

A **frequency distribution** is a table used to describe a data set. A frequency table lists intervals or ranges of data values called **data classes** together with the number of data values from the set that are in each class. This number is called the **frequency** of the class.

Lower Class Limit – The least value that can belong to a class.

Upper Class Limit – The greatest value that can belong to a class.

Class Width – The difference between the upper (or lower) class limits of consecutive *classes*. All classes should have the same class width.

Class Midpoint – The middle value of each data class. To find the class midpoint, average the upper and lower class limits.

$$\text{class midpoint} = \frac{\text{upper} + \text{lower}}{2}$$

Class Boundaries – The numbers that separate classes without forming gaps between them.

Range (of data) – The highest value – the lowest value

The **cumulative frequency** of a data class is the number of data elements in that class and all previous classes. (It can be ascending or descending.)

The **relative frequency** of a data class is the *percentage* of data elements in that class. We can calculate the relative frequency for each class as follows:

$$\text{relative frequency} = \frac{f}{n}$$

sum of frequencies should be 1:

$$\sum \frac{f}{n} = 1$$

Variables

x = data value

n = number of values in a sample data set

N = number of values in a population data set

f = frequency of a data class

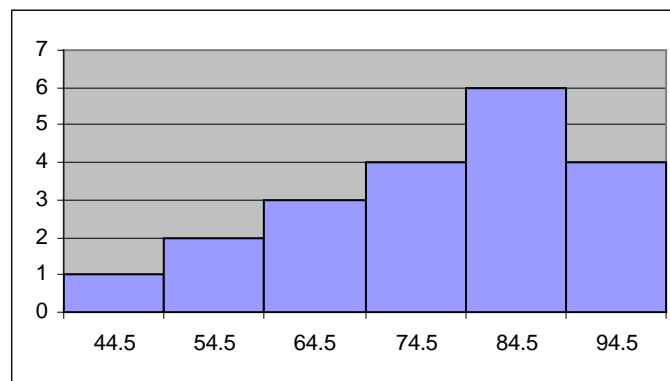
Symbol

Σ = the sum of all values for the following variable or expression.

A **histogram** is a graphical representation of the information in a frequency table using a bar graph.

The histogram should have the *variable being measured* in the data set as its *horizontal axis*, and the *class frequency* as the *vertical axis*. Each *data class* will be represented by a *vertical bar whose height is the frequency of the class and whose width is the class width*.

Example: Created in Excel from the data used in the previous examples.



Notice that the bar for each class is centered at the class midpoint, and the bars for successive classes touch.

A **frequency polygon** is a line graph representation of the information in a frequency table.

Like a histogram, the vertical axis represents frequency and the horizontal axis represents the variable being measured in the data set. To construct the graph, a *point is plotted for each class at its midpoint and with height given by the *frequency of the class**. The points are then connected by straight lines.

A **measure of central tendency** is a value used to represent the typical or “average” value in a data set.

Three Common Measures of Central Tendency:

- **Mean** – (average) the sum of all data values divided by the number of values in the data set. The mean of a sample data set is denoted by \bar{x} and the mean of a population data set by the Greek letter μ .

Sample data set:

$$\bar{x} = \frac{\sum x}{n}$$

Population data set:

$$\mu = \frac{\sum x}{N}$$

- **Median** – the value which separates the largest 50% of data values from the lowest 50%. To calculate the median, place data values in number order. If n is odd, the *middle value* is the median. If n is even, the *mean of the two middle values* is the median.

- **Mode** – the data value (or values) which appears the largest number of times in the set. If no data value is repeated, we say that there is no mode.

Exercise: Find the mode(s) of the quiz score data set.

- **Outlier** – a data entry far removed from the other entries in the data set.

A **weighted mean** is used when we want some data values in a set to factor more often into the calculation of the mean than others.

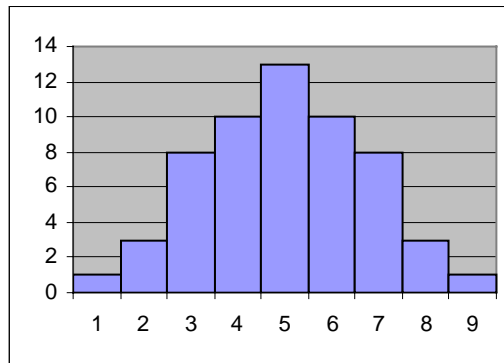
In this case, we attach a numerical **weight** (w) to each value and calculate the mean as follows:

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

Shapes of Data Distributions:

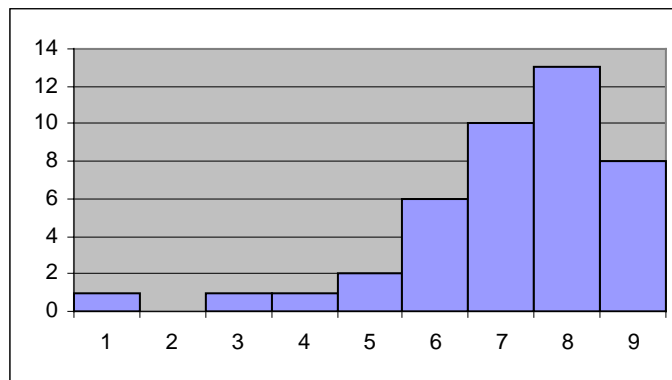
Symmetric – The data distribution is approximately the same shape on either side of a central dividing line.

The mean and median (and mode if unimodal) are equal in a symmetric distribution.



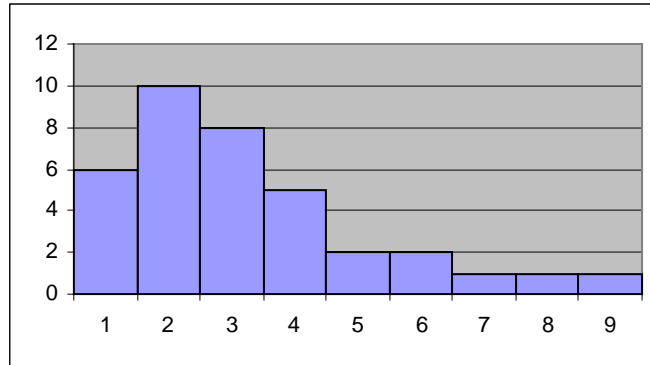
Left-Skewed – A few data values are much lower than the majority of values in the set. (Tail extends to the left)

Generally the mean is less (to the left) than the median (and mode) in a left-skewed distribution.

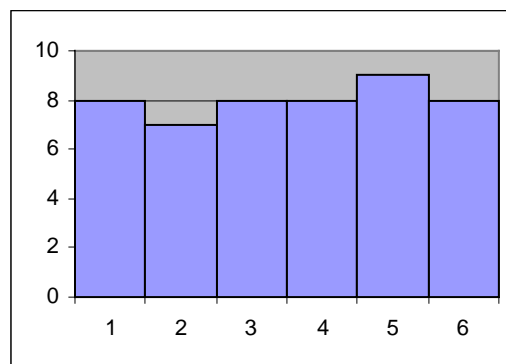


Right-Skewed – A few data values are much higher than the majority of values in the set. (Tail extends to the right)

Generally the mean is greater (to the right) than the median (and mode) in a right-skewed distribution.



Uniform – All data values are equally represented.



Variation in a data set is the amount of difference between data values.

In a data set with little variation, almost all data values would be close to one another. The histogram of such a data set would be narrow and tall.

Example: Quiz Scores: 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5

In a data set with a great deal of variation, the data values would be spread widely. The histogram of this data set would be low and wide.

Example: Quiz Scores: 1, 3, 4, 5, 6, 6, 7, 8, 8, 9, 10

Common Measures of Variation:

1. Range – the difference between the largest and smallest data values in a data set.

$$\text{range} = (\text{highest value} - \text{lowest value})$$

2. Standard Deviation – The most commonly used measure of variation. A measure of the “average” distance of a data value from the mean for the data set.

Standard deviation is calculated using two different formulae depending on whether the data set being considered is a population data set or a sample data set.

Population standard deviation, sigma σ , is calculated using the following formula:

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

Sample standard deviation, s , is calculated using the following formula:

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

3. Variance – the square of the standard deviation. Population variance is represented by σ^2 and sample variance by s^2

Theorems Involving Standard Deviation:

The standard deviation of a data set is an important quantity because it limits the number of data values that can be very far (high or low) from average.

The Empirical Rule (68-95-99.7 Rule)

- Applies only to *bell-shaped* distributions.
- Approximately 68% of data values must be within 1 standard deviation of the mean.
- Approximately 95% of data values must be within 2 standard deviation of the mean.
- Approximately 99.7% of data values must be within 3 standard deviation of the mean.