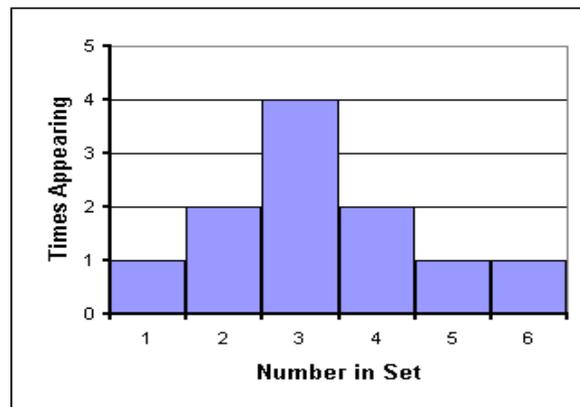# Histogram

A histogram, also known as a **frequency** distribution, is a **static** graphical display of data distribution. Time order sequence is not considered in a histogram and individual data is not displayed. Histograms summarize data by demonstrating the shape and range of a measure's values and estimating the probability distribution of a continuous variable. Tabulated **frequencies**, shown as adjacent rectangles, created over discrete intervals, with an area equal to the frequency of the observations in the interval are displayed. The categories are usually specified as consecutive, non-overlapping intervals of a variable. The categories (intervals) are adjacent and the rectangles of the histogram are drawn so that they touch each other to indicate that the original variable is continuous.

Graph groups of numbers according to **how often** they appear in the data set. A graph of the data set {1,2,2,3,3,3,3,4,4,5,6}, is seen below. This graph is easy to make and provides some useful data about the set. For example, the graph peaks at 3, which is also the median and the mode of the set. The shape of the graph demonstrates how the numbers in the set are distributed about the mean; the distribution of this graph is wide compared to size of the peak, indicating that values in the set are only loosely clustered around the mean.


When to Use a Histogram
- When you want to look at the distribution of data
    - Shape: normal, skewed
    - Spread/Range: tightly clustered, within set limits
- When you want to summarize and analyze extremely large datasets by reducing them to a single graph that can show peaks in data as well as give a visual representation of the statistical significance of those peaks. To get an idea, look at these histograms:

Histogram of Time to Room

Histogram of Time to Room2
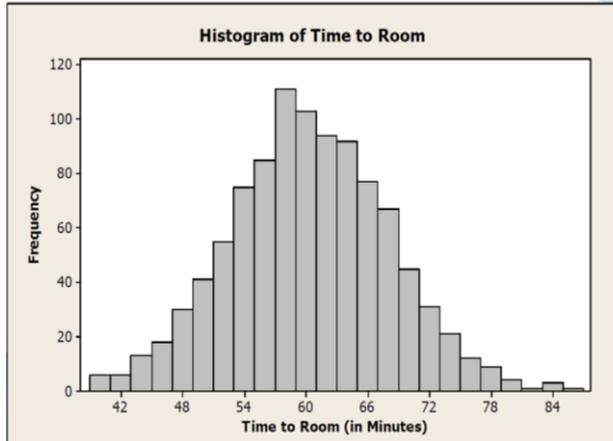
This plot represents data with a well-defined peak that is close in value to the median and the mean. While there are "outliers," they are of relatively low frequency. Thus it can be said that deviations in this data group from the mean are of low frequency.
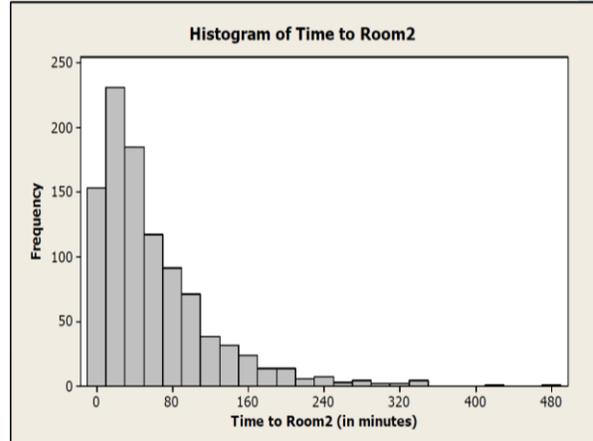
In this plot the data is skewed to the left with few outliers.

### How Histograms Are Constructed

1. Identify the values for the x axis

In most real data sets almost all numbers will be unique. Consider the set {3, 11, 12, 19, 22, 23, 24, 25, 27, 29, 35, 36, 37, 45, 49}. A graph which shows how many ones, how many twos, how many threes, etc. will be meaningless. Instead create bins of data by identifying practical, meaningful ranges (bin width). In this set, with a bin width of 10, the data can be grouped as below.

Note: Changing the size of the bin (range of included data values) changes the appearance of the graph and the conclusions you may draw from it.

2. Determine the frequency of values for each data range (bin) and adjust the scale of the y axis based on the range of frequency values.

Data Range	Frequency
0-10	1
0-20	3
20-30	6
30-40	4
40-50	2
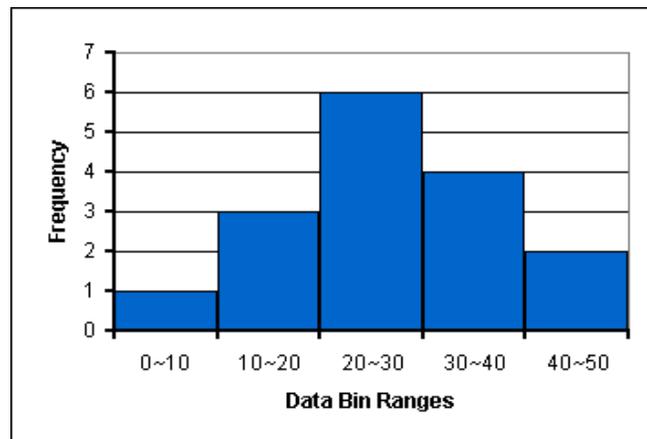
Note that the median is 25 and that there is no mode; the mean is 26.5.

A YouTube video on histograms can be viewed at https://www.youtube.com/watch?v=KCH_ZDygrm4

Christine M. Walsh-Kelly MD   9-13-2016