CHAPTER 4

Exploring Data: Variability

Summary

This chapter covers *variability*. Calculating and interpreting measures of variability such as the *standard deviation*, *range*, *interquartile range*, and *variance* give you a quantitative way to express familiar ideas about how consistent things are, from how long it takes to get food at a restaurant to the weather. Variability can give you context to better understand measures of central tendency, so calculating and using both together can be very helpful.

The first quantitative measure of variability in this chapter is the *range*, which is a simple and easily calculated statistic. The range is simply the numerical distance between the highest and lowest scores.

The *interquartile range* (IQR) is the range of scores that capture the middle 50% of the distribution. The IQR is an important measure of variability because the middle 50% of the distribution is more stable from sample to sample than the range because the range uses the two most extreme scores in a distribution. The IQR is the 75th percentile minus the 25th percentile. A location value (.25 x *N*) allows you to find the score at the 75th percentile (by working down from the top of the distribution) and the score at the 25th percentile (by working up from the bottom of the distribution).

The most popular measure of variability is the *standard deviation*. The standard deviation tells you the average amount that scores differ from the mean. After studying the standard deviation in this chapter and again in Chapter 7, you will find it a powerful and accurate way to express the concept of variability.

There are two ways to complete the math to identify the standard deviation. You first learned to use the *definitional formula*, which helps you understand what a deviation score is and helps you see the property of the mean $\sum(X - \bar{X}) = 0$ discussed in Chapter 3 in action. You then learned to use the *computational formula*, which can typically be used more quickly than the *definitional formula*. The *computational formula* will also turn up again in subsequent chapters.

The three standard deviations are:
1. σ, which is the standard deviation of a population (*N* in denominator)
2. ŝ, which is the standard deviation of a sample, used to estimate σ (*N*-1 in denominator)
3. *S*, which is the standard deviation of a sample when there is no interest in generalizing from the sample to its population (*N* in denominator)

Choosing among these three standard deviations depends on how the data were gathered (σ for population data and ŝ or *S* for sample data) and the purpose of the data gathering (if you have a sample you're trying to describe, calculate *S*; if you have a sample you're using to generalize to a population, calculate ŝ). If the issue of *N* or *N*-1 in the denominator has you shaking your head, we recommend that you work the exercise in the textbook described in the footnote of the section, "ŝ as an estimate of σ".

Lines called error bars extend one standard deviation above and below the mean and can be added to a graph of means. Standard deviation lines indicate the spread of scores around the mean.

The variance gets two short paragraphs in the text, which is not representative of its importance in the overall field of statistics. You will learn more about variance in Chapters 11, 12, and 13. As a descriptive index of variability, however, the variance is not nearly as useful as its square root, the standard deviation.

Multiple-Choice Questions

1. The standard deviation that estimates a population standard deviation from calculations on sample data is _____.

a. σ
b. ŝ
c. *S*
d. μ.

2. Two distributions that have the same mean must have the same _____.

a. range
b. standard deviation
c. variance
d. none of the above

3. You will get a negative number as a standard deviation _____.

a. when all scores are negative
b. when the mean of the scores is negative
c. when all deviation scores are negative
d. under no circumstances—standard deviations are always positive numbers

4. A researcher was interested in the variability of SAT scores in the new first-year class at her university. She gathered a sample of the first-year students' scores from the registrar. She should compute _____.

a. σ
b. ŝ
c. *S*
d. the 50th percentile

5. The standard deviation of the IQ scores of students at a high school for the gifted and talented should be _____the standard deviation of IQ scores of students at a high school that accepts everyone in its district.

a. less than
b. greater than
c. equal to
d. cannot be determined from the information given

6. Which of the following distributions is the most variable?

a. 1, 2, 3
b. 8, 9, 10
c. 1, 3, 5
d. all of the above are equally variable

7. The interquartile range always _____.

a. includes one-fourth of the range
b. becomes larger if more observations are added
c. contains 50% of the scores
d. all of the above

8. When is it the case that the sum of deviation scores will be zero?

a. when half the scores are negative
b. when the mean is zero
c. when the standard deviation is zero
d. always

9. From all the employees in a company, a small group was selected to participate in a study of employee satisfaction. The results of the study were going to be generalized to all the employees of the company. To find the variability in employee satisfaction, the investigator should compute _____.

    a. μ
    b. S
    c. $\hat{s}$
    d. σ

10. From the same employee satisfaction survey in the previous question, the following summary numbers were obtained: $\Sigma X = 30$, $\Sigma X^2 = 350$, $N = 12$. Calculate the appropriate standard deviation.

    a.   22.92
    b.   4.79
    c.   5.00
    d.   none of the above

11. On the first test of material on dinosaurs, a class of sixth graders had a mean score ($\bar{X}$) of 36 with a standard deviation (S) of 12. The teacher was disappointed and assigned six pages of homework on dinosaurs and scheduled a second test. The top one-fourth of the class studied the extra material and did much better on the second test. The other three-fourths ignored the material and made the same scores as before. Pick the mean and standard deviation that *might* be found, given the description.

    a. $\bar{X} = 36$; $S = 12$
    b. $\bar{X} = 36$; $S = 18$
    c. $\bar{X} = 42$; $S = 18$
    d. $\bar{X} = 42$; $S = 12$

12. Which of the following may be a negative number?

    a.   range
    b.   variance
    c.   deviation score
    d.   both b. and c.

13. Suppose you want to know how spread out your data are, in terms of the difference between the highest and lowest score. You should calculate_____.

    a. $\hat{s}$
    b. S
    c.   IQR
    d.   the range

14. The variance is _____ the standard deviation.

    a.   the sum of
    b.   the square of
    c.   the square root of
    d.   twice as large as

15. For any distribution of scores, which of the following will *always* be zero?

    a.   the sum of the deviation scores
    b.   range

c. standard deviation
    d. variance

16. The standard deviation of the height of a group of boys and girls aged 3-14 is _____ that of a group aged 13-14.

    a. greater than
    b. less than
    c. equal to
    d. cannot be determined from the information given

17. If the variance of a distribution is a number greater than one, the standard deviation will be _____ the variance.

    a. larger than
    b. smaller than
    c. either larger or smaller than
    d. equal to

18. Suppose you had three numbers with a mean of 2. You calculated a deviation score for each of the three numbers and then added the deviation scores. The sum should be _____.

    a. 1
    b. 2
    c. 3
    d. none of these choices

19. Which of the following is false?

    a. ŝ is used to estimate σ
    b. ŝ uses N-1 in the denominator
    c. ŝ < σ when both are calculated on the same numbers
    d. ŝ is based on a sample

Short-Answer Questions

1. When considering range, IQR, and standard deviation, which measure would be expected to vary the most from sample to sample, when all samples are taken from the same population? Why?

2. Identify the three types of standard deviations and the situations in which you'd use each one.

3. Suppose there are four equally convenient supermarkets and they all place advertisements in the newspaper. From these advertisements you have noted the prices for a pound of apples, a gallon of milk, and a box of your favorite cereal. You have discovered that the standard deviation of apples is $0.35, the standard deviation of milk is $0.02, and the standard deviation of your favorite cereal is $0.01. If your goal is to save as much money as possible, while only shopping at one store, should you go to the store with the cheapest price on apples, milk, or cereal?

Problems

1. Find the range and interquartile range of the following data set.

X	f

7	2
6	3
5	5
4	3
3	3
2	3

2. Calculate S using the definitional formula on the following numbers: 2, 5, 8

3. Calculate $\hat{s}$ using the definitional formula on the following numbers: 4, 3, 2, 1

4. Calculate σ using the computational formula on the following numbers: 5, 6, 7, 8, 9

5. Calculate $\hat{s}$ using the computational formula on the following numbers: 1, 3, 5, 7

6. Suppose that you are to choose the best machine from among three different brands. You have sample output from each of the three brands. One characteristic of a high quality machine is consistency, which would mean low variability.

For each brand, calculate $\hat{s}$ using the computational formula and use your answer to recommend the machine with the most consistent output.

Brand A	Brand B	Brand C
3	-2	-5
11	7	3
-8	3	2
-2	-6	1
-4	-2	-1

7. In Chapter 3, you used the data below to identify measures of central tendency about the age (in months) at which 90% of babies first stand without support. Find the sample standard deviation $\hat{s}$ for this set of data using the computational formula.

Age (months) X	f
16	4
15	4
14	6
13	10
12	6
11	5
10	2
9	1

8. In Chapter 3, you used the data below to identify measures of central tendency for the opinions of college students and older adults about the importance of social security as a source of income for the elderly. Find the range, IQR, and use the computational formula to find ŝ.

College Students' Level of Agreement X	f	Older Adults' Level of Agreement X	f
5	4	5	2
4	4	4	1
3	2	3	2
2	1	2	4
1	1	1	6

9. The Clock Test is a technique for studying human vigilance. In the Clock Test, a hand moves regularly at one step per second, but sometimes, at random intervals, it jumps two steps. The participant's task is to notice the two-step jumps and press a button. The numbers that follow are the percentages of two-step jumps that were missed by 5 participants during the first 15 minutes of a two-hour watch. Find the mean (μ) and σ using the computationalformula.

11	8	17	10	14

10. In the Clock Test experiment described in the last problem, the experimenter also recorded the percentages of missed jumps during the *last* 15 minutes of the two-hour watch. Find the mean (μ) and σ using the computational formula. Using this answer and your answer from the last problem, write a sentence about the effect of two hours of vigilance on variability.

25	12	38	21	29

ANSWERS

Multiple-Choice Questions

1. b. ŝ
2. d. none of the above
   Explanation: Knowing the mean does not tell us anything about the variability of a distribution.
3. d. under no circumstances—standard deviations are always positive numbers
   Explanation: Because finding the standard deviation involves squaring deviation scores, it is not mathematically possible to get a negative number. Given that the standard deviation is a measure of how different a set of scores are from each other, if all of the scores are the same, you would have a standard deviation of zero. And, you can never have less difference than a standard deviation of zero.
4. b. ŝ
5. a. less than
   Explanation: Because IQ is part of the criteria for identifying gifted and talented students, we would expect a group of students who meet the gifted and talented criteria to have less variability in their scores than the overall school population.
6. c. 1, 3, 5
7. c. contains 50% of the scores
8. d. always
   Explanation: An important mathematical property of the mean is $\sum(X - \bar{X}) = 0$.
9. c. ŝ
10. c. 5.00
    Explanation:

$$\hat{s} = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}} = \sqrt{\frac{350 - \frac{(30)^2}{12}}{11}} = \sqrt{\frac{350 - 75}{11}} = \sqrt{\frac{275}{11}} = \sqrt{25} = 5.00$$

   Remember, this study is using a sample to make decisions about all employees of the company, which means you should use *N*-1 in the denominator. If you answered b., it means you used *N* in the denominator.
11. c. $\bar{X} = 42, S = 18$
    Explanation: If 25% of the students did better, and everyone else stayed the same, we would expect two things. First, the mean would be pulled in the direction of the skew, towards higher scores. Second, the amount of variability would go up because 25% of the students have higher scores than before.
12. c. deviation score
    Explanation: Range, IQR, standard deviation, and variance are all positive numbers. But, the deviation score of one individual score being compared to the mean can be negative, which would happen when that one individual gets a score that is lower than the mean.
13. d. the range
14. b. the square of
15. a. the sum of the deviation scores about the mean
16. a. greater than
    Explanation: Because there is a larger age range, there will be more variability. In addition, children get a lot taller between 3-13, and that variability will increase the standard deviation of the data set.
17. b. smaller than
    Explanation: Remember the standard deviation is the square root of the variance, so it will be smaller than the variance when the variance is greater than 1.
18. d. none of these choices
    Explanation: The sum of deviation scores will always be zero. An important mathematical property of the mean is $\sum(X - \bar{X}) = 0$.
19. c. ŝ < σ when both are calculated on the same numbers
    Explanation: Using *N*-1 in the denominator, rather than *N*, makes the ŝ calculation come out to be greater than σ when both are calculated on the same numbers.

Short-Answer Questions

1. The range will vary the most from sample to sample, because it is calculated using only the two most extreme scores in the distribution, which are the most likely scores to vary from sample to sample. The IQR will vary less than the range. Even though it also uses only two scores in the distribution, those two scores come from closer to the middle of the distribution (they are the 75th percentile and 25th percentile). The standard deviation is the least variable because it accounts for each and every score in the distribution.
2. $\hat{s}$ is the standard deviation of a sample, used to estimate σ (N-1 in denominator)

    σ is the standard deviation of a population (N in denominator)

    S is the standard deviation of a sample when there is no interest in generalizing from the sample to its population (N in denominator)
3. You should go to the store with the cheapest price on apples. Because the standard deviations for milk and cereal are so small, all of the stores will have almost the same price. But, the larger standard deviation for apples means that prices vary more and you will save the most money by going to the store with the cheapest price on apples.

Problems

1. Range: 5, IQR: 6-3 = 3
    Explanation: You first need to identify the number of scores ($\Sigma f = N = 19$). To find the IQR position, multiply 0.25 times N. Because 0.25 x 19 = 4.75, you are looking for the scores 4.75 down from the top of the distribution (6) and 4.75 up from the bottom of the distribution (3).
2. $S = 2.45$

X	$\bar{X}$	$X - \bar{X}$	$(X - \bar{X})^2$
2	5	-3	9
5	5	0	0
8	5	3	9
		$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 = 18$

$$S = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} = \sqrt{\frac{18}{3}} = \sqrt{6} = 2.45$$

3. $\hat{s} = 1.29$

X	$\bar{X}$	$X - \bar{X}$	$(X - \bar{X})^2$
4	2.5	1.5	2.25
3	2.5	0.5	0.25
2	2.5	-0.5	0.25
1	2.5	-1.5	2.25
		$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 = 5$

$$\hat{s} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N-1}} = \sqrt{\frac{5}{3}} = \sqrt{1.67} = 1.29$$

4. $\sigma = 1.41$

$$\sigma = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N}} = \sqrt{\frac{255 - \frac{(35)^2}{5}}{5}} = \sqrt{\frac{255-245}{5}} = \sqrt{\frac{10}{5}} = \sqrt{2} = 1.41$$

5. $\hat{s} = 2.58$

$$\hat{s} = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}} = \sqrt{\frac{84 - \frac{(16)^2}{4}}{3}} = \sqrt{\frac{84-64}{3}} = \sqrt{\frac{20}{3}} = \sqrt{6.67} = 2.58$$

6. If you use a formula for ŝ:
   Brand A: ŝ = 7.31, Brand B: ŝ = 5.05, Brand C: ŝ = 3.16

   Brand C is the best because the standard deviation is the smallest, which means there is less variability in output and the product is the most consistent.

   Brand A

   $$\hat{s}_A = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}} = \sqrt{\frac{214 - \frac{(0)^2}{5}}{4}} = \sqrt{\frac{214-0}{4}} = \sqrt{\frac{214}{4}} = \sqrt{53.5} = 7.31$$

   Brand B

   $$\hat{s}_B = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}} = \sqrt{\frac{102 - \frac{(0)^2}{5}}{4}} = \sqrt{\frac{102-0}{4}} = \sqrt{\frac{102}{4}} = \sqrt{25.5} = 5.05$$

   Brand C

   $$\hat{s}_C = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}} = \sqrt{\frac{40 - \frac{(0)^2}{5}}{4}} = \sqrt{\frac{40-0}{4}} = \sqrt{\frac{40}{4}} = \sqrt{10} = 3.16$$

7. ŝ = 1.79

Age (months) X	f	fX	X^2	fX^2
16	4	64	256	1024
15	4	60	225	900
14	6	84	196	1176
13	10	130	169	1690
12	6	72	144	864
11	5	55	121	605
10	2	20	100	200
9	1	9	81	81
	$\sum f = N = 38$	$\sum fX = 494$		$\sum fX^2 = 6540$

$$\hat{s} = \sqrt{\frac{\sum fX^2 - \frac{(\sum fX)^2}{N}}{N-1}} = \sqrt{\frac{6540 - \frac{(494)^2}{38}}{37}} = \sqrt{\frac{6540-6422}{37}} = \sqrt{\frac{118}{37}} = \sqrt{3.19} = 1.79$$

8. College Students: Range = 4, IQR = 2, ŝ =1.29
   Older Adults: Range = 4, IQR = 2.25, ŝ = 1.44

College Students

College Students' Level of Agreement X	f	fX	fX^2

	4	20	100
5	4	20	100
4	4	16	64
3	2	6	18
2	1	2	4
1	1	1	1
	$\Sigma f = N = 12$	$\Sigma fX = 45$	$\Sigma fX^2 = 187$

$$\hat{s} = \sqrt{\frac{\Sigma fX^2 - \frac{(\Sigma fX)^2}{N}}{N-1}} = \sqrt{\frac{187 - \frac{(45)^2}{12}}{11}} = 1.29$$

Older Adults

Older Adults' Level of Agreement X	f	fX	fX^2
5	2	10	50
4	1	4	16
3	2	6	18
2	4	8	16
1	6	6	6
	$\Sigma f = N = 15$	$\Sigma fX = 34$	$\Sigma fX^2 = 106$

$$\hat{s} = \sqrt{\frac{\Sigma fX^2 - \frac{(\Sigma fX)^2}{N}}{N-1}} = \sqrt{\frac{106 - \frac{(34)^2}{15}}{14}} = 1.44$$

9. $\mu = 12$, $\sigma = 3.16$

X	X^2
11	121
8	64
17	289
10	100
14	196
$\Sigma X = 60$	$\Sigma X^2 = 770$

$$\mu = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N}} = \sqrt{\frac{770 - \frac{(60)^2}{5}}{5}} = \sqrt{\frac{770 - 720}{5}} = \sqrt{\frac{50}{5}} = \sqrt{10} = 3.16$$

10. $\mu = 25$, $\sigma = 8.6$

The error rate goes up from 12% in the first 15 minutes of the experiment to 25% in the last 15 minutes of the experiment. This means that people make more mistakes after two hours of sustained vigilance. In addition, the standard deviation goes from 3.16 at the beginning of the study to 8.6 at the end of the study. This means that sustained vigilance affects people differently and there is more variability in peoples' mistakes at the end of the study than at the beginning.

X	X^2
25	625
12	144
38	1444
21	441

29	841
$\sum X = 125$	$\sum X^2 = 3495$

$$\mu = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}} = \sqrt{\frac{3495 - \frac{(125)^2}{5}}{5}} = \sqrt{\frac{3495 - 3125}{5}} = \sqrt{\frac{370}{5}} = \sqrt{74} = 8.6$$