Statistics, Mathematics 104

Statistics

Statistics is the science of learning from data.
Mathematics is a part of statistics, however there is more.

Data are numbers, but they are not just numbers,
they are numbers with a context.

Where the data come from matters.

Examples:

An example is election polling.
Pollsters take a sample of who the population will vote for in an election, and use statistics to predict who the winner will be.

There are lots of potential flaws in polling.

The people polled might not tell the truth.
The people polled might change their mind.
The sample of people polled might not be random.

Most polls are done by calling people on the phone.
Until recently pollsters would only call people on a land line.
People who own land lines tend to be older, and therefore more conservative.
So the data is not random and may be skewed toward an incorrect results.

In Philadelphia, if you track ice cream sales and crime, when ice cream sales go up, crime goes up.
This suggests some strange possibilities.

Do ice cream sales cause crime?
Does crime increase ice cream sales?

Both of these possibilities seem unlikely. In fact both occur during summer months when the weather is warmer. People buy more ice cream during warmer weather. Criminals are more active when it is warm out.

Here, the time of year is what is called a hidden variable.

Those of you who take statistics, possible Math 108, will learn a lot about polling, hidden variables, and methodologies. We are going to look at some basic mathematics behind statistics.

Descriptive Statistics

Descriptive statistics allow you to characterize data based on its properties.

There are 5 main categories of descriptive statistics.

1. Measures of Frequency

This involves count, percent, and frequency

Example, count:

Let's use the gender of students in class as an example.

The gender of students is what is called a variable.

The number of male students in class today is: _____

The number of female students class today is: _____

Example, percent:

What is the percent of students in class today who are female:_____

Example, frequency:

What is the frequency of blue eyes in the class today:_____

2. Measures of Central Tendency

Often when you have a data set, you are interested in which data is in the middle.

There are three common measures we use, mean (or average) median, and mode.

Let's take a set of data as follows.

A class takes an exam, The scores are

65, 70, 72, 74, 79, 83, 85, 95 and 100

Average or Mean

We find the average by adding the scores together and dividing by the number of scores.

$$\frac{65+70+72+74+79+83+85+95+100}{9} = 80.3\overline{3}$$

This can often be a good measure of where the middle of the data is.
However in some cases this is not a good measure.
Consider the average income in the US

In 2016 the mean income was $46,550. However most people earned less than this.
This is because the wealthy earn so much more.

Median

The median of a dataset is the number that is in the middle. For our dataset that is 79. For the income in the US in 2016 it was $31,099, much lower than the mean. In this case it seems that the median is a better measure of the middle since half the people earn more and half earn less.

Note: with a small dataset if you have an even number of data points, there is no middle, so by convention we take the average of the two middle numbers.

Example:

5 10 25 30

The average is (5+10+25+40)/4 = 20
The median is (10+25)/2 = 17.5

Note that as a measure of center, Median is more resistant to outliers. Let's say we have a group of people of various ages.

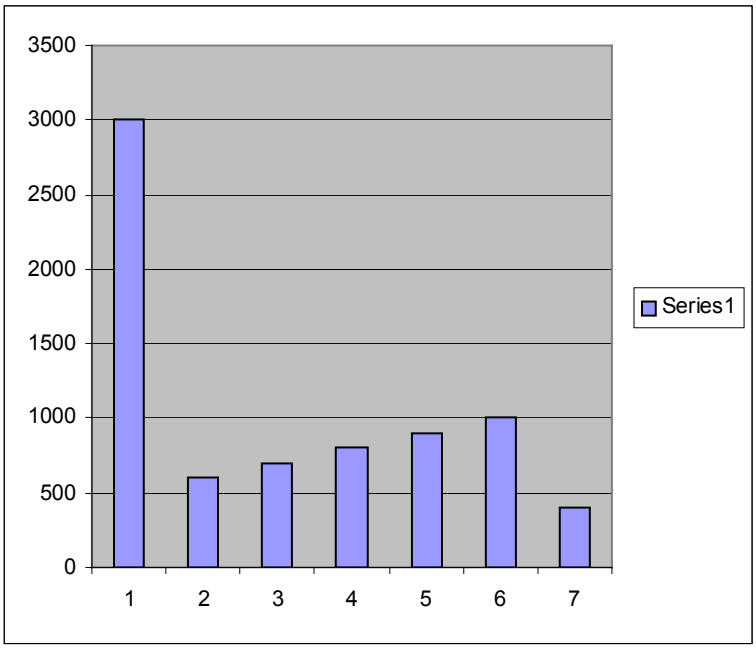
1, 2, 7, 12, 20, 21, 21, 21, 21, 24, 25, 25, 27, 30, 60.

The median and the average here is 21.

If we add a person aged 100, the average changes to 26, but the media is still 21.

The Mode

The mode is the value that appears most often. For small samples it does not have much meaning since it can be a random artifact of the data. In a large dataset, if the mode is different from the mean and average what does it tell you about the data?



3. Measures of Dispersion or Variation

The measure of dispersion or variation are the range, variance and standard deviation.

The range is just what values are possible.

Example:

The range of ages for all people is [0, 122] years.

The range of students at USF is probably about [16, 50]

Variance and Standard deviation are values you can calculate on a data set that describe how the data is spread out from the average.

For a set of data points, the standard deviation is calculated as

$$s = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)}{N-1}}$$

Where $\bar{x} = \frac{\sum_{i=1}^{N} x_n}{N}$, the average

The variance is just $\sigma = s^2 = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)}{N-1}$

The numbers tell how close to the average the data lies.

If the standard deviation is large, the data is spread out.
If the standard deviation is small, the data is all close to the average.

The standard deviation is a distance from the center of the data.
For a large data set, about 67% of the data is within 1 standard deviation of the center.
94% will be within 2 standard deviations.

The standard deviation is a complicated value to calculate, but most graphing calculators will have let you enter data and calculate it for you.

4. Measures of Position

Measures of position describe how data falls in relationship to itself.
Examples are percentile ranks, or quartile ranks.

Example:

For the SAT if you score

Score	Percentile
780-800	99
760-780	98
740-760	97
720-740	95
700-720	94

So if you scored 730, 5% of the people who took the exam scored higher than you, and 95% scored lower.

Examples:

Using the following data and a calculator, find the media, mean, mode and standard deviation of each set.

On a Ti-83/84 follow the following procedue

1) Press [STAT]
2) On EDIT press ENTER
3) Enter the data
4) Press [STAT}
5) Cursor over to CALC and press enter on "1-Var Stats"

Note the meaning of the symbols
$\bar{x}$ is the average
$\sum x$ is the sum of the numbers
$\sum x^2$ is the sum of the squares
Sx is the Standard Deviation
The numbers
minX, Q1, Med, Q3 and maxX are the

minimum, maxiumum, median and where the value at the first quartile and 3 quartile. These give a standard view of the distribution of the data

Let's look at the values for these datasets

1. 1 1 2 4 5 6 9 12 13 14 16 17 17

2. 1 7 7 7 8 8 9 10 10 11 11 11 17

3. 1 2 3 3 3 3 3 3 3 3 3 4 7 9 12

4. 1 5 7 12 12 12 18 79 100