2020 AP Statistics Review

Bryan Li

February 15, 2020

Contents

Chapter 1

Working with Data

1.1 Data, what is it?

In this section, we'll be working with data. Remember, data are **values** along with their **context**.

1.1.1 Context

When describing the context of data, it is important to consider the W's

- Who? - Who was the data collected from?
- What? - The variables recorded from each experimental unit, and the unit of measurement. What type is the variable? Quantitative or categorical?
- When? - When was the data collected?
- Where? - Where was the data collected?
- Why? - Why was the data collected?
- How? - How was the data collected?

Data without context is useless. For example, consider:

1.2, 3.7, 4.4, 2.2

Are these data? No. These are just numbers without any context. Numbers without any context mean nothing.

1.1.2 Types of Data

There are two types of data. These include *categorical* and *quantitative* data. These two types of data must be expressed differently.

Definition 1.1.1. Categorical Variable: A categorical variable is a variable that can take on one of a limited amount of possible values. Individuals are assigned to a group by some qualitative property. Examples: race, sex, eye color.

Definition 1.1.2. Quantitative Variable: A quantitative variable is a variable that is measured on a numeric scale. Examples: age, height, weight.

1.2 Displaying Data

When displaying data, it is important to follow the *area principle*.

Definition 1.2.1. Area Principle: The area principle states that the area of a graph should equal the magnitude of the data it represents.

Be careful, as many 3D graph types can violate the area principle.

1.2.1 Displaying Categorical Data

Categorical data are displayed primarily with 2 types of visual charts: Bar graphs and pie charts. It can also be expressed with a frequency table, like the one in Figure 1.1.

Figure 1.1: An example frequency table

Sex of Respondents

Sex	Male	Female	Total
Frequency	13	10	23

Bar Graphs

There are many types of bar graphs. They include normal bar graphs, frequency bar graphs, and stacked bar graphs. Below are examples of

different types of bar graphs. **It is important that the bars do not touch each other. This signifies the bars are ordered. Categorical variables cannot be ordered in a meaningful way, so this is wrong.**

- Normal Bar Graph: Displays the total number of cases in each category.
- Relative Frequency Bar Graph: Displays the relative frequency (percentage or decimal) of each category. All relative frequencies add up to 100%, or 1.
- Stacked Bar Graph: Shows comparisons between categories of data.

Figure 1.2: An example of a bar graph

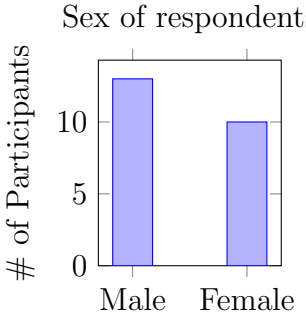


Figure 1.3: An example of a relative frequency bar graph

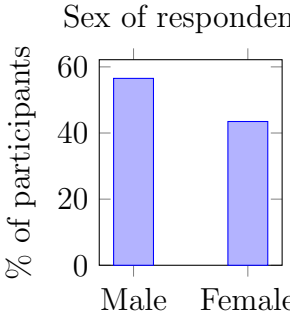
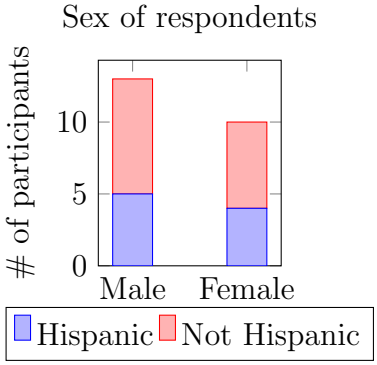


Figure 1.4: An example of a stacked bar graph

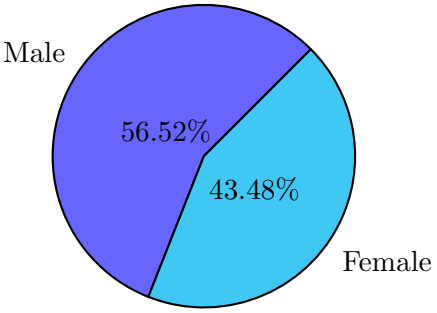


Pie Charts

While bar graphs are great for looking at numerical values, pie charts can be useful when determining the percent composition of the categories. Below is an example of a pie chart.

As you can see, the pie chart does not violate the area principle. Each "slice" of a pie chart grows in size. Polar area charts, which are similar to pie charts, do not obey the area principle. You should also see that the sum of the relative frequencies equals 100%, or 1.

Figure 1.5: An example pie chart



Contingency Tables

Similar to the stacked bar graph, these tables are used to display information about two variables.

Figure 1.6: An example contingency table

Sex of Respondents

	Male	Female	Total
Hispanic	5	4	9
Not Hispanic	8	6	14
Total	13	10	23

The *marginal frequency* of a variable is the value in the margins: the rightmost column and the bottommost row. It can be described as $P(A)$ where A is the value of the variable in the row or column the frequency is.

The *conditional frequency* of two variables indicates the frequency of which a variable occurs given a known value of the other. For example, $P(Male|Hispanic)$ is $\frac{5}{9}$, which is the probability that a respondent is male given that they are Hispanic.

1.2.2 Displaying Quantitative Data

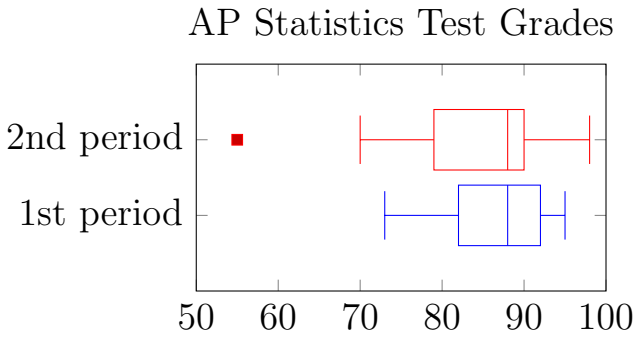
Displaying quantitative data is a bit different from displaying categorical data. Display methods are typically box-and-whisker plots, histograms, and stem plots.

Box and Whisker Plots

Box and Whisker plots are useful for determining these statistics:

- M/Q_2 (Median)
- Q_1 (First quartile)
- Q_3 (Third quartile)
- $minX$ (Minimum)
- $maxX$ (Maximum)
- IQR (Interquartile Range)
- R (Range)

Figure 1.7: An example box and whisker plot



And for determining any outliers. Outliers can be determined by individual data points that are not included in the whisker or box.

Definition 1.2.2. **Outlier:** As a general rule, outliers are data points that do not fall into the range $[Q_1 - 1.5 * IQR,\ Q_3 + 1.5 * IQR]$.

Histograms

Histograms are primarily useful in order to view a distribution's shape, and any gaps it might have.

A histogram's primary difference from a bar graph is that it is used to display quantitative data. This means that it has "bins", which are the bars, which touch each other.