MTH245 Unit 4 Module 1 Descriptive Statistics

Descriptive statistics are tools we can use to take a big pile of data and make sense of it.

For example, I gave a test to my online MTH111 class with the following results:

29 83 83 42 78 48 51 43 42 96 29 51 99 35 69 94 64 55 85 102 101 100 101 67 39

Looking at the numbers, what can you conclude…did the students do well? Did they do poorly? Looking at the raw data it is difficult to see what is going on.

Measures of Center are different ways to find the “center” of a data set. The most familiar measure of center is the **mean**, or **average**. This is calculated by adding the data values together and dividing by the number of data values in the set.

$$= \frac{29+\ 83+83+42+78+48+51+\cdots+\ 101+67+39}{25} = \frac{1686}{25} = 67.44$$

On this test the average score was 67.44 points.

Another measure of center is the **median** which is the middle number of the data set when it is ordered.

29 29 35 39 42 42 43 48 51 51 55 64 67 69 78 83 83 85 94 96 99 100 101 101 102

⬆

There are 25 values in this data so the 13th value–67–is the middle number in the ordered data set, which is the median. If there had been an even number of values in the data set, then the median would be calculated by taking the average of the two middle values.

Our third measure of center is called the **mode**, and is the value that occurs most frequently in the data set. Some data sets will not have a mode (such as a data set where no value occurs more than once) and other data sets may have more than one mode. Which is the case for our data set. The values 29, 42, 51, 83 and 101 all appear twice, so our data set has 5 modes and the mode is not a useful measure for this particular data set.

Why do we have 3 different measures of the center? If there is a very large or very small value in a data set it may distort what we consider the “middle” of a data set.

Example: Five students are asked what their total income was last year. They report their income as:

$0 $8000 $10,000 $12,000 $13,000

The average income is $8600, while the median income is $10,000. The student who had no income last year pulled down the average but not the median. In a data set where the average and median differ substantially the data is considered “skewed” and it is worth looking at a histogram of the data to see what is going on.

With a large data set we would not want to take the time to do these calculation by hand, once the values are entered in Excel the measures of center can easily be calculated using:

Measures of Center

Mode: Most frequent value. =Mode (number1, number2,…)

Median: Middle value. Half the data is smaller, half bigger. =Median(number1, number2,…)

Mean: Average value. = Average(number1, number2,…)

Note, in Excel, if a data set has more than one mode the Mode function will give you the smallest of the modes.

Another category of descriptive statistics are measures of spread, these include quartiles, percentiles and standard deviation.

Measure of Spread

Quartiles and **Percentiles**

First **Quartile**: 25% of the data is smaller and 75% is larger. It’s the cutoff for the bottom quarter of the data.

Second Quartile: 50% of the data is smaller and 50% is larger. It’s the same value as the median.

Third Quartile: 75% is smaller and 25% is larger. It’s the cutoff for the top quarter of the data.

The median cuts a data set in half, the quartiles will divide a data set into quarters.

In Excel we can also calculate a “zeroth quartile” which will give us the minimum value in a data set, and a “fourth quartile” which will be the maximum value of the data set.
In Excel: =Quartile(array, quart) where array is the data set and quart is which quartile we want, 0 through 4.

The five quartiles, 0th through 4th, form five “fences” that divide the data set into four groups.

29 29 35 39 42 42 43 48 51 51 55 64 67 69 78 83 83 85 94 96 99 100 101 101 102

⇧ ⇧ ⇧ ⇧ ⇧

Quartiles	
0	29
1	43
2	67
3	94
4	102

The bottom 25% of the class had scores from 29 and 43, while the top 25% of the class had scores from 94 through 102.

A variation on this is **Percentile**, which will let you fine tune what it is you are measuring. Percentiles are usually used when you have a very large data set, such as the scores of all students in the U.S. who took the SAT in a particular year. In Excel, =PERCENTILE (array, .10) would give you a value that 10% of data is smaller than and 90% is larger then. =PERCENTILE (array, .50) is the median again.

Standard Deviation is another measure of spread, if the mean is the middle of the data set, the standard deviation is the average distance between individual values in a data set and the mean. Say I gave two tests, both have an average of 70% but one class had a standard deviation of 5 while the other had a standard deviation of 20. What is happening in the two classes? In the class with the small standard deviation most of the students had a score very close to the average of 70%; there were no low scores and no high scores. For the class with the large standard deviation the class was spread out with many high scores and many low scores.

In Excel, =STDEV(number1, number2,…)

The measures of center and measures of spread are numerical summaries of a data set. We can also have graphical ways of summarizing data sets.

Graphical Summaries of Data

Going back to our data set of test scores:

29 83 83 42 78 48 51 43 42 96 29 51 99 35 69 94 64 55 85 102 101 100 101 67 39

We can make a quick tally of how many data values are in within a range of possible values:

100	2, 1, 0, 1
90	6, 9,4,
80	3, 3, 5,
70	8,
60	9, 4, 7
50	1, 1, 5,
40	2, 8,3, 2
30	5, 9
20	9, 9,

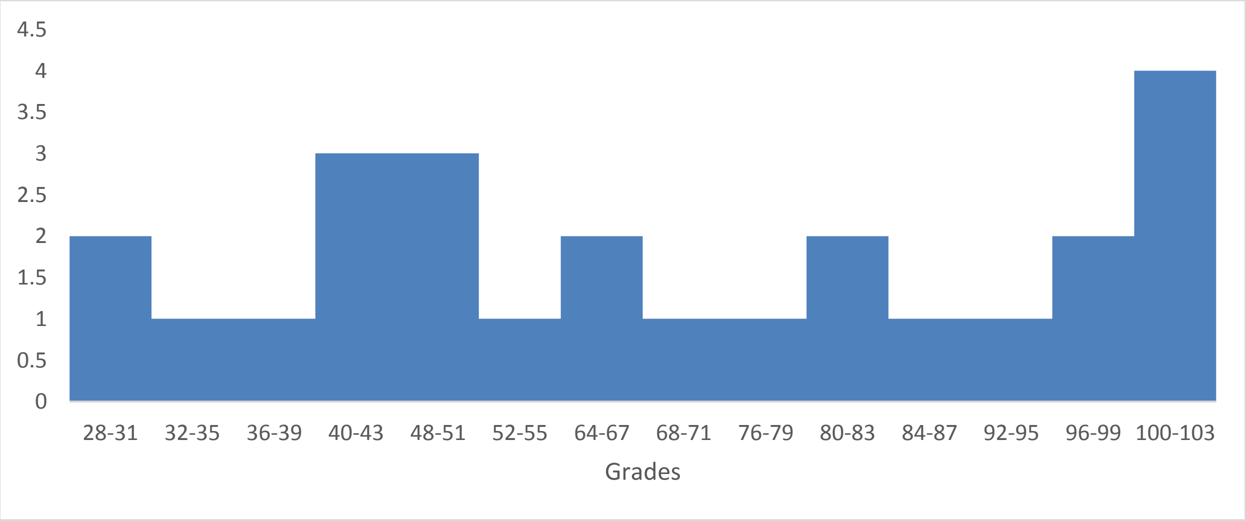
←This is quick and easy, actually a “stem and leaf” table.

The **frequency table** is a little more formal→

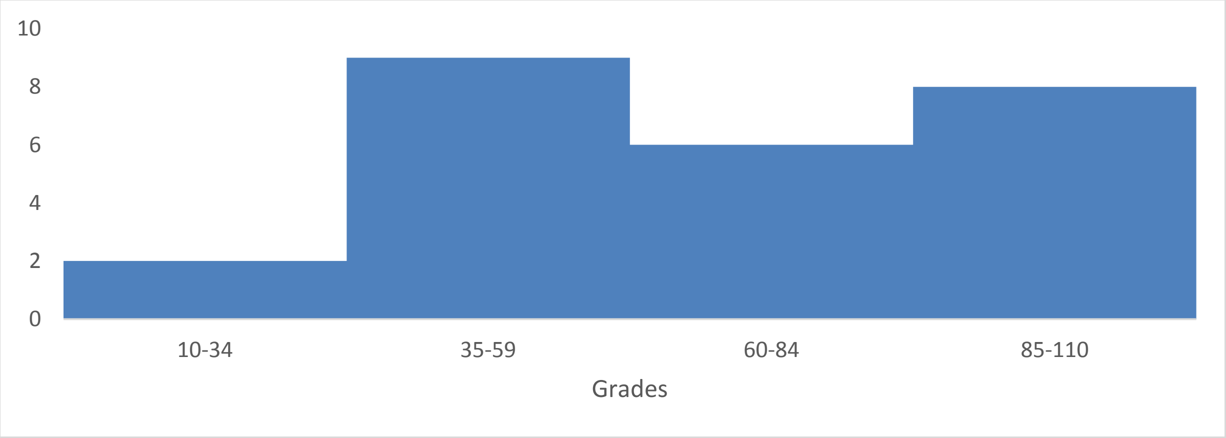
bin#	Start =>	End <	Freq	% of total
1	20	30	2	8%
2	30	40	2	8%
3	40	50	4	16%
4	50	60	3	12%
5	60	70	3	12%
6	70	80	1	4%
7	80	90	3	12%
8	90	100	3	12%
9	100	110	4	16%

The graphical version of this is a **histogram**. It is important to select appropriate intervals to use when sorting the data. Too small of intervals and you end up with little more than your original list of numbers, too large of intervals and you use all detail.

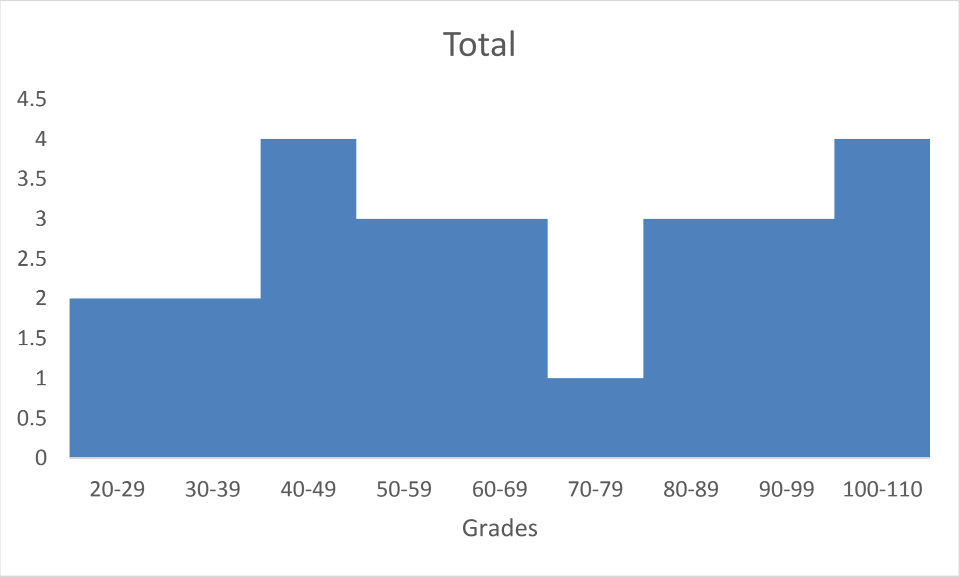
Here, the intervals are too narrow:



Here, the intervals are too broad:



With test scores, interval widths of 10 points often make sense:



(These histograms were constructed in Excel using the pivot table techniques covered in Unit 3 Module 3 Modeling Randomness)

With this picture of the data, it is easier to see that the class is really split into two different groups, there are 14 students really struggling with scores in the 20s through 60s, while there are 10 students really doing well with scores in the 80s to scores over 100.