Social Studies 201
Notes for March 18, 2005

Estimation of a mean, small sample size – Section 8.4, p. 501.

When a researcher has only a small sample size available, the central limit theorem does not apply to the distribution of sample means. In this case, if certain assumptions are made, the t-distribution can be used to describe the distribution of sample means. From this, an interval estimate of the population mean μ can be constructed.

The t-distribution

The t-distribution is sometimes referred to as Student's t-distribution. A table of the t-distribution is contained in Appendix I, p. 911, of the text. This distribution has a shape that is very similar to that of the normal distribution, and has the same interpretation and use as the normal distribution in that is is symmetrical about the centre, peaked in the centre, and trailing off toward the horizontal axis in each direction from centre.

The t-distribution has a mean of 0 and a standard deviation of 1 – the same as in the case of a standardized normal distribution. In the same way that Z-values are positions along the X-axis, so the t-values are measured along the horizontal or X-axis. Since the mean is 0 and the standard deviation is 1, the t-value associated with each point is also the number of standard deviations. For example, a t-value of 1.50 is associated with a point on the horizontal axis 1.5 standard deviations to the right of centre.

One difference between the t and normal distributions is that the t-distribution is a little more spread out than the normal. One way of picturing the t distribution is to imagine taking the normal distribution and stretching it to the left and right. See the diagram on p. 503 of the text, where one distribution is superimposed on the other.

> **Degrees of freedom**. Another difference between the t and normal distribution, is that there is a different t-distribution for each degree of freedom (df) – a new concept that is related to sample size. As a concept, degrees of freedom is a little difficult

> to explain at this stage of the course – it refers to how many sample values are free to vary and how many are constrained. In the case of estimation of a mean, there are $n - 1$ degrees of freedom, one less than the sample size of n. When estimating a mean from n sample values, any $n - 1$ values are free to vary but one value is fixed or constrained, by the fact that a particular value of a mean must result. If you find this confusing, for now just accept that in estimating the mean the degrees of freedom is the sample size minus one, that is, $df = n - 1$.

To return to the t-distribution, when there are few degrees of freedom, the distribution is very dispersed. For example, when the degrees of freedom are only 4, the middle 95% of the t-distribution requires including the area from -2.776 to +2.776. This is in contrast to the corresponding Z-value of ± 1.96, from the normal distribution.

But as the number of degrees of freedom increases, the t-distribution approaches the normal distribution. Going down the column of the t-table (p. 911) associated with the 95% confidence level, if there is a sample size of 25, meaning $df = n-1 = 25-1 = 24$ degrees of freedom, the t-value is 2.064, considerably less than the 2.776 for 4 degrees of freedom. As the sample size, and the corresponding degrees of freedom, becomes larger, the t-distribution actually approaches the normal distribution. To see this, examine the last row of the t-table. For a very large degrees of freedom (labelled infinite), the t-value for 95% confidence is 1.96, exactly the same as the corresponding Z-value from the table of the normal distribution.

For most purposes, when the sample size reaches 30, we use the normal distribution. For 29 degrees of freedom and 95% confidence, the t-value is 2.045, not much larger than the 1.96 associated with the normal distribution.

The table of the t-distribution on p. 911 lists various confidence levels across the top. You are thus restricted to obtaining confidence intervals for the confidence levels listed there. But the table provides the t-values associated with common confidence levels such as 80%, 90%, 95%, and 99%. To use the table, pick the proper confidence level and the associated degrees of freedom (sample size minus 1) and the t-values in the table provide the associated area under the t-distribution between the t-value and the negative of that t-value.

Distribution of the sample mean, small n

See text, p. 507.

Under certain assumptions, the t-distribution can be used to obtain interval estimates for the mean of certain distributions. This section outlines the conditions for this.

Strictly speaking, the t-distribution can only be used if sample is drawn from a normally distributed population. That is, if a researcher has some assurance that the characteristic of the population being examined is normally distributed, then small random samples from this population have a t-distribution. This result can be stated as follows.

> Suppose a normally distributed population has a mean of μ and a standard deviation of σ. If random samples of small sample size n (less than 30 cases) are drawn from this population, then the sample means $\bar{X}$ of these samples have a t-distribution with mean μ, standard deviation $s/\sqrt{n}$, and $n-1$ degrees of freedom, where s is the standard deviation obtained from the sample.

This can be stated symbolically. If

$$X \text{ is Nor } (\mu, \sigma)$$

where μ and σ are unknown, and if random samples of size n are drawn from this population,

$$\bar{X} \text{ is } t_d \left(\mu, \frac{s}{\sqrt{n}}\right).$$

where $d = n - 1$ is the degrees of freedom and $\bar{X}$ and s are the mean and standard deviation, respectively, from the sample.

When the sample size n is small, say less than $n = 30$, the t distribution should be used. If $n > 30$, then the t values become so close to the standardized normal values Z that the Central Limit Theorem can be used to describe the sampling distribution of $\bar{X}$. That is,

$$t \to Z \text{ as } n \to \infty.$$

This means that the t distribution is likely to be used only when the sample size is small. For larger sample sizes, $\bar{X}$ may still have a t distribution, but if

the sample size is large enough, the normal values are so close to the t values that the normal values are ordinarily used.

There are two assumptions associated with this result.

1. As is the case with larger sample size, the samples are to be random samples from the population. If the samples are not random samples, it is difficult to determine what the distribution of the sample means might be.

2. Unlike the central limit theorem, which generally holds regardless of the nature of the distribution of the variable, the t-distribution requires sampling from a normally distributed population. This is quite a restrictive assumption, since few populations are likely to be exactly normally distributed. In practice, the t-distribution is often used even when there is no assurance that the sample is drawn from a normally distributed population. If a researcher considers the population to be very different than normally distributed, perhaps the t-distribution should not be used. But if the population is not distributed all that differently from a normal distribution, little error may be introduced by using the t-distribution. I generally argue that, in the case of sample sizes less than 30, it is always better to use the t-distribution than the normal distribution, when conducting interval estimates or hypothesis tests. The reason for this is that the t-distribution is more dispersed than the normal distribution, so it gives a better picture of the precision of the sample. If a normal distribution is used, interval estimates may be reported as narrower than they really are in practice. By using the t-distribution, a researcher is less likely to make results look more precise than they really are.

Interval estimate for the mean – small sample size

The t distribution for $\bar{X}$ can be used to obtain interval estimates of a population mean μ. The method is the same for this small sample method as it is for the large sample method. That is, the same series of five steps can be used.

If the population mean μ is to be estimated, and the sample is a random sample of size n with sample mean $\bar{X}$ and sample standard deviation s, and

if the population from which this sample is drawn is normally distributed, then

$$\bar{X} \text{ is } t_d\left(\mu, \frac{s}{\sqrt{n}}\right).$$

where $d = n - 1$.

In order to obtain an interval estimate, the researcher picks a confidence level C% and uses this to determine the appropriate t-value from the t-table on p. 911. For d degrees of freedom, let t_d be the t value such that C% of the area under the t curve lies between $-t_d$ and t_d. The C% confidence interval is then

$$\bar{X} \pm t_d \frac{s}{\sqrt{n}}$$

or in interval form,

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

Note that this is the same formula as for the confidence interval when the sample size is large – the only difference is that t_d replaces the Z-value. Note that the sample standard deviation s is used in the formula, rather than σ. The latter was used when presenting the formula for the interval estimate in the case of the large sample size. But even in the case of a large sample size, σ is almost always unknown, so that in practice s is used as an estimate of σ in that formula.

The interpretation of the confidence interval estimate is also the same as earlier. That is, C% of the the intervals

$$\bar{X} \pm t_d \frac{s}{\sqrt{n}}$$

contain μ if random samples of size n are drawn from the population, where $d = n - 1$. Any specific interval which is constructed will either contain μ or it will not contain μ, but the researcher can be confident that C% of these intervals will be wide enough so that μ will be in the interval.

Example – wages of workers after plant shutdown

In "Bringing 'Globalization' Down to Earth: Restructuring and Labour in Rural Communities" in the August, 1995 issue of the *Canadian Review of Sociology and Anthropology*, the authors Belinda Leach and Anthony Winson examine changes in wages of workers after a plant shutdown. Before shutdown, mean male wages were \$13.76 per hour and mean female wages were \$11.80 per hour. After shutdown, some of the workers found new jobs and the data from small samples of such workers is contained in Table 1. Using data in this table, obtain 95% interval estimates for the true mean wages of (i) male workers after the plant shutdown, and (ii) female workers after the plant shutdown. From these interval estimates comment on whether there is strong evidence that the wages of male and female workers have declined.

Table 1: Data on Hourly Wages of Workers with Jobs, After Plant Shutdown

Type of Worker	Hourly Wage in Dollars Mean	St. Dev.	Sample Size
Male	12.20	3.27	12
Female	8.11	3.53	12

Answer

For the first part, the parameter to be estimated is μ, the true mean wage for all male workers who lost jobs because of the plant shutdown. Organizing the answer in terms of the five steps involved in interval estimation (see notes of November 10), the answer is as follows.

1. The sample mean $\bar{X}$, standard deviation s, and sample size n are given in Table 1. The sample size of $n = 12$ is small in this case, so it will be necessary to use the t-distribution.

2. Assuming the distribution of wages of all male workers who lost jobs in the shutdown is a normal distribution, the distribution of $\bar{X}$ is a t-distribution with mean μ and standard deviation $s/\sqrt{n}$ with $n - 1 =$

$12 - 1 = 11$ degrees of freedom. That is

$$\bar{X} \text{ is } t_d\left(\mu, \frac{s}{\sqrt{n}}\right).$$

where $d = n - 1 = 11$.

3. From the question, the confidence level is $C = 95\%$.
4. For 11 degrees of freedom and 95% confidence level, the t-value is 2.201.
5. The intervals are
$$\bar{X} \pm t_d \frac{s}{\sqrt{n}}$$
and using values from this sample, the intervals are
$$\bar{X} \pm 2.201 \frac{3.27}{\sqrt{12}}$$
$$\bar{X} \pm 2.201 \frac{3.27}{3.464}$$
$$\bar{X} \pm (2.201 \times 0.944)$$
$$\bar{X} \pm 2.078$$
For $\bar{X} = 12.20$, the interval is
$$12.20 \pm 2.078$$
Thus the 95% interval estimate for the true mean wage level for all males who have lost jobs because of the plant shutdown is ($10.12, $14.28).

For the female workers, the same steps yield the intervals

$$\bar{X} \pm 2.201 \frac{3.53}{\sqrt{12}}$$
$$\bar{X} \pm 2.201 \frac{3.53}{3.464}$$
$$\bar{X} \pm (2.201 \times 1.019)$$
$$\bar{X} \pm 2.243$$

For $\bar{X} = 8.11$, the interval is

$$8.11 \pm 2.243$$

Thus the 95% interval estimate for the true mean wage level for all females who have lost jobs because of the plant shutdown is $(\$5.87, \$10.35)$.

Comment on results. From the data in Table 1, the samples provide evidence that the hourly wages of both male and female workers has declined since the plant shutdown. The twelve males in the sample had a mean wage of \$12.20 after the shutdown, \$1.56 per hour less than the \$13.76 they were earning prior to the shutdown. On average, the twelve female workers suffered a decline of \$3.69 per hour, from \$11.80 prior to the shutdown to \$8.11 after the shutdown.

The interval estimates provide fairly strong evidence that all female workers suffered a decline in hourly wages, while the evidence for a decline is not so clear in the case of males. Consider first the female interval estimate. There is 95% certainty that the sample means from a sample of size twelve yields differ from the true mean by no more than \$2.24. In the case of this sample, the interval is from \$5.87 to \$10.35. While a researcher cannot be certain this interval contains the true mean hourly wage for females after the plant shutdown, it very likely does. But this interval lies well below the former mean hourly wage of \$11.80 per hour. As a result, it seems fairly certain that female wages after shutdown are lower than prior to the shutdown.

In contrast, the decline in male wages was less than that for females and the 95% estimate for this sample yields an interval for the males that contains the previous mean pay of \$13.76. Since the researcher is relatively certain that the true mean male hourly wage is in the interval from \$10.12 to \$14.28, it is possible that the true mean for all male workers is around the former mean hourly wage of \$13.76. While the sample mean is less than the previous mean, it is not a lot less – there is thus weak evidence for a decline of male hourly wages but the evidence is not as strong as in the female case.

The interval estimates do not provide direct tests of whether mean wages have changed or not – that will be provided later in the section on hypothesis testing. But the results of hypothesis tests will be shown to be consistent with the comments above – that is, there is evidence that female wages declined but insufficient evidence to prove that male wages declined.

Small and large sample sizes

Small samples generally result in fairly wide confidence intervals, thus providing less precise estimates of the mean than do larger samples. Comparing the intervals for small samples

$$\bar{X} \pm t_d \frac{s}{\sqrt{n}}$$

with those from larger samples

$$\bar{X} \pm Z \frac{s}{\sqrt{n}}$$

there are two differences that contribute to this.

1. If the sample has a small sample size, the t-distribution must be used, rather than the normal distribution. As noted earlier, and as can be seen by comparing the t-values with the corresponding normal values, t-values are always larger than the corresponding Z-values for any given confidence level. The $\pm$ section of the interval is thus larger for small n, since the larger t-value, rather than the smaller Z-value, must be used.

2. When the sample size is smaller, the square root of the sample size is also smaller. Since the square root of the sample size is smaller in this case, the denominator of $s/\sqrt{n}$ is smaller, so the fraction as a whole is larger.

Table 2 illustrates the differences between a small and a large sample size for samples from a hypothetical population. As in the case of the last example, a sample standard deviation of $s = 12$ is hypothesized for each of these samples. For the sample of size 16, the intervals are $\bar{X} \pm 6.4$, while for the sample of size 256, the intervals are $\bar{X} \pm 1.3$. From the table, it can be seen that this large difference in interval width emerges from the two factors mentioned above. For the small sample, the t-value of 2.131 is larger than the Z-value of 1.96 for the larger sample. In addition, the standard error, or standard deviation of the sample means is 3 in the case of the small sample and only 0.667 in the case of the large sample.

Table 2: Precision of estimates for 95% intervals, small sample of size 16 and large sample of size 256, with a sample standard deviation of $s = 12$

Size of sample	Sample size (n)	Characteristics of samples $\sqrt{n}$	$s/\sqrt{n}$	$t(s/\sqrt{n})$ or $Z(s/\sqrt{n})$
Small	16	4	$12/4 = 3$	$2.131 \times 3 = 6.393$
Large	256	16	$12/16 = 0.667$	$1.96 \times 0.667 = 1.31$

Given these different sized intervals that can emerge from different samples, the next section is devoted to determining appropriate sample size prior to the sample being selected.

Last edited March 18, 2005.