Tidy Data Analysis using rstatix and ggpubr

Dr. Austin Brown

Kennesaw State University

Introduction

- So far this semester, we have discussed a wide variety of R programming topics ranging from data manipulation and visualization to iteration.
- All of these topics are important and useful when programming using R.
- However, all of these topics, for us as statisticians/data scientists, are meant to help us in our primary aim: data analysis!

Introduction

- Many of you with familiarity in R programming already likely already know how to conduct some analyses using the software.

- However, what I want to show you today is how we can use some of the helper tools we've learned about so far to aid in conducting common types of analyses.

- Specifically, we will be learning how we can use the `ggpubr` and `rstatix` packages (as well as a couple of others) to perform traditional analyses in a more modern way.

Group Comparisons

- One incredibly common research design is the comparison of groups. For example, let's say that using the `Lahman::Batting` dataframe, that we want to compare batting averages for players who played in the 1985 and those who played in the 1995.
- Clearly, since we assume that batting averages across seasons and certainly across decades are independent of each other (typically done in sports data), an independent t-test seems appropriate.
- Let's look at the traditional way of solving this problem and then the method using a more modern approach.

Group Comparisons

- First, it's always best to state our hypotheses:

$$H_0 : \mu_{85} = \mu_{95}$$

$$H_1 : \mu_{85} \neq \mu_{95}$$

- And as we know, the independent t-test has a couple of assumptions besides mutual independence: (1) normality and (2) equality of variances.

Group Comparison

- Before we go through and make a decision with respect the H_0, we need to make sure our assumptions are reasonably met.

- Let's check normality first. What the assumption of normality in a t-test really means is that you're testing:

$$H_0 : F_1(\mu_1, \sigma_0) = F_2(\mu_2, \sigma_0)$$

$$H_1 : F_1(\mu_1, \sigma_0) \neq F_2(\mu_2, \sigma_0)$$

- This means that histograms/density plots can be a useful tool for assessing the normality assumption (of course, these should be paired with normality tests, too).

Group Comparison

- We can clearly extend the comparison of two independent groups to the comparison of several indepent groups, which is typically done using a one-way ANOVA model.
- Suppose now we want to add a third year to the batting average comparison, say 2005.
- Let's see how we can use `rstatix` and `ggpubr` to make our lives easier.

Group Comparison

- Of course, in most research, we're likely interested in more than just a single explanatory variable.
- Let's say, for example, that we want to see the effect a person's gender and education level have on job statisfaction. Since gender and education level are both categorical variables, a two-way ANOVA model seems like an appropriate choice.
- Let's take a look at how we can approach this problem with R.