Math 214 – Introductory Statistics **Summer 2008**
6-9-08 Class Notes

Sections 3.2, 3.3
3.2: 1-21 odd
3.3: 7-13, 35-39 odd

Measures of Central Tendency

Notation: Let N be the size of the population, n the size of the sample, and x a data value.

Definition: The ***mean*** (or arithmetic mean) is the sum of all data values divided by the number of data values. This is denoted by μ for the whole population and $\bar{x}$ for a sample. So in symbols,

$$\mu = \frac{\sum x}{N} \text{ and } \bar{x} = \frac{\sum x}{n}.$$

Definition: The ***median*** is the "middle" of the data set (when the data is arranged in increasing order). This is usually denoted by MD.

Definition: The ***mode*** is the most common data value. If there is only one such value, the data set is said to be ***unimodal***. If there are two data values that occur most often, the data set is said to be ***bimodal***. If there are more than two data values occurring most often, the data set is ***multimodal***.

Definition: The ***midrange*** is the mean of the highest and lowest data values. This is denoted by $MR = \frac{x_{\max} + x_{\min}}{2}$.

Example: Find the mean of the data set: 8, 7, 5, 3, 2, 3, 6, 9, 11.

$$\bar{x} = \frac{8+7+5+3+2+3+6+9+11}{9} = \frac{54}{9} = 6.$$

Example: Find the median of the data set: 8, 7, 5, 3, 2, 3, 6, 9, 11.

~~2, 3, 3, 5~~, 6, ~~7, 8, 9, 11~~

↑

middle value

Note that for this data set, the mean and median were the same. This does not happen in general.

Example: Find the mean and median of the data set: 20, 3, 4, 2, 4, 9, 13.

$$\bar{x} = \frac{20+3+4+2+4+9+13}{7} = \frac{55}{7} = 7.9$$

~~2, 3, 4~~, 4, ~~9, 13, 20~~

↑

middle value

Example: Find the median of the data set: 2, 2, 2, 5, 7, 2, 3, 6.

~~2, 2, 2~~, 2, 3, ~~5, 6, 7~~

↑

middle values

Thus the median is their mean $\frac{2+3}{2} = 2.5$.

Example: Find the mode for each data set:

(a) 1, 3, 3, 3, 5, 2, 7
(b) 1, 5, 7, 9, 12, 13, 15
(c) 11, 5, 5, 11, 5, 11, 7, 5, 11

(a) 3 is the most common value. So the mode is 3.
(b) There is no value that occurs more than once. So no mode.
(c) 5 and 11 both occur the most times, so the modes are 5 and 11. This set is bimodal.

Example: Find the mean and the midrange for the data set: 2, 3, 6, 7, 7, 8, 9, 9, 10.

$$\bar{x} = \frac{2+3+6+7+7+8+9+9+10}{9} = 6.8 \qquad MR = \frac{10+2}{2} = 6$$

The measure of central tendency that gives the most accurate representation of the data depends on the particular data set.

Example: Find the mean, median, and mode for the data set: 2, 2, 2, 3, 4, 5, 1000.

$$\bar{x} = \frac{2+2+2+3+4+5+1000}{7} = 145.4$$

~~2, 2, 2~~, 3, ~~4, 5, 1000~~

$MD = 3$

Mode=2

Other Types of Means

Sometimes the different data values are not equally weighted. For example, suppose you receive the following grades:

Course	Credit Hours	Grade	Grade Points
A	2	C	2
B	3	A	4
C	3	A	4
D	5	D	1
E	1	A	4
	14		

An arithmetic mean of your grade points would yield a "grade point average" of 3.00. But since the courses were not equally weighted, your actual grade point average is probably lower than that (since you received a D in that 5-hour course). This is an example of a *weighted mean*.

Definition: Let $x_1, x_2, \ldots, x_n$ be the data values and let $w_1, w_2, \ldots, w_n$ be the respective weights of each data value. The ***weighted mean*** is $\frac{\sum wx}{\sum w}$.

In the grade point example, your actual grade point average is:

$$\frac{2 \cdot 2 + 3 \cdot 4 + 3 \cdot 4 + 5 \cdot 1 + 1 \cdot 4}{14} = \frac{37}{14} = 2.64$$

In our first set of notes, we spent time trying to summarize data with a frequency distribution. How do we approximate the mean after we've done this? Consider our first frequency distribution:

Class (Hours per week)	Frequency (Number of students)
1-4	7
5-8	12
9-12	15
13-16	9
17-20	7
	50

To find the mean, we find the weighted mean of the midpoints of the classes (where the frequencies are the weights). So let's add a column for the midpoints.

Class	Frequency	Midpoint
1-4	7	2.5
5-8	12	6.5
9-12	15	10.5
13-16	9	14.5
17-20	7	18.5
	50	

An approximation for the mean of the data values is then

$$\frac{\sum wx}{\sum w} = \frac{7 \cdot 2.5 + 12 \cdot 6.5 + 15 \cdot 10.5 + 9 \cdot 14.5 + 7 \cdot 18.5}{50} = \frac{513}{50} = 10.3 .$$

Measures of Variance

Knowing an "average" for a set of data is very important, but it also useful to know how much the data varies from that average. Consider these two sets of data:

Group A		Group B	
65	73	42	77
66	74	54	77
67	77	58	85
68	77	62	93
71	77	67	100

For each group, the mean is 71.5. In fact they also have the same median (72), mode (77), and midrange (71). All measure of central tendency are equal.

But clearly Group B is more widespread and scattered than Group A. We now present three measures of variation that will quantify this characteristic, the first of which we have already seen.

Definition: The ***range*** of a set of data is the difference between the lowest data value and the highest data value. In symbols, $R = x_{\max} - x_{\min}$.

In our example above, the range of Group A was $R_A = 77 - 65 = 12$ and the range of Group B was $R_B = 100 - 42 = 58$. Clearly B is more widespread. But the range can be a misleading attribute since it only depends on the highest and lowest data values (and ignores all the rest).

Definition: The ***variance*** of a data set is a measure of how dispersed the data is. It is denoted by σ^2 for a population and s^2 for a sample. The formulas are as follows:

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N}, \text{ and } s^2 = \frac{\sum (x-\bar{x})^2}{n-1}.$$

Let's break those formulas down and see how to compute the sample variance.

(1) First find the sample mean $\bar{x}$.

(2) Subtract $\bar{x}$ from each data value to find $(x-\bar{x})$. This quantity is called *the deviation from the mean*.

(3) Square each deviation from the mean to get $(x-\bar{x})^2$.

(4) Add all these squared deviations from the mean to get $\sum (x-\bar{x})^2$.

(5) Divide by $n-1$.

There is a shortcut formula that is a little easier to compute. USE THIS FORMULA INSTEAD OF THE ONE ABOVE.

$$s^2 = \frac{n\sum\left(x^2\right) - \left(\sum x\right)^2}{n(n-1)}$$

Example: Find the sample variance for the data: 2, 3, 5, 5, 8.

x	x^2
2	4
3	9
5	25
5	25
8	64
23	127

So, $s^2 = \dfrac{5(127) - (23)^2}{5(4)} = \dfrac{635 - 529}{20} = 5.3$.

Definition: The ***standard deviation*** of a data set is the square root of the variance. It is (obviously) denoted by σ for a population and s for a sample.

In our worked example, the standard deviation is $\sqrt{5.3} = 2.30$.

As noted above, the larger the standard deviation, the more spread out the data is. In fact, we can state approximately how much of the data lies within one standard deviation of the mean, or two standard deviations, or three…

Theorem (Chebyshev): The proportion of data values that will fall within k standard deviations from the mean will be at least $1 - 1/k^2$ (where k is a number greater than 1).

Let $k = 2$. Then Chebyshev tells us that at least $1 - 1/4 = 3/4$ (or 75%) of the data value will fall within two standard deviations of the mean.

Example: A sample of hourly wages has a mean of \$6.45 and a standard deviation of \$0.32. Find the range in which at least 88.89% of the data lies.

According to Chebyshev, $.8889 = 1 - 1/k^2$ would give us a k value of 3. So we want 3 standard deviations from the mean. If we add and subtract the standard deviation to the mean three times, we'll get the required range: \$5.49-7.41.

Example: The average cost of a certain type of grass seed is \$4.00 a box with a standard deviation of \$0.10. Find the minimum percentage of data values that will lie in the range of \$3.82 to \$4.18.

The given range represents how many standard deviations? The high and low are both \$0.18 from the mean, the standard deviation is \$0.10, so we are looking at a range of 1.8 standard deviations. So $k = 1.8$. Substituting that into Chebyshev's Theorem, we get $1 - 1/(1.8)^2 = .6914$. So 69.14%.