Chapter five

How Typical Is the Typical Case?: Measuring Dispersion

Measures of dispersion

What Do They Tell Us About Our Data?

Measuring dispersion in nominal and ordinal scales: proportions, percentages, and the variation ratio

How are They Calculated?

What are Their Characteristics?

Measuring dispersion in interval scales: range, variance, and standard deviation

How are They Calculated?

What are Their Characteristics?

D. Weisburd and C. Britt, *Statistics in Criminal Justice*, DOI 10.1007/978-1-4614-9170-5_5,

MEASURES OF CENTRAL TENDENCY provide a snapshot of the typical case; however, the same statistic may be obtained from samples or populations that are in fact quite dissimilar. For example, a sample of police recruits with a mean or median age of 23 is not likely to include people younger than 18 or older than 30, because most police departments have age requirements for incoming officers. A sample of offenders with a mean or median age of 23, however, will include offenders younger than 18 and much older than 30. In both these samples, the average person studied is 23 years old. But the sample of offenders will include more younger and older people than the sample of police recruits. The ages of the offenders are dispersed more widely around the average age.

Measures of dispersion allow us to fill a gap in our description of the samples or populations we study. They ask the question: How typical is the typical case? They tell us to what extent the subjects we studied are similar to the case we have chosen to represent them. Are most cases clustered closely around the average case? Or, as with the sample of offenders above, is there a good deal of dispersion of cases both above and below the average?

Measures of Dispersion for Nominal- and Ordinal-Level Data

With nominal scales, we define the typical case as the category with the largest number of subjects. Accordingly, in Chapter 4 we chose "private attorney" as the modal category for legal representation for a sample of white-collar offenders. But how would we describe to what extent the use of a private attorney is typical of the sample as a whole? Put another way, to what degree are the cases concentrated in the modal category?

The Proportion in the Modal Category

The most straightforward way to answer this question is to describe the proportion of cases that fall in the modal category. Recall from Chapter 3 that a proportion is represented by the following equation:

$$\text{Proportion} = \frac{N_{\text{cat}}}{N_{\text{total}}}$$

Accordingly, we can represent the proportion of cases in the modal category using Equation 5.1:

$$\text{Proportion} = \frac{N_{\text{modal cat.}}}{N_{\text{total}}}$$

Equation 5.1

That is, we take the number of cases in the modal category and divide it by the total number of cases in the sample.

Taking the example of legal representation, we divide the N of cases in the modal category (private attorney) by the total N of cases in the sample (see Table 5.1):

Working It Out

$$\begin{aligned}\text{Proportion} &= \frac{N_{\text{modal cat.}}}{N_{\text{total}}}\\ &= \frac{380}{671}\\ &= 0.5663\end{aligned}$$

Following our earlier suggestions regarding rounding to the second decimal place, we say that the proportion of white-collar offenders in the modal category was about 0.57.

Table 5.1 Legal Representation for White-Collar Crime

CATEGORY	FREQUENCY (*N*)
No Attorney	20
Legal Aid	26
Court Appointed	92
Public Defender	153
Private Attorney	380
Total (Σ)	671

Table 5.2 Method of Execution in the United States, 1977–2000

CATEGORY	FREQUENCY (N)
Lethal Injection	518
Electrocution	149
Lethal Gas	11
Hanging	3
Firing Squad	2
Total (Σ)	683

Source: Tracy L. Snell, "Capital Punishment 2000," *Bureau of Justice Statistics Bulletin,* 2001, p. 12.

Table 5.2 presents information about the method of execution used on the 683 persons executed in the United States from 1977 to 2000. The modal category is lethal injection, so the proportion in the modal category is found by dividing the N of cases in that category by the total N of cases:

Working It Out

$$\begin{aligned}\text{Proportion} &= \frac{N_{\text{modal cat.}}}{N_{\text{total}}}\\ &= \frac{518}{683}\\ &= 0.7584\end{aligned}$$

Of persons executed in the United States from 1977 to 2000, the proportion killed through lethal injection was about 0.76.

The Percentage in the Modal Category

Alternatively, we may refer to the percentage in the modal category. Most people find percentages easier to understand than proportions. Recall that a percentage is obtained by taking a proportion and multiplying it by 100. Accordingly, we can take Equation 5.1 and multiply the result by 100 to get the percentage of cases in the modal category.

$$\text{Percentage} = \frac{N_{\text{modal cat.}}}{N_{\text{total}}} \times 100$$

For our legal representation example,

Working It Out

$$\begin{aligned}\text{Percentage} &= \frac{N_{\text{modal cat.}}}{N_{\text{total}}} \times 100\\ &= \frac{380}{671} \times 100\\ &= 56.6319\end{aligned}$$

That is, about 57% of the cases in the sample fall in the modal category.

Similarly, for the method of execution example, the percentage in the modal category is

$$\begin{aligned}\text{Percentage} &= \frac{518}{683} \times 100\\ &= 75.8419\end{aligned}$$

About 76% of all executions in the United States from 1977 to 2000 involved the use of lethal injection.

The Variation Ratio

Another way to describe the degree to which the modal category represents the cases in a sample is to use a statistic called the **variation ratio** (VR). The variation ratio is based on the same logic as a proportion, but it examines the extent to which the cases are spread outside the modal category, rather than concentrated within it. The proportion of cases in the modal category is subtracted from 1:

$$\text{VR} = 1 - \left(\frac{N_{\text{modal cat.}}}{N_{\text{total}}}\right) \qquad \textbf{Equation 5.2}$$

For the legal representation example, the variation ratio is

Working It Out

$$\begin{aligned}\text{VR} &= 1 - \left(\frac{N_{\text{modal cat.}}}{N_{\text{total}}}\right)\\ &= 1 - \left(\frac{380}{671}\right)\\ &= 0.4337\end{aligned}$$

The variation ratio for legal representation in this sample of white-collar offenders is about 0.43. But what does this say about the extent to which cases in the sample are clustered around the typical case? Is a variation ratio of 0.43 large or small? What rule can we use for deciding more generally whether the distribution we are examining is strongly clustered?

One approach is to define at the outset the upper and lower limits for the variation ratio or proportion for a particular measure. Obviously, the largest proportion, regardless of the study, is 1.0, which would mean that all of the cases were in the modal category. Having all of the cases in the modal category would lead to a variation ratio of 0, indicating no dispersion.

The smallest proportion (or largest VR) depends, however, on the number of categories in your measure. The mode is defined as the category in your measure with the most cases, so it must have at least one more case than any other category. If you have only two categories, then the modal category must include one more than half of the cases in your study. So, in the instance of two categories, the least possible concentration is just over 0.50 of the cases. The least possible dispersion, as measured by the variation ratio, would be 1 minus this proportion, or just under 0.50. If you have four categories, the modal category must have more than one-quarter of the cases. Accordingly, the smallest variation ratio would be a bit smaller than 0.75.

What about our example of legal representation? We have five categories and 671 cases. The smallest number of cases the modal category could have with these numbers is 135. In this instance, each of the other four categories would have 134 cases. This is the maximum amount of dispersion that could exist in this sample, and it amounts to about 20.12% of the total number of cases in the sample, or a variation ratio of 0.7988. As noted earlier, the greatest degree of concentration in the modal category would yield a proportion of 1 and a variation ratio of 0. The estimates we calculated for legal representation (proportion = 0.57; VR = 0.43) lie somewhere between these two extremes.

Is this dispersion large or small? As with many of the statistics we will examine, the answer depends on the context in which you are working. “Large” or “small” describes a value, not a statistical concept. Statistically, you know that your estimate falls somewhere between the largest possible degree of concentration and the largest possible degree of dispersion. But whether this is important or meaningful depends on the problem you are examining and the results that others have obtained in prior research.

For example, if, in a study of legal representation for white-collar crime in England, it had been found that 90% of the cases were concentrated in the private attorney category, then we might conclude that our results reflected a relatively high degree of dispersion of legal representation in the United States. If, in England, only 25% of the cases had been in the modal category, we might conclude that there was a relatively low degree of dispersion of legal representation in the United States.

The proportion and the variation ratio are useful primarily for describing dispersion with nominal-level measures. In some circumstances, however, they can be useful for describing ordinal-level variables as well. This is true primarily when there are just a few categories in a measure or when there is a very high degree of concentration of cases in one category. The problem in using a simple proportion or variation ratio for ordinal-level measures is that the mode, upon which these statistics are based, is often a misleading measure for ordinal scales. As discussed in Chapter 4, the mode does not take into account the positions of scores in a measure, and thus it may provide a misleading view of the average case.

Index of Qualitative Variation

One measure of dispersion that is not based on the mode—and that can be used for both nominal and ordinal scales—is the **index of qualitative variation** (IQV). The IQV compares the amount of variation observed in a sample to the total amount of variation possible, given the number of cases and categories in a study. It is a standardized measure. This means that whatever the number of cases or categories, the IQV can vary only between 0 and 100. An IQV of 0 means that there is no variation in the measure, or all of the cases lie in one category. An IQV of 100 means that the cases are evenly dispersed across the categories.

$$\text{IQV} = \left(\frac{\displaystyle\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} N_{\text{obs}_i} N_{\text{obs}_j}}{\displaystyle\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} N_{\text{exp}_i} N_{\text{exp}_j}} \right) \times 100 \qquad \textbf{Equation 5.3}$$

Equation 5.3 provides a guide for how to compute the IQV. You are already familiar with the summation symbols within the parentheses. Here we are summing not across cases, but across products of distinct categories. N_{obs} represents the number of cases we observe within a category in our study. N_{exp} represents the number of cases we would expect in a category if the measure were distributed equally across the categories. That is, it is the N we would expect if there were the maximum amount of dispersion of our cases. We use the subscripts i, j, and k as a shorthand way to say that we should multiply all of the potential pairs of categories. Here's how this works: k represents the total number of categories of a variable. In the legal representation example, $k = 5$. Subscripts i and j index the categories of the variable. Use of the subscripts i and j provides us with a way of keeping track and making sure that we have multiplied all possible pairs of observed frequencies from each of the categories.

For example, if a variable had three categories, then the numerator (the measure of observed variation) would be equal to

$$N_{\text{obs}_1}N_{\text{obs}_2} + N_{\text{obs}_1}N_{\text{obs}_3} + N_{\text{obs}_2}N_{\text{obs}_3}$$

If a variable had four categories, then the numerator would be equal to

$$N_{\text{obs}_1}N_{\text{obs}_2} + N_{\text{obs}_1}N_{\text{obs}_3} + N_{\text{obs}_1}N_{\text{obs}_4} + N_{\text{obs}_2}N_{\text{obs}_3} + N_{\text{obs}_2}N_{\text{obs}_4} + N_{\text{obs}_3}N_{\text{obs}_4}$$

A concrete example will make it much easier to develop this statistic in practice. Let's say that we wanted to describe dispersion of an ordinal-scale measure of fear of crime in a college class of 20 students. The students were asked whether they were personally concerned about crime on campus. The potential responses were "very concerned," "quite concerned," "a little concerned," and "not concerned at all." The responses of the students are reported under the "*N* observed" column in Table 5.3. As you can see, the cases are fairly spread out, although there are more students in the "very concerned" and "quite concerned" categories than in the "a little concerned" and "not concerned at all" categories. The expected number of cases in each category under the assumption of maximum dispersion is 5. That is, if the cases were equally spread across the categories, we would expect the same number in each. Following Equation 5.3, we first multiply the number of cases observed in each category by the number observed in every other category and then sum. We then divide this total by the sum of the number expected in each category

Table 5.3 Fear of Crime Among Students

CATEGORY	*N* OBSERVED	*N* EXPECTED
Not Concerned at All	3	20/4 = 5
A Little Concerned	4	20/4 = 5
Quite Concerned	6	20/4 = 5
Very Concerned	7	20/4 = 5
Total (Σ)	20	20

multiplied by the number expected in every other category. This amount is then multiplied by 100:

Working It Out

$$\text{IQV} = \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} N_{\text{obs}_i} N_{\text{obs}_j}}{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} N_{\text{exp}_i} N_{\text{exp}_j}} \times 100$$

$$= \left(\frac{(3 \times 4) + (3 \times 6) + (3 \times 7) + (4 \times 6) + (4 \times 7) + (6 \times 7)}{(5 \times 5) + (5 \times 5) + (5 \times 5) + (5 \times 5) + (5 \times 5) + (5 \times 5)}\right) \times 100$$

$$= \left(\frac{145}{150}\right) \times 100$$

$$= 96.6667$$

The observed variation is 145. The expected variation is 150, representing the maximum amount of dispersion possible for the measure. The IQV for this measure is 96.67, meaning that the cases studied are very dispersed among the categories of the measure.

Measuring Dispersion in Interval Scales: The Range, Variance, and Standard Deviation

A common method of describing the spread of scores on interval or higher scales is to examine the **range** between the highest and lowest scores. Take, for example, the distribution of cases in Table 5.4. Let's say that this was a distribution of crime calls at hot spots over a one-year period. In describing typicality in this distribution, we would report the mean number of calls for the 12 places, which is 21.50. In describing how dispersed the scores are, we would report that the scores range between 2 and 52, or that the range of scores is 50.

The range is very simple and easy to present. Its attraction lies precisely in the fact that everyone understands what a range represents. However, the range is an unstable statistic because it uses very little of the information available in interval-level scales. It bases its estimate of dispersion on just two observations, the highest and lowest scores. This means that a change in just one case in a distribution can completely

Table 5.4 Crime Calls at Hot Spots in a Year

HOT SPOT NUMBER	NUMBER OF CALLS
1	2
2	9
3	11
4	13
5	20
6	20
7	20
8	24
9	27
10	29
11	31
12	52

alter your description of dispersion. For example, if we changed the case with the most calls in Table 5.4 from 52 to 502, the range would change from 50 to 500.

One method for reducing the instability of the range is to examine cases that are not at the extremes of your distribution. In this way, you are likely to avoid the problem of having the range magnified by a few very large or small numbers. For example, you might choose to look at the range between the 5th and 95th percentile scores, rather than that between the lowest and highest scores. It is also common to look at the range between the 25th and 75th percentile scores or between the 20th and 80th percentile scores. But however you change the points at which the range is calculated, you still rely on just two scores in determining the spread of cases in your distribution. The range provides no insight into whether the scores below or above these cases are clustered together tightly or dispersed widely. Its portrait of dispersion for interval scales is thus very limited.

How can we gain a fuller view of dispersion for interval scales? Remember that we became interested in the problem of dispersion because we wanted to provide an estimate of how well the average case represented the distribution of cases as a whole. Are scores clustered tightly around the average case or dispersed widely from it? Given that we have already described the mean as the most appropriate measure of central tendency for such scales, this is the natural place to begin our assessment. Why not simply examine how much the average scores differ from the mean?

In fact, this is the logic that statisticians have used to develop the main measures of dispersion for interval scales. However, they are faced with a basic problem in taking this approach. As we discussed in Chapter 4, if we add up all of the deviations from the mean, we will always

Table 5.5 Deviations from the Mean for Crime Calls at Hot Spots in a Year

HOT SPOT NUMBER	NUMBER OF CALLS	DEVIATIONS FROM THE MEAN $(X_i - \overline{X})$
1	2	2 − 21.5 = −19.5
2	9	9 − 21.5 = −12.5
3	11	11 − 21.5 = −10.5
4	13	13 − 21.5 = −8.5
5	20	20 − 21.5 = −1.5
6	20	20 − 21.5 = −1.5
7	20	20 − 21.5 = −1.5
8	24	24 − 21.5 = 2.5
9	27	27 − 21.5 = 5.5
10	29	29 − 21.5 = 7.5
11	31	31 − 21.5 = 9.5
12	52	52 − 21.5 = 30.5
		Total (Σ) = 0.0

come up with a value of 0. You can see this again by looking at the data on crime calls at hot spots in Table 5.5. If we take the sum of the differences between each score and the mean, written in equation form as

$$\sum_{i=1}^{N}(X_i - \overline{X})$$

the total, as expected, is 0.

As discussed in Chapter 4, when we add up the deviations above and below the mean, the positive and negative scores cancel each other out. In order to use deviations from the mean as a basis for a measure of dispersion, we must develop a method for taking the sign, or direction, out of our statistic. One solution is to square each deviation from the mean. Squaring will always yield a positive result because multiplying a positive number or a negative number by itself will result in a positive outcome. This is the method that statisticians have used in developing the measures of dispersion most commonly used for interval scales.

The Variance

When we take this approach, the **variance** (s^2) provides an estimate of the dispersion around the mean. It is the sum of the squared deviations from the mean divided by the number of cases. Written in equation form, it is

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N} \qquad \textbf{Equation 5.4}$$

Table 5.6 Variance for Crime Calls at Hot Spots in a Year

HOT SPOT NUMBER	NUMBER OF CALLS	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$
1	2	$2 - 21.5 = -19.5$	380.25
2	9	$9 - 21.5 = -12.5$	156.25
3	11	$11 - 21.5 = -10.5$	110.25
4	13	$13 - 21.5 = -8.5$	72.25
5	20	$20 - 21.5 = -1.5$	2.25
6	20	$20 - 21.5 = -1.5$	2.25
7	20	$20 - 21.5 = -1.5$	2.25
8	24	$24 - 21.5 = 2.5$	6.25
9	27	$27 - 21.5 = 5.5$	30.25
10	29	$29 - 21.5 = 7.5$	56.25
11	31	$31 - 21.5 = 9.5$	90.25
12	52	$52 - 21.5 = 30.5$	930.25
		Total (Σ) = 0.0	**Total (Σ) = 1,839.00**

In practice, you must take the following steps to compute the variance (as we do for our example in Table 5.6):

1. Take each case and subtract the mean from it, to get the deviation from the mean. For our example of crime calls at hot spots, we first take the case with 2 calls and subtract the mean of 21.5 from it, to get a score of −19.5.
2. Square each of these scores. For the first case, our result is 380.25.
3. Sum the results obtained in step 2. For our example of hot spots of crime, this yields a total of 1,839.
4. Finally, divide this result by the number of cases in the study. For our 12 cases, this leads to a variance of 153.25.[1]

[1] If you are working with SPSS or another computer package, you will notice that the result you get computing the variance by hand using this formula and the result provided by the computer package are slightly different. For example, SPSS computes a variance of 167.18 for the distribution provided in Table 5.6. The difference develops from the computer's use of a correction for the bias of sample variances: 1 is subtracted from the *N* in the denominator of Equation 5.4. The correction is used primarily as a tool in inferential statistics and is discussed in Chapter 10. Though it is our view that the uncorrected variance should be used in describing sample statistics, many researchers report variances with the correction factor for sample estimates. When samples are larger, the estimates obtained with and without the correction are very similar, and thus it generally makes very little substantive difference which approach is used.

Working It Out

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}$$

$$= \frac{\sum_{i=1}^{12}(X_i - 21.5)^2}{12}$$

$$= \frac{1{,}839}{12}$$

$$= 153.25$$

As another example, consider the data presented in Table 5.7 on bail amounts required for a group of 15 defendants. The mean bail amount is \$3,263.33. Following the same procedure as before, we subtract the mean from each of the individual observations. These values are presented in the third column. The squared deviations from the mean appear in the fourth column, and the sum of the squared deviations appears at the bottom of the column. When we divide the total by the *N* of cases, we gain a variance of \$6,984,155.56 for the dollar amount of bail.

Table 5.7 Variance for Bail Amounts for a Sample of Persons Arrested for Felonies

DEFENDANT	BAIL AMOUNT	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$
1	500	−2,763.33	7,635,992.69
2	1,000	−2,263.33	5,122,662.69
3	1,000	−2,263.33	5,122,662.69
4	1,000	−2,263.33	5,122,662.69
5	1,200	−2,063.33	4,257,330.69
6	1,500	−1,763.33	3,109,332.69
7	2,500	−763.33	582,672.69
8	2,500	−763.33	582,672.69
9	2,500	−763.33	582,672.69
10	2,750	−513.33	263,507.69
11	5,000	1,736.67	3,016,022.69
12	5,000	1,736.67	3,016,022.69
13	5,000	1,736.67	3,016,022.69
14	7,500	4,236.67	17,949,372.69
15	10,000	6,736.67	45,382,722.69
		Total (Σ) = 0.05	**Total (Σ) = 104,762,333.33**

Working It Out

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}$$

$$= \frac{\sum_{i=1}^{15}(X_i - 3{,}263.33)^2}{15}$$

$$= \frac{104{,}762{,}333.33}{15}$$

$$= 6{,}984{,}155.56$$

With the variance, we now have a statistic for computing dispersion based on deviations from the mean. However, how can we interpret whether the variance for a distribution is large or small? If you are having trouble making sense of this from our two examples, you are not alone. While squaring solves one problem (the fact that the raw deviations from the mean sum to 0), it creates another. By squaring, we generally obtain numbers that are much larger than the actual units in the distributions we are examining.[2]

The Standard Deviation

Another measure of dispersion based on the variance provides a solution to the problem of interpretation. This measure, the **standard deviation,** is calculated by taking the square root of the variance. Accordingly, it reduces our estimate of dispersion, using a method similar to the one we employed to solve the problem of positive and negative differences from the mean adding to 0. The standard deviation (*s*) provides an estimate of dispersion in units similar to those of our original scores. It is described in equation form as

$$s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}}$$

Equation 5.5

[2] In the special case of a fraction, the result will be smaller numbers.

Computational Equations for the Variance and Standard Deviation

Although Equations 5.4 and 5.5 provide a useful way of conceptualizing and measuring the variance and standard deviation, you can also use a computing formula that has fewer steps and is less likely to result in computational error. In Table 5.6, we rounded the mean and then calculated squared deviations based on values that were rounded at each step. In an attempt to limit the amount of rounding, and consequently decrease the chances of a mistake, an alternative equation that can be used for the variance is

$$s^2 = \frac{\sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}}{N}$$

And an alternative for the standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}}{N}}$$

Let's reconsider the data in Table 5.6 on hot spots. The following table illustrates the key calculations:

HOT SPOT NUMBER	**NUMBER OF CALLS (X_i)**	X_i^2
1	2	4
2	9	81
3	11	121
4	13	169
5	20	400
6	20	400
7	20	400
8	24	576
9	27	729
10	29	841
11	31	961
12	52	2,704
Total (Σ)	258	7,386

The variance (s^2) is then calculated with the computational equation as

$$s^2 = \frac{7{,}386 - \frac{(258)^2}{12}}{12} = 153.25$$

And the standard deviation is simply the square root of the variance:

$$s = \sqrt{153.25} = 12.38$$

Similarly, let's revisit the bail data in Table 5.7 and compute the variance with the computational formula. The following table illustrates the key calculations.

DEFENDANT	BAIL AMOUNT (X_i)	X_i^2
1	500	250,000
2	1,000	1,000,000
3	1,000	1,000,000
4	1,000	1,000,000
5	1,200	1,440,000
6	1,500	2,250,000
7	2,500	6,250,000
8	2,500	6,250,000
9	2,500	6,250,000
10	2,750	7,562,500
11	5,000	25,000,000
12	5,000	25,000,000
13	5,000	25,000,000
14	7,500	56,250,000
15	10,000	100,000,000
Total (Σ)	48,950	264,502,500

The variance is

$$s^2 = \frac{264{,}502{,}500 - \frac{(48{,}950)^2}{15}}{15} = 6{,}984{,}155.56$$

And the standard deviation is

$$s = \sqrt{6{,}984{,}155.56} = 2{,}642.76$$

In calculating the standard deviation, we add one step to our calculation of variance: We take the square root of our result. For the example of crime calls at 12 hot spots (where the variance equaled 153.25), we obtain a standard deviation of $\sqrt{153.25} = 12.38$.[3] If you were to define, on average, how much the scores differed from the mean just by looking at these 12 cases, you would probably come to a conclusion close to that provided by the standard deviation. Similarly, if we take the square root of the variance for the bail example above, we come up with a figure that makes much more intuitive sense than the variance. In this case, the standard deviation is $\sqrt{6{,}984{,}155.56}$, or \$2,642.76.

The standard deviation has some basic characteristics, which relate generally to its use:

1. A standard deviation of 0 means that a measure has no variability. For this to happen, all of the scores on a measure have to be the same. For example, if you examine a group of first-time offenders, there will be no variation in the number of offenses in their criminal records. By definition, because they are all first-time offenders, the standard deviation (and the variance) will be 0.
2. The size of the standard deviation (and the variance) is dependent on both the amount of dispersion in the measure and the units of analysis that are used. When cases are spread widely from the mean, there is more dispersion and the standard deviation will be larger. When cases are tightly clustered around the mean, the standard deviation will be smaller.

   Similarly, when the units of analysis in the measure are large, the standard deviation will reflect the large units. For example, if you report the standard deviation of police salaries in a particular city in dollars, your standard deviation will be larger than if you reported those salaries in units of thousands of dollars. If the standard deviation is 3,350 in dollars, the standard deviation would be 3.35 using the unit of thousands of dollars.
3. Extreme deviations from the mean have the greatest weight in constructing the standard deviation. What this means is that here, as with the mean, you should be concerned with the problem of outliers. In this case, the effect of outliers is compounded because they affect not only the mean itself, which is used in computing the standard deviation, but also the individual deviations that are obtained by subtracting the mean from individual cases.

[3] As discussed in footnote 1, SPSS and many other computer packages would provide a slightly different result, based on the use of a correction of -1 in the denominator.

Table 5.8

Duncan SEI for Bribery and Antitrust Offenders

CATEGORY	N	$\overline{X}$	s
Bribery	83	59.27	19.45
Antitrust	112	61.05	11.13
Total (Σ)	195		

The standard deviation is a useful statistic for comparing the extent to which characteristics are clustered or dispersed around the mean in different samples. For example, in Table 5.8, a sample of offenders convicted of antitrust violations is compared to a sample of offenders convicted of bribery. The characteristic examined is social status, as measured by the interval-scale Duncan socioeconomic index (SEI).[4] The index is based on the average income, education, and prestige associated with different occupations. The mean Duncan scores for these two samples are very similar (61.05 for antitrust violators; 59.27 for bribery offenders), but the standard deviation for those convicted of bribery is about twice that of those convicted of antitrust violations.

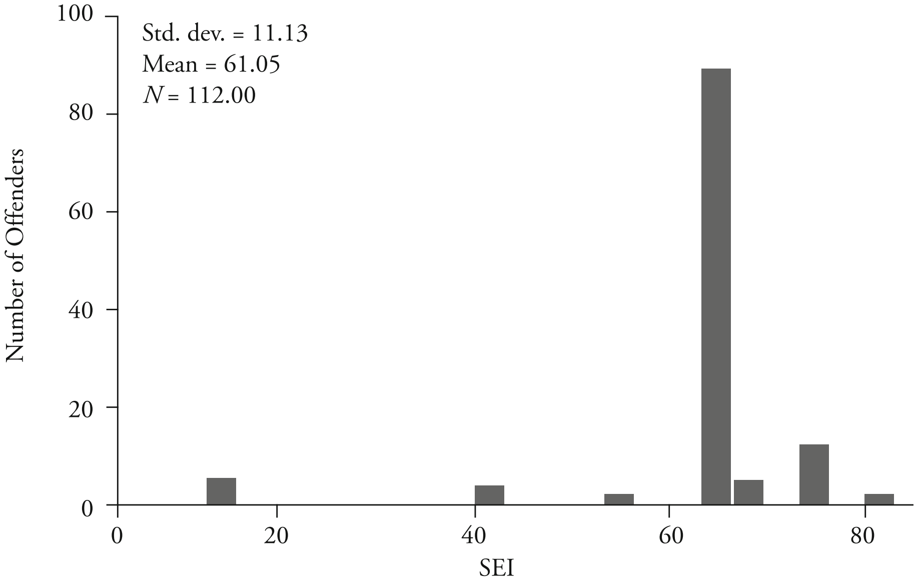
Figure 5.1 illustrates why these two samples yield similar means but very different standard deviations. The scores for most antitrust offenders are clustered closely within the range of 55 to 75. For bribery offenders, in contrast, the scores are much more widely spread across the distribution, including many more cases between 75 and 90 and below 50. What this tells us is that the antitrust sample includes a fairly homogeneous group of offenders, ranking on average relatively high on the Duncan socioeconomic index. Bribery is a much more diverse category. Although the means are similar, the bribery category includes many more lower- and higher-status individuals than does the antitrust category.

The Coefficient of Relative Variation

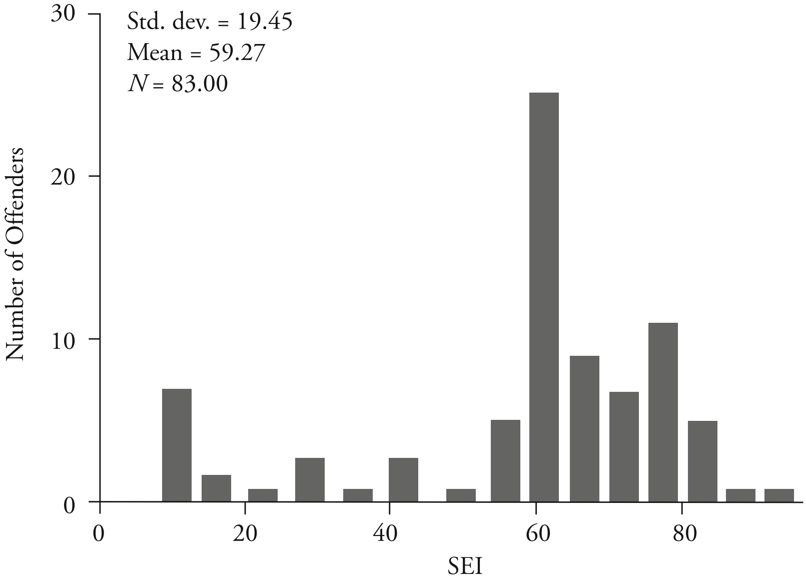
For the data on bribery and antitrust offenders in Table 5.8, in which the means of the two groups are fairly similar, a direct comparison of standard deviations provides a good view of the differences in dispersion. When the means of two groups are very different, however, this comparison may not be a fair one. If the mean Duncan score for one group was 10 and for the other was 50, we might expect a larger standard deviation in the latter group simply because the mean was larger and there was

[4] See Albert J. Reiss, *Occupations and Social Status* (New York: Free Press, 1961).

Figure 5.1 *Socioeconomic Indices*



(a) *Antitrust Offenders*



(b) *Bribery Offenders*

greater potential for dispersion. Similarly, if two measures use different units of analysis—for example, dollars and number of offenses—a direct comparison of standard deviations does not make sense.

One solution to this problem is to use the **coefficient of relative variation** (CRV). The coefficient of relative variation looks at the size of the standard deviation of a measure relative to the size of its mean:

$$\text{CRV} = \frac{s}{\bar{X}} \qquad \textbf{Equation 5.6}$$

In the example of the SEI for antitrust offenders, we divide the standard deviation (11.13) by the mean (61.05) to obtain a CRV of 0.18, meaning that the standard deviation is about one-fifth the size of the mean. Because the CRV expresses dispersion in a measure in a standardized form relative to the mean, we can compare the CRV across measures that have widely different means and standard deviations.

A measure that has a CRV of 1, for example, may be considered to include much greater relative variation than is found in our sample of antitrust offenders.

Working It Out

$$\begin{aligned}\text{CRV} &= \frac{s}{\bar{X}}\\ &= \frac{11.13}{61.05}\\ &= 0.1823\end{aligned}$$

A Note on the Mean Deviation

The standard deviation allows us to measure dispersion in interval scales, taking into account the deviation from the mean of each case in our sample or population. But it is not the only measure that allows us to do this. The **mean deviation** takes a similar approach, but relies on absolute values, rather than squaring, to overcome the fact that the sum of the deviations from the mean equals 0. When you take the absolute value of a number, you ignore its sign. Accordingly, −8 and 8 both have an absolute value of 8; in mathematical notation, $|-8| = |8| = 8$.

The equation for the mean deviation is similar to that for the variance. The only difference is that we take the absolute value of the

Table 5.9 Mean Deviation for Crime Calls at Hot Spots in a Year

HOT SPOT NUMBER	NUMBER OF CALLS	DEVIATIONS FROM THE MEAN $\lvert X_i - \overline{X}\rvert$
1	2	$\lvert 2 - 21.5 \rvert =$ 19.5
2	9	$\lvert 9 - 21.5 \rvert =$ 12.5
3	11	$\lvert 11 - 21.5 \rvert =$ 10.5
4	13	$\lvert 13 - 21.5 \rvert =$ 8.5
5	20	$\lvert 20 - 21.5 \rvert =$ 1.5
6	20	$\lvert 20 - 21.5 \rvert =$ 1.5
7	20	$\lvert 20 - 21.5 \rvert =$ 1.5
8	24	$\lvert 24 - 21.5 \rvert =$ 2.5
9	27	$\lvert 27 - 21.5 \rvert =$ 5.5
10	29	$\lvert 29 - 21.5 \rvert =$ 7.5
11	31	$\lvert 31 - 21.5 \rvert =$ 9.5
12	52	$\lvert 52 - 21.5 \rvert =$ 30.5
		Total (Σ) = 111.0

difference between each score and the mean, rather than the square of the difference:

$$\text{Mean deviation} = \frac{\sum_{i=1}^{N} \left| X_i - \overline{X} \right|}{N}$$

Equation 5.7

Using the data on crime calls in hot spots from Table 5.4, we take the following steps to obtain the mean deviation. We first take the absolute value of the difference between each score and the mean (see Table 5.9). We then sum up the 12 scores. Notice that we obtain a positive number now (111), and not 0, because we are taking the absolute values of the differences. Dividing this sum by the number of cases, *N*, we get a mean deviation of 9.25.

Working It Out

$$\begin{aligned}\text{Mean deviation} &= \frac{\sum_{i=1}^{N} \left| X_i - \overline{X} \right|}{N}\\ &= \frac{\sum_{i=1}^{12} \left| X_i - 21.5 \right|}{N}\\ &= \frac{111}{12}\\ &= 9.25\end{aligned}$$

The mean deviation and the standard deviation provide similar estimates of dispersion, but the mean deviation here is a bit smaller than the standard deviation of 12.38 that we calculated earlier. Which is the better estimate of dispersion? In some sense, the mean deviation is more straightforward. It simply looks at the average deviation from the mean. In obtaining the standard deviation, we first must square the deviations; then later, to return our result to units similar to those of the original distribution, we must take the square root of the variance.

Given our rule that we should use the least complex presentation that is appropriate to answering our research question, you may wonder why the standard deviation is almost always preferred over the mean deviation in criminal justice research. As you will see in the next few chapters, the answer is that the standard deviation is relevant to a number of other statistics that we use in analyzing and describing data.

Chapter Summary

Measures of dispersion describe to what extent cases are distributed around the measure of central tendency. They tell us just how typical the typical case is.

There are several measures of dispersion for nominal and ordinal scales. Proportions and percentages describe the extent to which cases are concentrated in the modal category. The **variation ratio** (VR) describes the extent to which cases are spread outside the modal category. A proportion of 1 (VR of 0) means that all the cases are in the modal category. This represents the least possible amount of dispersion. The value for the greatest possible dispersion can be determined by calculating the minimum possible value of the modal category and then translating that into a proportion or VR value. These measures can, in principle, be used with ordinal-level data, but the results may be misleading, as they take into account only the value of the mode. As an alternative, the **index of qualitative variation** (IQV) is a standardized measure that takes into account variability across all the categories of a nominal- or ordinal-level variable. An IQV of 0 means that there is no variation; an IQV of 100 means that there is maximum variation across the categories.

A different set of measures is used to measure dispersion for interval and ratio scales. The **range** measures the difference between the highest and lowest scores. It has the advantage of simplicity, but it uses very little information (only two scores) and the scores used are taken from the two extremes. It is also very sensitive to outliers. A researcher may instead choose to measure the range between, say, the 95th and the 5th percentile. Such measures, however, are still based on minimal information and thus are generally considered unstable statistics. A more stable statistic for measuring dispersion in interval-level scales is the **variance**. The variance is the sum of the squared deviations of each score from the mean divided by the number of cases. The **standard deviation** (*s*) is the square root of the variance. The advantage of the standard deviation over the variance is that the results are more easily interpreted. If all the scores in a sample are the same, *s* will be 0. The more widely the scores are spread around the mean, the greater will be the value of *s*. Outliers have a considerable impact on the standard deviation.

Comparing the standard deviations of means is problematic when the means are very different or when their units of measurement are different. An alternative measure, the **coefficient of relative variation** (CRV), enables comparisons among samples with different means. A less often used measure of dispersion for interval scales is the **mean deviation.** The mean deviation is computed by taking the sum of the absolute values of the deviations from the mean divided by the number of cases.

Key Terms

coefficient of relative variation A measure of dispersion calculated by dividing the standard deviation by the mean.

index of qualitative variation A measure of dispersion calculated by dividing the sum of the possible pairs of observed scores by the sum of the possible pairs of expected scores (when cases are equally distributed across categories).

mean deviation A measure of dispersion calculated by adding the absolute deviation of each score from the mean and then dividing the sum by the number of cases.

range A measure of dispersion calculated by subtracting the smallest score from the largest score. The range may also be calculated from specific points in a distribution, such as the 5th and 95th percentile scores.

standard deviation A measure of dispersion calculated by taking the square root of the variance.

variance (s^2) A measure of dispersion calculated by adding together the squared deviation of each score from the mean and then dividing the sum by the number of cases.

variation ratio A measure of dispersion calculated by subtracting the proportion of cases in the modal category from 1.

Symbols and Formulas

$N_{\text{modal cat.}}$	Number of cases in the modal category
N_{total}	Total number of cases
N_{obs}	Number of cases observed in each category
N_{exp}	Number of cases expected in each category
s	Standard deviation
s^2	Variance

To calculate the proportion of cases falling in the modal category:

$$\text{Proportion} = \frac{N_{\text{modal cat.}}}{N_{\text{total}}}$$

To calculate the percentage of cases falling in the modal category:

$$\text{Percentage} = \frac{N_{\text{modal cat.}}}{N_{\text{total}}} \times 100$$

To calculate the variation ratio:

$$\text{VR} = 1 - \left(\frac{N_{\text{modal cat.}}}{N_{\text{total}}}\right)$$

To calculate the index of qualitative variation:

$$\text{IQV} = \left(\frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} N_{\text{obs}_i} N_{\text{obs}_j}}{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} N_{\text{exp}_i} N_{\text{exp}_j}}\right) \times 100$$

To calculate the variance:

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}$$

To calculate the standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}}$$

To calculate the coefficient of relative variation:

$$\text{CRV} = \frac{s}{\overline{X}}$$

To calculate the mean deviation:

$$\text{Mean deviation} = \frac{\sum_{i=1}^{N} \left| X_i - \overline{X} \right|}{N}$$

Exercises

5.1 Police records for 105 rape victims were analyzed to determine whether any prior relationship existed between the victim and the offender. The results were as follows:

Spouse	41
Family member other than spouse	14
Acquaintance	22
No prior relationship	28

a. Calculate the modal proportion and the variation ratio.

b. What are the minimum and maximum possible values for the variation ratio?

c. Calculate the index of qualitative variation.

5.2 As part of a larger study on the influence of delinquent peers, a sample of high school youth were asked how much they wanted to be like their best friend. The responses were coded as follows: in every way, 26; in most ways, 36; in some ways, 41; and not at all, 8.

a. Calculate the variation ratio for these data.

b. Calculate the index of qualitative variation for these data.

5.3 People convicted of minor traffic offenses who appeared in the magistrate's court of a given locality on a given day were sentenced as follows: conditional discharge, 14; fine, 35; and license disqualification, 11.

a. Calculate the variation ratio.

b. Calculate the index of qualitative variation.

c. Why do these two results differ?

5.4 A sample of women was drawn from town A, and another sample was drawn from town B. All the women were asked how safe or unsafe they felt walking alone at night in their neighborhoods. The results were recorded on a scale as follows: totally unsafe (town A: 40; town B: 25), quite unsafe (town A: 29; town B: 23), quite safe (town A: 10; town B: 15), and totally safe (town A: 21; town B: 17).

a. For each town, describe the typical case, using an appropriate measure of central tendency. Explain why this is the best measure of central tendency for these data.

b. For each town, describe how typical the typical case is, using an appropriate measure of dispersion. Explain why this is the best measure of dispersion for these data.

c. In comparing the measures of central tendency and dispersion for the two towns, what conclusions may be drawn about the attitudes of the women?

5.5 For a sample of 12 offenders convicted of weapons violations, the length of prison sentence in months was recorded as:

6 6 2 12 36 48 60 24 24 20 18 15

a. Calculate the range for these data.

b. Calculate the mean and the variance for these data.

5.6 A group of 20 prisoners in a particular cell block were tested on their knowledge of the rules of the institution. The marks (out of a possible 70) were as follows:

31 28 27 19 18 18 41 0 30 27
27 36 41 64 27 39 20 28 35 30

a. Calculate the range.

b. Remove the largest and smallest scores. Calculate the range for the remaining cases.

c. How do you account for the difference between the values of the above two measures of dispersion?

5.7 Police crack a drug ring of 18 suppliers and discover that of the 18, only 4 have no previous convictions for drug- or theft-related offenses. Eight of those arrested have 1 previous conviction, and the others have 2, 3, 4, 5, 6, and 8, respectively.

a. Calculate the mean and the standard deviation of the 18 cases.

b. If each of the drug suppliers is convicted this time around, does the extra conviction on each of their criminal records affect the mean or the standard deviation in any way? Explain your answer.

5.8 Use the data collected from tests of prisoners' knowledge of institution rules in Exercise 5.6.

a. Calculate the mean and the standard deviation for these data.

b. If you remove the two most extreme scores, 0 and 64, what are the new mean and standard deviation?

c. How do you account for this effect?

5.9 When asked about how often in the last year they drank more than four beers in one evening, a sample of college students reported the following:

Number of Times	Frequency
0	187
1	213
2	162
3	94
4	71
5	55
6	39
7	12
8	9
9	5
10	13

a. Calculate an appropriate measure of dispersion for these data. Explain why this measure is most appropriate for these data.

b. Describe one way these data could be recoded to reduce the number of categories. Calculate an appropriate measure of dispersion for the recoded data and explain why this measure is most appropriate.

5.10 A researcher takes a sample of shop owners in Tranquiltown and a sample of shop owners in Violenceville and asks them to estimate the value of goods stolen from their shops in the past 12 months. The mean figure is \$11.50 (s = \$2.50) for Tranquiltown and \$4,754.50 (s = \$1,026.00) for Violenceville. When the study is published, the mayor of Violenceville protests, claiming that the mean sum for his town is a misleading figure. Because the standard deviation for Violenceville is much bigger than that for Tranquiltown, he argues, it is clear that the mean from Violenceville is a much less typical description of the sample than the mean from Tranquiltown.

a. What statistic might help the researcher to refute this criticism? Why?

b. Calculate this statistic for each town. What should the researcher conclude?

5.11 A researcher investigating differences in violence among preschool-age boys and girls found that the average number of violent acts per week was 7.6 ($s = 4.8$) for boys and 3.1 ($s = 1.9$) for girls.

a. Calculate the coefficient of relative variation for boys and for girls.

b. How can the coefficient of relative variation be used to compare these two groups? What does it tell you?

Computer Exercises

Similar to the measures of central tendency discussed in Chapter 4, there are several ways to obtain measures of dispersion for interval-level variables in SPSS and Stata (Neither program computes measures of dispersion, such as the index of qualitative variation, for nominal and ordinal variables.).

SPSS

The same two commands that we used to obtain measures of central tendency—DESCRIPTIVES and FREQUENCIES—will be used to obtain measures of dispersion. Since you have already used these commands, our discussion here is focused on the default measures of dispersion and additional options available for measuring dispersion.

The DESCRIPTIVES command allows you to compute the standard deviation and variance. The default is to compute the standard deviation, as well as minimum and maximum values. To obtain the variance and/or range, add the option /STATISTICS = to the command line:

```
DESCRIPTIVES VARIABLES = variable_names
   /STATISTICS = MEAN STDDEV VARIANCE RANGE MIN MAX.
```

The abbreviations should be clear, but STDDEV is the standard deviation, MIN is the minimum value, and MAX is the maximum value.

In computing the variance, SPSS uses a correction for the bias of sample measures of variance and dispersion: 1 is subtracted from the *N* in the denominator of Equations 5.4 and 5.5. The correction is used primarily as a tool in inferential statistics and is discussed in greater detail in Chapter 10. Though it is our view that the uncorrected variance and standard deviation should be used in describing sample statistics, many researchers report these statistics with the correction factor for sample estimates. When samples are larger, the estimates obtained with and without the correction are very similar, and thus it generally makes very little substantive difference which approach is used.

The FREQUENCIES command provides similar measures of dispersion through use of the /STATISTICS = option. You may request the standard deviation, variance, range, minimum, and maximum. In addition to these measures of dispersion, the FREQUENCIES command has the option of computing percentiles. Since calculation of percentiles by hand can be very difficult (and error prone), this is a nice feature. There are three options for calculating percentiles in SPSS: quartiles (25th, 50th, and 75th percentiles) or cut points for some number of equally spaced groups (e.g., 10th, 20th, …, 90th percentiles) by using the NTILES = option, as well as specific percentiles that you may be interested in (e.g., 5th, 95th, 99th) by using the /PERCENTILES = option. The general form, including all the various options, is:

```
FREQUENCIES VARIABLES = variable_names

  /PERCENTILES = list_of_percentiles

  /NTILES = number_of_equally_spaced_groups

  /STATISTICS = STDDEV VARIANCE RANGE MINIMUM MAXIMUM

  MEAN MEDIAN MODE.
```

If you are not interested in percentiles or equally spaced groups (e.g., quartiles, deciles, etc.), then omit those lines from the command.

Specific examples of the use of each of these commands are provided in the accompanying SPSS syntax file for Chapter 5 (Chapter_5.sps).

Stata

To obtain measures of dispersion in Stata, the same two commands used in obtaining measures of central tendency will again be used. To obtain detailed measures of dispersion, it is best to simply add the **detail** option to the command line:

```
summarize variable_names, detail
```

This will produce the variance, standard deviation, and numerous percentiles. Stata makes the correction to the computation of the variance in the same way that SPSS does by subtracting 1 from the total sample size.

The **tabstat** command is also useful in trimming out the unnecessary pieces of information and printing only those items we want:

```
tabstat variable_names, statistics (median mean min max
range sd var cv p#)
```

Where min is the minimum, max is the maximum, sd is the standard deviation, var is the variance, cv is the coefficient of variation (not available in SPSS), and p# refers to a specific percentile we may be interested in and included for as many percentiles as we would like reported. For example, if we were interested in the 5th, 10th, 90th, and 95th percentiles, we would include **p5 p10 p90 p95** on the command line (and within the parentheses).

Specific examples of the use of each of these commands are provided in the accompanying Stata do file for Chapter 5 (Chapter_5.do).

Problems

1. Enter the data from Exercise 5.6 on the 20 prisoners' test scores.

   a. What is the range?

   b. What are the 5th and 95th percentiles? What is the range between the 5th and 95th percentiles?

   c. How does your answer to part b compare to your answer to part b in Exercise 5.6?

2. Enter the data from Exercise 5.7. (Be sure that you have 18 lines of data, since there are 18 observations listed in the question.)

   a. What are the mean and the standard deviation? How does the standard deviation differ from the value you calculated in Exercise 5.7?

   b. To add 1 to each person's number of prior convictions, you will need to create a new variable.

   In SPSS:

   ```
   COMPUTE new_var_name = old_var_name + 1.
   EXECUTE.
   ```

   In Stata:

   gen new_var_name = old_var_name +1

   What are the mean and standard deviation for this new variable? What has changed? What has remained the same?

3. Open the NYS data file (nys_1.sav, nys_1_ student.sav, or nys_1.dta).

   a. Choose five of the delinquency measures. What are the quartiles for each of the five types of delinquency?

   b. What is the range between the 25th and 75th percentiles for each delinquency measure? (This difference is known as the "inter-quartile range.") What do the differences in inter-quartile ranges appear to indicate about the dispersion of these different measures self-reported delinquency?

c. What number of delinquent acts would mark the 15 % least delinquent youth? The 20 % most delinquent youth?

d. Compute the coefficient of relative variation for each delinquency variable (Remember, this measure is available in the **tabstat** command in Stata, but will need to be calculated by hand if using SPSS.). What do the values of the coefficient of relative variation for each delinquency variable indicate about the relative dispersion of these variables?