Determining the sample size

One of the most common questions any statistician gets asked is "How large a sample size do I need?" Researchers are often surprised to find out that the answer depends on a number of factors and they have to give the statistician some information before they can get an answer! As with all our exampes so far, the answers are essentially different depending on whether the study is a survey designed to find out the proportion of something, or is designed to find a sample mean. We consider these cases separately.

Sample size to estimate a proportion

Example: A professor in UNC's Sociology department is trying to determine the proportion of UNC students who support gay marriage. She asks, "How large a sample size do I need?"

To answer a question like this we need to ask the researcher certain questions, like

1. How accurately do you need the answer?
2. What level of confidence do you intend to use?
3. (possibly) What is your current estimate of the proportion of UNC students who support gay marriage?

Possible answers might be:

1. "We need a margin of error less than 2.5%". Typical surveys have margins of error ranging from less than 1% to something of the order of 4% — we can choose any margin of error we like but need to specify it.

2. 95% confidence intervals are typical but not in any way mandatory — we could do 90%, 99% or something else entirely. For this example, we assume 95%.

3. May be guided by past surveys or general knowledge of public opinion. Let's suppose answer is 30%.

Calculation of sample size:

We already know that the margin of error is 1.96 times the standard error and that the standard error is $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$. In general the formula is

$$\text{ME} = z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \qquad (\dagger)$$

where

- ME is the desired margin of error
- z is the z-score, e.g. 1.645 for a 90% confidence interval, 1.96 for a 90% confidence interval, 2.58 for a 99% confidence interval (see Table 8.2, page 369)
- $\widehat{p}$ is our prior judgment of the correct value of p.
- n is the sample size (to be found)

So in this case we set ME equal to 0.025, $z = 1.96$ and $\widehat{p} = 0.3$, and (†) becomes

$$0.025 \;=\; 1.96\sqrt{\frac{0.3 \times 0.7}{n}}$$

or

$$\frac{0.3 \times 0.7}{n} \;=\; \left(\frac{0.025}{1.96}\right)^2 \;=\; .0001627$$

which translates into

$$n \;=\; \frac{0.3 \times 0.7}{.0001627} \;=\; 1291.$$

So we would need a sample of about 1300 students.

We could clearly try varying any of the elements of this. For example, maybe the researcher would be satisfied with a 90% confidence interval, for which $z = 1.645$. In this case (†) becomes

$$0.025 \;=\; 1.645\sqrt{\frac{0.3 \times 0.7}{n}}$$

for which we can quickly find $n = 909$. If we are willing to accept a lower confidence level, we can get away with a smaller sample size.

A different type of variation is "What if we have no initial estimate of $\widehat{p}$?" In this case, the convention is to assume $\widehat{p} = 0.5$. The reason is that the standard error formula, $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$, is largest when $\widehat{p} = 0.5$, so this is a conservative assumption that allows for $\widehat{p}$ being unknown a priori.

If we repeat the calculation with $\widehat{p} = 0.5$ (but returning to $z = 1.96$), we find

$$0.025 \;=\; 1.96\sqrt{\frac{0.5 \times 0.5}{n}}$$

which results in $n = 1537$. The cost of $\widehat{p}$ being unknown is an increase in the sample size, though if $\widehat{p}$ were known and already quite close to 0.5 (as occurs in many election predictions where the result is close), this would not be too important a feature.

Sample size to estimate a population mean

The issues are similar if we are designing a survey or an experiment to estimate a population mean. In this case, the formula is

$$\text{ME} = t\frac{s}{\sqrt{n}} \qquad (\ddagger)$$

where

- ME is the desired margin of error
- t is the t-score that we use to calculate the confidence interval, that depends on both the degrees of freedom and the desired confidence level,
- s is the standard deviation,
- n is the sample size we want to find.

There is a complication here because the sample size affects t as well as n. However, when $n \geq 30$, the value of t is quite close to the value of z that we would get if we ignored the distinction between the normal and t distributions, so often we do ignore that distinction and just use the z value, e.g. 1.96 for a 95% confidence interval.

The second complication is the need to specify s. In practice, s will be the sample standard deviation, computed *after* the sample is taken. So we can't possibly know that in advance. But s is typically a guess, based either on past experience or on rough estimates of what sort of variability we would expect.

Example. We would like to estimate the mean teacher's salary in the Chapel Hill school district, with 99% confidence, to an accuracy within \$2,000. In this case we have literally no idea what s would be. But if you refer back to problem 2.120 on page 87 (this was part of HW3), there we deduced that among four possible values that were given, the likeliest was \$6,000. So in the absence of anything better, let's use that as our guess for s.

In this case the 99% confidence interval translates to a z or t of 2.58. Therefore (‡) becomes

$$2000 \;=\; \frac{2.58 \times 6000}{\sqrt{n}}$$

which solves to

$$n \;=\; \left(\frac{2.58 \times 6000}{2000}\right)^2 \;=\; 59.9$$

or 60 to the nearest whole number.

Other ideas

(no need to study in detail, but please read briefly)

1. Small sample estimation (pages 391–393): Idea of adding 2 to both the number of successes and the number of failures in the sample. This has been found to make the $\sqrt{\widehat{p}(1-\widehat{p})/n}$ formula work quite well even when n is small.

2. *Bootstrapping* (pages 395-397): Idea of generating new samples by *resampling* from current data. Actually, I have used this in some of the simulations I showed you in this course, though I didn't call it that at the time.

Some Worked Examples

8.95. A survey estimated that 20% of all Americans aged 16 to 20 drove under the influence of drugs or alcohol. A similar survey is planned for New Zealand. They want a 95% confidence interval to have a margin of error of 0.04.

(a) Find the necessary sample size if they expect to find results similar to those in the United States.

(b) Suppose instead they used the conservative formula based on $\widehat{p} = 0.5$. What is now the required sample size?

Solution:

(a) The general formula is

$$ME = z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

which also translates to

$$n = \frac{\widehat{p}(1-\widehat{p})z^2}{ME^2}$$

With $ME = 0.04, \widehat{p} = 0.2, z = 1.96$ we get

$$n = \frac{0.2 \times 0.8 \times 1.96 \times 1.96}{0.04 \times 0.04} = 384.2.$$

(b) With $ME = 0.04, \widehat{p} = 0.5, z = 1.96$ we get

$$n = \frac{0.5 \times 0.5 \times 1.96 \times 1.96}{0.04 \times 0.04} = 600.25.$$

The sample size is 384 for (a) and 600 for (b), showing the advantage in using the estimated $\widehat{p}$ (0.2) so long as we feel confident that this is roughly the right guess. Note that the choice $z = 1.96$ arises because this is the z value appropriate for a 95% confidence interval. If we were asked for a 99% confidence interval, for example, we would use $z = 2.58$.

8.97. A tax assessor wants to assess the mean property tax bill for all homeowners in Madison, Wisconsin. A survey ten years ago got a sample mean and standard deviation of $1400 and $1000.

(a) How many tax records should be sampled for a 95% confidence interval to have a margin of error of $100?

(b) In reality, the standard deviation is now $1500. Using the sample size you used in (a), would the margin of error for a 95% confidence interval be less than $100, equal to $100, or greater than $100?

(c) (Adapted.) Under (b), what is the true probability that the sample mean falls within $100 of the population mean?

Solution:

(a) The formula $ME = t\frac{s}{\sqrt{n}}$ translates to

$$n = \left(\frac{st}{ME}\right)^2.$$

With $s = 1000, t = 1.96, ME = 100$, we get $n = 384$.

(b) Since ME is proportion to s, if s increases from 1000 to 1500, then ME increases in the same proportion (to 150).

(c) $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ so with $\bar{x} - \mu = \pm 100, s = 1500, n = 384$ we get $t = \pm 1.31$. In this case with df=383, the t distribution is almost the same as the normal distribution, so we look this up in the standard normal table: the probability of getting a z score between -1.31 and $+1.31$ is $.9049 - .0951 = .8098$, i.e. the nominal 95% confidence interval in reality has about an 81% chance of including the true value.