DOI: 10.15415/mjis.2014.31005

On Sample Size Determination

K MURALIDHARAN

Department of Statistics, Faculty of Science Maharajah Sayajirao University of Baroda, Vadodara- 390 002, India.

Email: muralikustat@gmail.com

Received: December 2, 2013 | Revised: February 3, 2014 | Accepted: May 27, 2014

Published online: September 20, 2014

Abstract: One of the questions most frequently asked of a statistician is: How big should the sample be? Managers are anxious to obtain an answer to this fundamental question during the planning phase of the survey since it impacts directly on operational considerations such as the number of interviewers required. There is no magical solution and no perfect recipe for determining sample size. It is rather a process of compromise in which the precision requirements of the estimates are weighed against various operational constraints such as available budget, resources and time. In this article we revisit to the estimate of sample size for various project characteristics. Examples for each are supported numerically.

Keywords: Accuracy, Precision, Confidence interval, margin of error, mean, proportion, sampling error.

1. Introduction

Sample size refers to the number of observational units. A sample size can only be calculated for some specific aspect (value of interest) of the *population* (the entire set of units) to be estimated, usually related to some important response variable. The value of interest may be mean, proportion, variance, correlation etc. The procedure of drawing a sample from a population is called *random sampling* or simply *sampling*. Since sample forms the basics of all statistical decision making, it is essential to have a suitable sample size for any investigation. Such a sample size will depend on many factors like, population size, type of data (continuous or discrete), *precision* required, confidence level etc.

1.1. Accuracy and Precision

The size of sample that will require attaining a given precision for such an estimate depends on the variability of the population and on the extent to which it is possible to reduce the different components of this variability in the random sampling error. The measured value that has little deviation from the actual value is called *accuracy*. Accuracy is usually tested by comparing an average of repeated measurements to a known standard value for that unit. Mean, mode, median etc are all accuracy measures. The sta*ndard error* (standard deviation of the sample observations), a crude measure of the precision of an estimate obtained from a sample, is accurate enough to make the sample size calculation as it is a function of sample size.

Precision is how narrow you want the range to be for an estimate of a characteristic. For example, we quite often make the following statements about a process: the estimate of cycle time must be within 2 days, the estimate of the percent defective is within 3%, the estimate of the proportion should be within 5% of the total, the estimate the standard deviation can vary between ± 2 of the process mean etc. So what we mean by these statements is the amount of precision that can be allowed in the process specified according to the value of interest. Now consider the following statements: the 95% *confidence interval* (CI) of the cycle time is 40 ± 2 days (or the CI is (38, 42)), the 99% confidence interval of the sample proportion is 0.6 ± 0.15 (or the CI is (0.45, 7.5)), the 98% approximate confidence interval of the population variance is 12 ± 2.5 hrs (or the CI is (9.5, 14.5) etc. That is, we are specifying the allowable precision (or *margin of error*) admissible in each situation along with the characteristics of the value of interest. Note that, a confidence interval in general can be written as

$$\begin{aligned}\text{Confidence interval} &= \text{Accuracy measure} \pm \text{Precision} \\ &= \text{Point estimate} \pm \text{critical value*standard error}\end{aligned} \tag{1}$$

and therefore, precision is equal to half the width of a confidence interval. For example: a 95% confidence interval = (38, 42) for cycle time (in days) means we are 95% confident that the interval from 38 days to 42 days contains the average cycle time. Therefore, the width of the CI = 4 days and hence the precision is 2 days (i.e. the estimate is within ± 2 days). Thus if *UCL* and *LCL* are the upper confidence limit and lower confidence limit respectively, then

$$\text{Precision} = \frac{UCL - LCL}{2}, \tag{2}$$

Suppose the characteristics of interest is the process average, say μ, and if the process follow a normal distribution with known standard deviation, say σ, then the 95% confidence interval for μ is $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$, which is equivalent to $\bar{x} \pm k\frac{1}{\sqrt{n}}$, where $k = 1.96\sigma$. Similarly, the 95% confidence interval for population proportion p is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, which is equivalent to $\hat{p} \pm k_1 \frac{1}{\sqrt{n}}$, where $k_1 = 1.96\sqrt{\hat{p}(1-\hat{p})}$. Thus precision is a function of the sample size and standard error. Moreover, Precision (d) is inversely proportional to the square root of sample size n. That is

$$d \propto \frac{1}{\sqrt{n}} \tag{3}$$

Hence, the knowledge of precision can ease the calculation of sample size. Note that lesser the value of d more precise the estimator will be. So for good precision, we need to have a large sample size. As such, there is no clear-cut answer about how much precision you need; the answer depends on the business impact of using the estimate. Each situation is unique and should not be influenced by someone else's decision. But it is always possible to make a wild guess about precision. Also note that, to improve precision, you need to increase sample size (which incurs more cost). The converse may not be true.

The next important aspect of sample size calculation is to get some knowledge about the standard deviation. You need to have some idea of the amount of variation in the data because as the variability increases, the necessary sample size increases. There are many options for this:

- Find an existing data and calculate $\hat{\sigma} = s$ or
- Use a control chart (for individuals) from a similar process and get $\hat{\sigma} = s$ or
- Collect a small sample and calculate $\hat{\sigma} = s$ or
- Take an educated guess about s based on your process knowledge and memory of similar data

Below, we discuss the sample size computation for different characteristics and answer its statistical importance in estimation and inferential studies.

2. Sample Size When Characteristic of Interest is Mean

As seen above, if the characteristics of interest is the process mean, say μ, then the $(1-\alpha)100\%$ confidence interval for μ is $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. Here is called the significance level and $z_{\alpha/2}$ is the cut-off (or percentile) point of the area corresponds to $(\alpha/2)$ of the standard normal curve. Specifically, if the confidence level is sought for 95%, then $\alpha = 0.05$ and $z_{\alpha/2}$ = 1.96. Therefore,

$$d = 1.96\frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow n = \left(\frac{1.96\hat{\sigma}}{d}\right)^2 = \left(\frac{1.96s}{d}\right)^2 \tag{4}$$

The practitioners working on industrial applications generally consider the approximate value of the sample size is often considered as $n \cong \left(\frac{2\hat{\sigma}}{d}\right)^2 \cong \left(\frac{2s}{d}\right)^2$. Even in most of the six sigma applications, the black belts consider this simplified formula. This adjustment will only increase a few samples extra and will not harm the analysis in any way.

For many Biological, psychological and Social science research, the researcher's may look for more precision in their estimators as getting sample size many not be a difficult problem for them. In that case the precision may be improved upon by considering more level of standard errors. The formula derived in (4) is corresponding to a precision of one standard error of the mean. Similarly, the formula correspond to a precision of two standard errors (margin of error reduced by half) and three standard errors (margin of error reduced by one third) are respectively given by

$$\frac{d}{2} = 1.96\frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow n = \left(\frac{2x1.96\hat{\sigma}}{d}\right)^2 = \left(\frac{2x1.96s}{d}\right)^2 \tag{5}$$

and

$$\frac{d}{3} = 1.96\frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow n = \left(\frac{3x1.96\hat{\sigma}}{d}\right)^2 = \left(\frac{3x1.96s}{d}\right)^2 \tag{6}$$

Note that for more precise estimates, more sample size is required and vice versa. The sample size formula for different confidence level may be obtained similarly. The other value of α generally used in statistical research is α =

0.01 (with $z_{\alpha/2} = 2.58$) and $\alpha = 0.10$ (with $z_{\alpha/2} = 1.645$). These values may be inserted suitably in the formula discussed above as per the requirement. Tables 1-3 present the value of sample size according to various level of confidence, the estimate of sigma and precision levels (All entries are corresponds to the actual $z_{\alpha/2}$ values).

Table 1: Sample size for mean correspond to one-precision

Level of Confidence	*d*=1		*d*=2		*d*=4		*d*=6		*d*=10	
	s=5	s=10	s=5	s=10	s=5	s=10	s=5	s=10	s=5	s=10
90%	67	272	17	68	4	17	2	8	1	3
95%	96	384	24	96	6	24	3	11	1	4
98%	136	543	34	136	8	34	4	15	1	5
99%	166	666	42	166	10	42	5	18	2	7

Table 2: Sample size for mean correspond to two-precision

Level of Confidence	*d*=1		*d*=2		*d*=4		*d*=6		*d*=10	
	s=5	s=10	s=5	s=10	s=5	s=10	s=5	s=10	s=5	s=10
90%	269	1089	68	272	17	68	8	30	3	11
95%	384	1537	96	384	24	96	11	43	4	15
98%	543	2172	136	543	34	136	15	60	5	22
99%	666	2663	166	666	42	166	18	74	7	27

Table 3: Sample size for mean correspond to three-precision

Level of Confidence	*d*=1		*d*=2		*d*=4		*d*=6		*d*=10	
	s=5	*s*=10	*s*=5	*s*=10	*s*=5	*s*=10	*s*=5	*s*=10	*s*=5	*s*=10
90%	605	2450	153	613	38	153	17	68	6	25
95%	864	3457	216	864	54	216	24	96	9	35
98%	1222	4886	305	1222	76	305	34	136	12	49
99%	1498	5991	374	1498	94	374	42	166	15	60

3. Sample Size When Characteristic of Interest is Proportion

As seen above, if the characteristics of interest is the population proportion p, then the 95% confidence interval for p is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Therefore,

$$d = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$\Rightarrow n = \frac{(1.96)^2\,\hat{p}(1-\hat{p})}{d^2} \tag{7}$$

Similarly, the formula correspond to a precision of two (margin of error reduced by half) and three standard errors (margin of error reduced by one third) are respectively obtained as

$$n = \frac{(2x1.96)^2\,\hat{p}(1-\hat{p})}{d^2} \tag{8}$$

and

$$n = \frac{(3x1.96)^2\,\hat{p}(1-\hat{p})}{d^2} \tag{9}$$

The sample size formula for proportion for different confidence level may be obtained similarly. Tables 4-6 presents the value of sample size for estimating proportion according to various level of confidence, the estimate of proportion and precision levels (All entries are correspond to the actual $z_{\alpha/2}$ values).

Table 4: Sample size for proportion correspond to one-precision

Level of Confidence	*d*=5%			*d*=10%			*d*=20%			*d*=30%		
	p=0.2	p=0.5	p=0.7	p=0.2	p=0.5	p=0.7	p=0.2	p=0.5	p=0.7	p=0.2	p=0.5	p=0.7
90%	172	269	226	43	67	56	11	17	14	5	7	6
95%	246	384	323	61	96	81	15	24	20	7	11	9
98%	347	543	456	87	136	114	22	34	29	10	15	13
99%	426	666	559	107	166	140	27	42	35	12	18	16

Table 5: Sample size for proportion correspond to two-precision

Level of Confidence	d=5%			d=10%			d=20%			d=30%		
	p=0.2	p=0.5	p=0.7	p=0.2	p=0.5	p=0.7	p=0.2	p=0.5	p=0.7	p=0.2	p=0.5	p=0.7
90%	689	1076	904	172	269	226	43	67	56	19	30	25
95%	983	1537	1291	246	384	323	61	96	81	27	43	36
98%	1390	2172	1824	347	543	456	85	136	114	39	60	51
99%	1704	2663	2237	426	666	559	107	166	140	47	74	62

Table 6: Sample size for proportion correspond to three-precision

Level of Confidence	d=5%			d=10%			d=20%			d=30%		
	p=0.2	p=0.5	p=0.7	p=0.2	p=0.5	p=0.7	p=0.2	p=0.5	p=0.7	p=0.2	p=0.5	p=0.7
90%	1549	2421	2033	387	605	508	97	151	127	43	67	56
95%	2213	3457	2904	553	864	726	138	216	182	61	96	81
98%	3127	4886	4104	782	1222	1026	195	305	257	87	136	114
99%	3834	5991	5032	959	1498	1258	240	374	315	107	166	140

4. Sample Size When Characteristic of Interest is Counts

If the characteristics of interest is some average counts, say μ, then $\hat{\sigma}$ may be estimated as $\sqrt{\hat{\mu}}$, then as developed in section 2, we obtain the sample size as,

$$n = \left(\frac{1.96}{d}\right)^2 \hat{\mu}, \text{Correspond to one standard error precision} \qquad (10)$$

$$n = \left(\frac{2*1.96}{d}\right)^2 \hat{\mu}, \text{Correspond to two standard error precision} \qquad (11)$$

and

$$n = \left(\frac{3*1.96}{d}\right)^2 \hat{\mu}, \text{ Correspond to three standard error precision} \qquad (12)$$

Tables 7-9 presents the value of sample size for estimating counts according to various level of confidence, the estimate of count and precision levels (All entries are correspond to the actual $z_{\alpha/2}$ values).

Table 7: Sample size for counts correspond to one-precision

Level of Confidence	*d*=5%			*d*=10%			d=20%		
	μ=2	μ =4	μ =6	μ =2	μ =4	μ=6	μ =2	μ =4	μ =6
90%	2152	4303	6455	538	1076	1614	134	269	403
95%	3073	6147	9220	768	1537	2305	192	384	576
98%	4343	8686	13029	1086	2172	3257	271	543	814
99%	5325	10650	15975	1331	2663	3994	333	666	998

Table 8: Sample size for counts correspond to two-precision

Level of Confidence	*d*=5%			*d*=10%			*d*=20%		
	μ=2	μ =4	μ =6	μ =2	μ =4	μ=6	μ =2	μ =4	μ =6
90%	8607	17213	25820	2152	4303	6455	538	1076	1614
95%	12293	24586	36879	3073	6147	9220	768	1537	2305
98%	17372	34745	52117	4343	8686	13029	1086	2172	3257
99%	21300	42601	63901	5325	10650	15975	1331	2663	3994

Table 9: Sample size for counts correspond to three-precision

Level of Confidence	*d*=5%			*d*=10%			*d*=20%		
	μ=2	μ =4	μ =6	μ =2	μ =4	μ=6	μ =2	μ =4	μ =6
90%	19365	38730	58095	4841	9683	14524	1210	2421	3631
95%	27660	55319	82979	6915	13830	20745	1729	3457	5186
98%	39088	78176	117264	9772	19544	29316	2443	4886	7329
99%	47926	95852	143778	11982	23963	35945	2995	5991	8986

5. Sample Size When Characteristic of Interest is Difference of Means

As we know, if the characteristics of interest is the difference of population means $\mu_1 - \mu_2$, then for equal sample sizes, the 95% confidence interval for $\mu_1 - \mu_2$ is $(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{\sigma_1^{\ 2}}{n} + \frac{\sigma_2^{\ 2}}{n}}$. Therefore,

$$d = 1.96\sqrt{\frac{\sigma_1^{\ 2}}{n} + \frac{\sigma_2^{\ 2}}{n}} \tag{13}$$

$$\Rightarrow n = \frac{(1.96)^2(\sigma_1^{\ 2} + \sigma_2^{\ 2})}{d^2}$$

where σ_1^2 and σ_2^2 are the variances of populations under consideration. Similarly, the formula correspond to a precision of two (margin of error reduced by half) and three standard errors (margin of error reduced by one third) are respectively obtained as

$$n = \frac{(2x1.96)^2(\sigma_1^2 + \sigma_2^2)}{d^2} \tag{14}$$

and

$$n = \frac{(3x1.96)^2(\sigma_1^2 + \sigma_2^2)}{d^2} \tag{15}$$

The sample size formula for difference of means for different confidence level may be obtained similarly.

6. Sample size when characteristic of interest is difference of proportions

If the characteristics of interest is difference of population proportions $P_1 - P_2$, then for equal sample sizes, the 95% confidence interval for $P_1 - P_2$ is $(p_1 - p_2) \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}$. Therefore,

$$d = 1.96\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}$$
$$\Rightarrow n = \frac{(1.96)^2[p_1(1-p_1) + p_2(1-p_2)]}{d^2} \tag{16}$$

where p_1 and p_2 are the sample proportions of the study populations. Similarly, the formula correspond to a precision of two (margin of error reduced by half) and three standard errors (margin of error reduced by one third) are respectively obtained as

$$n = \frac{(2x1.96)^2[p_1(1-p_1) + p_2(1-p_2)]}{d^2} \tag{17}$$

and

$$n = \frac{(3x1.96)^2[p_1(1-p_1) + p_2(1-p_2)]}{d^2} \tag{18}$$

The sample size formula for difference of proportion for different confidence level may be obtained similarly.

References

[1] Lindsey, J. (1999). *Revealing Statistical Principles*. Arnold Publishers, London.

[2] Survey Methods and Practices. (2010). Statistics Canada, Editors: Sarah Franklin and Charlene Walker, (www.statcan.gc.ca).