Chapter-1

Data: collections of observations like survey

Statistics: obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data

Population: the complete collection of all individuals to be studied.

Census: Collection of data from every member of a population

Sample: Sub collection of members selected from a population

- Sample data must be collected by random selection.
- If sample data are not collected by random: the data may be so completely useless that no amount of statistical torturing can salvage them.

Conclusions: Make statements that are clear to those without an understanding of statistics and its terminology, and Avoid making statements not justified by the statistical analysis.

Parameter: characteristic of a population.

Statistic: characteristic of a sample.

Categorical(qualitative): consists of names or labels

Example: Color, Gender, Religion, Zip Code

- Mathematical operations are Meaningless

Quantitative(numerical): consists of numbers representing counts or measurements

Example: ages, weights, temp, Time

- Mathematical operations are Meaningful

Working with Quantitative Data:

❖ **Discrete data:** result is a finite number or a 'countable' number (usually A count)

Example

- Example: The number of eggs that a hen lays
- the number of possible values is 0, 1, 2, 3, . . .)

❖ **Continuous:** result from infinitely many possible values(not count) (usually A Measurement)

Example:

- The amount of milk that a cow produces; e.g. 2.343115 gallons per day
- Temp

Levels of Measurement:

1. **Nominal**: Categories Not ordered
   - Example: Survey responses yes, no, undecided
2. **Ordinal**: Categories with some order, Differences are Meaningless.
   - Example: Rank, Color, Course grades A, B, C, D, or F
3. **Interval**: ordered, difference are Meaningful (no natural zero).
   - Example: Temp, Years 1000, 2000
4. **Ratio**: Just Like Interval but with A Natural Zero.
   - Example: Prices of college textbooks, Amount of Money.

Collecting Data:

- **Observational study:** Measures specific Traits but Dose Not modify subjects
- **Experiment:** apply some treatment and then measure its effects on the subjects
- **Random**: Each Member of A Population Has An Equal chance Of Being Selected in the Sample
- **Simple Random Sample:** Each Group of size (n) Has an Equal chance Of Being Selected

Methods of Sampling or Techniques:

- **Random Sampling**: selection so that each individual member has an equal chance of being selected
- **Convenience Sampling:** use results that are easy to get. (Not Random)
- **Systematic Sampling:** Select some starting point and then select every kth element in the population.
- **Stratified Sampling:** Break population into Subgroups Based on Characteristic then Sample Each Subgroup.
- **Cluster Sampling**: divide the population area into sections (or clusters) randomly select some of those clusters; choose all members from selected clusters
- **Multistage Sampling**: Collect data by using some combination of the basic sampling methods

Types of Studies:

- **Cross sectional study:** data are observed, measured, and collected at one point in time.
- **Retrospective (or case control) study:** data are collected from the past by going back in time (examine records, interviews)
- **Prospective (or longitudinal or cohort) study:** data are collected in the future from groups sharing common factors (called cohorts)

Randomization: is used when subjects are assigned to different groups through a process of random selection. The logic is to use chance as a way to create two groups that are similar.

Replication: is the repetition of an experiment on more than one subject. Samples should be large enough so that the erratic behavior that is characteristic of very small samples will not disguise the true effects of different treatments. It is used effectively when there are enough subjects to recognize the differences from different treatments.

Blinding: is a technique in which the subject doesn't know whether he or she is receiving a treatment or a placebo.

Double-Blind: Blinding occurs at two levels:

1. The subject doesn't know whether he or she is receiving the treatment or a placebo
2. The experimenter does not know whether he or she is administering the treatment or placebo

Confounding: occurs in an experiment when the experimenter is not able to distinguish between the effects of different factors.

Controlling Effects of Variables

- **Completely Randomized Experimental Design:** assign subjects to different treatment groups through a process of random selection
- **Randomized Block Design** a block is a group of subjects that are similar, but blocks differ in ways that might affect the outcome of the experiment
- **Rigorously Controlled Design** carefully assign subjects to different treatment groups, so that those given each treatment are similar in ways that are important to the experiment
- **Matched Pairs Design** compare exactly two treatment groups using subjects matched in pairs that are somehow related or have similar characteristics

Errors:

- **Sampling error:** the difference between a sample result and the true population result; such an error results from chance sample fluctuations
- **Non-sampling error:** sample data incorrectly collected, recorded, or analyzed (such as by selecting a biased sample, using a defective instrument, or copying the data incorrectly)

Chapter-2

Important Characteristics of Data:

- **Center:** A representative or average value that indicates where the middle of the data set is located.
- **Variation:** A measure of the amount that the data values vary.
- **Distribution:** The nature or shape of the spread of data over the range of values (such as bell-shaped, uniform, or skewed).
- **Outliers:** Sample values that lie very far away from the vast majority of other sample values.
- **Time:** Changing characteristics of the data over time.

Frequency Distribution (or Frequency Table) Definitions:

- **Frequency Distribution:** A list of Values with Corresponding Frequency.
- **Class Width:** difference between Two Lower Class Limits
- **Lower Class Limit:** Smallest value belong To A Class
- **Upper Class Limits:** Largest value belong To A Class
- **Class Boundaries:** Numbers used to separate classes, but without the gaps created by class limits.
- **Class Midpoints:** the values in the middle of the classes. $\text{Class midpoint} = \frac{upper\ class\ limit + lower\ class\ limit}{2}$

Reasons for Constructing Frequency Distributions:

- Large data sets can be summarized.
- We can analyze the nature of data.
- We have a basis for constructing important graphs.

Steps for Constructing a Frequency Distribution:

1. Determine the number of classes (should be between 5 and 20).
2. Calculate the **class width** (round up). $\text{class width} \approx \frac{\text{(maximum value) – (minimum value)}}{\text{number of classes}}$
3. **Starting point:** Smallest value.
4. Create Classes using Class Width.
5. List the lower class limits in a vertical column First and proceed to enter the upper class limits.
6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to get the frequency.

Relative Frequency Distribution:

Includes the same class limits as a frequency distribution, but the frequency of a class is replaced with a relative frequency (a proportion) or a percentage frequency (a percent)

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$

$$\text{percentage frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}} \times 100\%$$

Cumulative Frequency Distribution: Adds Sequential classes' together. (Sum of the class and all classes below it in a frequency distribution)

Normal distribution:

- The frequencies start low, then increase to one or two high frequencies, then decrease to a low frequency.
- The distribution is approximately symmetric, with frequencies preceding the maximum being roughly a mirror image of those that follow the maximum.

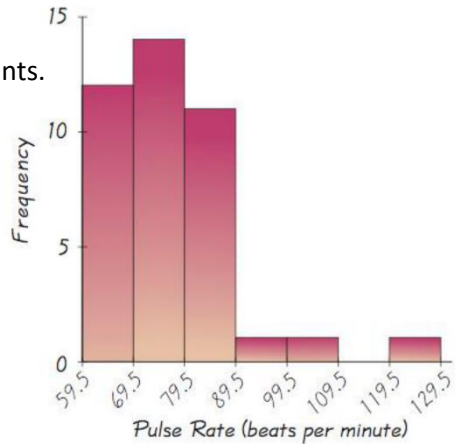
Gaps: The presence of gaps can show that we have data from two or more different populations.

Histogram:

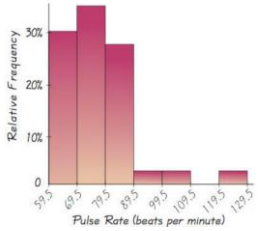
- is a graphical representation of the distribution of numerical data
- Touching bar chart
- A graph consisting of bars of equal width drawn adjacent to each other (without gaps).
- Basically a graphic version of a frequency distribution.

Histogram Scale

- Horizontal Scale for Histogram: Use class boundaries or class midpoints.
- Vertical Scale for Histogram: Use the class frequencies.



Relative Frequency Histogram: Has the same shape and horizontal scale as a histogram, but the vertical scale is marked with relative frequencies instead of actual frequencies



Characteristic of a normal distribution (bell shape) are:

- The frequencies increase to a maximum, and then decrease.
- symmetry, with the left half of the graph roughly a mirror image of the right half.

Frequency Polygon: Uses line segments connected to points directly above class midpoint values

Relative Frequency Polygon: Uses relative frequencies (proportions or percentages) for the vertical scale.

Ogive: A line graph that depicts cumulative frequencies.

Dot Plot: Consists of a graph in which each data value is plotted as a point (or dot) along a scale of values. Dots representing equal values are stacked.

Stem plot (or Stem-and-Leaf Plot): Represents quantitative data by separating each value into two parts: the stem (such as the leftmost digit) and the leaf (such as the rightmost digit)

Bar Graph: Uses bars of equal width to show frequencies of categories of qualitative data. Vertical scale represents frequencies or relative frequencies. Horizontal scale identifies the different categories of qualitative data.

Multiple Bar Graph: has two or more sets of bars, and is used to compare two or more data sets.

Pareto Chart: A bar graph for qualitative data, with the bars arranged in descending order according to frequencies

Pie Chart: A graph depicting qualitative data as slices of a circle, size of slice is proportional to frequency count

Scatter Plot (or Scatter Diagram): A plot of paired (x,y) data with a horizontal x-axis and a vertical y-axis. Used to determine whether there is a relationship between the two variables

Time-Series Graph: Data that have been collected at different points in time: time-series data

Bad Graphs:

- **Nonzero Axis:** Are misleading because one or both of the axes begin at some value other than zero, so that differences are exaggerated.
- **Pictographs:** are drawings of objects. Three-dimensional objects money bags, stacks of coins are commonly used to depict data. These drawings can create false impressions that distort the data.

Chapter-3

Measure of Center: The Middle of the Data Set.

- **Mean (Average):** Add all the values and dividing the total by the number of values
  - Advantages: Is relatively reliable, means of samples drawn from the same population don't vary as much as other measures of center,Takes every data value into account
  - Disadvantage: Is sensitive to every data value, one extreme value can affect it dramatically; is not a resistant measure of center
- **Median**: the middle value of A Data Set and Must be in order.
  - often denoted by $\tilde{x}$ (pronounced 'x-tilde')
  - is not affected by an extreme value - is a resistant measure of the center
  - Finding the Median:
    - If number of values is odd, The Median is the Middle number
    - If number of values is Even, The Median is the Mean of the two Middle number
- **Mode:** the value that occurs with the greatest frequency
  - Data set can have one, more than one, or no mode:
    - Bimodal: two data values occur with the same greatest frequency
    - Multimodal: more than two data values occur with the same greatest frequency
    - No Mode: no data value is repeated
  - Mode is the only measure of central tendency that can be used with nominal data

Notation:	
$\sum$	denotes the sum of a set of values.
x	is the variable usually used to represent the individual data values.
n	represents the number of data values in a sample.
N	represents the number of data values in a population.
$\bar{x}$	is pronounced 'x-bar' and denotes the mean of a set of sample values
μ	is pronounced 'mu' and denotes the mean of all values in a population

$$\bar{x} = \frac{\sum x}{n} \qquad \mu = \frac{\sum x}{N}$$

Midrange: the value midway between the maximum and minimum values in the original data set

- **Sensitive to extremes**: because it uses only the maximum and minimum values, so rarely used
- **Redeeming Features:**
  - very easy to compute
  - reinforces that there are several ways to define the center
  - Avoids confusion with median

Round-off Rule for Measures of Center: Carry one more decimal place than is present in the original set of values.

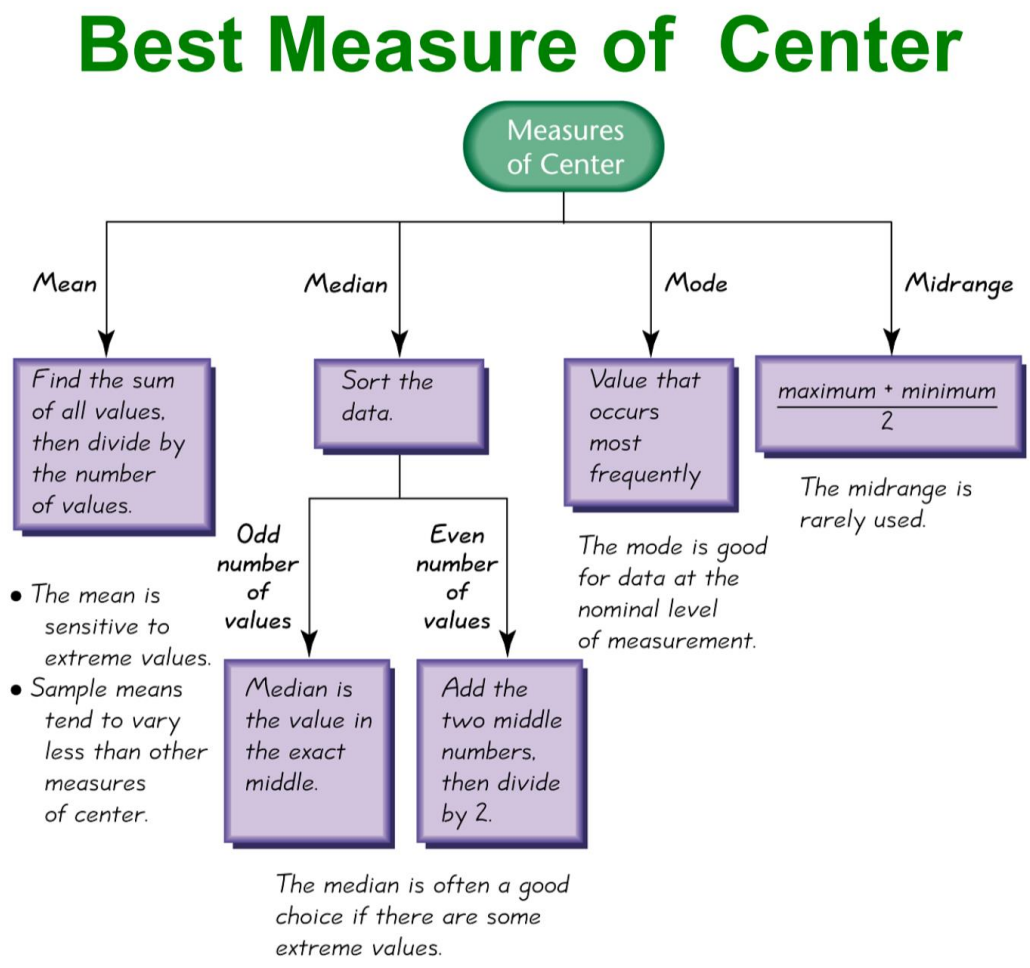
Mean from a Frequency Distribution:

- Assume that all sample values in each class are equal to the class midpoint.
- use class midpoint of classes for variable x

$$\bar{x} = \frac{\Sigma(f \cdot x)}{\Sigma f}$$

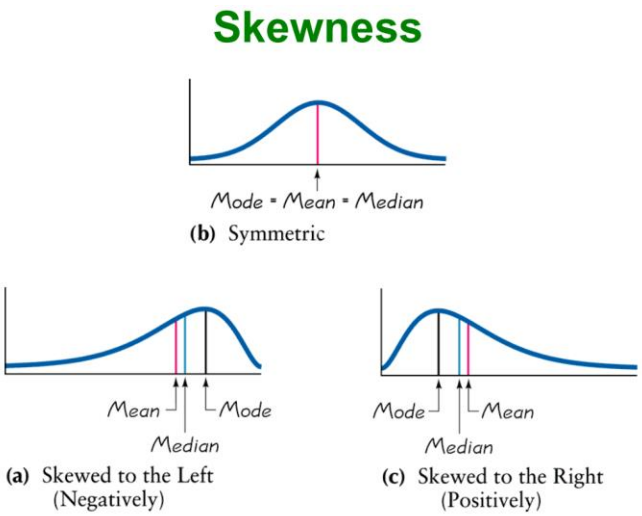
Weighted Mean: When data values are assigned different weights, we can compute a weighted mean.

$$\bar{x} = \frac{\Sigma (w \cdot x)}{\Sigma w}$$



- Symmetric: distribution of data is symmetric if the left half of its histogram is roughly a mirror image of its right half
- Skewed: distribution of data is skewed if it is not symmetric and extends more to one side than the other.
- Skewed to the left: (also called negatively skewed) have a longer left tail, mean and median are to the left of the mode
- Skewed to the right: (also called positively skewed) have a longer right tail, mean and median are to the right of the mode

Shape of the Distribution: The mean and median cannot always be used to identify the shape of the distribution.



Measures of Variation:

- Definition: The range of a set of data values is the difference between the maximum data value and the minimum data value.
- It is very sensitive to extreme values; therefore, not as useful as other measures of variation

Ways to measure Variation:

- **Range = (maximum value) – (minimum value)**
  - Easy to find
  - does not consider all values.
- standard deviation: Measures the Average Distance your data values are from the Mean
  - Never Negative and Never 0 unless All entries are the Same.
  - Greatly Affected by Outliers

Sample Standard Deviation Formula:

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

Sample Standard Deviation (Shortcut Formula):

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

Standard Deviation Important Properties:

- The standard deviation is a measure of variation of all values from the mean.
- The value of the standard deviation s is usually positive.
- The value of the standard deviation s can increase dramatically with the inclusion of one or more outliers (data values far away from all others).
- The units of the standard deviation s are the same as the units of the original data values.

Comparing Variation in Different Samples:

- It's a good practice to compare two sample standard deviations only when the sample means are approximately the same.
- When comparing variation in samples with very different means, it is better to use the coefficient of variation, which is defined later in this section.

Population Standard Deviation:

$$\sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{N}}$$

Variance: The variance of a set of values is a measure of variation equal to the square of the standard deviation.

- **Sample variance:** s^2 Square of the sample standard deviation s
- **Population variance:** σ^2 Square of the population standard deviation σ

Unbiased Estimator: The sample variance s^2 is an unbiased estimator of the population variance σ^2, which means values of s^2 tend to target the value of σ^2 instead of systematically tending to overestimate or underestimate σ^2.

Notation:	
s	sample standard deviation
s^2	sample variance
σ	population standard deviation
σ^2	population variance

Range Rule of Thumb: is based on the principle that for many data sets, the vast majority (such as 95%) of sample values lie within two standard deviations of the mean.

Range Rule of Thumb for Interpreting a Known Value of the Standard Deviation: Informally define usual values in a data set to be those that are typical and not too extreme. Find rough estimates of the minimum and maximum "usual" sample values as follows:

- **Minimum "usual" value = (mean) – 2 * (standard deviation)**
- **Maximum "usual" value = (mean) + 2 * (standard deviation)**

Range Rule of Thumb for Estimating a Value of the Standard Deviation s: To roughly estimate the standard deviation from a collection of known sample data use

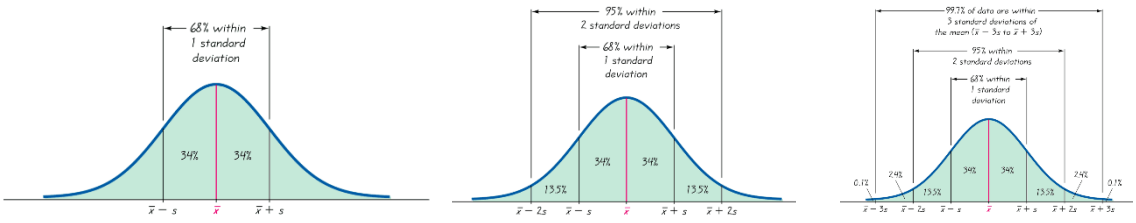
$$s \approx \frac{\text{range}}{4}$$

Properties of the Standard Deviation:

- Measures the variation among data values
- Values close together have a small standard deviation, but values with much more variation have a larger standard deviation
- Has the same units of measurement as the original data
- For many data sets, a value is unusual if it differs from the mean by more than two standard deviations
- Compare standard deviations of two different data sets only if the they use the same scale and units, and they have means that are approximately the same

Empirical (or 68-95-99.7) Rule: For data sets having a distribution that is approximately bell shaped, the following properties apply:

- About 68% of all values fall within 1 standard deviation of the mean.
- About 95% of all values fall within 2 standard deviations of the mean.
- About 99.7% of all values fall within 3 standard deviations of the mean.



Chebyshev's Theorem: The proportion (or fraction) of any set of data lying within K standard deviations of the mean is always at least $(1 - \frac{1}{K^2})$ where K is any positive number greater than 1.

- For K= 2: at least $\frac{3}{4}$ (or 75%) of all values lie within 2 standard deviations of the mean.
- For K= 3: at least 8/9 (or 89%) of all values lie within 3 standard deviations of the mean.

Rationale for using n – 1 versus n:

- There is only n – 1 independent values. With a given mean, only n – 1 values can be freely assigned any number before the last value is determined.
- Dividing by n – 1 yields better results than dividing by n. It causes s^2 to target σ^2 whereas division by n causes s^2 to underestimate σ^2.

Coefficient of Variation:

The coefficient of variation (or CV) for a set of nonnegative sample or population data, expressed as a percent, describes the standard deviation relative to the mean.

Sample	Population
$CV = \frac{s}{\bar{x}} \cdot 100\%$	$CV = \frac{\sigma}{\mu} \cdot 100\%$

Measures of Relative Standing: Measures of Relative Standing

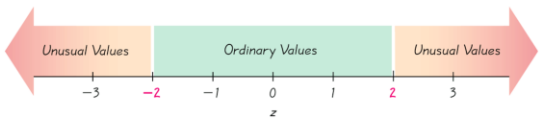
Z score: (or standardized value) the number of standard deviations that a given value x is above or below the mean

Measures of Position z Score:

Sample	Population
$z = \frac{x - \bar{x}}{s}$	$z = \frac{x - \mu}{\sigma}$

Round z scores to 2 decimal places

Interpreting Z Scores:



Whenever a value is less than the mean, its corresponding z score is negative

- Ordinary values: –2 ≤ z score ≤ 2
- Unusual Values: z score < –2 or z score > 2

Percentiles: are measures of location. There are 99 percentiles denoted P1, P2, . . . P99, which divide a set of data into 100 groups with about 1% of the values in each group.

Finding the Percentile of a Data Value:

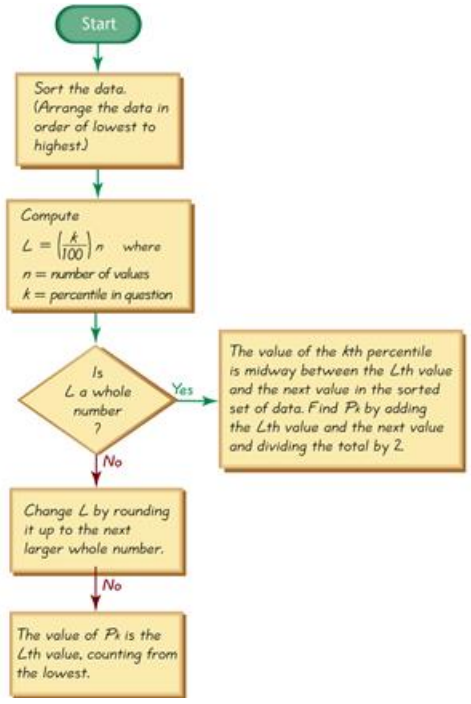
$$\text{Percentile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

Converting from the kth Percentile to the Corresponding Data Value:

Notation

$$L = \frac{k}{100} \cdot n$$

- n total number of values in the data set
- k percentile being used
- L locator that gives the position of a value
- P_k kth percentile

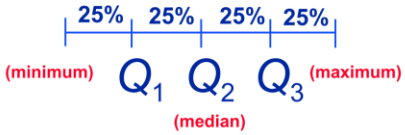


Quartiles: Are measures of location, denoted Q1, Q2, and Q3, which divide a set of data into four groups with about 25% of the values in each group.

- Q1 (First Quartile) separates the bottom 25% of sorted values from the top 75%.
- Q2 (Second Quartile) same as the median; separates the bottom 50% of sorted values from the top 50%.
- Q3 (Third Quartile) separates the bottom 75% of sorted values from the top 25%.

Q_1, Q_2, Q_3

divide ranked scores into four equal parts



Some Other Statistics:

- **Interquartile Range (or IQR):** $Q_3 - Q_1$
- **Semi-interquartile Range:** $\frac{Q_3 - Q_1}{2}$
- **Midquartile:** $\frac{Q_3 + Q_1}{2}$
- **10 - 90 Percentile Range:** $P_{90} - P_{10}$

5-Number Summary: For a set of data, the 5-number summary consists of the minimum value; the first quartile Q1; the median (or second quartile Q2); the third quartile, Q3; and the maximum value.

Boxplot: A boxplot (or box-and-whisker diagram) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile, Q1; the median; and the third quartile, Q3.

Outliers: An outlier is a value that lies very far away from the vast majority of the other values in a data set.

Important Principles:

- An outlier can have a dramatic effect on the mean.
- An outlier can have a dramatic effect on the standard deviation.
- An outlier can have a dramatic effect on the scale of the histogram so that the true nature of the distribution is totally obscured.

Outliers for Modified Boxplots:

- For purposes of constructing modified boxplots, we can consider outliers to be data values meeting specific criteria.
- In modified boxplots, a data value is an outlier if it is . . .
- above Q3 by an amount greater than 1.5 * IQR OR below Q1 by an amount greater than 1.5 * IQR