15.10.1 Whose Analysis Is It Anyway?

Jeremy has put in a huge amount of work, covered a huge amount of topics, and has really stepped into his leadership position.

Of course, he was hired for the leadership position, even without a strong R background, because of his experience at AutosRUs. That's because there is something even more important than knowing how the math works or knowing how to code in R when it comes to driving value for your company: *knowing the right questions to ask.*

When using data to make informed decisions in a professional environment, implementing a statistics function is not the biggest challenge. Rather, it's determining what questions to ask.

In data science, researchers use **retrospective analysis** to analyze and interpret a previously generated dataset where the outcome is already known. Retrospective analyses are helpful because there are no upfront costs to generate data and statistical results can be compared to the known outcomes. Depending on the dataset and input variables, there is a (potentially) limitless number of statistical questions that can be asked from the data:

- Are two groups statistically different? Use a t-test with one dichotomous independent variable and one continuous dependent variable.
- Can one continuous dependent variable be predicted using another independent variable? What about multiple independent variables and one dependent variable? Use regression analysis.

- Are there multiple categorical variables tightly linked in a dataset? Are the distributions of the different categorical variables equal? We can test with chi-squared.

In contrast, researchers can **design their own study** to answer their own specific questions. In this case, the data types and size of the dataset will be directly reflective of how complicated their hypotheses are, and what statistical analyses are required to answer the question.

For example, if we want to verify that a car battery ages at an appropriate rate, we would need to test our question with a regression model. If we were to use a multiple linear regression model, we would need to collect numerical variables, such as number of uses, time, battery capacity, tire tread, and engine horsepower. Once we select the variables to collect, we would estimate sample size based on how low of a significance level is necessary and how sensitive the measurements are.

Regardless, if we collect and measure the data ourselves, or if the data has been curated from a previous dataset, statistical tests can help us provide quantitative interpretation to the results.