Lecture 17: Paired Data and Difference of Two Means

Chapter 5.2, 5.1

Goals for Today

- Difference of means
- Paired differences of means
- Note on Practical vs Statistical Significance

The 8 Types of Questions

Here are the 8 broad types of questions we can answer with statistical methods (confidence intervals and hypothesis tests) in this class:

The 8 Types of Questions

Here are the 8 broad types of questions we can answer with statistical methods (confidence intervals and hypothesis tests) in this class:

1. What is the mean value μ?

The 8 Types of Questions

Here are the 8 broad types of questions we can answer with statistical methods (confidence intervals and hypothesis tests) in this class:

1. What is the mean value μ?
2. Are the means of two groups μ_1 and μ_2 equal or not?

The 8 Types of Questions

Here are the 8 broad types of questions we can answer with statistical methods (confidence intervals and hypothesis tests) in this class:

1. What is the mean value μ?
2. Are the means of two groups μ_1 and μ_2 equal or not?
3. What is the mean paired difference μ_{diff}?

The 8 Types of Questions

Here are the 8 broad types of questions we can answer with statistical methods (confidence intervals and hypothesis tests) in this class:

1. What is the mean value μ?
2. Are the means of two groups μ_1 and μ_2 equal or not?
3. What is the mean paired difference μ_{diff}?
4. What is the proportion p of "successes"?

The 8 Types of Questions

Here are the 8 broad types of questions we can answer with statistical methods (confidence intervals and hypothesis tests) in this class:

1. What is the mean value μ?
2. Are the means of two groups μ_1 and μ_2 equal or not?
3. What is the mean paired difference μ_{diff}?
4. What is the proportion p of “successes”?
5. Are the proportions of “successes” of two groups p_1 and p_2 equal or not?

The 8 Types of Questions

Here are the 8 broad types of questions we can answer with statistical methods (confidence intervals and hypothesis tests) in this class:

1. What is the mean value μ?
2. Are the means of two groups μ_1 and μ_2 equal or not?
3. What is the mean paired difference μ_{diff}?
4. What is the proportion p of "successes"?
5. Are the proportions of "successes" of two groups p_1 and p_2 equal or not?
6. Are the means $\mu_1, \ldots, \mu_k$ of k groups all equal or not?

The 8 Types of Questions

Here are the 8 broad types of questions we can answer with statistical methods (confidence intervals and hypothesis tests) in this class:

1. What is the mean value μ?
2. Are the means of two groups μ_1 and μ_2 equal or not?
3. What is the mean paired difference μ_{diff}?
4. What is the proportion p of "successes"?
5. Are the proportions of "successes" of two groups p_1 and p_2 equal or not?
6. Are the means $\mu_1, \ldots, \mu_k$ of k groups all equal or not?
7. Are we observing counts that we were expected?

The 8 Types of Questions

Here are the 8 broad types of questions we can answer with statistical methods (confidence intervals and hypothesis tests) in this class:

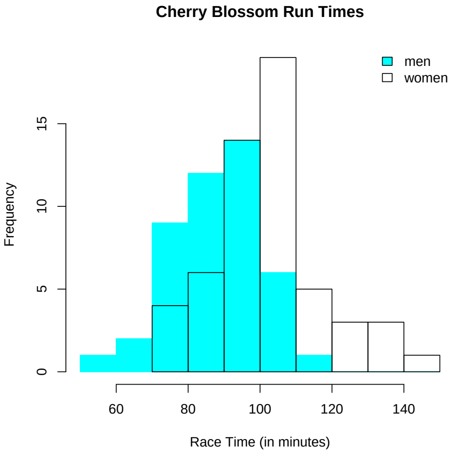
1. What is the mean value μ?
2. Are the means of two groups μ_1 and μ_2 equal or not?
3. What is the mean paired difference μ_{diff}?
4. What is the proportion p of "successes"?
5. Are the proportions of "successes" of two groups p_1 and p_2 equal or not?
6. Are the means $\mu_1, \ldots, \mu_k$ of k groups all equal or not?
7. Are we observing counts that we were expected?
8. Are two categorical variables independent?

Chapter 5.2: Are the means of 2 groups μ_1 and μ_2 equal?

Did men (n=45) run faster than women (n=55)?

Chapter 5.2: Are the means of 2 groups μ_1 and μ_2 equal?

Did men (n=45) run faster than women (n=55)?



Difference in Means

We are interested in the difference of two population means:

Difference in Means

We are interested in the difference of two population means:

- μ_w is the mean time for women
- μ_m is the mean time for men

Difference in Means

We are interested in the difference of two population means:

- μ_w is the mean time for women
- μ_m is the mean time for men

The data:

	men	women
$\overline{x}$	87.65	102.13
s	12.5	15.2
n	45	55

Difference of Means

We now recreate all the elements of Chapter 4 using this new population parameter $\mu_w - \mu_m$:

Difference of Means

We now recreate all the elements of Chapter 4 using this new population parameter $\mu_w - \mu_m$:

1. Determine a point estimate of $\mu_w - \mu_m$.

Difference of Means

We now recreate all the elements of Chapter 4 using this new population parameter $\mu_w - \mu_m$:

1. Determine a point estimate of $\mu_w - \mu_m$.
2. Show the normality of the sampling distribution: mean and SE

Difference of Means

We now recreate all the elements of Chapter 4 using this new population parameter $\mu_w - \mu_m$:

1. Determine a point estimate of $\mu_w - \mu_m$.
2. Show the normality of the sampling distribution: mean and SE
3. Build a confidence interval

Difference of Means

We now recreate all the elements of Chapter 4 using this new population parameter $\mu_w - \mu_m$:

1. Determine a point estimate of $\mu_w - \mu_m$.
2. Show the normality of the sampling distribution: mean and SE
3. Build a confidence interval
4. Conduct hypothesis tests

Difference of Means

We now recreate all the elements of Chapter 4 using this new population parameter $\mu_w - \mu_m$:

1. Determine a point estimate of $\mu_w - \mu_m$.
2. Show the normality of the sampling distribution: mean and SE
3. Build a confidence interval
4. Conduct hypothesis tests

First, the point estimate for $\mu_w - \mu_m$ is the sample difference of means

$$\bar{x}_w - \bar{x}_m = 102.13 - 87.65 = 14.48$$

Normality of Sampling Distribution

If the sample means $\overline{x}_1$ and $\overline{x}_2$

Normality of Sampling Distribution

If the sample means $\overline{x}_1$ and $\overline{x}_2$

- each meet the criteria for having nearly normal sampling distributions

Normality of Sampling Distribution

If the sample means $\overline{x}_1$ and $\overline{x}_2$

- each meet the criteria for having nearly normal sampling distributions
- also the observations from the two samples are independent

Normality of Sampling Distribution

If the sample means $\overline{x}_1$ and $\overline{x}_2$

- each meet the criteria for having nearly normal sampling distributions
- also the observations from the two samples are independent

then the difference in sample means $\overline{x}_1 - \overline{x}_2$ will also have a nearly normal sampling distribution...

Normality of Sampling Distribution

with

Normality of Sampling Distribution

with

- mean $\mu_1 - \mu_2$

Normality of Sampling Distribution

with

- mean $\mu_1 - \mu_2$
- estimated standard error

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Normality of Sampling Distribution

with

- mean $\mu_1 - \mu_2$
- estimated standard error

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Note the different s^2's and sample sizes.

Normality of Sampling Distribution

We verify the conditions:

Normality of Sampling Distribution

We verify the conditions:

- Because each sample consists of less than 10% of their respective populations (men: 45 of 7192 and women: 55 of 9732).

Normality of Sampling Distribution

We verify the conditions:

- Because each sample consists of less than 10% of their respective populations (men: 45 of 7192 and women: 55 of 9732).
- The observations for both groups don't look too skewed.

Normality of Sampling Distribution

We verify the conditions:

- Because each sample consists of less than 10% of their respective populations (men: 45 of 7192 and women: 55 of 9732).
- The observations for both groups don't look too skewed.
- Each sample has at least 30 observations (rule of thumb).

Normality of Sampling Distribution

We verify the conditions:

- Because each sample consists of less than 10% of their respective populations (men: 45 of 7192 and women: 55 of 9732).
- The observations for both groups don't look too skewed.
- Each sample has at least 30 observations (rule of thumb).
- The samples are independent (not paired or linked in any way).

Normality of Sampling Distribution

We verify the conditions:

- Because each sample consists of less than 10% of their respective populations (men: 45 of 7192 and women: 55 of 9732).
- The observations for both groups don't look too skewed.
- Each sample has at least 30 observations (rule of thumb).
- The samples are independent (not paired or linked in any way).

the sampling distribution is Normal with mean=$\mu_w - \mu_m$ and

$$SE_{\bar{x}_w - \bar{x}_m} = \sqrt{\frac{15.2^2}{55} + \frac{12.5^2}{45}} = 2.77$$

Confidence Interval

A 95% confidence interval for $\mu_1 - \mu_2$ is

$$(\text{point estimate for } \mu_1 - \mu_2) \pm 1.96 \times SE$$

Confidence Interval

A 95% confidence interval for $\mu_1 - \mu_2$ is

$$\begin{aligned}(\text{point estimate for } \mu_1 - \mu_2) &\pm 1.96 \times SE \\ (\bar{x}_1 - \bar{x}_2) &\pm 1.96 \times SE_{\bar{x}_1 - \bar{x}_2}\end{aligned}$$

Confidence Interval

A 95% confidence interval for $\mu_1 - \mu_2$ is

$$\begin{aligned}(\text{point estimate for } \mu_1 - \mu_2) \pm 1.96 \times SE \\ (\bar{x}_1 - \bar{x}_2) \pm 1.96 \times SE_{\bar{x}_1 - \bar{x}_2}\end{aligned}$$

So for the Cherry Blossom Run data, a 95% CI for $\mu_w - \mu_m$ is:

$$14.48 \pm 1.96 \times 2.77 = [9.05, 19.91]$$

Hypothesis Test

Chapter 5.1: Paired Data

Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Chapter 5.1: Paired Data

Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Examples:

- Cholesterol levels before and after some intervention

Chapter 5.1: Paired Data

Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Examples:

- Cholesterol levels before and after some intervention
- Disease rates amongst pairs of twins

Chapter 5.1: Paired Data

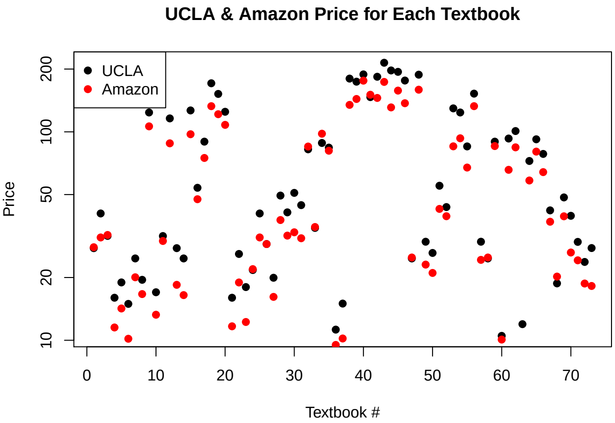
Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Examples:

- Cholesterol levels before and after some intervention
- Disease rates amongst pairs of twins
- In the text: price of the same textbook at the UCLA bookstore vs Amazon

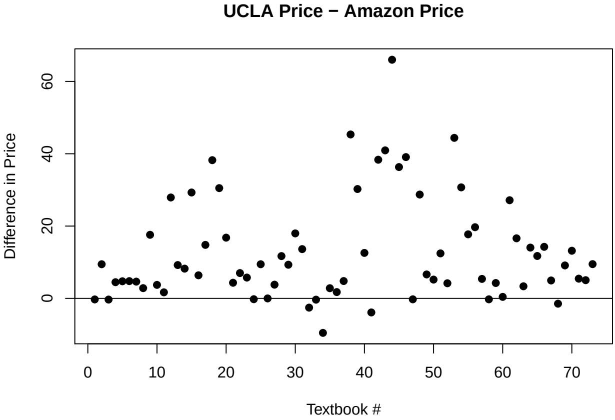
Paired Differences

The methodology for paired data remains the same, except our observations are the difference in pairs. Example, for the UCLA Bookstore vs Amazon book price example in the text



Paired Differences

The methodology for paired data remains the same, except our observations are the difference in pairs. Example, for the UCLA Bookstore vs Amazon book price example in the text



Paired Differences

We have

Paired Differences

We have

- population parameter is μ_{diff}

Paired Differences

We have

- population parameter is μ_{diff}
- point estimate $\overline{x}_{diff}$ of μ_{diff}

Paired Differences

We have

- population parameter is μ_{diff}
- point estimate $\overline{x}_{diff}$ of μ_{diff}
- Conditions: not on the original observations, but rather the differences: 10% rule, sample size n, and not too skewed *differences*.

Paired Differences

We have

- population parameter is μ_{diff}
- point estimate $\overline{x}_{diff}$ of μ_{diff}
- Conditions: not on the original observations, but rather the differences: 10% rule, sample size n, and not too skewed *differences*.
- If met, $\overline{x}_{diff}$ has a normal sampling distribution with mean μ_{diff} and $SE_{diff} = \frac{s_{diff}}{\sqrt{n_{diff}}}$.

Next Time

- One sample t-test