3. Summarizing Distributions

A. Central Tendency
    1. What is Central Tendency
    2. Measures of Central Tendency
    3. Median and Mean
    4. Additional Measures
    5. Comparing measures
B. Variability
    1. Measures of Variability
C. Shape
    1. Effects of Transformations
    2. Variance Sum Law I
D. Exercises

Descriptive statistics often involves using a few numbers to summarize a distribution. One important aspect of a distribution is where its center is located. Measures of central tendency are discussed first. A second aspect of a distribution is how spread out it is. In other words, how much the numbers in the distribution vary from one another. The second section describes measures of variability. Distributions can differ in shape. Some distributions are symmetric whereas others have long tails in just one direction. The third section describes measures of the shape of distributions. The final two sections concern (1) how transformations affect measures summarizing distributions and (2) the variance sum law, an important relationship involving a measure of variability.

What is Central Tendency?

by David M. Lane and Heidi Ziemer

Prerequisites

- Chapter 1: Distributions
- Chapter 2: Stem and Leaf Displays

Learning Objectives

1. Identify situations in which knowing the center of a distribution would be valuable
2. Give three different ways the center of a distribution can be defined
3. Describe how the balance is different for symmetric distributions than it is for asymmetric distributions.

What is "central tendency," and why do we want to know the central tendency of a group of scores? Let us first try to answer these questions intuitively. Then we will proceed to a more formal discussion.

Imagine this situation: You are in a class with just four other students, and the five of you took a 5-point pop quiz. Today your instructor is walking around the room, handing back the quizzes. She stops at your desk and hands you your paper. Written in bold black ink on the front is "3/5." How do you react? Are you happy with your score of 3 or disappointed? How do you decide? You might calculate your percentage correct, realize it is 60%, and be appalled. But it is more likely that when deciding how to react to your performance, you will want additional information. What additional information would you like?

If you are like most students, you will immediately ask your neighbors, "Whad'ja get?" and then ask the instructor, "How did the class do?" In other words, the additional information you want is how your quiz score compares to other students' scores. You therefore understand the importance of comparing your score to the class distribution of scores. Should your score of 3 turn out to be among the higher scores, then you'll be pleased after all. On the other hand, if 3 is among the lower scores in the class, you won't be quite so happy.

This idea of comparing individual scores to a distribution of scores is fundamental to statistics. So let's explore it further, using the same example (the pop quiz you took with your four classmates). Three possible outcomes are shown in Table 1. They are labeled "Dataset A," "Dataset B," and "Dataset C." Which of

the three datasets would make you happiest? In other words, in comparing your score with your fellow students' scores, in which dataset would your score of 3 be the most impressive?

In Dataset A, everyone's score is 3. This puts your score at the exact center of the distribution. You can draw satisfaction from the fact that you did as well as everyone else. But of course it cuts both ways: everyone else did just as well as you.

Table 1. Three possible datasets for the 5-point make-up quiz.

Student	Dataset A	Dataset B	Dataset C
You	3	3	3
John's	3	4	2
Maria's	3	4	2
Shareecia's	3	4	2
Luther's	3	5	1

Now consider the possibility that the scores are described as in Dataset B. This is a depressing outcome even though your score is no different than the one in Dataset A. The problem is that the other four students had higher grades, putting yours below the **center of the distribution**.

Finally, let's look at Dataset C. This is more like it! All of your classmates score lower than you so your score is above the center of the distribution.

Now let's change the example in order to develop more insight into the center of a distribution. Figure 1 shows the results of an experiment on memory for chess positions. Subjects were shown a chess position and then asked to reconstruct it on an empty chess board. The number of pieces correctly placed was recorded. This was repeated for two more chess positions. The scores represent the total number of chess pieces correctly placed for the three chess positions. The maximum possible score was 89.

	8	05
	7	156
	6	233
	5	168
330	4	06
9420	3	
622	2	

Figure 1. Back-to-back stem and leaf display. The left side shows the memory scores of the non-players. The right side shows the scores of the tournament players.

Two groups are compared. On the left are people who don't play chess. On the right are people who play a great deal (tournament players). It is clear that the location of the center of the distribution for the non-players is much lower than the center of the distribution for the tournament players.

We're sure you get the idea now about the center of a distribution. It is time to move beyond intuition. We need a formal definition of the center of a distribution. In fact, we'll offer you three definitions! This is not just generosity on our part. There turn out to be (at least) three different ways of thinking about the center of a distribution, all of them useful in various contexts. In the remainder of this section we attempt to communicate the idea behind each concept. In the succeeding sections we will give statistical measures for these concepts of central tendency.

Definitions of Center

Now we explain the three different ways of defining the center of a distribution. All three are called measures of central tendency.

Balance Scale

One definition of central tendency is the point at which the distribution is in balance. Figure 2 shows the distribution of the five numbers 2, 3, 4, 9, 16 placed upon a balance scale. If each number weighs one pound, and is placed at its

position along the number line, then it would be possible to balance them by placing a fulcrum at 6.8.

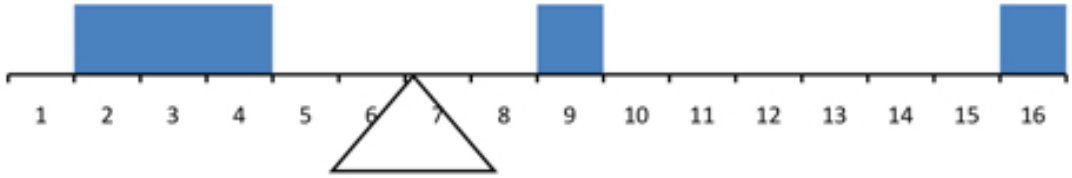


Figure 2. A balance scale.

For another example, consider the distribution shown in Figure 3. It is balanced by placing the fulcrum in the geometric middle.

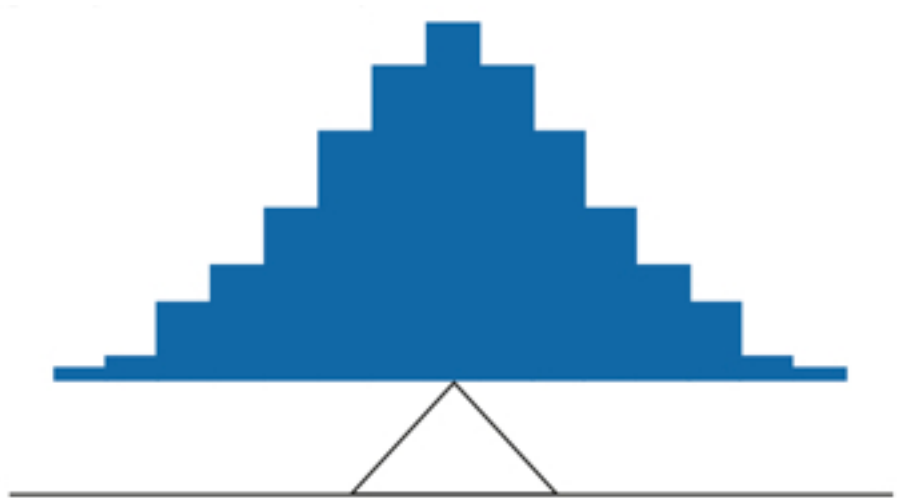


Figure 3. A distribution balanced on the tip of a triangle.

Figure 4 illustrates that the same distribution can't be balanced by placing the fulcrum to the left of center.

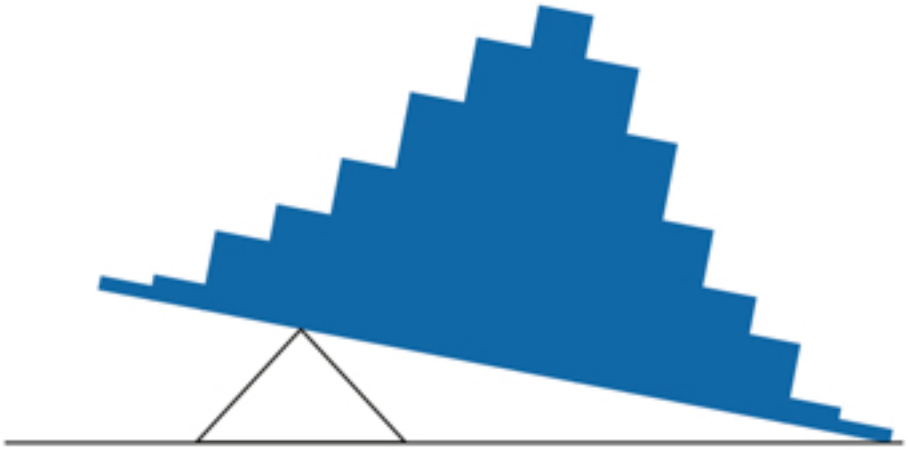


Figure 4. The distribution is not balanced.

Figure 5 shows an asymmetric distribution. To balance it, we cannot put the fulcrum halfway between the lowest and highest values (as we did in Figure 3). Placing the fulcrum at the “half way” point would cause it to tip towards the left.

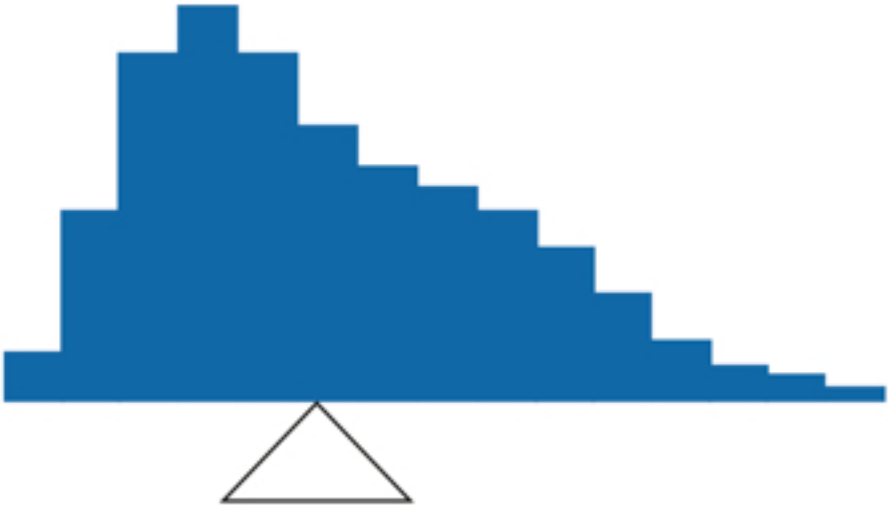


Figure 5. An asymmetric distribution balanced on the tip of a triangle.

The balance point defines one sense of a distribution's center.

Smallest Absolute Deviation

Another way to define the center of a distribution is based on the concept of the sum of the absolute deviations (differences). Consider the distribution made up of the five numbers 2, 3, 4, 9, 16. Let's see how far the distribution is from 10

(picking a number arbitrarily). Table 2 shows the sum of the absolute deviations of these numbers from the number 10.

Table 2. An example of the sum of absolute deviations

Values	Absolute Deviations from 10
2	8
3	7
4	6
9	1
16	6
Sum	**28**

The first row of the table shows that the absolute value of the difference between 2 and 10 is 8; the second row shows that the absolute difference between 3 and 10 is 7, and similarly for the other rows. When we add up the five absolute deviations, we get 28. So, the sum of the absolute deviations from 10 is 28. Likewise, the sum of the absolute deviations from 5 equals $3 + 2 + 1 + 4 + 11 = 21$. So, the sum of the absolute deviations from 5 is smaller than the sum of the absolute deviations from 10. In this sense, 5 is closer, overall, to the other numbers than is 10.

We are now in a position to define a second measure of central tendency, this time in terms of absolute deviations. Specifically, according to our second definition, the center of a distribution is the number for which the sum of the absolute deviations is smallest. As we just saw, the sum of the absolute deviations from 10 is 28 and the sum of the absolute deviations from 5 is 21. Is there a value for which the sum of the absolute deviations is even smaller than 21? Yes. For these data, there is a value for which the sum of absolute deviations is only 20. See if you can find it.

Smallest Squared Deviation

We shall discuss one more way to define the center of a distribution. It is based on the concept of the sum of squared deviations (differences). Again, consider the distribution of the five numbers 2, 3, 4, 9, 16. Table 3 shows the sum of the squared deviations of these numbers from the number 10.

Table 3. An example of the sum of squared deviations.

Values	Squared Deviations from 10
2	64
3	49
4	36
9	1
16	36
Sum	**186**

The first row in the table shows that the squared value of the difference between 2 and 10 is 64; the second row shows that the squared difference between 3 and 10 is 49, and so forth. When we add up all these squared deviations, we get 186. Changing the target from 10 to 5, we calculate the sum of the squared deviations from 5 as 9 + 4 + 1 + 16 + 121 = 151. So, the sum of the squared deviations from 5 is smaller than the sum of the squared deviations from 10. Is there a value for which the sum of the squared deviations is even smaller than 151? Yes, it is possible to reach 134.8. Can you find the target number for which the sum of squared deviations is 134.8?

The target that minimizes the sum of squared deviations provides another useful definition of central tendency (the last one to be discussed in this section). It can be challenging to find the value that minimizes this sum.

Measures of Central Tendency

by David M. Lane

Prerequisites
- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: Central Tendency

Learning Objectives
1. Compute mean
2. Compute median
3. Compute mode

In the previous section we saw that there are several ways to define central tendency. This section defines the three most common measures of central tendency: the mean, the median, and the mode. The relationships among these measures of central tendency and the definitions given in the previous section will probably not be obvious to you.

This section gives only the basic definitions of the mean, median and mode. A further discussion of the relative merits and proper applications of these statistics is presented in a later section.

Arithmetic Mean

The arithmetic mean is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers. The symbol "μ" is used for the mean of a population. The symbol "M" is used for the mean of a sample. The formula for μ is shown below:

$$\mu = \frac{\sum X}{N}$$

where ΣX is the sum of all the numbers in the population andN is the number of numbers in the population.

The formula for M is essentially identical:

$$M = \frac{\sum X}{N}$$

where ΣX is the sum of all the numbers in the sample and N is the number of numbers in the sample.

As an example, the mean of the numbers 1, 2, 3, 6, 8 is 20/5 = 4 regardless of whether the numbers constitute the entire population or just a sample from the population.

Table 1 shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season. The mean number of touchdown passes thrown is 20.4516 as shown below.

$$\mu = \frac{\sum X}{N} = \frac{634}{31} = 20.4516$$

Table 1. Number of touchdown passes.

37, 33, 33, 32, 29, 28,
28, 23, 22, 22, 22, 21,
21, 21, 20, 20, 19, 19,
18, 18, 18, 18, 16, 15,
14, 14, 14, 12, 12, 9, 6

Although the arithmetic mean is not the only "mean" (there is also a geometric mean), it is by far the most commonly used. Therefore, if the term "mean" is used without specifying whether it is the arithmetic mean, the geometric mean, or some other mean, it is assumed to refer to the arithmetic mean.

Median

The median is also a frequently used measure of central tendency. The median is the midpoint of a distribution: the same number of scores is above the median as below it. For the data in Table 1, there are 31 scores. The 16th highest score (which equals 20) is the median because there are 15 scores below the 16th score and 15

scores above the 16th score. The median can also be thought of as the 50th percentile.

Computation of the Median

When there is an odd number of numbers, the median is simply the middle number. For example, the median of 2, 4, and 7 is 4. When there is an even number of numbers, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 7, 12 is:

$$\frac{(4+7)}{2} = 5.5$$

When there are numbers with the same values, then the formula for the third definition of the 50th percentile should be used.

Mode

The mode is the most frequently occurring value. For the data in Table 1, the mode is 18 since more teams (4) had 18 touchdown passes than any other number of touchdown passes. With continuous data, such as response time measured to many decimals, the frequency of each value is one since no two scores will be exactly the same (see discussion of continuous variables). Therefore the mode of continuous data is normally computed from a grouped frequency distribution. Table 2 shows a grouped frequency distribution for the target response time data. Since the interval with the highest frequency is 600-700, the mode is the middle of that interval (650).

Table 2. Grouped frequency distribution

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Median and Mean

by David M. Lane

Prerequisites

- Chapter 3: What is Central Tendency
- Chapter 3: Measures of Central Tendency

Learning Objectives

1. State when the mean and median are the same
2. State whether it is the mean or median that minimizes the mean absolute deviation
3. State whether it is the mean or median that minimizes the mean squared deviation
4. State whether it is the mean or median that is the balance point on a balance scale

In the section "What is central tendency," we saw that the center of a distribution could be defined three ways: (1) the point on which a distribution would balance, (2) the value whose average absolute deviation from all the other values is minimized, and (3) the value whose squared difference from all the other values is minimized. The mean is the point on which a distribution would balance, the median is the value that minimizes the sum of absolute deviations, and the mean is the value that minimizes the sum of the squared deviations.

Table 1 shows the absolute and squared deviations of the numbers 2, 3, 4, 9, and 16 from their median of 4 and their mean of 6.8. You can see that the sum of absolute deviations from the median (20) is smaller than the sum of absolute deviations from the mean (22.8). On the other hand, the sum of squared deviations from the median (174) is larger than the sum of squared deviations from the mean (134.8).

Table 1. Absolute and squared deviations from the median of 4 and the mean of 6.8.

Value	Absolute Deviation from Median	Absolute Deviation from Mean	Squared Deviation from Median	Squared Deviation from Mean
2	2	4.8	4	23.04
3	1	3.8	1	14.44
4	0	2.8	0	7.84
9	5	2.2	25	4.84
16	12	9.2	144	84.64
Total	20	22.8	174	134.8

Figure 1 shows that the distribution balances at the mean of 6.8 and not at the median of 4. The relative advantages and disadvantages of the mean and median are discussed in the section "Comparing Measures" later in this chapter.

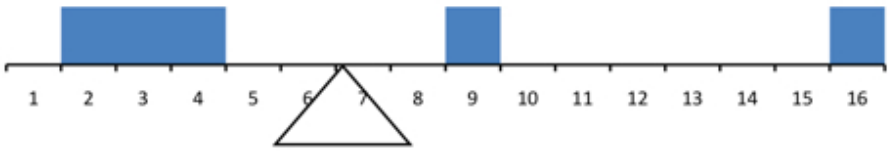


Figure 1. The distribution balances at the mean of 6.8 and not at the median of 4.0.

When a distribution is symmetric, then the mean and the median are the same. Consider the following distribution: 1, 3, 4, 5, 6, 7, 9. The mean and median are both 5. The mean, median, and mode are identical in the bell-shaped normal distribution.

Additional Measures of Central Tendency

by David M. Lane

Prerequisites

- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: What is Central Tendency
- Chapter 3: Measures of Central Tendency
- Chapter 3: Mean and Median

Learning Objectives

1. Compute the trimean
2. Compute the geometric mean directly
3. Compute the geometric mean using logs
4. Use the geometric to compute annual portfolio returns
5. Compute a trimmed mean

Although the mean, median, and mode are by far the most commonly used measures of central tendency, they are by no means the only measures. This section defines three additional measures of central tendency: the trimean, the geometric mean, and the trimmed mean. These measures will be discussed again in the section "Comparing Measures of Central Tendency."

Trimean

The trimean is a weighted average of the 25th percentile, the 50th percentile, and the 75th percentile. Letting P25 be the 25th percentile, P50 be the 50th and P75 be the 75th percentile, the formula for the trimean is:

$$Trimean = \frac{(P25 + 2P50 + P75)}{4}$$

Consider the data in Table 2. The 25th percentile is 15, the 50th is 20 and the 75th percentile is 23.

Table 1. Number of touchdown passes.

37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20, 19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6

Table 2. Percentiles.

Percentile	Value
25	15
50	20
75	23

The trimean is therefore :

$$\frac{(15 + 2 \times 20 + 23)}{4} = \frac{78}{4} = 19.5$$

Geometric Mean

The geometric mean is computed by multiplying all the numbers together and then taking the nth root of the product. For example, for the numbers 1, 10, and 100, the product of all the numbers is: 1 x 10 x 100 = 1,000. Since there are three numbers, we take the cubed root of the product (1,000) which is equal to 10. The formula for the geometric mean is therefore

$$\left(\prod X\right)^{\frac{1}{N}}$$

where the symbol Π means to multiply. Therefore, the equation says to multiply all the values of X and then raise the result to the 1/Nth power. Raising a value to the 1/Nth power is, of course, the same as taking the Nth root of the value. In this case, $1000^{1/3}$ is the cube root of 1,000.

The geometric mean has a close relationship with logarithms. Table 3 shows the logs (base 10) of these three numbers. The arithmetic mean of the three logs is 1. The anti-log of this arithmetic mean of 1 is the geometric mean. The anti-log of 1 is $10^1 = 10$. Note that the geometric mean only makes sense if all the numbers are positive.

Table 3. Logarithms.

X	Log10(X)
1	0
10	1
100	2

The geometric mean is an appropriate measure to use for averaging rates. For example, consider a stock portfolio that began with a value of $1,000 and had annual returns of 13%, 22%, 12%, -5%, and -13%. Table 4 shows the value after each of the five years.

Table 4. Portfolio Returns

Year	Return	Value
1	13%	1,130
2	22%	1,379
3	12%	1,544
4	-5%	1,467
5	-13%	1,276

The question is how to compute average annual rate of return. The answer is to compute the geometric mean of the returns. Instead of using the percents, each return is represented as a multiplier indicating how much higher the value is after the year. This multiplier is 1.13 for a 13% return and 0.95 for a 5% loss. The multipliers for this example are 1.13, 1.22, 1.12, 0.95, and 0.87. The geometric mean of these multipliers is 1.05. Therefore, the average annual rate of return is 5%. Table 5 shows how a portfolio gaining 5% a year would end up with the same value ($1,276) as shown in Table 4.

Table 5. Portfolio Returns

Year	Return	Value

1	5%	1,050
2	5%	1,103
3	5%	1,158
4	5%	1,216
5	5%	1,276

Trimmed Mean

To compute a *trimmed mean*, you remove some of the higher and lower scores and compute the mean of the remaining scores. A mean trimmed 10% is a mean computed with 10% of the scores trimmed off: 5% from the bottom and 5% from the top. A mean trimmed 50% is computed by trimming the upper 25% of the scores and the lower 25% of the scores and computing the mean of the remaining scores. The trimmed mean is similar to the median which, in essence, trims the upper 49+% and the lower 49+% of the scores. Therefore the trimmed mean is a hybrid of the mean and the median. To compute the mean trimmed 20% for the touchdown pass data shown in Table 1, you remove the lower 10% of the scores (6, 9, and 12) as well as the upper 10% of the scores (33, 33, and 37) and compute the mean of the remaining 25 scores. This mean is 20.16.

Comparing Measures of Central Tendency

by David M. Lane

Prerequisites

- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: What is Central Tendency
- Chapter 3: Measures of Central Tendency
- Chapter 3: Mean and Median

Learning Objectives

1. Understand how the difference between the mean and median is affected by skew
2. State how the measures differ in symmetric distributions
3. State which measure(s) should be used to describe the center of a skewed distribution

How do the various measures of central tendency compare with each other? For symmetric distributions, the mean, median, trimean, and trimmed mean are equal, as is the mode except in bimodal distributions. Differences among the measures occur with skewed distributions. Figure 1 shows the distribution of 642 scores on an introductory psychology test. Notice this distribution has a slight positive skew.

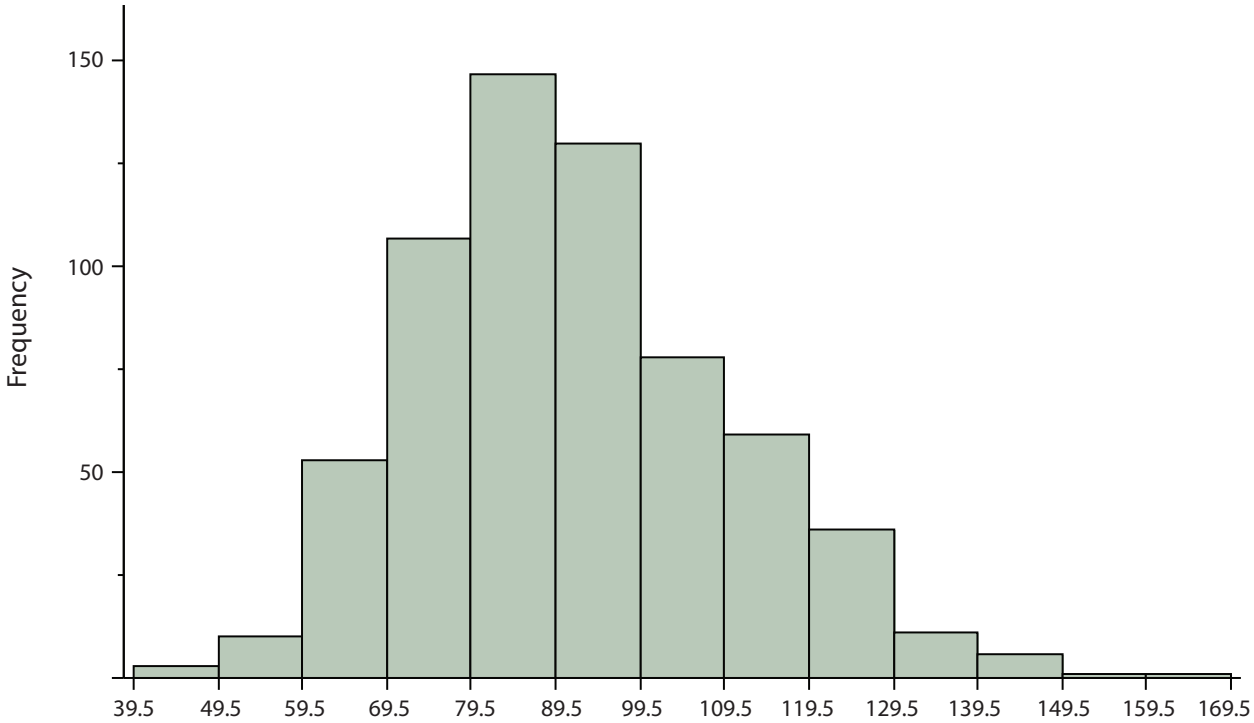


Figure 1. A distribution with a positive skew.

Measures of central tendency are shown in Table 1. Notice they do not differ greatly, with the exception that the mode is considerably lower than the other measures. When distributions have a positive skew, the mean is typically higher than the median, although it may not be in bimodal distributions. For these data, the mean of 91.58 is higher than the median of 90. Typically the trimean and trimmed mean will fall between the median and the mean, although in this case, the trimmed mean is slightly lower than the median. The geometric mean is lower than all measures except the mode.

Table 1. Measures of central tendency for the test scores.

Measure	Value
Mode	84.00
Median	90.00
Geometric Mean	89.70
Trimean	90.25
Mean trimmed 50%	89.81
Mean	91.58

The distribution of baseball salaries (in 1994) shown in Figure 2 has a much more pronounced skew than the distribution in Figure 1.

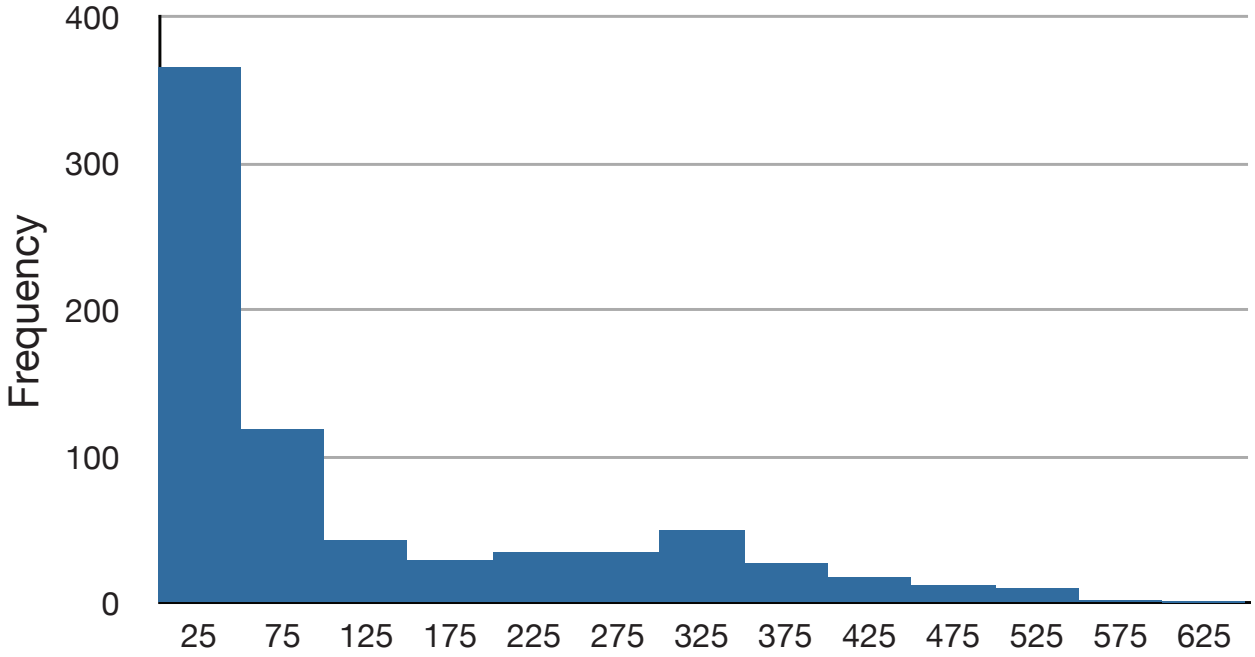


Figure 2. A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars).

Table 2 shows the measures of central tendency for these data. The large skew results in very different values for these measures. No single measure of central tendency is sufficient for data such as these. If you were asked the very general question: "So, what do baseball players make?" and answered with the mean of $1,183,000, you would not have told the whole story since only about one third of baseball players make that much. If you answered with the mode of $250,000 or

the median of $500,000, you would not be giving any indication that some players make many millions of dollars. Fortunately, there is no need to summarize a distribution with a single number. When the various measures differ, our opinion is that you should report the mean, median, and either the trimean or the mean trimmed 50%. Sometimes it is worth reporting the mode as well. In the media, the median is usually reported to summarize the center of skewed distributions. You will hear about median salaries and median prices of houses sold, etc. This is better than reporting only the mean, but it would be informative to hear more statistics.

Table 2. Measures of central tendency for baseball salaries (in thousands of dollars).

Measure	Value
Mode	250
Median	500
Geometric Mean	555
Trimean	792
Mean trimmed 50%	619
Mean	1,183

Measures of Variability

by David M. Lane

Prerequisites

- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: Measures of Central Tendency

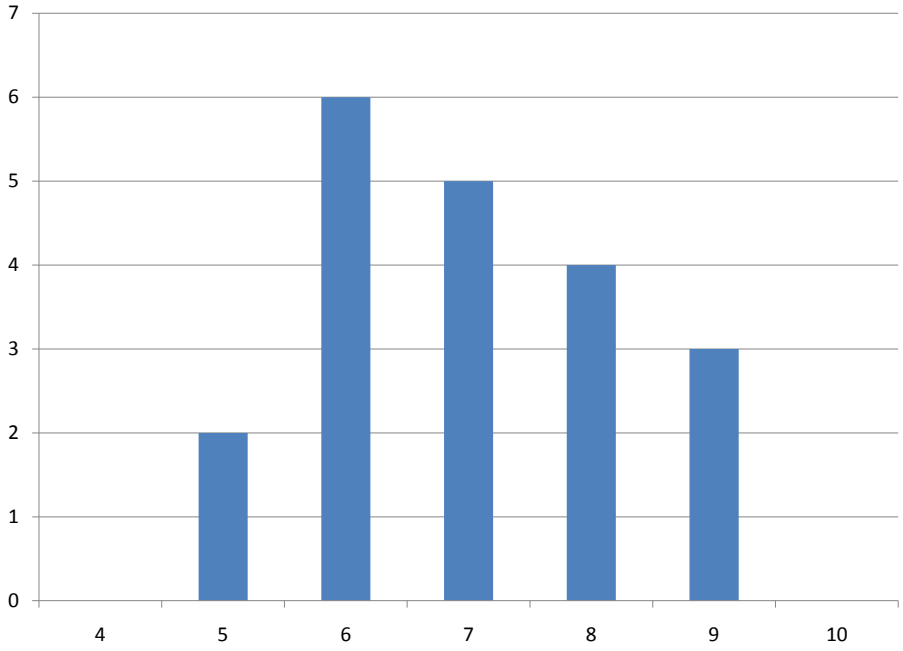
Learning Objectives

1. Determine the relative variability of two distributions
2. Compute the range
3. Compute the inter-quartile range
4. Compute the variance in the population
5. Estimate the variance from a sample
6. Compute the standard deviation from the variance

What is Variability?

Variability refers to how "spread out" a group of scores is. To see what we mean by spread out, consider graphs in Figure 1. These graphs represent the scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, you can see that the distributions are quite different. Specifically, the scores on Quiz 1 are more densely packed and those on Quiz 2 are more spread out. The differences among students were much greater on Quiz 2 than on Quiz 1.

Quiz 1



Quiz 2

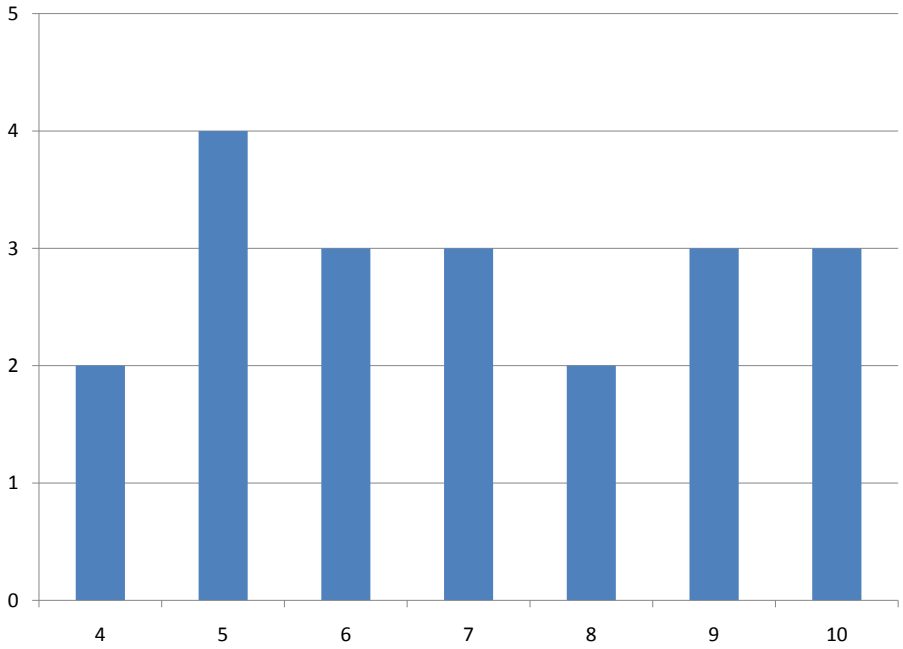


Figure 1. Bar charts of two quizzes.

The terms variability, spread, and dispersion are synonyms, and refer to how spread out a distribution is. Just as in the section on central tendency where we discussed measures of the center of a distribution of scores, in this chapter we will

discuss measures of the variability of a distribution. There are four frequently used measures of variability: range, interquartile range, variance, and standard deviation. In the next few paragraphs, we will look at each of these four measures of variability in more detail.

Range

The range is the simplest measure of variability to calculate, and one you have probably encountered many times in your life. The range is simply the highest score minus the lowest score. Let's take a few examples. What is the range of the following group of numbers: 10, 2, 5, 6, 7, 3, 4? Well, the highest number is 10, and the lowest number is 2, so 10 - 2 = 8. The range is 8. Let's take another example. Here's a dataset with 10 numbers: 99, 45, 23, 67, 45, 91, 82, 78, 62, 51. What is the range? The highest number is 99 and the lowest number is 23, so 99 - 23 equals 76; the range is 76. Now consider the two quizzes shown in Figure 1. On Quiz 1, the lowest score is 5 and the highest score is 9. Therefore, the range is 4. The range on Quiz 2 was larger: the lowest score was 4 and the highest score was 10. Therefore the range is 6.

Interquartile Range

The interquartile range (IQR) is the range of the middle 50% of the scores in a distribution. It is computed as follows:

```
IQR = 75th percentile - 25th percentile
```

For Quiz 1, the 75th percentile is 8 and the 25th percentile is 6. The interquartile range is therefore 2. For Quiz 2, which has greater spread, the 75th percentile is 9, the 25th percentile is 5, and the interquartile range is 4. Recall that in the discussion of box plots, the 75th percentile was called the upper hinge and the 25th percentile was called the lower hinge. Using this terminology, the interquartile range is referred to as the H-spread.

A related measure of variability is called the semi-interquartile range. The semi-interquartile range is defined simply as the interquartile range divided by 2. If a distribution is symmetric, the median plus or minus the semi-interquartile range contains half the scores in the distribution.

Variance

Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution. Using the mean as the measure of the middle of the distribution, the variance is defined as the average squared difference of the scores from the mean. The data from Quiz 1 are shown in Table 1. The mean score is 7.0. Therefore, the column "Deviation from Mean" contains the score minus 7. The column "Squared Deviation" is simply the previous column squared.

Table 1. Calculation of Variance for Quiz 1 scores.

Scores	Deviation from Mean	Squared Deviation
9	2	4
9	2	4
9	2	4
8	1	1
8	1	1
8	1	1
8	1	1
7	0	0
7	0	0
7	0	0
7	0	0
7	0	0
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1
5	-2	4
5	-2	4
Means		
7	0	1.5

One thing that is important to notice is that the mean deviation from the mean is 0. This will always be the case. The mean of the squared deviations is 1.5. Therefore, the variance is 1.5. Analogous calculations with Quiz 2 show that its variance is 6.7. The formula for the variance is:

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

where σ^2 is the variance, μ is the mean, and N is the number of numbers. For Quiz 1, $\mu = 7$ and $N = 20$.

If the variance in a sample is used to estimate the variance in a population, then the previous formula underestimates the variance and the following formula should be used:

$$s^2 = \frac{\sum(X-M)^2}{N-1}$$

where s^2 is the estimate of the variance and M is the sample mean. Note that M is the mean of a sample taken from a population with a mean of μ. Since, in practice, the variance is usually computed in a sample, this formula is most often used.

Let's take a concrete example. Assume the scores 1, 2, 4, and 5 were sampled from a larger population. To estimate the variance in the population you would compute s^2 as follows:

$$M = \frac{1+2+3+4+5}{4} = \frac{12}{4} = 3$$

$$s^2 = \frac{(1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2}{4-1} = \frac{4+1+1+4}{3} = \frac{10}{3} = 3.333$$

There are alternate formulas that can be easier to use if you are doing your calculations with a hand calculator:

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

and

$$s^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}$$

For this example,

$$\left(\sum X\right)^2 = \frac{(1+2+4+5)^2}{4} = \frac{144}{4} = 36$$

$$\sigma^2 = \frac{(46-36)}{4} = 2.5$$

$$s^2 = \frac{(46-36)}{3} = 3.333$$

as with the other formula.

Standard Deviation

The standard deviation is simply the square root of the variance. This makes the standard deviations of the two quiz distributions 1.225 and 2.588. The standard deviation is an especially useful measure of variability when the distribution is normal or approximately normal (see Chapter 7) because the proportion of the distribution within a given number of standard deviations from the mean can be calculated. For example, 68% of the distribution is within one standard deviation of the mean and approximately 95% of the distribution is within two standard deviations of the mean. Therefore, if you had a normal distribution with a mean of 50 and a standard deviation of 10, then 68% of the distribution would be between 50 - 10 = 40 and 50 +10 =60. Similarly, about 95% of the distribution would be between 50 - 2 x 10 = 30 and 50 + 2 x 10 = 70. The symbol for the population standard deviation is σ; the symbol for an estimate computed in a sample is s. Figure 2 shows two normal distributions. The red distribution has a mean of 40 and a standard deviation of 5; the blue distribution has a mean of 60 and a standard deviation of 10. For the red distribution, 68% of the distribution is between 45 and 55; for the blue distribution, 68% is between 40 and 60.

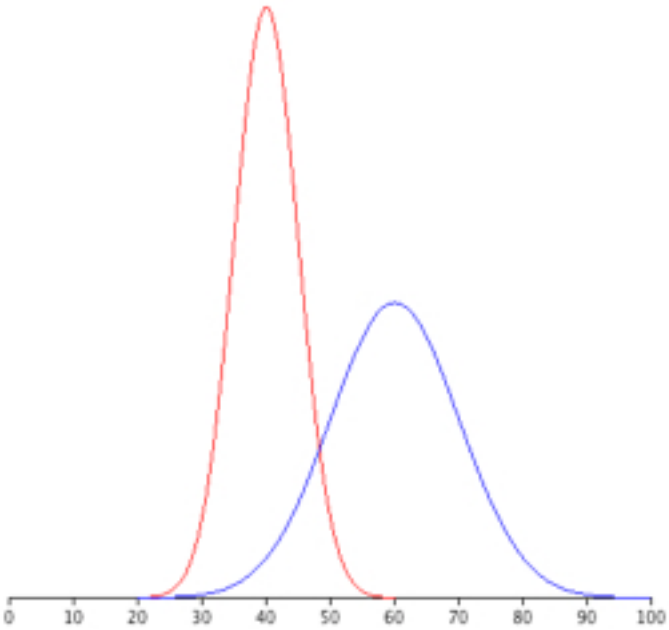


Figure 2. Normal distributions with standard deviations of 5 and 10.

Shapes of Distributions

by David M. Lane

Prerequisites

- Chapter 1: Distributions
- Chapter 3: Measures of Central Tendency
- Chapter 3: Variability

Learning Objectives

1. Compute skew using two different formulas
2. Compute kurtosis

We saw in the section on distributions in Chapter 1 that shapes of distributions can differ in skew and/or kurtosis. This section presents numerical indexes of these two measures of shape.

Skew

Figure 1 shows a distribution with a very large positive skew. Recall that distributions with positive skew have tails that extend to the right.

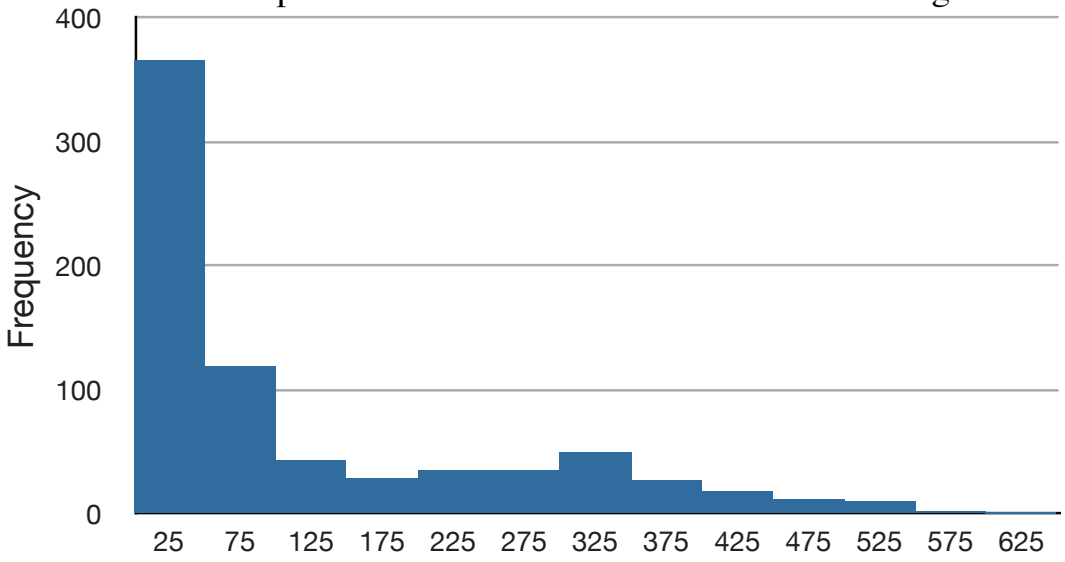


Figure 1. A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars).

Distributions with positive skew normally have larger means than medians. The mean and median of the baseball salaries shown in Figure 1 are $1,183,417 and $500,000 respectively. Thus, for this highly-skewed distribution, the mean is more than twice as high as the median. The relationship between skew and the relative size of the mean and median lead the statistician Pearson to propose the following simple and convenient numerical index of skew:

$$\frac{3(Mean - Median)}{\sigma}$$

The standard deviation of the baseball salaries is 1,390,922. Therefore, Pearson's measure of skew for this distribution is 3(1,183,417 - 500,000)/1,390,922 = 1.47.

Just as there are several measures of central tendency, there is more than one measure of skew. Although Pearson's measure is a good one, the following measure is more commonly used. It is sometimes referred to as the third moment about the mean.

$$\sum \frac{(X - \mu)^3}{\sigma^3}$$

Kurtosis

The following measure of kurtosis is similar to the definition of skew. The value "3" is subtracted to define "no kurtosis" as the kurtosis of a normal distribution. Otherwise, a normal distribution would have a kurtosis of 3.

$$\sum \frac{(X - \mu)^4}{\sigma^4} - 3$$

Effects of Linear Transformations

by David M. Lane

Prerequisites

• Chapter 1: Linear Transformations

Learning Objectives

1. Define a linear transformation
2. Compute the mean of a transformed variable
3. Compute the variance of a transformed variable

This section covers the effects of linear transformations on measures of central tendency and variability. Let's start with an example we saw before in the section that defined linear transformation: temperatures of cities. Table 1 shows the temperatures of 5 cities.

Table 1. Temperatures in 5 cities on 11/16/2002.

City	Degrees Fahrenheit	Degrees Centigrade
Houston	54	12.22
Chicago	37	2.78
Minneapolis	31	-0.56
Miami	78	25.56
Phoenix	70	21.11
Mean	54.000	12.220
Median	54.000	12.220
Variance	330.00	101.852
SD	18.166	10.092

Recall that to transform the degrees Fahrenheit to degrees Centigrade, we use the formula

$$C = 0.55556F - 17.7778$$

which means we multiply each temperature Fahrenheit by 0.556 and then subtract 17.7778. As you might have expected, you multiply the mean temperature in Fahrenheit by 0.556 and then subtract 17.778 to get the mean in Centigrade. That is, (0.556)(54) - 17.7778 = 12.22. The same is true for the median. Note that this

relationship holds even if the mean and median are not identical as they are in Table 1.

The formula for the standard deviation is just as simple: the standard deviation in degrees Centigrade is equal to the standard deviation in degrees Fahrenheit times 0.556. Since the variance is the standard deviation squared, the variance in degrees Centigrade is equal to 0.556^2 times the variance in degrees Fahrenheit.

To sum up, if a variable X has a mean of μ, a standard deviation of σ, and a variance of σ^2, then a new variable Y created using the linear transformation

$$Y = bX + A$$

will have a mean of $b\mu+A$, a standard deviation of $b\sigma$, and a variance of $b^2\sigma^2$.

It should be noted that the term "linear transformation" is defined differently in the field of linear algebra. For details, follow this link.

Variance Sum Law I

by David M. Lane

Prerequisites
- Chapter 3: Variance

Learning Objectives
1. Compute the variance of the sum of two uncorrelated variables
2. Compute the variance of the difference between two uncorrelated variables

As you will see in later sections, there are many occasions in which it is important to know the variance of the sum of two variables. Consider the following situation: (a) you have two populations, (b) you sample one number from each population, and (c) you add the two numbers together. The question is, "What is the variance of this sum?" For example, suppose the two populations are the populations of 8-year old males and 8-year-old females in Houston, Texas, and that the variable of interest is memory span. You repeat the following steps thousands of times: (1) sample one male and one female, (2) measure the memory span of each, and (3) sum the two memory spans. After you have done this thousands of times, you compute the variance of the sum. It turns out that the variance of this sum can be computed according to the following formula:

$$\sigma^2_{sum} = \sigma^2_M + \sigma^2_F$$

where the first term is the variance of the sum, the second term is the variance of the males and the third term is the variance of the females. Therefore, if the variances on the memory span test for the males and females respectively were 0.9 and 0.8, respectively, then the variance of the sum would be 1.7.

The formula for the variance of the difference between the two variables (memory span in this example) is shown below. Notice that the expression for the difference is the same as the formula for the sum.

$$\sigma^2_{difference} = \sigma^2_M + \sigma^2_F$$

More generally, the variance sum law can be written as follows:

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$$

which is read: “The variance of X plus or minus Y is equal to the variance of X plus the variance of Y.”

These formulas for the sum and difference of variables given above only apply when the variables are independent.

In this example, we have thousands of randomly-paired scores. Since the scores are paired randomly, there is no relationship between the memory span of one member of the pair and the memory span of the other. Therefore the two scores are independent. Contrast this situation with one in which thousands of people are sampled and two measures (such as verbal and quantitative SAT) are taken from each. In this case, there would be a relationship between the two variables since higher scores on the verbal SAT are associated with higher scores on the quantitative SAT (although there are many examples of people who score high on one test and low on the other). Thus the two variables are not independent and the variance of the total SAT score would not be the sum of the variances of the verbal SAT and the quantitative SAT. The general form of the variance sum law is presented in a section in the chapter on correlation.

Statistical Literacy

by David M. Lane

Prerequisites

- Chapter 3: Median and Mean

The playbill for the Alley Theatre in Houston wants to appeal to advertisers. They reported the mean household income and the median age of theatergoers.

What do you think?

What might have guided their choice of the mean or median?

> It is likely that they wanted to emphasize that theatergoers had high income but de-emphasize how old they are. The distributions of income and age of theatergoers probably have positive skew. Therefore the mean is probably higher than the median, which results in higher income and lower age than if the median household income and mean age had been presented.

Exercises

Prerequisites

- All material presented in the Summarizing Distributions chapter

1. Make up a dataset of 12 numbers with a positive skew. Use a statistical program to compute the skew. Is the mean larger than the median as it usually is for distributions with a positive skew? What is the value for skew?

2. Repeat Problem 1 only this time make the dataset have a negative skew.

3. Make up three data sets with 5 numbers each that have:

   (a) the same mean but different standard deviations.

   (b) the same mean but different medians.

   (c) the same median but different means.

4. Find the mean and median for the following three variables:

A	B	C
8	4	6
5	4	2
7	6	3
1	3	4
3	4	1

5. A sample of 30 distance scores measured in yards has a mean of 10, a variance of 9, and a standard deviation of 3 (a) You want to convert all your distances from yards to feet, so you multiply each score in the sample by 3. What are the new mean, variance, and standard deviation? (b) You then decide that you only want to look at the distance past a certain point. Thus, after multiplying the original scores by 3, you decide to subtract 4 feet from each of the scores. Now what are the new mean, variance, and standard deviation?

6. You recorded the time in seconds it took for 8 participants to solve a puzzle. These times appear below. However, when the data was entered into the statistical program, the score that was supposed to be 22.1 was entered as 21.2.

You had calculated the following measures of central tendency: the mean, the median, and the mean trimmed 25%. Which of these measures of central tendency will change when you correct the recording error?

Time (seconds)
15.2
18.8
19.3
19.7
20.2
21.8
22.1
29.4

7. For the test scores in question #6, which measures of variability (range, standard deviation, variance) would be changed if the 22.1 data point had been erroneously recorded as 21.2?

8. You know the minimum, the maximum, and the 25th, 50th, and 75th percentiles of a distribution. Which of the following measures of central tendency or variability can you determine?

   mean, median, mode, trimean, geometric mean, range, interquartile range, variance, standard deviation

9. For the numbers 1, 3, 4, 6, and 12:

   Find the value (v) for which $\Sigma(X-v)^2$ is minimized.

   Find the value (v) for which $\Sigma|x-v|$ is minimized.

10. Your younger brother comes home one day after taking a science test. He says that some- one at school told him that "60% of the students in the class scored above the median test grade." What is wrong with this statement? What if he had said "60% of the students scored below the mean?"

11. An experiment compared the ability of three groups of participants to remember briefly- presented chess positions. The data are shown below. The numbers represent the number of pieces correctly remembered from three chess

positions. Compare the performance of each group. Consider spread as well as central tendency.

Non-players	Beginners	Tournament players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

12. True/False: A bimodal distribution has two modes and two medians.

13. True/False: The best way to describe a skewed distribution is to report the mean.

14. True/False: When plotted on the same graph, a distribution with a mean of 50 and a standard deviation of 10 will look more spread out than will a distribution with a mean of 60 and a standard deviation of 5.

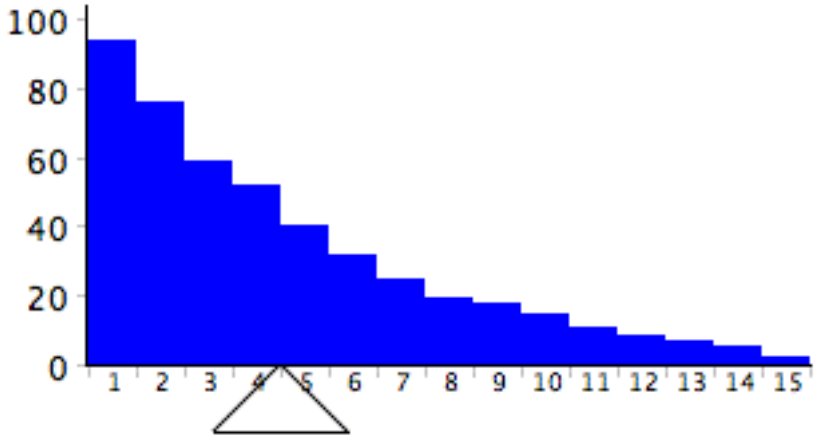
15. Compare the mean, median, trimean in terms of their sensitivity to extreme scores.

16. If the mean time to respond to a stimulus is much higher than the median time to respond, what can you say about the shape of the distribution of response times?

17. A set of numbers is transformed by taking the log base 10 of each number. The mean of the transformed data is 1.65. What is the geometric mean of the untransformed data?

18. Which measure of central tendency is most often used for returns on investment?

19. The histogram is in balance on the fulcrum. What are the mean, median, and mode of the distribution (approximate where necessary)?



Questions from Case Studies

Angry Moods (AM) case study

20. (AM) Does Anger-Out have a positive skew, a negative skew, or no skew?

21. (AM) What is the range of the Anger-In scores? What is the interquartile range?

22. (AM) What is the overall mean Control-Out score? What is the mean Control-Out score for the athletes? What is the mean Control-Out score for the non-athletes?

23. (AM) What is the variance of the Control-In scores for the athletes? What is the variance of the Control-In scores for the non-athletes?

Flatulence (F) case study

24. (F) Based on a histogram of the variable "perday", do you think the mean or median of this variable is larger? Calculate the mean and median to see if you are right.

Stroop (S) case study

25.(S) Compute the mean for “words”.

26. (S#2) Compute the mean and standard deviation for “colors”.

Physicians’ Reactions (PR) case study

27.(PR) What is the mean expected time spent for the average-weight patients? What is the mean expected time spent for the overweight patients?

28.(PR) What is the difference in means between the groups? By approximately how many standard deviations do the means differ?

Smiles and Leniency (SL) case study

29.(SL) Find the mean, median, standard deviation, and interquartile range for the leniency scores of each of the four groups.

ADHD Treatment (AT) case study

30.(AT) What is the mean number of correct responses of the participants after taking the placebo (0 mg/kg)?

31.(AT) What are the standard deviation and the interquartile range of the d0 condition?