Assignment 7  
November 14, 2016

Note: Remember that for questions involving the use of R, instructions on how to use all of the R commands mentioned in these questions are available on the course web site.

1. A researcher finds that there is a correlation of 1.0 between two variables, X and Y. The researcher then divides the subjects into two groups, based on their X scores. Subjects who score below the median are assigned to one group and subjects who score above the median are assigned to the other group (this is called a median split). Each subject in the first group is then assigned an X score of 1 and each subject in the second group is assigned an X score of 2. The researcher then computes a correlation coefficient between this revised set of X scores and the original Y scores. What is this new correlation going to be, relative to the original correlation of 1.0? Briefly explain your reasoning. (Hint: draw a scatterplot for the original situation and for the median split situation.)

2. A researcher obtains a random sample of 45 children at age 6 and gives them a test of extroversion. Four years later, these 45 children are again given the test and the researcher finds that there is a correlation of .68 between scores at age 6 and scores at age 10 for these 45 children. In addition, treating scores at age 6 as X and scores at age 10 as Y, the researcher found that \( M_X = 30, s_X = 6, M_Y = 34, \) and \( s_Y = 7. \) Now suppose that a new 6-year-old child is drawn at random from the population that provided the original sample of 45 children. This child has an extroversion score of 28. What is the probability that this child will have an extroversion score between 35 and 40 when he or she is 10 years old?

3. The correlation between intelligence and annual income in some population is found to be .376. What proportion of variability in annual income can be explained by differences in intelligence? At a minimum, how many subjects would have to be in the sample for this correlation to be significant with \( \alpha = .05 \) in a nondirectional test?

4. When working with linear regression, variability in Y (referred to as \( SS_Y \)) is composed of two parts:

\[
SS_Y = SS_{Y\hat{}} + SS_{\text{error}}
\]

Assume that as the size of the correlation between X and Y increases from 0 to 1.0, \( SS_Y \) remains constant. What changes occur in \( SS_{Y\hat{}} \) and \( SS_{\text{error}} \) as the correlation increases? That is, when \( r = 0 \), what values do these two quantities take on and what values to they approach as \( r \) approaches 1?

5. Working with the binomial distribution in the lecture notes associated with the test of Madam X's ability to predict the future, indicate which results (how many correct predictions in 12 coin flips) would lead to rejection of the null hypothesis if outcomes at either end of the distribution an be taken as evidence for rejection. Keep the criterion for rejecting the null hypothesis below .05.

6. A test of Madam X's ability can be done using a binomial distribution that indicates the probability of each possible number of correct predictions, or it can be done using the standard normal distribution to approximate the binomial distribution that is relevant to any particular test of her ability. For example, suppose that instead of making predictions
about 12 coin flips, Madam X were to make predictions for 100 coin flips. You can see that there are 101 possible outcomes, that is, anywhere from 0 to 100 correct predictions. Considering only outcomes that would indicate that Madam X is doing better than guessing, how many correct predictions should we require her to get for us to be willing to reject the null hypothesis that she is just guessing? We will use the usual criterion of .05 as our definition of unlikely (i.e., evidence for rejecting the null hypothesis). This problem can be solved by using the normal distribution to approximate the probability of obtaining various numbers of correct predictions under the null hypothesis. The mean of the distribution of possible numbers of correct predictions is computed as \( np \), where \( n \) is the number of trials (i.e., coin flips) and \( p \) is the probability of making a correct prediction under \( H_0 \) (i.e., .5 in the current example). So the mean is \( np = 100(.5) = 50 \). The standard deviation of this distribution of outcomes is \( \sqrt{n p q} \), where \( q \) is just \( 1 - p \). In the current problem, then, the standard deviation of the distribution of possible outcomes is \( \sqrt{100(.5)(.5)} = \sqrt{25} = 5 \). So now we have a normal distribution with mean = 50 and standard deviation = 5. Scores in the upper end of this distribution represent better than chance performance. We can estimate the minimum number of correct predictions needed to reject \( H_0 \) by determining what \( z \) score cuts off the upper 5% (or .05) of the distribution. Find that \( z \) score, then convert it into the corresponding number of correct predictions, to the nearest whole number, keeping in mind that the mean of the distribution is 50 and the standard deviation is 5.

7. In null hypothesis significance testing, the obtained \( p \) value tells us (select ONE option)
   (a) the probability that the alternative hypothesis is true
   (b) the probability that the null hypothesis is true
   (c) the probability of obtaining the observed data assuming the null hypothesis is true
   (d) the probability of obtaining the observed result or a more extreme result assuming \( H_0 \) is true
   (e) two of the above

8. Use the table of significant values for the correlation coefficient to answer the following question. Sketch a distribution of the distribution of possible values of the correlation coefficient when \( H_0 \) is true (i.e., \( r = 0 \)) and the sample consists of 32 subjects. Place a vertical mark at each end of the distribution corresponding to the values of \( r \) that cut off the most extreme 5% of the possible \( r \) values (2.5% at each end of the distribution). Label the value of \( r \) at each of these two points. Now place a vertical mark at the upper end of the distribution that cuts off the largest 5% of the possible \( r \) values. Label the value of \( r \) at this point.

9. Using R, read the data file posted for Assignment 5 (animacy.txt) into a variable. Use the R function cor.test to compute the correlation between the two variables in that file and to test the correlation for significance. Use the plot function to examine the scatter plot as well.

10. An additional data file called factor.txt contains a hypothetical set of data consisting of \( X \) and \( Y \) scores for each subject. Use R to plot the scatter plot for these data and then compute and test \( r \) for significance with the cor.test function. In null hypothesis
significance testing, what is the correct conclusion to reach given the results of the significant test?