Social Studies 201
Notes for November 22, 2006

Estimation of a population proportion – Section 8.5, p. 521.

For most of this semester, we have dealt with means and standard deviations this semester. This section examines proportions – using data from a random sample to estimate the proportion of a population with a particular characteristic. Using the same methods and procedures that are used for estimating means, it is also possible to obtain estimates of a population proportion.

Notation. Let p represent the proportion of a population with a particular characteristic and q denote the proportion of the population not having this characteristic. Since members of the population must either have this characteristic or not have this characteristic, $p + q = 1$. That is, the sum of the proportion with the characteristic and the proportion without this characteristic comprises the whole population. Since $p + q = 1$, the proportion of those without the characteristic is $q = 1 - p$. These are the values that will be estimated using the method of confidence interval estimates.

If a random sample of size n is selected from this population, let X represent the number of cases in the sample with this same characteristic. The sample proportion, or the proportion of cases in the sample with the characteristic, is X/n, and this is denoted by $\hat{p}$ in these notes. That is,

$$\hat{p} = \frac{X}{n}.$$

For example, in a sample of 800 respondents from Saskatchewan (see example of the election polls prior to the 2003 Saskatchewan provincial election, from November 8), suppose that 336 respondents support the NDP. In this sample, $n = 800$, $X = 336$, so that the proportion of those surveyed who support the NDP is

$$\hat{p} = \frac{X}{n} = \frac{336}{800} = 0.42.$$

The point estimate of the population proportion p is $\hat{p}$. For example, in opinion polls prior to an election, a pollster obtains estimates of the proportion of the population who state they will support each political party.

These proportions represent point estimates of the proportion of electors who actually will vote for the different political parties.

In the case of the CBC poll prior to the November 5, 2003 provincial election, Western Opinion Research reported that 42% of those surveyed said they would vote NDP while 39% said they would voted Saskatchewan Party. Converted into proportions of 0.42 and 0.39, these represent point estimates of the proportion who said they would vote NDP and Saskatchewan Party, respectively.

Sampling distribution of a proportion

To obtain interval estimates for a population proportion, it is necessary to know the sampling distribution of the sample proportion $\hat{p}$. That is, a researcher needs to know how $\hat{p}$ behaves as repeated random samples are drawn from a population. While this distribution can be obtained by considering the normal approximation to the binomial distribution (sections 6.3 and 6.5 of Chapter 6 of the text), another way is to consider a proportion as a special case of a mean. Then for large sample size, just as the sample mean $\bar{X}$ is normally distributed (by the central limit theorem), the sample proportion is also normally distributed. This result can be stated as follows.

> **Sampling distribution of a sample proportion $\hat{p}$.** If random samples of size n are drawn from a population with a proportion p of the population having a particular characteristic, and if the sample sizes are large, then the sample proportions $\hat{p}$ are normally distributed with mean p and standard deviation $\sqrt{pq/n}$. That is
>
> $$\hat{p} \text{ is Nor}\left(p, \sqrt{\frac{pq}{n}}\right).$$

This is essentially the same as the result from the central limit theorem, where large random samples yield a sampling distribution of sample means:

$$\bar{X} \text{ is Nor}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

In the above, the sample proportion $\hat{p}$ replaces $\bar{X}$. The statistic $\hat{p}$ is a special case of the mean when there are only two values for X – with the characteristic and without the characteristic. Similarly, the standard deviation of the sampling distribution is $\sqrt{pq/n}$ for a proportion, as opposed to $\sigma/\sqrt{n}$ for a mean.

Estimate of standard error of $\hat{p}$. In estimating the standard deviation of the sample mean, $\sigma/\sqrt{n}$, it was usually necessary to replace the unknown population standard deviation σ with the known sample standard deviation s. Similarly for an estimate of a proportion, the population proportion is not known, yet the standard error $\sqrt{pq/n}$ must be estimated. In order to provide this estimate, a researcher has two choices.

1. One possibility is to use $p = 0.5$ and $q = 0.5$ when estimating $\sqrt{pq/n}$. Since p and q must sum to one, it can be demonstrated that the maximum value of $p \times q$, when $p + q = 1$ occurs when $p = q = 0.5$. By selecting these values for p and q in the estimate of the standard error $\sqrt{pq/n}$, this produces the largest possible standard error for any given sample size. If a researcher wishes to ensure that he or she has not underestimated the sampling error, then it is best to use $p = q = 0.5$ in the expression $\sqrt{pq/n}$.

2. Another possibility, once a sample has been obtained, is to use $p = \hat{p}$ and $q = \hat{q} = 1 - \hat{p}$ in the expression $\sqrt{pq/n}$. This is comparable to using the sample standard deviation s as an estimate of σ when estimating the population mean. This tends to produce estimates of $\sqrt{pq/n}$ slightly smaller than the earlier approach. It also builds on the knowledge the researcher already has about the possible value of the population proportion, by using the point estimate $\hat{p}$.

Prior to conducting a sample, it is likely that the researcher would employ the first approach above; after the sample has been conducted, it is more common to use the second approach.

Large n. The sampling distribution of the sample proportion is normal so long as the sample is a random sample and the sample size is reasonably large. In the case of a proportion, a large sample size occurs when

$$n \geq \frac{5}{\text{smaller of p or q}}.$$

This rule is somewhat different than in the case of a large sample size for a sample mean, where the rule is a large n is more than $n = 30$. In the case of a proportion, if p is near 0.5, then a large sample size is any n larger than

$$\frac{5}{0.5} = 12.5.$$

However, if a characteristic of a population is uncommon, so the proportion of the population with this characeristic is small, say 0.01, or 1 in 100, then the sample size required is

$$\frac{5}{0.01} = 500$$

much larger than in the case of the central limit theorem.

Interval estimate for $\hat{p}$

The method of constructing a confidence interval estimate for the population proportion is the same as for the population mean. First, clearly define what population proportion p is being considered. Then the five steps are:

1. Obtain the sample size size n and the sample proportion $\hat{p}$. From $\hat{p}$, the sample proportion without the characteristic is $\hat{q} = 1 - \hat{p}$.

2. If the sample is a random sample with large n (more than 5 divided by the smaller of p or q), then

$$\hat{p} \text{ is Nor}\left(p, \sqrt{\frac{pq}{n}}\right).$$

3. Select a confidence level, C%.

4. Determine the Z-value associated with the confidence level of C%.

5. The interval estimates for the population proportion p are

$$\hat{p} \pm Z\sqrt{\frac{pq}{n}}$$

or

$$\left(\hat{p} - Z\sqrt{\frac{pq}{n}}\ ,\ \hat{p} + Z\sqrt{\frac{pq}{n}}\right).$$

That is, C% of the intervals constructed in this manner will contain the population proportion p.

Example – estimate of proportion supporting Saskatchewan Party

From *Saskatchewan Election Polls and Results* (page 2 of November 8, 2006 notes), the CBC poll, support for the Saskatchewan Party was 39% or, as a proportion, $\hat{p} = 0.39$. This is the point estimate of the proportion of Saskatchewan electors who support the Saskatchewan Party, p. A 95% interval estimate of p is obtained by using the five steps.

1. The sample size is $n = 800$ and the sample proportion of supporters is $\hat{p} = 0.39$, meaning that the sample proportion of non-supporters of the Saskatchewan Party is $\hat{q} = 1 - \hat{p} = 1 - 0.39 = 0.61$.

2. The sample size of $n = 800$ is large since it is much greater than 5 divided by an estimate of p. That is

$$\frac{5}{0.39} = 12.8$$

and this is much less than $n = 800$. The sample size of 800 is very large and more than sufficient to ensure that

$$\hat{p} \text{ is Nor}\left(p, \sqrt{\frac{pq}{n}}\right).$$

3. The handout from Western Opinion Research states the sampling error is ±3.5% nineteen times out of twenty. This is $19/20 \times 100\% = 95\%$ so the $C = 95\%$ confidence level is used.

4. For 95% confidence level and a normal distribution, the Z-value is 1.96 (95% of the area in the middle of the normal distribution).

5. The interval estimates are:

$$\begin{aligned} \hat{p} \pm Z\sqrt{\frac{pq}{n}} &= \hat{p} \pm 1.96\sqrt{\frac{0.39 \times 0.61}{800}} \\ &= \hat{p} \pm 1.96\sqrt{\frac{0.2379}{800}} \end{aligned}$$

$$\begin{aligned} &= \hat{p} \pm 1.96\sqrt{0.0002974} \\ &= \hat{p} \pm (1.96 \times 0.0172) \\ &= \hat{p} \pm 0.0338 \end{aligned}$$

and the interval estimate is 0.39 ± 0.034, that is, (0.356, 0.424) or 35.6% to 42.4%. This is slightly different than stated in the Western Opinion Research handout since the $\hat{p}$ and $\hat{q}$ were used in the estimate of standard error, rather than $p = q = 0.5$.

Note that the CBC poll provided a very accurate estimate of the proportion who actually voted for the Saskatchewan Party on November 5, with 39.35% voting this way. The interval from 35.6% to 42.4% certainly included this $p = 0.3935$.

Similar interval estimates for each of the other parties could also be obtained. The sampling error for the NDP and Liberal Party are each around ±3.5% as well, so that the actual percentages of voters who voted for the NDP lies within the respective 95% confidence interval estimate. The actual percentage of electors who voted Liberal is just outside the interval. For the Cutler poll, there are similar interval estimates and all the actual election results are within the respective confidence intervals.

Margin of error in Saskatoon. In their press release, Western Opinion Research also states that a sample of $n = 400$ voters was obtained in the city of Saskatoon, and the margin of error for this sample is ±4.9%, nineteen times out of twenty. In this case, the sample proportion $\hat{p}$ is not specified, yet it is possible to use the method of interval estimates to obtain the 4.9% sampling error.

The method used is exactly the same as for the province as a whole. For determining whether the sample size is large, an estimate of $p = 0.5$ can be used in the formula 5 divided by the smaller of p or q. That is, the formula for determining a large sample size uses the smaller of p or q. Using $p = q = 0.5$ avoids the issue of which of these two values is smaller. The sample size of $n = 400$ exceeds $5/0.5 = 12.5$ so the normal distibution for $\hat{p}$ can be used as before.

For purposes of estimating $\sqrt{pq/n}$ in the interval estimates, $p = q = 0.5$ can be used. That is, $\hat{p}$ and $\hat{q}$ are not specified, so an estimate of the maximum possible sampling error for a sample size of $n = 400$ is obtained by using

$p = q = 0.5$ in the estimate of $\sqrt{pq/n}$.

The interval estimates are

$$\begin{aligned}\hat{p} \pm Z\sqrt{\frac{pq}{n}} &= \hat{p} \pm 1.96\sqrt{\frac{0.5 \times 0.5}{400}} \\ &= \hat{p} \pm 1.96\sqrt{\frac{0.25}{400}} \\ &= \hat{p} \pm 1.96\sqrt{0.000625} \\ &= \hat{p} \pm (1.96 \times 0.025) \\ &= \hat{p} \pm 0.049\end{aligned}$$

or $\pm 4.9\%$, as stated in the handout.

Conclusion. From these results it can be seen that a sample of $n = 800$ results in a sampling error of approximately $\pm$ 3.5%, while a sample size of only $n = 400$ results in a sampling error of just under 5 per cent, both with 95% confidence. That is, if a researcher obtains such random samples, he or she can be confident that 95 out of 100 sample proportions will be within 3.5 percentage points of the population proportion if the sample size is 800. When random samples of size 400 are drawn from a population, a researcher can also be 95% sure that sample proportions are within about 5 percentage points of the population proportion.

Note that the interval will be wider:

1. the larger the confidence level – Z-value is larger and there is greater certainty that the intervals contain the population proportion.

2. the smaller the sample size – a smaller n reduces the denominator of $\sqrt{pq/n}$ and increases its overall value.

3. if the value for p and q are close to $p = q = 0.5$. If the researcher is fairly certain that the true proportion is either much greater or much less than 0.5, then $\hat{p}$ and $\hat{q}$ can be used in $\sqrt{pq/n}$, and this will generally produce a narrower interval.

Determining sample size for estimation of a population proportion – Section 8.6.2, p. 541.

As indicated in the notes for November 20, when sample size is larger, the interval estimate is narrower and sampling error is reduced, compared with smaller sample size. This section of the notes outlines how to obtain the sample size required to estimate a population proportion for any specified sampling error and confidence level.

Notation. Let p represent the proportion of a population with a particular characteristic and q denote the proportion of the population not having this characteristic. Since members of the population must either have this characteristic or not, $p + q = 1$ and $q = 1 - p$.

Let the size of the sampling error be given the symbol E. That is, the C% confidence level will result in the interval estimates of $\hat{p} \pm E$ if the required sample size is obtained. And if the required sample size is obtained, C% of these intervals will contain the population proportion p.

Note that the units for E are proportions. For example, if the proportion of population members with a particular characteristic is to be estimated to within ± 2 percentage points, the value of E will be 0.02. That is, the point estimate of p will be a proportion $\hat{p}$, and this will be accurate to within ± 0.02, so that the intervals will be $\hat{p} - 0.02$ to $\hat{p} + 0.02$.

Formula for deterining sample size

As with the interval estimates for a population proportion p, determining sample size begins by considering the sampling distribution of the sample proportion $\hat{p}$. Suppose that random samples of large sample size are taken from a population with a proportion p of members having a particular characteristic. The sample proportions $\hat{p}$ are normally distributed with mean p and standard deviation $\sqrt{pq/n}$. That is,

$$\hat{p} \text{ is Nor} \left(p, \sqrt{\frac{pq}{n}} \right).$$

This is the case so long as n exceeds 5 divided by the smaller of p or q.

Larger sample sizes yield normal distributions of $\hat{p}$ that are more concentrated, smaller sample sizes yield normal distributions of $\hat{p}$ that are more dispersed. For any given confidence level C and associated Z-value, the aim is to find a distribution where the confidence interval estimates

$$\hat{p} \pm Z\sqrt{\frac{pq}{n}}$$

match the intervals associated with the specified sampling error E:

$$\hat{p} \pm E.$$

That is, the C% intervals are constructed so that they are $Z\sqrt{pq/n}$ on either side of $\hat{p}$. But the researcher specifies these are to be intervals of amount E on either side of $\hat{p}$. The desired error of estimate E and the confidence intervals are the same when a sample size is selected so that

$$E = Z\sqrt{\frac{pq}{n}}.$$

When this latter expression is solved for n, the required sample size is

$$n = \left(\frac{Z}{E}\right)^2 pq$$

This is the formula for the required sample size for a specified error of estimate E and for a Z-value associated with the specified confidence level.

The procedure for estimating sample size is to select a confidence level C and an error of estimate E that the researcher wishes to obtain. From the confidence level the Z-value can be determined from the table of the normal distribution. Using the above formula, the only other parts in question are the values of p and q. As stated earlier, when $p + q = 1$, the maximum value of the product of p and q occurs when $p = q = 0.5$. If a researcher wishes to determine a sample size that is sufficient to obtain sampling error E with confidence level C, then this is obtained when $p = q = 0.5$. In this circumstance, the formula for obtaining the required sample size becomes simply

$$n = \left(\frac{Z}{E}\right)^2 \times 0.25$$

since $pq = 0.5 \times 0.5 = 0.25$.

If a researcher has some knowledge that p and q are quite different than 0.5 each, then these alternate estimates for p and q can be used in the formula

$$n = \left(\frac{Z}{E}\right)^2 pq.$$

This will result in a smaller required sample size and it may be easier or less costly for the researcher to obtain this smaller sample. The concern a researcher might have though is that this smaller sample size may not be sufficient to produce intervals with the required error of estimate. Resulting interval estimates may be wider than desired.

Examples.

Suppose a researcher wishes to estimate the proportion of a population who support legalizing marijuana, correct to within (a) 5 percentage points, or (b) 2 percentage points, with probability 0.99. What are the required sample sizes?

Answer. This is an estimate of a proportion – the proportion p of the population who support the legalization of marijuana. Since the sample size will likely be fairly large, it can be assumed that the sample proportions $\hat{p}$, of those who support legalization of marijuana, will be normally distributed. The distribution of the sample proportions

$$\hat{p} \text{ is Nor} \left(p, \sqrt{\frac{pq}{n}} \right).$$

The formula for sample size is

$$n = \left(\frac{Z}{E} \right)^2 pq$$

where $E = 0.05$ for part (a). The confidence level specified is 99% (0.99 probability) and the associated Z-value is 2.575. Letting $p = q = 0.5$, the required sample size is

$$n = \left(\frac{2.575}{0.05} \right)^2 0.5 \times 0.5 = (51.5)^2 \times 0.25 = 2,652.25 \times 0.25 = 663.1$$

The required sample size is 664.

For an accuracy of 2 percentage points, $E = 0.02$ and the required sample size is

$$n = \left(\frac{2.575}{0.02} \right)^2 0.5 \times 0.5 = (128.75)^2 \times 0.25 = 16,576.562 \times 0.25 = 4,144.1$$

or 4,145. This latter sample size is very large so it is unlikely that most research projects could obtain a sample with accuracy of ± 2 percentage points with probability 0.99.

Conclusion. A few concluding points concerning the determination of sample size for estimation of a proportion are as follows.

1. The formula for determining sample size in the case of estimation of a proportion
$$n = \left(\frac{Z}{E}\right)^2 pq$$
has advantages over the formula for estimating a population mean in that the values of p and q can always be set to 0.5 each. This will always produce a sample size sufficient to produce the required accuracy E at whatever confidence level the researcher specifies. In the case of estimating the sample mean, the researcher needed some knowlege of the variability of the population being sampled – that is, an estimate of σ was required in order to determine sample size. In the case of a proportion, this is not necessary; a researcher can always use $p = q = 0.5$ and be sure this will produce a large enough sample size.

2. All of the above results apply to random sampling from a population. While researchers consider larger sample size to be better than smaller sample sizes, strictly speaking this may be the case only if the samples are random, or chosen using the principles of probability. If samples are judgment or snowball samples, large samples may not be all that much better than smaller samples.

   If other forms of probability samples are used, for example, cluster or stratified samples, formula such as that used in this section can be developed. But the formula in this section applies only to random sampling.

3. If a researcher considers the sample size too large when $p = q = 0.5$, different estimates of p and q can be used. In the example, if a researcher thinks that only 15% of the population oppose the legalization of marijuana, so that the researcher is willing to work with $\hat{p} = 0.85$ and $\hat{q} = 0.15$ when estimating $\sqrt{pq/n}$, the required sample size for (b)

would be

$$n = \left(\frac{2.575}{0.02}\right)^2 0.85 \times 0.15 = (128.75)^2 \times 0.1275$$

$$n = 16,576.562 \times 0.1275 = 2,113.5$$

or 2,114. This is much less than the earlier sample size of $n = 4,145$. The only danger here is that if the proportions supporting or opposing legalization of marijuana are closer to 0.5 than 0.85 and 0.15, then this sample size may produce a confidence interval estimate that has a sampling error greater than 0.02.

4. Given that $p = q = 0.5$ can always be used in order to determine sample size, it is possible to construct tables of required sample size for different confidence levels C and accuracy of estimate E. Table 8.8, p. 544 of the text is reproduced here as Table 1. Using $p = q = 0.5$ and the above formula, you should be able to verify all the sample sizes in this table.

Table 1: Sample Sizes for a Proportion, Common Levels of Accuracy and Confidence

Level of	Confidence Level		
Accuracy (E)	90%	95%	99%
0.05	271	385	664
0.04	423	601	1,037
0.03	752	1,068	1,842
0.02	1,692	2,401	4,145
0.01	6,766	9,604	16,577

From Table 1, note that as the researcher is more demanding in terms of accuracy (smaller E), required sample size is greater. Similarly, as a researcher is more demanding in terms of requiring greater confidence that the intervals will contain the mean, sample size is again increased. In practice, the actual sample size selected is likely to be informed by the considerations of this section, but may depend more on the budget and time available for the researcher. With limited budget and time for a survey, a researcher may just have to live with the lesser accuracy associated with a smaller sample size.

Last edited November 22, 2006.