**MATH221, Week 1**

Dispersion of Data and Shape of Distributions

Dispersion of Data

Dispersion refers to how spread out the data points are from the middle of the data set. For example, if the middle of the data set was 50 and the data points went from a low of 40 up to 60, then there is not much dispersion in the data. If, however, the middle of the data set was 50 and the data points went form a low of 10 up to 90, then there would be much more dispersion in the data set.

Dispersion is measured through the range, the variance, or the standard deviation. The range is a simple calculation subtracting the smallest data value from the largest data value. For example, if this were the data set:

2 4 5 6 7 8 8 9

Then the range would be 9 minus 2 which equals a range of 7. The disadvantage of the range is that is can give a false idea of the dispersion if there is one data point with a value far different from the other. Consider this data set:

2 4 5 6 7 8 8 29

Now the range is 29 minus 2 which equals a range of 27. The data points in this data set are generally close together with little dispersion, except for this one data value. Therefore, different measures of dispersion were created.

Variance finds out how far each data point is from the mean, squares that difference, and essentially finds the average squared distance each point is from the mean. The variance then has the advantage in that it considers every data point in the data set to measure dispersion. The disadvantage is that variance is in squared units. For example, if the data set was of dollar amounts, variance would measure dispersion in squared dollars. No one knows what a squared dollar really is, so a different measure of dispersion was considered.

Standard deviation is the most common measure of dispersion in statistics. Standard deviation is the square root of the variance. This maintains the advantage of considering all of the data points, while converting the final measure of dispersion back into original units, such as dollars in the earlier example.

Two important reminders:

\* All three measures of dispersion (range, variance, and standard deviation) are calculated on the week 1 spreadsheet.

\* The variance and standard deviations calculations are different for a sample data set and for a population data set. The week 1 spreadsheet calculates both and then it is necessary to select the measurement that matches the problem.

Shape of Distributions

Below are three graphs where the values of the data are along the x axis (horizontal) and the frequency (how many times that value is in the data set) is on the y axis.

A close up of a clock

Description generated with high confidence

|  |
| --- |
| These are three histograms illustrating three different shapes of a data distribution. In each graph, the horizontal axis represents values of a variable named Data Entry. These values range from 1 to 9. There is a vertical bar above each value. The height of the bar indicates the frequency of the value in the data set.  The first graph represents a data set where all values are 5. Therefore, there is only one bar above the number 5 on the x axis. The bar shows that there are eight 5’s. It is stated on the graph that the mean is 5 and the standard deviation is 0.  The second graph represents a data set where the values of x are spread out between 3 and 7. There are bars that represent frequencies of all values. The frequencies are as follows: One three, two threes, four fives, two sixes, one seven. Combined, the bars form a symmetric bell-shaped distribution. It is stated on the graph that the mean is 5 and the standard deviation is 1.2.  The third graph represents a data set where the values of x are spread out between 2 and 8. There are bars that represent frequencies of all values. The frequencies are as follows: three twos, one three, zero fours, zero fives, zero sixes, one seven, three eights. Combined, the bars form a symmetric shape with a low middle and high bars on each side. It is stated on the graph that the mean is 5 and the standard deviation is 3.0. |

All three have a mean of 5. On the left, the values are all 5, so the standard deviation is 0. In other words, there is no variation. In the middle, the data points are more spread out, although many are still in the middle, near 5. In this case the standard deviation is 1.2. On the right, none of the data points have the value of the mean and all data points are spread out away from the mean. This data set has the largest standard deviation of the three pictures at 3.0.

To summarize, data sets with small standard deviations have little dispersion in the data values while those with large standard deviations have lots of dispersion and the data points are quite spread out.