Lecture 17: Paired Data and Difference of Two Means

Chapter 5.1-5.2

Goals for Today

- Note on Practical vs Statistical Significance
- Difference of Means

Terminology Recap (Page 192)

- Summary statistics are a single number summarizing a large amount of data.
  Ex: sample mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Point estimates use observations $x_1, \ldots, x_n$ to guess at the value of an unknown parameter.
  Ex: the sample mean $\overline{x}$ estimates the true population mean μ.
- A test statistic is a summary statistic used in hypothesis testing or for identifying the p-value.
  Ex: in the Reed sleep example, we used $\overline{x}$. Since $\overline{x}$ is approximately normal by the CLT, we use the z-score of $\overline{x}$ as the test statistic.

Hypothesis Testing Procedure

1. Construct your hypothesis testing framework:
   - Define H_0, H_A and if applicable a null value.
   - Set your significance level α
2. Verify that the conditions hold
3. Compute your test statistic
4. Compute the p-value
   - Identify the appropriate distribution to compare the test statistic to
   - Depending on H_A, determine what constitutes being more extreme and compute the p-value using the appropriate probability table.
5. If the p-value is $< \alpha$, reject H_0. Otherwise do not.

In General: Confidence Intervals

All confidence intervals have form:

$$[\text{point estimate} - z^* \times SE, \ \text{point estimate} + z^* \times SE]$$
$$\text{point estimate} \ \pm \ z^* \times SE$$
$$\text{point estimate} \ \pm \ \text{margin of error}$$

where z^* determines the confidence level.

The point estimate and SE will change depending on the context.

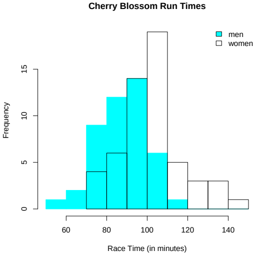
The 8 Types of Questions

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 what we were expecting?
8. Are two categorical variables independent?

Are the means of two groups μ_1 and μ_2 equal or not?

Example from Chapter 5.2: Did men (n=45) run faster than women (n=55)?



Difference in Means

We are interested in the difference of two population means $\mu_w - \mu_m$ where

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

The data:

	men	women
$\bar{x}$	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 $\mu_w - \mu_m$:

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

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

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

- 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:

- 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

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

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

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

Next Time

- Hypothesis test for differences in means
- Paired differences
- One sample t-test