Social Studies 201
Notes for November 8, 2006

Sampling distributions

Rest of semester

For the remainder 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 examining 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.

These notes discuss some aspects of sampling, in particular, random sampling. When a random sample is drawn from a population, it can be shown that the distribution of the sample mean has a normal distribution. This result is known as the Central Limit Theorem, a result that forms the basis for much of the statistical inference in the remaining parts of the semester. Before examining this, there is a short discussion of opinion polls conducted prior to an election. The example used here is that of the Saskatchewan Election Polls and Results from the 2003 provincial election – see Table 1.

Saskatchewan Election Polls and Results – 1999 and 2003

Table 1. Percentage of respondents, votes, and number of seats by party, November 5, 2003 Saskatchewan provincial election

Political Party	CBC Poll, October 20-26	Cutler Poll, October 29 to November 5	Election Result	Number of Seats
NDP	42	47	44.61	30
Saskatchewan Party	39	37	39.35	28
Liberal	18	14	14.17	0
Other	1	2	1.87	0
Total	100	100	100.00	58
Undecided	15%	16%		
Sample size	800	773		

Sources: CBC Poll results from Western Opinion Research, “Saskatchewan Election Survey for The Canadian Broadcasting Corporation,” October 27, 2003. Obtained from web site http://sask.cbc.ca/regional/servlet/View?filename=poll_one031028, November 7, 2003.

Cutler poll results provided by Fred Cutler and from the *Leader-Post*, November 7, 2003, p. A5.

Table 2. Percentage of vote and number of seats by political party, 1999 Saskatchewan provincial election

Political Party	Election Result	Number of Seats
NDP	38.73	29
Saskatchewan Party	39.61	25
Liberal	20.15	4
Other	1.51	0
Total	100.0	58

Source: http://www.mapleleafweb.com/education/spotlight/issue_42/recent-election-results.html, November 7, 2003.

Saskatchewan Election Results

Table 1 of Saskatchewan Election Results 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 other groups. 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 and surveyed. It is these sampling errors that form the main part of the discussion of Chapter 8.

In addition to the potential 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 polls of this type, there is also potential error in predicting the exact results because of those who are undecided – in this case fifteen or sixteen per cent of those polled were undecided just a few days prior to the election. Since a pollster can say little 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 manner similar to those who had decided how to vote when they talked to the pollsters.

A final issue is prediction of the number of seats won by each party. It is much more difficult to predict the distribution of seats than it is to predict the distribution of overall votes by party. This is because the provincial election is conducted in many constituencies so that, in essence, a provincial election is really a set of fifty-eight simultaneous elections – one in each constituency. 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 too expensive for most polling agencies so this is usually not done.

Random sampling and central limit theorem

One of the major reasons for conducting social research is to determine characteristics of populations that are unknown. For example, before an election, no one is certain how many votes there will be for each party or candidate, so pollsters conduct surveys of members of the population in an attempt to predict vote results. 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 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. If the sample is a random sample drawn from a population, it is possible to determine the probability associated with different levels of sampling error, $|\bar{X} - \mu|$. 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 population, 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
>
> $$\bar{X}\text{is Nor}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{when } n \text{ is large}$$

Four important results emerge from this theorem, a theorem that can be proven mathematically, but unless you have considerable expertise in mathematics, you 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 qualifications are that the sample must be a **random** sample from the population and the sample size must be reasonably large (see item 4 for guidelines).

2. **Normal distribution**. 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 and listed in the table of the normal distribution, you can use these to determine probabilities for different levels of sampling error.

   For example, 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. This is a best estimate of the true mean income of all household but, in fact, the researcher does not know what the true mean income is, since only a sample was surveyed. But from the central limit theorem, the researcher can calculate the probability that the sample mean income is in error by no more than \$100, or differs from the true mean by no more than \$100. Example 7.6.1, pp. 456-460 of the text provides an example of how probability this can be determined.

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**. While this is a misnomer in that it is not really an error, the standard error refers to the standard deviation of the distribution of the sampling error associated with the mean. The standard error is sometimes given the symbol $\sigma_{\bar{X}}$ and

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

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

4. **Large sample size**. While a sample size of $n = 30$ is sometimes regarded as large, there is disagreement among statisticians about the number of cases required in a random sample to ensure that the central limit theorem holds. Many researchers would likely agree that a random sample of size 100 or more is sufficient to ensure that the theorem holds. Some researchers might argue that a sample size of just over thirty cases is insufficient to ensure the theorem holds. 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, introduced in Chapter 8 of the text, p. 502.

   One result that emerges from the theorem is that the larger the size of the random sample, the smaller is the size of the standard error. For example, suppose that the standard deviation of income for a population is $1,500. Further suppose that two random samples are drawn from this population, one of size $n = 100$ and another of size $n = 2,500$. The standard errors associated with these samples are obtained and summarized in Table 3.

Table 3: Standard error for random samples of size 100 and 2,500 from a population with a 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 figure on page 446 of the text 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 random sample having a smaller sample size. The larger the sample size, the more precise are estimates obtained from the sample.

In this course, these results are next applied to the issue of estimating the mean of a population.

Last edited November 14, 2006.