Histograms in Excel

A histogram is a type of chart that uses vertical columns to show the distribution of data.

Key features of histograms:

- Data points are divided into equal intervals called **bins.** These bins are represented on the x-axis.
- The chart measures how many data points fit into each bin. This is called the **frequency**, which is plotted on the y-axis of the chart.
- A histogram plot generalizes the data. As a result, we can take an educated guess where the central value is located based on the total number of data points/frequency, but we can't calculate these from the histogram itself. Before you make a histogram, it's important to calculate the descriptive statistics yourself. This will help you interpret the data and verify the results.
- The number of bins is important because different conclusions can be made based on the intervals chosen to sort the data into. It's important to choose an interval based on the type of data that you're representing.
  - If you are not sure how many bins to use, take the square root of the number of data points and round to the nearest whole number. Then adjust the number of bins from there as needed. $\# \, bins = \sqrt{number \; of \; data \; points}$

The example below shows the grade distribution for 24 students on an algebra 1 final exam.

Student	Grade
1	59
2	40
3	67
4	89
5	94
6	99
7	63
8	97
9	88
10	84
11	86
12	79
13	58
14	92
15	80
16	64
17	74
18	41
19	63
20	49
21	69
22	90
23	76
24	87

Mean	74.50 %
Median	77.50 %
Mode	63.00 %
Max.	99.00 %
Min.	40.00 %
St. Dev.	17.19 %
Range	59.00 %
Count	24

Bin	*Freq.*
40 -49	3
50 - 59	2
60 - 69	5
70 -79	3
80 - 89	6
90 -99	5

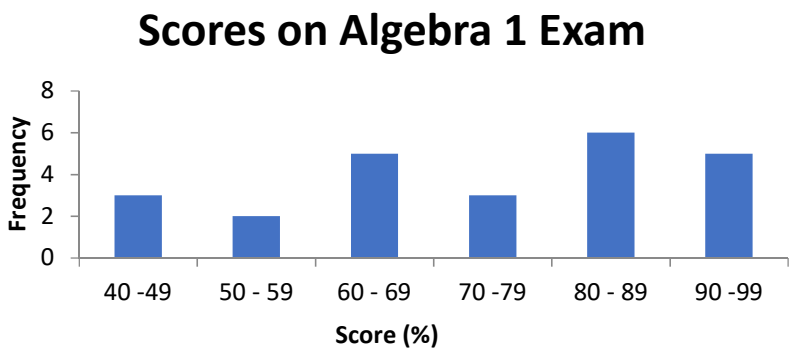
To find the number of bins: $\# \, bins = \sqrt{24} = 4.9 \approx 5$

The range of the data set is 59. Use this to find the interval size:

$interval = range \div \# \, bins = 59 \div 5 = 11.8 \approx 12$

Based on this general rule, our bins will start at the minimum score of 40% and increase by 12 until it reaches the maximum score. However, an interval size of 12 might not be the most useful in this case.

The collected data represents test scores, which are already broken down by letter grade. We may make the decision to start the bins at 40% with an interval of 10 until we reach 100%, for a total of 6 bins instead.



Histograms in Excel

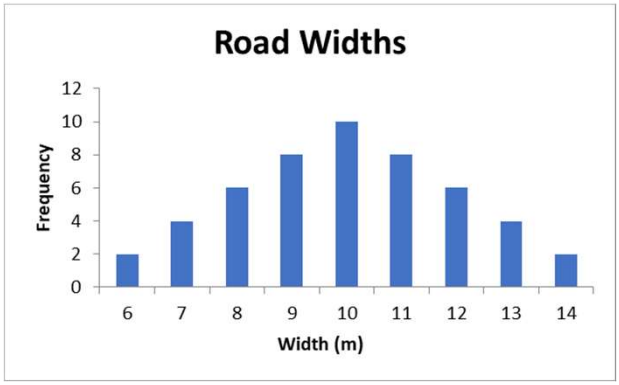
There are four different types of distribution that often show up in histograms.

Normal Distribution
Data is evenly distributed around the center value.

Also called a bell curve.

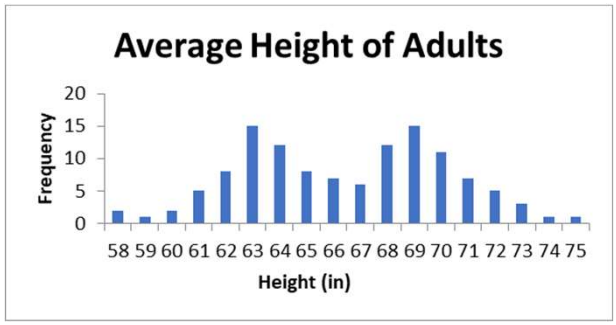
Mean, median, and mode are all the same value.

Mean	10
Median	10
Mode	10
Max.	14
Min.	6
St. Dev.	2.02
Range	8
Count	50



Bimodal Distribution
Data is distributed around two values that appear most often.

In this example, the two different peaks represent heights of males and females. It might make more sense to separate this data into two data sets.

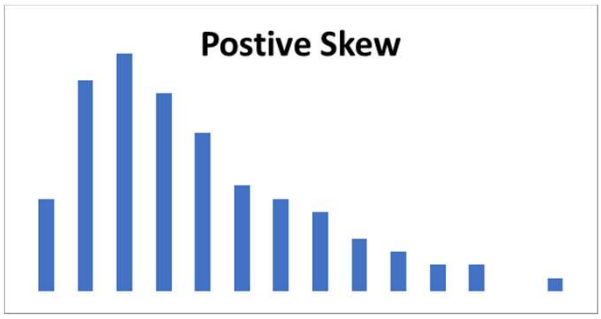


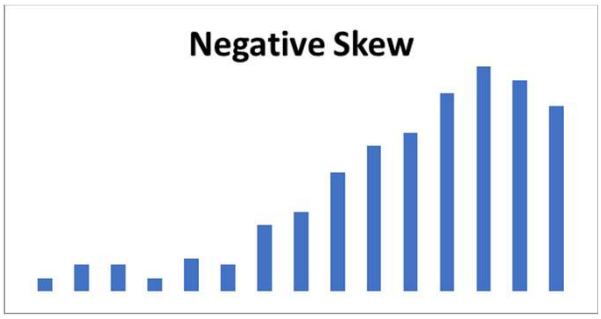
Positive Skew
Long tail of values on the right side of the histogram (also called skewed right).

Mean is greater than the median.

Negative Skew
Long tail of values on the left side of the histogram (also called skewed left).

Mean is less than the median.





Skews are often a result of a restriction on data.
(e.g. a maximum or minimum value like a maximum battery percentage of 100%)

"Skewness" or the amount of skew can be calculated using =SKEW() in Excel.

NOTE: Sometimes a distribution can look one way but the descriptive statistics would suggest its not. For example, a distribution can look normal but it's only normal if the measures of center are equal values.