2.6: Measures of Dispersion

Learning Objectives

- Explore several measures of dispersion in data
- Develop measures of dispersion:
  - Range
  - Interquartile Range
  - Mean Absolute Deviation
  - Variance
  - Standard Deviation
- Compute various measures of dispersion

Section 2.6 Excel File (contains all of the data sets for this section)

What is Dispersion?

Consider the two histograms in Figure 2.6.1 representing scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, we can see that the distributions are quite different. The scores on Quiz 1 are more densely packed than the scores on Quiz 2. The differences of scores were much greater on Quiz 2 than on Quiz 1.

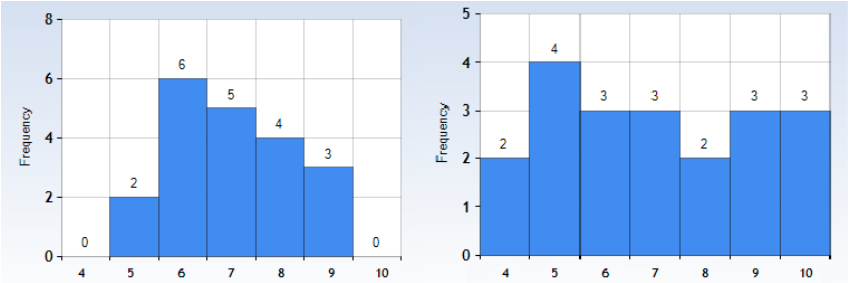

Figure 2.6.1: Histograms for Quiz 1 (left) and Quiz 2 (right)

The terms variability, spread, and dispersion are synonyms. They refer to how varied data are in a data set or how spread out the distribution of the data is. In this section we will discuss measures of the dispersion of a distribution. We seek a single number to describe how spread out or dispersed the data is. There are many ways to measure "dispersion," and no single measure gives complete insight into the data's dispersion. We will examine five frequently used measures of dispersion: the range, interquartile range, mean absolute deviation, variance, and standard deviation.

Range

The range is the most straightforward measure of dispersion to calculate. As a warning, the term "range" is used in multiple ways, so do not confuse the statistical use of this word with other uses, such as in algebra. Recall that our summary measures tend to be given as a single value so, in statistics, the **range** is simply the highest data value minus the lowest data value, that is:

$$\text{range} = \text{maximum} - \text{minimum}.$$

Since we are subtracting data values, we must work with interval or ratio-level data for the range to have meaning; we do not have a range measure in nominal or ordinal level data.

Text Exercise 2.6.1

Determine the range of the following group of numbers.

$$10, 2, 5, 6, 7, 3, 4$$

Answer

The highest number is 10, and the lowest number is 2, so $10 - 2 = 8$. The range is 8. These values are within 8 units from each other.

? Text Exercise 2.6.2

Now consider the two quizzes in Figure 2.6.1. What is the range of each quiz?

Answer

On Quiz 1, the lowest score is 5 and the highest score is 9. Therefore, the range is 4.

The range on Quiz 2 is larger: the lowest score is 4 and the highest score is 10. Therefore, the range is 6.

Since Quiz 1 has a smaller range, we can say that Quiz 2 is more spread out than Quiz 1.

The range is a quick way to get a rough idea of the spread of the data. However, it is a very coarse measure since it depends on only two data points. The sets $\{0,5,5,5,5,10\}$ and $\{0,0,0,10,10,10\}$ have the same range but would not be called equally dispersed. We must investigate other measures.

Interquartile Range

A similar measure to the range is the **interquartile range (IQR)**. The IQR is the range of the middle 50% of the scores in a distribution. We can understand this visually as the length of the box in the box plot.

$$\text{IQR} = 75^{th} \text{ percentile} - 25^{th} \text{ percentile} = Q_3 - Q_1$$

? Text Exercise 2.6.3

In Section 2.4, we looked at data from men and women on the Stroop Test. In the women's data, the 25^{th} percentile is 17, the 50^{th} percentile is 18.5, and the 75^{th} percentile is 20. For the men, the 25^{th} percentile is 19, the 50^{th} percentile is 22, and the 75^{th} percentile is 24. Calculate the IQR for men and women.

Answer

Women: $\text{IQR} = 20 - 17 = 3$

Men: $\text{IQR} = 24 - 19 = 5$

Measures of dispersion relay how spread out the data is. The range of a data set gives the distance between the minimum and maximum values. Similarly, the IQR of a data set gives a distance, but this time the distance is the smallest length of an interval such that the central 50% of observations could fall into the interval. Larger values of range and IQR indicate that the data set is more spread out. One can have data with a large range and small (\text{IQR},\) indicating large dispersion in the entire data set, and yet the central 50% of the data is not varied in comparison.

As we mentioned earlier, these measures do not provide a complete understanding of the data set since they depend on only a few values. As we progress, we will discuss measures that incorporate all data values into the calculation, but even these measures of variability will not provide a complete understanding: better, yes; complete, no.

? Text Exercise 2.6.4

For each measure of dispersion discussed so far, range and IQR, construct two data sets (each with 10 values) that have the same measure of dispersion value but starkly different degrees of spread when looking at the data.

Answer

It is good to remember that there are many different solutions to these questions.

Let us begin with range. If our two data sets are to have the same range, the difference between maximum and minimum values must be the same. We can pick that value freely; let us say 20. We could have both minimum and maximum values be the same, or we could have them be different. Let us keep them the same.

$$\{10, \ , \ , \ , \ , \ , \ , \ , \ , 30\}$$
$$\{10, \ , \ , \ , \ , \ , \ , \ , \ , 30\}$$

Here again, we had the freedom to pick at least one of our values. Once 30 was chosen as a maximum, 10 was forced to be our minimum.

We now need to think about how we could have different degrees of spread. We could have values spread fairly evenly from 10 to 30 in one data set and have them closely packed around one value in the other.

$$\{10, 13, 15, 17, 19, 21, 23, 25, 27, 30\}$$
$$\{10, 20, 20, 20, 20, 20, 20, 20, 20, 30\}$$

Now let us look at IQR. If our two data sets are to have the same IQR, the difference $Q_3 - Q_1$ needs to be the same. As with range, we could have Q_1 and Q_3 be the same or different. Since we have one example where we had the same values, we will make them different. Let us again choose 20 for our value. Since $n = 10$, Q_1 is the third value in our ordered data set, and Q_3 is the eighth value in our ordered data set.

$$\{\ ,\ , 10,\ ,\ ,\ ,\ , 30,\ ,\ \}$$
$$\{\ ,\ , 15,\ ,\ ,\ ,\ , 35,\ ,\ \}$$

What sort of differences could we have in our data sets to elicit different degrees of spread? We already considered a reasonably uniform spread and one centered on a single value. We could have a greater spread outside than inside our box as opposed to having two clusters centered at Q_1 and Q_3.

$$\{-50, -40, 10, 14, 20, 22, 26, 30, 80, 90\}$$
$$\{13, 14, 15, 16, 16, 34, 34, 35, 36, 37\}$$

Deviation

Measures of dispersion give us an idea about how spread apart our data are. Our previous measures incorporated only some of the data values. One way to include all of the data is to compare how far away each piece of data, call it x, is from some specific value v. We call the difference $x - v$ the **deviation from v.** A data value's distance from v, which would be $|x - v|$, is called the **absolute deviation from v.** There are many possible options for v. The most common choice of v is the mean. Once we have all the deviations, we must decide what to do with them since a measure of dispersion is a single value. Two options are summing up and averaging the deviations.

Consider the distribution of the five numbers 2,3,4,9,17.

Table 2.6.2: An example of various deviations

Values	Deviations from the mean	Deviations from the median	Absolute deviations from the mean	Absolute deviations from the median
2	$2-7=-5$	$2-4=-2$	$\|2-7\|=5$	$\|2-4\|=2$
3	$3-7=-4$	$3-4=-1$	$\|3-7\|=4$	$\|3-4\|=1$
4	$4-7=-3$	$4-4=0$	$\|4-7\|=3$	$\|4-4\|=0$
9	$9-7=2$	$9-4=5$	$\|9-7\|=2$	$\|9-4\|=5$
17	$17-7=10$	$17-4=13$	$\|17-7\|=10$	$\|17-4\|=13$
Sum	0	15	24	21
Average Deviation	0	3	$\frac{24}{5}=4.8$	$\frac{21}{5}=4.2$

Perhaps the final value of 0 in the "Deviations from the mean" column surprised some of us. After some thought, it should make sense that some values are above the mean and others are below the mean.

Examining the table can help us gain a deeper understanding. The last two columns look at the absolute values of the deviations rather than the deviations themselves. Deviation is similar to displacement, and the absolute value of deviation is similar to distance. When we sum our deviations, values below our central value contribute negatively, while values above our central value contribute positively and cancel each other out.

Note: Displacement and Distance

Displacement is the difference between the initial position and the final position.

Distance is the path length from our initial position to our final destination.

We are summing up all of the deviations from the mean. If we put this in summation notation, we arrive at

$$\sum (x - \bar{x})$$

Hopefully, we recognize this from our discussion on central tendency. The mean as the balance point is the value that makes this sum equal 0; this will be the case regardless of the data set. Thus, the average deviation from the mean is always 0.

We want to avoid the cancellation with summing deviations; there is more than 0 spread in the data set. The absolute value of the deviations finds the distance each observation is from the central value. The sum of the absolute deviations can be considered the total distance our observations are from our central value. Since the total distance is affected by the number of observations, we prefer to use the **mean absolute deviation from v ($\mathbf{MAD}$).** We can understand the MAD as the average distance the various data values are from the central value. Again, larger MAD values indicate that the data is spread to a greater degree.

$$\mathrm{MAD} = \frac{1}{n}\sum |x_i - v|$$

We will not prove this, but if v is chosen to be the median, the MAD is minimized. Try to convince yourself that this is true using Excel.

? Text Exercise 2.6.5

Consider the spread of the two data sets and compute the MAD from the mean.

$$\{-2, -2, 2, 2\}$$
$$\{-3, -1, 1, 3\}$$

Answer

Our first intuition is to look at the ranges, 4 and 6, and subsequently say that the second data set is more widely dispersed. Both data sets have a mean of 0. The MAD from the mean for the first data set is thus $\frac{2+2+2+2}{4} = 2$, and MAD from the mean for the second data set is $\frac{3+1+1+3}{4} = 2$. So both data sets have the same MAD from the mean. Note that every observation in the first data set is 2 units away from the mean, but in the second data set, two are closer than the average distance, and two are farther than the average distance. The MAD from the mean fails to distinguish between these two sets.

Variance

Recall that the difference between deviation and absolute deviation can be understood as the difference between displacement and distance. An advantage of using distance is that values do not cancel when summing. Using absolute values computes distances in a single dimension. There are many notions of distance in higher dimensions; hopefully, we are familiar with our standard two-dimensional distance formula $d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. Without going into the details, we can understand our next measures of dispersion from the perspective of a higher dimensional distance between our central value v and our data set $x_1, x_2, \ldots, x_n$.

$$d = \sqrt{(x_1 - v)^2 + (x_2 - v)^2 + \cdots + (x_n - v)^2}$$

Note that we sum over all the deviations from v and that each deviation from v is being squared. The details here are beyond the scope of this course, but hopefully, we have built at least an initial intuition as to why we might now consider the **squared deviation from v.** We do so using the same data from the previous section.

Table 2.6.3: An example of various squared deviations

Values	Squared deviations from the mean	Squared deviations from the median
2	$(2-7)^2 = (-5)^2 = 25$	$(2-4)^2 = (-2)^2 = 4$
3	$(3-7)^2 = (-4)^2 = 16$	$(3-4)^2 = (-1)^2 = 1$
4	$(4-7)^2 = (-3)^2 = 9$	$(4-4)^2 = (0)^2 = 0$
9	$(9-7)^2 = (2)^2 = 4$	$(9-4)^2 = (5)^2 = 25$
17	$(17-7)^2 = (10)^2 = 100$	$(17-4)^2 = (13)^2 = 169$
Sum	154	199
Mean	$\frac{154}{5} = 30.8$	$\frac{199}{5} = 39.8$

Interpreting the sums and means of these squared deviations is more complicated. We were dealing with displacements and distances previously; now, we have squared distances. If we had units attached to our data, such as cm, the units on these measures would be units^2,

such as cm^2, but the intuition that we have been building remains consistent. Larger values of these measures indicate greater degrees of spread. We shall encounter an associated measure with a more intuitive interpretation soon.

When considering the MAD, the median minimized the sum of the absolute deviations. We note that with the squared deviations, the median does not minimize the sum because the sum for the squared deviations from the mean is smaller! Indeed, the sum of the squared deviations from the mean is the smallest possible (if you have a background in calculus, see if you can show that the mean minimizes the sum of squared deviations). For this reason, among many others, we define the variance of a population data set as the average of the squared deviations from the mean.

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

We denote the variance with σ^2, where σ is the lower case Greek letter sigma. Recall that using Greek letters for a descriptive statistic indicates a population parameter. This is indeed the case here; this formula is for **population variance**.

? Text Exercise 2.6.6

Consider the quiz data from the beginning of this section in the table below. We are only interested in the performance of these particular students and, therefore, treat the data as population data. Compute the variance for both Quiz 1 and Quiz 2.

Table 2.6.4: Scores from Quiz 1 and Quiz 2.

Quiz 1	5	5	6	6	6	6	6	6	7	7	7	7	7	8	8	8	8	9	9	9
Quiz 2	4	4	5	5	5	5	6	6	6	7	7	7	8	8	9	9	9	10	10	10

Answer

To find the population variance, we follow these steps:

1. List each data value.
2. Calculate the mean.
3. Calculate the deviation from the mean for each score.
4. Square the deviations from the mean.
5. Average the squared deviations from the mean.

Table 2.6.5: Calculation of variance for Quiz 1 scores

Quiz 1 **Scores**	**Deviations from the Mean**	**Squared Deviations**
9	$9-7=2$	$2^2=4$
9	2	4
9	2	4
8	$8-7=1$	$1^2=1$
8	1	1
8	1	1
8	1	1
7	$7-7=0$	$0^2=0$
7	0	0
7	0	0
7	0	0
7	0	0
6	$6-7=-1$	$(-1)^2=1$
6	-1	1
6	-1	1

6	-1	1
6	-1	1
6	-1	1
5	$5-7=-2$	$(-2)^2=4$
5	-2	4
$\mu=\frac{\sum x}{N}=7$		$\sigma_1^2=\frac{\sum(x-\mu)^2}{N}=\frac{30}{20}=1.5$

Table 2.6.6: Calculation of variance for Quiz 2 scores

Quiz 2 **Scores**	**Deviations from the Mean**	**Squared Deviations**
10	$10-7=3$	$3^2=9$
10	3	9
10	3	9
9	$9-7=2$	$2^2=4$
9	2	4
9	2	4
8	$8-7=1$	$1^2=1$
8	1	1
7	$7-7=0$	$0^2=0$
7	0	0
7	0	0
6	$6-7=-1$	$(-1)^2=1$
6	-1	1
6	-1	1
5	$5-7=-2$	$(-2)^2=4$
5	-2	4
5	-2	4
5	-2	4
4	$4-7=-3$	$(-3)^2=9$
4	-3	9
$\mu=\frac{\sum x}{N}=7$		$\sigma_2^2=\frac{\sum(x-\mu)^2}{N}=\frac{78}{20}=3.9$

The variance for Quiz 1 is 1.5 and the variance for Quiz 2 is 3.9. From the histograms, we knew that Quiz 2 was more spread out, which we see in Quiz 2 having a larger variance.

? Text Exercise 2.6.7

Recall the two data sets from Text Exercise 2.6.5 (treat them as population data), which were indistinguishable using MAD from the mean. Compute the variance for each data set and explain how the variance can distinguish them while the MAD from the mean cannot.

$$\{-2,-2,2,2\}$$
$$\{-3,-1,1,3\}$$

Answer

Recall that both sets have a mean of 0. So, the squared deviations from the mean are just the values squared. The variance of the first data set is $\sigma_1^2 = \frac{4+4+4+4}{4} = 4$, and the variance of the second data set is $\sigma_1^2 = \frac{9+1+1+9}{4} = 5$. The two data sets are different using variance because the squaring action puts greater weight on values farther from the mean. While -3 and -2 are just one unit away from each other, the -3 contributes 9 to the sum in the variance while the -2 contributes only 4. And similarly, even though we did not see this in our example, if deviations are small, less than 1, their weight is even less by squaring.

We are more interested in the larger population when we have sample data. We use sample statistics to estimate population parameters. Statisticians have found that when calculating the average of squared deviations from the mean using sample data, the computation tends to underestimate the variance of the larger population significantly. Thinking intuitively, this makes sense as there is usually greater variability in a large group than in some subset of that group. Imagine the population was $1,2,3,4,5,6,7,8,9,10$; take a sample from this population, say, 2, $3,5,7$. Notice the sample is less dispersed than the population. The sample will, more often than not, have less variance than the population. If we want to estimate the population variance based on a sample, using a number larger than the value obtained from the above formula is better. Because of this, statisticians have adjusted the averaging process when dealing with sample data to result in a better inferential measure; the solution is to divide by $n-1$ rather than n. The **sample variance** is defined as follows.

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

Notice the use of s, a Latin letter, in s^2; the Latin letter again reminds us that we are dealing with sample data. In practice, variance is usually computed in a sample, so this formula is often used.

? Text Exercise 2.6.8

Let's take a concrete example. Consider a random sample of 10 quiz scores from Quiz 2. We constructed a random sample from the table on the previous problem: $\{5,5,5,6,7,7,7,8,10,10\}$. Calculate the sample variance for this sample data. Compare the sample variance to the population variance.

Answer

$$\bar{x} = \frac{5+5+5+6+7+7+7+8+10+10}{10} = \frac{70}{10} = 7$$

$$s^2 = \frac{[(5-7)^2+(5-7))^2+(5-7))^2+(6-7))^2+(7-7))^2+(7-7))^2+(7-7))^2+(8-7))^2+(10-7))^2+(10-7))^2]}{(10-1)}$$

$$= \frac{(4+4+4+1+0+0+0+1+9+9)}{9}$$

$$= \frac{32}{9} = 3+\frac{5}{9}$$

$$\approx 3.5556$$

The population variance σ^2 is 3.9 while the sample variance s^2 is about 3.5556.The values are off by about 0.3444,which is much closer than it would have been if we had divided by 10 rather than 9 in our computation. Without adjusting the average for sample variance, we would compute 3.2,which is 0.7 away from the population variance.

Standard Deviation

Variance is a powerful measure and is the basis for much of statistics. We struggle to interpret this value because of the squared units. Consequently, we introduce another measure closely related to the variance: the **standard deviation**. The standard deviation is simply the square root of the variance. The population standard deviation is the square root of the population variance, and the sample standard deviation is the square root of the sample variance. This makes the units on the measure the same as those of the original data; for example, if the original data was in cm, then the standard deviation will also be measured in cm. We can understand the standard deviation loosely as a measure of the distance a typical value is from the mean. Our natural choices of symbols are the bases for our two variances: σ for population and s for sample.

? Text Exercise 2.6.9

Compute the standard deviations for both Quiz 1 and Quiz 2.

Answer

Since we have already computed the variances for Quiz 1 and Quiz 2 in text exercise 2.6.5.We only need to take the square root of each.

$\sigma_1 = \sqrt{1.5} \approx 1.225$

$\sigma_2 = \sqrt{3.9} \approx 1.975$

? Text Exercise 2.6.10

What is the smallest value that standard deviation can take on? Construct a data set of 5 observations with such a standard deviation.

Answer

As the square root of variance, we must analyze the square root function and variance as a measure. Square roots return nonnegative values. The smallest nonnegative value is 0. The square root is equal to 0, $\sqrt{d} = 0$, only when the input d is 0. Can the variance of a data set be 0? An average is 0 only when the sum of the values is 0. We are adding up squared deviations, which are all nonnegative. The only way a sum of nonnegative values is 0 is if they are all 0. The deviation from the mean is 0 only when the observation is equal to the mean. There is no variability in the data values. An example of such a data set would be $\{1,1,1,1,1\}$.

It is worth noting that all of the measures discussed here (range, IQR, MAD, variance, standard deviation) are always non-negative. Moreover, they will all be 0 on constant data sets, like $\{1,1,1,1,1\}$; conversely, the range, MAD, variance, and standard deviation will not be 0 if there are at least two different data values.
