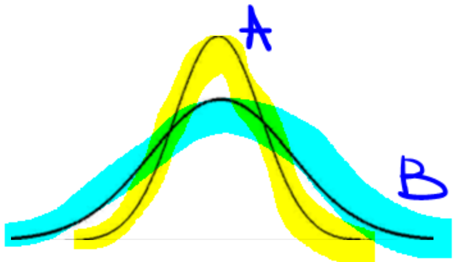
Section 1.3
Variance and Standard Deviation

Another important question we want to answer about data is about its spread or dispersion. If the outcomes cover a wide range, the spread is larger. If the outcomes are clustered around a single value, the spread is smaller.

Compare the two normal curves below, which variance would be larger? B



The **population variance**, σ^2 (read sigma-squared), is the average of the squared differences of the data values from the mean. The **population standard deviation**, σ, is the square root of the **population variance**.

Population Variance: $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

Population Standard Deviation: $\sigma = \sqrt{\sigma^2}$

where N is the number of values in our population, x_i represents each data value in the population, and μ is the population mean.

As stated above the standard deviation measures the same thing as the variance. Since we square the $(x_i - \mu)$ in the variance formula the units of x are squared, so to remedy this we take the square root.

Most of the time, however, we will not be working with the entire population but with a sample of the population. We compute the **sample variance**, s^2, and **sample standard deviation**, s, a little differently.

sample mean

Sample Variance: $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\underline{n-1}}$ **Sample Standard Deviation:** $s = \sqrt{s^2}$

where n is the number of values in our sample, x_i represents each data value in the sample, and $\bar{x}$ is the sample mean.

The commands in R for variance and standard deviation are:

var("name of data") sd("name of data")

Example 1: A teacher wants to decide whether or not to curve an exam. From her class of 300 students, she chose a sample of 10 students and their grades were:

72, 88, 85, 81, 60, 54, 70, 72, 63, 43

Find the mean, variance and standard deviation for this sample.

Commands:
grades=c(72,88,85,81,60,54,70,72,63,43)
mean(grades)
var(grades)
sd(grades)

Results:
$\bar{x} = 68.8$
$s^2 = 199.7333$
$s = 14.1327$

Now suppose the teacher decides to curve the grades by **adding 10 points to each score**. What is the new mean, variance and standard deviation?

Commands:
new = grades + 10
mean(new)
var(new)
sd(new)

Results:
$\bar{x} = 78.8$
$s^2 = 199.7333$
$s = 14.1327$

Conclusion: Since variance and standard deviation measure "spread" of the data set, if we added a constant value to each data value of the set, neither would not change. But, the mean would shift.

If we **multiplied each data value by the same amount**, it would affect the mean, variance and standard deviation.

Recall the population variance formula: $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

Note $(x_i - \mu)^2$ part of the formula. So if all of the data values are equal, then the variance (and standard deviation) would be zero.

Sometimes we want to compare the variation between two groups. The coefficient of variation can be used for this. The **coefficient of variation** is the ratio of the standard deviation to the mean. A smaller ratio will indicate less variation in the data.

$\frac{sd}{mean}$

Example 2: The following statistics were collected on two different groups of stock prices. What can be said about the variability of each portfolio?

	Portfolio A	Portfolio B
sample size	10	15
sample mean	$52.65	$49.80
sample std. dev.	$6.50	$2.95

Again, we can use R as a calculator.

Commands:	coefficient of variation:
Port. A: 6.50/52.65	0.1235
Port. B: 2.95/49.8	0.0592

Which one has less variability? B

If you were to make a decision on which to invest in, which one would it be?

B, since less variability