Chapter 2: Descriptive Statistics: Tabular and Graphical Displays

Data visualization: use of graphical displays to summarize and present information about a data set.

2.1 Summarizing Data for a Categorical Variable

Frequency distribution: tabular summary of data (categorical or quantitative) showing the number (frequency) of observations in each of several nonoverlapping categories or classes.
Relative frequency distribution: tabular summary of data (categorical or quantitative) showing the relative frequency (proportion) of observations in each of several nonoverlapping categories or classes.
Percent frequency distribution: tabular summary of data (categorical or quantitative) showing the percent frequency (percentage) of observations in each of several nonoverlapping categories or classes.

Relative frequency of a category (class) = frequency of a category (class)/total in all categories (classes)
Percent frequency of a category (class) = relative frequency x 100

A **bar chart**, where each bar represents one category (class) in a distribution of categorical data, can be used to graphically display categorical data summarized in a frequency, relative frequency, or percent frequency distribution. The horizontal axis specifies the labels used for the categories (classes). The vertical axis specifies the frequency, relative frequency, or percent frequency scale. Separate bars, of fixed width, are drawn above each category (class) label. The height of each bar indicates the frequency, relative frequency or percent frequency of each category (class).
A **pie chart** can be used to graphically display categorical data summarized in a frequency, relative frequency or percent frequency distribution. A pie chart is a circle subdivided into sectors, or parts, each with a surface area, relative to the surface area of the whole circle, corresponding to the relative frequency or percent frequency of each category (class). The numerical values shown for each sector can be frequencies, relative frequencies, or percent frequencies.

2.2 Summarizing Data for a Quantitative Variable

When constructing a **frequency distribution** for quantitative data, classes are formed by specifying ranges that will be used to group the data. A general guideline is to use **between 5 and 20 classes**, of **equal width**, by setting **class limits** so as to have nonoverlapping classes. The goal is to use enough classes to show the variation in the data, but not so many classes that some contain only a few data items. To determine an approximate class width, divide the range by the number of classes you'd like, where the range is largest data value minus the smallest data. Ultimately, the analyst uses judgment to determine the combination of the number of classes and class width that provides the best frequency distribution for summarizing the data. Class limits must be specified so that each data item belongs to one and only one class. Note that the class width is equal to the difference between the lower class limits of adjacent classes.

In some applications, we want to know the midpoints of the classes in a frequency distribution for quantitative data—the **class midpoint** is the value halfway between the lower and upper class limits.

SKIP: Dot Plot & Stem-and-Leaf Display

A **histogram**, where each bar represents a class in a distribution of quantitative data, can be used to graphically display quantitative data summarized in a frequency, relative frequency, or percent frequency distribution. The horizontal axis specifies each class' lower and upper limits, below each bar. On that horizontal axis, the bars display classes in ascending order, from left to right. The vertical axis specifies the frequency, relative frequency, or percent frequency scale. Unlike a bar chart, the bars, in a histogram, touch each other (they are not separated by space). Like in a bar chart, the bars are of fixed width. The height of each bar indicates the frequency, relative frequency or percent frequency of each class.

A histogram is said to be **skewed to the left** if its tail extends farther to the left, indicating that a lower proportion or percentage of the data fall into lower classes.
A histogram is said to be **skewed to the right** if is tail extends farther to the right, indicating that a lower proportion or percentage of the data fall into higher classes.
A histogram is said to be **symmetric** if the left tail mirrors the shape of the right tails, indicating that a similar proportion or percentage of the data fall into lower and higher classes.

Cumulative frequency distribution: tabular summary of quantitative data showing the number (frequency) of observations with values less than or equal to the upper class limit of each class. **Cumulative relative frequency distribution**: tabular summary of quantitative data showing the relative frequency (proportion) of observations with values less than or equal to the upper class limit of each class.
Cumulative percent frequency distribution: tabular summary of quantitative data showing the percent frequency (percent) of observations with values less than or equal to the upper class limit of each class.

2.3 Summarizing Data for Two Variables Using Tables

Crosstabulation: tabular summary of data for two variables (categorical or quantitative).
Simpson's Paradox: conclusions drawn from two or more separate crosstabulations that can be reversed when the data are aggregated into a single crosstabulation.

2.4 Summarizing Data for Two Variables Using Graphical Displays

Scatter diagram: graphical display of the relationship between two quantitative variables or plot of observations with values on two quantitative variables. Values on one variable are illustrated on the horizontal axis, values of the other variable are illustrated on the vertical axis; those values become each observation, or point, Cartesian coordinates.
Trendline: line that provides an approximation of the relation between two quantitative variables.
Side-by-side bar chart: a graphical display for depicting multiple bar charts on the same display.
Stacked bar chart: a bar chart in which each bar is broken into rectangular segments of different colors showing the relative frequency of each class in a manner similar to a pie chart.

2.5 Data Visualization: Best Practices in Creating Effective Graphical Displays

2.5.1 Creating Effective Graphical Displays

- Give the display a clear & concise title.
- Keep the display simple.
- Clearly label each axis & provide the units of measure.
- If using different colors, make sure they are clearly distinct.
- If using different colors or line types, use a legend to define them.
- Minimize the need for screen scrolling.
- Use borders between charts to improve readability.

2.5.2 Choosing the Type of Graphical Display

- Displays to show data distribution: bar & pie charts, histograms.
- Displays to make comparisons: side-by-side & stacked bar charts.
- Displays to show relationships: scatter diagram & trendline.

2.5.3 Data Dashboards

Data dashboard: set of visual displays that organizes & presents timely summary information that is used to monitor the performance of a company or organization in a manner that is easy to read, understand, and interpret.

2.5.4 Data Visualization in Practice: Cincinnati Zoo and Botanical Garden

NOTES:

- Most popular data visualization software available: Gognos, JMP, Spotfire & Tableau.
- Geographic Information System (GIS) is a powerful tool for visualizing geographic data.