Social Studies 201
Notes for November 12 and 15, 2004

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 1 summarizes the types of estimates examined in Chapter 8.

Table 1: 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	$\hat{P}$

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

> 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 confidence 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 is 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 $Z = \pm 1.96$).

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, $Z_C = Z_{95} = 1.96$ 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. Obtain $\bar{X}$, s and n from the sample. In problems in this class, these are ordinarily provided in the question. When working on a data set on the computer using SPSS, these can be obtained from the *Frequencies*, *Descriptives*, or *Means* procedures.

2. If the sample is a random sample and the sample size, n, is large – say 30 or more – the sampling distribution of the sample mean, $\bar{X}$, is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. This result comes from 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 4.

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 2: 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

$$\begin{aligned} \bar{X} \pm Z\frac{s}{\sqrt{n}} &= 16.11 \pm 1.96\ \frac{8.31}{\sqrt{76}} \\ &= 16.11 \pm 1.96\ \frac{8.31}{8.718} \\ &= 16.11 \pm 1.96 \times 0.953 \\ &= 16.11 \pm 1.87 \end{aligned}$$

   or (14.24 , 17.98). Rounded to one decimal, the 95% confidence interval for the mean is from 14.2 to 18.0 cigarettes smoked daily.

   $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). Rounded to one decimal, the interval estimate for the mean number of cigarettes smoked daily is from 13.6 to 16.1.

2. Table 3 provides a summary of the interval estimates of the mean number of cigarettes smoked daily for the two income groups. The interval estimate for the 60 plus group is left blank since the sample size ofr that group is only $n = 27$ (see Table 4).

Table 3: 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	

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.

Example continued

The following question comes from the above example, using the data in Table 4.

Question. Using the data in Table 4, obtain 95%, 90% and 99% interval estimates for the true mean number of cigarettes smoked daily by all

Saskatchewan smokers.

Table 4: 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

Answer

For each part of this question, the value that is to be estimated is μ, the true mean number of cigarettes smoked daily by Saskatchewan smokers. These answers also assume that the sample is a random sample of Saskatchewan adults who are smokers.

The sample of all Saskatchewan adults who smoke has a sample size of $n = 238$, a large sample size. Thus 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 $n = 238$ is large. The three interval estimates are as follows.

1. **95% interval estimate**. The 95% interval estimate is

$$\begin{aligned} \bar{X} \pm Z\frac{s}{\sqrt{n}} &= \bar{X} \pm 1.96\ \frac{8.16}{\sqrt{238}} \\ &= \bar{X} \pm 1.96\ \frac{8.16}{15.427} \\ &= \bar{X} \pm (1.96 \times 0.529) \\ &= \bar{X} \pm 1.04 \end{aligned}$$

For the sample mean of $\bar{X} = 15.21$, the interval is thus

$$\bar{X} \pm 1.04 = 15.21 \pm 1.04$$

or from 14.17 to 16.25. The 95% interval estimate for the mean for all Saskatchewan smokers is (14.2 , 16.2) cigarettes smoked per day.

2. **90% interval estimate**. For the 90% interval estimate, the procedure is the same as for the 95% estimate, but the Z-value changes. For 90% confidence, the Z-values are those associated with 90% of the area in the middle of the normal distribution. If there is 90% of the area in the middle of the normal distribution, this means $90/2 = 45$ per cent, or 0.4500, of the area on each side of centre. For an A area of 0.4500, the Z-value is 1.64 or 1.65 – in this case, the Z-value exactly half-way between these is $Z = 1.645$. The intervals are:

$$\begin{aligned} \bar{X} \pm Z\frac{s}{\sqrt{n}} &= \bar{X} \pm 1.645\ \frac{8.16}{\sqrt{238}} \\ &= \bar{X} \pm 1.645\ \frac{8.16}{15.427} \\ &= \bar{X} \pm (1.645 \times 0.529) \\ &= \bar{X} \pm 0.870 \end{aligned}$$

   For the sample mean of $\bar{X} = 15.21$, the interval is thus

$$\bar{X} \pm 1.04 = 15.21 \pm 0.87$$

   or 14.34 to 16.08. Rounded to one decimal, the 90% confidence interval for the mean is (14.3 , 16.1) cigarettes per day. This is very similar to the 95% interval, but is slightly narrower as a result of the smaller confidence level associated with this interval.

3. **99% interval estimate**. For the 99% interval estimate, the procedure is the same as for the previous two estimates, but again the Z-value changes. For 99% confidence, the Z-values are those associated with 99% of the area in the middle of the normal distribution. If there is 99% of the area in the middle of the normal distribution, this means $99/2 = 49.5$ per cent, or 0.4950, of the area on each side of centre. For an A area of 0.4950, the Z-value is 2.57 or 2.58 – in this case, the Z-value exactly half-way between these is $Z = 2.575$. The intervals are:

$$\begin{aligned} \bar{X} \pm Z\frac{s}{\sqrt{n}} &= \bar{X} \pm 2.575\ \frac{8.16}{\sqrt{238}} \\ &= \bar{X} \pm 2.575\ \frac{8.16}{15.427} \end{aligned}$$

$$\begin{aligned} &= \bar{X} \pm (2.575 \times 0.529) \\ &= \bar{X} \pm 1.362 \end{aligned}$$

For the sample mean of $\bar{X} = 15.21$, the interval is thus

$$\bar{X} \pm 1.04 = 15.21 \pm 1.36$$

or 13.85 to 16.57. Rounded to one decimal, the interval is (13.8 , 16.6), again similar to the 95% interval, but wider because of the higher confidence level.

The three interval estimates are summarized in Table 5.

Table 5: Interval estimates of mean number of cigarettes smoked daily by Saskatchewan smokers – three confidence levels

Confidence level	Interval
90%	(14.3, 16.1)
95%	(14.2, 16.2)
99%	(13.8, 16.6)

From these three interval estimates, it should be apparent that, for any given sample, the larger the confidence level, the wider the interval. A higher confidence level means that the researcher or analyst is more certain that any intervals contain the true mean of the population. In order to have this higher level of confidence, it is necessary to include a larger percentage of the area under the normal distribution. This means a larger Z-value and, for any given standard deviation and sample size, a wider interval. That is, for any given sample, a higher confidence level is associated with a wider interval. The following section contains some guidelines concerning choice of an appropriate confidence level.

Confidence level – see p. 493

There are a number of considerations about what is the proper confidence level for a researcher or analyst to select. While the selection of a confidence level is somewhat arbitrary, there are a number of guidelines and conventions about confidence levels. Some of these are as follows.

1. **Report a confidence level**. The first guideline is that an interval estimate must always have a confidence level associated with it, otherwise it is meaningless. For example, a statement such as "The interval estimate for mean income is from \$38,000 to \$42,000" is meaningless without a probability or confidence level associated with it. Each interval estimate obtained from a random sample of a population has a certain probability or confidence level associated with it – make sure that you report the confidence level.

2. **95% and other commonly used levels**. Confidence levels are generally large values, such as 90%, 95%, or 99%, representing large probabilities that the intervals contain the population mean, This is so that the researcher can be relatively certain that the interval contains the true population mean. A larger confidence level is associated with a larger Z-value, meaning that there is a greater probability that intervals

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

   contain the population mean.

   By far the most commonly level used level is the 95% confidence level. This is the same as the 19 in 20 times often reported for opinion polls ($19/20 \times 100 = 95\%$).

3. **Use the level requested**. If you are requested to report a particular confidence level on a problem set or examination, use the level requested. If no confidence level is requested, but you are expected to provide an interval estimate, the 95% level is always acceptable. As noted in items 4 and 5, other levels may be more appropriate.

4. **Comparison with other research**. If a particular confidence level has been used in a research report or journal article, and you wish to compare your results with this research, use the same level as was used in the other research reports.

5. **Health and safety issues**. Much social research, dealing with issues of general concern in the social world, is not as exact or demanding as are conditions related to personal health and safety. In the above

example concerning cigarette consumption, whether a researcher uses the 90%, 95%, or 99% confidence level is largely a matter of researcher preference. The connection between income levels and smoking may be interesting, but it is not a crucial and immediate matter of life and death. In contrast, issues such as ensuring the safety of a bridge crossed by thousands of people daily, or establishing whether a prescription drug is safe, are of great concern for individuals. Errors made in constructing a bridge, that would make it unsafe, could endanger the lives of many commuters crossing the bridge. Or an unsafe drug could cause serious health problems, increasing chances of death for users. In these latter cases, most users would consider probabilities of safety such as 95%, or even 99%, to be insufficient. Users would like to be 99.99999% sure of safety, or perhaps even more demanding. In these matters of close connection to personal health, or to life and death, research should be much more demanding, ensuring that interval estimates are such that use is safe, to within very demanding standards.

While the above provide some guidelines concerning appropriate confidence levels, each researcher may select a different confidence level, using his or her experience and considering the potential uses of the interval estimates.

One additional point to note is that with any given sample, a higher confidence level is associated with a wider interval. Once data have been produced, the researcher cannot simultaneously produce a narrower interval with a larger confidence level. Consider the interval

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

Once a sample mean, sample standard deviation, and sample size have been obtained, the only part of the formula the researcher can control is Z, and the determination of Z depends on the confidence level selected.

If a researcher wishes to obtain a narrower confidence interval (producing a more precise estimate of the population mean) with high confidence, the only means to do this is to obtain a larger sample size. If the researcher can obtain more cases from the population, then n is increased and the size of

$$\pm Z\frac{\sigma}{\sqrt{n}}$$

can be reduced. This may be difficult, or impossible, to accomplish, if the survey has already been completed. In that case, the values of $\bar{X}$, s, and n must be accepted and the only discretionary item for the researcher or analyst is to select the appropriate confidence level.

Next day – estimate of mean, small sample size. t-distribution.