JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

An Introduction to Hypothesis Testing: The Paired t-Test

John McGready
Johns Hopkins University

Lecture Topics

- Comparing two groups: the paired data situation
- Hypothesis testing: the null and alternative hypotheses
- Relationships between confidence intervals and hypothesis testing when comparing means
- p-values: definition, calculations, and more information

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section A

The Paired t-Test; the Confidence Interval Component

Comparison of Two Groups

- Are the population means different? (continuous data)

- Paired design
  - Before-after data
  - Twin data
  - Matched case-control

- Two independent sample design
  - Randomized trial
  - Smokers to non-smokers

Paired Design—Example: Before vs. After

- Why pairing?
  - Control extraneous noise
  - Each observation acts as a control
  - Good way to get preliminary data/estimates to be used to develop further research

Paired Design—Example: Before vs. After

- Ten non-pregnant, pre-menopausal women 16–49 years old who were beginning a regimen of oral contraceptive (OC) use had their blood pressures measured prior to starting OC use and three-months after consistent OC use[1]

- The goal of this small study was to see what, if any, changes in average blood pressure were associated with OC use in such women

- The data on the following slides shows the resulting pre- and post-OC use systolic BP measurements for the 10 women in the study

[1] Data taken from Rosner B Fundamentals of Biostatistics, 6th ed. (2005) Duxbury Press.

Blood Pressure and Oral Contraceptive Use

	BP Before OC	BP After OC	After-Before
1.	115	128	13
2.	112	115	3
3.	107	106	-1
4.	119	128	9
5.	115	122	7
6.	138	145	7
7.	126	132	6
8.	105	109	4
9.	104	102	-2
10.	115	117	2

$\bar{x}_{before} = 115.6$ $\bar{x}_{after} = 120.4$ $\bar{x}_{diff} = 4.8$

Blood Pressure and Oral Contraceptive Use

- The sample average of the differences is 4.8
  - Also note $\bar{x}_{diff} = \bar{x}_{after} - \bar{x}_{before}$

- The sample standard deviation (s) of the differences is s_{diff} = 4.6
  - Standard deviation of differences found by using the formula:

$$s_{diff} = \sqrt{\frac{\sum_{i=1}^{n} (x_{diff} - \bar{x}_{diff})^2}{n-1}}$$

  - Where:
    - Each x_{diff_i} represents an individual difference and
    - x_{diff_i} is the mean difference

Blood Pressure and Oral Contraceptive Use

	BP Before OC	BP After OC	After-Before
1.	115	128	13
2.	112	115	3
3.	107	106	-1
4.	119	128	9
5.	115	122	7
6.	138	145	7
7.	126	132	6
8.	105	109	4
9.	104	102	-2
10.	115	117	2

$\bar{x}_{before} = 115.6$ $\bar{x}_{after} = 120.4$ $\bar{x}_{diff} = 4.8$

Note

- In essence, what we have done is reduce the BP information on two samples (women prior to OC use, women after OC use) into one piece of information: information on the differences in BP between the samples

- This is standard protocol for comparing paired samples with a continuous outcome measure

Confidence Interval Approach

- Want to draw a conclusion about a population parameter
  - In a population of women who use oral contraceptives, is the average (expected) change in blood pressure (after-before) 0 or not?

- Sometimes the term *expected* is used for the population average

- μ is the expected (population) mean change in blood pressure

- CI approach allows us to create a range of possible values for μ using data from a single, imperfect (paired) sample

95% Confidence Interval

- 95% confidence interval for mean change in BP in population of women taking oral contraceptives, after starting OC use compared to before OC use

$$\bar{x}_{diff} \pm t_{.95,9} \times S\hat{E}(\bar{x}_{diff})$$

$$\bar{x}_{diff} \pm t_{.95,9} \times \frac{s_{diff}}{\sqrt{10}}$$

$$4.8 \pm 2.26 \times \left(\frac{4.6}{\sqrt{10}} \right)$$

$$1.5\ mmHG\ to\ 8.1\ \ mmHg$$

95% Confidence Interval

- 95% confidence interval for mean change in BP in population of women taking oral contraceptives, after starting OC use compared to before OC use using *cii* in Stata

```
. cii 10 4.8 4.6

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |         10         4.8    1.454648        1.509358    8.090642
```

95% Confidence Interval

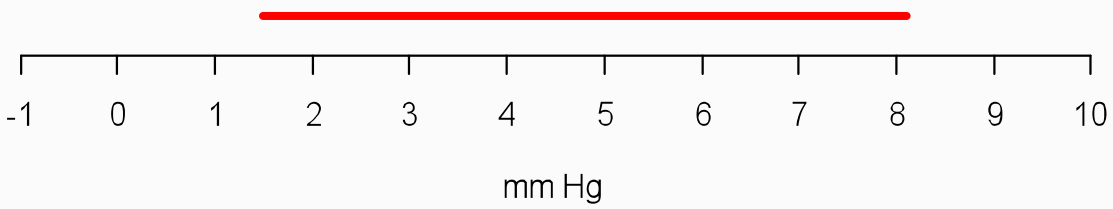
- 95% confidence interval for mean change in BP in population of women taking oral contraceptives, after starting OC use compared to before OC use using *cii* in Stata

```
. cii 10 4.8 4.6

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |         10         4.8    1.454648       1.509358    8.090642
```

Notes

- The number 0 is **NOT** in confidence interval (1.5–8.1)



Notes

- The number 0 is **not** in confidence interval (1.5–8.1)
  - Because 0 is not in the interval, this suggests there is a non-zero change in BP over time
  - The phrase “statistically significant” change is used to indicate a non-zero mean change

Notes

- The BP change could be due to factors other than oral contraceptives
  - Changes in weather over pre- and –post period
  - Changes in personal stress
  - Other changes?

- A control group of comparable women who were not taking oral contraceptives would strengthen this study
  - This is an example of a *pilot study*—a small study done just to generate some evidence of a possible association
  - This can be followed up with a larger, more scientifically rigorous study

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section B

The Paired t-Test; the Hypothesis Testing Component

Hypothesis Testing Approach

- Want to draw a conclusion about a population parameter
  - In a population of women who use oral contraceptives, is the average (expected) change in blood pressure (after-before) 0 or not?

- Sometimes the term *expected* is used for the population average

- μ is the expected (population) mean change in blood pressure

- Hypothesis testing approach allows us to choose between two competing possibilities for μ using a single imperfect (paired) sample

Hypothesis Testing

- Two, mutually exclusive, exhaustive possibilities for "truth" about mean change
  - Null hypothesis: represented by H_o: ("h-knot" or "h-oh")
    - H_o: $\mu = 0$
  - Alternative hypothesis
    - H_A: $\mu \neq 0$

- We will use the results from our study to choose between the null and alternative hypotheses

The Null Hypothesis, H_0

- **Null:** typically represents the hypothesis that there is "no association" or "no difference"
  - For example, there is no association between oral contraceptive use and blood pressure
    - H_o: $\mu = 0$

- **Alternative:** the very general complement to the null
  - For example, there is an association between blood pressure and oral contraceptive use
    - H_A: $\mu \neq 0$

Hypothesis Testing

- We are testing both hypotheses at the same time
  - Our result will allow us to either:
    - "Reject H_0"
      or
    - "Fail to reject H_o"

- We start by assuming the null (H_o) is true, and asking:
  - How likely is the result we got from our sample if H_o is the truth —i.e., no change in mean blood pressure after taking OCs?
  - $\bar{x}$ would have to be far from zero to claim H_A is true
    - But is $\bar{x}$ = 4.8mmHg big enough to choose H_A ?

Hypothesis Testing Question

- Is our sample result "unlikely" when H_o is true—and therefore we should H_0 in favor of H_A?
  - We need some measure of how probable the result from our sample is, if the null hypothesis is true
  - Need the probability of having gotten such an extreme sample mean as 4.8 if the null hypothesis (H_0: $\mu = 0$) was true?
    - This probability is called the *p-value*

Hypothesis Testing Question

- Does our sample result allow us to reject H_0 in favor H_A?
  - If that probability (p-value) is small, it suggests the observed result cannot be easily explained by chance

- So, what can we turn to evaluate how unusual our sample statistic is when the null is true?
  - We need a mechanism that will explain the behavior of the sample mean across many different random samples of 10 women—when the truth is that oral contraceptives do not affect blood pressure
  - Luckily, we've already defined this mechanism: it's the *sampling distribution of the sample mean!*

Sampling Distribution

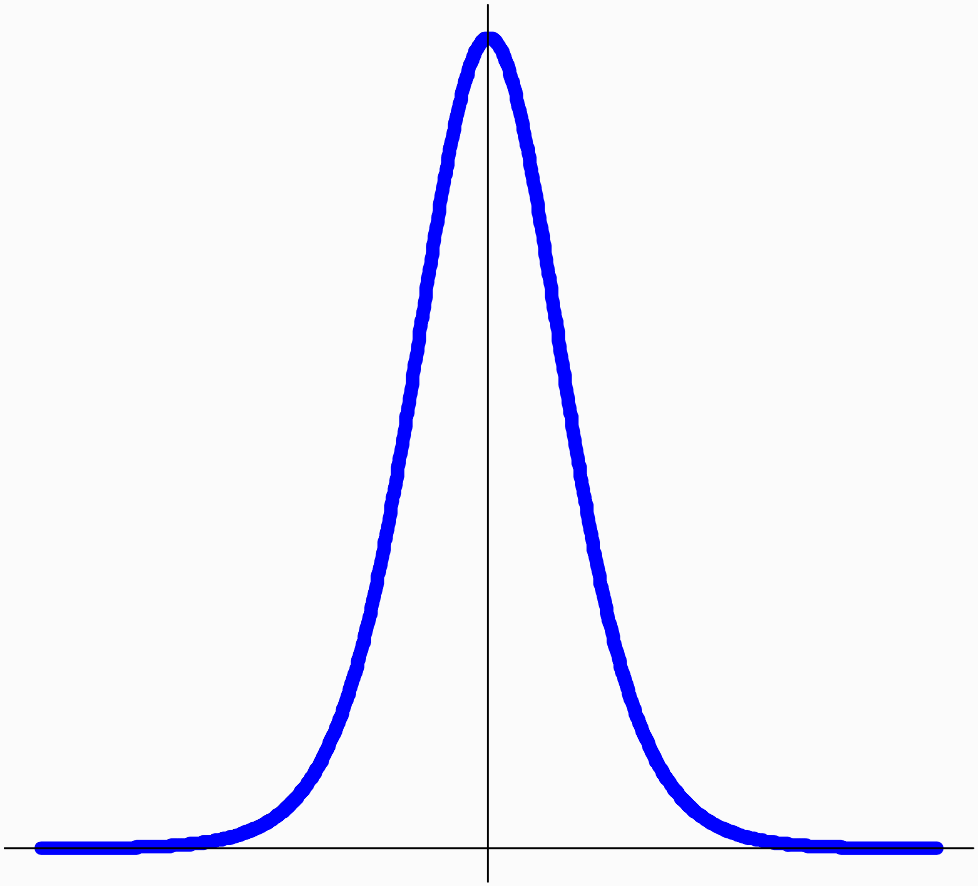
- *Sampling distribution of the sample mean* is the (theoretical) distribution of all possible values of $\bar{x}$ from samples of same size, *n*

- For BP example, theory tells us it is a *t_9 distribution*

- Recall, the sampling distribution is centered at the "truth," the underlying value of the population mean, μ
  - In hypothesis testing, we start under the assumption that H_0 is true—so the sampling distribution under this assumption will be centered at μ_0, the null mean

Sampling Distribution

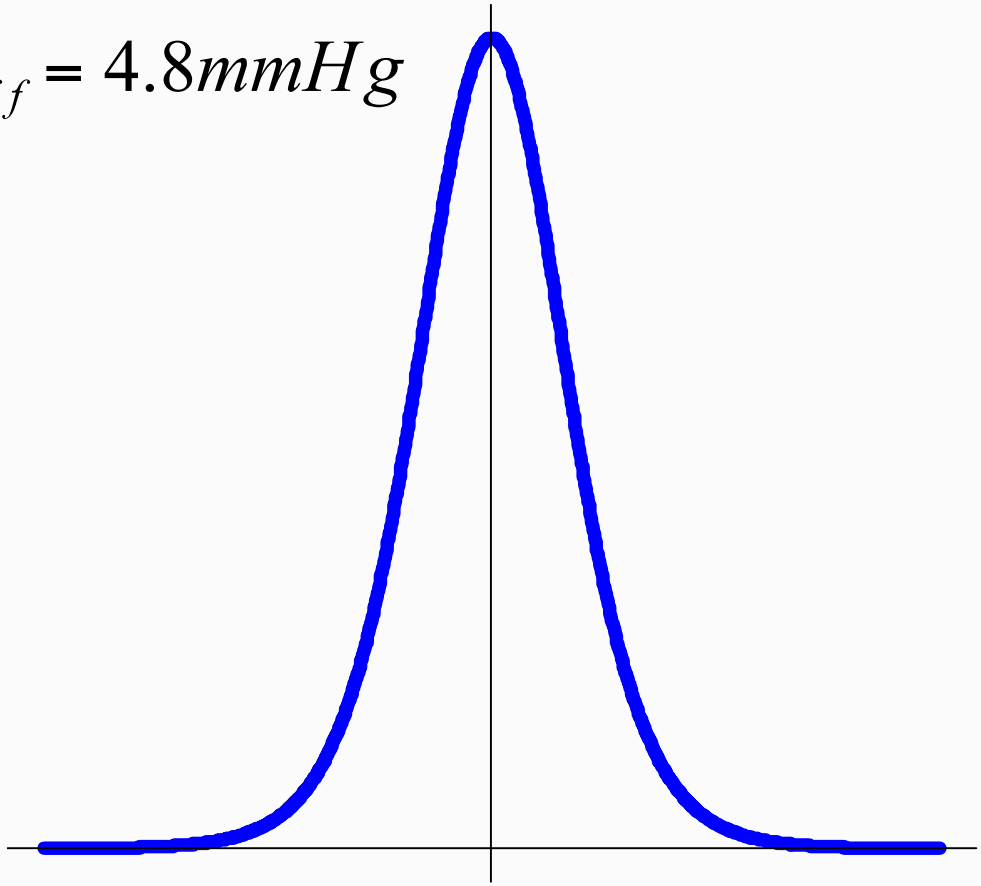
- Sampling distribution of sample mean differences (after–before) in BP, from samples of size n=10



Getting a p-Value

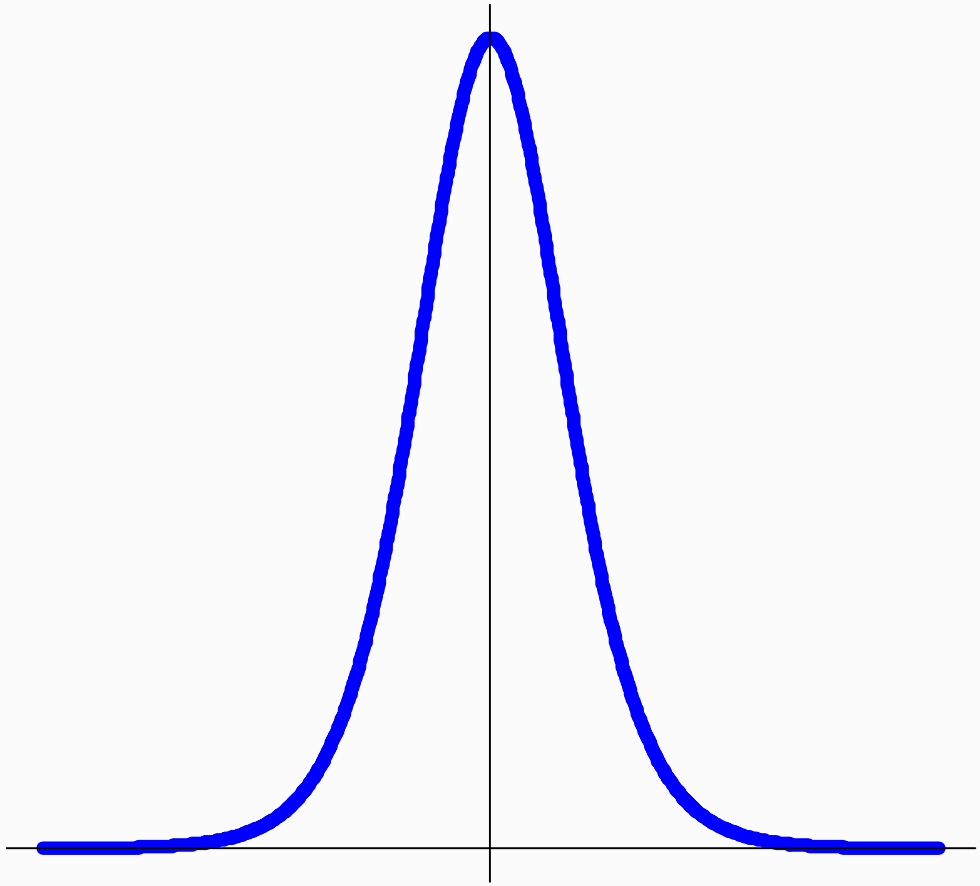
- To compute a p-value, we need to find our value of $\bar{x}_{diff}$ on the graph and figure out how "unusual" it is

- Recall: $\bar{x}_{diff} = 4.8 mmHg$



Getting a p-Value

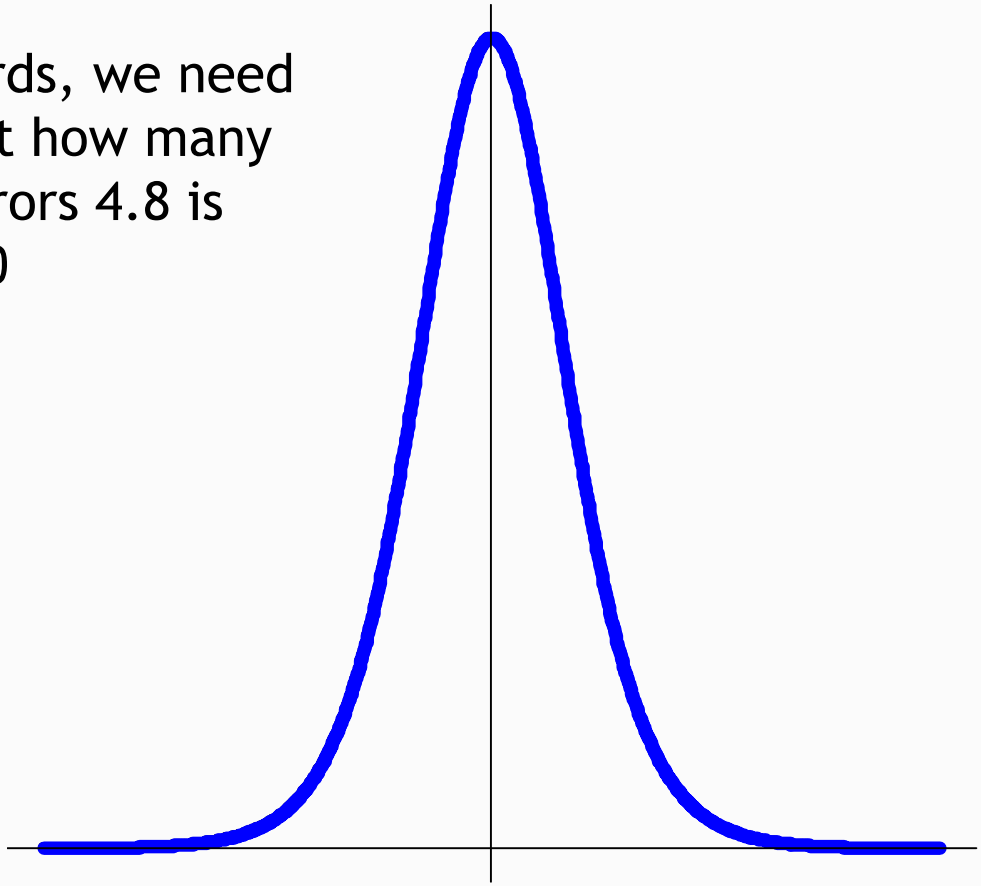
- Where is $\bar{x}_{diff} = 4.8 mmHg$ under the curve?



Getting a p-Value

- We need to figure out how “far” our result—4.8—is from 0, in “standard statistical units”

- In other words, we need to figure out how many standard errors 4.8 is away from 0



How Are p-Values Calculated?

- Calculate the distance in standard errors
  - Called a *t-statistic,* but synonymous with *z-score,* normal score, etc.—think of it as a distance

$$t = \frac{\bar{x}_{diff} - 0}{S\hat{E}(\bar{x})}$$

$$t = \frac{4.8 - 0}{4.6/\sqrt{10}} = \frac{4.8}{1.45} \approx 3.3$$

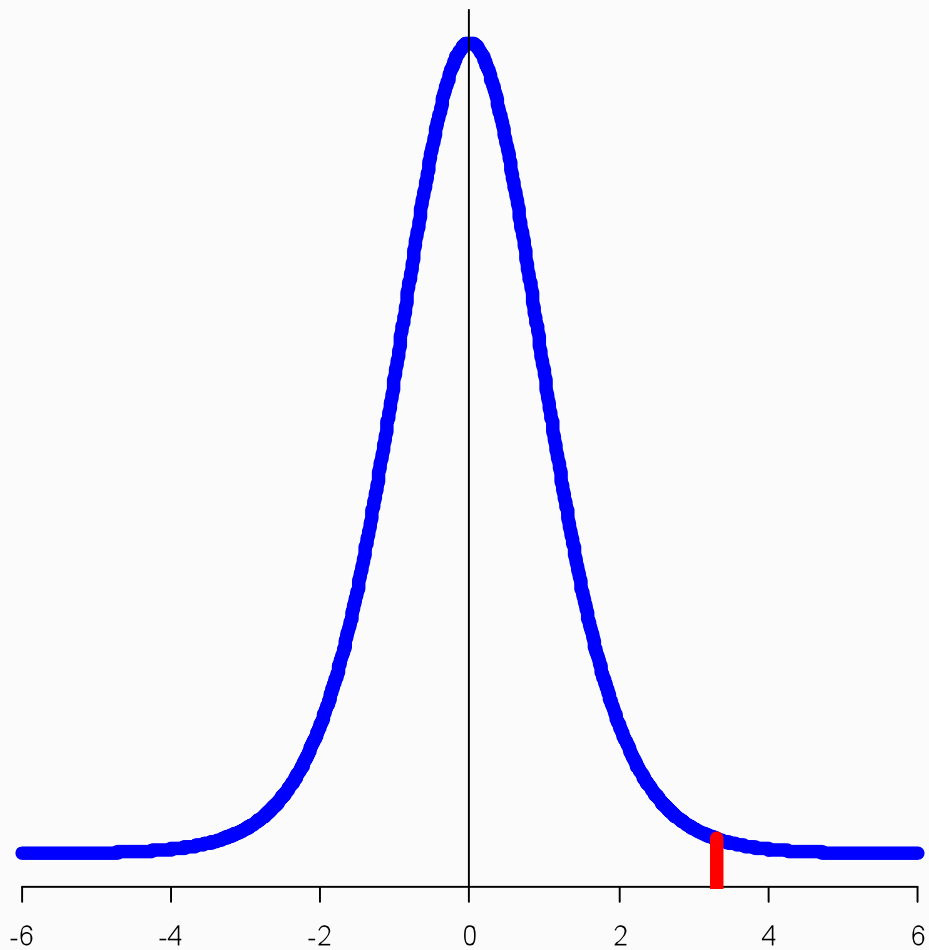
How Are p-Values Calculated?

- We observed a sample mean that was 3.3 standard errors of the mean (SEM) away from what we would have expected the mean to be if OC use were not associated with blood pressure

- Is a result 3.3 standard errors above its mean unusual?
  - Lets see where it falls on the sampling distribution

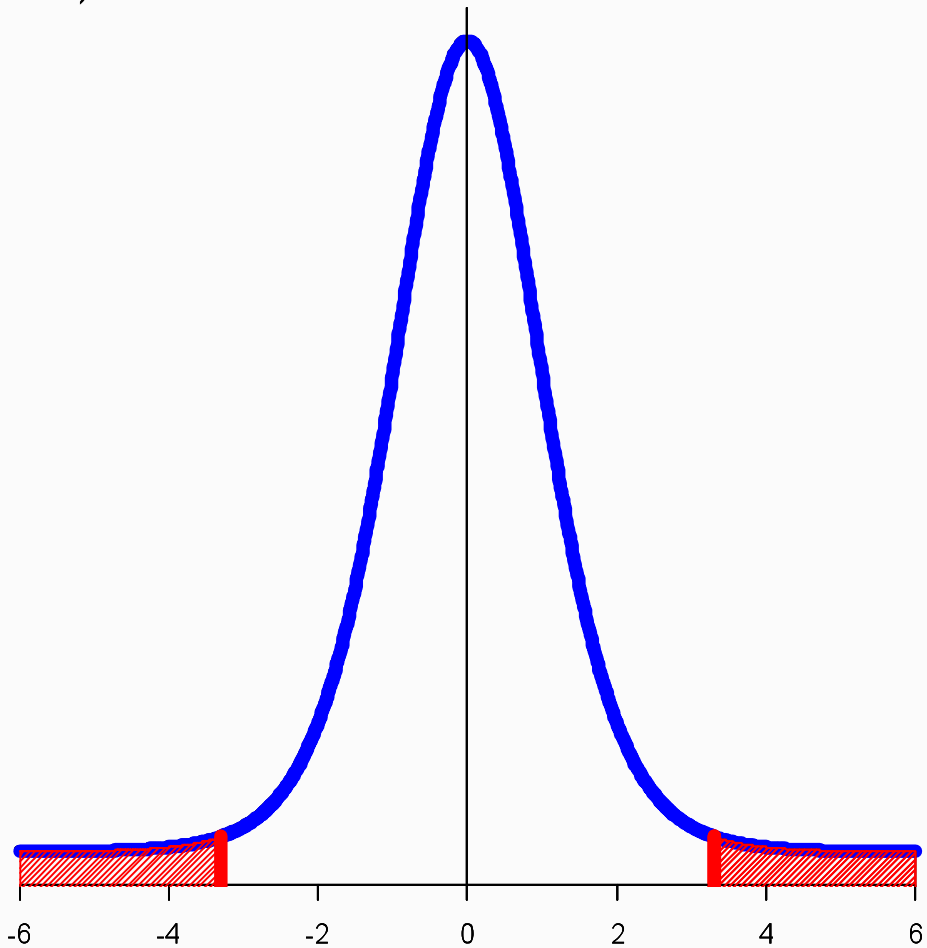
How Are p-Values Calculated?

- 3.3 on the sampling distribution (t_9)



How Are p-Values Calculated?

- The p-value is the probability of getting a sample result as (or more) extreme than what you observed (3.3) away from $\mu_0 = 0$ (in either direction from 0)



How Are p-Values Calculated?

- We could look this up in a t-table . . .
- Better option—let Stata do the work for us!

How to Use STATA to Perform a Paired t-Test

- At the command line:

  `ttesti` n s $\bar{x}_{diff}$ $\mu 0$

- For the BP-OC data:

```
ttesti  10  4.8  4.6  0
```

Stata Output

- Using *ttesti*

```
. ttesti 10 4.8 4.6 0

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      10         4.8    1.454648         4.6    1.509358    8.090642
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   3.2998
Ho: mean = 0                                     degrees of freedom =        9

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.9954         Pr(|T| > |t|) = 0.0092          Pr(T > t) = 0.0046
```

Stata Output

- 95% CI

```
. ttesti 10 4.8 4.6 0

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      10         4.8    1.454648         4.6    1.509358    8.090642
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   3.2998
Ho: mean = 0                                     degrees of freedom =        9

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.9954         Pr(|T| > |t|) = 0.0092          Pr(T > t) = 0.0046
```

Stata Output

- p-value

```
. ttesti 10 4.8 4.6 0

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      10         4.8    1.454648         4.6    1.509358    8.090642
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   3.2998
Ho: mean = 0                                     degrees of freedom =        9

    Ha: mean < 0                 Ha: mean != 0                  Ha: mean > 0
 Pr(T < t) = 0.9954         Pr(|T| > |t|) = 0.0092          Pr(T > t) = 0.0046
```

Interpreting the p-Value

- The p-value in the blood pressure/OC example is **.0092**
  - Interpretation: if the true before OC/after OC blood pressure difference is 0 among all women taking OCs, then the chance of seeing a mean difference as extreme/more extreme as 4.8 in a sample of 10 women is **.0092**

Using the p-Value to Make a Decision

- We now need to use the p-value to choose a course of action: either reject H_0, or fail to reject H_0
  - We need to decide if our sample result is unlikely enough to have occurred by chance if the null was true
    - Our measure of this "unlikeliness" is $p = 0.0092$

Using the p-Value to Make a Decision

- Establishing a cutoff
  - In general, to make a decision about what p-value constitutes "unusual" results, there needs to be a cutoff, such that all p-values less than the cutoff result in rejection of the null
  - Standard cutoff is .05—this is an arbitrary value
  - Cut off is called *alpha-level* of the test

Using the p-Value to Make a Decision

- Establishing a cutoff
  - Frequently, the result of a hypothesis test with a p-value less than .05 (or some other arbitrary cutoff) is called *statistically significant*
  - At the .05 level, we have a statistically significant blood pressure difference in the BP/OC example

Blood Pressure: Oral Contraceptive Example

- Statistical method
  - The changes in blood pressures after oral contraceptive use were calculated for 10 women
  - A paired t-test was used to determine if there was a statistically significant change in blood pressure, and a 95% confidence was calculated for the mean blood pressure change (after-before)

Blood Pressure: Oral Contraceptive Example

- Result
  - Blood pressure measurements increased on average 4.8 mmHg with standard deviation 4.6 mmHg
  - The 95% confidence interval for the mean change was 1.5 mmHg–8.1 mmHg
  - The blood pressure measurements after oral contraceptive use were statistically significantly higher than before oral contraceptive use (p=.009)

Blood Pressure: Oral Contraceptive Example

- Discussion
  - A limitation of this study is that there was no comparison group of women who did not use oral contraceptives
  - We do not know if blood pressures may have risen without oral contraceptive usage

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section C

The Paired t-Test; Two More Examples

Clinical Agreement by Two Diagnosing Physicians

- Two different physicians assessed the number of palpable lymph nodes in 65 randomly selected male sexual contacts of men with AIDS or AIDS-related conditions[1]

	Doctor 1	Doctor 2	Difference
Mean ($\bar{x}$)	**7.91**	**5.16**	**-2.75**
sd (*s*)	**4.35**	**3.93**	**2.83**

[1]Example based on data taken from Rosner, B. (2005). *Fundamentals of Biostatistics*, sixth. ed. Duxbury Press. (Based on research by Coates, et al. (1988). Assessment of generalized … *Journal of Clinical Epidemiology*, 41(2).

95% Confidence Interval

- 95% CI for difference in mean number of lymph nodes, Doctor 2 compared to Doctor 1

$$\bar{x}_{diff} \pm 2 \times S\hat{E}(\bar{x}_{diff})$$

$$\bar{x}_{diff} \pm 2 \times \frac{s_{diff}}{\sqrt{65}}$$

$$2.75 \pm 2 \times \left(\frac{2.83}{\sqrt{65}} \right)$$

$$-3.45 \ \ to \ \ -2.05$$

Getting a p-Value

- Hypotheses
  - H_o: $\mu_{diff} = 0$
  - H_A: $\mu_{diff} \neq 0$

- First, start by "assuming" null is true and computing distance (in SEs) between $\bar{x}_{diff}$ and 0
  - Sample result is 7.8 SEs below 0—*is this unusual?*

$$t = \frac{\bar{x}_{diff} - 0}{S\hat{E}(\bar{x})} = \frac{-2.75}{2.83/\sqrt{65}} = -7.8$$

Getting a p-Value

- Sample result is 7.8 SEs below 0—*is this unusual?*
  - See where this falls on sampling distribution of all possible mean differences based on random samples of 65 patients
    - Theory tells us this is normal

- The p-value is probability of being 7.8 or more standard errors from 0 under a standard normal curve
  - Without looking up, we know p <<< .001!

Everything with Stata

- ttesti 65 -2.75 2.83 0

```
. ttesti  65 -2.75 2.83 0

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      65       -2.75    .3510183        2.83    -3.45124    -2.04876
------------------------------------------------------------------------------
    mean = mean(x)                                                t =  -7.8343
Ho: mean = 0                                     degrees of freedom =       64

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Oat Bran and LDL Cholesterol

- Cereal and cholesterol: 14 males with high cholesterol given oat bran cereal as part of diet for two weeks, and corn flakes cereal as part of diet for two weeks

	Corn Flakes	Oat Bran	Difference
Mean ($\bar{x}$)	**4.44 mmol/dL**	**4.08**	**0.36**
sd (*s*)	**1.0**	**1.1**	**0.40**

[1]Example based on data taken from Pagano, M. (2000). *Principles of Biostatistics*, 2nd ed. Duxbury Press. Based on research by Anderson J, et al. (1990). Oat Bran Cereal Lowers … *American Journal of Clinical Nutrition*, 52.

95% Confidence Interval

- 95% CI for difference in mean LDL, corn flakes vs. oat bran

$$\bar{x}_{diff} \pm t_{.95,13} \times S\hat{E}(\bar{x}_{diff})$$

$$\bar{x}_{diff} \pm 2 \times \frac{s_{diff}}{\sqrt{14}}$$

$$0.36 \pm 2 \times \left(\frac{.040}{\sqrt{14}} \right)$$

$$0.13 \ \ to \ \ 0.60 \ mmol/dL$$

Getting a p-Value

- Hypotheses
  - H_o: $\mu_{diff} = 0$
  - H_A: $\mu_{diff} \neq 0$

- First, start by "assuming" null is true, and computing distance (in SEs) between $\bar{x}_{diff}$ and 0
  - Sample result is 3.3 SEs above 0—*is this unusual?*

$$t = \frac{\bar{x}_{diff} - 0}{S\hat{E}(\bar{x})} = \frac{.036}{.04/\sqrt{14}} \approx 3.3$$

Getting a p-Value

- Sample result is 3.3 SEs above 0—*is this unusual?*
  - See where this falls on sampling distribution of all possible mean differences based on random samples of 14 patients: theory tells us this is t_{13}

- The p-value is probability of being 3.3 or more standard errors from 0 under a t_{13} curve: look up in table or go to Stata

Everything with Stata

- cii 14 .36 .40 0

```
. ttesti 14 .36 .40 0

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      14         .36    .1069045          .4    .1290469    .5909531
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   3.3675
Ho: mean = 0                                     degrees of freedom =       13

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.9975         Pr(|T| > |t|) = 0.0050          Pr(T > t) = 0.0025
```

Direction of Comparison is Arbitrary

- Does not impact overall results at all, direction changes, so signs of mean diff and CI endpoints change; but message exactly the same

```
. ttesti 14 -.36 .40 0

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      14        -.36    .1069045          .4   -.5909531   -.1290469
------------------------------------------------------------------------------
    mean = mean(x)                                                t =  -3.3675
Ho: mean = 0                                     degrees of freedom =       13

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.0025         Pr(|T| > |t|) = 0.0050          Pr(T > t) = 0.9975
```

Summary: Paired t-Test

- Designate null and alternative hypotheses
- Collect data
- Compute difference in outcome for each paired set of observations
  - Compute $\bar{x}_{diff}$, sample mean of the paired differences
  - Compute *s*, sample standard deviation of the differences

Summary: Paired t-Test

- Compute 95% (or other level) CI for true mean difference between paired groups compared
  - "Big *n*" ($n > 60$)

$$\bar{x}_{diff} \pm 2 \times \frac{s_{diff}}{\sqrt{n}}$$

  - "Small *n*" ($n \leq 60$)

$$\bar{x}_{diff} \pm t_{.95,n-1} \times \frac{s_{diff}}{\sqrt{n}}$$

Summary: Paired t-Test

- To get p-values
  - Start by assuming H_o true
  - Measure distance of sample result from μ_o

$$t = \frac{\bar{x}_{diff} - \mu_o}{S\hat{E}(\bar{x}_{diff})}$$

  - Usually, μ_o=0, so:

$$t = \frac{\bar{x}_{diff}}{S\hat{E}(\bar{x}_{diff})} = \frac{\bar{x}_{diff}}{s_{diff}/\sqrt{n}}$$

Summary: Paired t-Test

- Compare test statistics (distance) to appropriate distribution to get p-value
  - Reminder: p-value measures how likely your sample result (and other result less likely) are if null is true

Summary: Paired t-Test/Paired Data Situations

- Example 1
  - The blood pressure/OC example
- Example 2
  - Degree of clinical agreement, each patient received two assessments
- Example 3
  - Single group of men given two different diets at in two different time periods
  - LDL cholesterol levels measured at end of each diet

Summary: Paired t-Test/Paired Data Situations

- Twin study

- Matched case control scenario
  - Suppose we wish to compare levels of a certain biomarker in patients with a given disease versus those without

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section D

The p-Value *in Even More Detail!*

p-Values

- p-values are probabilities (numbers between 0 and 1)

- Small p-values mean that the sample results are unlikely when the null is true

- The p-value is the probability of obtaining a result as extreme or more extreme than you did by chance alone assuming the null hypothesis H_0 is true
  - How likely your sample result (and other result less likely) are if null is true

p-Values

- The p-value is not the probability that the null hypothesis is true!

- The p-value alone imparts no information about scientific/ substantive content in result of a study

- Example: from Example 3, the researchers found a statistically significant (p=0.005!) difference in average LDL cholesterol levels in men who had been on a diet including corn flakes versus the same men on a diet including oat bran cereal
  - Which diet showed lower average LDL levels?
  - How much was the difference; does it mean anything nutritionally?

p-Values

- If the p-value is small either a very rare event occurred and
  - H_0 is true

    or

  - H_0 is false

- Type I error
  - Claim H_A is true when in fact H_0 is true
  - The probability of making a Type I error is called the *alpha-level (α-level)* or *significance level*

Note on the p-Value and the Alpha-Level

- If the p-value is less then some pre-determined cutoff (e.g., .05), the result is called *statistically significant*

- This cutoff is the *α-level*
  - α-level is the probability of a type I error
  - It is the probability of falsely rejecting H_0 when H_0 true

- Idea: to keep the chance of "making a mistake" when the H_0 is true low and only reject if the sample result is "unlikely"
  - Unlikeliness threshold is determined by *α-level*

Note on the p-Value and the Alpha-Level

- Truth versus decision made by hypothesis testing

	TRUTH	
	H_o	H_A
Reject H_o	Type I Error alpha-level	Power 1-beta
Not Reject H_o		Type II Error beta

Note on the p-Value and the Alpha-Level

- Truth versus decision made by hypothesis testing

	TRUTH	
	H_o	H_A
Reject H_o	Type I Error alpha-level	Power 1-beta
Not Reject H_o		Type II Error beta

Note on the p-Value and the Alpha-Level

- Truth versus decision made by hypothesis testing

	TRUTH	
	H_o	H_A
Reject H_o	Type I Error alpha-level	Power 1-beta
Not Reject H_o		Type II Error beta

Note on the p-Value and the Alpha-Level

- Truth versus decision made by hypothesis testing

	TRUTH	
	H_o	H_A
Reject H_o	Type I Error alpha-level	Power 1-beta
Not Reject H_o		Type II Error beta

More on p-Value: One-Sided vs. Two-Sided Controversy

- Two-sided p-value (BP/OC: p = .009)
  - Probability of a result as or more extreme than observed (either positive or negative)

- One-sided p-value
  - Probability of a more extreme positive result than observed or a more extreme negative result: only considers extremes in one direction of null when evaluation how likely your sample result is (and results less likely)
  - If the direction of the alternative hypothesis in the one-sided test is the same as the direction of the sample result in terms of above/below the null, then the one-sided p-value with be half the two-sided p-value

Stata Output

- One-sided alternative: true mean difference >0
  - Sample mean difference was greater than 0

```
. ttesti 10 4.8 4.6 0

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      10         4.8    1.454648         4.6    1.509358    8.090642
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   3.2998
Ho: mean = 0                                     degrees of freedom =        9

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.9954         Pr(|T| > |t|) = 0.0092          Pr(T > t) = 0.0046
```

Stata Output

- One-sided alternative: true mean difference <0
  - Sample mean difference was greater than 0

```
. ttesti 10 4.8 4.6 0

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      10         4.8    1.454648         4.6    1.509358    8.090642
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   3.2998
Ho: mean = 0                                     degrees of freedom =        9

    Ha: mean < 0                 Ha: mean != 0                  Ha: mean > 0
 Pr(T < t) = 0.9954         Pr(|T| > |t|) = 0.0092          Pr(T > t) = 0.0046
```

More on the p-Value

- In some cases, a one-sided alternative may not make scientific sense
  - In the absence of pre-existing information, in evaluating the BP/OC relationship, wouldn't either result be interesting and useful? (i.e., negative or positive association?)

- In some cases, a one-sided alternative often makes scientific sense
  - For example: not really interested if new treatment is worse than old treatment—only care whether it's better

- However: because of "culture of p-value" and sanctity of ".05," one-sided p-values are viewed with suspicion

- In this course, we will use two-sided p-values exclusively

Connection: Hypothesis Testing and CIs

- The confidence interval gives plausible values for the population parameter
  - "Data take me to the truth"

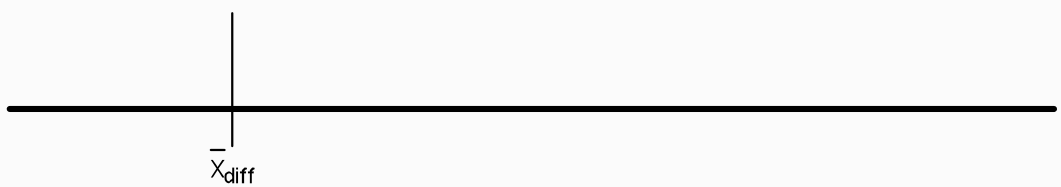
- Hypothesis testing postulates two choice for the population parameter
  - "Here are two possibilities for the truth; data help me choose one"

95% Confidence Interval

- If 0 is not in the 95% CI, then we would reject H_0 that $\mu = 0$ at level $a = .05$ (the p-value < .05)

- Why?

- With confidence interval we start at sample mean difference and go two standard errors in either direction (or slightly more in small samples)

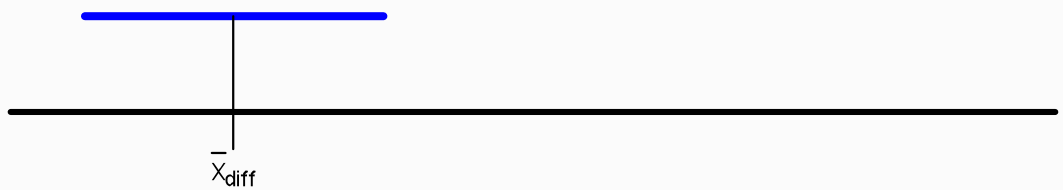


95% Confidence Interval

- If 0 is not in the 95% CI, then we would reject H_0 that $\mu = 0$ at level $a = .05$ (the p-value < .05)

- Why?

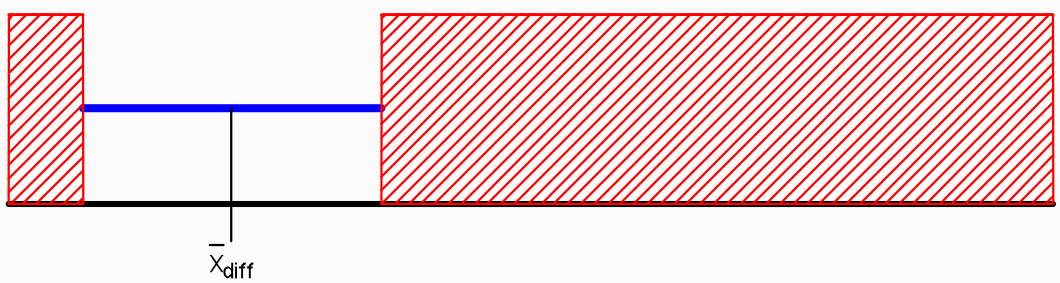
- With confidence interval we start at sample mean difference and go two standard errors in either direction (or slightly more in small samples)



95% Confidence Interval

- If 0 is not in the 95% CI, then this must mean $\bar{x}$ is > 2 standard errors away from 0 (either above or below)

- Hence, the distance (*t*) will be > 2 or < -2: and the resulting p-value <.05



95% Confidence Interval and p-Value

- In the BP/OC example, the 95% confidence interval tells us that the p-value is less than .05, but it doesn't tell us that it is p = .009
- The confidence interval and the p-value are complementary
- However, you can't get the exact p-value from just looking at a confidence interval, and you can't get a sense of the scientific/ substantive significance of your study results by looking at a p-value

More on the p-Value

- Statistical significance does not imply/prove causation
- For example: in the blood pressure/oral contraceptives example, there could be other factors that could explain the change in blood pressure
- A significant p-value is only ruling out random sampling (chance) as the explanation
- Need a comparison group to better establish causality
  - Self-selected (may be okay)
  - Randomized (better)

More on the p-Value

- *Statistical significance* is not the same as *scientific significance*

- Hypothetical example: blood pressure and oral contraceptives:
  - Suppose:
    - n = 100,000; $\bar{x}_{diff}$ = .03 mmHg; *s* = 4.6 mmHg
    - p-value = .04

- Big *n* can sometimes produce a small p-value, even though the magnitude of the effect is very small (not scientifically/ substantively significant)

- Very important
  - Always report a confidence interval
  - 95% CI: 0.002–0.058 mmHg

More on the p-Value

- *Lack of statistical significance* is not the same as *lack of scientific significance*
  - Must evaluate in context of study, sample size

- Small *n* can sometimes produce a non-significant even though the magnitude of the association at the population level is real and important (our study just can't detect it)

- *Low power* in small sample studies makes not rejecting hard to interpret

- Sometimes small studies are designed without power in mind just to generate preliminary data