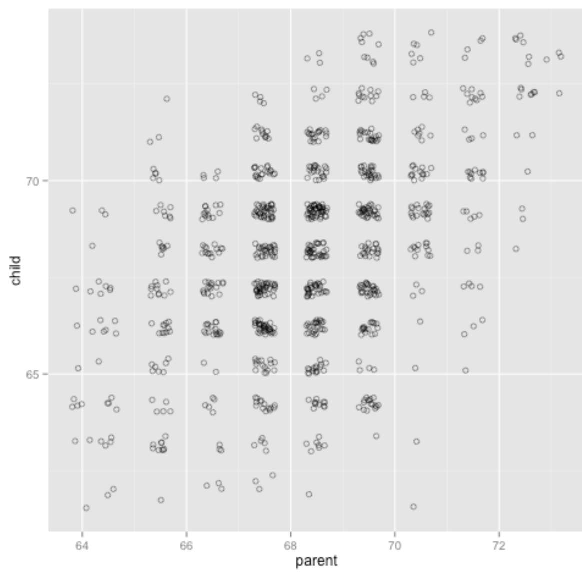
Assignment #9: Correlation (bootstrap interval and randomization test)

Scenario: Do taller parents typically have taller children? In 1885, Francis Galton recorded the heights of 928 children along with the average height of each child's parents. A scatterplot of this data is displayed below (with some *jitter* added to separate identical measurements):



You can download this data at:
http://www.bradthiessen.com/html5/data/galton.csv

1. Copy the data and paste it into the bootstrap confidence interval applet:
   http://lock5stat.com/statkey/bootstrap_2_quant/bootstrap_2_quant.html

   Record the correlation coefficient for this data: r = _______________

2. Generate at least 10,000 bootstrap samples and record the 95% confidence interval: __________________________

3. This time, paste the data into the randomization applet:
   http://www.rossmanchance.com/applets/RegShuffle.htm?hideExtras=2

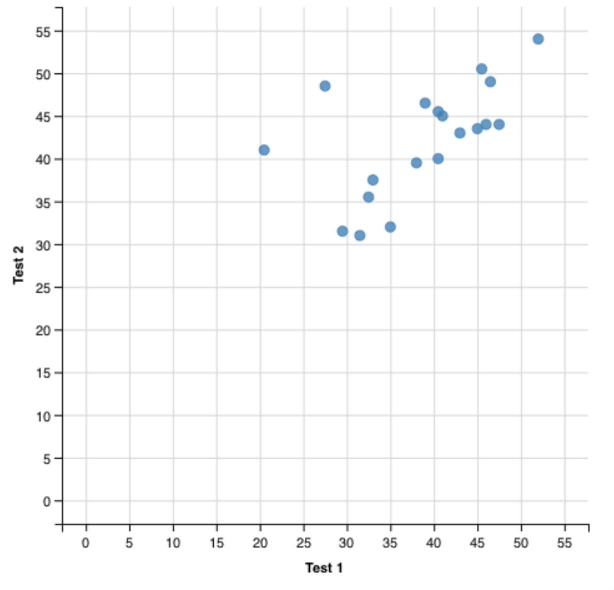
   Check the CORRELATION COEFFICIENT box to verify the correlation you recorded above. Then, check the SHOW SHUFFLE OPTIONS box, shuffle the data at least 10,000 times, and record the p-value.

   p = ______________

4. What can we conclude from all of this?

Scenario: The following table displays the unit 1 and unit 2 test scores for students in this course in 2012.

Student	Test 1	Test 2
1	20.5	41.0
2	27.5	48.5
3	29.5	31.5
4	31.5	31.0
5	32.5	35.5
6	33.0	37.5
7	35.0	32.0
8	38.0	39.5
9	39.0	46.5
10	40.5	40.0
11	40.5	45.5
12	41.0	45.0
13	43.0	43.0
14	45.0	43.5
15	45.5	50.5
16	46.0	44.0
17	46.5	49.0
18	47.5	44.0
19	52.0	54.0
Means	38.63	42.18
Std Dv	8.04	6.52



From this data, I calculated the following:

Correlation	95% Confidence Interval	p-value
Pearson's r = 0.5886	(0.258, 1.000)	0.00401
Spearman's rho = 0.6119		
Kendall's tau = 0.4765		

You can download this data at: http://www.bradthiessen.com/html5/data/testdata.csv

5. It looks like the scores from test 1 and test 2 have (roughly) a linear relationship. On the scatterplot displayed above, sketch the line you think best fits the data. Estimate the slope and y-intercept of that line and write the formula here:

y = mx + b –> test 2 = ________________ (test 1) + ________________
(slope) (y-intercept)

6. Every student who answers the previous question will (probably) have different values for the slope and y-intercept. How could we decide which line (from all possible lines students could sketch) is best? We'll learn one approach (the *least squares criterion*) in the next activity.

As we'll find out, the line that best fits this data can be written as $y = b_0 + b_1 x$. To calculate the slope and y-intercept of this best-fitting line by hand, we'll derive the following formulas:

$b_1 = r\frac{s_y}{s_x}$ and $b_0 = \overline{Y} - b_1\overline{X}$, where r is Pearson's r and s represents a standard deviation. If we let X = test 1 and Y = test 2, calculate this best-fitting line:

predicted test 2 = ________________ (test 1) + ________________
(slope) (y-intercept)

7. Using the formula for the best-fitting line you just calculated, predict the following:

Predicted score on test 2 for a student with test 1 = 45: ____________________

Predicted score on test 2 for a student with test 1 = 15: ____________________

In which prediction do you have more confidence? Explain why:

____

8. I calculated this best-fitting line using R. The output is pasted below (so you can check your answer to #6). Interpret this slope and y-intercept. What do they represent in this scenario (regarding test scores)?

```
lm(formula = Test2 ~ Test1)

Coefficients:
(Intercept)       Test1
    23.7461       0.4773
```

The slope (0.4773) represents: ____________________

The y-intercept (23.7461) represents: ____________________

9. In the next activity, we'll also learn about the *coefficient of determination*, R^2. Calculate this coefficient in this example by simply squaring the correlation coefficient. This coefficient can be interpreted in much the same way as we interpreted eta-squared when we conducted ANOVA. Go ahead and try to interpret this coefficient of determination in this scenario.

R^2 = ____________ Interpretation: ____________________

____

10. You may have heard the phrase *correlation does not imply causation*. Some examples of this appear on the following two websites:

    Correlation vs. Causation: http://jfmueller.faculty.noctrl.edu/100/correlation_or_causation.htm

    Spurious correlations: http://tylervigen.com

    Go to the first link (Correlation vs Causation) and choose one article that might interest you (e.g., Dogs walked by men are more aggressive).

    For that article, do the following:

    a) Very briefly summarize the correlation implied by the article
    b) Briefly explain why that correlation does not imply causation. Identify potential reasons why the two variables in the article would have a positive correlation (and, if you can, hypothesize what other variable might be causing that correlation).
    c) Briefly describe an experiment you could do to test if the correlation implied by the article is, in fact, causation.