

UNIT 4

Numerical Summaries

Objectives:

- To obtain and interpret the mean and median for quantitative data
- To obtain and interpret the range and the interquartile range

The three major characteristics of the distribution for a quantitative variable that are of primary interest are the center of the distribution, the amount of dispersion in the distribution, and the shape of the distribution. Our study of these three characteristics thus far has been mainly through the use of graphical displays. However, we shall find it useful and necessary to be able to describe these characteristics with numerical summaries. A *numerical summary* is a number used to describe a specific characteristic about a data set.

What numerical summaries could we use to describe the center of a distribution? One natural choice is Q_2 , the second quartile in the five-number summary, since this is a value which divides the ordered array for the data in half. Since we think of Q_2 as a measure of the center of a distribution, we give it a special name in statistical terminology; we call Q_2 the *median*.

Another possible choice for a numerical summary to describe the center of a distribution would be the average of the observations. The average for a list of numbers is most likely not a new concept to you. To obtain the average of observations of a quantitative variable, we simply add the observations, and divide this sum by the number of observations. We give this arithmetic average a special name in statistical terminology; we call the average the *mean*. The mean is so important and widely used in classical statistics, that it has been given its own special symbol; the special symbol \bar{x} (read as "x-bar") is used to denote the mean.

Recall how letters are often used in mathematical notation to represent numerical values. We shall find it convenient to let the letter n represent the number of observations in a data set. In statistics, we frequently use subscripted letters to represent a list of n numerical values. In other words, a data set consisting of n quantitative observations can be represented by x_1, x_2, \dots, x_n . It is common practice in mathematics to use the upper case Greek letter sigma Σ is used to denote summation. To understand how Σ is used, consider the variable "Number of Children" in the SURVEY DATA, displayed as data set A.1 in the appendix. Focusing only on the $n=10$ individuals with an urban residence, we let

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

The sum of the x_i s is denoted as Σx_i , or more simply just as Σx . (That is, for the sake of brevity, we often delete the subscript i with the sigma notation.) You can easily check that

$$\Sigma x = \Sigma x = x_1 + x_2 + \dots + x_n = 2+3+3+1+0+2+2+4+3+1 = 21.$$

The mean \bar{x} of n quantitative observations x_1, x_2, \dots, x_n is simply the sum Σx divided by the number of observations n . Using the summation notation, we can write

$$\bar{x} = \frac{\Sigma x}{n} .$$

Since we found that $\Sigma x = 21$ for the variable "Number of Children" in the SURVEY DATA when focusing only on the $n=10$ individuals with an urban residence, the mean "Number of Children" for these urban residents is

$$\bar{x}_U = \frac{21}{10} = 2.1 \text{ children.}$$

The subscript "U" on the \bar{x}_U indicates that the mean is calculated for urban residents.

To find the median (Q_2) for urban residents, we need to average the two middle observations in the ordered array for the $n=10$ individuals:

$$0 \quad 1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 4 .$$

We find the median number of children for the urban area to be $(2+2)/2 = 2$ children.

From the SURVEY DATA, find the mean and median number of children for each of the other two areas of residence (rural and suburban). You should find that the mean and median for the rural area are respectively 4.6 and 3.5, and that the mean and median for the suburban area are respectively 3.0 and 3.0.

Since the mean and the median are each measures of the center of a distribution, it is reasonable to wonder whether or not they will always be close to each other in value. Compare the mean and the median number of children for each of the three areas of residence. In the rural area the mean and median appear to be quite different (4.6 and 3.5), whereas in the suburban area the mean and median are equal to one another (3.0 and 3.0), and in the urban area the mean and median appear to be quite close (2.1 and 2.0). To see what characteristic about a distribution can affect whether or not the mean and median will be close in value, return to the dotplots of Figure A.1-1(d).

Remember that a positively skewed distribution is one where the observations in the upper half of the distribution show considerably more dispersion than the observations in the lower half of the distribution, and a negatively distribution is one where the observations in the lower half of the distribution show considerably more dispersion than the observations in the upper half of the distribution. We see that the distribution of "Number of Children" is positively skewed in the rural area, is symmetric in the suburban area, and is a little negatively skewed in the urban area. This suggests to us that the difference in value between the mean and median can be influenced by skewness. The fact that there are a few unusually large observations of "Number of Children" in the rural area explains why the mean is so much larger than the median. Since the mean depends on the sum of the observations, it is influenced by a few unusually large or unusually small observations.

To illustrate this point further, let us return to the ordered array for "Number of Children" for urban residents:

0 1 1 2 **2** **2** 3 3 3 4

Earlier, we found the mean to be $\bar{x}_U = 2.1$ children, and we found the median to be $(2+2)/2 = 2$ children.

The two middle observations of the ordered array, from which the median is obtained, are boldfaced. Suppose, however, that one of the 3s in the ordered array was replaced by a 23; that is, suppose there was one urban resident in the SURVEY DATA with 23 children. (Okay, perhaps this is a bit far fetched, but just imagine it is true for a minute for the sake of the point we are trying to make.) The ordered array would then be as follows:

0 1 1 2 **2** **2** 3 3 4 23

It is easy to see that the median would still be equal to 2 children, but the sum of the observations would now be 41, making the mean $41/10 = 4.1$ children. The mean would actually be larger than every observation except one! This illustrates how the mean can give us a misleading impression about the center of a distribution.

The fact that the mean can be sensitive to a few unusually large, or unusually small, observations, but not the median, leads us to two noteworthy facts. First, with a positively or negatively skewed distribution, the median is often a more reasonable measure of center than the mean. Second, the values of the mean and median together not only give us information about the center of a distribution, but also give us some information about the shape of a distribution. When a distribution is nearly symmetric, the mean and median will be close in value; when the mean and median are not close in value, the distribution will be positively or negatively skewed. If the mean is smaller than the median, the distribution is negatively skewed; if the mean is larger than the median, the distribution is positively skewed.

We find that the mean and median are equal for the suburban area and close for the urban area. This reinforces our earlier statements about the dotplots of Figure A.1-1(d), that the distribution of "Number of Children" appears to be symmetric in the suburban area and only a little negatively skewed in the urban area. On the other hand, the difference in mean and median for the rural area reinforces our earlier statement that the distribution of "Number of Children" appears in Figure A.1-1(d) to be positively skewed for the rural area.

We now have two numerical summaries to describe the center of a distribution: the mean and the median. What numerical summaries could we use to describe the amount dispersion in a distribution? One natural choice is the *range*, which we have previously defined to be *Max-Min*. Obviously, though, the range is sensitive to one unusually large, or one unusually small, observation. Another possible measure of dispersion is $Q_3 - Q_1$, the difference between the third and first quartiles; this difference is called the *interquartile range*, which we shall abbreviate as *IQR*.

You should easily be able to check that *IQR* for the variable "Number of Children" in the SURVEY DATA is $3 - 1 = 2$ for the urban area of residence. Find *IQR* for each of the other two areas of residence (rural and suburban). You should find that *IQR* = 5 for the rural area, and that *IQR* = 2 for the suburban area. This reinforces the fact that when we look at the dotplots of Figure A.1-1(d), we that there appears to be considerably more dispersion in the rural area than in either of the suburban or urban areas.

Self Test Problem 4-1. Use the data from the variable "Yearly Income" in the SURVEY DATA, displayed as data set A.1 in the appendix to do each of the following:

- (a) Find the mean and median for males; then find the mean and median for females.
- (b) Do the means and medians found in part (a) suggest that the distribution of "Yearly Income" is centered at a different value for males and females? If yes, for which sex does the center of the distribution appear to be greater?
- (c) Do the means and medians found in part (a) suggest that the distribution of "Yearly Income" has a different shape for males and females? If yes, how does the shape appear to differ between the two sexes?
- (d) Find the range and interquartile range for males; then find the range and interquartile range for females.
- (e) Do the ranges and interquartile ranges found in part (d) suggest that the distribution of "Yearly Income" shows a different amount of dispersion for males and females? If yes, for which sex does the amount of dispersion appear to be greater?

Self Test Problem 4-2. The "Yearly Income" in thousands of dollars for the 10 Republicans in the SURVEY DATA, displayed as data set A.1 in the appendix, are as follows:

34 55 53 45 30 29 33 61 41 64.

The "Yearly Income" in thousands of dollars for the 8 Democrats in the SURVEY DATA are as follows:

28 75 26 78 40 60 49 39

- (a) Find the mean and median of the incomes for the Republicans; then find the mean and median of the incomes for the Democrats.
- (b) Do the means and medians found in part (a) suggest that the distribution of "Yearly Income" is centered at a different value for Republicans and Democrats? If yes, for which party does the center of the distribution appear to be greater?
- (c) Do the means and medians found in part (a) suggest that the distribution of "Yearly Income" has a different shape for Republicans and Democrats? If yes, how does the shape appear to differ between the two parties?
- (d) Find the range and interquartile range for Republicans; then find the range and interquartile range for Democrats.
- (e) Do the ranges and interquartile ranges found in part (d) suggest that the distribution of "Yearly Income" shows a different amount of dispersion for Republicans and Democrats? If yes, for which party does the amount of dispersion appear to be greater?
- (f) In part (a), the mean income for the 10 Republicans was found to be 44.5 thousand dollars, and the mean income for the 8 Democrats was found to be 49.375 thousand dollars. You now are told that the mean income is 42.5 thousand dollars for 4 Independents, and that the mean income is 44 thousand dollars for 8 Others. Find the mean income for all 30 individuals.

We have considered obtaining the mean and median 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 mean and median. 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.

To find the mean, we need to divide the sum $\sum x$ by the number of observations $n=40$. From Table 4-1, you should realize 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 observations, but this would not be particularly useful. Instead, we realize that we could obtain the sum of the 40 observations simply by multiplying 0 by 10, multiplying 1 by 13, multiplying 2 by 7, multiplying 3 by 5, multiplying 4 by 3, multiplying 5 by 2, and summing these

Table 4-1
Absences for Last Year

<u>Days of Absences</u>	<u>Raw Frequency</u>
0	10
1	13
2	7
3	5
4	3
5	2

results. This is because multiplying 0 by 10 is the same as summing ten 0s, multiplying 1 by 13 is the same as summing thirteen 1s, multiplying 2 by 7 is the same as summing seven 2s, etc. Complete the calculation to find the mean. (You should find the mean days of absence to be $\bar{x} = 1.6$.)

To find the median, we need the two middle observations in the ordered array of 40 observations; these are the 20th and 21st observations. From Table 4-1, you should realize that the first ten observations in the ordered array are 0s, and the next 13 observations are 1s. Consequently, the 20th and 21st observations in the ordered array must each be 1, which implies that the median is $(1+1)/2 = 1$ day. Notice that the mean appears to be a bit larger than the median. This, of course, reflects the fact that the distribution of days absent is positively skewed, which you can see just by looking at the raw frequencies and picturing the shape of a histogram for this frequency distribution.

To find the interquartile range, we must first find Q_1 and Q_3 . We obtain Q_1 by averaging the two middle observations among the lower 20 observations in the ordered array; these are the 10th and 11th observations. We obtain Q_3 by averaging the two middle observations among the upper 20 observations in the ordered array; these are the 30th and 31st observations. Find Q_1 and Q_3 from Table 4-1. You should find that $Q_1 = (0+1)/2 = 0.5$ days and $Q_3 = (2+3)/2 = 2.5$ days. We then have $IQR = 2.5 - 0.5 = 2$ days

Table 4-2 is frequency distribution constructed from overtime in the last week for each of 175 employees in a particular office. Unlike Table 4-1, Table 4-2 does not display the actual individual values of the observations. Earlier, when we were working from Table 4-1, we could actually write down a list of the observations. In Table 4-2, however, classes have been defined. You should realize from Table 4-2 that the data contain 21 observations which are greater than 0 up to and including 3 hours, but we cannot say precisely what any of these observations are equal to. Similar comments can be made about each of the other classes. In other words, we could not actually write down a list of the 175 observations.

<u>Overtime Hours</u>	<u>Raw Frequency</u>
Above 0 to 3	21
Above 3 to 6	35
Above 6 to 9	49
Above 9 to 12	42
Above 12 to 15	28

To find the median, we need the middle observation in the ordered array of 175 observations; this is the 88th observation. We can tell from Table 4-2 that 21 observations in the ordered array are less than or equal to 3 hours, $21+35 = 56$ observations in the ordered array are less than or equal to 6 hours, and $21+35+49 = 105$ observations in the ordered array are less than or equal to 9 hours. Consequently, the 88th observation in the ordered array, which is the median, must each be somewhere between 6 and 9 hours. By looking at the raw frequencies and picturing the shape of a histogram for this frequency distribution, we can tell that the distribution of overtime in the last week is close to symmetric (or perhaps just a bit negatively skewed). Consequently we might make a guess that the mean is also somewhere between 6 and 9 hours. If we were interested in a more accurate approximation of the mean and median, we could work with the midpoints of the classes and make use of interpolation; we shall not do this here, however.

Self Test Problem 4-3. Fifty households in a certain area are surveyed, and the number of cars owned in each is recorded. 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.

- (a) In the construction of a frequency distribution for this data, why would it not be necessary to define classes?
- (b) Find the mean number of cars per household and the median number of cars per household.
- (c) What does the difference between the mean and median say about the shape of the distribution?
- (d) Find the range and the interquartile range for the number of cars per household.

Self Test Problem 4-4. Potato yield in pounds per acre is recorded for each of 150 different plots. The results are organized into the frequency distribution displayed as Table 4-3.

- (a) Why was it necessary to define classes in the construction of the frequency distribution for this data?
- (b) What does the shape of the distribution suggest about the difference between the mean and median potato yield?
- (c) In which class does the median potato yield lie?
- (d) What should you tell someone who tries to obtain the mean potato yield by dividing $17+20+24+24+22+21+22$ by 7?
- (e) In which class does Q_1 lie, and in which class does Q_3 lie, for these potato yields?
- (f) Explain why it is not possible for the mean of the observations used to create Table 4-3 to be equal to 23.

Table 4-3
Potato Yield

<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

Answers to Self Test Problems

- 4-1 (a) For males, $\bar{x}_M = 53.4$ thousand dollars and median = 55 thousand dollars; for females, $\bar{x}_F = 37.4$ thousand dollars and median = 34 thousand dollars. (b) Since the mean and median for males are each more than 10 thousand dollars greater than the mean and median for females, it would appear that the distribution of incomes for males is centered at a higher value. (c) The distribution of yearly income appears to be a little negatively skewed for males and somewhat positively skewed for females. (d) For males, range = $78 - 30 = 48$ thousand dollars, and $IQR = 65 - 39 = 26$ thousand dollars; for females, range = $71 - 25 = 46$ thousand dollars, and $IQR = 44 - 28 = 16$ thousand dollars. (e) The ranges and interquartile ranges do not suggest a drastic difference in the amount of dispersion for males and females.
- 4-2 (a) For Republicans, $\bar{x}_R = 44.500$ thousand dollars and median = 43.0 thousand dollars; for Democrats, $\bar{x}_D = 49.375$ thousand dollars and median = 44.5 thousand dollars. (b) Since the mean and median for Democrats are each respectively greater than the mean and median for Republicans, it would appear that the distribution of incomes for Democrats is centered at a higher value. (c) The distribution of yearly income appears to be more positively skewed for Democrats than for Republicans. (d) For Republicans, range = $64 - 29 = 35$ thousand dollars, and $IQR = 55 - 33 = 22$ thousand dollars; for Democrats, range = $78 - 26 = 52$ thousand dollars, and $IQR = 67.5 - 33.5 = 34$ thousand dollars. (e) The ranges and interquartile ranges suggest that there is less dispersion in incomes among Republicans than among Democrats. (f) The sum of the incomes for all 30 individuals is $\sum x = (10)(44.5) + (8)(49.375) + (4)(42.5) + (8)(44) = 1362$, from which we find the mean income for all 30 individuals to be $1362/30 = 45.4$ thousand dollars.
- 4-3 (a) Classes would not be necessary, since the data consist of a few different values with many repetitions. (b) The mean number of cars per household is 3.22; the median number of cars per household is 3.5. (c) The mean is a little smaller than the median, since the distribution is somewhat negatively skewed. (d) range = $6 - 0 = 6$ cars, and $IQR = 4 - 2 = 2$ cars.
- 4-4 (a) Classes would have been necessary, since the data most likely consists of many distinct values with few repetitions. (b) The fact that distribution roughly appears to be a uniform distribution, which is symmetric, suggests that the mean and median are close in value. (c) The median potato yield is between 160 and 180 lbs./acre. (d) The numbers being averaged are the raw frequencies, not the potato yields. (e) Q_1 is very close to 140, and Q_3 is between 200 and 220 lbs./acre. (f) Since all the observations used to create Table 4-3 must be between 100 and 240 lbs./acre, it is not possible for the mean to be equal to 23.

Summary

A *numerical summary* is a number used to describe a specific characteristic about a data set. The mean and median are two numerical summaries to describe the center of a distribution for a quantitative variable; the range and interquartile range are two numerical summaries to describe dispersion in a distribution for a quantitative variable.

The *median* of n quantitative observations x_1, x_2, \dots, x_n is Q_2 , the second quartile in the five-number summary. The *mean* of n quantitative observations x_1, x_2, \dots, x_n is simply the arithmetic average calculated by dividing the sum of the observations by the number of observations; using the summation notation, we can

say that $\bar{x} = \frac{\sum x}{n}$ is the mean. The mean can be sensitive to a few unusually large, or unusually small,

observations, but not the median. When a distribution is nearly symmetric, the mean and median will be close in value; when the mean and median are not close in value, the distribution will be positively or negatively skewed. If the mean is smaller than the median, the distribution is negatively skewed; if the mean is larger than the median, the distribution is positively skewed. With a positively or negatively skewed distribution, the median is often a more reasonable measure of center than the mean.

The *range* is *Max-Min*, the difference between the largest and smallest observations. The *interquartile range*, abbreviated *IQR*, is $Q_3 - Q_1$, the difference between the third and first quartiles. The range is sensitive to one unusually large, or one unusually small, observation.