Models to Compare Two Population Means

In this class, you will work with your group to explore how to fit and interpret models to compare two population means. While you work through this activity, make sure that all group members are engage and contribute ideas and follow the code.

The R Manual has useful R code for today's activities.

Part 1: Understanding the study design

Before you begin analyzing the games data discussed in Chapter 2, you need to review the study design and consider how the design impacts the conclusions that you can draw.

Task 1: Review the study design on page 31 and complete activities 1-4.

To complete the "individual values plot" for activity 4, check out the R Manual.

Part 2: Statistical models for the two-sample t-test

As you saw before class, the underlying model for a two-sample t-test is given by

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \text{for } i = 1, 2, \quad j = 1, \ldots, n_i, \quad \text{where } \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

Let group 1 be the color distractor group and group 2 be the standard game group.

Task 2. Calculate estimates for the two population means. To do this, think back to the EDA activity from Monday.

Part 3: Checking assumptions

The necessary assumptions/conditions for the two-sample t-test are:

1. the error terms are i.i.d.
2. the error terms follow a normal distribution
3. the error terms have mean 0
4. the population variance is the same for each group

Before conducting inference, we should check these assumptions/conditions. If they are violated, then our conclusions could be suspect.

Data plot

As you saw earlier, you can assess some aspects of the error term distribution using a plot of the data (i.e., an individual value plot).

- Is the spread the same between groups? Is it drastically different?
- Are there extreme outliers (unusual observations)?

Residual plots

Most of the assumptions can be checked by estimating the error terms (i.e., calculating the residuals) and creating sensible plots.

- Histogram: could the residuals follow a normal distribution?
- Histogram by group: how does the spread compare between groups? Are both groups centered around 0?
- Boxplots by group: is the spread (IQR) about the same between groups?
- Residuals vs. time: is there

Comparing sample variances

In addition to graphically checking whether the spread of the residuals is similar between the groups, you can also conduct an informal comparison of the sample variances.

You book recommends that $\max(s_1^2, s_2^2)/\min(s_1^2, s_2^2) < 2$, but this is not a strict rule.

Task 3: Complete activities 7-10. The R Manual has useful code.

Part 4: Fitting a regression model to compare means

As you learned in preparation for this class, a simple linear regression model can be used in place of the two-sample procedure you just considered. The regression model will be of the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{for } i = 1, 2, \ldots, n \quad \text{where } \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Recall further that x_i will need to be an **indicator variable**.

Task 4: Complete activities 11-13. The R Manual has useful code.

Part 5: Check the model

Before you can trust your inferential results, you should verify that the necessary assumptions/conditions are met. That is you need to check that

1. the error terms are i.i.d.
2. the error terms follow a normal distribution
3. the error terms have mean 0
4. the come from a single population with variance σ^2 (i.e., the variance doesn't depend on the value of x_i)

Luckily, you can use the same graphical tools as you did for the two-sample model after you extract the residuals.

Task 5: Complete activity 14.

Part 6: Communicate the results

If your model seems adequate, then you should clearly communicate your findings. This can be in the form of a graphic, an equation, or a couple sentences.

Task 6: Complete activity 15.