9.2: Estimating a Population Mean

Recall: A *parameter* is a numerical summary of a population; a *statistic* is a numerical summary of a sample. (For example, the population mean and population standard deviation are parameters; the sample mean and sample standard deviation are statistics.)

Definition: A *point estimate* is the value of a statistic that estimates the value of a parameter.

Because it is usually unrealistic to measure or observe the entire population of interest, we use samples to gain information about the population. It seems reasonable to use a sample statistic to estimate a population parameter. However, we would not expect the sample statistic to exactly match the population parameter. How close should we expect them to be?

Confidence intervals:

Definition: A confidence interval for an unknown parameter is an interval of numbers generated by a point estimate for that parameter.

Definition: The *confidence level* (usually given as a percentage) represents how confident we are that the confidence interval contains the parameter.

If a large number of samples is obtained, and a separate point estimate and confidence interval are generated from each sample, then a 95% confidence level indicates that 95% of all these confidence intervals contain the population parameter.

A confidence interval is obtained by placing a *margin of error* on either side of the point estimate of the parameter.

In other words, the confidence interval consists of: Point estimate $\pm$ margin of error

Point estimate for the population mean:

The point estimate of the population mean μ is the sample mean $\bar{x}$.

So, for every sample, the sample mean will be in the center of the confidence interval. If we use E to indicate the margin of error, the confidence interval is $\bar{x} \pm E$, or $(\bar{x} - E, \bar{x} + E)$

Simulations:

http://rpsychologist.com/d3/CI/
(Created by Kristoffer Magnussen; permission to use granted by Creative Commons License)

http://onlinestatbook.com/stat_sim/conf_interval/index.html
(Rice Virtual Lab in Statistics; public domain resource partially funded by the National Science Foundation; creation led by David Lane of Rice University)

Example 1: Suppose $(125,138)$ is the 95% confidence interval for μ generated by a sample. Find the sample mean $\bar{x}$ and the margin of error *E*.

$\bar{x} = \frac{125+138}{2} = 131.5$

margin of error $= E = 138 - 131.5 = 6.5$

Because the margin of error on each side of $\bar{x}$ will be the same, we should be able to write the confidence interval as $\left(\bar{x} - z_c\sigma_{\bar{x}}, \bar{x} + z_c\sigma_{\bar{x}}\right)$, where $\sigma_{\bar{x}}$ is the standard deviation of the sampling distribution of the sample means, and z_c is a multiplier that tells us how many standard deviations (of the sampling distribution of the sample means) lie between the sample mean $\bar{x}$ and the edge of the confidence interval. We call this z_c the *critical value* for a *z*-score in the sampling distribution of the sample means.

Recall: The standard deviation of the sampling distribution of the sample means is called the *standard error*. It is calculated by dividing the population standard deviation by the square root of the sample size.

$$\text{Standard error: } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Problem: We typically do not know the population standard deviation, σ.

Our only option is to use the sample standard deviation, *s*, to estimate the population standard deviation. However, the sample standard deviation will generally be larger than the population standard deviation.

From the Central Limit Theorem, we know that the *z*-score of $\bar{x}$, $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is normally distributed.

provided *n* is sufficiently large.

However, $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ is NOT normally distributed (although for very large sample sizes, it approaches normality).

The Student *t*-distribution:

William S. Gosset (1876-1937) was a mathematician and chemist who worked for the Guinness Brewery in Dublin, Ireland. He discovered that using the sample standard deviation resulted in incorrect confidence intervals. He showed that $(\bar{x}-\mu)/(s/\sqrt{n})$ did not follow a normal distribution, but instead followed a different distribution, which eventually became known as the *t*-distribution. The brewery had very tight restrictions on what its scientists could publish; Gosset obtained permission to publish his results, but he had to use a pseudonym: Student.\

More information on William Gosset, known as Student:
http://blogs.sas.com/content/jmp/2013/10/07/celebrating-statisticians-william-sealy-gosset-a-k-a-student/
http://scepticemia.com/2012/09/21/william-gosset-a-true-student/
http://www-history.mcs.st-and.ac.uk/Biographies/Gosset.html

Student's *t*-distribution:

If a random sample of size *n* is drawn from a normal population, the distribution of

$$t=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

follows the *t*-distribution with $n-1$ degrees of freedom. The set of *t*-distributions is a family of distributions with the following properties:

- Changing the degrees of freedom changes the distribution.
- Area under the curve is 1.
- Distribution is symmetric about 0.
- The distribution is bell-shaped, but with more area in the tails than the normal distribution.
- As the number of degrees of freedom increases, the *t*-distribution more closely resembles the standard normal distribution.

Areas under intervals of the *t*-distribution can be found in Table IV, on page A-13.

Figure from
http://onlinestatbook.com/2/estimation/t_distribution.html

Home page:
http://onlinestatbook.com/2/index.html

Primary author and editor:
David Lane of Rice University

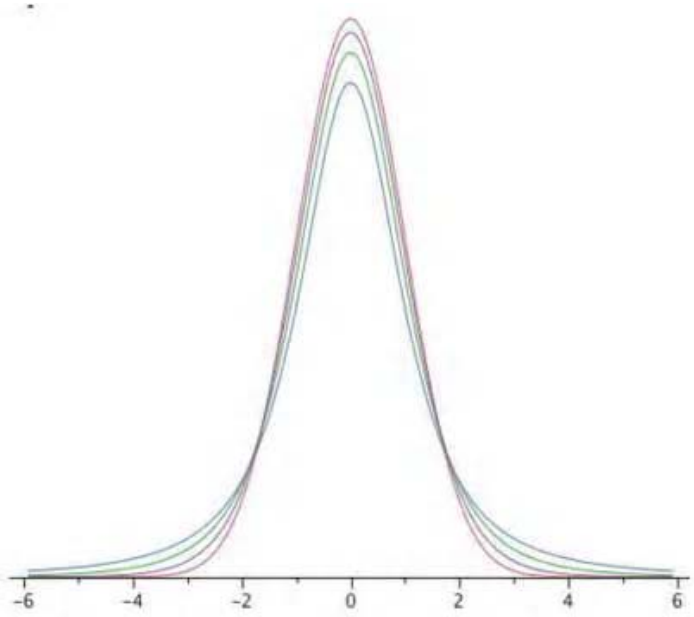
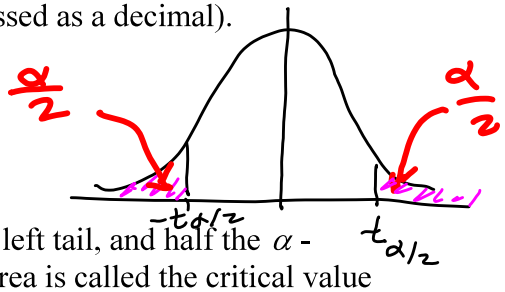


Figure 1. A comparison of t distributions with 2, 4, and 10 df and the standard normal distribution. The distribution with the lowest peak is the 2 df distribution, the next lowest is 4 df, the lowest after that is 10 df, and the highest is the standard normal distribution.

α greek letter alpha

The alpha-value, α, is equal to 1 minus the confidence level (expressed as a decimal). For example, a 90% confidence level results in $\alpha = 0.10$.
A 95% confidence level results in $\alpha = 0.05$.



Constructing the confidence interval for the mean:

To construct the confidence interval, we put half the α-value in the left tail, and half the α-value in the right tail. The boundary value adjacent to the right tail area is called the critical value of t, and is denoted $t_{\alpha/2}$.

Note: The *t*-distribution assumes that the variable of interest is normally distributed in the underlying population. However, the procedure is relatively *robust* in regards to departures from normality. If the sample size is sufficiently large, we can often use the *t*-distribution even if the population is not normal. A common rule of thumb is that we can apply the *t*-distribution to a non-normal population if $n \geq 30$ and if there are not many outliers.

Note: If the sample is more than 5% of the population, you should multiply the standard error by a finite population correction factor, $\sqrt{\frac{N-n}{n-1}}$. (In this class, I do not anticipate that we will encounter this situation.)

Procedure:

1. Verify that the population is normal, or that the sample size is sufficiently large that the departure from normality can be neglected.
2. Determine the confidence level, $1-\alpha$.
3. Determine the degrees of freedom, $n-1$.
4. Sketch the *t*-distribution, placing $\alpha/2$ in each tail.
5. Use Table VI in Appendix A, Page A-13 (or a similar table, or a computer program) to determine the critical value $t_{\alpha/2}$.
6. Estimate the standard error, $\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$
7. Multiply $t_{\alpha/2}$ by the estimated standard error $\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$ to obtain the margin of error.
8. Add and subtract the margin of error from the sample mean to obtain the lower and upper bounds of the confidence interval:

   Lower bound: $\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$

   Upper bound: $\bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$

Ex: confidence level 95% corresponds to $\alpha = 0.05$

confidence level 90% corresponds to $\alpha = 0.10$

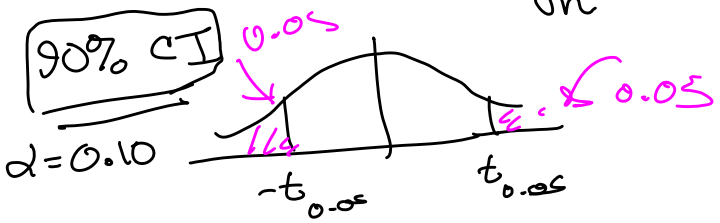
Note: If the correct value for degrees of freedom does not appear in Table IV, use the next closest value for degrees of freedom. Or, for the homework, use an online *t*-score calculator, such as this one from StatTrek: http://stattrek.com/online-calculator/t-distribution.aspx

Example 2: Suppose that 40 American college students were surveyed about the number of hours outside of class they spent studying. The mean weekly study time was 11.9 hours and the standard deviation of the weekly study time was 9.6 hours. Construct and interpret the 90% and the 95% confidence intervals.

$\bar{x} = 11.9$

std dev of sample: $s = 9.6$

Standard error: $\frac{s}{\sqrt{n}} = \frac{9.6}{\sqrt{40}} = 1.518$



Degrees of freedom: $n - 1 = 40 - 1 = 39$

From Table on A-13, $t_{0.05} = 1.685$

see next page

Ex 2 continued:

Confidence interval: upper bound: $\bar{x} + 1.685\frac{s}{\sqrt{n}} = 11.9 + 1.685(1.518)$

$= 14.458$

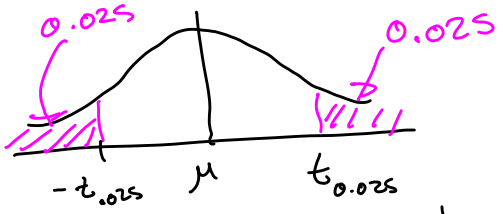
Lower bound: $\bar{x} - 1.685\frac{s}{\sqrt{n}} = 11.9 - 1.685(1.518)$

$= 9.342$

90% confidence interval: (9.34, 14.46)

We are 90% confident the population mean is in this interval.

For 95% confidence Interval:



$\alpha = 1 - 0.95 = 0.05$

So $\frac{\alpha}{2} = 0.025$. Put 0.025 in each tail.

Standard error: $\frac{s}{\sqrt{n}} = \frac{9.6}{\sqrt{40}} = 1.5179$

Use Table on page A-13 to look up t for 39 degrees of freedom and area 0.025: $t_{0.025} = 2.023$

Upper bound of 95% CI: $\bar{x} + t_{.025}\frac{s}{\sqrt{n}}$

$= 11.9 + 2.023(1.518) = 14.971$

Lower bound of 95% CI: $\bar{x} - t_{.025}\frac{s}{\sqrt{n}}$

$= 11.9 - 2.023(1.518) = 8.829$

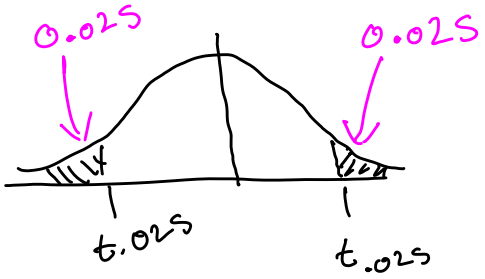
95% Confidence Interval: (8.829, 14.97)

We are 95% confident the population mean falls in this interval.

Example 3: Based on experience, a fast-food restaurant manager believes that drive-through service times follow a normal distribution. A sample of 24 drive-through transactions results in a mean service time of 3.7 minutes, with a standard deviation of 1.6 minutes. Construct and interpret the 95% and 99% confidence intervals.

$n = 24, \quad \bar{x} = 3.7, \quad s = 1.6$

For 95% Confidence Interval: $\alpha = 0.05$



Degrees of freedom: $df = n - 1 = 23$

Area in right tail: 0.025

From t-table, $t_{\alpha/2} = t_{.025} = 2.069$

This is the critical value of t

Estimate the standard error: $\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}} = \frac{1.6}{\sqrt{24}} = 0.3266$

Upper bound: $\bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 3.7 + 2.069(0.3266)$

$= 4.376$

Lower bound: $\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 3.7 - 2.069(0.3266)$

$= 3.024$

95% Confidence Level

95% CI: $(3.024, 4.376)$

We are 95% confident that the population mean is in this interval.

Sample size needed to estimate the population mean within a given margin of error:

The margin of error is $E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$. We can solve this for n:

CI: $(\bar{x} - E, \bar{x} + E)$

$E\sqrt{n} = t_{\alpha/2} \cdot s$

$\sqrt{n} = \frac{t_{\alpha/2} \cdot s}{E} \Rightarrow n = \left(\frac{t_{\alpha/2} \cdot s}{E}\right)^2$

See next page for 99% CI

However, the value of $t_{\alpha/2}$ is dependent on the sample size, which is what we are trying to find.

So, because the *t*-distribution approaches the normal distribution for large *n*, we use $z_{\alpha/2}$ instead.

($z_{\alpha/2}$ is based on the standard normal distribution and does not depend on sample size.)

Ex 3 2nd Part

Find the 99% Confidence Interval

Standard error is still approximately

$$\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}} = \frac{1.6}{\sqrt{24}} \approx 0.3266$$

99% CI $\Rightarrow \alpha = 0.01$

In table: use df = 24 − 1 = 23

Area in right tail: 0.005

Critical value of t: $t_{.005} = 2.807$

Upper bound: $\bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 3.7 + 2.807(0.3266)$

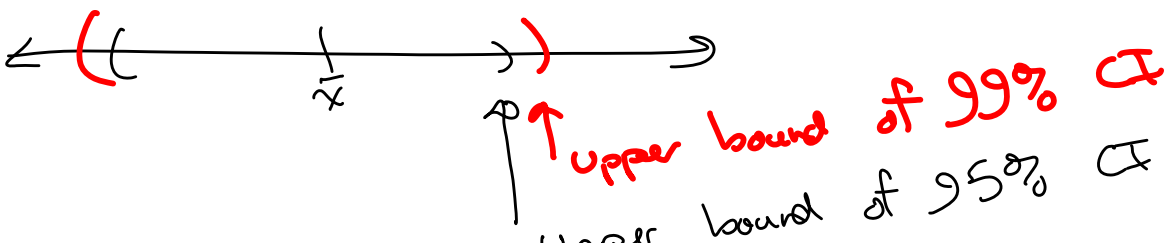
$= 4.617$

Lower bound: $\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 3.7 - 2.807(0.3266)$

$= 2.783$

99% CI: (2.783, 4.617)

Note: The 95% CI is a subset of the 99% CI.



Required sample size for estimation of the population mean:

For a specified α associated with a confidence level, the sample size required to estimate the population mean within E units is

$$n = \left(\frac{z_{\alpha/2} s}{E} \right)^2 .$$

Because this n is considered a minimum threshold, we round the calculated value of n up to the nearest whole number *above*.

Note: To use this formula, we would need to have a value for the sample standard deviation. Typically we would use an estimated value from a pilot study, or from other published research studies.

Example 4: Suppose a researcher wishes to estimate the number of hours that people spend using their computers each day. The researcher wants the estimate to be accurate to within 20 minutes. Several earlier studies of daily computer use had standard deviations of about 2.3 hours.

a) Estimate the required sample size needed to construct the 95% confidence interval.
b) If the researcher decided she could be satisfied with an estimate that was accurate within 40 minutes, how does that change the required sample size?

(a) For 95% CI, $\alpha = 0.05$

$$n = \left(\frac{z_{.025} \cdot s}{E} \right)^2$$

$$n = \left(\frac{1.96(2.3)}{0.3333} \right)^2 = 182.935$$

Use a sample of at least 183.

$E = 20 \text{ min} = \frac{1}{3} \text{ hour} \approx 0.3333$

Use $s = 2.3$ hours

Use z-table to find area = 0.475 in z-table: corresponds to $z = 1.96$