Math 107
Statistical measures of center and spread

Suppose I come into the room the day after we take a test and of course what is it that you want to know? You want to know how you did on the test. Normally the way I will communicate this to you is to give the average. But the average (like much of statistics) can be a tricky idea!

Definition: The mean (or arithmetic mean or average) of a set of data points $x_1, x_2, x_3, ... x_n$ is given by the formula:

$$\bar{x} = \frac{\text{sum of the data points}}{\text{number of data points}} = \frac{x_1 + x_2 + ... + x_n}{n}$$

Calculating the average is easy – usually easy enough to warrant doing it by hand:

Example: The high temperatures for Spokane last week

47	44	58	55	62	40	52

Find the average

Example: Salaries of communication majors for a certain year one year after graduations

23,500	31,000	28,500	19,750	23,500	3,500,000

Find the average

As we can see from this example – the mean is sometimes not the best measure to use for an average – certainly it is not representative of the real “average salary” of communication majors. One reason the average (or mean) is not always used is that it is sensitive to “outliers” which are data points which are much higher or much lower than almost all the other values. Another measure is called the median

The median is the middle value in a sorted data set. To calculate the median

1. Sort the data
2. If there are an odd number of data points, take the value in the middle
3. If there are an even number of data points, take the average of the two middle values

Example: Calculate the median for the two data sets given above

Another “average” is called the mode – this is the value that occurs most frequently

- If no value occurs more frequently than others (they all occur once) there is NO mode
- If there is a tie for the mode – if there are two data points – it is bimodal, if more than two we say multimodal

Example: Find the mode of the examples given above

Example: Find the mean median and mode of the following data (this data represents the ages of students in a math in society course)

23	18	33	28	24	20
19	22	25	43	21	30
22	25	19	19	20	21
27	41	20	22	23	23
19	29	35	31	29	25

This is hard to do by hand – how can our TI-83/84 or Excel help us?

Measures of Variation

Calculate the mean, median and mode for each of the following datasets

Dataset A:

50	60	70	70	80	90

Dataset B:

68	69	70	70	71	72

These datasets have the same measures of center, yet are quite different - we need a measure of spread - We will use many measures of spread - one measure is the range.

Range = high value - low value

Calculate the range for each of the data sets above

Range is great but not always adequate - consider these datasets

DATASET #1

2	15	15	15	15	15

DATASET #2

21	25	27	29	31	34

These datasets have the same range, BUT they are clearly distributed differently. A better measure of spread is the standard deviation

$$\sigma^2 = \frac{\sum (x - \overline{x})^2}{n}$$ variance

$$\sigma = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$ standard deviation

To calculate the standard deviation - manually - do the following

1) Make a chart place the original data in the first column
2) Subtract the mean from the original data and place this number in the second column
3) Square these results place answers in third column
4) Add the third column and divide by n - this gives the variance
5) Take the square root - this is the standard deviation

Data	Data-mean	(data-mean) squared
50		
60		
70		
70		
80		
90		

Sum of last column

Divided by n

Square root of this result

Now we will do the second dataset on your calculator

While standard deviation gives us a measure of spread - sometimes we want a measure of dispersion that does not depend on all data values. Consider the following salaries for mathematics graduates including the professional basketball player Joe "Hoops" Jackson

30000	24500	22700	19850	35675	21000	2575000	15800

We sometimes want to use the five number summary - which is
-The minimum value (min)
- First Quartile (Q_1)
- Median
-Third Quartile (Q_3)
- Maximum
Finding the min and max are easy and you already know how to find the median - to find (Q_1) and (Q_3) first sort the data and Q_1 is the median of the first half of the data and Q_3 is the median of the upper half of the data

15800	19850	21000	22700	24500	30000	35675	2575000

Min = 15800
Q1 = (19850+21000)/2= 20425
Med = 23600
Q3=32837.5
Max = 2575000

You might compare with the standard deviation for this dataset -

If we want a visualization of the five number summary - we draw a boxplot. Consider this data set

50	55	60	63	70	70	70	75	82	89	90

Calculate the five number summary

Draw a boxplot below

All of these can be done on your Ti-83/84 or Excel. As I will illustrate in the next example

Find the standard deviation, variance, five number summary and draw a boxplot for the following dataset

45	78	90	88	85	75	64	77	85	93
80	78	65	96	80	74	60	53	80	92
75	65	80	82	75	71	61	52	47	94