

Objectives:

- To learn to calculate certain statistics (*range*, *interquartile range*, *Standard deviation*) for summarizing the variability of a distribution of data
- To discover the *five-number summary* of a distribution
- To explore the *boxplot* as another convenient and informative visual display of a distribution
- To investigate and determine properties of these summary statistics
- To understand the *empirical rule* as a means for interpreting the value of standard deviation for certain types of distributions
- To appreciate the applicability of calculating *z-scores* for comparing distributions of different variables
- To recognize some of the limitations of these measures of variability

Three commonly used measures of variability of a distribution:

- The **range** is simply the difference between the largest and the smallest value in the distribution
- The **interquartile range (IQR)** is the difference between the upper and lower quartiles (75th and 25th percentiles). A simple way to do this is to first find the median of all the observations. This divides the observations into two groups. Then find the median of each of the two groups formed (If N is odd then the median is not included in either group) the median of the upper group is the 75th percentile; the lower is the 25th percentile. Find the difference between these two values to find the **IQR**.

Lower Quartile (Q_1 or Q_L) the value that 25% of the values fall below and 75% fall above.

Upper Quartile (Q_3 or Q_U) the value that 75% of the values fall below and 25% fall above.

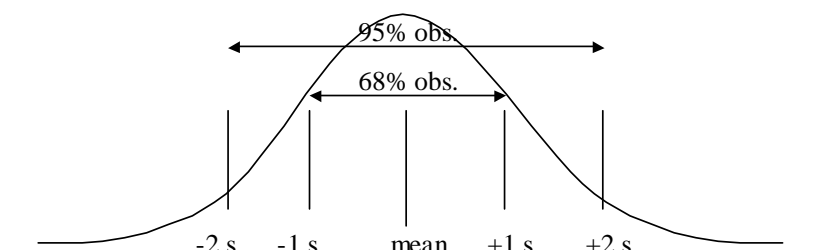
IQR represents the middle 50% of the data (the “average Joe”)

- The **standard deviation** is a more complicated but more telling measure of distribution. It is based on the squared deviations from the mean.
 1. Calculate the mean of the observations
 2. Find the difference between each observation and the mean
 3. Square each difference
 4. Add up all the differences
 5. Divide the resulting sum by $n-1$
 6. Finally, find the square root.

The following is the formula for finding the standard deviation (if you don't remember the notation look at Topic 3 notes)

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

Empirical rule The empirical rule states that dev if the distribution is **normal**, approximately 68% of all observations in a distribution will fall within +/- 1 std dev of the mean and 95% of the observations will fall within +/- 2 std dev. and virtually all (99.7%) of the observations will fall within +/- 3 std dev.



Z-score or **standardized score** are used to find where an observation is placed with relation to the rest of the observations. Standardizing also allows us to compare like data with different measurement techniques. *e.g.* we can not make comparisons between the scores on an SAT and ACT since the techniques are completely different but we can make a comparison of the scores after they are standardized and we can see how far from the mean we are. If John obtained a 29 on the ACT while Sue obtained a 1250 on her SAT, We cannot say anything about John compared to Sue. But if we say John's score had a z-score of 1.3 while Sue had a z-score of 1.4 now we can say Sue scored slightly better than John. The tests are scored differently yet they measure essentially the same thing and now we can find out who did better only after standardizing the scores.

Finding a **z-score**: first find the difference between the observation and the mean then divide by the std dev. This is the z-score. This should only be used when the distribution of the observations is approximately *normal*.

$$z - score = \frac{x - \bar{x}}{s} = \frac{observation - mean}{stddev} = \frac{x - \mu}{\sigma}$$