

Lecture 17: Paired Data and Difference of Two Means

Chapter 5.2, 5.1

Goals for Today

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

6 Types of Questions

Here are the 6 broad types of questions about **population parameters** we'll be answering with statistical methods: confidence intervals and hypothesis tests

1. What is the mean value μ ?
2. Are the means μ_1 and μ_2 of two groups different?
3. What is the mean paired difference μ_{diff} ?
4. What is the proportion p of “successes”?
5. Are the proportions of “successes” p_1 and p_2 of two groups different?
6. Are the means μ_1, \dots, μ_k of k groups different?

Today we look at 3 and 2.

General Outline

We now generalize what we did in Chapter 4:

1. Define the population parameter and determine its point estimate
2. Show that the sampling distribution of the point estimate is Normal
 - ▶ Verify CLT & any additional conditions
 - ▶ Find the SE

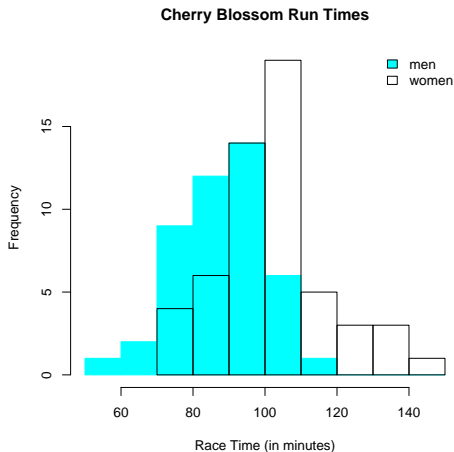
Then we either:

- ▶ Build a confidence interval: point estimate $\pm z^* SE$
- ▶ Conduct a hypothesis test with test statistic: z-score of the point estimate

$$z = \frac{\text{point estimate} - \text{null value}}{SE}$$

Chapter 5.2: Are Two Means μ_1 & μ_2 Different?

We randomly sample 45 men (of 7192) and 55 women (of 9732) runners in the 2012 Cherry Blossom Run. Did men run faster than women?



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

Difference in Means

We want the difference of two population means:

- ▶ μ_w : mean time for women
- ▶ μ_m : mean time for men

Thus:

- ▶ Population parameter: $\mu_w - \mu_m$.
i.e. if men run faster, this is positive
- ▶ Point estimate: $\bar{x}_w - \bar{x}_m = 102.13 - 87.65 = 14.48$
i.e. difference of sample means

Normality of Sampling Distribution

If two sample means \bar{x}_1 and \bar{x}_2

- ▶ each satisfy the 3 CLT conditions
- ▶ Additionally: the two samples are independent from each other

Then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will be approximately normal with

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

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Normality of Sampling Distribution

We verify the conditions:

1. Each sample consists of $\leq 10\%$ of their respective populations.
2. Both histograms don't look too skewed.
3. Each sample has at least 30 observations (rule of thumb).
4. **Additionally:** the samples are independent (not paired or linked in any way).

Thus 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

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

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

For $\alpha = 0.001$ (i.e. we want reject with high confidence) we test

- ▶ $H_0 : \mu_w - \mu_m = 0$
- ▶ $H_A : \mu_w - \mu_m > 0$

Test statistic: z-score of $\bar{x}_w - \bar{x}_m$ under H_0 :

$$\begin{aligned}\frac{\text{point estimate} - \text{null value}}{SE} &= \frac{(\bar{x}_w - \bar{x}_m) - \text{null value}}{SE_{\bar{x}_1 - \bar{x}_2}} \\ &= \frac{14.48 - 0}{2.77} = 5.23\end{aligned}$$

The p-value is 0, hence we reject H_0 and declare that men ran significantly faster than women.

Practical vs Statistical Significance

When rejecting H_0 , we call this a statistically significant result. But statistically significant results aren't always practically significant.

Say for very large n_M & n_F we observe $\bar{x}_M = 87.65$ and $\bar{x}_F = 87.651$ and reject H_0 .

The point estimate of the difference $\bar{x}_M - \bar{x}_F = 0.001$. Near negligible!

However, the 95% CI might be:

$$[0.0005, 0.0015]$$

Practical vs Statistical Significance

Moral of the story

- ▶ Hypothesis tests with “rejections of H_0 ” focus almost entirely on statistical significance.
- ▶ Confidence intervals allow you to also focus on practical significance.

Chapter 5.1: Paired Data

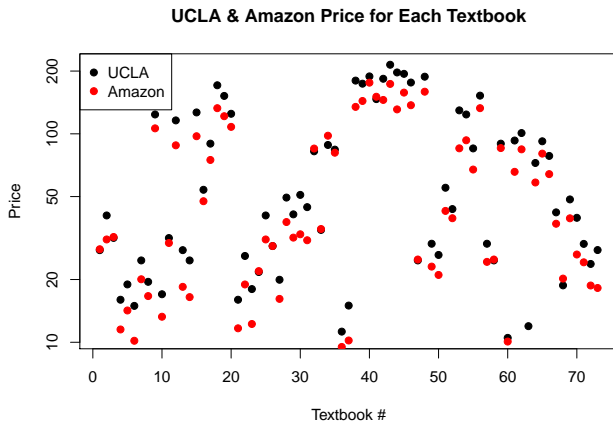
Two sets of observations are **paired** if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Examples:

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

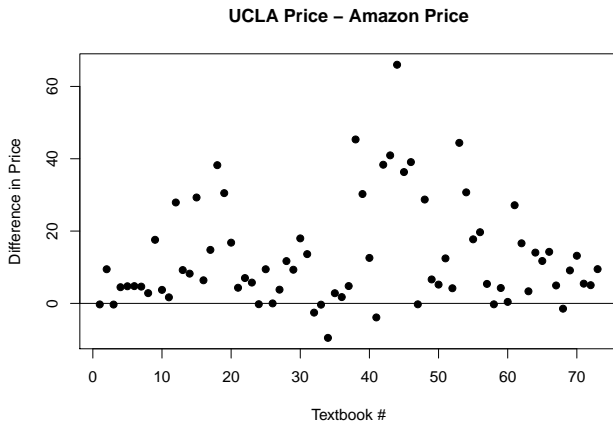
Paired Differences

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

- ▶ population parameter is μ_{diff} with point estimate \bar{x}_{diff}
- ▶ Check the conditions not on the original observations, but rather the **differences**.
- ▶ If met, \bar{x}_{diff} has a normal sampling distribution
 - ▶ mean μ_{diff}
 - ▶ $SE_{diff} = \frac{\sigma_{diff}}{\sqrt{n_{diff}}} \approx \frac{s_{diff}}{\sqrt{n_{diff}}}$

Next Time

- ▶ t-test