

Social Studies 201**Notes for November 17 and 19, 2004****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 it is symmetrical about the centre, peaked in the centre, and trailing off toward the horizontal axis in each direction from centre. Like the normal distribution, the t-distribution has a mean of 0 and a standard deviation of 1. t-values are measured along the horizontal axis, and with a standard deviation of 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 from the normal distribution is that the t-distribution is a little more spread out than the normal – one way of picturing it is to take the normal distribution and pull it out a bit at each end. See the diagram on p. 503 of the text, where one distribution is superimposed on the other.

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

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.

If a population has a mean of μ and a standard deviation of σ , and if small random samples of 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,

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

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		Sample Size
	Mean	St. Dev.	
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: 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	$12/16 = 0.667$	$1.96 \times 0.667 = 1.31$

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.

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.

Sample size for estimating a population mean

By using the central limit theorem, prior to obtaining the sample, it is possible to specify the sample size required to achieve a given degree of accuracy for an estimate of the mean. In addition, the confidence level for the interval estimate must be specified and the researcher must have some knowledge of the variability of the population from which the sample is drawn. Since a larger sample size generally takes more time and effort, costs more, and may disturb the population more, a researcher wants to select the smallest possible sample size consistent with obtaining the required accuracy and confidence level. But this may be quite a large sample size, and this section provides the rationale for determining this sample size. Such a sample must also be a random sample – other methods of sampling may be associated with different required sample sizes.

Since the required sample size is usually large, the central limit theorem can be used to describe the distribution of sample means. This provides the researcher with a way to determine the variability of sample means prior to obtaining the sample. From this theorem, the normal distribution is generally used, rather than the t-distribution, since the sample size that will be required is usually larger than 30.

Notation. The required accuracy for the estimate is denoted by E , so that the interval constructed will be $\bar{X} \pm E$ after the sample data have been obtained. Note that this is an interval of $\pm E$ on either side of the sample mean \bar{X} , so the interval width is $W = 2E$. The letter E is used here to denote “error,” that is, sampling error. This amount E is equivalent to the sampling error of the sample. As before, the confidence level is $C\%$, with the corresponding value from the normal table given the symbol Z_C . That is, $\pm Z_C$ are the Z -values such that $C\%$ of the distribution is between them.

Derivation of the formula for sample size

If the mean μ of a population is to be estimated, the central limit theorem describes the distribution of sample means \bar{X} . The theorem states that when random samples are drawn from a population with mean μ and standard

deviation σ ,

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

when a random sample from the population has more than thirty cases. As noted earlier (Table 2) and as can be seen by examining the formula, a larger n produces a smaller standard error (standard deviation of the mean) than does a smaller sample size. That is, there are different normal distributions for each different sample size. The trick is to select the normal distribution that will result in an interval estimate of required accuracy E , that is, an interval of $\bar{X} \pm E$.

Since the sample means are normally distributed, for confidence level C , the corresponding Z -value from the normal distribution is Z_C . $C\%$ of the normal distribution lies within Z_C standard deviations of the true mean μ . Since one standard deviation for the distribution of sample means is σ/\sqrt{n} , this means that Z_C standard deviations amount to

$$Z_C \frac{\sigma}{\sqrt{n}}.$$

From this, it should be possible to see that, for the distribution of sample means (with mean μ and standard deviation σ/\sqrt{n}), $C\%$ of the area under the distribution is within the interval from

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

$C\%$ of the sample means \bar{X} also lie within these limits.

Now if the interval estimate is to be accurate to within $\pm E$, this means that the sample mean must be within E of the population mean μ . What we need to match the accuracy of the estimate with the interval of the previous paragraph is a sample size such that:

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

and

$$\mu \pm E$$

match. This occurs when:

$$E = Z_C \frac{\sigma}{\sqrt{n}}.$$

This can be obtained by solving this expression for n . Rearranging and solving this expression for n (see p. 533 of the text), gives

$$n = \frac{Z^2 \sigma^2}{E^2} = \left(\frac{Z\sigma}{E} \right)^2$$

where I have simply used Z , rather than Z_C . If a random sample of the size specified by this formula is obtained, then the confidence interval estimate obtained by the researcher should be of accuracy E , that is, it should be approximately $\bar{X} \pm E$.

In order to see that it is practical to use this formula, all the terms on the right side of this equation can be obtained prior to the sample being obtained. That is, the accuracy of the estimate desired, E , can be specified by the researcher prior to conducting the sample. The Z -value can be determined from the table of the normal distribution once a confidence level is given. Finally, a researcher has to have some estimate of the variability of the population from which the sample is to be drawn, so that σ can be specified in the formula. Some guidelines concerning this are contained later in these notes, and on pp. 538-9 of the text.

Example – sample size to estimate mean wage

Using the data above concerning the wages of workers who lost their jobs because of a plant shutdown, the problem was that with samples of size 12, the intervals were fairly wide, just over $\pm \$2$ for each of male and female workers. In this example, the sample size required in order to determine the mean wage correct to within (a) one dollar, and (b) fifty cents will be determined. For this example, the 95% confidence level is used.

Answer

As with any problem of this sort, the first step is to be clear concerning what is to be estimated. In this case, the parameter to be estimated is μ , the current mean wage of all workers who lost jobs because of the plant shutdown. Since this is an estimate of a mean, and since we expect the required sample size to be reasonably large, the central limit theorem can be used to describe the distribution of sample means. As shown above, from this, the required sample size is

$$n = \left(\frac{Z\sigma}{E} \right)^2$$

where E is the accuracy required of the estimate.

For the first part of the example, the accuracy required is one dollar, or $E = 1$. If 95% confidence is to be used, this means $Z = 1.96$, since 95% of the area under a normal distribution lies between $Z = -1.96$ and $Z = +1.96$. While the true standard deviation of hourly wages for all workers is not known, the small samples from Table 1 provide an idea of the variability of hourly wages. Since female wages vary more than male wages, the female standard deviation of $s = 3.53$ will be used to provide an estimate of σ in the formula for sample size. That is, the standard deviation from the more variable group will be used, to ensure that a large enough sample size is obtained.

From these values the determination of sample size is

$$\begin{aligned} n &= \left(\frac{Z\sigma}{E} \right)^2 \\ n &= \left(\frac{1.96 \times 3.53}{1.00} \right)^2 \\ n &= 6.9188^2 = 47.87 \end{aligned}$$

or a sample size of $n = 48$. A random sample of $n = 48$ workers should provide an estimate of mean hourly wage correct to within $\pm \$1$, with probability 0.95 or 95% confidence.

For an interval estimate accurate to within fifty cents, or $E = \$0.50$, the same formula is used, but with $E = 0.50$ replacing $E = 1$. The sample size is

$$\begin{aligned} n &= \left(\frac{Z\sigma}{E} \right)^2 \\ n &= \left(\frac{1.96 \times 3.53}{0.50} \right)^2 \\ n &= 13.8376^2 = 191.479 \end{aligned}$$

or a sample size of $n = 192$. A random sample of $n = 192$ workers should provide a sample mean \bar{X} that differs from the population mean μ by no more than \$0.50.

Additional notes on sample size

1. **Round up.** In the above example, where there were decimals for the sample size n , these were always rounded up to the next integer when reporting the required sample size. In order to specify a large enough sample size, the answer should always be rounded up to the next integer.
2. **Units.** When using the formula above, make sure that E and the estimate of σ are in the same units. In the example above, everything was converted into dollars, to ensure consistency.
3. **Trade-off.** There is often a trade-off between the budget for a survey and the accuracy of the results. A larger sample size produces greater accuracy but this may cost much more and take much more time and effort. As a result, a researcher may not be able to obtain the sample size specified by the formula, and may have to live with the less accurate results from a sample size smaller than desired.
4. **Factors associated with larger required n .** A careful look at the formula

$$n = \left(\frac{Z\sigma}{E} \right)^2$$

shows that the required sample size, n , increases as Z increases, σ increases, and E decreases. This can be summarized as follows:

- (a) A larger confidence level, $C\%$, produces a larger Z -value and results in a larger required sample size.
 - (b) A more variable population, with larger σ , means a larger sample size is required to achieve the given level of accuracy. In contrast, populations where members are similar to each other in the characteristics being examined, do not require such large sample sizes to achieve the required accuracy of estimate.
 - (c) The greater the accuracy required, the smaller is the value of E , and the larger the required sample size.
5. **Population size not important.** The required sample size does not depend on the size of the population from which the sample is

being drawn, unless the required sample size is a large proportion of the population. Suppose the above formula leads to a sample size of 200, but the population size is 10,000 people. Then a random sample from this population gives the accuracy required. If the size of the population is 100,000, or one million, the sample size is the same – a random sample with a size of $n = 200$ is required in each case.

The only exception to this is when the population is relatively small. Say the population size is 1,000, so the sample size recommended is $200/1,000 = 0.2$, or 20%, of the population size. In this case, the required sample size may be reduced somewhat, since the sample size is a considerable portion of the total population. But if the sample size is less than, say, 5% of the population size, the above formula holds. The reason for this apparent paradox are within probability theory – consult a text on the mathematical principles involved in sampling if you are interested in this issue.

6. **Estimate of σ .** In order to calculate required sample size, some estimate of σ , the variability of the population, is required. Some methods of obtaining a prior estimate of σ are as follows (see pp. 538-9 for a fuller discussion).
 - (a) **Small sample.** As in the above example, a researcher may have a small sample and, from this, an initial idea of the variability of the population from which a larger sample is to be drawn.
 - (b) **Other studies and other populations.** Other researchers may have obtained surveys from a population, or similar population, and the sample standard deviations from these surveys may provide a reasonable estimate of σ .
 - (c) **Range.** Recall that the standard deviation may be close to one-quarter of the range for a variable. This is not very exact but, in the absence of much knowledge of the variability of a population, may provide a quick and rough estimate of σ .
 - (d) **Sampling method.** It may be possible to devise a sampling method so the sample size can be increased later, after the initial sample is drawn. That is, conduct an initial random sample and if the results are not accurate enough, randomly select more cases.