Determination of Sample Size

Prakash Dixit

Department of Statistics, Modern College of Arts Science and Commerce, Shivajinagar, Pune 5.

Abstract : *One of the crucial questions in researchers mind is how big my sample should be. There are several methods of estimation of sample size. We classify them in two categories. The first one is the forecasting, estimation or prediction and the second is the testing of hypotheses or comparison. In each case the results are reliable or precise if sampling is proper and the size of sample is adequate. We give the formulae to determine the size of sample in each of the situations which will be of immense use to the researchers.*

Keywords : *Sample, population, margin of error, confidence coefficient, hypothesis, level of significance, power of test, ANOVA, normal distribution, control chart, degrees of freedom,chi-square test, contingency table.*

I. INTRODUCTION

Many researchers conduct survey using sampling techniques. One of the important questions they face is the size of the sample. The answer to the question is not so simple or unique. The complexity of the answer is mainly due to the lack of knowledge of population or universe (Population is the group of objects under study). The researchers are always criticized that the size of sample is inadequate. There are two misunderstandings about the sample size. The first one is, larger the sample size better are the results. The second one is, the sample size should be at least 10% of the population. However the results are reliable if the sample is proper representative of population rather than it is unexpectedly larger in size. Merely increasing the size of sample, we do not achieve the precision. It does not increase proportionately as the sample size increases. The sample should be large enough to include all the important facts about the population. It should be the cross section of the population. On the other hand it should take into account the constraints of money, time, manpower. Increasing the sample size beyond certain limit does not increase substantial accuracy, however it may be marginal. Hence the sample size should be optimal.

The point is illustrated with the help of following examples. Suppose we want to test blood sugar level of an individual. We test only 2 or 3 ml of blood out of the entire blood in the body. Such a small sample also yields reliable results. The results cannot be improved just by testing large amount of blood. In this situation small sample is sufficient because the population (the entire blood in the body) is completely homogeneous.

Suppose we want to estimate objectively the number of legs of cow. Suppose the population of cow in India is in some crores. A sample of size one is enough in this case. Sample of large size will not yield more accuracy.

In estimation problems the sample size is based on two quantities, one is the margin of error (the gap between true value and the estimate). We denote it by 'e'. The second is the confidence coefficient (the percentage of cases where the gap is less than the margin of error) .We denote it by (1-α). Then the equation given below determines the sample size as a **Rule of Thumb**.

$$E = z_{\alpha/2}.(\text{Standard error of the estimator}) \quad (1)$$

Where $z_{\alpha/2}$ is value of standard normal variable X such that, $P(X > z_{\alpha/2}) = \alpha/2$.

Situation 1. Estimation of mean :

Suppose we want to estimate the mean of the population under study, using the sample mean, the above equation reduces to

E = $z_{\alpha/2}$ (Standard error of the sample mean)

$e=z_{\alpha/2}\sigma/\sqrt{n}$ Which gives $n = \frac{(z_{\alpha/2}\sigma)^2}{e^2}$ (2)

Illustration : Suppose $\alpha = 0.95$, e=10, σ=50. Hence $z_{\alpha/2} = 1.96$.

$n= (1.96)^2 (50)^2/(10)^2=96.04$

The equation (2) assumes that the population is infinite, however it is not so always. The formula is modified using finite population correction as follows, assuming N as the population size.

$$n = \frac{n_0}{1+\frac{n_0}{N}} \quad \text{Where} \quad n_0 = \frac{(z_{\alpha/2}\sigma)^2}{e^2} \text{ as}$$

the first approximation.

If N=2000 then

$$n = \frac{n_0}{1+\frac{n_0}{N}} = \frac{96.04}{1+\frac{96.04}{2000}} = 91.64$$

N	n
100	48.99
500	80.56
1000	87.62
2000	91.64
4000	93.79

The above table shows that sample size does not increase proportionately as N increases.

Situation 2. Estimation of proportion:

To estimate the proportion of certain objects in the population under study using the sample proportion, the above equation (1) reduces to

e= $z_{\alpha/2}$. (Standard error of the sample proportion)

$$e = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Which gives $n = \frac{z^2_{\alpha/2}\,p(1-p)}{e^2}$ (3)

Illustration : Suppose $\alpha = 0.95$, $p = 0.1$, e = 0.07 . Hence $z_{\alpha/2} = 1.96$ and

q = 1- p = 0.9.Using (3) we get

$$n = \frac{1.96^2.(0.1)(0.9)}{0.07^2} = 70.56$$

The equation (3) assumes that the population is infinite, however if it is not then the formula is modified using finite population correction as follows, assuming N as the population size.

$$n = \frac{n_0}{1+\frac{n_0-1}{N}} \text{ Where } \quad n_0 = \frac{z^2_{\alpha/2}\,p(1-p)}{e^2} \text{ as}$$

the first approximation .

If N=500 then

$$n = \frac{n_0}{1+\frac{n_0-1}{N}} = \frac{70.56}{1+\frac{69.56}{500}} = 61.94$$

Note : If the value of p is unknown then we assume that it is 0.5 which gives the maximum sample size

$$n = \frac{z^2_{\alpha/2}}{4e^2}$$

Situation 3. Estimation of proportion of rare events :

Suppose the proportion of certain objects is very small, in such cases sample size should be enough large so that it gives a chance to include such cases in sample. For example an occurrence of twin births at a certain place is 1%. If the sample size is 10 then it may not include such case. So we assume in this case the number of twin births (X) follows Poisson probability distribution. We find a sample of size *n* so that it contains at least one twin in 90% samples.

It gives an equation to find n as follows

$P(X>1) > 0.9$ or $P(x=0) < 0.1$ Suppose X follows Poisson distribution with mean $\lambda = np$. Here the occurrence of twins is 1%, hence $p = 0.01$.

We get $P(X=0) = e^{-\lambda} = e^{-(0.01)n} < 0.1$

$-(0.01)n < \log_e(0.1)$. Hence, $n > 230$

Thus a sample of size 230 will guarantee that it will contain at least one twin in 90% samples.

Situation 4. Estimation of sample size for testing population mean :

Researchers require to test the hypothesis $H_0: \mu = \mu_0$ against $H_1: \mu = \mu_1 = \mu_0 + \Delta$. In this situation, we need to find the required sample size so that test will have given level of significance (α) and the power of test ($1-\beta$). Assuming that the characteristic under study follows normal distribution with mean μ and the standard deviation σ. The basic equation about the test statistic Z involving the sample size n is

$$P\left(Z > -Z_{\alpha/2} + \frac{|\Delta|\sqrt{n}}{\sigma}\right) = 1-\beta \qquad (4)$$

Where Z is assumed to follow standard normal distribution. The relation (4) gives after simplification

$$n \geq \frac{[z_{\alpha/2} + |Z_{1-\beta}|]^2 \sigma^2}{\Delta^2} \qquad (5)$$

Illustration: Suppose we have to find the sample size to test $H_0: \mu = 50$ against

$H_1: \mu = 52$ for a normal population with standard deviation 5. It is given that the level of significance is 5% and the power of test is 80%.

Here we have $\alpha = 0.05$ hence $z_{\alpha/2} = 1.96$, $1-\beta = 0.8$ gives $Z_{1-\beta} = -0.8416$, $\sigma = 5$,

$\Delta = \mu_1 - \mu_0 = 2$, hence we get

$$n \geq \frac{[1.96 + |-0.8416|]^2 5^2}{2^2} = 49.056 \approx 49$$

Note: If σ is unknown we use $t_{\alpha/2}$ and $t_{1-\beta}$ with (n-1) degrees of freedom in place of $z_{\alpha/2}$ and $Z_{1-\beta}$ respectively for computing n using (5) as the first approximation.

Situation 5. Estimation of sample size for testing equality of two population means :

Suppose we need to test $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 - \mu_2 = \Delta$. Let the sample size of group1 and group 2 be n_1 and n_2 respectively. We denote $r = n_1/n_2$ and $n = n_1 + n_2$. The expression for n used in this case with little modification is as follows.

$$n \geq \left(\frac{r+1}{r}\right)\frac{[z_{\alpha/2} + |Z_{1-\beta}|]^2 \sigma^2}{\Delta^2} \qquad (6)$$

If we have $\alpha = 0.05$ hence $z_{\alpha/2} = 1.96$, $1-\beta = 0.8$ gives $Z_{1-\beta} = -0.8416$, $\sigma = 10$,

$\Delta = \mu_1 - \mu_0 = 3$, $r = n_1/n_2 = 1.5$, hence we get

$$n \geq \frac{[1.96 + |-0.8416|]^2 10^2}{3^2} = 130.82 \approx 131$$

We divide 131in 3:2, (Since r = 1.5), it gives $n_1 = 87$ and $n_2 = 44$.

Note: A similar expression for $n_1$and n_2 can be obtained for testing equality of population propotions ($H_0: P_1 = P_2$ against $H_1: P_1 - P_2 = \Delta$)

$$n \geq \left(\frac{r+1}{r}\right)\frac{[z_{\alpha/2} + |Z_{1-\beta}|]^2 P(1-P)}{\Delta^2}$$

Where $P = (rP_1 + P_2)/(r+1)$

In this case we need to ensure whether for the n_1 and n_2 obtained justifies that the test statistic Z follows normal distribution .

Situation 6 : Mead's resource equation

In analysis of variance (ANOVA) the sample size should be large enough so that the degrees of freedom associated with error is at least 12, thus it gives

n > (number of treatments) +
(number of blocks) +12

Situation 7 : Chi-square test of goodness of fit

Sample size should be large enough so that expected frequency of every cell is greater than 5.

Situation 8 : Chi-square test of independence of attributes.

In testing the independence of two attributes we use mxn contingency table, the sample size should be large enough to have each cell frequency larger than 5.

Situation 9 : Control chart for the number of defects

In the construction of control chart for number of defects, we use C chart. The sample size should be so large to have lower control limit (LCL) positive.

$$LCL = \lambda - 3\sqrt{\frac{\lambda}{n}} > 0 \quad \text{gives} \quad n > 9/\lambda$$

Situation 9 : Control chart for the fraction defectives.

In the construction of control chart for the fraction defectives we use P chart. The sample size should be so large to have lower control limit (LCL) positive.

$$LCL = p - 3\sqrt{\frac{p(1-p)}{n}} > 0 \text{ gives } n > 9(1-p)/p$$

If p = 0.3 then we require
n>(9) (0.7)/0.3 = 21.

Situation 10 : Control chart for the fraction defectives and catching of shift.

Suppose we want to construct control chart for the fraction defectives (P chart) so that the probability of catching the shift of amount d in fraction defective within the first sample is at least 0.5 .It gives

$$P\left(Z > 3 - \frac{d\sqrt{n}}{\sqrt{P(1-P)}}\right) \geq 0.5$$

It is possible if $n > 9P(1-P)/d^2$.

Situation 11: Allocation problem in stratified sampling:

Under the stratified sampling we need to allocate the fraction of the total sample size n to different strata. If the i th stratum has the population size is N_i , the stratum standard deviation is S_i and the cost of sampling per unit is C_i then we get the i th stratum sample size (n_i) under the following allocation methods as

(a) **Proportional Allocation** :

$$n_i = n\frac{N_i}{N} \qquad Where \qquad N = \sum N_i$$

(b) **Neyman's allocation** : $n_i = n\frac{N_i S_i}{\sum N_i S_i}$

(c) **Optimum allocation** : Suppose the cost function is $C=C_0+ \Sigma n_i C_i$ then the sample size n_i so that the total cost of sampling is C_i is given by

$$n_i = (C - C_0)\frac{N_i S_i / \sqrt{C_i}}{\sum N_i S_i \sqrt{C_i}}$$

REFERENCES

[1] Chocran W. G. :Sampling Techniques ,second edition. Newyork John Wiley and sons Inc.

[2] Montgomery D.C.: Introduction to Statistical Quality Control. Sixth edition. Wiley India Private Limited.

[3] www.ykhoa.net

[4] www.surveysystem.com

[5] www.wikipedia.org

[6] www.statsdirect.com