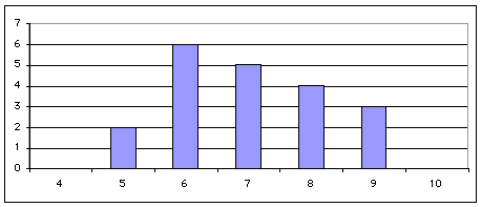
Measures of Variability*

David Lane

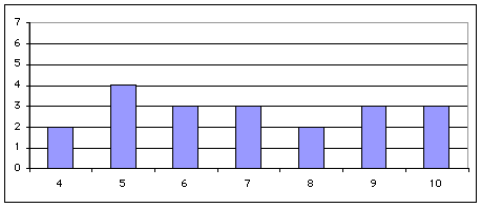
This work is produced by OpenStax-CNX and licensed under the
Creative Commons Attribution License 1.0†

1 What is Variability?

Variability refers to how "spread out" a group of scores is. To see what we mean by spread out, consider graphs in Figure 1. 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.



(a)



(b)

Figure 1: Bar charts of two quizzes. (a) Quiz 1 (b) Quiz 2

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. There are four frequently used

*Version 2.3: Jul 7, 2003 12:00 am -0500
†http://creativecommons.org/licenses/by/1.0

measures of variability, the range, interquartile range, variance, and standard deviation. In the next few paragraphs, we will look at each of these four measures of variability in more detail.

2 Range

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. 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 $99 - 23 = 76$; the range is 76. Now consider the two quizzes shown in Figure 1. On Quiz 1, the lowest score was 5 and the highest score was 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.

3 Interquartile Range

The **interquartile range** (IQR) is a range that contains the middle 50% of the scores in a distribution. It is computed as follows:

$$\text{IQR} = 75\text{th percentile} - 25\text{th percentile}$$

For Quiz 1, the 75th percentile is 8 and the 25th percentile is 6. The interquartile range is therefore 2. For Quiz 2, which has greater spread, the 75th percentile is 9, the 25th percentile is 5, and the interquartile range is 4. Recall that in the discussion of boxplots, the 75th percentile was called the upper hinge and the 25th percentile was called the lower hinge. Using this terminology, the interquartile range is referred to as the **H-spread**.

A related measure of variability is called the **semi-interquartile range**. The semi-interquartile range is defined simply as the interquartile range divided by 2. If a distribution is symmetric, the median plus or minus the semi-interquartile range contains half the scores in the distribution.

4 Variance

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 Quiz 1 are shown in Table 1: Calculation of Variance for Quiz 1 scores.. The mean score is 7.0. Therefore, the column "Deviation from Mean" contains the score -7. The column "Squared Deviation" is simply the previous column squares.

Calculation of Variance for Quiz 1 scores.

	Scores	Deviation from Mean	Squared Deviation
	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
Mean	7	0	1.5

Table 1

One thing that is important to notice is that the mean deviation from the mean is 0. This will always be the case. The mean of the squared deviations is 1.5. Therefore, the variance is 1.5. Analogous calculations with Quiz 2 show that it's variance is 6.7. The formula for the variance is:

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

where σ^2 is the variance, μ is the mean, and N is the number of numbers. For Quiz 1, $\mu = 7$ and $N = 20$.

If the variance in a sample is used to estimate the variance in a population, then the previous formula underestimates the variance and the following formula should be used:

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

where s^2 is the estimate of the variance and M is the sample mean. Note that M is the mean of a sample taken from a population with a mean of μ. Since, in practice, the variance is usually computed in a sample,

this formula is most often used. The simulation "estimating variance[1]" illustrates the bias in the formula with N in the denominator.

Let's take a concrete example. 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:

$$M = \frac{1+2+4+5}{4} = \frac{12}{4} = 3$$

$$s^2 = \frac{(1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2}{4-1} = \frac{4+1+1+4}{3} = \frac{10}{3} = 3.333$$

There are an alternate formulas that can be easier to use if you are doing your calculations with a hand calculator:

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

and

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

For this example,

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

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

$$\sigma^2 = \frac{46-36}{4} = 2.5$$

and

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

as with the other formula.

5 Standard Deviation

The **standard deviation** is simply the square root of the variance. This makes the standard deviations of the two quiz distributions 1.225 and 2.588. The standard deviation is an especially useful measure of variability when the distribution is normal or approximately normal (see Probability[2]) because the proportion of the distribution within a given number of standard deviations from the mean can be calculated. For example, 68% of the distribution is within one standard deviation of the mean and approximately 95% of the distribution is within two standard deviations of the mean. Therefore, if you had a normal distribution with a mean of 50 and a standard deviation of 10, then 68% of the distribution would be between $50-10 = 40$ and $50+10 = 60$. Similarly, about 95% of the distribution would be between $50 - 2 \times 10 = 30$ and $50 + 2 \times 10 = 70$. The symbol for the population standard deviation is σ; the symbol for an estimate computed in a sample is s. Figure 2 shows two normal distributions. Both distributions have means of 50. The blue distribution has a standard deviation of 5; the red distribution has a standard deviation of 10. For the blue distribution, 68% of the distribution is between 45 and 55; for the red distribution, 68% is between 40 and 60.

[1] "Estimating Variance Simulation" <http://cnx.org/content/m11210/latest/>
[2] "Remarks on the concept of "Probability"" <http://cnx.org/content/m10953/latest/>

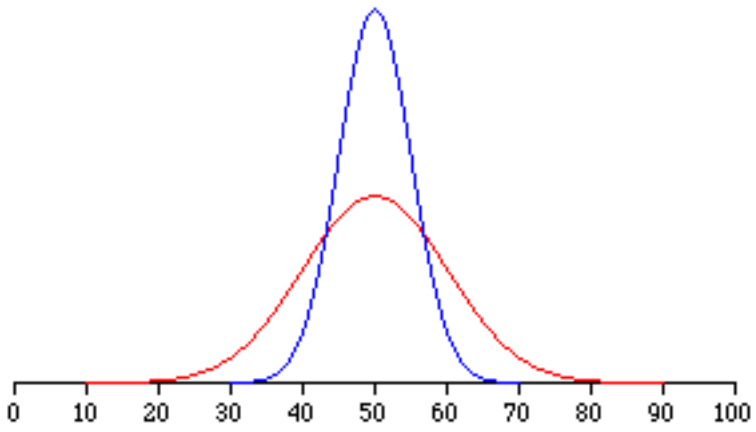


Figure 2: Normal distributions with standard deviations of 5 (blue line) and 10 (red line).