3.6: Introduction to Standard Deviations and Calculations

Sum of Squares

Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution. Using the mean as the measure of the middle of the distribution, we can see how far, on average, each data point is from the center. The data from a fake Quiz 1 are shown in Table 3.6.1. The mean score is 7.0:

$$\frac{\Sigma X}{N} = \frac{140}{20} = 7$$

Therefore, the column "$X - \overline{X}$" contains deviations (how far each score deviates from the mean), here calculated as the score minus 7. The column "$(X - \overline{X})^2$" has the "Squared Deviations" and is simply the previous column squared.

There are a few things to note about how Table 3.6.1 is formatted, as this is the format you will use to calculate standard deviation. The raw data scores (X) are always placed in the left-most column. This column is then summed at the bottom to facilitate calculating the mean (simply divided this number by the number of scores in the table). Once you have the mean, you can easily work your way down the middle column calculating the deviation scores. This column is also summed and has a very important property: it will always sum to 0 (or close to zero if you have rounding error due to many decimal places). This step is used as a check on your math to make sure you haven't made a mistake. THIS IS VERY IMPORTANT. When you mis-calculate an equation, it is often because you did some simple math (adding or subtracting) incorrectly. It is very useful when equations have these self-checking points in them, so I encourage you to use them. If this column sums to 0, you can move on to filling in the third column of squared deviations. This column is summed as well and has its own name: the **Sum of Squares (abbreviated as** SS and given the formula $\sum(X - \overline{X})^2$). As we will see, the Sum of Squares appears again and again in different formulas – it is a very important value, and this table makes it simple to calculate without error.

Table 3.6.1: Calculation of Variance for Quiz 1 scores.

X	$X - \overline{X}$	$(X - \overline{X})^2$
9	2	4
9	2	4
9	2	4
8	1	1
8	1	1
8	1	1
8	1	1
7	0	0
7	0	0
7	0	0
7	0	0
7	0	0
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1

X	$X-\overline{X}$	$(X-\overline{X})^2$
6	-1	1
5	-2	4
5	-2	4
$\Sigma = 140$	$\Sigma = 0$	$\Sigma = 30$

The calculations in Table $3.6.1$ can be done by hand, but it is also very easy to set up the data in any spreadsheet program and learn the simple commands to make the spreadsheet do the simple math. As long as you tell it what to do with the correct numbers, then your results will be correct. You can also use the memory function in graphing calculators to save the the data set, and run some of the more common mathematical functions. Using spreadsheets and your graphing calculator to do the math also saves problems with rounding since the devices keep all of the decimals so you only have to round your final result. This statistics textbook will not go into explanations on how to use software (like spreadsheets, calculators, or more sophisticated statistical programs), but much that is easily accessible (like spreadsheets on Excel or Google) are relatively easy to learn to use

Variance (of a Sample)

Now that we have the Sum of Squares calculated, we can use it to compute our formal measure of average distance from the mean, the variance. The variance is defined as the average squared difference of the scores from the mean. We square the deviation scores because, as we saw in the Sum of Squares table, the sum of raw deviations is always 0, and there's nothing we can do mathematically without changing that.

The population parameter for variance is σ^2 ("sigma-squared") and is calculated as:

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

Notice that the numerator that formula is identical to the formula for Sum of Squares presented above with $\overline{X}$ replaced by μ. Thus, we can use the Sum of Squares table to easily calculate the numerator then simply divide that value by N to get variance. If we assume that the values in Table $3.6.1$ represent the full population, then we can take our value of Sum of Squares and divide it by N to get our population variance:

$$\sigma^2 = \frac{30}{20} = 1.5$$

So, on average, scores in this population are 1.5 squared units away from the mean. This measure of spread is much more robust (a term used by statisticians to mean resilient or resistant to) outliers than the range, so it is a much more useful value to compute.

But we won't do much with variance of a population. Instead, we'll focus on variance of a sample. The sample statistic used to estimate the variance is s^2 ("s-squared"):

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

This formula is very similar to the formula for the population variance with one change: we now divide by $N-1$ instead of N. The value $N-1$ has a special name: the **degrees of freedom** (abbreviated as df**)**. You don't need to understand in depth what degrees of freedom are (essentially they account for the fact that we have to use a sample statistic to estimate the mean ($\overline{X}$) before we estimate the variance) in order to calculate variance, but knowing that the denominator is called df provides a nice shorthand for the variance formula: SS/df.

Going back to the values in Table $3.6.1$ and treating those scores as a sample, we can estimate the sample variance as:

\[s^{2}=\dfrac{30}{20-1}=1.58 \n\nonumber \]

Notice that this value is slightly larger than the one we calculated when we assumed these scores were the full population. This is because our value in the denominator is slightly smaller, making the final value larger. In general, as your sample size N gets bigger, the effect of subtracting 1 becomes less and less. Comparing a sample size of 10 to a sample size of 1000; 10 – 1 = 9, or 90% of the original value, whereas 1000 – 1 = 999, or 99.9% of the original value. Thus, larger sample sizes will bring the estimate

of the sample variance closer to that of the population variance. This is a key idea and principle in statistics that we will see over and over again: larger sample sizes better reflect the population.

The Big Finish: Standard Deviation

The standard deviation is simply the square root of the variance. This is a useful and interpretable statistic because taking the square root of the variance (recalling that variance is the average squared difference) puts the standard deviation back into the original units of the measure we used. Thus, when reporting descriptive statistics in a study, scientists virtually always report mean and standard deviation. Standard deviation is therefore the most commonly used measure of spread for our purposes.

The sample statistic follows the same conventions and is given as s:

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

The sample standard deviation from Table $3.6.1$ is:

$$s = \sqrt{\frac{30}{20 - 1}} = \sqrt{\frac{30}{19}} = \sqrt{1.58} = 1.26$$

We'll practice calculating standard deviations, then interpreting what the numbers mean. Because in behavioral statistics, it's not about the numbers. We never end with a number, we end with a conclusion (which can be as simple as a sentence, or can be several paragraphs). Social scientists want to know what the numbers *mean* because we use statistical analyses to answer real questions.

Contributors

- Foster et al. (University of Missouri-St. Louis, Rice University, & University of Houston, Downtown Campus)
- Dr. MO (Taft College)