

DEFINITIONS AND DESCRIPTIVE STATISTICS

For demonstrations, our data set will consist of the following list: 2, 2, 3, 5, 5, 5, 7, 8, 9, 9, 12, 13 and has been ordered from smallest to largest. Σ means to sum, n = number of data points, x_i = any one particular data point

MEASURES OF CENTER

ARITHMETIC MEAN $\bar{x} = \frac{1}{n} \Sigma x_i$ Example: $\frac{2+2+3+5+5+5+7+8+9+9+12+13}{12} = \frac{80}{12} \approx 6.7$

MEDIAN: Data point in the middle of a data set that has been ordered from smallest (labeled X_1) to biggest (labeled X_{12}). (This is also the second quartile Q_2) (see below)

Example: $n = 12$, so the median lies between X_6 and X_7 in the sorted data:

$$\frac{5+7}{2} = 6 = \text{median}$$

Note on MEDIAN: If the number in the data set is odd, then the middle number is used. For example, if we dropped the 13 from our data set, we would have 11 data points from 2 to 11. X_5 would then be the MEDIAN.

MODE: Data point with most occurrence. In our example data, 5 is the *mode* of our set. Data sets with two modes are *bimodal*, and data sets with more than two modes are *multimodal*.

MEASURES OF SPREAD

RANGE: The range of the data is the MAX – MIN. In our data set, MAX = 13 and MIN = 2, so our RANGE is 13-2 = 11

QUARTILES: The first quartile Q_1 is in the middle of the start of the data and the median (Q_2). In our example:

2, 2, 3, ($Q_1 = \frac{3+5}{2} = 4$) 5, 5, 5, ($Q_2=6$) 7, 8, 9, 9, 11, 13

The third quartile Q_3 is between the middle of the data set and the end of the data set.

2, 2, 3, 5, 5, 5, ($Q_2=6.5$) 7, 8, 9, ($Q_3 = \frac{9+9}{2} = 9$) 9, 11, 13

Note that min, Q1, Median, Q3 and max are known as the FIVE NUMBER SUMMARY.

PERCENTILES: If p is the percentile desired, then the percentile in the data set can be found by:

$$p * n = L$$

where L is the location- given by the subscript i of X_i - in the ordered list.

Example: P_{90} is located at $0.90 * 12 = 10.8$, which we round off at 11. Therefore $P_{90} = X_{11} = 12$ Note that $Q_1 = P_{25}$ and $Q_3 = P_{75}$.

INTERQUARTILE RANGE (IQR): $IQR = Q_3 - Q_1$ Example: $9 - 4 = 5$ so $IQR = 5$

OUTLIER: A data point is a suspected outlier if it falls more than $1.5 * IQR$ above the third quartile or below the first quartile.

Example: $1.5(5) = 7.5$ So an outlier would be below $2-7.5 = -5.5$ or $9+7.5 = 16.5$, therefore our data set has no outliers.

VARIANCE and STANDARD DEVIATION: VARIANCE is the measure of data spread. It is an unbiased estimator (ie, its expected value and its true value are the same) whereas its square root, STANDARD DEVIATION, is a biased estimator. However, the STANDARD DEVIATION is in the same units as the mean and the data set.

VARIANCE is calculated by: $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$

STANDARD DEVIATION is the square root of VARIANCE.

Example:

$$s^2 = \frac{1}{12-1} \left[\begin{array}{l} (2 - 6.7)^2 + (2 - 6.7)^2 + (3 - 6.7)^2 + (5 - 6.7)^2 + (5 - 6.7)^2 + (5 - 6.7)^2 + (7 - 6.7)^2 \\ + (8 - 6.7)^2 + (9 - 6.7)^2 + (9 - 6.7)^2 + (11 - 6.7)^2 + (13 - 6.7)^2 \end{array} \right] \approx 13.33$$

therefore $S = \sqrt{13.33} = 3.65$