

Activity 9: Correlation

In the last activity, we used a chi-square test to determine if two categorical variables were independent. In other words, we tested the relationship between two categorical variables. In this activity, we'll learn how to measure and evaluate the relationship between two continuous variables using correlation coefficients.

NFL Malevolence: According to many movies, good guys wear white and bad guys wear black. Does the color of a person's clothing have a significant impact on that person's behavior? To address this question, researchers collected data measuring:

- The perceived malevolence of NFL team uniforms/logos
- The number of yards each NFL team was penalized (converted to a z-score)

A sample of the data is displayed in the following table:

NFL Team	Malevolence Rating X	Penalty Yards (z-score) Y
 1. L.A. Raiders	5.10	1.19
 5. Chicago Bears	4.68	0.29
 13. Green Bay Packers	4.00	-0.73
 15. Minnesota Vikings	3.90	-0.81
 24. Detroit Lions	3.38	0.04
 28. Miami Dolphins	2.80	-1.60

You can download this data at <http://www.bradthiessen.com/html5/data/nflm.csv>

1. Before we begin measuring the relationship between X and Y, let's look once again at the concept of variance:

$$\sigma_x^2 = E\left[(X - E[X])^2\right] = E\left[(X - \bar{X})^2\right] = \sum \frac{(X - \bar{X})^2}{n}$$

If we expand our notation, we can rewrite this as:

$$\sigma_{xx} = E\left[(X - E[X])(X - E[X])\right] = E\left[(X - \bar{X})(X - \bar{X})\right] = \sum \frac{(X - \bar{X})(X - \bar{X})}{n}$$

With this notation, what does a variance represent?

Source: <http://www.psych.cornell.edu/sec/pubPeople/tdg1/Frank%20&%20Gilo%2088.pdf>

Frank, M.G., & Gilovich, T. (1988). The dark side of self- and social perception: black uniforms and aggression in professional sports. *Journal of Personality and Social Psychology*, 54(1), 74-85.

2. We can take this concept of variation and apply it to two variables. When we do this, we're defining the *covariance*:

$$\sigma_{xy} = \text{Cov}(x, y) = E[(X - E[X])(Y - E[Y])] = E[(X - \bar{X})(Y - \bar{Y})] = \sum \frac{(X - \bar{X})(Y - \bar{Y})}{n}$$

Calculate the covariance for the subset of the NFL Malevolence data listed below. What does this covariance represent? Sketch a quick scatterplot of this data.

Malevolence X	Penalty Yards Y
5.10	1.19
4.68	0.29
4.00	-0.73
3.90	-0.81
3.38	0.04
2.80	-1.60
Mean = 3.977	Mean = -0.27

$$\sigma_{xy} = \sum \frac{(X - \bar{X})(Y - \bar{Y})}{n} = \frac{(5.10 - 3.977)(1.19 + 0.27) + \dots + (2.80 - 3.977)(-1.60 + 0.27)}{6} = 0.5741$$

3. Interpret the covariances for each of the following tables. Sketch a scatterplot for each table.

Malevolence Rating X	Wins Y
5.10	4
4.68	10
4.00	11
3.90	10
3.38	4
2.80	7
Mean = 3.997	Mean = 7.667

$$\sigma_{xy} = 0.0656$$

Penalty Yards (z-score) Y	Wins Y
1.19	4
0.29	10
0.04	4
-0.73	11
-0.81	10
-1.60	7
Mean = -0.270	Mean = 7.667

$$\sigma_{xy} = -1.182$$

4. Suppose we took each number in that last table and multiplied them by 100. What would happen to the value of the covariance? What's the smallest possible value for a covariance? What's the largest possible value? What would a covariance of zero represent?

5. While the covariance does measure the relationship between two continuous variables, it's measured in (units of X) x (units of Y). So, from our examples on the previous page, we found covariances of:

0.5741 malevolence penalty yards 0.0656 malevolence wins -1.182 penalty yard wins

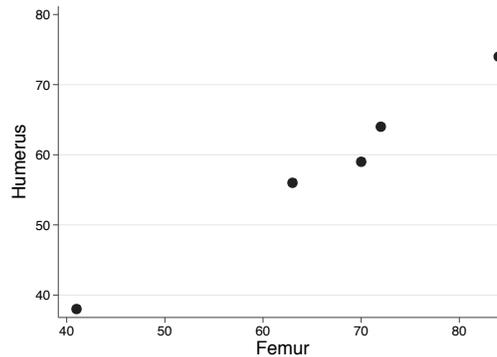
Because they're all measured in different units, we can't compare them. It would be nice to have a standardized measure of the relationship between two variables that only ranges from -1 to 1.

To get this, we can define the *correlation coefficient*: $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \sigma_{z_x z_y}$

As the formula shows, we divide the covariance by the product of the standard deviations of each variable. This, in essence, is the same as calculating the covariance of two variables after we convert each variable into z-scores.

Let's calculate a correlation coefficient by hand (and never calculate another one by hand again). Suppose we've unearthed fossilized femur and humerus bones from 5 unknown dinosaurs. The length of each bone is displayed in the following table:

Femur (cm)	Humerus (cm)
38	41
56	63
59	70
64	72
74	84
Mean = 58.2	Mean = 66.0
Std.Dev = 13.2	Std.Dev = 15.9



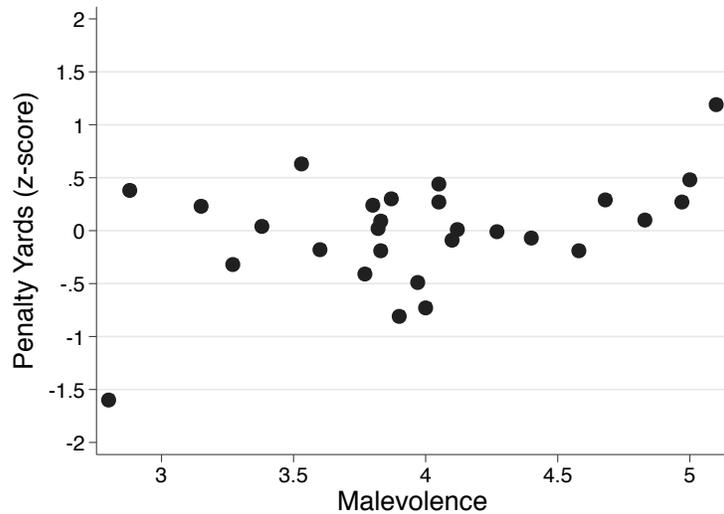
First, we convert each observation into a z-score. Fill-in the missing mean and standard deviation values.

Femur (cm)	Humerus (cm)	Zx * Zy
$(38-58.2) / 13.2 = -1.53$	$(41-66) / 15.9 = -1.57$	$-1.53 * -1.57 = 2.4021$
$(56-58.2) / 13.2 = -0.67$	$(63-66) / 15.9 = -0.19$	$-0.67 * -0.19 = 0.1273$
$(59-58.2) / 13.2 = 0.06$	$(70-66) / 15.9 = 0.25$	$0.06 * 0.25 = 0.0150$
$(64-58.2) / 13.2 = 0.44$	$(72-66) / 15.9 = 0.38$	$0.44 * 0.38 = 0.1672$
$(74-58.2) / 13.2 = 1.20$	$(84-66) / 15.9 = 1.13$	$1.20 * 1.13 = 1.3560$
Mean = _____	Mean = _____	Sum = 3.977
Std.Dev = _____	Std.Dev = _____	$r = 3.977 / 4$ $= 0.99425$

With these z-scores, we can calculate: $r_{xy} = \frac{\sum Z_x Z_y}{n-1}$

6. The correlation coefficient we just calculated is called the *Pearson product-moment correlation*. While you could look online for different formulas used to calculate this type of correlation coefficient, you should really use technology to speed things up. Use a computer or calculator to verify the correlation coefficient for the (full) NFL Malevolence data listed below:

Malevolence	Penalty Yards
5.10	1.19
5.00	0.48
4.97	0.27
4.83	0.10
4.68	0.29
4.58	-.19
4.40	-.07
4.27	-.01
4.12	0.01
4.10	-.09
4.05	0.44
4.05	0.27
4.00	-.73
3.97	-.49
3.90	-.81
3.87	0.30
3.83	-.19
3.83	0.09
3.82	0.02
3.80	0.24
3.77	-.41
3.60	-.18
3.53	0.63
3.38	0.04
3.27	-.32
3.15	0.23
2.88	0.38
2.80	-1.6



$$r_{xy} = 0.4298$$

How can we interpret this correlation coefficient? Is there a relationship between the perceived malevolence of an NFL team's uniform and their total yards penalized? If so, what's the nature of that relationship? Does having more malevolent uniforms cause NFL teams to be penalized more?

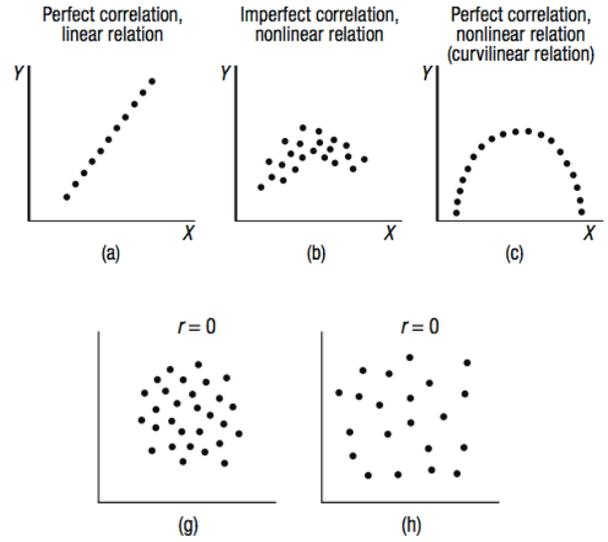
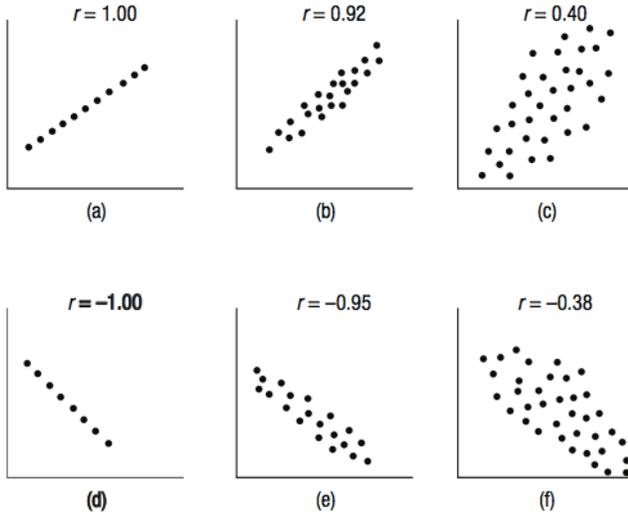
7. In 2002, Alan Lockwood published a letter in *Diabetes Care* detailing how he calculated a correlation coefficient between diabetes rates and pollution levels for each of the 50 states. He found the correlation to be $r = 0.54$. How would you interpret this correlation? Does pollution increase diabetes rates?

Lockwood was somewhat careful in writing his conclusion, stating, "... the correlation between air emissions and the prevalence of diabetes does not prove a cause-and-effect relationship; the significance of the relationship demands attention." In response, Mark Nicolich questioned whether the relationship even "demands attention." Using the same diabetes data, Nicolich found the following correlations:

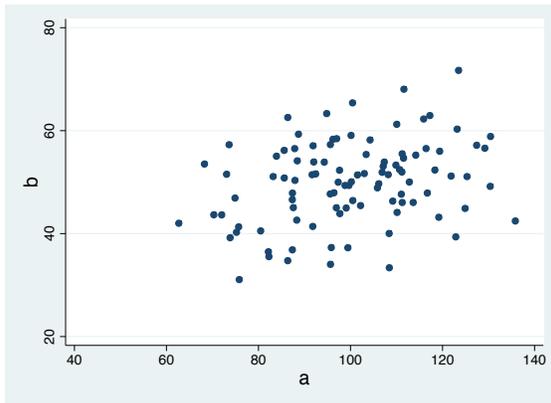
$$r_{\text{diabetes, alphabetized rank of states}} = 0.49 \quad \text{Explain the point in telling you this story.}$$

$$r_{\text{diabetes, latitude of state capital}} = -0.54$$

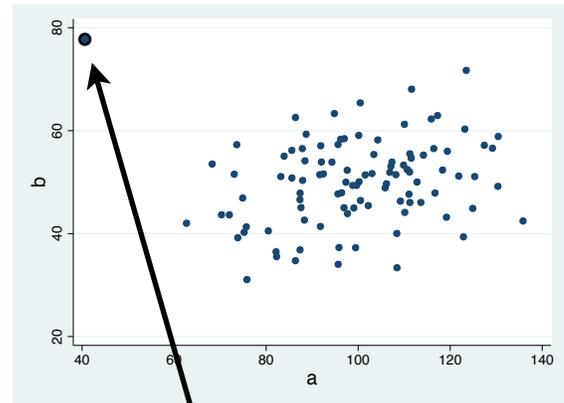
8. The correlation coefficient r measures the strength of a **linear** relationship. The following scatterplots display various values of correlations:



9. Outliers can greatly impact the value of a correlation coefficient. To demonstrate this, I simulated a dataset with a correlation of $r = 0.30$. I then added one outlier, displayed in the graph to the right. What do you think happened to the value of the correlation coefficient?

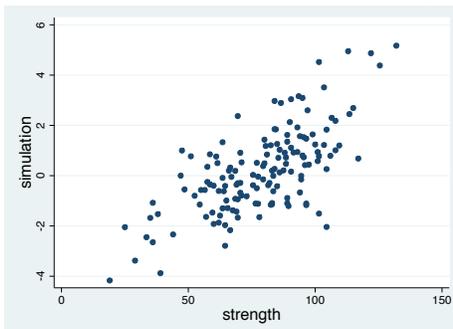


original data set
 $r = 0.30$

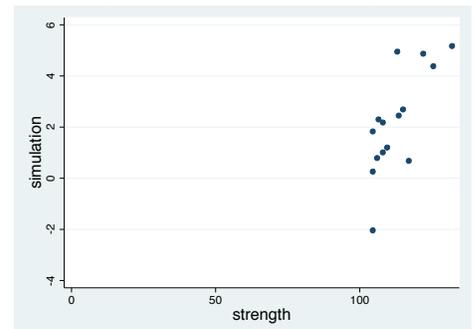


outlier added
 $r = ???$

10. Another way to manipulate the magnitude of a correlation coefficient is through range restriction. As an example, suppose we're interested in the relationship between the physical strength of construction workers and their job performance. The following scatterplot shows this data collected from 147 construction workers. The correlation was found to be $r = 0.686$. What do you think happens to the correlation if we only use data from the top 10% in physical strength?

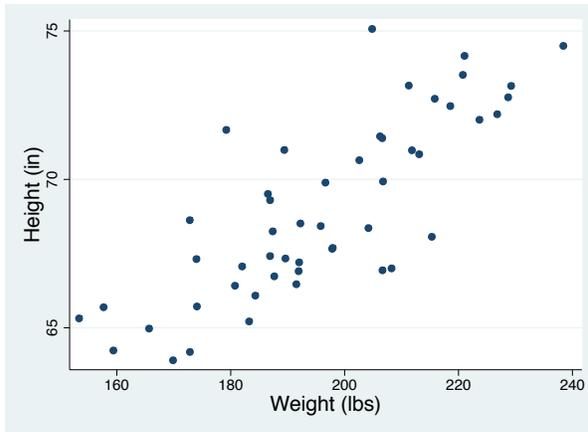


physical strength
(as shown on the right)?

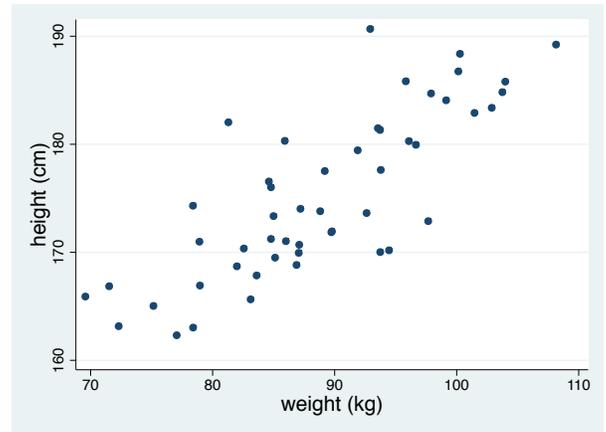


11. What happens to the correlation coefficient if we transform our variables? To investigate this, I simulated some data. The scatterplot on the left displays the relationship between the height and weight of 50 (fictitious) adult males. The data were simulated with a correlation of 0.80.

The scatterplot on the right shows the same data, but the units have been converted to metric.



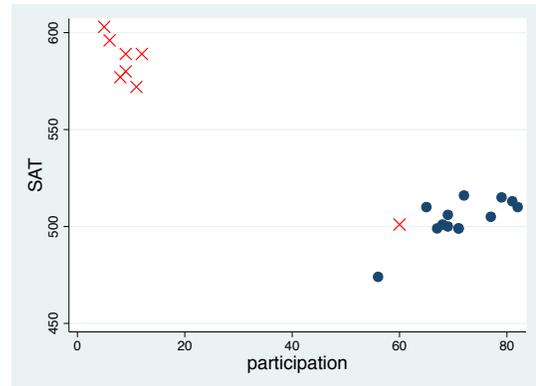
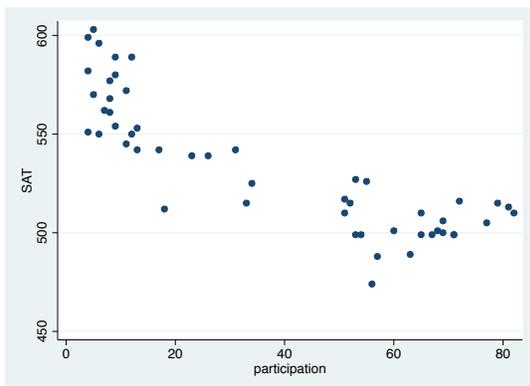
$r = 0.80$



$r = \underline{\hspace{2cm}}$

12. Newspapers often rank states according to their average SAT score (thereby "proving" that some states have better educational systems than others). For example, the average SAT math score in Iowa (where 5% of high school seniors take the exam) is 610. The average math score for Connecticut (where 82% of high school seniors take the exam) is only 510. Does this prove that Iowa has a superior educational system to Connecticut?

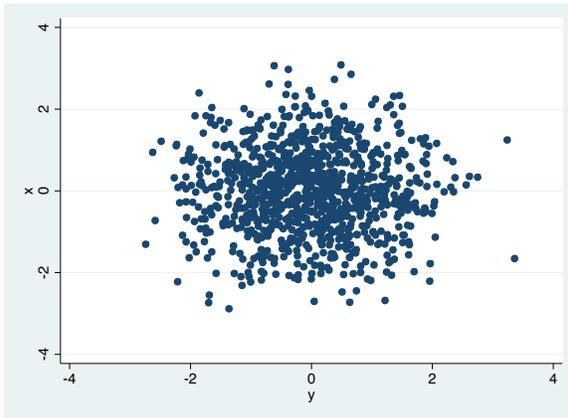
The following scatterplot displays the relationship between SAT scores and SAT participation rates for each of the 50 states. How would you describe the relationship between participation rates and SAT scores? What do you estimate the correlation to be?



On the right is the same data set. The only difference is that I labeled midwestern states as X and northeastern states as O. What conclusions can you draw? Does living in the Northeast cause you to have lower SAT scores? Does having a high percentage of students take the SAT cause your state to have low SAT scores?

13. Suppose I calculate a correlation between the height of 50 adult males and 50 randomly chosen numbers. What is the expected value of that correlation coefficient? Would you really expect to find this value for the correlation?

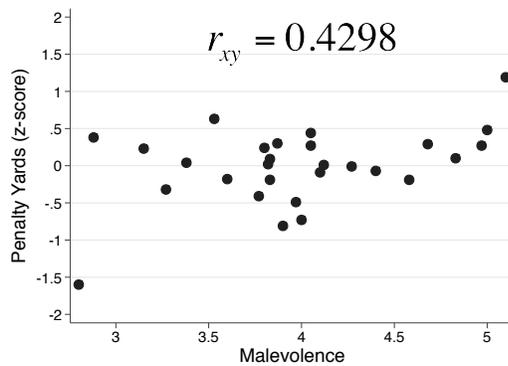
14. To see if our intuition is correct, I generated two sets of 1000 random numbers from a standard normal distribution. From the scatterplot, it appears as though there's no relationship between these two sets of random numbers. When I calculate a correlation coefficient, however, I find $r = -0.0185$.



From this, you should be convinced that any two sets of numbers will have a nonzero correlation coefficient.

Suppose we have two variables that we think are related. How large does the correlation of these two variables have to be to convince us that the correlation didn't simply happen by chance? In a future activity, we'll learn how to calculate a t-test for correlation coefficients. For now, let's briefly look at a randomization-based test.

15. Let's go back to the first example in this activity - the relationship between malevolence and penalty yards for NFL teams. As a reminder, we calculated a correlation of 0.4298 for this data.



Does the magnitude of this correlation coefficient convince us that the relationship between malevolence and penalty yards is statistically significant? In other words, if there were NO relationship between these two variables, is it possible for us to have found a correlation of 0.4298 (or higher) in our sample data?

Let's write out a null hypothesis in terms of the correlation coefficient that indicates a scenario in which malevolence and penalty yards have no relationship.

16. Under this null hypothesis, we're assuming penalty yards have no relationship with malevolence. If this were true, then there's no real reason for the Raiders, with the highest malevolence rating, to have 1.19 penalty yards. With no relationship between these variables, the Raiders were just as likely to have 0.48, 0.27, or -1.6 penalty yards.

Likewise, take a look at the Chicago Bears data. Their malevolence rating was 4.68. If malevolence has no relationship with penalty yards, the Bears were just as likely to get 1.19, 0.10, -0.81, or any other penalty yards as they were to get 0.29.

Malevolence	Penalty Yards	Team
5.10	1.19	Raiders
5.00	0.48	
4.97	0.27	
4.83	0.10	
4.68	0.29	Bears
4.00	-.73	Packers
3.97	-.49	
3.90	-.81	Vikings
3.38	0.04	Lions
2.80	-1.6	Dolphins

(not all data is displayed here)

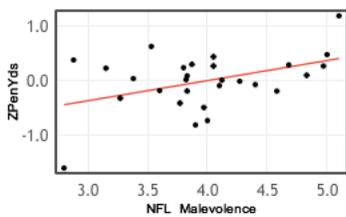
As we've seen before, this is the key concept behind a randomization test. Once we have our hypothesis, we randomize our data to reflect this hypothesis. We repeat this process thousands of times and then determine the likelihood of our statistic of interest. In this case, we're interested in a correlation coefficient of 0.4298 or higher.

I conducted this randomization test at http://lock5stat.com/statkey/randomization_2_quant/randomization_2_quant.html

For each randomization, I have the computer randomly randomly assign the penalty yards to each team. Here's what I got:

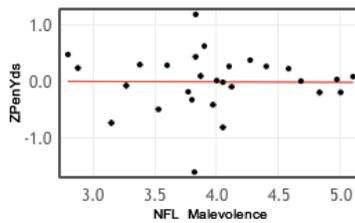
Original Sample

$n = 28, r = 0.43, slope = +0.368, intercept = -1.471$



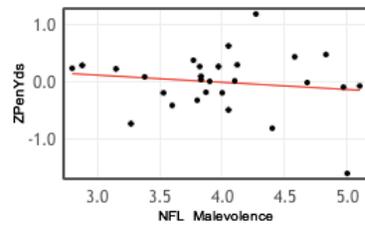
original
 $r = 0.4298$

$n = 28, r = -0.0093, slope = -0.0079, intercept = +0.028$



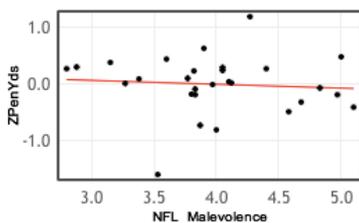
randomization #1
 $r = -0.0093$

$n = 28, r = -0.147, slope = -0.126, intercept = +0.498$



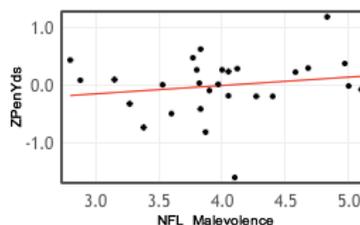
randomization #2
 $r = -0.147$

$n = 28, r = -0.08, slope = -0.069, intercept = +0.269$



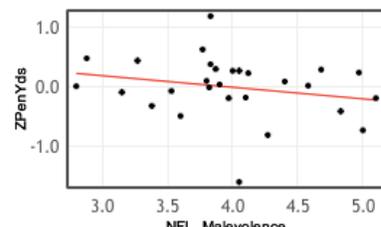
randomization #3
 $r = -0.08$

$n = 28, r = 0.169, slope = +0.145, intercept = -0.582$



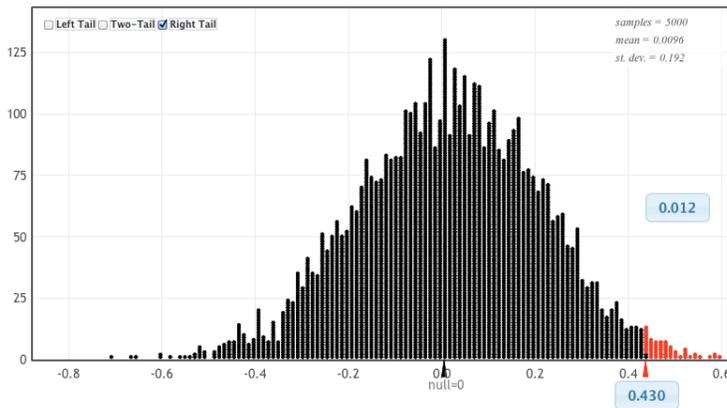
randomization #4
 $r = 0.169$

$n = 28, r = -0.227, slope = -0.195, intercept = +0.772$



randomization #5
 $r = -0.227$

17. I continued this process 5,000 times and then graphed all the correlation coefficients from my randomizations. How did I estimate the p-value? What conclusions can I make?



18. We could also use bootstrap methods to estimate a confidence interval for the correlation.

In the randomization-based process, we shuffle one of the variables (Y) while holding the other (X) constant.

With a bootstrap method, we keep our (X, Y) pairs together and randomly sample pairs with replacement. In this example, we would randomly sample 28 pairs of penalty yards and malevolence ratings.

Malevolence	Penalty Yards	Team
5.10	1.19	Raiders
4.68	0.29	Bears
5.10	1.19	Raiders
4.00	-.73	Packers
3.90	-.81	Vikings
3.38	0.04	Lions
4.68	0.29	Bears

(notice that we can get the same team multiple times)

With that bootstrap sample, we would calculate a correlation. We'd repeat this process many times. To estimate a 95% confidence interval, we'd identify the bootstrap correlations that cut-off the middle 95%. I conducted this bootstrap process in R and found:

```

lower_bound correlation upper_bound
value 0.02550679 0.429796 0.6931392

```

What can we conclude from this?

19. We could also run a t-test to test if a population correlation differs from zero. We'll see where this formula comes from in the next activity. For now, interpret the output from this t-test.

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

```

t = 2.4272, df = 26, p-value = 0.01122
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
0.1299369 1.0000000

```

20. As I wrote earlier, the correlation coefficient we've been learning about is called *Pearson's correlation coefficient*. We'll take a little bit of time to introduce two other correlation coefficients. These indices are commonly referred to as nonparametric correlation indices, because they do not require the data to follow any specific distribution.

The first index we'll investigate is *Spearman's Rho*. It's calculated using exactly the same formula as Pearson's coefficient. The only difference is we must first convert the data to ranks.

8 MBA graduates are studied to measure the strength of the relationship between their score on the GMAT, which they took prior to entering graduate school, and their grade point average while they were in the MBA program. Convert these scores into ranks. Then enter these ranks into your calculator and calculate the correlation.

<u>Student</u>	<u>GMAT</u>	<u>GMAT rank</u>	<u>GPA</u>	<u>GPA rank</u>
1	710		4.0	
2	610		4.0	
3	640		3.9	
4	580		3.8	
5	545		3.7	
6	560		3.6	
7	610		3.5	
8	530		3.5	

21. Another fairly popular nonparametric correlation coefficient is Kendall's Tau.

A new worker is assigned to a machine that manufactures bolts. Each day, a sample of bolts is examined and the percent defective is recorded. Do the following data indicate a significant improvement over time for that worker?

<u>Day</u>	<u>Percent</u>	<u>Concordant Below</u>	<u>Discordant Below</u>
1	1.6		
2	7.5		
3	7.7		
4	5.9		
5	5.2		
6	6.2		
7	5.3		
8	4.5		
9	4.9		
10	4.6		
11	3.0		
12	4.0		
13	3.7		

Project ideas: Distance correlation, Hoeffding's D; Polychoric, Polyserial, Biserial correlations