CONCEPTS OF STATISTICS II
ASSIGNMENT #2
SETH HOWELLS
07/19/2020

Examining Data

The first step in any analysis begins with examining the data for errors, missing data, and outliers. ***Missing data*** refers to information not available for a case, which in return leads to skewed outputs and therefore misinterpretations regarding the data suspects under investigation. An outlier on the other hand has a different impact on results. While missing data is the absence of a value, an ***outlier*** is the presence of an extreme value – both extremely high or extremely low. The ultimate goal is to provide reasonably clean data so that the results can be validly and reliably interpreted.

Graphical Displays

In most analysis, the quickest way to gain the perspective of a variable is to create a visual graph. A ***histogram*** is a type of graphical display that provides a single variable's distribution of values. The height of the bars relates to the frequency, or count, while the x-axis values are the classes, or bins, within the range of the variable. The result is a visual display of the variable's distribution. Applying a normal curve on the histogram will quickly show how the variable's distribution relates to a normal mode. For example, *FIG-1* has a relatively normal distribution as indicated by the normal curve overlay on the histogram, while *FIG-2* displays non-normal distributions.

Although histograms are limited to a single variable, profiling bivariate relationships can be visualized through a ***scatter plot*** where the two quantitative variables are assigned to the corresponding X and Y axis. As a result, each record has an (X, Y) coordinate graphed as a single point on the scatter plot. Collectively, four distinct patterns can be identified as displayed in *FIG-3*.

FIG-3: Characteristics

- **Top left:** indicates that as X increases, Y also increases.
- **Top right**: indicates that as X increases, Y decreases.
- **Bottom left**: indicates that X is void of any relationship with Y.
- **Bottom right**: indicates that X will not bring the same change in the Y variable.

The importance of bivariate relationships is to gain inference whether or not the variables under investigation possess linear qualities. Understanding these types relationships and how to relate to linearity plays a critical role in forecasting or overall prediction.

Missing Data Types and Impacts

Establishing a valid and reliable analysis requires understanding of missing data types, impacts of missing data, and techniques to mitigate and impact of missing data. Missing data can be classified into three main categories:

- **Missing completely at random (MCAR):** probability of missing data is equal for all cases – not systematic.
  - *Example*: a manufacturer's weigh station ran out of battery and therefore did not capture the information simply due to bad luck.
- **Missing at Random (MAR):** probability of missing data is the same amongst groups of observed data – systematic.
  - *Example*: a manufacturer's weight station does not capture the information when weighed on a soft surface but captures the information on hard surfaces.

- **Missing not at Random (MNAR):** probability of missing data varies for unknown reasons.
    - *Example*: a manufacturer's weigh station depreciates over time and as a result, produces more missing data as time passes.

Missing data can create various problems in any analysis. For example, missing data impacts the representativeness in the sample, reduce the statistical power, or can cause biased estimates in the results. While these various problems can occur with missing data, it is important to note that the next step is to analyze the degree to which missing data impacts the output. A few missing values in a large sample size will have less of an impact on the result than many missing values on a smaller sample size. The ultimate goal is to identify any patterns or trends of the missing values. Using tools like the Missing Value Patterns Chart, as shown in *FIG-4,* can help to quickly visualize such patterns.

Handling Missing Data Strategies

The three main strategies for handling the presence of missing data is listed below with advantages, disadvantages, and rules of thumb for appropriate usage.

1. **Complete-Case Analysis (Listwise deletion):** excludes any record that contains missing data. Best to use when missing data count is low.
    a. *Advantages*: quick and easy completion time
    b. *Disadvantages*: reduces sample size and statistical power
    c. *Rule of Thumb*: missing data is under 10%
2. **Pairwise deletion** only excludes data missing for the specific analysis.
    a. *Advantages*: more accurate than listwise deletion, quick and easy to implement
    b. *Disadvantages*: less reduction in sample size
    c. *Rule of Thumb*: missing data is between 10-20%
3. **Imputation**: estimating the missing value based on the aggregate of similar cases.
    a. *Advantages*: does not affect sample size and provides most accurate analysis
    b. *Disadvantages*: time-consuming
    c. *Rule of Thumb*: missing data is over 20%

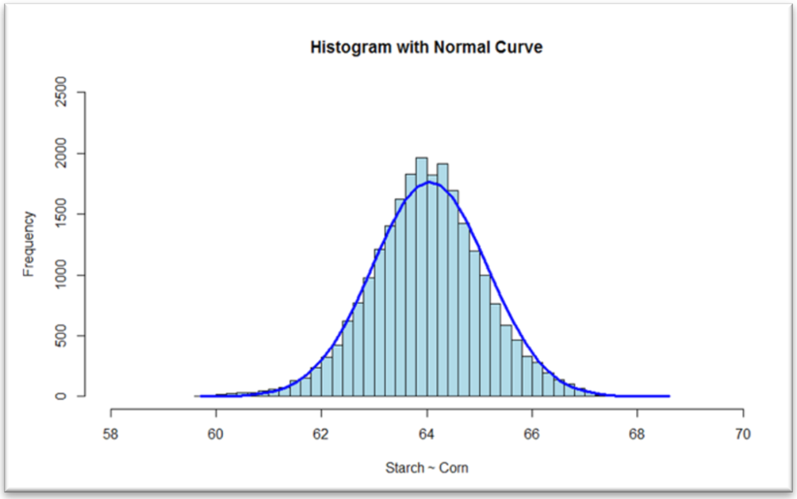

FIG 1: Normal Distribution

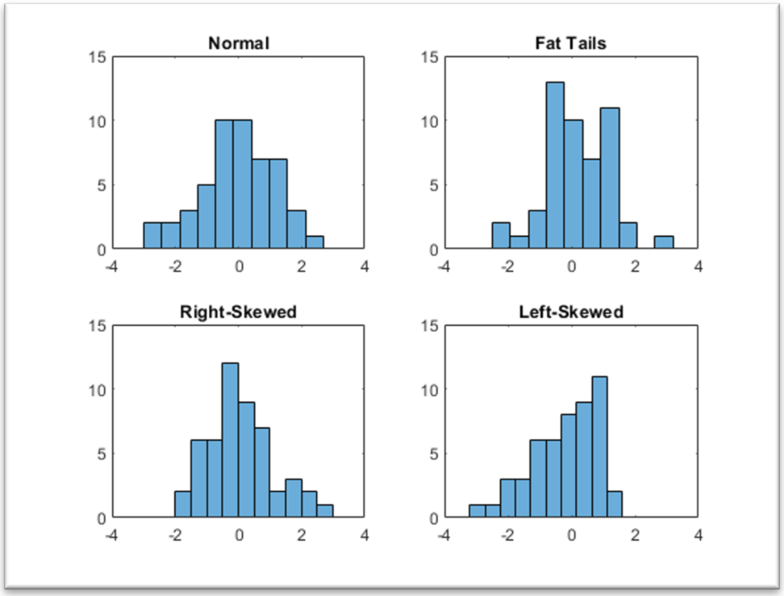

FIG 2: Types of distribution

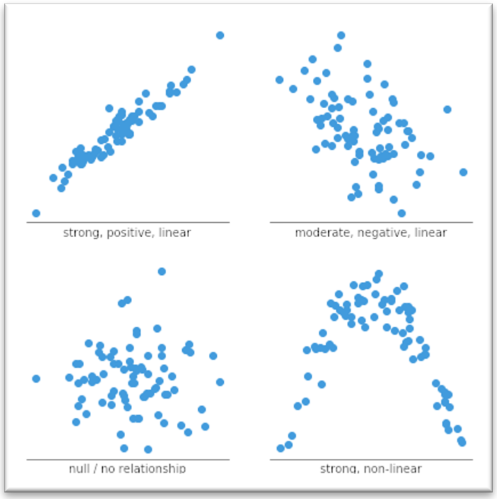

FIG-3: Types of scatter plots – linearity

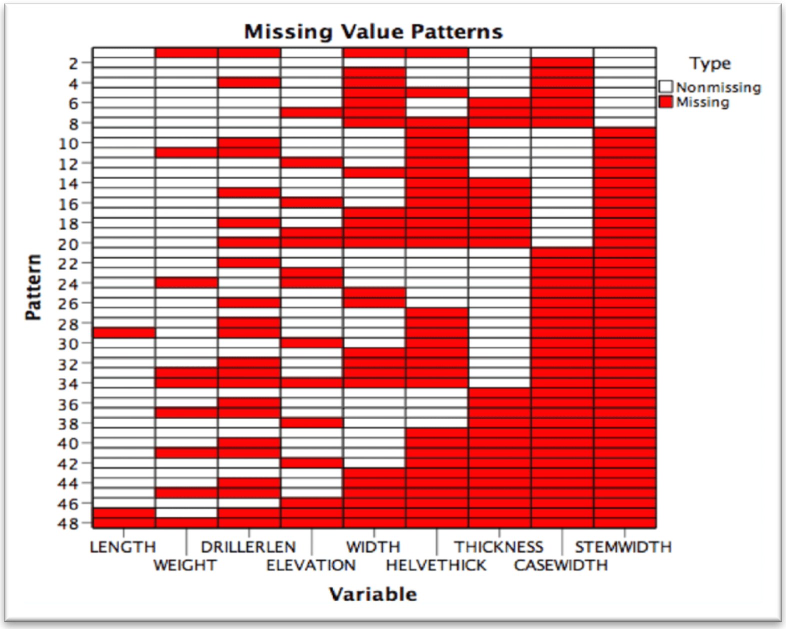

FIG-4: Missing Values Pattern Chart