MEASURES OF DISPERSION (VARIABILITY)

While measures of *central tendency* indicate what value of a variable is (in one sense or other, e.g., mode, median, mean), "average" or "central" or "typical" in a set of data, measures of *dispersion* (or *variability* or *spread*) indicate (in one sense or other) the extent to which the observed values are "spread out" around that center — how "far apart" observed values typically are from each other or from some average value (in particular, the mean). Thus:

(a) if all cases have identical observed values (and thereby also all have the average value), dispersion is zero;

(b) if most cases have observed values that are quite "close together" (thereby also quite "close" to the average value), dispersion is low (but greater than zero); but

(c) if many cases have observed values that are quite "far apart" from many others (or from the average value), dispersion is high.

A *measure of dispersion* provides a summary statistic that indicates the magnitude of such dispersion and, like a measure of central tendency, is a univariate statistic.

Because dispersion is concerned with how "close together" or "far apart" observed values are (i.e., with the magnitude of the *intervals* between them), it should be apparent that the notion of dispersion make sense — and measures of dispersion are defined — only for *interval* (or ratio) variables. (There is one exception: a very crude measure of dispersion called the *variation ratio,* which is defined for ordinal and even nominal variables. It will be discussed briefly in the Answers & Discussion to PS #7.)

There are two principal types of measures of dispersion: *range measures* and *deviation measures*.

Range Measures

Range measures are based on the distance between (relatively) "extreme" values observed in the data and are conceptually connected with the *median* as a measure of central tendency (See the data illustrating *Percentiles, the Median, and Ranges* on the back page of the Handout #6 on *Measures of Central Tendency*.)

The ("total" or "simple") *range* is the *maximum* (highest) value observed in the data (the value of the case at the *100th percentile*) minus the *minimum* (lowest) value observed in the data (the value of the case at the *0th percentile*) — that is, the "distance" or "interval" between the values of these two extreme cases. (Note that this may be less than the range of the *possible values* of the variable, since logically possible extreme values may not be observed in actual data; for example, the variable LEVEL OF TURNOUT has logically possible values ranging from 0% to 100%, but in U.S. Presidential elections, the range of observed values [as conventionally measured, i.e., as Total Vote for President divided by Voting Age Population] over the past 60 years or so ranges from a minimum observed of about 48% (in 1996) to about 64% (in 1960). The problem with the (total or simple) range as a measure of dispersion is that it depends on the values of just two cases — cases that by definition have atypical (and perhaps extraordinarily atypical) values. In particular, the range

makes no distinction between a *polarized* distribution in which almost all observed values are close to either the minimum or maximum values and a distribution in which almost all observed values are bunched together but there are a few extreme *outliers*. Also the range is undefined for theoretical distributions that are "open-ended" (the technical term is *asymptotic*), like the *normal distribution* (that we will take up in the next topic) or the upper end of an income distribution type of curve (see PS #5C). Therefore other variants of the range measure that do not reach entirely out to the extremes of the frequency distribution are often used in place of the total range.

The *interdecile range* is the value of the case that stands at the *90th percentile* of the distribution minus the value of the case that stands at the *10th percentile* — that is, the "distance" or "interval" between the values of these two less extreme cases. In like manner, the *interquartile range* is the value of the case that stands at the *75th percentile* of the distribution minus the value of the case that stands at the *25th percentile*. (The *first quartile* is the median observation among all cases that lie *below* the overall median and the *third quartile* is the median observation among all cases that lie *above* the overall median. In these terms, the interquartile range is third quartile minus the first quartile.)

We have previously used a range measure in a special context. The handout on Random Sampling said the following:

> Suppose the Gallup Poll takes a random sample of *n* respondents and reports that the President's current approval rating is 62% and that this sample statistic has a margin of error of ± 3 %. Here is what this means: if (hypothetically) Gallup were to take a great many random samples of the same size n from the same population (e.g., the American VAP on a given day), the different samples would give different statistics (approval ratings), but 95% of these samples would give approval ratings within 3 percentage points of the true population parameter.

Thus, if our data is the list of sample statistics produced by the (hypothetical) "great many" random samples, the margin or error specifies the *range* between the value of the sample statistic that stands at the *97.5th percentile* minus the sample statistic that stands at the *2.5th percentile* (so that 95% of the sample statistics lie within the range). Specifically (and letting *P* be the value of the population parameter) this range is $(P + 3\%) - (P - 3\%) = 6\%$, i.e., twice the margin error.

Deviation Measures

Deviation measures are based on average deviations from some average value. (Recall the discussion of *Deviations from the Average* in Handout #6 on *Measures of Central Tendency*.) Since we are dealing with interval variables, we can calculate means, and deviation measures are typically based on the mean deviation from the mean value. Thus the usual deviation measures are conceptually connected with the *mean* as a measure of central tendency.

Suppose we have a variable *X* and a set of cases numbered 1,2, . . . , *n*. Let the observed value of the variable in each case be designated x_1, x_2, etc. Thus:

$$\textit{mean of } X \;=\; \bar{x} \;=\; \frac{x_1 + x_2 + \ldots + x_n}{n} \;=\; \frac{\sum x}{n} \;.$$

The deviation from the mean for a representative case i is $(x_i - \bar{x})$. If almost all of these deviations are small (if almost all cases are close to the mean value), dispersion is small; but if many of these deviations are large (if many cases are much above or below the mean), dispersion is large. This suggests we could construct a measure D of dispersion that would simply be the average (mean) of all the deviations:

$$D = \frac{(x_1 - \bar{x}) + (x_2 - \bar{x}) + ... + (x_n - \bar{x})}{n} = \frac{\sum (x_i - \bar{x})}{n} .$$

But this will not work, because some of the deviation are positive and others are negative and, as we saw earlier (Handout #6, point (d) under *Deviations from the Average*), these positive and negative deviations necessarily "balance out" and add up to zero, i.e., for any distribution of observed values $\sum(x_i - \bar{x}) = 0$.

A practical way around this problem is simply to ignore the fact that some deviations are negative while others are positive by averaging the *absolute values* of the deviations (in effect, by ignoring the negative sign before each negative deviation):

$$MD = \frac{\sum |x_i - \bar{x}|}{n} .$$

This measure (called the *mean deviation*) tells us *the average* (mean) *amount that the values for all cases deviate* (regardless of whether they are higher or lower) *from the average* (mean) *value*. Indeed, this is an intuitive, understandable, and perfectly reasonable measure of dispersion, and it is occasionally used in research.

However, statisticians are mathematicians, and they dislike this measure because the formula is mathematically messy by virtue of being "non-algebraic" (in that it ignores negative signs). Therefore statisticians, and most researchers, use another slightly different deviation measure of dispersion that is "algebraic," and that makes use of the fact that the *square* of *any (positive or negative)* number (i.e., the number multiplied by itself) other than zero is itself always *positive*. This formula is based on finding the average of the *squared deviations*; since these are all non-negative, they do not "balance out." This measure of dispersion is called the *variance* of the variable.

$$\textit{Variance of } X = Var(X) = s^2 = \frac{\sum (x_i - \bar{x})^2}{n} .$$

That is, the variance is the *average squared deviation from the mean*. Remember from Handout #6 (point (e) under *Deviations from the Average*) that the average squared deviation from the mean value of X is smaller than the average squared deviation from any other value of X. The variance is the usual measure of dispersion in *statistical theory*, but it has a drawback when researchers want to *describe the dispersion in data in a practical way*. Whatever units the original data (and its average values and its mean dispersion) are expressed in, the variance is expressed in the *square* of those units, and thus it doesn't make much intuitive or practical sense. This can be remedied by finding the (positive) *square root* of the variance (which takes us back to the original units). This measure of dispersion is called *standard deviation* of the variable:

$$\textit{Standard Deviation of } X = SD(X) = s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}.$$

In order to interpret a standard deviation, or to make a plausible estimate of the SD of some data, it is useful to think of the mean deviation because (i) it is easier to estimate the magnitude of the mean deviation and (ii) *the standard deviation has approximately the same numerical magnitude as the mean deviation*. More precisely, given any distribution of data, the standard deviation is *never less* than the mean deviation; it is equal to the mean deviation if the data is distributed in a maximally "polarized" fashion; otherwise the SD is somewhat larger — typically about 20-50% larger.

Sample Estimates of Population Dispersion

Random sample statistics that are percentages or averages provide *unbiased* estimates of the corresponding population parameters. However, sample statistics that are dispersion measures provide estimates of population dispersion that are *biased* (at least slightly) downward. This is most obvious in the case of the range; it should be evident that a *sample range* is almost always smaller, and can never be larger, than the corresponding *population range*. The sample standard deviation (or variance) is also biased slightly downward . (While the SD of a particular sample can be larger than the population SD, sample SDs are *on average* slightly smaller than the corresponding population SDs). However, the sample SD can be adjusted to provide an *unbiased* estimate of the population SD; this adjustment consists of dividing the sum of the squared deviations by $n - 1$, rather than by n. (Clearly this adjustment makes no practical difference unless the sample is quite small. Notice that if you apply the SD formula in the event that you have just a *single* observation in your sample, i.e., $n = 1$, it must give SD = 0 regardless of what the observed value is. More intuitively, you can get no sense of how much dispersion there is in a population with respect to some variable until you observe at least two cases and can see how "far apart" they are.) This is why you will often see the formula for the variance and SD with an $n - 1$ divisor (and scientific calculators often build in this formula). However, for POLI 300 problem sets and tests, you should use the formula given in the previous section of this handout.

Dispersion in Ratio Variables

Given a ratio variable (e.g. income), the interesting "dispersion question" may pertain not to the *interval* between two observed values or between an observed value and the mean value but to the *ratio* between the two values. (For example, one household "poverty level" is defined as *one half* the median household income, and households with more than *twice* the median income are sometimes characterized as "well off." The average compensation of CEOs today is about *250 times* that of the average worker, whereas 50 years it was only about *40 times* that of the average worker.) The degree of dispersion in ratio variables can naturally be referred to as the degree *inequality*. One ratio measure of dispersion/inequality is the *coefficient of variation*, which is simply *the standard deviation divided by the mean*. Another is the *Gini Index of Inequality*, which is based on a comparison between the actual *cumulative distribution* when cases are ranked ordered from lowest

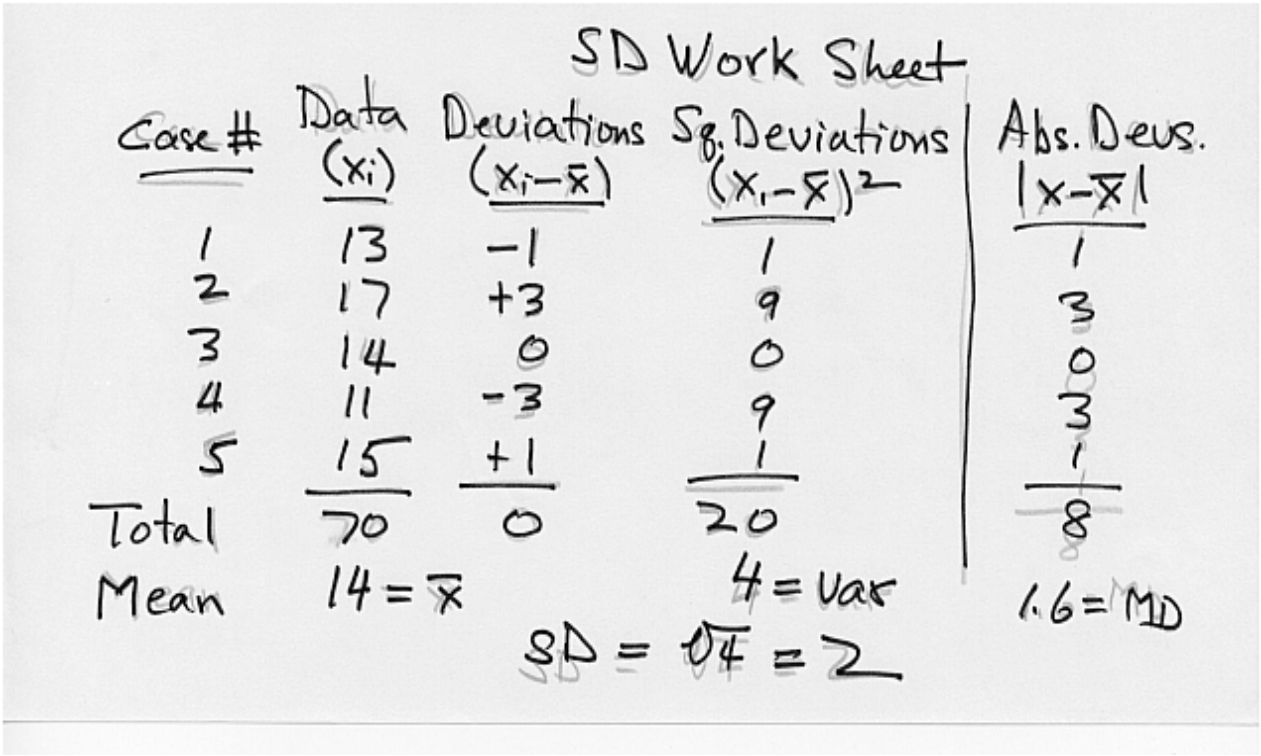
to highest value (e.g., from poorest to richest) and the cumulative distribution that would exist if all cases had the same value.

How to Compute a Standard Deviation

The formula for the standard deviation is: $SD(X) = s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$.

Here is how to use the formula.

1. Set up a worksheet like the one shown below.
2. In the first column, list the values of the variable X for each of the n cases. (This is the raw data.)
3. Find the *mean* value of the variable in the data, by adding up the values in each case and dividing by the number of cases.
4. In the second column, subtract the mean from each value to get, for each case, the *deviation from the mean*. Some deviations are positive, others negative, and (apart from rounding error) they must add up to zero; add them up as an arithmetic check.
5. In the third column, *square* each deviation from the mean, i.e., multiply the deviation by itself. Since the product of two negative numbers is positive, every squared deviation is non-negative, i.e., either positive or (in the event a case has a value that coincides with the mean value).
6. Add up the squared deviations over all cases.
7. Divide the sum of the squared deviations by the number of cases; this gives the *average squared deviation from the mean*, commonly called the *variance*.
8. The *standard deviation* is the (positive) *square root* of the variance. (The square root of x is that number which when multiplied by itself gives x.)


SD Work Sheet
Case # | Data (x_i) | Deviations ($x_i - \bar{x}$) | Sq. Deviations $(x_i - \bar{x})^2$ | Abs. Devs. $|x - \bar{x}|$
1 | 13 | −1 | 1 | 1
2 | 17 | +3 | 9 | 3
3 | 14 | 0 | 0 | 0
4 | 11 | −3 | 9 | 3
5 | 15 | +1 | 1 | 1
Total | 70 | 0 | 20 | 8
Mean | 14 = $\bar{x}$ | | 4 = Var | 1.6 = MD
SD = $\sqrt{4}$ = 2