**Social Studies 201**
**Notes for November 10, 2006**

**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 | $\mu$ | $\bar{X}$ |
| Variance | $\sigma^2$ | $s^2$ |
| Standard Deviation | $\sigma$ | $s$ |
| Proportion | $p$ | $\hat{p}$ |
| Percentage | $P$ | $\hat{P}$ |

In Table 1, the population values, or parameters, are characteristics of the population. However, these are generally unknown and the task of the researcher is to obtain data so a conclusion can be made concerning the possible population value. In order to do this, a researcher conducts a survey, sample, or experiment and, from this, obtains statistics. That is, the first step in obtaining an estimate of a characteristic of a population, such as the

mean of the population or the proportion or members with a particular characteristic, 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. In the case of a researcher attempting to determine the mean of a population, 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 he or she constructs actually contains the population mean. But the researcher can construct an interval estimate and attach a probability to the interval estimate – the probability that the interval estimates 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 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 is taken from a population with unknown mean $\mu$, the issue addressed here is how to obtain an estimate of $\mu$. The sample mean $\bar{X}$ is generally regarded as providing a good point estimate of $\mu$. 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 $\mu$.

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 $\pm 1.96$. 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 $\mu$ 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 $\mu$. 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 $\mu$ 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 $\sigma$ 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 $\sigma$ 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 $\mu$.**

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 $\mu$, 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 $\mu$ and standard deviation $\sigma$.

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 $\mu$ 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 $\sigma$ is unknown, as is usually the case, construct the interval by replacing $\sigma$ 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 $\mu$.

**Example – Age at first marriage**

Statistics Canada's 2001 General Social Survey, *Cycle 15: Family History* asked respondents about a variety of family history issues, as well as personal and family characteristics. Data from this survey, concerning the age at first marriage for Saskatchewan males and females, are in Table 2.

1. For each of males and females, obtain 95% interval estimates for the true mean age at first marriage for all Saskatchewan males and females, respectively. (2 interval estimates).

2. Write a short note explaining whether the true mean age for males and females differ.

Table 2: Statistics of age at first marriage, Saskatchewan, 2001

| Sex | n | $\bar{X}$ | $s$ |
|---|---|---|---|
| Male | 321 | 27.52 | 8.07 |
| Female | 352 | 24.67 | 7.68 |

**Answer**

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

1. **Males**. Let $\mu$ be the true mean age at first marriage for all Saskatchewan males in 2001. From Table 2, the sample size is $n = 321$ and the test statistic is $\bar{X}$. Given this large sample size (well over 30 cases), $\bar{X}$ has a normal distribution with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$, where the sample standard deviation $s = 8.07$ is used as an estimate of $\sigma$. For 95% confidence, the $Z$ value is 1.96 and the 95% interval estimate is:

$$\begin{aligned} \bar{X} \pm Z\frac{\sigma}{\sqrt{n}} &= \bar{X} \pm 1.96\frac{8.07}{\sqrt{321}} \\ &= \bar{X} \pm 1.96\frac{8.07}{17.916} \\ &= \bar{X} \pm (1.96 \times 0.450) \\ &= \hat{X} \pm 0.883 \\ &= 27.52 \pm 0.88 \end{aligned}$$

   or from 26.6 to 28.4 years (rounded to the nearest tenth of a year). The interval estimate can be written as (26.6 , 28.4) years.

   **Females**. For females, the method is the same. Again the sample size $n = 352$ is large, so the sample mean has a normal distribution. Also, since $n$ is large, $s$ is used as an estimate of $\sigma$, and the 85% interval estimate is

$$\bar{X} \pm Z\frac{\sigma}{\sqrt{n}} = \bar{X} \pm 1.96\frac{7.68}{\sqrt{352}}$$

$$\begin{aligned} &= \bar{X} \pm 1.96\frac{7.68}{18.762} \\ &= \bar{X} \pm (1.96 \times 0.409) \\ &= \hat{X} \pm 0.802 \\ &= 24.67 \pm 0.80 \end{aligned}$$

The interval estimate is (23.9 , 25.5) years, or from 23.9 to 25.5 years (again rounded to the nearest tenth of a year).

2. Table 3 provides a summary of the interval estimates for the mean age at first marriage for all Saskatchewan males and females. All the ages have been rounded to the nearest tenth of a year.

Table 3: 95% interval estimates of mean age at first marriage for Saskatchewan males and females, 2001

| Income Group | Sample Mean | Interval |
|---|---|---|
| Males | 27.5 | (26.6 , 28.4) |
| Females | 24.7 | (23.9 , 25.5) |

From the sample means, it appears that males are approximately three years older than females when they first marry. However, these are only the sample values, so that it is necessary to look at the interval estimates before concluding that the age at first marriage for all males exceeds that for all females.

For males, the 95% interval estimate is that the true mean age at first marriage is between 26.6 and 28.4 years. For females, the 95% interval estimate is from 23.9 to 25.5 years. Since these intervals do not overlap, a researcher can be quite certain that the mean age at first marriage for all Saskatchewan males is greater than that for all females. While there is some small probability that this conclusion is not correct, it is very small, since there is a 0.95 probability that each of the interval estimates contain the respective true mean.