Lecture Notes

# Chapter 4: Measures of Variability

## Learning Objectives

1. Explain the importance of measuring variability.
2. Calculate and interpret the index of qualitative variation, range, interquartile range, the variance, and the standard deviation.
3. Identify the relative strengths and weaknesses of the measures.

## Chapter Outline

1. Introduction
	1. There are five **measures of variability** (numbers that describe diversity or variability in the distribution):
		1. The index of qualitative variation
		2. The range
		3. The interquartile range
		4. The standard deviation
		5. The variance.
2. The Importance of Measuring Variability
	1. Important to look at variation and diversity.
	2. One form of stereotyping is treating a group as if it were totally characterized by its central value, ignoring the diversity within the group.
	3. The concept of variability has implications not only for describing the diversity of social groups but also for issues that are important in everyday life.
	4. Reconstructing curriculum to make it more responsive to the needs of students.
		1. Suppose that a university committee is examining how to better respond to the needs of students.
		2. To evaluate statistics courses offered in different departments, the committee compares the grading policy in two courses.
		3. However, we need to look more closely into how the grades are distributed in each of the classes.
		4. As this example demonstrates, information on how scores are spread from the center of a distribution is as important as information about the central tendency in a distribution.
3. The Index of Qualitative Variation
	1. The **index of qualitative variation** (IQV): A measure of variability for nominal variables. It is based on the ratio of the total number of differences in the distribution to the maximum number of possible differences within the same distribution.
	2. The index can vary from 0.00 to 1.00.
		1. When all the cases in the distribution are in one category, there is no variation (or diversity) and the IQV is 0.00.
		2. When the cases in the distribution are distributed evenly across the categories, there is maximum variation (or diversity) and the IQV is 1.00.
	3. Steps for Calculating the IQV
		1. To calculate the IQV, we use this formula: 

where

*K* = the number of categories,

 = the sum of all squared percentages in the distribution.

* 1. The steps to calculate the IQV:
		1. Construct a percentage distribution.
		2. Square the percentages for each category.
		3. Sum the squared percentages.
		4. Calculate the IQV using the formula.
	2. Expressing the IQV as a Percentage
		1. To express the IQV as a percentage rather than a proportion, simply multiply the IQV by 100.
		2. Expressed as a percentage, the IQV would reflect the percentage of differences relative to the maximum possible differences in each distribution.
1. Statistics in Practice: Diversity in U.S. Society
	1. By the middle of the 21st century, the United States will no longer be a predominantly White society--it will be a minority–majority nation.
	2. There are geographic concentrations of minority groups in specific regions and metropolitan areas; demographers refer to this as chain migration.
	3. IQV can be used to measure the amount of racial and ethnic diversity in different regions.
2. The Range
	1. The simplest and most straightforward measure of variation is the **range**, which measures variation in interval-ratio variables.
	2. It is the difference between the highest (maximum) and the lowest (minimum) scores in the distribution.
	3.  To find the ranges in a distribution, simply pick out the highest and the lowest scores in the distribution and subtract.
	4. Although the range is simple and quick to calculate, it is a rather crude measure because it is based on only the lowest and the highest scores.
	5. These two scores might be extreme and rather atypical, which might make the range a misleading indicator of the variation in the distribution.
3. The Interquartile Range
	1. To remedy the limitation of the range, we can employ an alternative, the **interquartile range**.
	2. The interquartile range (IQR): a measure of variation for interval-ratio and ordinal variables is the width of the middle 50% of the distribution.
	3. It is defined as the difference between the lower and upper quartiles (*Q*1 and *Q*3).
		1. 
		2. The first quartile (*Q*1) is the 25th percentile, the point at which 25% of the cases fall below it and 75% above it.
		3. The third quartile (*Q*3) is the 75th percentile, the point at which 75% of the cases fall below it and 25% above it.
		4. The IQR, therefore, defines variation for the middle 50% of the cases.
	4. Like the range, the IQR is based on only two scores. However, because it is based on intermediate scores, rather than on the extreme scores in the distribution, it avoids some of the instability associated with the range.
	5. Steps for calculating the IQR:
		1. To find *Q*1 and *Q*3, order the scores in the distribution from the highest to the lowest score or vice versa.
		2. Identify the first quartile, *Q*1 or the 25th percentile. To find *Q*1, we multiply *N* by 0.25: .
		3. To find Q3, or the third quartile, we multiply *N* this time by 0.75:
			1. .
		4. We are now ready to find the IQR. 
		5. It may be more useful to report the full IQR rather than the single value.
4. The Box Plot
	1. A graphic device called the box plot can visually present the range, the IQR, the median, the lowest (minimum) score, and the highest (maximum) score.
	2. The box plot provides us with a way to visually examine the center, the variation, and the shape of distributions of interval-ratio variables.
	3. To construct a box plot, we use the lowest and the highest values in the distribution, the upper and lower quartiles, and the median.
	4. We can easily draw a box plot by hand following these instructions:
		1. Draw a box between the lower and upper quartiles.
		2. Draw a solid line within the box to mark the median.
		3. Draw vertical lines (called whiskers) outside the box, extending to the lowest and highest values.
	5. From creating a box plot, we can obtain a visual impression of the following properties:
		1. The center of the distribution is easily identified by the solid line inside the box.
		2. Since the box is drawn between the lower and upper quartiles, the IQR is reflected in the height of the box.
		3. The length of the vertical lines drawn outside the box (on both ends) represents the range of the distribution.
		4. Both the IQR and the range give us a visual impression of the spread in the distribution.
		5. The relative position of the box and the position of the median within the box tell us whether the distribution is symmetrical or skewed.
			1. A perfectly symmetrical distribution would have the box at the center of the range as well as the median in the center of the box.
			2. When the distribution departs from symmetry, the box and/or the median will not be centered; it will be closer to the lower quartile when there are more cases with lower scores or to the upper quartile when there are more cases with higher scores.
5. The Variance and the Standard Deviation
	1. The variance and the standard deviation are two closely related measures of variation that increase or decrease based on how closely the scores cluster around the mean.
	2. The **variance** is the average of the squared deviations from the center (mean) of the distribution, and the **standard deviation** is the square root of the variance.
	3. Both measure variability in interval-ratio and ordinal variables.
	4. Calculating the Deviation from the Mean
6. While calculating the difference between the average difference of the selected values, from the total average (the mean), it makes sense to first look at the difference between each value and the mean.
7. This difference, called a deviation from the mean, is symbolized as . The sum of these deviations can be symbolized as .
8. Each value has either a positive or a negative deviation score.
9. The deviation is positive when the percentage change in the selected value is above the mean.
10. It is negative when the percentage change is below the mean.
11. We cannot calculate the average of these deviations by simply adding up the deviations and dividing them, because the sum of the deviations of scores from the mean is always zero, or algebraically, .
12. In other words, if we were to subtract the mean from each score and then add up all the deviations, the sum would be zero, which in turn would cause the average deviation (i.e., average difference) to compute to zero.
13. This is always true because the mean is the center of gravity of the distribution.
14. Mathematically, we can overcome this problem by
15. Ignoring the plus and minus signs, using instead the absolute values of the deviations.
16. Squaring the deviations, that is, multiplying each deviation by itself to get rid of the negative sign.
17. Since absolute values are difficult to work with mathematically, the latter method is used to compensate for the problem.
18. The sum of the squared deviations is symbolized as .
19. By squaring the deviations, we end up with a sum representing the deviation from the mean, which is positive.
	1. Calculating the Variance and the Standard Deviation
		1. The average of the squared deviations from the mean is known as the variance.
		2. The variance is symbolized as *s*2.
		3. We are interested in the average of the squared deviations from the mean.
		4. Therefore, we need to divide the sum of the squared deviations by the number of scores (*N*) in the distribution.
		5. However, unlike the calculation of the mean, we will use *N* − 1 rather than *N* in the denominator.
		6. The formula for the variance can be stated as , where

*s*2 = the variance,

 = the deviation from the mean,

 = the sum of the squared deviations from the mean,

*N* = the number of scores.

* + 1. This formula means that the variance is equal to the average of the squared deviations from the mean.
		2. Follow these steps to calculate the variance:
1. Calculate the mean .
2. Subtract the mean from each score to find the deviation .
3. Square each deviation 
4. Sum the squared deviations 
5. Divide the sum by *N* – 1.
6. The answer is the variance: .
	* 1. One problem with the variance is that it is based on squared deviations and therefore is no longer expressed in the original units of measurement.
		2. This figure is expressed in squared percentages. Thus, we often take the square root of the variance and interpret it instead.
		3. This gives us the standard deviation, symbolized as *s,* is the square root of the variance, or 
		4. The formula for the standard deviation uses the same symbols as the formula for the variance, .
		5. As we interpret the formula, we can say that the standard deviation is equal to the square root of the average of the squared deviations from the mean.
		6. The advantage of the standard deviation is that unlike the variance, it is measured in the same units as the original data.
		7. The actual number tells us very little by itself, but it allows us to evaluate the dispersion of the scores around the mean.
7. In a distribution where all the scores are identical, the standard deviation is zero (0). Zero is the lowest possible value for the standard deviation; in an identical distribution, all the points would be the same, with the same mean, mode, and median. There is no variation or dispersion in the scores.
8. The more the standard deviation departs from zero, the more variation there is in the distribution. There is no upper limit to the value of the standard deviation.
9. The standard deviation can be considered a standard against which we can evaluate the positioning of scores relative to the mean and to other scores in the distribution.
10. Considerations for Choosing a Measure of Variation
	1. We have considered five measures of variation:
		1. The IQV
		2. The range
		3. The IQR
		4. The variance
		5. The standard deviation.
	2. Each measure can represent the degree of variability in a distribution.
	3. However, in general, we tend to use only one measure of variation and the choice of the appropriate one involves several considerations. These considerations and how they affect our choice of the appropriate measure are presented in the form of a decision tree.
	4. As in choosing a measure of central tendency, one of the most basic considerations in choosing a measure of variability is the variable’s level of measurement. Valid use of any of the measures requires that the data are measured at the level appropriate for that measure or higher.
11. Reading the Research Literature: Community College Mentoring
	1. According to a study by Myron Pope, students of color enrolled in community colleges response best to multiple levels of mentoring--formal and informal methods from different sources.
	2. To improve these mentoring opportunities, community college administrators must begin more aggressively to recruit future faculty and administrators from various ethnic backgrounds.