Applied Regression and Time Series Analysis
Lab 1: Probability and Statistics Basics
Spring 2016

Jeffrey Yau and Paul Laskowski

January 15, 2016

Instructions

- **Due Date:** Beginning of Week 3 Live Session.
- **Submission:** **Submit 2 files. Missing one of the two files will result in a 50% reduction in grade.**
    1. A report (in pdf format) detailing your answers and all the steps to arrive at your answers
    2. A well-documented R-script, jupyter notebook, or Rmd file detailing all of the codes used to arrive at your answers.
- Late submission will not receive any credit.
- Answers need to be typed. Latex, which you will likely have to use in R markdown, is great for typesetting documents with mathematical symbols.
- All the steps used to arrive at your final answers need to be shown clearly. These steps are as important as the final answer.
- The final answer of each question needs to be very easy identified; the use of bold fonts, highlights, or circling will help.
- This is a group project. Form a group with 3 or 4 people.
- Although this is a group project, we encourage you to attempt all of the exercises before discussing with your teammates. Do not use the "division-of-labor" approach. Each of the students in a group is expected to make sufficient contribution to the lab. If any of your teammate does not make sufficient contribution, please contact your instructor.
- DO NOT copy and paste or even leverage on the solutions we gave to the students in previous semesters. Violation will be reported to the Director of the MIDS program and the Office that oversees UC Berkeley Academic Integrity. In any case, the lab has various subtle changes that make those answers not directly applicable.

Part I: Marginal, Joint, and Conditional Probabilities

Question 1 (6 points)

In a team of data scientists, 36 are expert in machine learning, 28 are expert in statistics, and 18 are awesome. 22 are expert in both machine learning and statistics, 12 are expert in machine learning and are awesome, 9 are expert in statistics and are awesome , and 48 are expert in machine learning or statistics or are awesome. Suppose you are in a cocktail party with this group of data scientists and you have an equal probability of meeting any one of them.

1. What is the probability of meeting a data scientist who is an expert in both machine learning and statistics and is awesome?
2. Suppose you meet a data scientist who is an expert in machine learning. Given this information, what is the probability that s/he is not awesome?
3. Suppose the you meet a data scientist who is awesome. Given this information, what is the probability that s/he is an expert in either machine learning or statistics?

Question 2 (4 Points)

Suppose for events A and B, $Pr(A) = p \leq \frac{1}{2}$, $Pr(B) = q$, where $\frac{1}{4} < q < \frac{1}{2}$. These are the only information we have about the events.

1. What are the maximum and minimum possible values for $Pr(A \cup B)$?
2. What are the maximum and minimum possible values for $Pr(A|B)$?

Part II: Random Variables, Expectation, Conditional Expectation

Question 3 (6 Points)

Suppose the life span of a particular server is a continuous random variable, t, with a uniform probability distribution between 0 and k year, where $k \leq 10$ is a positive integer.

The server comes with a contract that guarantees a full or partial refund, depending on how long it lasts. Specifically, if the server fails in the first year, it gives a full refund denoted by θ. If it lasts more than 1 year but fails before $\frac{k}{2}$ years, the manufacturer will pay $x = \$A(k-t)^{1/2}$, where A is some positive constant equal to 2 if $t \leq \frac{k}{2}$. If it lasts between $\frac{k}{2}$ and $\frac{3k}{4}$ years, it pays $\frac{\theta}{10}$.

1. Given that the server lasts for $\frac{k}{4}$ years without failing, what is the probability that it will last another year?
2. Compute the expected payout from the contract, $E(x)$.
3. Compute the variance of the payout from the contract.

Question 4 (9 Points)

Continuous random variables X and Y have a joint distribution with probability density function $f(x,y) = 2e^{-x}e^{-2y}$ for $0 < x < \infty$, $0 < y < \infty$ and 0 otherwise.

1. Compute $P(X > a, Y < b)$, where a, b are positive constants and $a < b$.
2. Compute $P(X < Y)$.
3. Compute $P(X < a)$.

Question 5 (6 Points)

Let X be a random variable and x be a real number. A linear function of the squared deviation from x is another random variable, $Y = a + b(X-x)^2$, where a and b are some positive constant.

1. Find the value of x that minimizes $E(Y)$. Show that your result is really the minimum.
2. Find the value of $E(Y)$ for the choice of x you found in (1)?
3. Suppose $Y = ax + b(X-x)^2$. Find the values of x that minimizes $E(Y)$. Show that your result is really the minimum.

Question 6 (4 Points)

Suppose X and Y are independent continuous random variables, where both of which are uniformly distributed between 0 and 1. Let random variable $Z = X + Y$.

1. Choose a value of z between 0 and 2, and draw a graph depicting the region of the $X - Y$ plane for which Z is less than z.

2. Derive the probability density function, $f(z)$.

Question 7 (10 Points)

In a casino, you pay the following game. A pair fair, ordinary 6-faced dices is rolled. If the sum of the dice is $2, 3$, or 12, the house wins. If it is 7 or 11, you win. If it is any other number x, the house rolls the dice again until the sum is either 7 or x. If it is 7, the house wins. If it is x, you win. A game ends if one of the two players wins. Let Y be the number of rolls needed until the game ends.

1. Is the expected number of rolls given that you win *more than, equal to, or less than* the expected number of rolls given that house wins (in a game)? The steps to arrive at your answer numerically need to be clearly shown.

2. Suppose it takes \$20 to pay, and the payoff is $\$100, \$80, \$60, \$40, \$0$ if you win in the $1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}$ round, respectively. That is, if you win in the 1^{st} round, you are paid \$100 (so your net profit is \$80, if you win in the 2^{nd} round, you are paid \$80, etc. Derive the expected payoff function of a game.

Part III: Statistical Estimation and Statistical Inference

In classical statistics, parameters are unknown constants whereas estimators are functions of samples and are random variables. The questions in this section are designed to clarify the relationship between parameters and estimators, and explore the properties that different estimators may have.

Question 8 (10 Points)

Let $Y_1, \ldots, Y_n$ be n random variables, such that any two of them are uncorrelated, and all share the same mean μ and variance σ^2. Let $\bar{Y}$ be the average Y_i, which is also a random variable.

Define the class of linear estimators of μ by

$$W = \sum_{i=1}^{n} a_i Y_i$$

where the a_i are constants.

1. What restriction on the a_i is needed for W to be an unbiased estimator of μ?
2. Find $Var(W)$.
3. Given a set of numbers $a_1, a_2, ...a_n$, the following inequality holds:

$$\frac{1}{n}\left(\sum_{i=1}^{n} a_i\right)^2 \leq \sum_{i=1}^{n} a_i^2$$

   Use this inequality, along with the previous parts of this question, to show that $Var(W) \geq Var(\bar{Y})$ whenever W is unbiased. We say that $\bar{Y}$ is the best linear unbiased estimator (BLUE).

Question 9 (10 Points)

Let $\bar{Y}$ denote the average of n independent draws from a population distribution with mean μ and variance σ^2. Consider two alternative estimators of μ: $W_1 = \left(\frac{n-1}{n}\right)\bar{Y}$ and $W_2 = k\bar{Y}$, where $0 < k < 1$.

1. Compute the biases of both W_1 and W_2. Which estimator is consistent?
2. Compute $Var(W_1)$ and $Var(W_2)$. Which estimator has lower variance?

Question 10 (10 Points)

Given a random sample $Y_1, Y_2, \ldots, Y_n$ from some distribution $F(.)$ with mean μ and variance σ^2, where both μ and σ^2 are unknown parameters.

Let $\bar{Y}$ be the average of the sample. Consider the following estimator for σ^2:

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(Y - \bar{Y}\right)^2$$

1. Show that $E(\bar{Y}) = E(Y_i)\ \forall i \in 1, 2, \ldots, n$
2. Show that $Var(\bar{Y}) = \frac{1}{n}Var(Y_i)\ \forall i \in 1, 2, \ldots, n$
3. Compute the expectation of $\widehat{\sigma^2}$ in terms of n and σ^2. In your derivation, make sure make use of the *i.i.d.* property and identify where you use it.
4. Is this an unbiased estimator for σ^2?
5. If not, what function of $\widehat{\sigma^2}$ produce an unbiased estimator?

Question 11 (5 Points)

Wooldridge's textbook: Appendix C, Question $4i, ii, iii$.

Question 12 (10 Points)

Wooldridge's textbook: Appendix C, Question 6.

Question 13 (10 Points)

Wooldridge's textbook: Appendix C, Question 8. In addition, answer the following questions:

1. Define Type I error.
2. What is the probability of Type I error of this test?
3. Define Type II error.
4. What is the probability of Type II error when using this decision rule, assuming the "true" population proportion is $theta^* = 0.45$.
5. Define the power of the test (in general terms).
6. Calculate the power of this test, again assuming the "true" population proportion is $\theta^* = 0.45$.