1 Exploring One-Variable Data

1.1 Representing Categorical and Quantitative Variables with Graphs

Data contains information about a group of individuals. The information is organized using variables.

Individuals are objects described by a set of data. Individuals may be people but may be animals or inanimate objects.

Variables are characteristics of individuals. A variable may take on different values of different variables. Variables can be split into two types: categorical or quantitative.

Categorical variables place individuals into specific groups.

Quantitative variables takes on numerical values for which it makes sense to do arithmetic operations like adding and averaging. Quantitative variables fall into two categories: discrete and continuous.

Be careful - just because it is a number doesn't make it quantitative.

Discrete variables are numerical values where counting makes sense; in other words, decimals would not be an appropriate way to record the data.

Continuous variables are numerical values where decimals are appropriate; it usually involves some form of measuring.

The difference between discrete and continuous isn't always clear. An example of this would be age.

One of the easiest ways to display categorical data is with a table.

Count is the amount of that category in a table and relative count is

$$\frac{\text{count}}{\text{total}}$$

If you wanted to display two categorical variables at a time, we could make a two way table.

To better visualize the data, there are graphs that we can make from the data. We want to visualize the graphs to get a better idea of the distribution.

Distribution of a variable tells us what values the variable takes and how often it takes these values.

Bar Graphs have the following characteristics:

- Label each axis clearly
- The x-axis will contain the categorical variable and the y-axis will display teh counts
- Each category has its own bar and the bars cannot touch
- Order is not important when creating the x-axis

To make a histogram, we need to put the data into even intervals that capture our data. We will do this first by hand by counting how many data scores are in each bin.

To find the interval width, we can use the formula

$$\frac{\text{max-min}}{\#\text{ of wanted intervals}}$$

To make the histogram:

- Draw rectanges for each interval with height representing the count
- Bars must touch
- Label the x-axis with the lower bound values of each interval

1.2 Representing Quantitative Variables with Graphs

Stemplots (or stem and leaf plots) are an alternate way to illustrate data using a semi-graph. It is similar to a histogram, but the data isn't lost. If the data has two digits, the stem is the first digit and the leaf is the second. If the data has 3 digits, the stem is the first two digits. You must always add a key to the graph.

Back-to-Back Stemplots are created when you can separate the data into two categories. The stems are the same, but the data can be split into different categories. You still must have a key for both sides.

Split Stemplots are the last type of stem and leaf plots. You can split the stem; in a similar way to creating more bins on a histogram if the bin width resulted in a skyscraper.

Dot plots are a very simple type of graph that involves plotting data values with dots above the values on a number line.

To construct a dot plot:

- Label your axis and title your graph. Draw a horizontal line and label it with the variable.
- Scale the axis based on the values of the variable.
- Mark a dot above the number on the horizontal axis corresponding to each data value.

Cumulative relative frequency graphs display percentiles. A percentile will tell you what percent of data falls above a value.

In order to create one of these graphs, you must make a table of the cumulative relative frequencies in order to graph it. This can be done by first finding the relative frequencies and then you add them together to get the cumulative relative frequency. This graph is also called an ogive.

1.3 Describing Distributions of Quantitative Variables

We can describe the distribution of a quantitative variable by its shape, outliers, center, and spread.

After graphing the distribution, the first thing we identify is the shape. A unimodal shape has one peak, a bimodal shape has two peaks, a uniform shape has no peak. If a graph is skewed it will be skewed in the direction in which the tail is located.

There are three measures of center: mean, median, and mode. Mean is defined as

$$\bar{x} = \frac{\sum x_i}{n}$$

, where $\bar{x}$ is the mean of sample, x_i is each individual observation, and n is the number of observations. Mean is called non-resistant because the mean is strongly influenced by outliers.

Median is a different measure of center that is resistant. The median is found by ordering the data and finding the middle value in the list. The location of the median is

$$\frac{n+1}{2}$$

Mode is the most occurring value in a data set.

To summarize, a unimodal symmetric graph will have the median, mean, and mode similar to each other. A unimodal, left-skewed graph will have mean $<$ median $<$ mode, and a unimodal, right-skewed graph will have mode $<$ median $<$ mean.

There are three measures of spread: range, standard deviation, and IQR.

Range is the difference between the maximum number and the minimum number.

The standard deviation is the average deviation of an observation from the mean of the data set. It is calculated by

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n-1)}}$$

, where s_x is the standard deviation, and s_x^2 is the variance of sample.

The variance is the average squared deviation. The quantitative variable typically varied by the mean by "standard deviation units".

Standard deviation has a few properties:

- The standard deviation is always positive.
- The standard deviation is always 0 when all observations are equal.
- The standard deviation has the same units of measure as the original variable measured.
- The standard deviation is non-resistant.
- The greater the standard deviation, the greater the distribution.

IQR is inter-quartile range and it uses percentiles to describe the spread of distribution. The 0th percentile is the minimum, the 25th percentile is Quartile 1, the 50th percentile is the median, the 75th percentile is Quartile 3, and the 100th percentile is the maximum.

An outlier is an individual piece of data that falls outside the overall pattern of the distribution. We can determine if a point is an outlier:

- First, find the five-number summary (the percentiles mentioned above)
- Find the IQR by subtracting the value of Quartile 1 from the value of Quartile 3
- Compute Quartile 1 - (1.5 * IQR). Any data above this number is an outlier.
- Compute Quartile 3 - (1.5 * IQR). Any data above that number is an outlier.

Using this data, we can make box plots or modified box plots depending on if the data set as an outlier determined.

1.4 Comparing Distributions of Quantitative Variables

In the world of statistics, it is not enough to just report the graph or just report the summary statistics.

Given a set of data, first you must create a graph for the data. If the data is categorical, use a bar graph. If the data is quantitative, if it's discrete, use a dot plot, stem plot, or boxplot. If it is continuous, use a histogram or box plot.

You want to summarize your findings after this. First, describe the shape. Next, describe if there are any outliers. Then, use the mean or median for the center of data. Lastly, describe the spread using range or IQR if you described the center with the median, or standard deviation if you described the center with mean.

1.5 Z-Scores and the Empirical Rule

Normal distributions are appropriate for many distributions whoes shapes are unimodal and approximately symmetric.

In the normal distribution, Mean $= \mu$, Standard deviation $= \sigma$, and this is written as

$$N(\mu, \sigma)$$

A normal distribution curve has these properties:

- Symmetric around the mean
- mean = median = mode
- 50% of observations are greater than the mean and 50% are less than the mean
- The standard deviation tells us how measurements for a group of observations are spread out from the mean
- In a normal distribution, approximately all the data lies 3 standard deviations above and below the mean

In a normal distribution, 68% of the data is likely to be found within 1 standard deivation from the mean, 95% found 2 standard deviations away, and 99.7% 3 standard deviations away.

To calculate how many standard deviations away from the mean an observation is, we use a z-score.

$$z = \frac{x - \mu}{\sigma}$$

Observations larger than the mean have positive z-scores, and observations smaller than the mean have negative z-scores.

1.6 The Standard Normal Curve

Using a calculator, we can determine what percentage of data falls above, below, or between specific z-scores.

Using the calculator commands: 2ND $\rightarrow$ Vars $\rightarrow$ 2:normalcdf(), we can find the percentage of data in between two values.

We can also find the z-score that corresponds to a percentile.

Using the calculator commands: 2ND $\rightarrow$ Vars $\rightarrow$ 3:invNorm(), we can find this.