Print

Identifying Associations in ANOVA with Box Plots

An association exists between two variables when the expected value of one variable changes at different levels of the other variable. So if you were to find that the average sale price of homes with an attached garage is quite different from that of homes with a detached garage or no garage, this would imply a possible association between these variables.

Let's begin by exploring associations between variables in the context of one-way ANOVA, when the response is continuous and the predictor is categorical. A simple way to look for possible associations in your data is to create box plots. In a box plot, the response variable is typically on the Y axis, and the predictor variable is on the X axis. The box is drawn to represent the interquartile range, or the middle 50% of the data. The upper bound of the box is the 75th percentile and the lower bound is the 25th percentile. The diamond within each box is the mean of Y, and the horizontal line in each box represents the median. Extending from the top and bottom of each box are whiskers that represent the spread of the data.

You can include a line on your box plot to connect the means of Y at each value of X. A horizontal line indicates that there's no association between X and Y. In other words, knowing the value of X doesn't tell you anything about the value of Y. So for each value of X, your best guess as to the value of Y would simply be the mean of Y. If the line is not horizontal, then there might be an association between X and Y. In other words, the expected value of Y differs at different levels of X. In this example, participants over age 30 have a higher average response than those under 30. Thus, knowing age provides additional information when you're predicting the response.

Close