SOME CONSIDERATIONS OF AN OPTIMUM SAMPLE SIZE FOR A ONE-STAGE SAMPLING PROCEDURE

by

Daniel Zakich

Thesis submitted to the Graduate Faculty of the
Virginia Polytechnic Institute
in candidacy for the degree of

MASTER OF SCIENCE
in
STATISTICS

APPROVED:	APPROVED:
Louis A. Pardue Director of Graduate Studies	Boyd Harshbarger Head of Department
E. Burke Johnston Dean of Applied Science and Business Administration	Paul N. Somerville Major Professor

August 1954

Blacksburg, Virginia

LD
5655
V855
1954
Z344
c.2

TABLE OF CONTENTS

I. INTRODUCTION

Many cases arise when a manufacturer wishes to make a choice between two methods of production, about whose relative merits little or nothing is known, but some decision as to which method to accept for future production must be made. If the manufacturer is not interested in his loss due to experimentation, then the procedure will be to manufacture n items using each method, then manufacture the required amount N, using that method proved the better on the initial 2n. The optimum sample size for this problem has been discussed by Somerville (1954).

Suppose instead the manufacturer is interested in his loss due to experimentation and the sample size must come from the total amount N, to be produced. Then the procedure will be to manufacture n items using each method, producing the remaining (N-2n) items using that method proved the better on the initial 2n.

This thesis will deal with some considerations of applying a minimax solution for the optimum sample size in the latter case, taking into account the amount to be produced, the cost of sampling and the cost of making the wrong decision.

II. THE LOSS FUNCTION

If there is no reason to believe that one method will give a greater yield than the other, it is reasonable that an equal number, say n, should be produced using each method. Further it is reasonable to accept that method giving the greater yield in the preliminary sample for further production. This has been justified by Badahur (1950).

Let the yields from the two methods be normally distributed with unknown means $\mu_o \geq \mu_1$, respectively, and a common variance σ^2, known at least approximately. Let the sample means from the two methods (populations) be $\bar{x}_o$ and $\bar{x}_1$, respectively. Define π_o to be the population (unknown) with the larger mean. Let the cost of the preliminary sample be $C(n) = c_1 n + c_o$, where c_1 is the cost of sampling one unit and c_o the cost of setting up the experiment. Then the total expected loss may be expressed as

$$L = (N - 2n)(\mu_o - \mu_1)\, p_1 + n(\mu_o - \mu_1) + C(n), \qquad (2.1)$$

where $p_1 = P\left\{\text{choosing } \pi_1\right\} = P\left\{\bar{x}_o < \bar{x}_1\right\}$.

Now $P\left\{\bar{x}_o < \bar{x}_1\right\} = P\left\{\bar{x}_o - \bar{x}_1 < 0\right\}$

$$= P\left\{\frac{(\bar{x}_o - \bar{x}_1)-(\mu_o - \mu_1)}{\sqrt{2\sigma^2/n}} < -\frac{\mu_o - \mu_1}{\sqrt{2\sigma^2/n}}\right\}.$$

Set $$\gamma = \frac{\mu_o - \mu_1}{\sqrt{2\sigma^2/n}}, \qquad (2.2)$$

then the expected loss may be written as

$$L = (N-2n)\,\sigma\sqrt{\frac{2}{n}}\,\gamma\int_{-\infty}^{-\gamma} N(0,1)dx + \sigma\gamma\sqrt{2n} + C(n), \quad (2.3)$$

where $\int_{-\infty}^{-\gamma} N(0,1)dx$ is the standard normal integral.

2.1 Maximum and Minimum.

To find the critical values of γ, we solve for γ in $\frac{\partial L}{\partial \gamma} = 0$.

Hence,

$$\frac{\partial L}{\partial \gamma} = \frac{-\gamma e^{-\frac{\gamma^2}{2}}}{\sqrt{2\pi}} + \int_{-\infty}^{-\gamma} N(0,1)dx + \frac{n}{N-2n} = 0,$$

or

$$\frac{\gamma e^{-\frac{\gamma^2}{2}}}{\sqrt{2\pi}} - \int_{-\infty}^{-\gamma} N(0,1)dx - \frac{n}{N-2n} = 0. \quad (2.4)$$

Now, it is obvious from (2.4) that no absolute maximum or minimum exists independently of n and N. However, for any given values of γ, it is possible to obtain the value of the ratio $\frac{n}{N-2n}$ which corresponds to a critical value of L. A few of these values are tabulated in Table 1. It is seen that in order for L to have a maximum or minimum, then $\frac{n}{N-2n}$ must be less than .1289 and that $\gamma = \sqrt{2}$ corresponds to this value.

Table 1

Maximizing and Minimizing Values of γ

γ	$\frac{n}{N-2n}$
.80	.0199
.90	.0554
1.00	.0833
1.10	.1040
1.20	.1179
1.30	.1260
1.40	.1287
$\sqrt{2}$	.1289
1.50	.1275
1.60	.1227
1.70	.1153
1.80	.1062
1.90	.0960
2.00	.0852
2.10	.0745

2.2 Inflection Point.

To find the inflection point, we solve for γ in $\frac{\partial^2 L}{\partial\gamma^2} = 0$. Hence,

$$\frac{\partial^2 L}{\partial\gamma^2} = \frac{\gamma^2 e^{-\frac{\gamma^2}{2}}}{\sqrt{2\pi}} - \frac{2e^{-\frac{\gamma^2}{2}}}{\sqrt{2\pi}} = 0,$$

or

$$e^{-\frac{\gamma^2}{2}} [\gamma^2 - 2] = 0,$$

which gives $\gamma = \sqrt{2}$ as the inflection point.

Now, from Table 1, when $\gamma = \sqrt{2}$, $\frac{n}{N-2n} = .1289$, which gives $n = .1025$ N. Thus, from the preceding discussion, when $n < .1025$ N, $\frac{n}{N-2n} < .1289$ and L will have a relative maximum; when $n > .1025$ N, then $\frac{n}{N-2n} > .1289$ and L will be monotone increasing.

To determine whether we have a maximizing or minimizing value of γ, when $\frac{n}{N-2n} < .1289$, we may use the following rule:

$\gamma < \sqrt{2}$, a maximizing value,

$\gamma > \sqrt{2}$, a minimizing value.

This rule is justified since

$$\frac{\partial^2 L}{\partial \gamma^2} < 0 \text{ for } \gamma < \sqrt{2}$$

and

$$\frac{\partial^2 L}{\partial \gamma^2} > 0 \text{ for } \gamma > \sqrt{2}.$$

III. SPECIAL CASES

3.1 $\underline{\gamma < \sqrt{2}}$.

If some a priori assumption is made that $\gamma < \sqrt{2}$, then a minimax solution will be shown to exist.

Considerable calculation indicates that by taking $\gamma = 1$ as the unique maximizing value, then the expected loss, L, departs little from the actual loss for the true γ. Hence assuming $\gamma = 1$, then

$$L_{max} = (N-2n)\,\sigma\sqrt{\frac{2}{n}}\int_{-\infty}^{-1} N(0,1)dx + \sigma\sqrt{2n} + c_1 n + c_o,$$

$$= .1587\, N\sigma\sqrt{2}\, n^{-\frac{1}{2}} + .6826\,\sigma\sqrt{2}\, n^{\frac{1}{2}} + c_1 n + c_o.$$

Now,

$$\frac{\partial L_{max}}{\partial n} = n^{-3/2} - \frac{4.3012}{N} n^{-1/2} - \frac{8.9112c_1}{\sigma N} = 0, \qquad (3.1)$$

which upon solution for n yields

$$n_{opt} = \sigma^{2/3} N^{2/3}\left[\sqrt[3]{d_1c_1+\sqrt{d_2c_1^2-d_3\frac{\sigma^2}{N}}} + \sqrt[3]{d_1c_1-\sqrt{d_2c_1^2-d_3\frac{\sigma^2}{N}}}\right]^{-2} \qquad (3.2)$$

where $d_1 = 4.4556$,

$d_2 = 19.8524$,

and $d_3 = 2.9472$.

The proof is as follows:

In (3.1), let $y = n^{-1/2}$,

$$a = \frac{-4.3012}{N},$$

and $b = \dfrac{-8.9112c_1}{\sigma N}$;

then we have $y^3 + ay + b = 0.$ (3.3)

Now, if $\frac{(8.9112)^2 c_1^2}{4\sigma^2 N^2} > \frac{(4.3012)^3}{27N^3}$, the equation will possess only one real root. This will be the case since in most practical considerations, N will be large compared to σ and c_1.

Cardan's solution to (3.3), gives

$$y = A + B,$$

where

$$A = \frac{1}{\sigma^{1/3} N^{1/3}} \sqrt[3]{d_1 c_1 + \sqrt{d_2 c_1^2 - d_3 \frac{\sigma^2}{N}}},$$

$$B = \frac{1}{\sigma^{1/3} N^{1/3}} \sqrt[3]{d_1 c_1 - \sqrt{d_2 c_1^2 - d_3 \frac{\sigma^2}{N}}},$$

and $d_1 = 4.4556,$

$d_2 = 19.8524,$

$d_3 = 2.9472.$

Hence,

$$n = y^{-2} = \sigma^{2/3} N^{2/3} \left[\sqrt[3]{d_1 c_1 + \sqrt{d_2 c_1^2 - d_3 \frac{\sigma^2}{N}}} + \sqrt[3]{d_1 c_1 - \sqrt{d_2 c_1^2 - d_3 \frac{\sigma^2}{N}}} \right]^{-2}.$$

If the ratio $\frac{\sigma^2}{N}$ is very small then, since $d_1 = \sqrt{d_2}$, (3.2) reduces to

$$n_{opt} \approx .2326\, \sigma^{2/3} N^{2/3} c_1^{-2/3} \qquad (3.4)$$

3.2 $\underline{\gamma \geq \sqrt{2}}$.

For $\gamma \geq \sqrt{2}$, the loss function will possess two values of γ, corresponding to a relative maximum and a "largest value". A general solution of this case has been attempted but with no satisfactory results. However, it is easily seen that if γ is sufficiently large, the loss function becomes

$$L' = n(\mu_o - \mu_1) + c_1 n + c_o \qquad (3.5)$$

since $\gamma \int_{-\infty}^{-\gamma} N(0,1)dx$ approaches 0.

Now, (3.5) is a linear function of n and hence L' is monotone increasing, and to minimize the maximum loss, we should not sample at all. This is an unsatisfactory result and indicates that if the manufacturer is not willing to assume that $\frac{\mu_o - \mu_1}{\sigma}$ is small, then a minimax solution is not reasonable and thus no satisfactory one-stage solution seems to exist.

Perhaps a two-stage sampling procedure may give a reasonable solution but no work on this possibility has been attempted. Another possibility is assigning a distribution to the difference $\mu_o - \mu_1$ and attempting to resolve the problem by minimizing an average loss function.

3.3 No Loss Due to Sampling.

Suppose the manufacturer is not concerned with the loss from sampling, then the expected loss will be

$$L^* = (N-2n)\ \sigma\sqrt{\frac{2}{n}}\ \ \gamma \int_{-\infty}^{-\gamma} N(0,1)dx + c_1 n + c_o.$$

It can be easily shown that

$$L^*_{Max} = .1700\ (N-2n)\ \sigma\sqrt{\frac{2}{n}} + c_1 n + c_o$$

for all possible values of μ_o and μ_1.

Now,

$$\frac{\partial L^*_{Max}}{\partial n} = -\frac{.1700}{\sqrt{2}} N\sigma n^{-3/2} - \frac{.3400}{\sqrt{2}} \sigma n^{-1/2} + c_1 = 0,$$

or

$$n^{-3/2} + \frac{2}{N} n^{-1/2} - \frac{8.3188 c_1}{\sigma N} = 0, \tag{3.7}$$

which has, as its only real root,

$$n_{opt} = \sigma^{2/3} N^{2/3} \left[\sqrt[3]{g_1 c_1 + \sqrt{g_2 c_1^2 + g_3 \frac{\sigma^2}{N}}} + \sqrt[3]{g_1 c_1 - \sqrt{g_2 c_1^2 + g_3 \frac{\sigma^2}{N}}} \right]^{-2}, \tag{3.8}$$

where $g_1 = 4.1594$,

$g_2 = 17.3006$,

and $g_3 = .2963$.

The proof is similar to that in Section 3.1, by letting

$$y = n^{-1/2},$$

$$a = \frac{2}{N},$$

and $$b = \frac{-8.3188 c_1}{\sigma N}.$$

Then (3.7) becomes

$$y^3 + ay + b = 0.$$

Now, $\frac{(-8.3188)^2 c_1^2}{4\sigma^2 N^2} + \frac{8}{27N^3} > 0$, hence the equation possesses only one real root.

Following Cardan's solution, as indicated previously, (3.8) is found to be the solution.

Now, again, if $\frac{\sigma^2}{N}$ is very small, then, since $g_1 = \sqrt{g_2}$, (3.8) reduces to

$$n_{opt} \approx .2435\ \sigma^{2/3}\ N^{2/3}\ c_1^{-2/3}. \qquad (3.9)$$

This result, (3.9), is the solution discussed by Somerville (1954), in the case where sampling is not from the amount to be produced.

IV. DISCUSSION OF THE SOLUTION

The preceding theory discusses the case when γ is known, at least approximately, to be within certain ranges. Since γ itself is a function of the sample size, it would be difficult, if not impossible, for the manufacturer to always ascertain within what range γ will lie. However, it is conceivable that he may have some idea of what the difference $\mu_o - \mu_1$ is expected to be.

When $\gamma < \sqrt{2}$, then this implies that the difference $\mu_o - \mu_1$ will be small. Hence for small differences, it seems feasible to use the solution for the case $\gamma < \sqrt{2}$. Also when γ is large, it would seem to imply that the difference $\mu_o - \mu_1$ would be large. Hence, for large differences, the case $\gamma \geq \sqrt{2}$ would seem to be applicable. Just how "small" or how "large" the difference $\mu_o - \mu_1$ must be, cannot be resolved very easily.

It seems practical to assume for a given "small" difference of $\mu_o - \mu_1$ that $\gamma < \sqrt{2}$ and proceed to find the optimum sample size under this assumption. Then by substituting the sample size obtained, in the equation for γ, (2.2), the assumption can be validated or refuted. If upon substitution, the result $\gamma \geq \sqrt{2}$ is found, then some other method of solution must be attempted.

Consider an elementary example. A manufacturer concerned with the production of radiation shields has two processes submitted for producing these shields. He is willing to assume that the mean difference in months, before the shields become radioactive, does not exceed .5 months. If he desires to produce a total of 100 shields and is confident that $\sigma \approx 1$, then he would try

$$n_{opt} = .2326\ N^{2/3}\ \sigma^{2/3}\ c_1^{-2/3}$$

which upon substitution would give

$$n_{opt} = 5.01\ c_1^{-2/3}.$$

Now, the cost of sampling must be expressed in the same units as is the loss expected. Hence the manufacturer must convert the monetary sampling cost, which he is assumed to know, in terms of months. Suppose, in this case, $c_1 = 1$ (month). Then his optimum sample would be 5 shields.

Now, when $n = 5$, then

$$\begin{aligned}\gamma &= \frac{.5}{\sqrt{2/5}}\\ &= .7906,\end{aligned}$$

hence the assumption $\gamma < \sqrt{2}$ is valid.

V. SUMMARY

The purpose of this work is to discover an optimum sample size to be used for deciding between two methods (populations) to choose for future production. The procedure involves the formulation of a loss function, expressing the expected loss due to choosing the population with the smaller mean, as a function of the difference between the population means, the amount to be produced and the cost of sampling. A minimax procedure is applied to obtain the optimum sample size.

Since the function does not lend itself conveniently to mathematical considerations, special cases involving the difference between the means are considered and an optimum sample size is found for these cases. In all cases, the optimum sample size is an explicit function of the amount to be produced, the cost of sampling and the standard deviation.

VI. ACKNOWLEDGEMENTS

The author wishes to express his sincerest appreciation to Professor P. N. Somerville for his guidance and his many ideas and suggestions which have made this paper possible. Appreciation is also expressed for Professor Boyd Harshbarger's encouragement and criticisms.

The author is also indebted to Mrs. Marianne Byrd for preparing the final copies for presentation.

VII. REFERENCES

Badahur, R. R., "On a Problem in the Theory of k Populations." Annals of Math. Stat., Vol. XXI, p. 362-375, 1950.

Federal Works Agency, Tables of Probability Functions. Vol. II, Mathematical Tables Project, New York, 1942.

Hall, H. S. and Knight, S. R., Higher Algebra. MacMillan and Co., Ltd., 4th Edition, 1936.

Somerville, P. N., "Some Problems in Optimum Sampling." To be published in Biometrika, 1954.

Wald, A., Statistical Decision Functions. John Wiley and Sons, New York, 1950.

VIII. VITA

Son of John and Dana Zakich, born January 4, 1928 at Akron, Ohio. Graduated from South High School, Akron, Ohio in June, 1946 and entered the U. S. Army, being discharged July, 1947.

Entered the University of Akron in 1948, receiving the Bachelor of Science degree in Mathematics in June, 1952, and became a candidate for the Master of Science degree in Statistics at Virginia Polytechnic Institute in September, 1952. During an absence of 12 months, acted as Development Engineer in Quality Control with the Firestone Tire and Rubber Company, Akron, Ohio, returning to Virginia Polytechnic Institute in January, 1954 as a teaching fellow.

A member of Phi Sigma Kappa, Omicron Delta Kappa, VPI Mathematics Club, associate member of Sigma Xi and listed in Who's Who in American Colleges and Universities.

Daniel Zakich