Social Studies 201
Notes for November 10, 2003

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.

Estimation – Chapter 8

Estimation and hypothesis testing are the two parts of inferential statistics. Chapter 8 discusses methods of constructing estimates of various characteristics of populations. Chapters 9 and 10 examine hypothesis tests.

Characteristics of a population are generally not known prior to conducting a survey of the population – one aim of such a survey is to obtain estimates of the characteristics of the population. Estimates are of two types, **point estimates** and **interval estimates**. While Chapter 8 is primarily concerned with constructing interval estimates, obtaining a point estimate or statistic is the first step in constructing the interval estimate. Table 2 summarizes the types of estimates examined in Chapter 8.

That is, the first step in obtaining an estimate of a characteristic of a population, such as the mean or proportion, is to determine the corresponding characteristic from a sample. For example, in the case of the Saskatchewan election, the percentages reported by the pollsters are statistics or point estimates of the true percentages of votes obtained by the different political parties.

The interval estimate is an interval that is constructed around the statistic or point estimate. The purpose of the interval estimate is to construct an interval that hopefully contains or brackets the population mean. Given that the population mean is unknown, the researcher can never be certain that the interval actually contains the population mean. But the researcher can attach a probability to the interval estimate – the probability that the interval estimates, constructed from the samples, contain the mean. For example, in researching household income, after conducting a survey, a researcher might report something like

Table 2: Parameters and Point Estimates

Measure	Population value or parameter	Statistic or point estimate
Mean	μ	$\bar{X}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Percentage	P	$\hat{P}$

> The probability is 0.95 that the intervals $\bar{X} \pm \$1,000$ contain the true mean income of the population. From my sample, the point estimate of mean income is \$40,000 and the corresponding interval is from \$39,000 to \$41,000. I am relatively confident that the mean income for the population is in the interval (\$39,000, \$41,000).

Note that the researcher cannot say for sure what the true mean income is, but he or she can construct an interval associated with the estimate and feel relatively confident that the true mean is somewhere in the interval. And there is a probability that is attached to each such interval.

For the most part, results from polls and surveys report point estimates – statistics or single values that are estimates of characteristics of a population. In this section of the course, we examine how to obtain the corresponding interval estimates.

Terminology. Interval estimates are sometimes referred to as confidence intervals, reflecting the fact that the researcher has a certain confidence that the interval contains the true mean. To each interval is attached a probability, or confidence level, reflecting the probability or confidence that the interval contains the population mean. These interval estimates can be termed confi-

dence intervals or confidence interval estimates – all of these are just different terms for the same type of estimate.

Instead of probabilities, confidence levels in percentages are usually used. That is, an interval may have a probability of 0.95 that it contains the mean. But rather than using such a probability, it is more common to report that the confidence level is 95% that the interval contains the mean.

Confidence levels are usually large values, such as 90%, 95%, or 99%, so that the researcher is quite confident that the interval is wide enough to contain the population mean.

Interval estimate for a mean, large sample size – Section 8.3

If a random sample of n cases are taken from a population with unknown mean μ, the issue addressed here is how to obtain an estimate of μ. The sample mean $\bar{X}$ is generally regarded as providing a good point estimate of μ. If the sample is a random sample from a population, the larger the sample size n, the closer $\bar{X}$ will be to the true population mean μ.

If the sample is a random sample and the sample size, $n > 30$, then the central limit theorem states that

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

From this, the distribution of the sample means behave according to the normal distribution, so that C% of the sample means lie within the intervals

$$\mu \pm Z_C \frac{\sigma}{\sqrt{n}}$$

where Z_C is the Z-value so that C% of the distribution is within $\pm Z_C$ of the population mean.

For example, if $C = 95$, so the researcher specifies a range such that 95% of samples are within this limit, then this is the 95% confidence interval. For 95% of the area under a normal distribution, the appropriate Z-values are ± 1.96. That is, 95% of the normal distribution is between $Z = -1.96$ and $Z = +1.96$. (For 95%, or 0.95, of the area within the middle portion of the normal distribution, look for an A area of $0.95/2 = 0.475$ and this is at $Z = 1.96$).

Now the above interval cannot be constructed, since μ is unknown. But what can be constructed is the interval

$$\bar{X} \pm Z_C \frac{\sigma}{\sqrt{n}}$$

and it can be shown that C% of intervals constructed in this manner contain the true mean μ. For $C = 95\%$ confidence, $Z_C = 1.96$ so the interval is

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}.$$

What this means is that 95% of the random samples selected from the population with mean μ will yield intervals that contain this mean. In other words, the researcher is 95% confident that the interval constructed will contain the population mean.

Unknown population standard deviation. One further complication is that the standard deviation σ of the population is generally not known prior to conducting the survey (see pp. 486-7). From the sample selected, the standard deviation of the sample, s, can be calculated. If the sample is a random sample, and if the sample size is reasonably large ($n \geq 30$), then many researchers are willing to use the sample standard deviation s as a reasonable estimate of σ in the above formula.

The steps involved in constructing a confidence interval of this sort are as follows.

Procedure for constructing an interval estimate of μ.

There is a series of five steps involved in constructing a confidence interval estimate for the mean. The population characteristic, or parameter, being estimated is μ, the true mean of the population being examined. Here I am assuming that the sample is a random sample drawn from the population with mean μ and standard deviation σ.

The steps used in obtaining an interval estimate are (p. 488 of text):

1. Calculate $\bar{X}$, s and n from the sample.

2. If n is large, say 30 or more, the sampling distribution of $\bar{X}$ is normal, using the central limit theorem of Chapter 7.

3. Determine the confidence level C, either by choosing a level, or using the confidence level given in the problem.

4. Use the table of the normal distribution to determine the Z-value associated with an area of C% in the middle of the normal distribution.

5. Put all these values in

$$\left(\bar{X} - Z\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X} + Z\frac{\sigma}{\sqrt{n}}\right)$$

and this is the C% confidence interval. If σ is unknown, as is usually the case, construct the interval by replacing σ with s. The interval estimates are thus

$$\left(\bar{X} - Z\frac{s}{\sqrt{n}}\ ,\ \bar{X} + Z\frac{s}{\sqrt{n}}\right)$$

The interval that results is the C% confidence interval. A researcher can be confident that C% of these intervals contain the true population mean μ.

Example – Smoking by income level

Statistics Canada's 1996 General Social Survey, *Cycle 11: Social and Communtiy Support* asked respondents about a variety of issues. Data from this survey, concerning the smoking behaviour of Saskatchewan adults at various income levels, are in Table 3.

1. For each of the two lowest income levels, obtain 95% interval estimates for the true mean number of cigarettes smoked daily by Saskatchewan adults who smoke (2 interval estimates).

2. Write a short note explaining whether the true mean number of cigarettes smoked daily differs among Saskatchewan adults in the three income groups.

Answer.

The following answers assume that the sample is a random sample of Saskatchewan adults.

Table 3: Statistics concerning number of cigarettes smoked daily for Saskatchewan adults who smoke, by household income

Income	n	$\bar{X}$	s
Under $20,000	76	16.11	8.31
$20 - 59,999	135	14.86	7.31
$60,000 plus	27	14.41	11.32
Total	238	15.21	8.16

1. **Under $20,000**. Let μ be the true mean number of cigarettes smoked daily by Saskatchewan adults who smoke, who are in the under $20,000 income group. A sample size of $n = 76$ cases is large (greater than 30), so the sample mean $\bar{X}$ has a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. s can be used as an estimate of σ since $n = 76$ is large. The 95% interval estimate is

$$\begin{aligned} \bar{X} \pm Z\frac{s}{\sqrt{n}} &= 16.11 \pm 1.96\ \frac{8.31}{\sqrt{76}} \\ &= 16.11 \pm 1.96\ \frac{8.31}{8.719} \\ &= 16.11 \pm 1.96 \times 0.953 \\ &= 16.11 \pm 1.57 \end{aligned}$$

   or (14.24 , 17.98).

   $20-60,000. Again, let μ be the true mean for the $20 - 60 income group. $n = 135$ and again by the central limit theorem, $\bar{X}$ is normally distributed and the interval estimate is

$$\begin{aligned} \bar{X} \pm Z\frac{s}{\sqrt{n}} &= 14.86 \pm 1.96\ \frac{7.31}{\sqrt{135}} \\ &= 14.86 \pm 1.96\ \frac{7.31}{11.619} \\ &= 14.86 \pm 1.96 \times 0.6291 \\ &= 14.86 \pm 1.23 \end{aligned}$$

   or (13.63 , 16.09).

Table 4: 95% Interval estimates of mean number of cigarettes smoked daily by income group

Income Group	Sample Mean	Interval
Under 20	16.11	(14.2 , 18.0)
20 - 60	14.86	(13.6 , 16.1)
60 plus	14.41	

2. Table 4 provides a summary of the interval estimates of the mean number of cigarettes smoked daily for the two income groups. The sample mean for the under $20,000 income group is greater than for the middle income group. However, the 95% interval estimates overlap rather considerably, and the intervals are wide enough so that the sample means for each group are within the range of each interval. Since the true population mean for each of the income groups is likely to be anywhere within the intervals, this means that the true means may not be all that different from each other – or there is not strong evidence that the true means for the two income groups differ. On the other hand, the progression of means, from the largest mean at lowest income, to lower mean for the middle income group, to the lowest mean for the high income group (in the original table), supports the view that the true mean may be highest for the lowest income group, and lower for other income groups.