Social Studies 201
Notes for December 1, 2004

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

As indicated in the notes for November 17, 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.

Test of a proportion, large sample size –section 9.4, p. 622.

An hypothesis test for a population proportion can be conducted using the same principles of hypothesis testing as for a population mean. Recall from the notes of November 17 that a proportion is a special case of a mean. When considering the proportion of the population that takes on a particular characteristic, the variable takes on only two values, those without the characteristic and those with the characteristic, so the mean of this variable is equal to the proportion of the population with the characteristic.

Also recall that when a random sample is selected from a population with the proportion p of members having a particular characteristic, the sample proportions are normally distributed.

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

For this result, a large sample size means a sample size n greater than 5 divided by the smaller of p or $q = 1 - p$.

Using this distribution for the sample proportion $\hat{p}$, the number of standard deviations this sample proportion $\hat{p}$ is from the hypothesized proportion

of p is

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}.$$

When conducting an hypothesis test for p, if this Z-value is in the critical region, we reject H_0, but if it is not in the critical region there is insufficient evidence to reject the null hypothesis.

Using the above result, an hypothesis test can be constructed for a population proportion, using the same six steps as used for a test of a mean. The principles involved in an hypothesis test of a proportion are identical to those for testing a mean.

Example – election polls

In the 1999 Saskatchewan provincial election, the NDP received 38.73% of the of the total vote and the Saskatchewan party received 39.61% of the total vote. The CBC poll of 800 respondents, conducted about two weeks prior to the November 5, 2003 Saskatchewan provincial election reported that 42% of voters would vote NDP and 39% would vote for the Saskatchewan party. From these results, can you conclude that

1. Support for the NDP has increased? (0.10 level of significance).
2. Support for the Saskatchewan party has changed? (0.10 significance level).
3. Comment on the conclusions.

1. Test for support for the NDP

Let p be the true proportion of Saskatchewan voters who supported the NDP just before the November 5, 2003 election. The hypothesis test is as follows.

1. **Hypotheses**. The question is whether support for the NDP has increased. Since the null hypothesis must be an equality, it makes sense to hypothesize no change in support for the NDP and reject this hypothesis only if there is evidence of some increase in support. Thus the hypotheses are:

$$\text{Null hypothesis} \quad H_0 : p = 0.3873$$

$$\text{Alternative hypothesis} \quad H_1 : p > 0.3873$$

That is, the null hypothesis posits that the proportion of voters supporting the NDP has not changed since 1999; the alternative hypothesis is that the proportion supporting the NDP has increased since 1999.

2. **Test statistic**. The appropriate test statistic is $\hat{p}$, the proportion of those polled who express support for the NDP.

3. **Distribution of test statistic**. Since the sample size of $n = 800$ is large,
$$\hat{p} \text{ is Nor} \left(p, \sqrt{\frac{pq}{n}} \right).$$
To check that n is large, check to see whether n exceeds 5 divided by the smaller of p or $q = 1 - p$. For determining this, use $\hat{p} = 0.42$ for the estimate of p and $5/0.42 = 11.9 < 800$, so the sample size is large and $\hat{p}$ can be assumed to have a normal distribution.

4. **Critical region**. The question asks for an $\alpha = 0.10$ level of significance and the alternative hypothesis is one-directional, so the critical region for rejecting H_0 is at the extreme right end of the normal distribution. For a B area of $\alpha = 0.10$, the Z-value is 1.28. The region of rejection for the null hypothesis is all Z-values greater than 1.28.
$$\text{Region of rejection of } H_0 : \ Z > 1.28$$
$$\text{Area of nonrejection of } H_0 : \ Z \leq 1.28$$

5. **Conclusion**. The final step involved in the hypothesis test is to determine whether the sample proportion $\hat{p} = 0.42$ is in the critical region. This is accomplished by obtaining the Z-value associated with $\hat{p}$, that is,
$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}.$$
The hypothesized proportion of NDP supporters is $p = 0.3873$ so this value and the corresponding $q = 1 - p = 0.6127$ can be used to provide an estimate of pq in $\sqrt{pq/n}$, the standard deviation of $\hat{p}$.

$$\begin{aligned} Z &= \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \\ &= \frac{0.0327}{\sqrt{\frac{0.3873 \times 0.6127}{800}}} \\ &= \frac{0.0327}{\sqrt{\frac{0.2373}{800}}} \\ &= \frac{0.0327}{\sqrt{0.000297}} \\ &= \frac{0.0327}{0.0172} \\ &= 1.8986 > 1.28 \end{aligned}$$

As a result, the Z-value differs from $Z = 0$ enough to be in the critical region. This means that $\hat{p} = 0.42$ differs enough from $p = 0.3873$ to reject the null hypothesis. The null hypothesis is rejected and the alternative hypothesis, that support for the NDP has increased, is accepted at the 0.10 level of significance.

2. Test for support for the Saskatchewan Party

Let p be the true proportion of Saskatchewan voters who supported the Saskatchewan Party just before the November 5, 2003 election. The hypothesis test is as follows.

1. **Hypotheses**. The question is whether support for the Saskatchewan Party has changed. In this case the null hypothesis is no change in support for the Saskatchewan Party while the alternative hypothesis is that there has been a change. The hypotheses are:

$$\text{Null hypothesis} \quad H_0 : p = 0.3961$$

$$\text{Alternative hypothesis} \quad H_1 : p \neq 0.3961$$

That is, the null hypothesis posits that the proportion of voters supporting the Saskatchewan Party has not changed since 1999; the alternative hypothesis is that the proportion has changed since 1999.

2. **Test statistic**. The appropriate test statistic is $\hat{p}$, the proportion of those polled who express support for the Saskatchewan Party.

3. **Distribution of test statistic**. Since the sample size of $n = 800$ is large,
$$\hat{p} \text{ is Nor}\left(p, \sqrt{\frac{pq}{n}}\right).$$
To check that n is large, check to see whether n exceeds 5 divided by the smaller of p or $q = 1 - p$. For determining this, use $\hat{p} = 0.39$ for the estimate of p and $5/0.39 = 12.8 < 800$, so the sample size is large and $\hat{p}$ can be assumed to have a normal distribution.

4. **Critical region**. The question asks for an $\alpha = 0.10$ level of significance and the alternative hypothesis is two-directional, so the critical region for rejecting H_0 is the extreme 0.05 of the distribution at the left end of the distribution plus the extreme 0.05 of the distribution at the right end. For a B area of $\alpha = 0.05$, the Z-values are ± 1.645.
$$\text{Region of rejection of } H_0: \ Z < 1.645 \text{ or } Z > +1.645$$
$$\text{Area of nonrejection of } H_0: \ -1.645 \leq Z \leq +1.645.$$

5. **Conclusion**. The final step involved in the hypothesis test is to determine whether the sample proportion $\hat{p} = 0.39$ is in the critical region. This is accomplished by obtaining the Z-value associated with $\hat{p}$, that is,
$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}.$$
The hypothesized proportion of Saskatchewan Party supporters is $p = 0.3961$ so this value and the corresponding $q = 1 - p = 0.6039$ can be used to provide an estimate of pq in $\sqrt{pq/n}$, the standard deviation of $\hat{p}$.

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

$$
\begin{aligned}
&= \frac{0.39 - 0.3961}{\sqrt{\frac{0.3961 \times 0.6039}{800}}} \\
&= \frac{0.0061}{\sqrt{\frac{0.2392}{800}}} \\
&= \frac{0.0061}{\sqrt{0.000299}} \\
&= \frac{0.0061}{0.0172} \\
&= 0.3528 > -1.645 \text{ and } < +1.645.
\end{aligned}
$$

As a result, the Z-value is not different enough from $Z = 0$, or $\hat{p} = 0.39$ is not different enough from $p = 0.3961$, to reject the null hypothesis at the 0.01 level of significance.

3. Comments

From the CBC poll 800 respondents in late October 2003, there is initial evidence that support for the NDP increased by a few percentage points (from 38.73% to 42%) and support for the Saskatchewan party declined very slightly (from 39.61% to 39%). From the above tests, at the 0.10 level of significance, it can be concluded that support for the NDP increased and support for the Saskatchewan Party was unchanged.

At the time the poll was conducted, there was the possibility of Type I error in the conclusion concerning the NDP and the possibility of type II error in the conclusion about the Saskatchewan Party. Since the proportion of voters who would ultimately vote NDP was not known at the time of the CBC poll, there was the possibility that the poll sampled a set of voters who were more likely, than the population as a whole, to vote NDP. If there had been no change in support for the NDP, this could have resulted in type I error, rejecting the null hypothesis of no change in NDP support and concluding that support had increased. In fact, given the election results, support for the NDP had increased to 44.61% by election day, November 5, 2003. While there was the possibility of Type I error, such does not appear to have occurred here.

For the hypothesis test about support for the Saskatchewan Party, there was the possibility of type II error, that is, there may have been a change in support for the Saskatchewan Party, even though the CBC poll result was

consistent with the conclusion of no change in support. There very likely was type II error, so that Saskatchewan Party support was not exactly equal to the 39.61% they obtained in 1999. But on election day, November 5, 2003, the Saskatchewan Party received 39.61% of the popular vote, very little different than the 39.35% they received in 1999. While support changed, it did not change very much, so that the consequence of type II error were minimal here.

The conclusions from the hypothesis tests turned out to be correct, as demonstrated by the results on election day. It may be that opinions shifted slightly between the time of the poll and election day, so the poll could not have been expected to be an accuarate prediction of popular vote on November 5. But the poll came very close to predicting the popular vote on election day.

Last edited December 1, 2004.