Social Studies 201
Notes for November 12, 2003

Example continued

The following questions come from the example and method used in the last class – see the notes for November 10.

Question. Using the data in Table 1, obtain 95%, 90% and 99% interval estimates for the true mean number of cigarettes smoked daily by all Saskatchewan smokers.

Table 1: 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

$$\bar{X} \pm Z\frac{s}{\sqrt{n}} \quad = \quad \bar{X} \pm 1.96\ \frac{8.16}{\sqrt{238}}$$

$$\begin{aligned} &= \bar{X} \pm 1.96 \, \frac{8.16}{15.427} \\ &= \bar{X} \pm (1.96 \times 0.529) \\ &= \bar{X} \pm 1.04 \end{aligned}$$

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

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

or (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 (14.3 , 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:

$$\begin{aligned} \bar{X} \pm Z\frac{s}{\sqrt{n}} &= \bar{X} \pm 2.575\ \frac{8.16}{\sqrt{238}} \\ &= \bar{X} \pm 2.575\ \frac{8.16}{15.427} \\ &= \bar{X} \pm (2.575 \times 0.529) \\ &= \bar{X} \pm 1.362 \end{aligned}$$

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

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

or (13.8 , 16.6), again similar to the 95% interval, but wider because of the higher confidence level.

The three interval estimates are summarized in Table 2.

Table 2: 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 social life, 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 the safety of a bridge crossed by thousands of people daily or whether a prescription drug is safe for users 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.