_____Jieao Zhu_____

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.
   a. What is the null hypothesis corresponding to your belief?

      The die is not biased.

   b. 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.

If the die is not biased, if it rolls a 6, we mark 1, otherwise we mark 0, the distribution will be a binomial distribution, where $p = \frac{1}{6}$ **.**
The sampling distribution is a normal distribution where $\mu = \frac{100}{6},\ \sigma = \sqrt{\frac{500}{36}}$ **.**
Suppose that if we have more than m **or less than** n **6s, we will reject null hypothesis:** $1 - P(\frac{m-\mu}{\sigma}) + P(\frac{n-\mu}{\sigma}) = 0.05$
We will have : $n = 9.37,\ m = 23.97$
We will not reject null hypothesis because $23 < 23.97$ **.**

c. 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]

mean(aluminum)-mean(plastc) = 21.50
95% confidence interval of this difference: From 3.03 to 39.97
t = 2.4453
The two-tailed P value equals 0.0250,
this difference significant.
That is, aluminum performs better than plastic

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.
   a. Using least squares regression (the direct method) calculate β_0 and β_1 where the dependent variable is salary and the plus-minus is the predictor.

Assume that: $Y = \beta_0 + \beta_1 * X$, where Y is Salary, X is Plusminus.

$$Y = \begin{bmatrix} 2.5 \\ 10 \\ 8.5 \\ 4 \\ 1.5 \\ 14 \end{bmatrix}, \quad X = \begin{bmatrix} -1 \\ 4 \\ 3 \\ 4 \\ -3 \\ 6 \end{bmatrix}, \quad \hat{\beta} = (X^T X)^{-1} X^T Y = \begin{bmatrix} 4.14 \\ 1.2 \end{bmatrix}$$

$\beta_0 = 4.14$; $\beta_1 = 1.2$

b. What is the Pearson Product-Moment Correlation Coefficient between the two variables.

- $\sum X_i = 13$, $\bar{X} = 2.16$, $SS_X = 58.83$
- $\sum Y_i = 40.5$, $\bar{Y} = 6.75$, $SS_Y = 119.38$
- $r = \dfrac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{SS_X \cdot SS_Y}} = 0.84$

$r =$ 0.84

c. The relation is hypothesized to be positive, what is the corresponding p-value?

$t = \dfrac{\hat{\beta}}{\sqrt{\dfrac{S^2}{\sum (x-\bar{x})^2}}}$, where: $S^2 = \dfrac{\sum (Y-\hat{Y})^2}{n-2} = 8.5$, since: DF = 5

so: $t = \dfrac{1.2}{\sqrt{\frac{8.5}{59}}} = 3.16$, so: $P = 0.0251$

$p =$ 0.0251

d. 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?

$\bar{X} = 6.8$, $\bar{Y} = 6.75$

$Cov(X, Y) = \frac{1}{n-1} \sum (X-\bar{X})(Y-\bar{Y}) = 4.37$

$\sigma_X = 3.13$, $\sigma_Y = 4.46$

$r = \dfrac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y} = 0.31$

$r =$ 0.31

e. Using *standardized* multiple linear regression: What is the unique effect of plus-minus on salary, holding years of experience constant? What is the unique effect of years on salary, holding plus-minus constant?

$$\text{Year} = \begin{bmatrix} 2 \\ 12 \\ 5 \\ 6 \\ 9 \\ 7 \end{bmatrix} \xrightarrow{\text{stand}} \begin{bmatrix} -1.54 \\ 1.65 \\ -0.158 \\ -0.26 \\ 0.69 \\ 0.05 \end{bmatrix}, \quad \text{Plusminus} = \begin{bmatrix} -1 \\ 4 \\ 3 \\ 4 \\ -3 \\ 6 \end{bmatrix} \rightarrow \begin{bmatrix} -1 \\ 0.6 \\ 0.26 \\ 0.58 \\ -1.65 \\ 1.23 \end{bmatrix}, \quad \text{Salary} = \begin{bmatrix} 2.5 \\ 10 \\ 8.5 \\ 4 \\ 1.5 \\ 14 \end{bmatrix} \rightarrow \begin{bmatrix} -0.95 \\ 0.73 \\ 0.39 \\ -0.62 \\ -1.18 \\ 1.63 \end{bmatrix}$$

then do the same calculation as in 2.a

$\beta_{\text{plusminus}} = 0.82$; $\beta_{\text{years}} = 0.16$

f. 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)?

Assume that : $\beta_0 + \beta_1 \cdot \text{Plusminus} = Y$, $Y \in (0, 1)$

step 1: $\beta = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $p = \begin{bmatrix} 0.5 \\ \vdots \\ 0.5 \end{bmatrix}$, $W = \begin{bmatrix} 0.25 & & \\ & \ddots & \\ & & 0.25 \end{bmatrix}$, $z = \begin{bmatrix} -2 \\ \vdots \\ -2 \end{bmatrix}$, $\beta = 1.534$

step 2: $\beta = \begin{bmatrix} -0.67 \\ 0.98 \end{bmatrix}$, $W = \begin{bmatrix} 0.134 & & \\ & 0.249 & \\ & & \ddots \\ & & 0.232 \end{bmatrix}$, $z = \begin{bmatrix} -2.8 \\ -2 \\ 2 \\ -2 \\ -3.4 \\ 2.12 \end{bmatrix}$, $\beta = 1.53$

step 3: $\beta = \begin{bmatrix} -1.28 \\ 1.89 \end{bmatrix} \cdots\rightarrow$ step 4: $\beta = \begin{bmatrix} -1.367 \\ 2.01 \end{bmatrix}$

$\beta_{\text{plusminus}} = 2.01$

Part II. Programming (60 points) - Differential Topic Analysis

Task Overview: To find linguistic topics correlated with age and occupation, controlled for gender, using various statistical tests and plotting the results.

Necessary Files:

- the Blog Authorship Corpus: http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm
  (bandwidth limited; may take a while to download)
- topic model posteriors (probabilistic model of words likely to appear in similar messages).
  - p(topic| word): http://wwbp.org/downloads/public_data/wwbpFBtopics_condProb.csv
    ("category" is the topic id, "term" is the word, and "weight" is p(topic|word))
  - most prevalent words per topic:
    http://wwbp.org/downloads/public_data/2000topics.top20freqs.keys.csv
- happierfuntokenizing.py: http://wwbp.org/downloads/public_data/happierfuntokenizing.zip
  (a tokenizer splits a sentence into words; this one works well with social media and blogs)

Your code should be self-contained in a file called "a2_studentid.py" which can be run in python 2.7 by typing: `python a2_studentid.py blogs_directory wwbpFBtopics_condProb.csv 2000topics.top20freqs.keys.csv`
Assume happierfuntokenizing.py is in the same directory as your python code.

Python libraries permitted: default file IO libraries, csvreader, pandas (no statistical methods), numpy (only for arrays and algebra, no statistical tests or regression), scipy.stats (only for pdf, pmf, cdf lookup), and happierfuntokenizing. For question 5: any libraries are permitted for producing the plots.

1. **Read and tokenize the corpus.**
   Unzip the corpus file and look at the contents before reading on. Each file in the corpus is named according to information about the blogger: `user_id.gender.age.industry.star_sign.xml`
   Read each file using an xml parser or simply a regular expression searching for "<POST>" (other fields are not necessary). As you do so, record the user-id, gender, age, industry, and then tokenize each blog post of each user in the corpus using the "*tokenize*" method in happierfuntokenizing.py. *tokenize* takes in a string (i.e. a single blog post) and returns a list of strings (i.e. the words in the blog post). Record the total number of times each user mentions each word -- recommend storing this in a dictionary:
   { user1: {word1: count1, word2: count2, ...},
   user2: {word1: count1, word2: count2, ...},

…}

or a pandas DataFrame per user (storing as one large matrix may max out your ram).

output: The quoted portion below indicates text that should prepend your output for each item:

a. the total number of blog posts across all users: "1. a) posts: XXX"
b. the total number of users: "1. b) users: XXX"
c. the total number of words across all users: "1. c) words: XXX"
d. the number of users per industry: "1. d) Communication: XXX,
Consulting: XXX,
…"

2. **Calculate users' probability of mentioning each topic ("topic usage").**

Hint: *Do this while processing individual users information for step 1 so that you don't have to store all words across all users.*

The probability of a topic for each user is given by:

$$p(topic|user) = \sum_{word \in topic} p(topic|word)p(word|user)$$

where p(topic|word) comes from wwbpFBtopics_condProb.csv and

$$p(word|user) = \frac{count(user, word)}{\sum_{word \in allwords} count(user, word)}$$

a. print out the probabilities that the 3 users with the lowest user_id (i.e. users_id 73 is less than user_id 105) mention topics numbered 463, 963, and 981: "2. a) <user_id>: 463: .XXXX, 963: .XXXX, 981: .XXXX"
hint: these probabilities should be quite small

3. **Correlate each topic usage with user age, adjusting for gender.**

Use standardized multiple linear regression to find the unique effect of each topic on age, holding gender constant (i.e. as a covariate; you will run 2000 regressions with 2 predictors, rather 1 regression with >2000 variables). Test for significance using the t-test for multiple regression and then adjust for the fact that you are testing 2000 hypotheses (one for each topic) by using the Bonferroni correction (you must implement these methods -- no methods may be called beyond those for algebraic calculations or looking up pdf / cdf values).

a. print the top 10 most positively correlated topics and values: "3. a) topic_id: XXX, correlation: .XXX, p-value: .XXX, signficant after correction? (y/n)"

b. print the top 10 most negatively correlated topics and values: "3. b) topic_id: XXX, correlation: -.XXX, p-value: .XXX, signficant after correction? (y/n)"

4. **Correlate each topic usage with user industry, adjusting for gender and age.**
   Use standardized multiple logistic regression to find the unique effect of each topic on industry, adjusting for gender and age. Test for significance using the permutation test then adjust for the fact that you are testing 2000 hypotheses (one for each topic) by using the Benjamini-Hochberg false-discovery rate correction (you must implement these methods -- no methods may be called beyond those for algebraic calculations).
   a. print the top 5 most positively correlated topics and values per industry:
      "4. a) industry: XXX, topic_id: XXX, coefficient: .XXX, p-value: .XXX, signficant after correction? (y/n)"
   b. print the top 5 most positively correlated topics and values per industry:
      "4. b) industry: XXX, topic_id: XXX, coefficient: -.XXX, p-value: .XXX, signficant after correction? (y/n)"

5. **Plot topics by industry x age.**
   Limit this analysis to topics making it in the top 5s for questions 4a and 4b. Let the x axis represent correlation with industry, adjusted for age and gender (from question 4). Let the y axis represent the mean age of the top 25% of users according to how often they use the topic. For each topic, plot the top 4 words given by 2000topics.top20freqs.keys.csv at the coordinate dictated correlation with industry and mean age of top 25%. Produce one plot per industry all saved to one png (use a multiple panels in .
   a. print the name of the file with "5. a) 5a_industry_plots.png"
      (an example plot will be provided here)

Testing: Our test will use a random subset of the Blog Authorship Corpora (i.e. many files in the blogs_directory will be left out). **It is recommended that you develop your code over a random subset of ~500 users in order to be able to test and make progress quickly. Then, scale it up to the whole dataset at the end. We will use two test sets: a random subset of 2,000 users as our primary test set (worth 50 points), a secondary test set consisting of a random 15,000 users will test that your approach works at scale (worth 10 points).** You can count on the test machine having at least 8GB of ram, a 1.5+ghz processor, and allowing at least 30 minutes for your code to complete.

Questions / Clarifications: Please post questions on Piazza, so other classmates may see the answers. Questions posted within 72 hours of the deadline, are not guaranteed a response before the deadline.

Academic Integrity: As with all assignments (sans the team project), although you may discuss concepts with others, you must work independently and insure your work and code is not visible to any classmates.