

Unit 6

More Numerical Summaries

Objectives:

- To obtain and interpret several measures of center for the distribution of a quantitative variable
- To obtain and interpret several measures of dispersion for the distribution of a quantitative variable
- To obtain and interpret a measure of skewness for the distribution of a quantitative variable

We have defined several numerical summaries for the purpose of describing the center for the distribution of a quantitative variable and the amount of dispersion in the distribution of a quantitative variable. The mean (\bar{x}) and median (Q_2) are measures of center for the distribution of a quantitative variable, and the range and interquartile range (*IQR*) are measures of dispersion for the distribution of a quantitative variable. We shall now introduce one additional measure of dispersion, based on the amounts by which each of the observations in a data set differ from the mean.

A *deviation from the mean* is defined to be an observation minus the mean of the observations. Recall that we represent a list of n numerical values by x_1, x_2, \dots, x_n . We would then represent the deviations by

$$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}.$$

For the sake of brevity, subscripts can be omitted when there is no confusion, and deviations are simply represented by $x - \bar{x}$. A deviation will be positive when the observation is greater than the mean and negative when the observation is less than the mean.

It should certainly come as no surprise that the sum of the deviations will always be zero; in other words, the negative deviations and the positive deviations will balance each other out, so to speak. Using our summation notation, we would say that $\Sigma(x - \bar{x}) = 0$ is always true. To illustrate, we consider once again the variable "Number of Children" in the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1. Focusing only on the $n = 10$ individuals with an urban residence, we let

$$x_1 = 2, x_2 = 3, x_3 = 3, x_4 = 1, x_5 = 0, x_6 = 2, x_7 = 2, x_8 = 4, x_9 = 3, x_{10} = 1.$$

Previously, we have found that $\bar{x}_U = 2.1$. You should be able to verify easily that the deviations are

$$-0.1, +0.9, +0.9, -1.1, -2.1, -0.1, -0.1, +1.9, +0.9, -1.1.$$

It is now an easy matter to check that these deviations indeed sum to zero.

Since deviations from the mean always sum to zero, we must ignore the signs of the deviations in order to obtain a measure of dispersion. The most natural and intuitively pleasing measure of dispersion would be the average of the absolute values of the deviations, which can be called the *absolute mean deviation*.

Unfortunately, life is often not simple, and for somewhat complex mathematical reasons which we do not state here, the absolute mean deviation is almost never used; instead, another measure of dispersion which we shall now introduce is used.

Much of classical, statistical theory is based on a numerical summary called the *variance*. The variance is the sum of the squares of the deviations divided by one less than the number of observations. Squaring the deviations is one way of ignoring the signs of the deviations (although using absolute values may seem simpler). Dividing by one less than the number of observations instead of just dividing by the number of observations seems strange, but this is related to the fact that we must have at least two observations before we can measure dispersion. (One observation by itself does not give us a measure of variation.)

With summation notation, we use $\Sigma(x - \bar{x})^2$ to represent the sum of the squares of the deviations, on which the variance is based. It is important to realize that this notation implies that we first square each deviation, and then sum the results. This is quite different from summing the deviations first and then squaring this sum (which of course will always be zero, since, as we noted earlier, the sum of the deviations is always zero). Verify that $\Sigma(x - \bar{x})^2 = 12.9$ for "Number of Children" with the $n = 10$ individuals in the urban area of residence in the SURVEY DATA.

Because the variance is so widely used as a measure of dispersion, the special symbol s^2 is used to denote the variance. Using summation notation, we can write the following:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} .$$

The only way the variance can be equal to zero is if every observation in a data set is equal to the same value, implying that every observation will be equal to the mean. The more dispersion there is in a data set, that is, the more the observations tend to differ from each other and from the mean, the larger the variance will be. Since you have just found that $\Sigma(x - \bar{x})^2 = 12.9$ for "Number of Children" with the $n = 10$ individuals in the urban area of residence in the SURVEY DATA, we then find that the variance is

$$s_U^2 = \frac{12.9}{10 - 1} \approx 1.433 .$$

The subscript "U" on the s_U^2 indicates that the variance is being calculated for urban residents.

Since the computation of the variance involves squaring observations, its unit of measurement is not the same as that of the observations. For instance, if we are measuring distance in feet, then the variance will be in terms of square feet. While square feet is a meaningful unit of measurement, it is not the same unit of measurement as that of the observations. Often, the unit of measurement for the variance is not even meaningful, such as when we are observing number of children where the variance would be in terms of square number of children! In order that we have a measure of dispersion which has the same unit of measurement as the observations (as the mean and median do), we define the *standard deviation* to be the square root of the variance, and we represent the standard deviation with the symbol s . For the $n = 10$ individuals in the urban area of residence in the SURVEY DATA, the standard deviation of the variable "Number of Children" is $s_U = \sqrt{1.433} \approx 1.197$.

You previously found that for the variable "Number of Children" In the SURVEY DATA, $\bar{x}_R = 4.6$ and $\bar{x}_S = 3.0$. Now, find the sum of squared deviations from the mean, find the variance, and find the standard deviation for the individuals with a rural residence and for the individuals with a suburban residence. (For the rural area, you should find that $\Sigma(x - \bar{x})^2 = 64.4$, $s_R^2 \approx 7.111$, and $s_R \approx 2.667$; for the suburban area, you should find that $\Sigma(x - \bar{x})^2 = 30$, $s_S^2 \approx$, and $s_S \approx 1.826$.)

Notice that the standard deviation is smallest for the urban area and largest for the rural area. This reinforces what we see in the dot plots for Self-Test Problem 3-1 (displayed in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of Unit 3), which is that there appears to be more dispersion in the distribution of "Number of Children" in the rural area than in the urban area.

We have previously seen how positive or negative skewness in data can affect the mean but not the median. Because of this, we can base a measure of skewness on the distance between the mean and the median. In particular, we can compare the distance between the mean and median to the standard deviation by calculating the *skewness ratio* which we shall define to be $(\bar{x} - \text{median})/s$, which represents the number of standard deviations of difference between the mean and the median. This ratio will always be between -1 and 1 ; in other words, the mean and median will never be more than one standard deviation apart. Note that the skewness ratio will negative if the mean is smaller than the median and will be positive if the mean is larger than the median. Roughly speaking, a distribution can be considered very negatively skewed if the skewness ratio is less than -0.3 , and a distribution can be considered very positively skewed if the skewness ratio is more than $+0.3$. Of course, for a distribution which has a shape close to being symmetric, the skewness ratio will be close to zero.

From the SURVEY DATA, find the skewness ratio of the variable “number of children” for each of the three areas of residence (rural, suburban, urban); all the means and standard deviations you need appear earlier in this unit, and the medians were found in Unit 4 to be 3.5, 3, and 2 for the rural, suburban, and urban areas respectively. You should find that the skewness ratio is 0.41, 0, and 0.08 for the rural, suburban, and urban areas respectively. These results reinforce what we see in the dot plots for Self-Test Problem 3-1 (displayed in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of Unit 3), which is that the distribution of "Number of Children" is very positively skewed in the rural area, is symmetric in the suburban area, and is close to symmetric in the urban area.

Self-Test Problem 6-1. The variable "Yearly Income" in thousands of dollars for the 10 Republicans in the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, is recorded as follows:

34 55 53 45 30 29 33 61 41 64 .

In Self Test Problem 4-2, we found the mean to be 44.5 thousand dollars and the median to be 43.0 thousand dollars.

- Find the variance and standard deviation of the incomes for the Republicans.
- Find the skewness ratio for the incomes of the Republicans, and indicate what this suggests about the shape of the distribution.
- How could the 10 original incomes be altered so that the mean remains the same, but the standard deviation is larger?
- How could the 10 original incomes be altered so that the mean remains the same, but the standard deviation is smaller?

Self-Test Problem 6-2. In the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, following numerical summaries (which you may decide to verify for practice) apply to yearly income (\$1000s):

Republicans

$$\bar{x}_R = 44.500$$

$$median = 43.0$$

$$s_R = 13.083$$

Democrats

$$\bar{x}_D = 49.375$$

$$median = 44.5$$

$$s_D = 19.957$$

- How does the center of the distribution of yearly incomes seem to compare for the two parties?
- How does the dispersion in the distribution of yearly incomes seem to compare for the two parties?
- How does the shape of the distribution of yearly incomes seem to compare for the two parties?

Self-Test Problem 6-3. Figure 2-7 displays a box plot of yearly salaries for each of three mythical corporations. Indicate for which corporation(s) the standard deviation of salaries is likely to be the largest, and for which corporation(s) the standard deviation of salaries is likely to be the smallest.

We have considered obtaining the standard deviation only from raw data. If quantitative observations have been summarized into a frequency distribution with the raw data unavailable, we can still obtain, or at least approximate, the standard deviation. Table 4-1 is frequency distribution constructed from the number of absences in the last year for each of 40 employees in a particular office. We previously found from this frequency distribution that the mean days of absence is $\bar{x} = 1.6$.

To find the variance, we need to divide the sum of squared deviations $\Sigma(x - \bar{x})^2$ by one less than the number of observations (i.e., divide by $n - 1 = 40 - 1 = 39$). We easily obtained the mean from Table 4-1 by first realizing that the data contain ten 0s, thirteen 1s, seven 2s, five 3s, three 4s, and two 5s. If we wished, we could actually write down a list of the 40 deviations from the mean, but this would not be particularly useful.

Instead, we realize that we could obtain the sum of squared deviations $\Sigma(x - \bar{x})^2$ simply by multiplying $(0 - 1.6)^2$ by 10, multiplying $(1 - 1.6)^2$ by 13, multiplying $(2 - 1.6)^2$ by 7, multiplying $(3 - 1.6)^2$ by 5, multiplying $(4 - 1.6)^2$ by 3, multiplying $(5 - 1.6)^2$ by 2, and summing these results. Complete the calculation to find the variance. (You should find the variance and standard deviation for days of absence to be $s^2 = 2.092$ and $s = 1.446$ respectively.)

Self-Test Problem 6-4. In Self Test Problem 4-3, 50 households in a certain area were surveyed, and the number of cars owned in each was recorded. It was found that three households have no cars, four have 1 car, six have 2 cars, twelve have 3 cars, eighteen have 4 cars, five have 5 cars, and two have 6 cars. It was also found that the mean number of cars per household is 3.22 cars and the median number of cars per household is 3.5.

- (a) Find the standard deviation of cars per household for this data.
- (b) Find the skewness ratio, and indicate what this suggests about the shape of the distribution.

Self-Test Problem 6-5. Figure 2-8 displays a box plot of weights for each of five mythical types of oranges. Indicate what difference, if any, there is likely to be in the standard deviation between each of the following pairs:

- (a) Type A and Type B,
- (b) Type C and Type D,
- (c) Type D and Type E.

Self-Test Problem 6-6. Potato yield in pounds per acre is recorded for each of 300 different plots, where variety X potatoes are used in 150 plots, and variety Y potatoes are used in 150 plots. The results are organized into the two frequency distributions displayed as Tables 6-1 and 6-2.

- (a) Indicate what difference, if any, there is likely to be in the standard deviation of potato yield between the two varieties.
- (b) Indicate what difference, if any, there is likely to be in the skewness ratio of potato yield between the two varieties.

We defined both the mean and the median for the distribution of a quantitative variable, since we generally cannot calculate either of these with a qualitative variable. For instance, it makes no sense to compute a mean or median for the variable "Political Party Affiliation" in

<u>Yield (lbs./acre)</u>	<u>Raw Frequency</u>
Above 100 to 120	17
Above 120 to 140	20
Above 140 to 160	24
Above 160 to 180	24
Above 180 to 200	22
Above 200 to 220	21
Above 220 to 240	22

<u>Yield (lbs./acre)</u>	<u>Raw Frequency</u>
Above 100 to 120	7
Above 120 to 140	10
Above 140 to 160	34
Above 160 to 180	44
Above 180 to 200	32
Above 200 to 220	11
Above 220 to 240	12

the SURVEY DATA. With a qualitative variable, however, we can define a numerical summary known as the *mode*, which is defined to be the most frequently occurring observation(s) if any exist. The mode might be considered an analog to a measure of center for the distribution of a quantitative variable.

Table 3-1, Figure 3-1, and Figure 3-2 each display the number of individuals in each of the four categories for "Political Party Affiliation." The mode is the category with the largest frequency, which is the Republican Party. When no observation occurs more often than any other, we can call every observation a mode, or we might say that there is no mode. This is the case with the variable "Residence" in the SURVEY DATA, where each of the three categories (rural, suburban, urban) is observed exactly ten times.

With a quantitative variable, we can of course obtain a mode, but the mean and median are often of more interest than the mode. Table 3-2, Figure 2-1, and Figure 3-4 each display frequencies for the variable "Number of Children," from which we find the mode to be 2 children. If we focus only on urban residents, we find from the corresponding dot plot among the dot plots for Self-Test Problem 3-1, that there are two modes, 2 children and 3 children, since both occur the same number of times but more than any other value in the data. Now, you find the mode(s), if any exist, for the rural residents and the suburban residents. (You should find that the mode for the rural area is 2 children, and that there are three modes for the suburban area: 2 children, 3 children, and 4 children.)

Answers to Self-Test Problems

- 6-1** (a) $s^2 = 171.167$ and $s = 13.083$. (b) $(44.5 - 43.0) / 13.083 = 0.11$; this suggests that the distribution of incomes for the Republicans is a little positively skewed. (c) Change the smallest income of 29 thousand dollars to 19 thousand dollars, and change the largest income of 64 thousand dollars to 74 thousand dollars. (d) Change the smallest income of 29 thousand dollars to 39 thousand dollars, and change the largest income of 64 thousand dollars to 54 thousand dollars.
- 6-2** (a) It appears that the distribution of incomes for Democrats is centered at a higher value than for Republicans, since both the median and mean are higher for Democrats. (b) It appears that the distribution of incomes for Democrats is more dispersed than for Republicans, since the standard deviation is higher for Democrats. (c) Since the skewness ratio is 0.11 for Republicans and 0.24 for Democrats, it appears that the distribution of incomes is a little positively skewed for the Republicans but considerably more positively skewed for Democrats.
- 6-3** The standard deviation is likely to be the smallest for the Phi Corporation. The standard deviation will most likely be equal, or very close, for the Beta Corporation and the Gamma Corporation.
- 6-4** (a) The standard deviation of cars per household is $\sqrt{102/58/49} = 1.447$. (b) $(3.22 - 3.5) / 1.447 = -0.19$; this suggests that the distribution is a little negatively skewed.
- 6-5** (a) Type A is likely to have a larger standard deviation than Type B. (b) Type C and Type D are likely to have close to the same standard deviation. (c) Type E is likely to have a larger standard deviation than Type D.
- 6-6** (a) Variety X is likely to have a larger standard deviation than variety Y. (b) The skewness ratio will probably be close to zero for each variety.

Summary

For the distribution of a quantitative variable, the mean (\bar{x}) and median (Q_2) are measures of center, and the range and interquartile range (*IQR*) are measures of dispersion. Other measures of dispersion can be defined based on *deviations from the mean*:

$$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}.$$

A widely used measure of dispersion is the *variance*, denoted by s^2 . The variance is defined to be the sum of the squared deviations from the mean divided by one less than the number observations, that is,

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

The only way the variance can be equal to zero is if every observation in a data set is equal to the same value, implying that every observation will be equal to the mean. The more dispersion there is in a data set, that is, the more the observations tend to differ from each other and from the mean, the larger the variance will be. The *standard deviation*, denoted by s , is defined to be the square root of the variance; the standard deviation provides us with a measure of dispersion in the same unit of measurement as in the original data.

We define the *skewness ratio* to be $(\bar{x} - \text{median})/s$. Roughly speaking, a distribution can be considered very negatively skewed if the skewness ratio is less than -0.3 , and a distribution can be considered very positively skewed if the skewness ratio is more than $+0.3$. For a distribution which has a shape close to being symmetric, the skewness ratio will be close to zero.

The *mode* is defined to be the most frequently occurring observation(s) if any exist. With a quantitative variable, the mean and median are often of more interest than the mode. The mode is more useful with a qualitative variable, where computing a mean or median makes no sense as a general rule.