AP Statistics Chapter 1 - Exploring Data

1.1: Displaying Distributions with Graphs

Individuals and Variables

- **Individuals** are objects described by a set of data. Individuals may be people, but they may also be animals or things.
- A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

Categorical and Quantitative Variables

- A **categorical variable** places an individual into one of several groups or categories.
- A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

Distribution

The **distribution** of a variable tells us what values the variable takes and how often it takes these variables.

Overall Pattern of a Distribution

To describe the overall pattern of a distribution:

- Give the center and the spread
- See if the distribution has a simple shape that you can describe in a few words.

Outliers

An outlier in any graph of data is an individual observation that falls outside the overall pattern of the graph.

Symmetric and Skewed Distributions

- A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.
- A distribution is **skewed to the right** if the right side of the histogram (containing half of the observations with larger values) extend much farther out than the left side.
- A distribution is **skewed to the left** if the left side of the histogram extends much farther out than the right side

Percentile

The *p*th **percentile** of a distribution is the value such that *p* percent of the observation fall at or below it.

Time Plot

- A **time plot** of a variable plots each observation against the time at which it was measured.
- Always mark the time scale on the horizontal axis and the variable of interest on the vertical axis. If there are not too many points, connecting the points by lines helps show the pattern of changes over time.

1.2: Describing Distributions with Numbers

The Mean ($\overline{x}$)

To find the **mean** of a set of observations, add their values and divide by the number of observations. If the n observations are x1, x2, …, xn, their mean is:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

or the more compact notation,

$$\overline{x} = \sum_{i=1}^{n} x_i$$

The Median (M)

- The median M is the midpoint of distribution, the number such that half the observations are smaller and the other half are larger. To find the median of distribution:
- Arrange all observation in order of size, from smallest to largest.
- If the number of observations n is odd, the median M is the center observation in the ordered list.
- If the number of observations n is even, the median M is the mean of the two center observations in the ordered list.

The Quartiles (Q_1 and Q_3)

- To calculate the quartiles Arrange the observations in increasing order and locate the median M in the ordered list of observations.
- The first quartile Q1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
- The third quartile Q3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

The Interquartile Range (IQR)

The interquartile range (IQR) is the distance between the first and third quartiles,
IQR = Q3 - Q1

Outliers: The 1.5 x IQR Criterion

Call an observation an outlier if it falls more than 1.5 x IQR above the third quartile or below the first quartile.

- Low outlier cutoff: $Q_1 - 1.5 \times IQR$
- High outlier cutoff: $Q_3 + 1.5 \times IQR$

The Five-Number Summary

The five-number summary of a data set consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is:

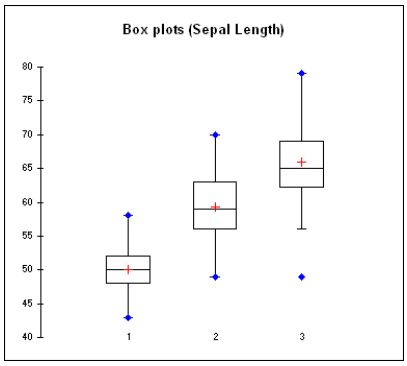
Minimum – Q_1 – M – Q_3 – Maximum

Boxplot (Modified)

A modified boxplot is a graph of the five-number summary, with outliers plotted individually.

- A central box spans the quartiles.
- A line in the box marks the median.
- Observations more than 1.5 x IQR outside the central box are plotted individually.
- Lines extend from the box out to the smallest and largest observations, not the outliers.

Example: The 3rd boxplot below shows an outlier marked as a separate point. This is a modified boxplot.



The Standard Deviation s

The standard deviation of a set of observations is the average of the squares of the deviations of the observations from their mean. The formula for the standard deviation of n observations $x_1, x_2, \ldots, x_n$ is:

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

Linear Transformations of Data

A linear transformation on data is of the form $x_{new} = a + bx$, where *a* and *b* are numbers.

Rules for the effect of a linear transformation on measures of center and spread:

- Adding the same number *a* to each observation adds *a* to measures of center but does not change measures of spread.
- Multiplying each observation by a positive number *b* multiplies both measures of center and spread by *b*.