

CHAPTER OVERVIEW

16: Correlation, Similarity, and Distance

Introduction

We continue with our discussion and introduction of inferential statistics. Recall that as we analyze a data set, we generally want to begin by describing it (central tendency, measures of variability), and we also want to plot the data. To begin our introduction to correlation and regression, first we describe how to produce graphs to help show **linear association** or in some cases, cause and effect — the latter perhaps the primary reason for using regression.

Graphical representation

The previous statistical procedures we have examined have used one or more **categorical** or **qualitative** variables ([Chapter 3](#)). For example,

1. Chi-Square Analyses: variables are all categorical, including the response variable ([Chapter 9](#)).
2. T-tests: one categorical (Factor) variable and one (Dependent, Outcome, Response) variable that was continuous or interval scale ([Chapter 8.5, 10](#)).
3. ANOVA Analyses: one or more variables are categorical (Factors, the independent variables) and one (Dependent, Outcome, Response) variable that was continuous or interval scale ([Chapter 12, 14](#)).

The convention in graphing ANOVA (or Chi-Square) is to use the Factor or Independent variables as the X-axis and to have the dependent variable (Response) as the Y-axis. We called these **bar charts** ([Chapter 4.1](#)).

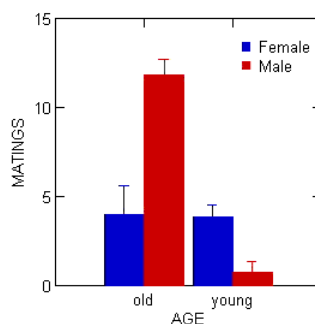


Figure 16.1: Bar chart with error bars.

Box plots ([Chapter 4.3](#)) are also useful, and perhaps the preferred choice to display this type of comparison (one involving groups) (Fig. 16.2).

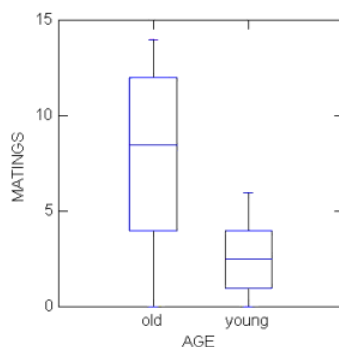


Figure 16.2: Box plots.

In correlation (and regression) analyses we will have two or more continuous or interval scale variables. To show relationships among continuous variables, a **scatter plot**, also called an X-Y plot, works well ([Chapter 4.5](#)).

In correlation, no causation is implied, so either variable can be placed on the X-axis. The convention of graphing in regression is to place the independent variable as the **X-axis** and the dependent variable as the **Y-axis** (Fig. 16.3). Another consideration: if one

variable is considered **fixed** and the other **random**, then the fixed variable would be assigned to the horizontal axis.

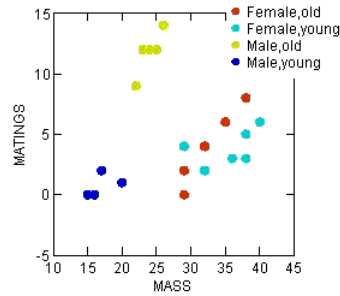


Figure 16.3: Scatterplot with groups.

To produce a scatterplot (also called an X-Y plot) in **Rcmdr**, select **Graph** → **Plot** → and select the Y and X variables. Use a combination of Options, Frame, and Edit Attributes selections to modify the default graph.

[16.1: Product-moment correlation](#)

[16.2: Causation and partial correlation](#)

[16.3: Data aggregation and correlation](#)

[16.4: Spearman and other correlations](#)

[16.5: Instrument reliability and validity](#)

[16.6: Similarity and distance](#)

[16.7: References and suggested readings](#)

This page titled [16: Correlation, Similarity, and Distance](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by [Michael R Dohm](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.