Chapter 7: Dispersion Of Data & Standard Scores

Objectives for This Chapter

- Enter the second dimension of statistics.
- Understand the concept of dispersion among scores.
- Define and compute the measures of dispersion covered in the chapter–the range, variance, and standard deviation (definitional and computational formulas).
- Understand how our statistics estimate population parameters.
- Calculate variance and standard deviation using your calculator's statistics functions.
- Define and compute *z* scores or standard scores.
- Understand the importance of z or standard scores.

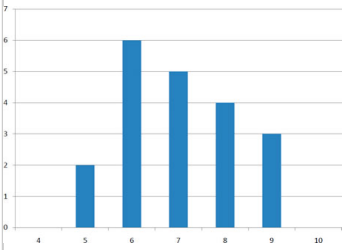
The second Dimension of Statistics: The Dispersion Zone

If you go out on the street and ask people if they understand the concept of *average*, most of them will give an explanation that's pretty good. Everybody knows about the first dimension. It's corporeal. We talk about averages so much in our world that they're almost concrete. Ahhh...to walk in a world where averages are good enough. I concede that it's quite blissful, but it's one dimensional thinking and highly misleading, as this video explains.

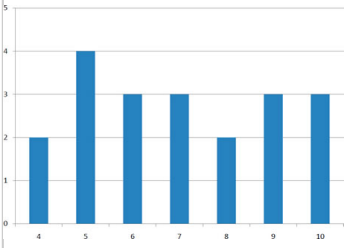
1/12 Variability And Rainfall

It's time to go deeper in the world of statistics. You're traveling through the second dimension -- a dimension not only of sight and sound but of mind. A journey into a wondrous land whose boundaries are that of imagination. That's a signpost up ahead. Your next stop--the Dispersion Zone! It's not for the faint of heart, but you know what you signed up for.

What does dispersion really mean?



(../../images/dispersion/spread1.jpg)What is dispersion? Dispersion refers to how "spread out" a group of scores is. To see what we mean by spread out, consider these graphs. These graphs represent the scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, you can see that the distributions are quite different. Specifically, the scores on Quiz 1 are more densely packed, and those on Quiz 2 are more spread out. The differences among students was much greater on Quiz 2 than on Quiz 1. Click on images to enlarge.



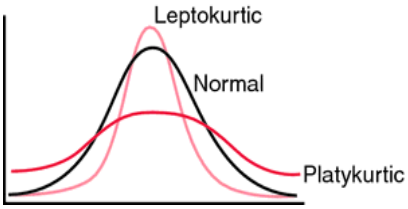
(../../images/dispersion/spread2.jpg)The terms *variability, spread*, and *dispersion* are synonyms, and refer to how spread out a distribution is. Just as in the section on *central tendency* we discussed measures of the center of a distribution of scores, in this chapter we will discuss measures of the variability of a distribution. The most common measures of dispersion are *range, variance* and *standard deviation*. Each of these measures is a descriptive statistic that can be used to help others understand what is going on in a group of numbers. In the next few sections, we will look at each of these three measures of variability in more detail.

Now, some variability is good, it's natural and normal. In fact, if we give a performance test to a group of people, we expect dispersion. If the test is too easy and everyone gets the highest score, then their is no dispersion. Similarly, if the test is too hard and everyone gets a 0, then there again is no dispersion. We say that such tests are afflicted with the dreaded *ceiling* and *floor* effects. Not only are such effects unnatural (kind of like the extreme

mullet hair-cut from the 80's), but they make it difficult to analyze data. As we move farther in this course, you will see that variability is a necessary element if we're going to perform inferential statistical analyses.

When we collect data, what we're usually hoping for is a data set that could be described with a symmetric, bell-shaped curve. In other words, a normal distribution with normally distributed data. Bellisimo!

Of course, the sad truth is that variability can take on a number of different looks. As we saw in a previous chapter, the shape of a distribution can lean to one side or another--a trend that we refer to as skew. But there is also an issue that we refer to as kurtosis. Sometimes there is (../../images/dispersion/kurtosis.png)very low variability in a distribution. This results in what we call a *leptokurtic* curve. This is a very high curve wherein all of the data points are tightly compacted. In this image, you see an example of a leptokurtic trend.



On the other hand, data can be spread out from the lowest to the highest possible scores in a distribution. We refer to such a data trend as a *platykurtic* distribution. There are lots of reasons for platykurtic distributions, but here are the two most important things to remember about excessive dispersion among the scores in a data set:

1. When dispersion increases, so to do our measures of dispersion (i.e., variance, standard deviation)
2. When dispersion is really large, it probably indicates increased *error* in the scores of the distribution. Where does the error come from? Lots of places--unreliable research conditions, invalid survey items, interviewer bias, participant characteristics, and on and on. More on this later.

For now, just remember that some variance...GOOD! Too much or too little variance...BAD! This is why it is so vital to understand dispersion.

How do we calculate range and what does it mean?

The *range* is the simplest measure of variability to calculate, and one you have probably encountered many times in your life. The range is simply the highest score minus the lowest score. Let's take a few examples. That's it.

What is the range of the following group of numbers: 10, 2, 5, 6, 7, 3, 4? Well, the highest number is 10, and the lowest number is 2, so 10 - 2 = 8. The range is 8. Let's take another example. Here's a dataset with 10 numbers: 99, 45, 23, 67, 45, 91, 82, 78, 62, 51. What is the range? The highest number is 99 and the lowest number is 23, so 99 - 23 equals 76; the range is 76. Now consider the two quizzes shown in the previous section. On Quiz 1, the lowest score is 5 and the highest score is 9. Therefore, the range is 4. The range on Quiz 2 was larger: the lowest score was 4 and the highest score was 10. Therefore the range is 6.

Hey, wait a minute! I thought that the average was the same for both of those quizzes. Exactly! You always have to keep in mind that measures of *central tendency* and *dispersion* describe different elements in a distribution. One is not tied to the other. This is why it's all the more important to include both types of measures when describing your data and why using only one can be misleading.

What is Variance and how do we calculate it?

You are definitely going to regret asking that question. But since you did ask...

X	X - M	$(X - M)^2$
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
6	-1	1
5	-2	4
5	-2	4
N=20	Σ=0	Σ=30

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, the **variance** is defined as *the average squared difference of the scores from the mean*. The data from a quiz are shown in this table. The mean score is 7.0. Therefore, the column "X-M" contains the score minus 7. The column "$(X-M)^2$" is simply the previous column squared.

One thing that is important to notice is that the average deviation from the mean is 0. This will always be the case. So, there is not a whole lot we can do with that column. What we can do is square that *differences column*, as you see in the table. To calculate variance for a population, we add up the squared-differences column and that gives us 30. When we divide by N, we get 1.5. Therefore, the variance is 1.5.

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

This is the formula for population variance where σ^2 (*little sigma squared*) is the variance, μ is the mean, and N is the number of numbers. For Quiz 1, μ = 7 and N = 20. I sometimes refer to this as the ***definitional formula***.

I won't show you a lot of population formulas in this web book, because these aren't the formulas that we use and having extra formulas tends to be more confusing than enlightening--for most students. However, it's good for you to recognize that *population parameters* and *sample statistics* are different, while they derive from very similar looking formulas.

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

A big visual distinction between the population and sample formulas is usually that population formulas have Greek symbols, whereas sample formulas usually have regular, old letters from the English alphabet. Compare this formula with the previous one. In the numerator, we still add up squared differences between scores and the sample mean. The big mathematical difference is in the denominator. This is what we will refer to as the ***definitional formula*** for sample variance.

A quick note on the numerator. The numerator in this formula is referred to as the **sum of squares** (SS) or *sum of squared devations from the mean*. That is, we find the difference between each score and the mean, square each of those differences and then add them up. This term (sum of squares) is commonly used in statistics books and I may use it from time to time in this class.

If the variance in a sample is used to **estimate** the variance in a population, then the population formula (the one

with Greek symbols) will *underestimate the variance* that is actually in the population. In other words, applying the population formula to data from a small group will yield a smaller variance number than we would have gotten if we had collected and summarized the data from the entire population. That's why we subtract 1 from the denominator in the sample formula. This slightly increases the variance number that we end up with, and theoretically our estimated variance is closer to the actual variance in the population. You will do a lot with N-1 as you learn how to apply statistics. Enlarge and watch this video for more insights from the Khan Academy (http://www.khanacademy.org).

4/12 Variance Definitional Formula

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

There is an alternate formula for calculating estimated variance. I'm going to call it the ***computational formula***. It looks a little more beastly, but it works a little better on some calculators and actually is easier to calculate by hand. Instead of needing three columns, you need only two. Let's go through a simple application of this formula.

The computational formula for estimated variance tells us that we need just a few pieces of information to complete the calculation.

1. **ΣX^2** tells us that we need to square the X values and then add those squared scores. To do this, we just create an X^2 column and add the numbers in that column.
2. **$(\Sigma X)^2$** tells us that we need to add up all of the scores and then square that sum.
3. We divide our answer in Step 2 by the number of scores (N) in our sample.
4. Then, subtract the answer you got in Step 3 from your answer to Step 1.
5. Divide your answer for Step 4 by N-1.

Note that in the above listed steps, numbers 1 through 4 comprise the *sum of squares* for the computational formula. Looks a little different than the previous sum of squares, but the resulting value is the same--at least if your calculations are correct.

Let's walk through a brief concrete example with this formula. Assume the scores 1, 2, 4, and 5 were sampled from a larger population. To estimate the variance in the population you would compute s^2 as follows:

$$\sum X^2 = 1^2 + 2^2 + 4^2 + 5^2 = 46$$

$$\frac{(\sum X)^2}{N} = \frac{(1+2+4+5)^2}{4} = \frac{144}{4} = 36$$

$$s^2 = \frac{(46-36)}{3} = 3.333$$

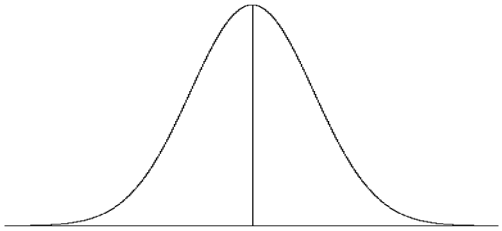
That's it! After you do a few of these, they become very easy. Just make sure that you organize your information properly in rows and columns and you set yourself up for success! Go into the next section and I'll have a step-by-step walkthrough that helps you calculate both variance and standard deviation, using the computational formula.

Oh...one little thing. This formula--the estimated variance--*there's only so much that we can do with it*. We really won't use it too much. We'll talk about this when we get to standard deviation.

Are we having fun yet?

No doubt, your mind is now blown. Your frontal and pre-frontal cortices are shredded. You are contemplating the wisdom of your current academic path. Hey, it's all good. Take a breather. Think about the really good times in your life--your last birthday, your first kiss, that guy at the end of your street who keeps calling you by the wrong name. And when you're ready...read on!

Okay. This is super important. Are you with me? You're not with me! Stop texting and close Facebook! There. When we calculate our sample statistics--mean, median, variance, standard deviation, etc.--**WE ARE CALCULATING ESTIMATES!** When we calculate the mean from a sample, this mean is just an estimate of the population mean. When we calculate variance from a sample, it is an estimate of the actual variance within the population. Aside from N-1, there are other things that we can do (more on this in a later chapter) to improve the accuracy of our estimates, but almost always there will be a difference between our estimate and what's actually going on in the population. But guess what...our estimate is the best that we've got.



Now, take a step back and remind yourself of what dispersion is all about. It's about how spread out a group of data is, relative to the average of that group. Keep this in mind. The formulas can seem a little complicated, but it always comes back to this basic idea. The *dispersion estimates that we are calculating simply tell us about the shape of a distribution*. Simple and important. As you learn more about statistics, this very basic idea will become more and more important. Nonethless, it will remain a very simple idea that you can easily picture.

What is standard deviation and how do we calculate it?

The **estimated standard deviation** (usually abbreviated as *s*) is simply the *square root of the variance*. Calculating variance is usually a step along the way to calculating the standard deviation. As you see in the formulas for s, the only difference (compared to the corresponding variance formulas) is the addition of the square-root sign. That's it. So long as you can remember that variance is the larger value and standard deviation is the smaller value, you're half way there.

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

Why not just stick with variance? Here's the problem: When we square the differences between the scores and the mean, our variance is no longer expressed in the units of our original group of scores. Just as the mean of the distribution tells us where a distribution is in balance, so too should our dispersion value accurately describe the spread of our scores. And squared units simply don't allow for this. So, we take the square root and arrive at standard deviation.

$$s = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N - 1}}$$

Expand and watch this tutorial to see a manual calculation of standard deviation, using the definitional formula.

3/12 Standard Deviation Definitional Formula

A couple of important things to keep in mind in your s calculations.

1. Standard deviation is *never negative*.
2. If our data are approximately a normal (bell-shaped) curve, then s will be close to *range divided by 4* (R/4). Dividing range by four (after you have completed your s calculations) can be a really helpful way to make sure that your value for s is "in the ballpark."

Calculating Standard Deviation From Frequency Distributions

One issue that we did not address in the section on variance was that of *frequency*. In the chapter on frequency distributions, you learned about different ways to organize data. Sometimes, calculating standard deviation from a frequency distribution is necessary or desirable. Compare this s formula with the one above. There are, of course, a few differences where score frequency or f is inserted into the formula.

Remember, that Big Sigma is in charge and always tells you to add whatever follows. Here are the steps that we would take to work through this:

$$s = \sqrt{\frac{\Sigma fX^2 - \frac{(\Sigma fX)^2}{N}}{N-1}}$$

1. **ΣfX^2** tells us that we need to square the X values, THEN multiply each of those squared values by their associated frequency and then add. To do this, we will have to make sure that we create FIRST, an X^2 column and SECONDLY, an fX^2 column.
2. **$(\Sigma fX)^2$** tells us that we need to multiply each score by its frequency, THEN add those products up and lastly, square our sum. To complete this step manually, you will have to create an *f*X column that can be added and squared.
3. We divide our answer in Step 2 by the number of scores (N) in our sample.
4. Then, subtract the answer you got in Step 3 from your answer in Step 1.
5. Divide your answer from Step 4 by N-1.
6. Take the square root of your answer to Step 5.

Expand and watch this interactive tutorial. In this tutorial, I talk about the computational and definitional formulas for estimated s. Additionally, I show the distinctions between manual calculations using a frequency distribution and not using a frequency distribution.

Personally, I'm not a big fan of calculating s using a frequency distribution. if you don't like calculating s with the frequency formula, all you have to do is expand your X column, so that each score is listed as many times as it should and then just delete the f column. For a lot of people, this is easier.

Okay. Reality check. What's the main point of all of this? To get a number (standard deviation) that provides an indication of how much spread we have in our group of scores. It's all about shape.

The Standard Deviation...On Drugs

In 2011, I was working out in the backyard and jammed my hand into a golden barrel cactus. Brilliant! I know. Anyhow, to get some of the thorns out, I had to actually go in for surgery. After being administered a general anesthetic, my wife thought that it would be amusing to see if my statistics perspicacity was still in tact. I thought it was. But really, when I tried to articulate the formula for calculating the standard deviation of a distribution, I butchered it. If you can figure out where I messed up and correct it, you are a true statistician, indeed!

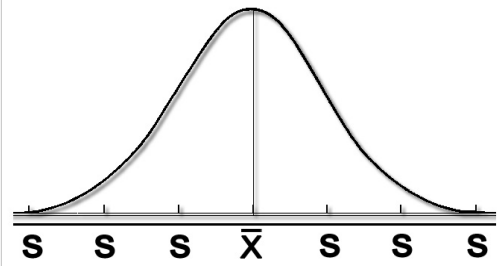
I will give **two extra credit points** to you, if you can: 1) identify where in the video I make a mistake articulating the computational, standard deviation formula, and 2) figure out how to correct that mistake. Write it up and submit no later than exam day for this module. Shouldn't take you more than a few sentences.

And now...enjoy Dr. B at his finest!

Standard Deviation On Drugs!

Where do Z Scores fit in to all of this?

A *standard score* or ***z* score** is the *deviation of a raw score from the mean in standard deviation units*. Each standard deviation unit represents a specific distance, expressed in the units of the sample scores. When we have normally distributed data, our deviation units will go out about three up and three down before we almost run out



of curve. You can see this in the picture.

Z scores can also be positive or negative. The sign of the *z* score tells the direction of the score relative to the mean: *Negative z*scores represent raw scores below the mean, and *positive z* scores indicate scores above the mean.

Calculating z-scores is a fairly straightforward process. To calculate the z score for a specific raw score, we just subtract the sample average and then divide the difference by our sample standard deviation. That's it. The process is easy but the z score itself can be applied in so many ways.

$$z = \frac{X - \bar{X}}{s}$$

Z scores are really powerful tools for accomplishing two things. First, z scores can be calculated for every score in a sample. What the z score tells us is the relative score of each individual's raw score. This is important, because sometimes a raw score of 50 on a math test is really bad. For example, if the class average was 90 and you got a score of 50, that would indicate that you need to spend more time studying. But what if the class average on the math test was 20? Then a score of 50, relatively speaking, would be pretty good.

Second, because z scores express one's performance relative to one's group, we refer to such scores as standard scores. Because the scores are standardized, we are able to compare between groups, with some degree of validity. For example, If someone gets a -1.5 z score on a math test and then a +2.3 on a French test, we could conclude that to some extent, this person is a better French than math student. Why? Because relative to the person's classes, he/she performed worse than average in math and significantly better than average in French.

$$X = zs + \bar{X}$$

Sometimes, we have to work the other way. That is, we know a z score and want to figure out the specific raw score with which it equates. In this case, we would follow this formula. To calculate, all we do is multiply the known z score by the known standard deviation value, and then add that product to the sample average.

It's kind of like converting money from one currency to another. For example, if we wanted to know which is more, 9,000 lire or 50 francs, we would have to convert both to a standard, using some economically derived factor. Let's use the U.S. dollar as the standard. There are about 5 francs to the dollar and 1,500 lire to the dollar. So, we can convert the lire to dollars by dividing 9,000 by 1,500 and the francs to dollars by dividing 50 by 5. Our lire convert to $6 and our francs to $10. So, we conclude that 50 francs is worth more than 9,000 lire.

Turns out that because they're so powerful, you will see z scores all over the place. Political polls, your cholesterol printout from the doctor's, sports statistics. Z scores really do get used a lot. If you want to be a wise and super consumer of statistical information, then you need to know about z scores. Even my good friend, Phil Zimbardo (the original Dr. Phil), thinks so.

Self Test

- Self-test for chapter (self_test_spread.pdf)
- Answers to self-test (answers_spread.pdf)

[<< back to top]

(index.html)

Some content adapted from other's work. See home page for specifics.

LAST UPDATED: 2015-09-30 6:32 PM

Mesa Community College | 1833 W. Southern Ave. Mesa, AZ 85202 | E-mail Address: dborman@mesacc.edu | Phone: (480) 461-7181 |

DEREK BORMAN: **PSYCHOLOGICAL SCIENCE**
MCC **PSYCHOLOGICAL SCIENCE** HOMEPAGE