(a) Compute $g/d = 1.5$.

(b) Refer to the corresponding n on the chart. At the $1 - \alpha = .90$ probability level $n = 26$; at the $1 - \alpha = .95$ level $n = 36$; at the $1 - \alpha = .99$ level $n = 62$.

The example also illustrates that we do not place an interval on μ until after s^2 has been computed. Thus g^2 has no effect on the final confidence interval.

CASE B

We desire an n such that:

(1) $P(|\bar{X} - \mu| > \frac{w}{2}) = 1 - \alpha$

(2) $P(w \leq d) = \beta$

(1) has been solved in case A and we obtained

$$w = 2t_{\alpha} \, s/\sqrt{n}.$$

We now desire the $P(w \leq d) = \beta$.

Substituting for w

$$P(2t_{\alpha} \, s/\sqrt{n} \leq d) = \beta,$$

hence

$$P(s \leq \frac{d\sqrt{n}}{2t_{\alpha}}) = \beta$$

and

$$P(\frac{s^2 (n - 1)}{\sigma^2} \leq \frac{d^2 n (n - 1)}{4t_{\alpha}^2 \sigma^2}) = \beta.$$

It is known that $\frac{s^2 (n - 1)}{\sigma^2}$ is distributed as the central chi-square with n - 1 degrees of freedom.

Hence for $\chi^2_{1-\beta}$ we obtain

$$\chi^2_{1-\beta} \leq \frac{d^2 n(n-1)}{4t_\alpha^2 \sigma^2} .$$

Since g^2 is used instead of σ^2 we set up the ineqality

$$\frac{4\chi^2_{(1-\beta)} t_\alpha^2 g^2}{n(n-1)} \leq d^2.$$

The value of n is obtained by an iterative process such that values of $\chi^2_{(1-\beta)}$, $4t_\alpha^2$, and n (n - 1) combine to form the fraction that is the largest value possible which is less than d^2. It is readily observable that g^2 will not influence α. However, g^2 will have some effect on β. This effect will be discussed later.

Example 3:

$$d = 2; g^2 = 9; \alpha = .05; \beta = .99$$

$$\frac{4(67.5)(4)(9)}{(50)(49)} \leq 4$$

$$n = 50$$

$$P(|\bar{X} - \mu| \geq 2) = .95$$

or

$$P(\bar{X} - 1 \leq \mu \leq \bar{X} + 1) = .95$$

and $$P(w \leq 2) = .99.$$

The influence of g^2. If g^2 is larger than σ^2 then β will be larger than specified and n will be larger than necessary. If g^2 is smaller than σ^2 β will be smaller than specified and n will be smaller than necessary.

Stein's Two Step Method: Methods mentioned previously in this chapter give no assurance that the confidence interval width will be equal to a desired width. The reason is that the value of g^2 is not ordinarily accurate. If the experimenter wants the width of the interval to be less than or equal to a specified value, he can use Stein's method with successful results. In Stein's method, any information about the variance is derived from the population itself. Stein's method has many applications in determining sample size and is a valuable aid to the researcher. Stein's technique assumes a normal population.

The sample is taken in two steps. The first set of values is of size n_1. A confidence interval is placed on μ with

$$w_1 = \frac{2t_\alpha s}{\sqrt{n_1}}$$

where t_α is Students' distribution with $n_1 - 1$ degrees of freedom.

If w_1 is less than or equal to d, the sample is of suitable size. The $1 - \alpha$ probability statement is exact. That is

$$P(\bar{X}_1 - w_1/2 < \mu < \bar{X}_1 + w_1/2) = 1 - \alpha.$$

If w_1 exceeds d then n_2 additional observations are taken until

$$n_2 = \frac{4s^2 t_\alpha^2}{d^2} - n_1.$$

This can be summarized as the following theorem.

The value of n can be obtained in a two step sequence such that

$$P(\bar{X} - w/2 < \mu < \bar{X} + w/2) = 1 - \alpha$$

and

$$P(w \leq d) = 100\%$$

by selecting n equal to either

(1) n_1

or

(2) $\dfrac{4t_\alpha^2 s_1^2}{d^2}$

where $\bar{X}_1 = \dfrac{\sum_{i=1}^{n_1} X_i}{n_1}$, $s_1^2 = \dfrac{\sum_{i=1}^{n_1} (X_i - \bar{X}_1)^2}{n_1 - 1}$, n_1 is the size of the first sample, and t_α has $n_1 - 1$ degrees of freedom.

Proof. Assume X_i is normally and independently distributed with mean μ and variance σ^2. The first sample consists of observations $X_1, X_2, \ldots X_{n_1}$. The sample has a computed mean of $\bar{X}_1$ and variance s^2. A confidence interval is placed on μ. Then

$$P(\bar{X}_1 - w_1/2 < \mu < \bar{X}_1 + w_1/2) = 1 - \alpha$$

where

$$w_1 = \frac{2 s t_\alpha}{\sqrt{n_1}} .$$

If μ is less than d, $n = n_1$ and the specified requirements are met.

If w_1 is larger than d, the additional observations $X_{n_1 + 1}$ $X_{n_1 + 2}, \ldots X_{n_1 + n_2}$ that have mean $\bar{X}_2$ are taken.

Let

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n}$$

where

$$n = n_1 + n_2$$

and

$$n_2 = [\, f(s)].$$

Let

$$y = \sqrt{n}\ (\bar{X} - \mu).$$

If s is fixed then n_2 is fixed and

$$f(y \mid s) = \frac{1}{\sqrt{2\pi}\ \sigma}\ e^{-y^2/2\sigma^2}.$$

Since $f(y \mid s)$ does not involve s then

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

Because

$$f(y \mid s) = f(y),$$

y and s are independently distributed.

It is known that if

(1) y is distributed normally with mean zero and variance σ^2,

(2) s is independent of y,

(3) $\frac{s^2(n_1 - 1)}{\sigma^2}$ is the chi-square distribution with $n_1 - 1$ degrees of freedom,

then y/s follows Student's "t" distribution with $n_1 - 1$ degrees of freedom.

Hence $\frac{\sqrt{n}\ (\bar{X} - \mu)}{s}$ has Students' distribution with $n_1 - 1$ degrees of freedom.

Therefore

$$P(\bar{X} - w/2 < \mu < \bar{X} + w/2) = 1 - \alpha$$

where

$$w = \frac{2t_{\alpha}\, s}{\sqrt{n}} .$$

If we desire $d \leq w$ then we get

$$n = \frac{4t_{\alpha}\, s^2}{d^2}$$

and the theorem is proved.

Example 4:

$$d = 2;\ 1 - \alpha = .95;\quad s^2 = 12;\quad n_1 = 20$$

$$w = (2)\ (2.093)\ (3.464)/4.472 = 3.2$$

is greater than d; hence

$$n = (4)\ (2.093)^2\ (12)/(2)^2 = 53.$$

Therefore 53 - 20 or 33 more samples are needed so that

$$P(\bar{X} - 1 < \mu < \bar{X} + 1) = .95.$$

Summary: Stein's method assures us a desired length of confidence interval 100 per cent of the time. This is distinct

from the probability $1 - \alpha$ that μ is included in the interval. There is a disadvantage in the use of this method when there is a change in the variance of the population (e.g. height of certain plants). The change of variance would contradict the assumption that the variance of the X_i was σ^2.

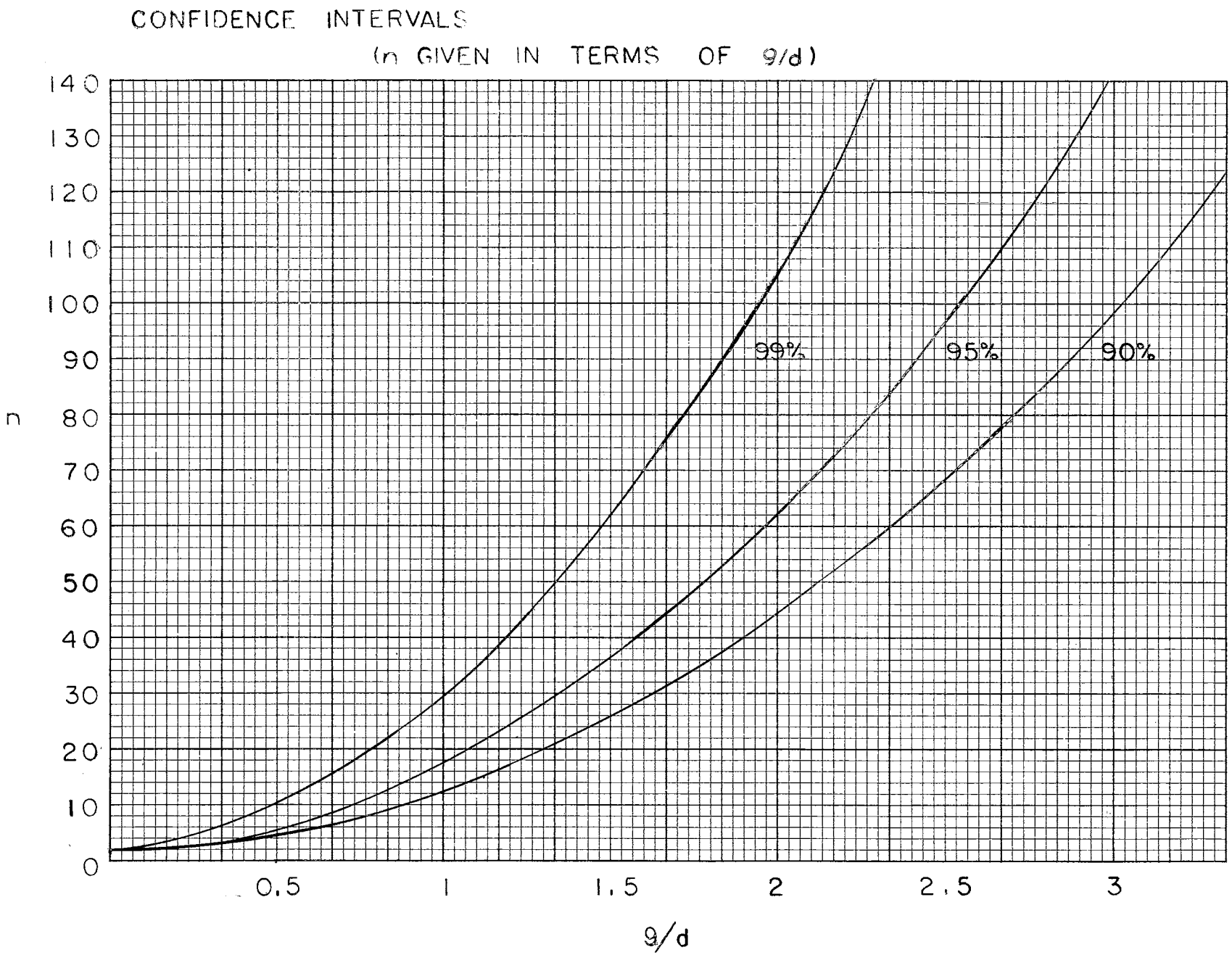
The problem of how many samples should be taken in the first step is answered in part in the next section.

Taking the first sample in Stein's Method. Stein's method leaves the selection of the size of the first sample to the discretion of the experimenter. This might be troublesome if he has no idea as to the size of n_1. Seelbinder (1953) has provided tables that aid in making the decision. These tables enable the experimenter to determine the expected total sample size given the size of the first sample. In order to use the tables, d and α are specified, and σ is known to be contained in a certain interval. The minimax technique is utilized so that the experimenter can take the smallest sample size and still have good results. This technique minimizes the maximum loss in additional observations because of the unknown value of σ^2. If σ^2 were known then no additional observation would be necessary. The maximum differences of the value of n (when σ^2 is known) and the expected value of n (given n_1) are set up in a table. Corresponding to the minimum of the maximum differences, a value of n_1 may be obtained.

TABLE I

SUMMARY OF SELECTING A SAMPLE SIZE CONSIDERING FOUR POSSIBLE ALTERNATIVES

σ^2 KNOWN	σ^2 UNKNOWN BUT ESTIMATED BY g^2 Case A: $E(w) = d$	σ^2 UNKNOWN BUT ESTIMATED BY g^2 Case B: $P(w < d) = \beta$	σ^2 UNKNOWN Stein's Method
Method $n = \frac{4Z_\alpha^2 \sigma^2}{d^2}$ Z_α is standard normal deviate at $1 - \alpha$ level. σ^2 is known variance. d is desired width of interval.	$E(w) = d = \frac{2t_\alpha g C_2}{\sqrt{n-1}}$ t_α is Student's "t" with $n - 1$ degrees of freedom. g is estimate given by the experimenter. c_2 is found in tables.	$\frac{\chi^2_{(1-\beta)} 4t_\alpha^2 g^2}{n(n-1)} = d^2$ $\chi^2_{1-\beta}$ is tabular at β level with $n - 1$ degrees of freedom. t is Student's t with $n - 1$ df. g^2 is estimate of σ^2, d is desired difference.	Sample n_1 observations; compute interval on μ; $w < d$ take no more samples. If $w \geq d$ compute $n = \frac{4t^2 s^2}{d^2}$ using t and s^2 from first sample (t with $n_1 - 1$ df). Take $n - n_1$ samples and compute interval using t and s.
Advantages n can be found such that interval will be exactly what is desired. Computation is easy.	If experimenter is familiar enough with data gives excellent idea of sample size. Computation rather easy.	Experimenter may state with given probability that width is less than desired interval provided g^2 is close estimate.	Width obtained is always less than or equal to d. Easy to compute.
Disadvantages Chances of knowing σ^2 are slim; Might tempt researcher into desiring too small d thereby obtaining n larger than capabilities.	Cumbersome to compute n. If g^2 less than σ^2 then n is too small. Desired results will not be obtained; If g^2 smaller than σ^2 then n is too large.	Rather difficult to compute n. If g^2 less than σ^2 then actual probability will be less than β; hence n is too small. g^2 greater than σ^2 implies β too large and n too large.	If population changes then statistics will be inaccurate.


CONFIDENCE INTERVALS
(n GIVEN IN TERMS OF g/d)
n
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0.5
1
1.5
2
2.5
3
g/d
99%
95%
90%

CHAPTER III

DETERMINATION OF SAMPLE SIZE IN A FINITE POPULATION

Introduction: Determining the sample size necessary for a desired confidence interval on the mean while sampling from a finite population can be handled most readily by the first of the three methods given; however it is quite possible that the results might require utilization of an appreciable amount of the population. In fact, the results may call for n larger than the size of the population itself! For this reason, we introduce the finite population correction (fpc). We define N to be the size of the finite population, n to be the desired sample size and the fpc = $\frac{N - n}{N}$.

Condition given σ^2 Known: This condition is quite probable in finite sampling. Usually the experimenter has a good knowledge of his data and can define σ^2. We procede in the same manner as in determining a sample size from an infinite population except we let n_1 be the value obtained. That is:

$$n_1 = \frac{4Z_\alpha^2 \sigma^2}{d^2} .$$

If the ratio n_1/N is appreciably large (to the experimenter) then we shall use the formula

$$n = \frac{n_1}{1 + n_1/N} .$$

If n_1/N were too small then the experimenter need not worry about the fpc.

Example 5:

$$N = 450 \quad ; \quad \sigma^2 = 90 \quad ; \; d = 4$$

$$n_1 = \frac{(4)\,(1.96)^2\,(90)}{16} = 86.4 \text{ or } 87$$

Since n_1/N does not use a negligible part of the population we compute

$$n = \frac{87}{1 + 87/450} = 72.5 \text{ or } 73.$$

The probability statement is

$$P(|\bar{X} - \mu| < 2) = .95.$$

Conclusion: A sample size of 73 will yield a confidence interval on μ of width less than 4.

Brief proof of method used to determine n: In the above discussion we have used μ and σ^2 to represent the population parameters. In the proof which follows we will use the more common definitions.

$\bar{X}$: Finite Population Mean

$\bar{x}$: Sample mean

S^2 : Population variance

s^2 : sample variance

Computations:

$$\bar{X} = \sum_{i=1}^{N} x_i/N \quad ; \quad \bar{x} = \sum_{i=1}^{n} x_i/n \quad S^2 = \sum_{i=1}^{N} (x_i - \bar{X})^2/N - 1;$$

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \qquad n_1 = \frac{4Z_\alpha^2 \; S^2}{d^2}$$

The variance of $\bar{X}$ is given as

$$s^2_{\bar{x}} = S^2 (N - n)/Nn.$$

Under the assumption $\bar{x}$ is asymptotically distributed normally with mean $\bar{X}$ and variance $S^2_{\bar{x}}$ we obtain

$$d = 2S \sqrt{(N - n)/Nn}.$$

Hence

$$n = \frac{(2tS/d)^2}{(1 + 4t^2 S^2/N d^2)} .$$

From our definition of n_1

$$n = \frac{n_1}{1 + n_1/N} .$$

CHAPTER IV

DETERMINATION OF SAMPLE SIZE IN A TEXTILE EXPERIMENT

Synopsis

A textile experimenter is interested in the mean breaking strength of white muslin cloth. He would like to take a conservative number of swatches from a bolt of cloth and still have confidence in the results. Two methods for obtaining a desirable sample size are presented. Method I will be used to obtain a sample size such that the confidence interval width will be smaller than a specified number. Method II will be used to obtain a sample size such that there is a specified probability that a confidence interval width is less than or equal to a specified value.

Introduction. Suppose a textile experimenter wanted to estimate the mean breaking strength of a population of muslin cloth. Suppose he gathered all of the muslin cloth that existed. He would have a population of breaking strengths if he were to take every possible swatch of cloth of a particular size and then measured a breaking strength from each swatch. The experimenter desires an estimate of the mean or average value of this population. It is possible for him to obtain an estimate by taking a sample of say n measurements. The computation would be to sum the sample values and divide by n.

He would like to know if the calculated sample mean $\bar{X}$ is reasonably close to the population mean μ. That is, he desires to say with a specified probability that a certain interval will include μ. For example, suppose the following 95% confidence interval had been computed: $45 < \mu < 55$. The probability that the mean breaking strength is between 45 and 55 lbs is equal to 95%. This probability statement means that if the experimenter selects many samples and sets a 95% confidence interval on μ for each sample, then 95% of those intervals will contain μ. We will denote the probability as the confidence coefficient $1 - \alpha$. The interval is informative but if too long might be useless. If this is the case, generally the best way to shorten the interval is to take a larger sample. To do this we would like to know how large a sample to take to find a specified interval width. Consequently we will reformulate the problem. The experimenter desires to take a minimum sample size so that the average length of the confidence interval will be less than d per cent of the true mean in length. The experimenter also wants to work with confidence coefficients other than 95% (e.g. 90% or 99%). Thus the problem can be stated as follows: How large a sample is necessary in order to place a $1 - \alpha$ confidence interval on μ such that the average length of the intervals will be less than d% of μ? "The average length of the intervals will be less than d% of μ" means that if the experimenter selects many samples;

sets a $1 - \alpha$ interval on μ for each sample; then the average width of all these intervals will be less than $d\mu/100$.

Another textile experimenter may feel that knowing the average interval is going to be less than or equal to $\frac{d\mu}{100}$ is not adequate. He feels that often the computed interval might be wider than he desires. Therefore, he would like to have an interval such that the probability that the interval width is less than a specified d is equal to β where β is a given per cent (e.g. 90% or 99%).

We will discuss both methods.

Method I: Average length of confidence intervals less that d per cent of μ. This requires the experimenter to know the coefficient of variation, g (g is defined as the population standard deviation divided by the mean.) If the experimenter has an estimate of what interval will include g, the method of determining n will prove to be quite satisfactory. The experimenter specifies g, d, and $1 - \alpha$. A chart is given for three values of $1 - \alpha$. The procedure is to compute g/d and read n from the approximate curve. If he believes g to be in a certain interval he can determine a range of n values and select an average n.

As an example, we will show how this method applies to a textile experiment. An experimenter desires to know the mean breaking strength (in pounds) of muslin cloth after it has been washed thoroughly and then dried in the sun for a period of four hours. After cutting a bolt of cloth into say n swatches, he washes and dries

the cloth and then takes a measurement of breaking strength (in pounds) from each swatch or sample. His problem is: How many samples (swatches) should be taken in order to place a 95% confidence interval on the mean breaking strength such that the average length of the intervals will be less than 4, 6, 8, or 10 per cent of μ.

The experimenter specifies $g = .012$. This value was obtained from an earlier experiment. By use of the graph, we can set up the following table:

VALUES OF n GIVEN g/d AND α

d	g/d	α 90%	95%	99%
.04	.30	3	3	6
.06	.20	2	2	4
.08	.15	2	2	3
.10	.12	2	2	2

Thus at the 95% level the experimenter would select either 2 or 3 samples depending upon the interval width desired. If he wanted to work at the 99% level he would select n equal to 2, 3, 4, or 6. His estimate of g could be wrong by as much as 10% with little effect on n.

Method II. The probability that the $1 - \alpha$ confidence interval width is less than a specified d is equal to β where β is a given per cent (e.g. 90% or 95%). This requires that the experimenter has an estimate of σ^2. We will call this value g^2. We will find n by using the relationship

$$\frac{4\chi^2_{1-\beta}\, t^2_{\alpha}\, g^2}{n\,(n-1)} = d^2$$

where t is the appropriate value of Students' distribution with n - 1 degrees of freedom, χ^2 is the appropriate value of the chi-square distribution with n - 1 degrees of freedom, and d and g are specified.

For example, suppose the textile researcher would like to have the probability that the 95% confidence interval width less than 6 equal to 99%. By a previous experiment is found an unbiased estimate of σ^2 to be 0.216. Then if n = 2

$$\frac{(.86)\,(9.21)\,(4.303)^2}{(2)\,(1)} > 36$$

and if n = 3

$$\frac{(.86)\,(11.3)\,(3.182)^2}{(3)\,(2)} < 36.$$

Thus, n = 3 since 3 is the largest integer such that satisfies the inequality.

If the experimenter would prefer 1 - α = .99 then if n = 3

$$\frac{(.86)\,(9.21)(9.925)^2}{(3)\,(2)} > 36$$

and if n = 4

$$\frac{(.86)\,(11.3)\,(5.841)^2}{(4)\,(3)} < 36.$$

Thus n = 4 since 4 is the largest integer that satisfies the inequality.

The major disadvantages in Method II are that the value of n is rather difficult to find and the value of g^2 could cause inaccurate results because it is an estimate of σ^2. In conclusion, it should be noted that n is always rounded to the largest integer so that the given expression is the maximum value less than or equal to d.

SUMMARY

This thesis is a collection of methods available to determine a minimum sample size that will enable the experimenter to have a desired confidence in the results. The main topics included finding a sample size when the variance is known, unknown, and approximated.

Further work that may be done along these lines include

(1) tables to find n when the probability that w is less than d equals a specified β,

(2) finding the sample size when sampling from other than a normal distribution,

(3) determining the sample size in Non-parametric Statistics,

(4) use of a two step method when the expected value of w is less than d,

(5) graphs of sample size curves under specified conditions (e.g. when the variance is known).

BIBLIOGRAPHY

(1) Cochran, Wm. Sampling Techniques. New York: John Wiley and Sons, Inc., 1953.

(2) Mood, Alexander. Introduction to the Theory of Statistics. New York: McGraw-Hill Book Company, Inc., 1950.

(3) Seelbinder, B. M. "On Stein's Two Stage Sampling Scheme." Annals of Mathematical Statistics, (December, 1953) 24, 640-647.

(4) Stein, Charles. "A Two Sample Test for a Linear Hypothesis Whose Power is Independent of the Variance." Annals of Mathematical Statistics, (June, 1945) 16, 243-258.

VITA

Aaron Sampson Goldman
candidate for the degree of
Master of Science

Thesis: DETERMINATION OF SAMPLE SIZE

Major: Mathematics

Minor: Statistics

Biographical and Other Items:

Born: February 8, 1932 at Red Lion, Pennsylvania

Undergraduate Study: O.A.M.C., 1950-1954

Graduate Study: V.P.I., 1954
O.S.U., 1956-1958

Experiences: U. S. Air Force, 1607th ATW(H) at Dover AFB, Dover Delaware, 1954-56; Statistician, Los Alamos Scientific Laboratory, Los Alamos, New Mexico, 1955; Graduate Assistant in Mathematics Department 1954-55, 1956-57.

Member of Pi Mu Epsilon

Date of Final Examination: May, 1958