II. SUMMARIZING DATA

Location Measures (Measures of Central Tendency):

A *location measure* or *measure of central tendency* for a variable is a single value or number that is taken as representing all the values of the variable. Different location measures are appropriate for different types of data.

A. The Mean

The mean is an appropriate location measure for interval or ratio variables. Suppose there are N individuals in the population and x denotes an interval or ratio variable. Let the value of x for the i^{th} individual be denoted by x_{i}. The mean of x is the number

$$\bar{x} = \frac{1}{N}(x_1 + x_2 + \cdots + x_N)\,.$$

Sometimes the Greek letter μ (mu) is used as a symbol for the mean of a population variable.

Some of the values of x in the population might be repeated. That is, it might happen that several individuals have the same value of x. Suppose there are M distinct values of x and we designate these values by $x'_1, x'_2, \cdots, x'_M$. Denote the number of times the value x'_j occurs by $n_{j.}$ This is called the *frequency* of the value x'_j .

Using the frequencies, the mean of x can be calculated another way.

$$\bar{x} = \frac{1}{N}(n_1 x'_1 + n_2 x'_2 + \cdots n_M x'_M)$$

Exercise: For each day of the past two weeks, the weather person's prediction of the daily high temperature was subtracted from the actual high temperature to obtain his or her error of prediction. The distinct values of the difference x are given in the table below. The bottom row of the table gives the frequencies of those values. Find the mean of x. Hint: What is N?

x'_j	-2	1	3	4	6

n_j	2	1	3	5	3

Exercise: The sum of all the frequencies is equal to *N*, the number of individuals. Why is this true?

A third way of calculating the mean involves the *relative frequencies*. The relative frequency of the value x'_j is defined as

$$f_j = \frac{n_j}{N}.$$

This is just the fraction of all the individuals who have an *x*-value of x'_j. The third expression for the mean is

$$\bar{x} = f_1 x'_1 + f_2 x'_2 + \cdots + f_M x'_M .$$

All three of these formulas give the same answer. Which one you use is strictly a matter of convenience.

Exercise: In the table above, calculate the relative frequencies and use them to recalculate the mean.

Exercise: The sum of all the relative frequencies is 1. Why is this true?

B. The Median

Roughly speaking, the median of a variable is the middle value when its values are arranged in order. Let *x* be an interval or ratio variable. A number *m* is a median of *x* if at least half the individuals *i* in the population have $x_i \geq m$ and at least half of them have $x_i \leq m$.

Examples 2.1:

1. Suppose the values of *x* arranged in order are –2.0, 1.5, 2.2, 3.1, and 5.7 and that there are no repetitions. The median is 2.2 because at least half of the 5 individuals in this population have $x_i \geq 2.2$ and at least half of them have $x_i \leq 2.2$.

2. Suppose the values of variable *y* are –2.0, 1.5, 3.1, 3.1, and 3.1. Notice that the value 3.1 is repeated. The median of *y* is 3.1 because at least half the individuals have $y_i \leq 3.1$ and at least half of them have $y_i \geq 3.1$.

3. Suppose the values of the variable *w* are –2.0, 1.5, 3.1, 5.7, 5.9, and 7.1. The numbers 3.1 and 5.7 are both medians of *w*. So is any number between 3.1 and 5.7.

In each of the first two examples above, there was only one value that qualified as a median according to our definition. This will always be the case when the number of individuals is odd. If the number of individuals is even, there might be more than one median. For numerical data, there is always a smallest median m_1 and a largest median m_2. Any number between m_1 and m_2 is also a median. If it is important that a single number *m* be chosen as the median, it is customary to take the midpoint between these two numbers:

$$m = \frac{m_1 + m_2}{2}.$$

If we use this convention, the median in Example 2.1-3 is $m = \frac{3.1 + 5.7}{2} = 4.4.$

Exercise: In Example 3, change the value 7.1 to 71. What is the effect of this change on the mean and the median? This shows that the median is less sensitive than the mean to changes in a few values.

The median is a useful location measure for interval and ratio data. It can sometimes be useful for ordered categories. This is illustrated by the following example.

Example 2.2:

The letter grades for a class of 100 students are tabulated below. The numbers in the second row of the table are the frequencies.

A	A-	B+	B	B-	C+	C	C-	D+	D	D-	F
8	5	10	18	18	15	14	6	4	1	1	0

The median grade is B- because at least half the students had a B- or better and at least half had a B- or below.

C. The Mode

For nominal variables or ordered categorical variables, a *mode* is a value that has the greatest frequency. There may be more than one mode. The mode is not especially useful for interval and ratio variables because their values are not often repeated if they are measured with precision. For strictly nominal variables, the mode is the only one of these location measures that is important.

Example 2.3:

In the grade table above, the modes are B and B-.

Exercise: Examine some of the data sets provided. Identify some variables of interval or ratio type. Use the spreadsheet program or a calculator to calculate the means and medians. If the mean and median seem to differ significantly, see if you can discover why. Is it because of only a few data values? Add one or two extremely large or extremely small values to a variable. What is the effect on the mean and the median?

Cumulative Frequencies and Percentiles:

Let x be a numeric variable of ordinal, interval, or ratio type. As before, let $x'_1, x'_2, \cdots, x'_M$ be the distinct values of x, let $n_1, n_2, \cdots, n_M$ be their frequencies, and let $f_1, f_2, \cdots, f_M$ be their relative frequencies. Assume that the distinct values have been arranged in order: $x'_1 < x'_2 < \cdots < x'_M$. The *cumulative frequencies* N_j and *cumulative relative frequencies* F_j are defined as follows.

$$\begin{array}{ll}
N_1 = n_1 & F_1 = f_1 \\
N_2 = n_1 + n_2 & F_2 = f_1 + f_2 \\
N_3 = n_1 + n_2 + n_3 & F_3 = f_1 + f_2 + f_3 \\
. & . \\
. & . \\
. & . \\
N_M = n_1 + n_2 + \cdots + n_M & F_M = f_1 + f_2 + \cdots + f_M
\end{array}$$

Exercise: In the last two equations just above, N_M is actually equal to N, the total number of individuals, and $F_M = 1$. Why?

Observe that $N_2 = N_1 + n_2$, $N_3 = N_2 + n_3$, $N_4 = N_3 + n_4$, and so on.
Similarly, $F_2 = F_1 + f_2$, etc.

Example 2.4:

The table below shows cumulative frequencies and relative frequencies for the weather person's daily prediction errors over a two-week period.

x'_j	-2	1	3	4	6
n_j	2	1	3	5	3
N_j	2	3	6	11	14
f_j	.1429	.0714	.2143	.3571	.2143
F_j	.1429	.2143	.4286	.7857	1.000

Exercise: From the table above, what fraction of the data is less than 1? What fraction is greater than 3? What fraction is greater than or equal to 3?

Let x be an interval or ratio variable and let p be a number between 0 and 100. A number a is a p^{th} *percentile* of x if at least p% of the values of x are less than or equal to a and at least (100-p) % of the values of x are greater than or equal to a. The 25th percentile is called the first quartile of x and the 75th percentile is the third quartile of x. The 50th percentile is the second quartile or median.

As with the definition of the median, there is some ambiguity in this definition. This is usually resolved by taking the smallest number that qualifies as a p^{th} percentile, except for the median, where we usually take the middle number that qualifies. Do not be alarmed if you see the ambiguity resolved by some other convention. Exactly how it is resolved is less important than that it be done in a consistent way.

Example 2.5:

For the weather person's errors, the 25th percentile is 3. The 50th percentile and third quartile are both 4.

Exercise: With a spreadsheet program or calculator, find the quartiles and the median of some of the variables in the data sets provided. Find the 10th and 90th percentiles.

Measures of Variability

Statisticians are not only interested in describing the values of a variable by a single measure of location. They also want to describe how much the values of the variable are dispersed about that location.

A. The Standard Deviation

To define the standard deviation we must first define another quantity called the variance. The standard deviation and variance are meaningful for interval or ratio variables. Let the mean of a population variable *x* be denoted by μ. The *variance* of *x* is the quantity

$$\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_N-\mu)^2}{N}.$$

The *standard deviation* is defined to be the square root of the variance: $\sigma = \sqrt{\sigma^2}$. The standard deviation is expressed in the same units as the values of *x*. It represents a “typical” difference between values of *x* and their mean, without regard for sign.

When the values of x come from a sample rather than the entire population, the definition of the variance and standard deviation are slightly different. If *n* is the number of individuals in the sample and $\bar{x}$ is the mean of the values of x for the sample, then the *sample variance* is defined as

$$s^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n-1}$$

For the sample variance, we divide by $n-1$ rather than *n*. This is so that the theoretical expected value of the random variable s^2 will be equal to the true variance σ^2. The technical terminology we use is that s^2 is an *unbiased estimator* of the parameter σ^2. This would not be true if we were to divide by *n* in the definition of s^2. However, if *n* is reasonably large it makes very little numerical difference whether we divide by *n* or *n* - 1. The *sample standard deviation* is the square root of the sample variance: $s = \sqrt{s^2}$. We refer to the variance and standard deviation using all the individuals in the population as the *population variance* σ^2 and the *population standard deviation* σ.

If the variable has repeated values, there are alternative formulas for the population variance:

$$\sigma^2 = \frac{n_1(x_1' - \mu)^2 + n_2(x_2' - \mu)^2 + \cdots + n_M(x_M' - \mu)^2}{N}$$

and

$$\sigma^2 = f_1(x_1' - \mu)^2 + f_2(x_2' - \mu)^2 + \cdots + f_M(x_M' - \mu)^2 .$$

Example 2.6:

You have already found the mean of the variable tabulated below.

x_j'	-2	1	3	4	6
n_j	2	1	3	5	3
N_j	2	3	6	11	14
f_j	.1429	.0714	.2143	.3571	.2143
F_j	.1429	.2143	.4286	.7857	1.000

The mean μ is 3.143. Using the first formula above, we have

$$\sigma^2 = \frac{2(-2-3.143)^2 + (1-3.143)^2 + 3(3-3.143)^2 + 5(4-3.143)^2 + 3(6-3.143)^2}{14}$$

$$\sigma^2 = 6.1224$$

$$\sigma = \sqrt{6.1224} = 2.474$$

This data is more likely to be considered a sample of 14 measurements from a larger population. Thus, we would calculate the sample variance and standard deviation and get slightly different answers.

$$s^2 = \frac{2(-2-3.143)^2 + (1-3.143)^2 + 3(3-3.143)^2 + 5(4-3.143)^2 + 3(6-3.143)^2}{13}$$

$$s^2 = 6.5934$$

$$s = \sqrt{6.5934} = 2.568$$

Exercise: Use a calculator or spreadsheet program to calculate the population variance and standard deviation of some of the variables in the data sets provided.

The interpretation of the standard deviation is straightforward. The greater the standard deviation, the more widely dispersed the values of *x* are about their mean. There is a general rule derived from a famous inequality called Chebyshev's inequality. We will not state the inequality completely but it implies the following: At least 3/4 of the values of a variable are within 2 standard deviations of its mean. At least 8/9 of the values are within 3 standard deviations of the mean. At least 15/16 of the values are within 4 standard deviations of the mean. These statements are true for any distribution of values whatever. Later, we will discuss normal distributions. For normally distributed variables, about 68% of the values are within 1 standard deviation of the mean, about 95% of the values are within 2 standard deviations of the mean, and almost all of the values are within 3 standard deviations of the mean.

For data that may be subject to large errors of recording or measurement, the standard deviation is possibly misleading as a measure of dispersion. Only a few extremely large or extremely small values of *x* can cause the standard deviation to be large, even though most of the values of *x* are close together. The next measure of variability does not have this problem.

B. The Interquartile Range

The *interquartile range* for an interval or ratio variable *x* is the difference between its 3rd and 1st quartiles. It is given by

$$IQR = Q_3 - Q_1,$$

where Q_1 and Q_3 are, respectively, the first and third quartiles. Since the interquartile range is determined by the middle 50% of the data, it is not affected by a few extreme values. There is no fixed relationship between the interquartile range and the standard deviation. However, for data that is normally distributed (a property that will be introduced later), the *IQR* is about 1.35 times the standard deviation. A comparison of the two measures may indicate that there are extremely large or extremely small values of *x*. These are called *outliers*. If the *IQR* is less than about half the standard deviation and the number of individuals is reasonably large, there may be outliers. A statistician might want to investigate any outlier individuals to see what makes them unusual.

Exercise: In the weather forecast data, what is the interquartile range? Compare it to the standard deviation.

Based on your response to the preceding exercise, you might think of outliers. Indeed, the two errors with an x-value of -2 do stand apart from the rest. However, this is a small data set and the rule of thumb given above should not be taken too seriously.

Exercise: With a spreadsheet program or calculator find the standard deviations and interquartile ranges of some variables in the data provided. Compare them. Is there any reason to suspect extreme outliers? Add one or two extremely large and small values and observe the effect on the standard deviation and the interquartile range.

C. The Range

The *range* of an interval or ratio variable is the difference between its largest and smallest values. The range is a relatively crude measure of variability because the range of values from a sample is unreliable as an indication of the range of values for the population. However, it is sometimes desirable for other reasons to know the range of the data.