Cheat Sheet: Describing Data Graphically

Essential Concepts

- Visual representations of categorical variables include:
    - Frequency tables show the frequency, or count, for each category of the variable. A frequency is the number of times a value of the data occurs.
    - Bar graphs are visual displays of data in which the frequency of each category listed across the horizontal axis is indicated by the height of its corresponding rectangular bar (or the length if the graph is displayed horizontally)
    - Pie charts display data in a round graph, split into “pie pieces,” or wedges, each representing the proportion of the quantity or frequency it represents and having a relative size to match. In other words, pie charts display the relative frequencies of your data set
- A relative frequency is the ratio (i.e., fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find relative frequency we divide the frequency of the specific category by the total count of all data.
- A contingency table is a table that displays the results of two categorical variables simultaneously. It is also called a two-way table
- Pie charts and bar graphs/charts are good visual representations of a categorical variable from a single population or group. But what can we do if we want to compare a categorical variable across multiple groups?
    - Stacked bar charts display values for different categories but rather than showing a different bar for each category, they display sub-categories as segments within each bar.
    - Side-by-side bar graphs present data for two categorical variables from more than one group by creating two bars on the chart for each group — one bar for each variable.
- Visual representations of quantitative variables include:

    - Dot plots show each data value as a dot or point above a number line. When multiple data values are the same, the dots are stacked vertically above that value. Dot plots are especially useful for small data sets and for comparing distributions.
    - Histograms group quantitative data into intervals called bins and display the frequency of values in each bin using adjacent rectangular bars. The height of each bar represents how many data values fall into that interval.

- When presented with large data sets, the dotplot is sometimes difficult to put together. In addition, it may not be the clearest way to present the data. For large data sets, a histogram can represent the numerous data points more simply as bars instead of the immense amount of data points that would be needed in a dotplot.

- In general:

    - Histograms are good choices for displaying data sets that have a large number of observations because they group observations into equal-size “bins”.
    - Dotplots display how many individual observations there are of each value observed. Each observation in the data set appears as its own dot on the graph.

- The features used to describe the distribution of a quantitative variable are the shape, center, spread, and presence of outliers.

    - To describe the shape of a distribution, imagine sketching the outline of the data to emphasize the general trend.

        - Left-skewed: A cluster of data on the right with a tail of data tapering off to the left.
        - Symmetric with a central peak (also called bell-shaped): The left and right sides of the distribution closely mirror each other. If you drew a vertical line down the center of the distribution and folded the distribution in half, the left and right sides would closely match one another.
        - Right-skewed: A cluster of data on the left with a tail of data tapering off to the right.

    - The shape can also be described from the number of peaks:

        - Unimodal: There is one prominent peak.
        - Bimodal: There are two prominent peaks.
        - Multimodal: There are three or more prominent peaks.
        - Uniform: There are no prominent peaks. A rectangular shape, the same amount of data for each variable value.

    - The center describes the location of the middle of the distribution.

    - The spread is a measure of how much the values in a data set tend to differ or vary from one another.

        - One way we can measure the spread is by finding the range. The range is the difference between the minimum and maximum values in the data. The minimum represents the smallest value in the data and the maximum represents the largest value.
    - Outliers are observations in the data that are unusual and outside the general pattern of the rest of the observations in the distribution.

Key Equations

$$relative\ frequency\ = \frac{subgroup}{total}$$

Glossary

bar graph (bar chart)

a graph where categories are represented by bars that are separated from each other

bell-shaped

the left and right sides of the distribution closely mirror each other

bimodal

two prominent peaks in the data

center

the location of the middle of the distribution

contingency table (two-way table)

a table that displays the results of two categorical variables simultaneously

dotplot

displays how many individual observations there are of each value observed

frequency table

lists the number of observations (the frequency) of each unique value of a categorical variable

histogram

displays datasets that have a large number of observations by grouping into equal-size “bins"

multimodal

three or more prominent peaks in the data

observational units

contains information about a group of individuals or things

outliers

observations in the data that are unusual and outside the general pattern of the rest of the observations in the distribution

pie chart

a graph where categories are represented by wedges in a circle and are proportional in size to the percentage of individuals or items in each category

population

people or entities such as animals or objects

range

the difference between the minimum and maximum values in the data

relative frequency

represents the proportion of observations that are in a particular category and can be expressed as a decimal or a percentage

shape

overall patterns and the number of peaks in the data

side-by-side bar chart, stacked bar chart

an extension of a bar chart that allows us to conduct comparison between multiple datasets

skewed-left

a cluster of data on the right with a tail of data tapering off to the left

skewed-right

a cluster of data on the left with a tail of data tapering off to the right

spread (variance)

measure of how much the values in a dataset tend to differ or vary from one another

uniform

no prominent peaks in the data

unimodal

one prominent peak in the data

variables

observational units are recorded