Typical Problem

The following data set are the weights (mg) of thymus glands from five chick embryos after 14 days of incubation.

The data was collected as part of a study on development of the thymus gland.

> thymus
[1] 29.6 21.5 28.0 34.6 44.9

If we model this data as having been sampled at random from a population of chick embryos with similar conditions, what can we say about the population mean weight?

Standard Error of the Mean

- We know that SD of the sampling distribution of the sample mean $\bar{y}$ can be computed by this formula.
  
  $$ \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} $$

- But if we only observe sample data $y_1, \ldots, y_n$, we do not know the value of the population SD $\sigma$, so we cannot use the formula directly.

- However, we can compute the sample standard deviation $s$, which is an estimate of the population standard deviation $\sigma$.

- The expression
  
  $$ SE_{\bar{y}} = \frac{s}{\sqrt{n}} $$

  is called the standard error of the sample mean and is an estimate of the standard deviation of the sampling distribution of the sample mean. (You can understand why statisticians gave this concept a shorter name.)

Statistical Estimation

- Statistical inference is inference about unknown aspects of a population based on treating the observed data as the realization of a random process.

- We focus in this course on inference in the setting of random samples from populations.

- Statistical estimation is a form of statistical inference in which we use the data to estimate a feature of the population and to assess the precision the estimate.

- Chapter 6 introduces these ideas in the setting of estimating a population mean $\mu$. 
**Example (cont.)**

- Here is some R code to compute the mean, standard deviation, and standard error for the example data.

```r
> m = mean(thymus)
> m
[1] 31.72
> s = sd(thymus)
> s
[1] 8.72909
> n = length(thymus)
> n
[1] 5
> se = s/sqrt(n)
> se
[1] 3.903767
```

- The **sample standard deviation** is an estimate of how far individual values differ from the population mean.
- The **standard error** is an estimate of how far sample means from samples of size \( n \) differ from the population mean.

**Confidence intervals**

The basic idea of a confidence interval for \( \mu \) is as follows.

- We know that the sample mean \( \bar{y} \) is likely to be close (within a few multiples of \( \sigma/\sqrt{n} \)) to the population mean \( \mu \).
- Thus, the unknown population mean \( \mu \) is likely to be close to the observed sample mean \( \bar{y} \).
- We can express a confidence interval by centering an interval around the observed sample mean \( \bar{y} \) — those are the possible values of \( \mu \) that would be most likely to produce a sample mean \( \bar{y} \).

**Derivation of a Confidence Interval**

From the sampling distribution of \( \bar{Y} \), we have the following statement

\[
Pr \left\{ \mu - z \frac{\sigma}{\sqrt{n}} \leq \bar{Y} \leq \mu + z \frac{\sigma}{\sqrt{n}} \right\} = 0.9
\]

if we let \( z = 1.645 \), because the area between \(-1.645\) and \(1.645\) under a standard normal curve is 0.9. Different choices of \( z \) work for different confidence levels.

The first inequality is equivalent to

\[
\mu \leq \bar{Y} + z \frac{\sigma}{\sqrt{n}}
\]

and the second is equivalent to

\[
\bar{Y} - z \frac{\sigma}{\sqrt{n}} \leq \mu
\]

which are put together to give

\[
Pr \left\{ \bar{Y} - z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z \frac{\sigma}{\sqrt{n}} \right\} = 0.9
\]

**Derivation of a Confidence Interval**

This recipe for a confidence interval is then

\[
\bar{Y} \pm z \frac{\sigma}{\sqrt{n}}
\]

- This depends on knowing \( \sigma \).
- If we don’t know \( \sigma \) as is usually the case, we could use \( s \) as an alternative.
- However, the probability statement is then no longer true.
- We need to use a different multiplier to account for the extra uncertainty.
- This multiplier comes from the \( t \) distribution.
The $t$ Distributions in R

- The functions `pt` and `qt` find areas and quantiles of $t$ distributions in R.
- The area to the right of 2.13 under a $t$ distribution with 4 degrees of freedom is:
  ```r
  > 1 - pt(2.13, 4)
  [1] 0.04286382
  ```
- To find the 95th percentile of the $t$ distribution with four degrees of freedom, you could do the following:
  ```r
  > qt(0.95, df = 4)
  [1] 2.131847
  ```
- This R code checks the values of the 0.05 upper tail probability for the first several rows of the table.
  ```r
  > round(qt(0.95, df = 1:10), 3)
  ```
- You can use R to find values not tabulated.
  ```r
  > qt(0.95, 77)
  [1] 1.664885
  ```

Statistics 371, Fall 2004

Sampling Distributions

$Z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}}$

$T = \frac{\bar{y} - \mu}{s / \sqrt{n}}$

- If the population is normal, the statistic $Z$ has a standard normal distribution.
- If the population is not normal but $n$ is sufficiently large, the statistic $Z$ has approximately a standard normal distribution (by the Central Limit Theorem).
- The distribution of the statistic $T$ is more variable than that of $Z$ because there is extra randomness in the denominator.
- The extra randomness becomes small as the sample size $n$ increases.

Mechanics of a confidence interval

A confidence interval for $\mu$ takes on the form

$$\bar{y} \pm t \times \frac{s}{\sqrt{n}}$$

where $t$ is selected so that the area between $-t$ and $t$ under a $t$ distribution curve with $n - 1$ degrees of freedom is the desired confidence level.

In the example, there are $df = n - 1 = 4$ degrees of freedom. A 90% confidence interval uses the multiplier $t = 2.132$. A 95% confidence interval would use $t = 2.776$ instead.

We are 90% confident that the mean thymus weight in the population is in the interval 31.72 ± 8.32 or (23.40, 40.04).

We are 95% confident that the mean thymus weight in the population is in the interval 31.72 ± 10.84 or (20.88, 42.56).

Student's $t$ Distribution

- If $Y_1, \ldots, Y_n$ are a random sample from any normal distribution and if $\bar{Y}$ and $S$ are the sample mean and standard deviation, respectively, then the statistic

$$T = \frac{\bar{Y} - \mu}{S / \sqrt{n}}$$

is said to have a $t$ distribution with $n - 1$ degrees of freedom.
- All $t$ distributions are symmetric, bell-shaped, distributions centered at 0, but their shapes are not quite the same as normal curves and they are spread out a more than the standard normal curve.
- The spread is largest for small sample sizes. As the sample size (and degrees of freedom) increases, the $t$ distributions become closer to the standard normal distribution.
- The Table in the back cover of your textbook provides a few key quantiles for several different $t$ distributions.
Another Example

The diameter of a wheat plant is an important trait because it is related to stem breakage which affects harvest. The stem diameters (mm) of a sample of eight soft red winter wheat plants taken three weeks after flowering are below.

2.32 6.2 4.2 2.2 3.2 5.1 9.2 0

The mean and standard deviation are $\bar{y} = 2.275$ and $s = 0.238$.

(a) Find a 95% confidence interval for the population mean.

(b) Interpret the confidence interval in the context of the problem.

Mechanics of a confidence interval

Notice that these multipliers 2.132 and 2.776 are each greater than the corresponding $z$ multipliers 1.645 and 1.96.

Had the sample size been 50 instead of 5, the $t$ multipliers 1.677 and 2.01 would still be larger than the corresponding $z$, but by a much smaller amount.

Interpretation of a confidence interval

In our real data example, we would interpret the 90% confidence interval as follows.

We are 90% confident that the mean thymus weight (mg) of all similar chick embryos that had been incubated under similar conditions would be between 23.4 and 40.04.

Notice that the interpretation of a confidence interval

- states the confidence level;
- states the parameter being estimated;
- is in the context of the problem, including units; and
- describes the population.

It is generally good practice to round the margin of error to two significant figures and then round the estimate to the same precision.
How big should \( n \) be?

- When planning a study, you may want to know how large a sample size needs to be so that your standard error is at least as small as a given size.
- Solving this problem is a matter of plugging in a guess for the population SD and solving for \( n \).

\[
\text{Desired SE} = \frac{\text{Guessed SD}}{\sqrt{n}}
\]

After solving for \( n \), we have this.

\[
n = \left( \frac{\text{Guessed SD}}{\sqrt{\text{Desired SE}}} \right)^2
\]

Conditions for Validity

- The most important condition is that the sampling process be \textit{like} \textit{simple} \textit{random} \textit{sampling}.
- If the sampling process is \textit{biased}, the confidence interval will greatly overstate the true confidence we should have that the confidence interval contains \( \mu \).
- If we have random sampling from a non-normal population, the confidence intervals are approximately valid if \( n \) is large enough so that the sampling distribution of \( \bar{Y} \) is \textit{approximately} normal.
- The answer to this question depends on the degree of non-normality.
- Specifically, \textit{strongly} \textit{skewed} \textit{distributions} require large \( n \) for the approximations to be good.
- Non-normal population shapes that are nonetheless symmetric converge to normal looking sampling distributions for relatively small \( n \).