

# Chapter 6

## 6.2-D: Distribution of a Mean

Recall that in chapter 3.1 we already learned about the distribution of the sample proportion.

A Sampling Distribution is the distribution of sample statistics computed for different samples of the same size from the same population. A sampling distribution shows us how the sample statistic varies from sample to sample.

Properties of Sampling distribution:

- **Center:** If samples are randomly selected, the sampling distribution will be centered around the population parameter. (for population mean:  $\mu$ )
- **Shape:** For most of the statistics we consider, if the sample size is large enough the sampling distribution will be symmetric and bell-shaped.
- In Chapter 5 we saw the generic formula for the sampling distribution:  
Statistic  $\sim N(\text{parameter}, SE)$
- We can change the generic formula to describe the sampling distribution for a sample mean:

$$\bar{x} \sim N(\mu, SE)$$

Notice that we need the standard error. The formula for the standard error when describing the sampling distribution of a sample mean is ...

$$SE = \frac{\sigma}{\sqrt{n}}$$

If  $n$  is sufficiently large:

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

### The Sampling Distribution of the Sample Mean $\bar{X}$ :

Central Limit Theorem:

1. For large  $n$  (roughly 30 or more), the sampling distribution  $\bar{X}$  is approximately normal even if the population distribution  $x$  is not.
2. If the population distribution  $x$  is approximately normal, then the sampling distribution  $\bar{X}$  is approximately normal for **all** sample sizes.

3. mean of  $\bar{X}$ :  $\mu_{\bar{x}} = \mu$       standard error of  $\bar{X}$ :  $SE = \frac{\sigma}{\sqrt{n}}$

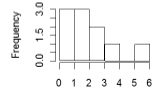
## CLT for a Mean

Population

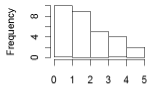


Distribution of Sample Data

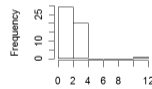
$n = 10$



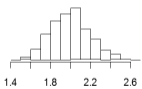
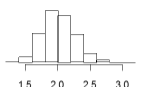
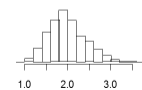
$n = 30$



$n = 50$



Distribution of Sample Means



## CLT for $\bar{x}$

If  $n$  is sufficiently large:

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- A normal distribution is usually a good approximation as long as  $n \geq 30$
- Smaller sample sizes may be sufficient for symmetric distributions, and 30 may not be sufficient for very skewed distributions or distributions with high outliers

## Standard Error

$$SE = \frac{\sigma}{\sqrt{n}}$$

- Usually, we don't know the population standard deviation  $\sigma$ , so estimate it with the sample standard deviation,  $s$

$$SE = \frac{s}{\sqrt{n}}$$

## t-distribution

- Replacing  $\sigma$  with  $s$  changes the distribution of the z-statistic from normal to  $t$
- The  $t$  distribution is very similar to the standard normal, but with slightly fatter tails to reflect this added uncertainty

## t-distribution

- If a population with mean  $\mu_0$  is approximately normal or if  $n$  is large ( $n \geq 30$ ), the standardized statistic for a mean using the sample  $s$  follows a  $t$ -distribution with  $n - 1$  degrees of freedom:

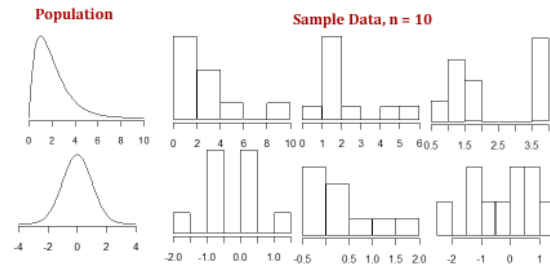
$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

## Normality Assumption

- Using the  $t$ -distribution requires that the data comes from a *normal distribution*
- Note: this assumption is about the population data, *not* the distribution of the statistic
- For large sample sizes we do not need to worry about this, because  $s$  will be a very good estimate of  $\sigma$ , and  $t$  will be very close to  $N(0,1)$
- For small sample sizes ( $n < 30$ ), we can only use the  $t$ -distribution if the distribution of the data is approximately normal

## Normality Assumption

- One small problem: for small sample sizes, it is very hard to tell if the data actually comes from a normal distribution!



Statistics: Unlocking the Power of Data

Lock<sup>5</sup>

## Small Samples

- If sample sizes are small, only use the  $t$ -distribution if the data look reasonably symmetric and do not have any extreme outliers.

- Even then, remember that it is just an approximation!

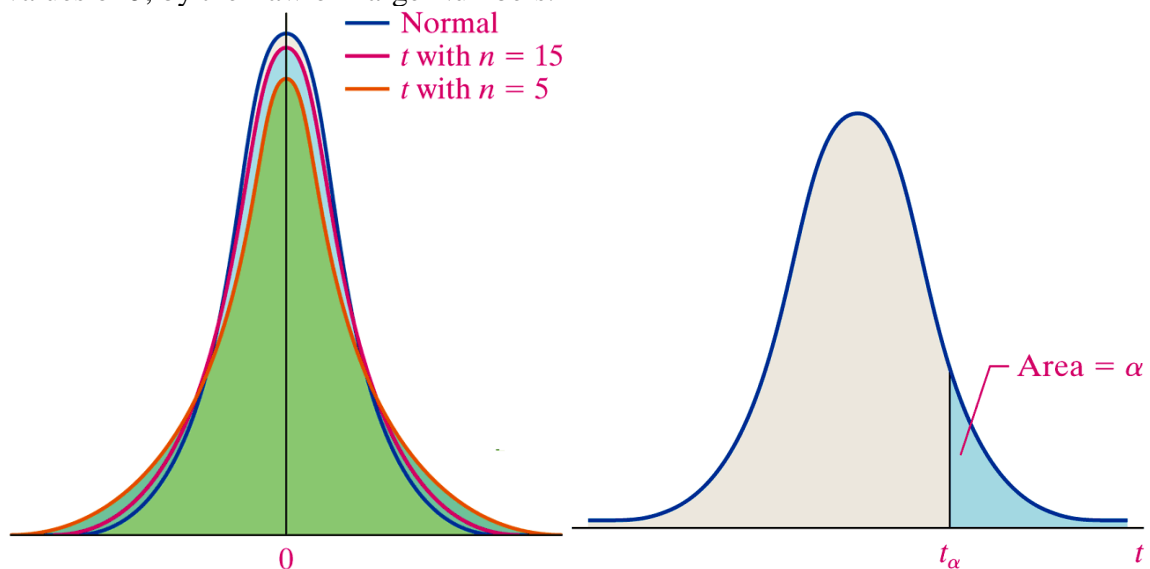
- In practice/life, if sample sizes are small, you should just use simulation methods (bootstrapping and randomization)

Statistics: Unlocking the Power of Data

Lock<sup>5</sup>

## Properties of the $t$ Distribution

- The  $t$ -distribution is bell shaped and symmetric about 0
- The probabilities depend on the degrees of freedom,  $df=n-1$ . The  $t$ -distribution has thicker tails than the standard normal distribution, i.e., it is more spread out. The additional spread in the distribution of  $t$  can be attributed to the fact that we use  $s$  to find  $t$  instead of  $\sigma$ . Because the sample standard deviation is itself a random variable (rather than a constant such as  $\sigma$ )
- The area under the curve is 1. The area under the curve to the right of 0 equals the area under the curve to the left of 0, which equals 1/2.
- As  $t$  increases or decreases without bound, the graph approaches, but never equals, zero.
- The area in the tails of the  $t$ -distribution is a little greater than the area in the tails of the standard normal distribution, because we are using  $s$  as an estimate of  $\sigma$ , thereby introducing further variability into the  $t$ -statistic.
- The higher the degrees of freedom, the closer the  $t$ -distribution is to the standard normal. This result occurs because, as the sample size  $n$  increases, the values of  $s$  get closer to the values of  $\sigma$ , by the Law of Large Numbers.



Ex: Find the  $t$ -value such that the area under the  $t$ -distribution to the right of the  $t$ -value is 0.2 assuming 10 degrees of freedom. That is, find  $t_{0.20}$  with 10 degrees of freedom.

Use Ti-83 calculator:  $\text{invT}(\text{area to the left}, \text{df}) = \text{invT}(.8, 10) = .8791$

Or use  $t$ -table

### **Example 1: Salaries of Major League Baseball Players**

There were 855 major league baseball players in 2012 and their mean salary was  $\mu = 3.44$  million dollars with standard deviation  $\sigma = 4.70$  million dollars. If we take random samples of size  $n = 30$  players and calculate the mean salary, in millions of dollars, of each sample, describe the shape, center, and standard error of the distribution of sample means.

*The distribution will be bell-shaped and centered at a mean of 3.44 million dollars. The standard error will be*

$$SE = \frac{4.70}{\sqrt{30}} = 0.858.$$

### **Example 2: More on Salaries of Major League Baseball Players**

Using the same data as Example 1, but now taking samples of size  $n = 75$ , describe the shape, center, and standard error of the distribution of sample means. Compare your answers with those of Example 1.

*The distribution will be bell-shaped and centered at a mean of 3.44 million dollars. The standard error will be*

$$SE = \frac{4.70}{\sqrt{75}} = 0.543.$$

*Notice that the shape and center don't change as the sample size gets larger, but the variability goes down.*

### **Quick Self-Quiz: Using the $t$ -Distribution**

- (a) Find endpoints of a  $t$ -distribution with 5% beyond them in each tail if the sample has size  $n = 18$ .
- (b) Find the area in a  $t$ -distribution to the right of 2.30 if the sample has size  $n = 15$ .  
*The degrees of freedom are  $df = 14$  and the right-tail area is 0.019.*  
*Or use Calculator  $\text{tcdf}(2.30, 100, 14) = 0.0187$*
- (c) Find the area in a  $t$ -distribution to the left of  $-1.22$  if the sample has size  $n = 50$ .  
*The degrees of freedom are  $df = 49$  and the left-tail area is 0.114.*  
*Or use Calculator  $\text{tcdf}(-100, -1.22, 49) = .1142$*

## **6.2-CI: Confidence Interval for a Mean**

In Chapter 5 we saw the generic formula for a confidence interval:

$$\text{Statistic} \pm \text{Critical Value} \cdot SE$$

Notice that we need the standard error. The formula for the standard error when describing a confidence interval for a population mean is ...

$$\frac{s}{\sqrt{n}}$$

Since we are using the sample standard deviation, we need a new critical value.

$$t^*$$

Therefore, the formula to describe a test statistic for a population mean is:

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

### Confidence Interval

- The general formula for a confidence interval is  $statistic \pm z^* \cdot SE$
- For means, replacing  $\sigma$  with  $s$  causes us to use the  $t$ -distribution instead of the standard normal
- For means:  $statistic \pm t^* \cdot SE$

Statistics: Unlocking the Power of Data

Lock<sup>5</sup>

### Confidence Interval for $\bar{x}$

$$statistic \pm t^* \cdot SE$$

If the population is approximately normal or the  $n$  is large ( $n \geq 30$ ), then a confidence interval for  $\mu$  can be computed by

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

Statistics: Unlocking the Power of Data

Lock<sup>5</sup>



### Margin of Error

$$ME = t^* \cdot \frac{s}{\sqrt{n}}$$

You can choose your sample size in advance, depending on your desired margin of error!

Given this formula for margin of error, solve for  $n$ .

$$n = \left( \frac{t^* s}{ME} \right)^2$$

Statistics: Unlocking the Power of Data

Lock<sup>5</sup>

### Margin of Error

$$n = \left( \frac{z^* s}{ME} \right)^2$$

- Problem 1: For  $t^*$ , need to know  $n$ .
  - Solution: Use  $z^*$  instead of  $t^*$  (they are usually close)
- Problem 2: For  $s$ , need data.
  - Solution: estimate  $s$ .
    1. Use data from a previous study or similar population
    2. Take a small pre-sample to estimate  $s$
    3. Estimate the range ( $\max - \min$ ) and use  $s \approx \text{range}/4$
    4. Make a reasonable guess.

Statistics: Unlocking the Power of Data

Lock<sup>5</sup>

<https://www.youtube.com/watch?v=bFefxSE5bmo> (Confidence Intervals for One Mean: Sigma Not Known (t Method))

<https://www.youtube.com/watch?v=bMPd9-XOLUQ> (minimum sample size to Estimate Population mean)

### Example 1: Dark Chocolate for Good Health

Eleven people were given 46 grams (1.6 ounces) of dark chocolate every day for two weeks, and their vascular health was measured before and after the two weeks. Larger numbers indicate

greater vascular health, and the mean increase for the participants was 1.3 with a standard deviation of 2.32. Assume a dotplot shows the data are reasonably symmetric with no extreme values. Find and interpret a 90% confidence interval for the mean increase in this measure of vascular health after two weeks of eating dark chocolate. Can we be 90% confident that the mean change for everyone would be positive?

For a 90% confidence interval with  $11 - 1 = 10$  degrees freedom we find  $t^* = 1.812$ .

$$\begin{aligned} & \text{Statistic} \pm t^* \cdot \text{SE} \\ & \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}} \\ & 1.3 \pm 1.812 \cdot \frac{2.32}{\sqrt{11}} \\ & 1.3 \pm 1.268 \\ & 0.032 \text{ to } 2.568 \end{aligned}$$

We are 90% sure that the mean change in this measure of vascular health for people who eat dark chocolate for two weeks is between 0.032 and 2.568. Since all values in the interval are positive, we are 90% confident that the mean change is positive (improving vascular health.)

### Example 2: Sample Size and Margin of Error for Dark Chocolate

(a) What is the margin of error for the confidence interval found in Example 1?

The margin of error is 1.268.

(b) What sample size is needed if we want a margin of error within  $\pm 0.5$ , with 90% confidence? (Use the standard deviation from the original sample to estimate  $\sigma$ .)

$$n = \left( \frac{z^* \hat{\sigma}}{ME} \right)^2 = \left( \frac{1.645 \cdot 2.32}{0.5} \right)^2 = 58.26.$$

We should use a sample size of at least 59 to achieve this level of accuracy.

### Quick Self-Quiz: Cell Phone Calls

A survey of 1,917 cell phone users in May 2010 asked “On an average day, about how many cell phone calls do you make and receive on your cell phone?” The mean number of calls was 13.10, with a standard deviation of about 10.2. Find and interpret a 99% confidence interval for the mean number of cell phone calls for all cell phone users.

We are 99% sure that the mean number of calls per day for all cell phone users is between calls and calls.

## 6.2-HT: Hypothesis Test for a Mean

In Chapter 5 we saw the generic formula for a test statistic:

$$\text{Test Statistic} = \frac{\text{Statistic} - \text{Value under } H_0}{\text{SE}}$$

We can change the generic formula to describe a test statistic for a single mean:

$$\text{Test Statistic} = \frac{\bar{x} - \mu_0}{\text{SE}}$$

**The population standard deviation is unknown**

- Thus instead of the test statistic knowing  $\sigma$

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

we calculate a test statistic using  $s$  follow **Student's  $t$ -distribution** with  $n - 1$  degrees of freedom.

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

This is the appropriate test statistic to use when  $\sigma$  is unknown

- Step 1: Assumptions**
  - The data are obtained using randomization
  - The population distribution is approximately normal:

Case 1: the population is normal

Case 2: the sample size is large ( $n \geq 30$ ).

In those cases, the distribution of the sample mean  $\bar{x}$  is normal with mean  $\mu$  and standard deviation  $s / \sqrt{n}$

**Step 2: Hypotheses:**

Form	Null and alternative hypotheses
Right-tailed test	$H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$
Left-tailed test	$H_0: \mu \geq \mu_0$ versus $H_a: \mu < \mu_0$
Two-tailed test	$H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$

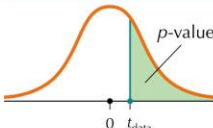
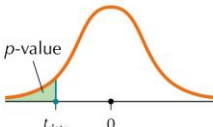
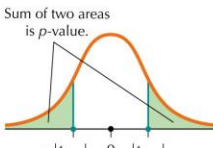
- Step 3: Test Statistic**

The test statistic measures how far the sample mean falls from the null hypothesis value  $\mu_0$ , as measured by the number of standard errors between them

The test statistic is:

$$t_{data} = t_0 = \frac{(\bar{x} - \mu_0)}{(s / \sqrt{n})} \quad \text{Degrees of freedom: } n - 1$$

- Step 4: P-value**

Form of test	The $p$ -value equals...	
<b>Right-tailed test</b> $H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$P(t > t_{\text{data}})$ Area to the right of $t_{\text{data}}$	
<b>Left-tailed test</b> $H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$P(t < t_{\text{data}})$ Area to the left of $t_{\text{data}}$	
<b>Two-tailed test</b> $H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$	$P(t >  t_{\text{data}} ) + P(t < - t_{\text{data}} )$ $= 2 \cdot P(t >  t_{\text{data}} )$ Sum of the two tail areas	

Right-tailed test:  $P\text{-value} = \text{tcdf}(t, 100, n-1)$

Left-tailed test:  $P\text{-value} = \text{tcdf}(-100, t, n-1)$

Two-tailed test:  $P\text{-value} = 2\text{tcdf}(\ , \ , n-1)$

- Step 5: Conclusion**

We summarize the test by reporting and interpreting the  $P$ -value

Reject  $H_0$  when the  $p\text{-value} < \alpha$ .

Do not reject the null hypothesis if the  $P\text{-value} > \alpha$

- Interpreting the Conclusion**

If you reject  $H_0$ , the interpretation is “There is evidence that [whatever  $H_a$  says].”

If you do not reject  $H_0$ , the interpretation is “There is insufficient evidence that [whatever  $H_a$  says].”

**Example:** In the 2010-11 National Hockey League (NHL) regular season, the number of penalty minutes per game for each of the 30 teams ranged from a low of 8.8 for the Florida Panthers to a high of 18.0 for the New York Islanders. The mean for all 30 teams is 12.20 penalty minutes per game with a standard deviation of 2.25. If we assume that this is a sample of all teams in all seasons, test to see if this provides evidence that the mean number of penalty minutes per game for a hockey team is less than 13. Show all details of the test.

We are testing  $H_0: \mu = 13$  vs.  $H_a: \mu < 13$  where  $\mu$  represents the mean number of penalty minutes per game for all hockey teams in the NHL.

$$\text{The test statistic is } t = \frac{\text{Statistic} - \text{Null}}{SE} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{12.20 - 13}{\frac{2.25}{\sqrt{30}}} = -1.947.$$

This is a left-tail test, and we use a  $t$ -distribution with  $df = 29$  to find the  $p$ -value. We see that the area to the left of  $-1.947$  in a  $t$ -distribution with  $df = 29$  is 0.031, so the  $p$ -value is 0.031.

At a 5% significance level, we reject  $H_0$  and conclude that there is evidence that the mean number of penalty minutes is less than 13. However, the results are not strong enough to be significant at the 1% level.