*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

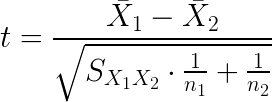
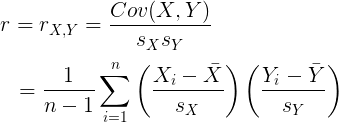
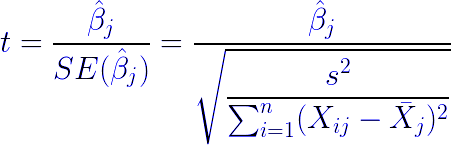
*Probability and Statistics for Data Science* name

*Stony Brook University  
CSE594 - Spring 2016*

Assignment 2: Discovery and Research Methods

Assigned: 3/9/2016; Due: 3/22/2016, 3:45pm EST

Part I. Problem Solving (40 points)

1. **Hypothetical Gambling. (15 points)** You believe a 6-sided die is weighted such that it will roll a 6 more often than any other number.
   1. What is the null hypothesis corresponding to your belief?  
        
      H0: That 6 is not more likely than other numbers.
   2. You test your hypothesis by rolling the die 100 times. You ended up with a 6, 23 times. If you choose an alpha of .05, can you reject the null? Show your work, including the rejection region for the null.  
        
      Yes. Let X ~ Binomial(100, ⅙) be the distribution of 6s rolled if the die is fair. Since we are only concerned if 6 is more common, then we are interested in a one-tail rejection region corresponding to when P(X < count) = 0.95. Using the ppf function (“ss.binom(100, 1/float(6)).ppf(.95)”), we determine count is 23.   
        
      (note: If the hypothesis was also concerned with 6s being rolled less often, assumed 6 is used more or less, then a two-tailed test would be used and we couldn’t reject the null)
   3. Dungeons and Dragons Incorporated, manufacturer of dice, believes it can create better performing dice out of aluminum rather than plastic. They perform a tests on 10 prototype aluminum dice compared to 10 plastic dice to compare performance (i.e. the number of rolls until a die becomes biased). They have a special machine that rolls and tests dice for bias. Given the following number of rolls until the dice become biased, can you conclude that it is 95% probable that aluminum performs better? (hint: may assume rolls until biased is well approximated as a Normal).  
      *aluminum rolls until biased* = [136, 73, 118, 122, 114, 103, 149, 118, 113, 105]  
      *plastic rolls until biased* = [129, 89, 97, 94, 124, 77, 85, 86, 86, 69]  
        
      Here we are looking at whether the mean for aluminum is significantly higher than the mean for plastic. Considering the sample standard deviations for the two variables are very similar but they are in fact from two independent dice, we assume they are independent with the same variance and use the corresponding t-test:   
      The sample statistics are as follows:  
       means: 115.100, 93.600  
       stds: 19.175, 18.112  
       ns: 10.000, 10.000  
       dfs: 9.000, 9.000  
      The pooled variance, SX1X2 = 347.865 = (df1\*sd1\*\*2 + df2\*sd2\*\*2) / (df1 + df2)  
      Plugging these into the two sample t-test, we get t = 2.5776. Considering this is again a 1-tailed question (since only asking if aluminum is better), we look up when the ppf of T(df1+df2) is .95: t > 1.73. Thus, we can reject the null (the p-value, given from the 1 - cdf(t) is .0095). (Note: If this was two-tailed, we would have looked up when the ppf > 97.5% or < .025; or we would multiply the p-value by 2).
2. **Valuable Hoops. (25 points)** In the NBA, players salary is thought to reflect their performance on the team. One way to assess performance is a player’s “plus-minus” the difference between the points scored by his team and the points scores by the other team while he is on the floor. Six players salary (in millions) and plus-minus are given: (2.5, -1), (10, 4), (8.5, 3), (4, 4), (1.5, -3), (14, 6). For all questions below, show all work and do not use a calculator except for addition and multiplication.
   1. Using least squares regression (the direct method) calculate 𝛽0 and 𝛽1 where the dependent variable is salary and the plus-minus is the predictor.  
        
      Let X = constant (for intercept) and the salary = [[ 1 -1] [ 1 4] [ 1 3] [ 1 4] [ 1 -3] [ 1 6]]. Y = [ 2.5 10. 8.5 4. 1.5 14. ]  
      (XTX)-1XTY = [ 4.14447592 1.20254958]  
        
       𝛽0 = 4.14 ; 𝛽1 =1.20
   2. What is the Pearson Product-Moment Correlation Coefficient between the two variables.  
      Using the same variables as above, except dropping the constant from X, we need to calculate:   
      Plugging in n and standard deviations: *r =*https://latex.codecogs.com/gif.latex?%5Clarge%20%5Cfrac%7B1%7D%7B6-1%7D%5Csum_%7Bi%20%3D%201%7D%5En%5Cleft%20%28%20%5Cfrac%7BX_i%20-%20%5Cbar%7BX%7D%7D%7B3.43%7D%20%5Cright%20%29%5Cleft%20%28%20%5Cfrac%7BY_i%20-%20%5Cbar%7BY%7D%7D%7B4.89%7D%5Cright%20%29Calculating the standard score for each x and y we get:  
      (⅕)\*sum( [-0.923, 0.534, 0.243, 0.534, -1.506, 1.118]\*[-0.87, 0.665, 0.358, -0.563, -1.074, 1.484])  
       *r* = .844
   3. The relation is hypothesized to be positive, what is the corresponding p-value?  
      We use t-test on the linear regression coefficient, 𝛽1 =1.20  
       = 3.15, looking up the cdf for t(df = 4) we get...  
      (Note one could also use: )  
       *p* = .017
   4. Years of experience in the NBA also plays a part, since experienced players are more reliable and provide wisdom off the court to the younger players. The six players have the following years of experience (same order as before): 2, 12, 5, 6, 9, 7. What is the Pearson Prod-Mom Correl Coef between years and salary?  
        
      Same equation as b. Plugging in our standard scores for X and Y we get:   
      (⅕)\*sum([-1.409, 1.506, -0.534, -0.243, 0.632, 0.049]\*[-0.87, 0.665, 0.358, -0.563, -1.074, 1.484])  
        
      *r =.313*
   5. Using *standardized* multiple linear regression: What is the unique effect (coefficient) of plus-minus (a predictor) on salary (the dependent variable), holding years of experience constant (another predictor)? What is the unique effect of years on salary, holding plus-minus constant?  
      (you may use a computer to solve matrix linear algebra operations, but report the resulting matrix or vector)  
      First, we setup X and y, and standardize:

|  |  |
| --- | --- |
| Original X = [[-1 2]  [ 4 12]  [ 3 5]  [ 4 6]  [-3 9]  [ 6 7]] | Original y = [[ 2.5]  [ 10. ]  [ 8.5]  [ 4. ]  [ 1.5]  [ 14. ]] |
| Standardized X =  [[-0.92315712 -1.40902929]  [ 0.53445939 1.50620373]  [ 0.24293609 -0.53445939]  [ 0.53445939 -0.24293609]  [-1.50620373 0.63163382]  [ 1.11750599 0.04858722]] | Standardized y =  [[-0.86979567]  [ 0.66513786]  [ 0.35815116]  [-0.56280896]  [-1.07445347]  [ 1.48376908]] |

Thus, (XTX)-1XTY = [ 0.81410113 0.15871045]

𝛽plusminus = .814 ; 𝛽years = .158

* 1. Being left-handed is also thought to be related to performance. The third and sixth players are left handed: (i.e. left\_handed = [0, 0, 1, 0, 0, 1]). Using *standardized* **logistic** regression: What is the logistic correlation coefficient on plusminus (the predictor) for being left-handed (the dependent variable)?   
     (you may use a computer to solve matrix linear algebra operations, but report the resulting matrix or vector)  
     Hint: your answer should be nearly converged at .01 precision after 3 iterations.  
     Here we have X = [standardized plus-minus, inercept] and y = left\_handed?

|  |  |
| --- | --- |
| X = [[-0.92315712 1. ]  [ 0.53445939 1. ]  [ 0.24293609 1. ]  [ 0.53445939 1. ]  [-1.50620373 1. ]  [ 1.11750599 1. ]] | Y = [[0]  [0]  [1]  [0]  [0]  [1]] |

We walk through the reweighted least squares algorithm:   
iteration 0 p: [ 0.5 0.5 0.5 0.5 0.5 0.5], w-diag: [ 0.25 0.25 0.25 0.25 0.25 0.25], z: [[-2. -2. 2. -2. -2. 2.]], betas: [[ 1.088 -0.667]]  
  
iteration 1 p: [ 0.16 0.48 0.4 0.48 0.09 0.63], w-diag: [ 0.13 0.25 0.24 0.25 0.08 0.23], z: [[-2.86 -2. 2.09 -2. -3.41 2.13]], betas: [[ 1.693 -1.042]]  
  
iteration 2 p: [ 0.07 0.47 0.35 0.47 0.03 0.7 ], w-diag: [ 0.06 0.25 0.23 0.25 0.03 0.21], z: [[-3.68 -2.01 2.25 -2.01 -4.62 2.28]], betas: [[ 2.098 -1.296]]  
//THIS IS ALL YOU NEEDED TO SHOW but for completion, below converges...

iteration 3 p: [ 0.04 0.46 0.31 0.46 0.01 0.74], w-diag: [ 0.04 0.25 0.22 0.25 0.01 0.19], z: [[-4.27 -2.01 2.41 -2.01 -5.47 2.4 ]], betas: [[ 2.239 -1.384]]  
  
iteration 4 p: [ 0.03 0.45 0.3 0.45 0.01 0.75], w-diag: [ 0.03 0.25 0.21 0.25 0.01 0.19], z: [[-4.48 -2.02 2.48 -2.02 -5.77 2.45]], betas: [[ 2.251 -1.391]]  
  
iteration 5 p: [ 0.03 0.45 0.3 0.45 0.01 0.75], w-diag: [ 0.03 0.25 0.21 0.25 0.01 0.19], z: [[-4.5 -2.02 2.48 -2.02 -5.79 2.45]], betas: [[ 2.251 -1.392]]  
  
 𝛽plusminus = 2.251 (+- 0.2 is ok)