**MATH221, Week 1**

Middle of Distributions and Distribution Shapes

Distributions

Distributions describe or show how the data values are related to each other. In a distribution, data values would cover the range from the smallest data value to the largest data value. For example, the ages of residents of a retirement home would have a distribution from about 65 to 105. Most of the ages would likely be at the smaller values of this distribution, meaning more of the residents were near 65 years of age with fewer closer to 105.

Middle of Distributions

Often in statistics, it is important to find the middle of the distribution or the data set. The middle can be measured with the mean, the median, or the mode. The mean is the most comment measurement in statistics of the middle. The mean is the same thing as the average. To find the mean, all the data values are added together and then the sum is divided by the number of data points in the data set. The week 1 spreadsheet calculates the mean of the data values typed into the green cells. Importantly, since all data points are included in this calculation, the mean can be noticeably impacted by one data point that is much larger or much smaller than the others.

The median is another way to consider the middle of the data set. To find the median, the data points are lined up from smallest to largest. Then looking at that list, the median is the data point in the middle of the list. If there is an odd number of data points, then the median is the data point that divides the data set into two equal parts on either side of it. If there is an even number of data points, the median is the average of the two values in the middle. The median is not affected by one data point that is much larger or much smaller than the others, as the median only considers the value in the middle of the data set.

The mode is the data value that occurs most frequently in the data set. Modes are bit unique in that there can be one mode, two modes, three modes, there can even be no modes. Consider this data set:

1 2 2 3 3 4 4 4

In this data set, the mode is 4 since it shows up three times and that is more frequently than any other data value. Consider this data set:

1 2 2 3 3 3 4 4 4

Here, the modes are 3 and 4 as those show up more than the other values and show up the same number of times. Consider this data set:

1 1 1 2 2 2 3 3 3 4 4 4

In this data set, all the values show up three times. Since ALL the data values show up the same number of times, there is no mode.

Distribution Shapes

To consider the shape of a distribution, you need to graph the data values, with the values on the x or horizontal axis and the frequency (how many times that value shows up) on the y or vertical axis. This means that the highest point in the distribution is the most frequent data value as height is frequency. This is important: Height is frequency.

Here is a normal distribution, also called a symmetrical distribution.

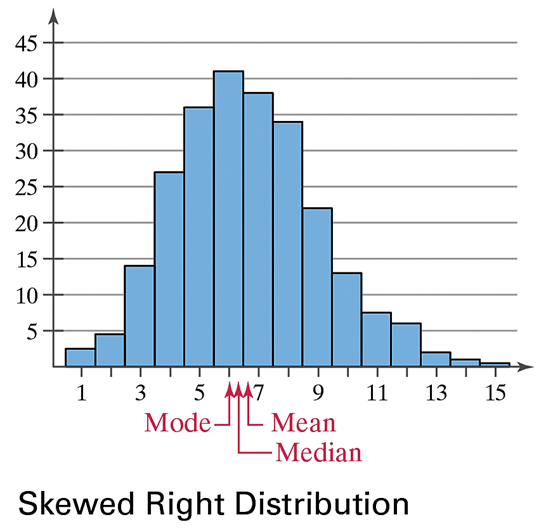
A close up of a logo

Description generated with very high confidence

|  |
| --- |
| This diagram shows a histogram. The variable x is represented on the horizontal axis. The values of x range from 1 to 15. There is a vertical bar that represents the frequency of each x value (i.e. the bar shows how many times that value of x appears in our data). The heights of the bars create a symmetric shape with the highest frequency belonging to the value of 10, which is in the middle, and the lowest frequency belonging to values of 1 and 15 which are to extreme left and right sides of the graph. |

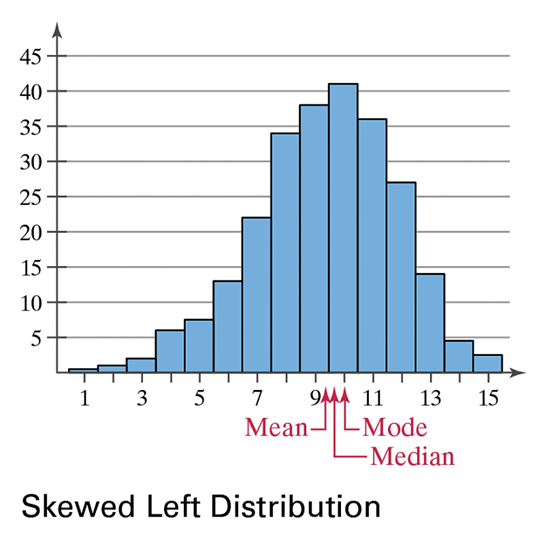
In a symmetric distribution, the mean, median, and mode all land in the middle of the data set. This is due to it symmetrical nature.

Here is a right skewed distribution.



In a right skewed distribution, there are some values that are noticeably higher or larger than the other data points. There is a “tail” of data off to the right. These larger data points pull the mean higher.

Finally, here is a left skewed distribution.



Here there are values that are noticeably smaller than the other data points. This pulls the mean smaller.