JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

The Paired t-test and Hypothesis Testing

John McGready
Johns Hopkins University

Lecture Topics

- Comparing two groups—the paired data situation
- Hypothesis testing—the null and alternative hypotheses
- p-values—definition, calculations, and more information

Section A

The Paired t-Test and Hypothesis Testing

Comparison of Two Groups

- Are the population means different? (continuous data)

Comparison of Two Groups

Two Situations

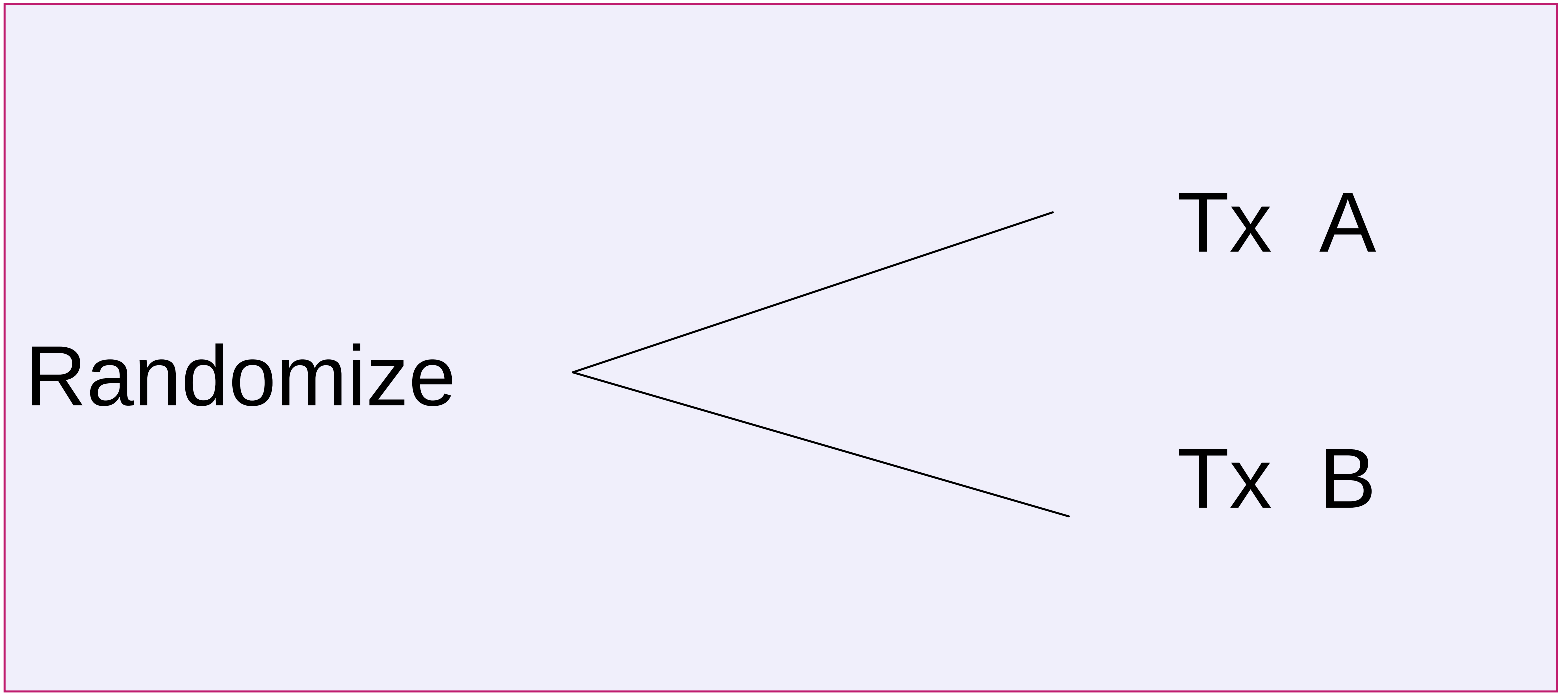
1. Paired Design
   - Before-after data
   - Twin data
   - Matched case-control

Continued

Comparison of Two Groups

Two Situations

2. Two Independent Sample Designs



Paired Design

Before vs. After

- Why pairing?
  - Control extraneous noise
  - Each observation acts as a control

Example: 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

Continued

Example: Blood Pressure and Oral Contraceptive Use

	BP Before OC	BP After OC	After-Before
7.	126	132	6
8.	105	109	4
9.	104	102	-2
10.	115	117	2
Sample Mean:	115.6	120.4	4.8
	$\overline{X}_{before}$	$\overline{X}_{after}$	$\overline{X}_{diff}$

Continued

Example: Blood Pressure and Oral Contraceptive Use

	BP Before OC	BP After OC	After-Before
7.	126	132	6
8.	105	109	4
9.	104	102	-2
10.	115	117	2
Sample Mean:	115.6	**120.4**	4.8
	$\overline{X}_{before}$	$\overline{X}_{after}$	$\overline{X}_{diff}$

Continued

Example: Blood Pressure and Oral Contraceptive Use

	BP Before OC	BP After OC	After-Before
7.	126	132	6
8.	105	109	4
9.	104	102	-2
10.	115	117	2
Sample Mean:	**115.6**	120.4	4.8
	$\bar{X}_{before}$	$\bar{X}_{after}$	$\bar{X}_{diff}$

Continued

Example: Blood Pressure and Oral Contraceptive Use

	BP Before OC	BP After OC	After-Before
7.	126	132	6
8.	105	109	4
9.	104	102	-2
10.	115	117	2
Sample Mean:	115.6	120.4	**4.8**
	$\bar{X}_{before}$	$\bar{X}_{after}$	$\bar{X}_{diff}$

Continued

Example: Blood Pressure and Oral Contraceptive Use

- The sample average of the differences is 4.8
- The sample standard deviation (s) of the differences is s = 4.6

Example: Blood Pressure and Oral Contraceptive Use

- Standard deviation of differences found by using the formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

- Where each X_i represents an individual difference, and $\overline{X}$ is the mean difference

Continued

Example: Blood Pressure and Oral Contraceptive Use

- Notice, we can get $\overline{X}_{diff}$ by
  $\overline{X}_{after} - \overline{X}_{before}$
  (120.4 – 115.6 = 4.8)
- However, we need to compute the individual differences first, to get s_{diff}

Note

- In essence, what we have done is reduced 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

95% Confidence Interval

- 95% confidence interval for mean change in BP
- $4.8 \pm t_9 \times \text{SEM}$
- Where SEM = $\dfrac{s}{\sqrt{n}}$

95% Confidence Interval

- 95% confidence interval for mean change in BP

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

Continued

95% Confidence Interval

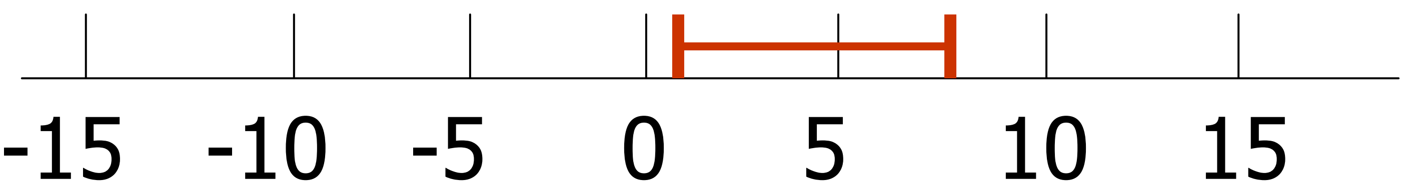
- 95% confidence interval for mean change in BP

$$4.8 \pm 2.26 \times 1.45$$

1.52 to 8.07

Notes

- The number 0 is **NOT** in confidence interval (1.53–8.07)



Continued

Notes

- The number 0 is **NOT** in confidence interval (1.53–8.07)
  - Because 0 is not in the interval, this suggests there is a non-zero change in BP over time

Continued

Notes

- The BP change could be due to factors other than oral contraceptives
  - A control group of comparable women who were not taking oral contraceptives would strengthen this study

Hypothesis Testing

- 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?

Continued

Hypothesis Testing

- Sometimes statisticians use the term **expected** for the population average
- μ is the expected (population) mean change in blood pressure

Continued

Hypothesis Testing

- Null hypothesis:
- Alternative hypothesis:

H_0: $\mu = 0$

H_A: $\mu \neq 0$

- We reject H_0 if the sample mean is far away from 0

The Null Hypothesis, H_0

- Typically represents the hypothesis that there is "no association" or "no difference"
- It represents current "state of knowledge" (i.e., no conclusive research exists)
  - For example, there is no association between oral contraceptive use and blood pressure

$$H_0: \ \mu = 0$$

The Alternative Hypothesis H_A (or H_1)

- Typically represents what you are trying to prove
  - 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_0"

Continued

Hypothesis Testing

- We start by assuming the null (H_0) is true, and asking . . .
  - "How likely is the result we got from our sample?"

Hypothesis Testing Question

- Do our sample results allow us to reject H_0 in favor of H_A?
  - $\overline{X}$ would have to be far from zero to claim H_A is true
  - But is $\overline{X} = 4.8$ big enough to claim H_A is true?

Continued

Hypothesis Testing Question

- Do our sample results allow us to reject H_0 in favor of H_A?
  - Maybe we got a big sample mean of 4.8 from a chance occurrence
  - Maybe H_0 is true, and we just got an unusual sample

Continued

Hypothesis Testing Question

- Does our sample results allow us to reject H_0 in favor H_A?
  - We need some measure of how probable the result from our sample is, if the null hypothesis is true

Continued

Hypothesis Testing Question

- Does our sample results allow us to reject H_0 in favor H_A?
  - What is 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)

Continued

Hypothesis Testing Question

- Does our sample results 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

The p-value

- So what can we turn to evaluate how unusual our sample statistic is when the null is true?

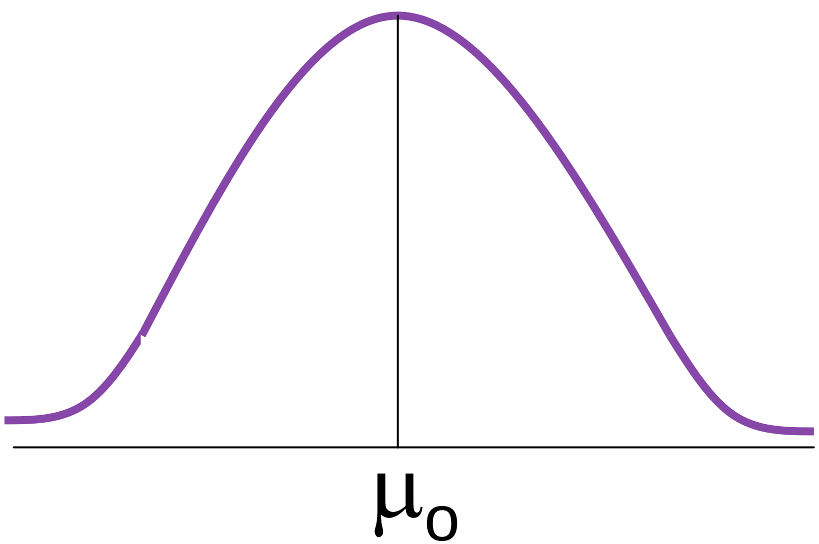
Continued

The p-value

- 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!

Sampling Distribution

- *Sampling distribution of the sample mean* is the distribution of all possible values of $\overline{X}$ from samples of same size, n



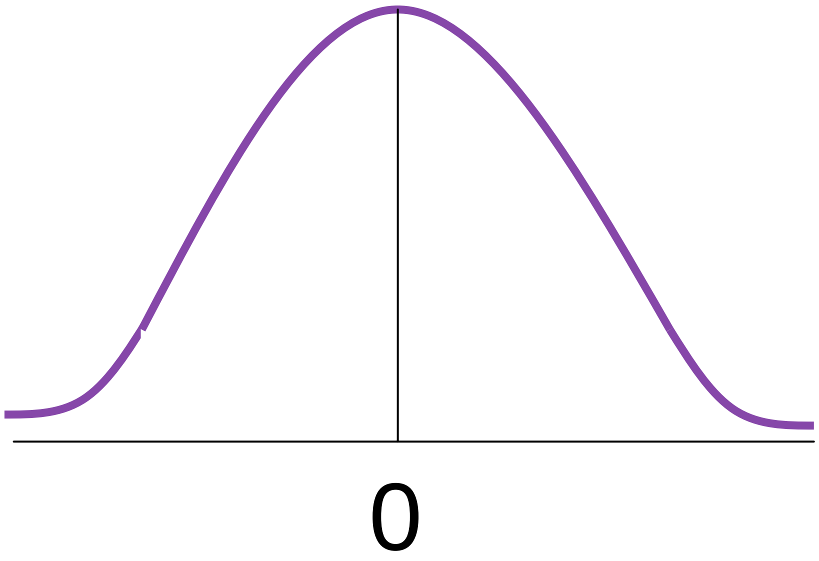
Continued

Sampling 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

Blood Pressure-OC Example

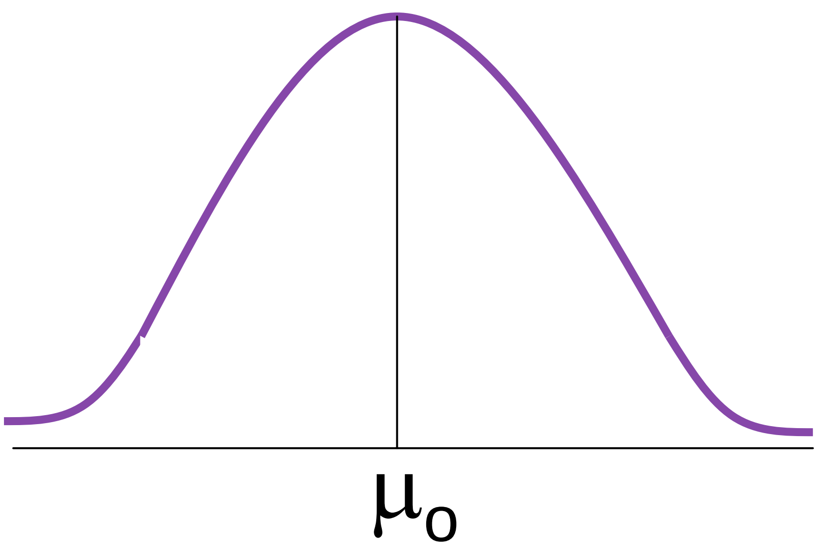
- *Sampling distribution* is the distribution of all possible values of $\bar{X}$ from random samples of 10 women each



Continued

Getting a p-value

- To compute a p-value, we need to find our value of $\overline{X}$, and figure out how "unusual" it is



Continued

Getting a p-value

- In other words, we will use our knowledge about the sampling distribution of $\overline{X}$ to figure out what proportion of samples from our population would have sample mean values as far away from 0 or farther, than our sample mean of 4.8

Section A

Practice Problems

Practice Problems

1. Which of the following examples involve the comparison of paired data?
   – If so, on what are we pairing the data?

Practice Problems

a. In Baltimore, a real estate practice known as "flipping" has elicited concern from local/federal government officials
   - "Flipping" occurs when a real estate investor buys a property for a low price, makes little or no improvement to the property, and then resells it quickly at a higher price

Continued

Practice Problems

a. This practice has raised concern, because the properties involved in "flipping" are generally in disrepair, and the victims are generally low-income
   – Fair housing advocates are launching a lawsuit against three real estate corporations accused of this practice

Continued

Practice Problems

a. As part of the suit, these advocates have collected data on all houses (purchased by these three corporations) which were sold in less than one year after they were purchased
   - Data were collected on the purchase price and the resale price for each of these properties

Continued

Practice Problems

a. The data were collected to show that the resale prices were, on average, higher than the initial purchase price
   - A confidence interval was constructed for the average profit in these quick turnover sales

Continued

Practice Problems

b. Researchers are testing a new blood pressure-reducing drug; participants in this study are randomized to either a drug group or a placebo group
   – Baseline blood pressure measurements are taken on both groups and another measurement is taken three months after the administration of the drug/placebo

Practice Problems

b. Researchers are curious as to whether the drug is more effective in lowering blood pressure than the placebo

Continued

Practice Problems

2. Give a one sentence description of what the p-value represents in hypothesis testing

Section A

Practice Problem Solutions

Solutions

1(a). The “flipping” example

- In this example, researchers were comparing the difference in resale price and initial purchase price for each property in the sample
- This data is paired and the “pairing unit” is each property

Continued

Solutions

1(b). "Miracle" blood pressure treatment

- Researchers used "before" and "after" blood pressure measurements to calculate individual, person-level differences

Continued

Solutions

1(b). "Miracle" blood pressure treatment

- To evaluate whether the drug is effective in lowering blood pressure, the researchers will want to test whether the mean differences are the same amongst those on treatment and those on placebo
- So the comparison will be made between two different groups of individuals

Continued

Solutions

2. The p-value is the probability of seeing a result as extreme or more extreme than the result from a given sample, if the null hypothesis is true

Section B

The p-value in Detail

Blood Pressure and Oral Contraceptive Use

- Recall the results of the example on BP/OC use from the previous lecture
  - Sample included 10 women
  - Sample Mean Blood Pressure Change—4.8 mmHg (sample SD, 4.6 mmHg)

How Are p-values Calculated?

- What is the probability of having gotten a sample mean as extreme or more extreme then 4.8 if the null hypothesis was true (H_0: $\mu = 0$)?
  - The answer is called the p-value
  - In the blood pressure example, $p = .0089$

Continued

How Are p-values Calculated?

- 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

Continued

How Are p-values Calculated?

$$t = \frac{\text{sample mean} - 0}{\text{SEM}}$$

$$t = \frac{4.8}{1.45} = 3.31$$

- The value t = 3.31 is called the **test statistic**

Continued

How Are p-values Calculated?

- We observed a sample mean that was 3.31 standard errors of the mean (SEM) away from what we would have expected the mean to be if OC use was not associated with blood pressure

Continued

How Are p-values Calculated?

- Is a result 3.31 standard errors above its mean unusual?
  - It depends on what kind of distribution we are dealing with

Continued

How Are p-values Calculated?

- The p-value is the probability of getting a test statistic as (or more) extreme than what you observed (3.31) by chance if H_0 was true
- The p-value comes from the sampling distribution of the sample mean

Sampling Distribution of the Sample Mean

- Recall what we know about the sampling distribution of the sample mean, $\overline{X}$
  - If our sample is large (n > 60), then the sampling distribution is approximately normal

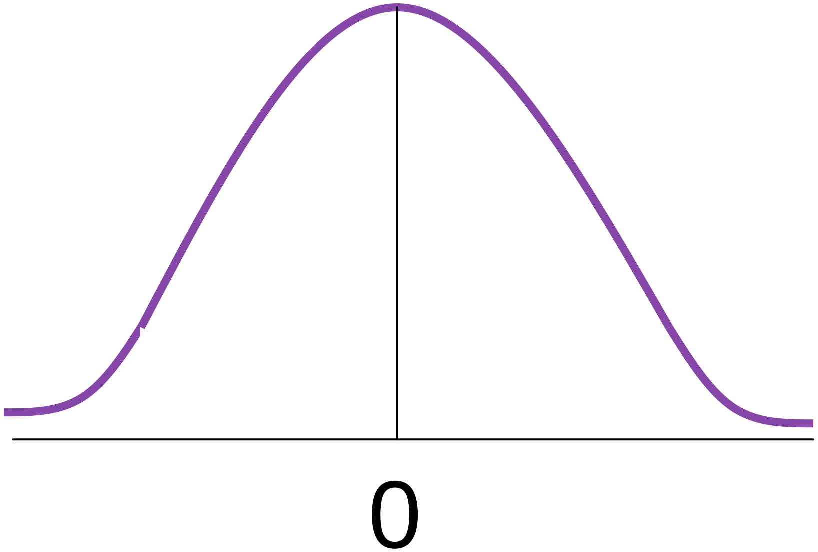
Continued

Sampling Distribution of the Sample Mean

- Recall what we know about the sampling distribution of the sample mean, $\overline{X}$
  - With smaller samples, the sampling distribution is a t-distribution with n-1 degrees of freedom

Blood Pressure and Oral Contraceptive Use

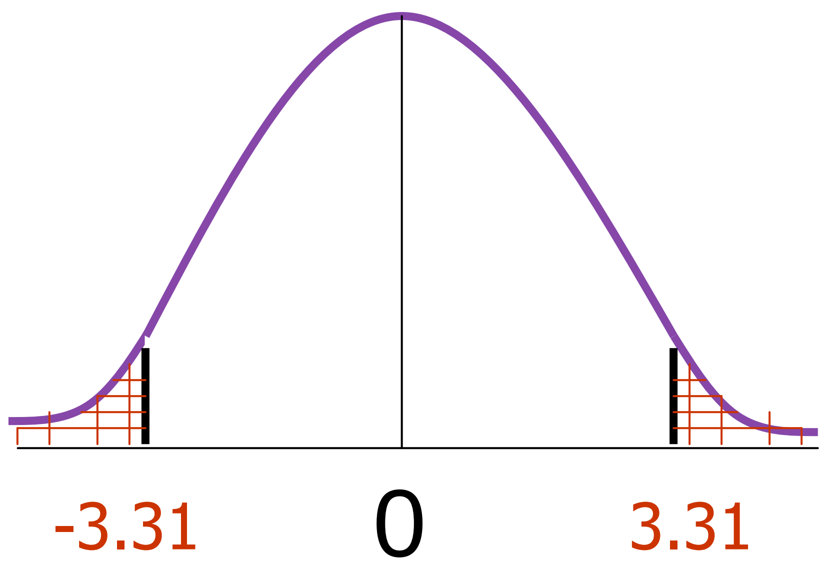
- So in the BP/OC example, we have a sample of size 10, and hence a sampling distribution that is t-distribution with 10 - 1 = 9 degrees of freedom



Continued

Blood Pressure and Oral Contraceptive Use

- To compute a p-value, we would need to compute the probability of being 3.31 or more standard errors away from 0 on a t_9 curve



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 $\bar{X}$ s μ_0

- For the BP-OC data:

ttesti 10 4.8 4.6 0

Stata Output

```
. 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
------------------------------------------------------------------------------
Degrees of freedom: 9

                               Ho: mean(x) = 0

     Ha: mean < 0               Ha: mean != 0                 Ha: mean > 0
       t =   3.2998                t =   3.2998                 t =   3.2998
   P < t =   0.9954          P > |t| =   0.0089             P > t =   0.0045
```

Interpreting Stata Output

s

```
. 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
------------------------------------------------------------------------------
Degrees of freedom: 9
```

n - 1

Interpreting Stata Output

H_o

H_A

```
                         Ho: mean(x) = 0

 Ha: mean < 0             Ha: mean != 0             Ha: mean > 0
   t =   3.2998             t =   3.2998              t =   3.2998
P < t =   0.9954      P > |t| =   0.0089          P > t =   0.0045
```

p-value

Note: “!=” is computer speak for “not equal”

Interpreting the p-value

- The p-value in the blood pressure/OC example is .0089
  - Interpretation—If the true before OC/after OC blood pressure difference is 0 amongst all women taking OC's, then the chance of seeing a mean difference as extreme/more extreme as 4.8 in a sample of 10 women is .0089

Using the p-value to Make a Decision

- Recall, we specified two competing hypotheses about the underlying, true mean blood pressure change, μ

$$H_0: \mu = 0$$

$$H_A: \mu \neq 0$$

Continued

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 $p = 0.0089$

Continued

Using the p-value to Make a Decision

- Establishing a cutoff
  - In general, to make a decision about what p-value constitute “unusual” results, there needs to be a cutoff, such that all p-values less than the cutoff result in rejection of the null

Continued

Using the p-value to Make a Decision

- Establishing a cutoff
  - Standard cutoff is .05—this is an arbitrary value
  - Cut off is called "α-level" of the test

Continued

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)

Continued

Blood Pressure

Oral Contraceptive Example

- **Result**
  - Blood pressure measurements increased on average 4.8 mm Hg with standard deviation 4.6 mmHg
  - The 95% confidence interval for the mean change was 1.5 mmHg - 8.1 mmHg

Continued

Blood Pressure

Oral Contraceptive Example

- **Result**
  - The blood pressure measurements after oral contraceptive use were statistically significantly higher than before oral contraceptive use (p=.009)

Continued

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

Summary: Paired t-test

- The paired t-test is a useful statistical tool for comparing mean differences between two populations which have some sort of "connection" or link

Continued

Summary: Paired t-test

- Example one
  - The blood pressure/OC example
- Example two
  - Study comparing blood cholesterol levels between two sets of fraternal twins—one twin in each pair given six weeks of diet counseling

Continued

Summary: Paired t-test

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

Continued

Summary: Paired t-test

- Designate null and alternative hypotheses
- Collect data

Continued

Summary: Paired t-test

- Compute difference in outcome for each paired set of observations
  - Compute $\bar{X}$, sample mean of the paired differences
  - Compute s, sample standard deviation of the differences

Continued

Summary: Paired t-test

- Compute test statistic

$$t = \frac{\overline{X} - \mu_o}{\textbf{SEM}}$$

- Usually, just:

$$t = \frac{\overline{X}}{\textbf{SEM}}$$

Continued

Summary: Paired t-test

- Compare test statistic to appropriate distribution to get p-value

Section B

Practice Problems

Practice Problems

- Eight counties were selected from State A
- Each of these counties was matched with a county from State B, based on factors, e.g.,
  - Mean income
  - Percentage of residents living below the poverty level
  - Violent crime rate
  - Infant mortality rate (IMR) in 1996

Continued

Practice Problems

- Information on the infant mortality rate in 1997 was collected on each set of eight counties
- IMR is measured in deaths per 10,000 live births
- A pre- and post-neonatal care program was implemented in State B at the beginning of 1997

Continued

Practice Problems

- This data is being used to compare the IMR rates in States A and B in 1997
  - This comparison will be used as part of the evaluation of the neonatal care program in State B, regarding its effectiveness on reducing infant mortality

Continued

Practice Problems

- The data is as follows:

Pair	IMR -State A	IMR - State B
1	80	76
2	130	112
3	88	97
4	98	67
5	103	107
6	121	116
7	83	94
8	93	78

Continued

Practice Problems

1. What is the appropriate method for testing whether the mean IMR is the same for both states in 1997?
2. State your null and alternative hypotheses
3. Perform this test by hand
4. Confirm your results in Stata

Continued

Practice Problems

5. What would your results be if you had 32 county pairs and the mean change and standard deviation of the changes were the same?

Section B

Practice Problem Solutions

Solutions

1. What is the appropriate test for testing whether the mean IMR is the same for both states?
   - *Because the data is paired, and we are comparing two groups, we should use the paired t-test*

Solutions

2. State your null and alternative hypotheses
   – *Three possible ways of expressing the hypotheses . . .*

$H_o: \mu_A = \mu_B$ $\quad$ $H_o: \mu_B - \mu_A = 0$ $\quad$ $H_o: \mu_{diff} = 0$

$H_A: \mu_A \neq \mu_B$ $\quad$ $H_A: \mu_A - \mu_B \neq 0$ $\quad$ $H_A: \mu_{diff} \neq 0$

Continued

Solutions

2. State your null and alternative hypotheses
   - *Three possible ways of expressing the hypotheses . . .*

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

Continued

Solutions

2. State your null and alternative hypotheses
   – *Three possible ways of expressing the hypotheses . . .*

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

Continued

Solutions

2. State your null and alternative hypotheses
   – *Three possible ways of expressing the hypotheses . . .*

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

Continued

Solutions

3. Perform this test by hand
   - *Remember, in order to do the paired test, we must first calculate the difference in IMR with in each pair*
   - *I will take the difference to be IMRB – IMRA*

Solutions

3. Perform this test by hand
   – *Once the differences are calculated, you need to calculate* $\bar{X}_{diff}$ *and* s_{diff}

$\bar{X}_{diff}$ = – 6.13 (deaths per 10,000 live births)

s_{diff} = 14.5 (deaths per 10,000 live births)

Continued

Solutions

3. Perform this test by hand
   – *To calculate our test statistic . . .*

$$t = \frac{\bar{X}_{diff} - \mu_o}{s_{diff} \big/ \sqrt{n}} = \frac{-6.13 - 0}{14.5 \big/ \sqrt{8}} = \frac{-6.13}{5.12} = -1.2$$

Continued

Solutions

3. Perform this test by hand
   - *We need to compare our test-statistic to a t-distribution with 8–1=7 degrees of freedom. Consulting our table, we see we must be at least 2.3 standard errors from the mean (below or above) for the p-value to be .05 or less*
   - *We are 1.2 SEs below; therefore, our p-value will be larger than .05*

Continued

Solutions

3. Perform this test by hand
   – *Since $p > .05$, we would fail to conclude there was a difference in mean IMR for State A and State B*
   – *This is as specific as we can get about the p-value from our t-table*

Continued

Solutions

4. Confirm your results in Stata

```
. ttesti 8 -6.13 14.5 0

------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |       8       -6.13    5.126524        14.5    -18.2523    5.992303
------------------------------------------------------------------------------
Degrees of freedom: 7

                          Ho: mean(x) = 0

   Ha: mean < 0              Ha: mean ~= 0              Ha: mean > 0
    t =  -1.1957              t =  -1.1957             t =  -1.1957
  P < t =   0.1354          P > |t| =   0.2707        P > t =   0.8646
```

Continued

Solutions

4. Confirm your results in Stata

```
. ttesti 8 -6.13 14.5 0

------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |       8       -6.13    5.126524        14.5   -18.2523    5.992303
------------------------------------------------------------------------------
Degrees of freedom: 7

                               Ho: mean(x) = 0

    Ha: mean < 0               Ha: mean ~= 0               Ha: mean > 0
     t =  -1.1957               t =  -1.1957             t =  -1.1957
  P < t =   0.1354          P > |t| =   0.2707         P > t =   0.8646
```

Solutions

4. Confirm your results in Stata

```
. ttesti 8 -6.13 14.5 0

------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |       8       -6.13    5.126524        14.5   -18.2523    5.992303
------------------------------------------------------------------------------
Degrees of freedom: 7

                          Ho: mean(x) = 0

    Ha: mean < 0              Ha: mean ~= 0              Ha: mean > 0
     t =  -1.1957              t =  -1.1957              t =  -1.1957
  P < t =   0.1354          P > |t| =   0.2707          P > t =   0.8646
```

Continued

Solutions

5. What would your results be if you had 32 county pairs and the mean change and standard deviation of the changes were the same?

Continued

Solutions

```
. ttesti 32 -6.13 14.5 0

------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      32       -6.13    2.563262        14.5   -11.35781   -.9021925
------------------------------------------------------------------------------
Degrees of freedom: 31

                               Ho: mean(x) = 0

     Ha: mean < 0               Ha: mean ~= 0               Ha: mean > 0
       t =  -2.3915              t =  -2.3915              t =  -2.3915
   P < t =   0.0115          P > |t| =   0.0230          P > t =   0.9885
```

Section C

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/or more extreme than you did by chance alone assuming the null hypothesis H_0 is true

Continued

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

Continued

p-values

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

H_0 is true

OR

H_0 is false

Two Types of Errors in Hypothesis Testing

- Type I error
  - Claim H_A is true when in fact H_0 is true
- Type II error
  - Do not claim H_A is true when in fact H_A is true

Two Types of Errors in Hypothesis Testing

- The probability of making a Type I error is called the α-level
- The probability of **NOT** making a Type II error is called the power (we will discuss this later)

Two Types of Errors in Hypothesis Testing

DECISION	TRUTH: H_o	TRUTH: H_A
Reject H_o	Type I Error α-level	Power 1- β
Not Reject H_o		Type II Error β

Continued

Two Types of Errors in Hypothesis Testing

		TRUTH	
		H_o	H_A
DECISION	Reject H_o	Type I Error α-level	Power $1-\beta$
	Not Reject H_o		Type II Error β

Continued

Two Types of Errors in Hypothesis Testing

		TRUTH	
		H_o	H_A
DECISION	Reject H_o	Type I Error α-level	Power $1- \beta$
	Not Reject H_o		Type II Error β

Continued

Two Types of Errors in Hypothesis Testing

DECISION		TRUTH: H_o	TRUTH: H_A
	Reject H_o	Type I Error α-level	Power $1- \beta$
	Not Reject H_o		Type II Error β

Continued

Two Types of Errors in Hypothesis Testing

DECISION	TRUTH: H_o	TRUTH: H_A
Reject H_o	Type I Error α-level	Power 1- β
Not Reject H_o		Type II Error β

Continued

Two Types of Errors in Hypothesis Testing

DECISION		TRUTH: H_o	TRUTH: H_A
	Reject H_o	Type I Error α-level	Power 1- β
	Not Reject H_o		Type II Error β

Continued

Two Types of Errors in Hypothesis Testing

DECISION		TRUTH: H_o	TRUTH: H_A
	Reject H_o	Type I Error α-level	Power $1-\beta$
	Not Reject H_o		Type II Error β

Note on the p-value and the α-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**
- The **α-level** is the probability of a type I error
- It is the probability of falsely rejecting H_0

Notes on Reporting p-value

Incomplete Options

- The result is "statistically significant"
- The result is statistically significant at $\alpha = .05$
- The result is statistically significant ($p < .05$)

Note of the p-value and the α-level

- Best to give p-value and interpret
  - The result is significant ($p = .009$)

More on the p-value

The One-Sided Vs Two-Sided Controversy

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

Continued

More on the p-value

- One-sided p-value ($p = .0045$)
  - Probability of a more extreme positive result than observed

More on the p-value

- You never know what direction the study results will go
  - In this course, we will use two-sided p-values exclusively
  - The "appropriate" one sided p-value will be lower than its two-sided counterpart

Stata Output

```
. 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
------------------------------------------------------------------------------
Degrees of freedom: 9

                               Ho: mean(x) = 0

     Ha: mean < 0              Ha: mean != 0              Ha: mean > 0
       t =   3.2998              t =   3.2998              t =   3.2998
   P < t =   0.9954          P > |t| =   0.0089          P > t =   0.0045
```

Two-sided p-value in Stata

- Always from "middle" hypothesis

```
. 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
------------------------------------------------------------------------------
Degrees of freedom: 9

                               Ho: mean(x) = 0

     Ha: mean < 0               Ha: mean != 0                 Ha: mean > 0
       t =   3.2998                t =   3.2998                 t =   3.2998
   P < t =   0.9954          P > |t| =   0.0089             P > t =   0.0045
```

Connection Between Hypothesis Testing and Confidence Interval

- The confidence interval gives plausible values for the population parameter

95% Confidence Interval

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

Continued

95% Confidence Interval

- So, in this example, the 95% confidence interval tells us that the p-value is less than .05, but it doesn't tell us that it is $p = .009$

Continued

95% Confidence Interval

- 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 Causation

- Blood pressure example
  - There could be other factors that could explain the change in blood pressure

Blood Pressure Example

- A significant p-value is only ruling out random sampling (chance) as the explanation

Continued

Blood Pressure Example

- Need a comparison group
  - Self-selected (may be okay)
  - Randomized (better)

More on the p-value

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

More on the p-value

- Example: Blood Pressure and Oral Contraceptives
  - n = 100,000; $\overline{X}$ = .03 mmHg; **s** = 4.57
  - p-value = .04

Continued

More on the p-value

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

Continued

More on the p-value

- **Very Important**
  - Always report a confidence interval

    95% CI: 0.002 - 0.058 mmHg

The Language of Hypothesis (Significance Testing)

- Suppose p-value is .40
- How might this result be described?
  - Not statistically significant ($p = .40$)
  - Do not reject H_0

Continued

The Language of Hypothesis (Significance Testing)

- Can we also say?
  - Accept H_0
  - Claim H_0 is true
- Statisticians much prefer the double negative
  - "Do not reject H_0"

More on the p-value

- Not rejecting H_0 is not the same as accepting H_0

Continued

More on the p-value

- Example: Blood Pressure and Oral Contraceptives (sample of five women)
  - $n = 5$; $X = 5.0$; $s = 4.57$
  - p-value = .07
- We cannot reject H_0 at significance level $\alpha = .05$

Continued

More on the p-value

- But are we really convinced there is no association between oral contraceptives on blood pressure?
- Maybe we should have taken a bigger sample?

Continued

More on the p-value

- There is an interesting trend, but we haven't proven it beyond a reasonable doubt
- Look at the confidence interval
  - 95% CI (-.67, 10.7)

Section C

Practice Problems

Practice Problems

1. Why do you think there is such a controversy regarding one-sided versus two-sided p-values?
2. Why can a small mean difference in a paired t-test produce a small p-value if *n* is large?

Continued

Practice Problems

3. If you knew that the 90% CI for the mean blood pressure difference in the oral contraceptives example did NOT include 0, what could you say about the p-value for testing . . .

$$H_o: \mu_{diff} = 0$$

vs.

$$H_A: \mu_{diff} \neq 0 \ ?$$

Continued

Practice Problems

4. What if the 99% CI for mean difference did NOT include 0?
   - What could you say about the p-value?

Section C

Practice Problem Solutions

Solutions

1. Why do you think there is such a controversy regarding one-sided versus two-sided p-values?

- *If the "appropriate" one-sided hypothesis test is done (the one that best supports the sample data), the p-value will be half the p-value of the two sided test*

Continued

Solutions

1. Why do you think there is such a controversy regarding one-sided versus two-sided p-values?

- *This allows for situations where the two sided p-value is not statistically significant, but the one-sided p-value is*

Solutions

2. Why can a small mean difference in a paired t-test produce a small p-value if *n* is large?

- *When* n *gets large (big sample), the SEM gets very small. When SEM gets small, t gets large*

$$t = \frac{\overline{X}_{diff} - \mu_o}{SEM}$$

Continued

Solutions

3. If you knew that the 90% CI for the mean blood pressure difference in the oral contraceptives example did *not* include 0, what could you say about the p-value for testing:

$$H_o: \mu_{diff} = 0$$

$$\textit{vs.}\quad H_A: \mu_{diff} \neq 0?$$

The p-value is less than .10 ($p < .10$). This is as as specific as we can be with the given information.

Continued

Solutions

- What if the 99% CI for mean difference did not include 0? What could you say about the p-value?
- *The p-value is less than .01 ($p < .01$)*