Math 217: Module 2 Review

To analyze the distribution of a quantitative variable, we describe the overall pattern of the data (shape, center, spread), and any deviations from the pattern (outliers).

Three types of graphs to analyze the distribution of a quantitative variable:

- **Dotplots**
  - Individual variable values are visible, particularly when the data set is small.
  - Descriptions of shape, center, and spread are not affected by how the dotplot is constructed.
  - We can accurately calculate the overall range (largest value – smallest value).

- **Histograms**
  - Individual variable values are not visible.
  - Groups individuals into bins of equal-sized intervals. This is particularly useful when analyzing large data sets.
  - We can easily use percentages, also called relative frequencies, to describe the distribution.

- **Boxplots**
  - Displays the five-number summary (min, Q1, med, Q3, and max).
  - Highlights any points that are considered outliers.
  - Commonly used to compare two data sets.

Four ways to described the shape of a distribution:

- **Skewed left**
- **Skewed right**
- **Bell-Shaped**
- **Uniform**
  
  Note: Not all distributions have a simple shape that fits into one of these categories.

The center of a distribution is a typical value that represents the group. Two measurements for determining the center of a distribution:

- **Mean**: Calculate the mean by adding the data values and dividing by the number of individual data points. The mean is also referred to as “the balancing point” of a distribution. If we measure the distance between each data point and the mean, the distances are balanced on each side of the mean.
  
  \[ \bar{x} = \frac{\sum x}{n} \]

- **Median**: The physical center of the data when we make an ordered list. It has the same number of values above it as below it. It is either the middle number or the average of the 2 middle numbers.

**General guidelines for choosing a measure of center**:

- Always plot the data. We need to use a graph to determine the shape of the distribution. By looking at the shape, we can determine which measure of center best describes the data.
- Use the mean as a measure of center for distributions that are reasonably symmetric with a central peak. When outliers are present, the mean is not a good choice.
- Use the median as a measure of center for all other cases.
The spread of a distribution is a description of how the data varies. We studied 3 ways to measure spread:

- **Range** = max – min
- **Interquartile range (IQR)** = Q3 – Q1: measures the variability in the middle half of the data
  - Five number summary {min, Q1, Median, Q3, max}
- **Standard deviation (SD)**: approximately the average distance of the data from the mean
  - Deviation from the mean: \( x - \bar{x} \)
  - Variance: \( s^2 = \frac{\text{sum of the squared deviations}}{n-1} \)
  - Empirical Rule for bell-shaped data
    - 68% of the data is within 1 standard deviation of the mean
    - 95% of the data is within 2 standard deviations of the mean
    - 99.7% of the data is within 3 standard deviations of the mean
  - **z-scores for data values** - how many standard deviations above or below the mean
    \[
    z - \text{score} = \frac{\text{data value} - \bar{X}}{s}
    \]

**Outliers**:
“Extreme values” They fall outside the fence.

Fence
\[
\begin{align*}
Q1 - 1.5(IQR) \\
Q3 + 1.5(IQR)
\end{align*}
\]

“Extreme values” Have a z-score greater than 3 or less than -3.

Note: Mean and standard deviation go together, and median and IQR go together. The mean and SD are strongly affected by outliers, while median and IQR are not.