Social Studies 201
Notes for November 8, 2004

Introduction to estimation

Rest of semester

For the rest of the semester, we will be studying and working with inferential statistics – estimation and hypothesis testing. This week and part of next week will be devoted to methods of estimating an unknown population mean or proportion. These methods are found in Chapter 8 of the text. Following that, we will study hypothesis testing – Chapters 9 and 10 of the text.

There will be one more problem set – Problem set 5 – that I will hand out on Friday, November 14 or Monday, November 17. It will be due around the end of the month. There will be some computer work in the labs of November 11, 18, and 25. I will also provide an extra set of optional problems so that those who wish to raise their grade a few points can attempt these.

Saskatchewan Election Results

The Saskatchewan Election Results handout demonstrates how pollsters can fairly accurately predict the popular vote for election results. Both the CBC and Cutler poll provided very close predictions of the per cent of the total vote obtained by the NDP, Saskatchewan Party, and the other group. Cutler came very close to predicting the Liberal vote but the CBC poll overestimated this by almost 4 percentage points (18 per cent predicted and 14 per cent in actuality). Apart from this, the prediction error was no more than 2.6 percentage points in all cases – the CBC underestimated the NDP vote by $|42 - 44.6| = 2.6$ percentage points. Much of the prediction error associated with these polls was likely due to sampling error – the potential error introduced because only a sample of electors, rather than the whole population, was selected. It is these sampling errors that form the main part of the discussion of Chapter 8.

In addition to the error due to sampling, pollsters face the problem that respondents may be undecided or unwilling to say which party they will favour. Or, on election day, they may vote differently than what they told the pollster a few days earlier. These nonsampling errors make it difficult

for a pollster to predict the exact election result. In the case of these polls, error is possibly introduced because there were fifteen or sixteen per cent undecided. Since a pollster cannot say anything about how these people will vote, the existence of a large undecided group can play havoc with predicting. If all the fifteen per cent undecided had decided to vote Saskatchewan Party, that party would have won with a landslide. If all of these fifteen per cent had decided to vote NDP, the NDP might have shut out all the other parties. In fact, it appears that either the undecided did not vote or split their votes in a similar manner to those who told pollsters how they would vote.

A final issue is prediction of the number of seats won by each party. This is a much more difficult matter, since the provincial election is, in essence, a series of fifty-eight simultaneous elections. That is, the electors of each constituency vote and a winner is decided in each of the fifty-eight constituencies. While predicting the popular vote can help predict the number of seats won, in order to provide an accurate prediction of the number of seats each party will win, a pollster would have to obtain a large random sample in each constituency. This would be much to expensive so is usually not done.

Importance of random sampling and central limit theorem

One of the major reasons for conducting social research is that the characteristics of populations are unknown. For example, before an election, it is not clear how the vote will go, so pollsters poll the population in an attempt to determine this. Much social research is also devoted to attempting to determine the mean value of various characteristics of a population – mean income, mean alcohol consumption, mean student debt, and so on. To provide good estimates of the unknown mean, μ, of a population, it is often useful to obtain a large random sample of the population. As will be argued below, the mean of the cases selected in the random sample, $\bar{X}$, provides a relatively accurate estimate of the mean μ of the whole population. In addition, if the sample is random, the probability of different levels of sampling error, $|\bar{X} - \mu|$, can also be determined. The rationale for these results is provided by the central limit theorem (p. 442). The theorem is as follows:

> **Central limit theorem**. If X is a variable with a mean of μ and a standard deviation of σ, and if random samples of size n are drawn from this pouplation, then the sample means from these samples, $\bar{X}$, have a mean of μ and a standard deviation of $\sigma/\sqrt{n}$.

If the sample sizes of these samples are reasonably large, say 30 or more, then the sample means are also normally distributed. Symbollically, this can be written

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

There are four important results that emerge from this theorem, a theorem that can be proven mathematically, but that we will have to accept.

1. **Any population**. For all practical purposes, the type of population, or distribution of a variable, from which a sample is drawn does not matter. That is, regardless of the nature of the population, the central limit theorem describes the way the sample means, $\bar{X}$, are distributed. The only real qualification is that the sample must be a **random** sample and the sample size must be reasonably large.

2. **Normal**. From the theorem, the distribution of sample means has a normal distribution. That is, the way the sample means are distributed is fairly predictable – it is not just that the sample means are centred at the population mean μ, but the sample means have the well-known pattern of a normal distribution. Since the areas, or probabilities, associated with a normal curve are known, a researcher can use these to determine probabilities for different levels of sampling error. Suppose a researcher is attempting to estimate the mean income of a population. After selecting a random sample of members of the population, a researcher may find that the sample mean household income is $40,000. While the researcher does not know what the true mean income is, from the central limit theorem, the researcher can determine the probability that the income is in error by no more than $5,000.

3. **Standard error**. The theorem states that the distribution of sample means has a standard deviation of $\sigma/\sqrt{n}$. This standard deviation is sometimes referred to as the **standard error of the mean**. That is, this standard deviation of the distribution of the sampling error of the mean is sometimes called the standard error and is sometimes given the following symbol:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

This is further described on p. 441 of the text.

4. **Large sample size**. While a sample size of $n = 30$ is often regarded as large, there is some disagreement about the size of the random sample that is required to ensure that the central limit theorem holds. Most researchers would likely agree that a random sample of size 100 or more is sufficient to ensure that the theorem holds. Some researchers may argue that a sample size of just over thirty cases is insufficient to ensure the theorem holds, but for this course we will accept the rule that 30 or more cases constitute a large sample. For samples that have sample size smaller than 30 cases, we will use the t-distribution.

   One result that is clear from the theorem is that the larger the size of the random sample, the smaller is the size of the standard error. For example, the standard deviation of income for a population is \$1,500, consider a random sample of size 100 from this population, and another random sample of size 2,500. These samples and their corresponding standard errors are summarized in Table 1.

Table 1: Standard error of mean for random samples of size 100 and 2,500 from a population with standard deviation of \$1,500

Sample size	Standard error
$n = 100$	$\sigma/\sqrt{n} = 1,500/\sqrt{100} = 1,500/10 = 150$
$n = 2,500$	$\sigma/\sqrt{n} = 1,500/\sqrt{2,500} = 1,500/50 = 30$

For the sample of size $n = 2,500$, the standard error is only \$30, whereas the standard error is \$150 for the sample of size 100. That is, the sample means from the samples of size 2,500 have a small standard error and are thus concentrated around the actual population mean. This implies that the probability of a large sampling error is relatively small. In contrast, for the smaller random samples of size 100, the standard deviation of the sample means is larger, meaning that the sample means are more likely to differ from the population

mean. The diagram on page 446 shows how the distribution of sample means differs for three different sample sizes.

As a result, when a larger random sample is available, it is preferred over a smaller random sample. The larger the sample size, the more precise are the estimates of the mean of the population.

These results are now applied to the issue of estimating the mean of a population.