STAT 515 Lec 13

Sample size calculations

Karl Gregory

Sample size calculations

A common question asked by researchers is how large a sample they should take. How many subjects should they include in their experiment? Intuition tells us that larger samples carry more information about the population than smaller samples, so we generally wish for n as large as possible. Of course, taking samples and running experiments costs money, so there is a trade-off between how much information the sample will carry and the expense of conducting the study.

We will focus on choosing the sample size when it is of interest to estimate a mean μ or a proportion p. Our strategy will be to consider how the sample size affects the width of the confidence interval one wishes to construct. Researchers typically know "how closely" they wish to estimate a population parameter, that is, they know they would like to estimate μ or p *within plus or minus something*, and they have a level of confidence in mind. These pieces of information can be put together to choose a reasonable sample size.

Concerning the mean μ

We first recall the form of a $(1-\alpha)100\%$ confidence interval for μ when σ is known. If the population is Normal or if the sample size is greater than 30, then we prescribe the interval

$$\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

as a $(1-\alpha)100\%$ confidence interval for μ. Now, for confidence intervals of the form

$$\text{estimator} \pm \text{something},$$

we refer to the "something" which we add and subtract as the *margin of error*. So the *margin of error* associated with the confidence interval above is

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Note that the width of a confidence interval is twice the margin of error. Now, suppose a researcher wishes to build a $(1-\alpha)100\%$ confidence interval for μ with a margin of error of at most M. Then we can compute the sample size necessary by writing

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq M,$$

and then rearranging it to get

$$n \geq \left(z_{\alpha/2}\frac{\sigma}{M}\right)^2.$$

This expression tells us that every sample size n greater than the right-hand side will lead to a confidence interval with a margin of error less than or equal to the desired margin of error M. The right-hand side of this expression may not be integer-valued, so we take the smallest whole number n for which it is satisfied; this amounts to computing the right-hand side and rounding up.

Note: We do not in general know σ, so we must replace it with something else in the above calculation. One option is to use data from previous studies to get an estimate, say $\hat{\sigma}^{\text{prev}}$, of σ. Then $\hat{\sigma}^{\text{prev}}$ can be used in the sample size calculation. If no data from any previous studies is available the researcher may conduct a small pilot study, spending a small amount of money to take a small sample just for the purpose of getting an estimate, say $\hat{\sigma}^{\text{pilot}}$ of σ. Then $\hat{\sigma}^{\text{pilot}}$ can be used in the sample size calculation to determine the sample size for the full-size study.

To summarize:

Result: Sample size for estimating μ

For a maximum desired margin of error M and confidence level $1-\alpha$, take

$$n = \left\lceil \left(\frac{z_{\alpha/2}\cdot\sigma}{M}\right)^2 \right\rceil.$$

In the above, $\lceil x \rceil$ is the smallest integer greater than or equal to x ("round up").

To use the above formula we need to substitute for σ an estimate from a previous or pilot study.

Example. Researchers would like to estimate the mean gestation period of a species of tortoises in a certain habitat. Specifically, they would like the margin of error at the 95% confidence level to be no greater than 0.5 months. Previous studies of tortoise gestation periods have resulted in an estimated standard deviation of 2 months. How many tortoises should the researchers include in their study?

Answer: The researchers want the margin of error to be less than $M = 0.5$. We have from previous studies the estimated standard deviation $\hat{\sigma}^{\text{prev}} = 2$. Since a 95% confidence interval

is to be constructed, we use $z_{0.05/2} = z_{0.025} = 1.96$. Thus we get

$$n \geq \left(z_{\alpha/2} \frac{\hat{\sigma}^{\text{prev}}}{M} \right)^2 = \left(1.96 \frac{2}{0.5} \right)^2 = 61.4656,$$

leading to the choice of sample size $n = 62$.

Concerning the proportion p

As with the mean μ, we begin by recalling the form of a $(1-\alpha)100\%$ confidence interval for p. We choose to work with the "large-sample" version

$$\hat{p}_n \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}},$$

which is the version we may use when $\min\{n\hat{p}_n, n(1-\hat{p}_n)\} \geq 15$. Suppose now that we want the margin of error of this confidence interval to be less than some researcher-specified value M. Then we may write

$$z_{\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leq M,$$

and rearrange it to get

$$n \geq \left(z_{\alpha/2} \frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{M} \right)^2.$$

This does not, however, immediately help us because we do not yet have a $\hat{p}_n$! Putting, for the moment, p instead of $\hat{p}_n$, we have this:

Result: Sample size for estimating p

For a maximum desired margin of error M and confidence level $1-\alpha$, take

$$n = \left\lceil \left(z_{\alpha/2} \cdot \frac{\sqrt{p(1-p)}}{M} \right)^2 \right\rceil.$$

To use the above formula we need to substitute for p

1. an estimate p from some previous or pilot study, or
2. simply the value $p = 1/2$, getting

$$n \geq \left(z_{\alpha/2} \frac{\sqrt{1/2(1/2)}}{M} \right)^2.$$

The value $p = 1/2$ maximizes $p(1-p)$, leading to the most conservative, i.e. the largest choice of n. So using $p = 1/2$ means erring towards a larger n. This is recommended when no previous studies are available and one does not wish to conduct a pilot study.

Example. It is of interest to estimate with 99% confidence the proportion of registered voters who will vote for a particular candidate. A margin of error of 2 percentage points is desired. How many people should be polled?

Answer: If it is a tight race between two candidates, the true proportion of registered voters favoring one of the candidates will be close to 0.5, so we should use $p = 0.5$ in our calculations. At the 99% confidence level $\alpha = 0.01$, so we need $z_{0.01/2} = z_{0.005} = 2.576$. The desired margin of error is $M = 0.02$. We thus get

$$n \geq \left(2.576 \frac{\sqrt{1/2(1/2)}}{0.02} \right)^2 = 4{,}147.36.$$

Se we should poll $n = 4{,}148$ people.