Stat 111:
Introduction to Statistical Inference, 2017-18
Susan A. Murphy & Neil Shephard
Assignment 3

Credit will be given to the best grades from 3 individually marked questions (each question gets equal credit). Answer as many questions are you would like.

1. On Canvas you can find a subset of the data from a randomized trial concerning the use of acupuncture for people who suffer from chronic headaches. The file is labelled `RCT.csv`. People randomized to acupuncture were eligible to receive up to 12 acupuncture sessions over 3 months. People randomized to control were eligible for their usual, standard care from their practitioners. The primary outcome, Y_i is self-report headache score at 1 year for the ith person. Y is labeled `pf5` in the data set. This subset contains data from 301 people.
   (a). Create a list of the people for whom we have observations of $Y_i(1)$.
   (b). What would be a good proxy for p the probability with which a person is randomized to acupuncture? Calculate this proxy.
   (c). Assume the randomization probability is $p = 1/2$; estimate the average causal treatment effect for the people in the study using this data.
   Some further information for you! It turns out that this study is much more complicated. Full details can be found at
   `http://www.bmj.com/content/328/7442/744`
   The full data is given in the article:
   `https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-7-15#MOESM1`.

2. This quesiton has two parts.

   (a) Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from the geometric distribution with probability of success, p. Derive the joint probability mass function (PMF); justify your answer.

   (b) Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from an exponential distribution with parameter λ, so $\mathrm{E}(Y_1) = \lambda^{-1}$. Derive the joint PDF; justify your answer.

3. Recall the linear regression model

$$\mathrm{E}\left(Y_i | X_i = x_i\right) = \theta x_i, \quad Y_i | X_i = x_i \overset{ind}{\sim}, \quad i = 1, 2, ..., n,$$

   along with the estimator

$$\widehat{\theta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

   Assume homoskedasticity: $\mathrm{Var}\left(Y_i | X_i = x_i\right) = \sigma^2$. Two more estimators of θ are

$$\widehat{\theta}_1 = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}$$

   and

$$\widehat{\theta}_2 = \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{x_i}.$$

   a. Show that $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are unbiased.
   b. Calculate the mse of these two estimators.
   c. It turns out that $\widehat{\theta}$ has the lowest mse of the three estimators. Prove this.

4. (This problem is modeled after 7.20, 7.21 in Casella and Berger (1990)). Consider the Auto MPG Data Set, `https://archive.ics.uci.edu/ml/datasets/Auto+MPG` In this problem Y_i is the MPG of the i-th auto and X_i is the weight of this auto. Define θ by $\mathrm{E}(Y_i|X_i = x_i) = \theta x_i$.
a. Estimate θ using the three estimators from the previous exercise (e.g. $\widehat{\theta}, \widehat{\theta}_1, \widehat{\theta}_2$).
b. Using the bootstrap approximate the variance of each of these 3 estimators. Use $R = 1,000$ bootstrap replications. Are your results consistent with the findings of the previous problem? Hint: when you do your bootstrap sampling you need to draw the pair (Y_i, X_i) together (i.e. bootstrap the rows of the data) not separately.

5. Solve Exercise 61, Chapter 4, of Blitzstein and Hwang (2015) which is about unbiased estimators. This is reproduced on the next page of this pdf.

6th February 2018
To be handed in by 1.05pm, Tuesday, 20th February (week 5).

(c) Suppose that m, n, N are such that EY is an integer. If the sampling is done with a fixed sample size equal to EY rather than sampling until exactly m tagged elk are obtained, find the expected number of tagged elk in the sample. Is it less than m, equal to m, or greater than m (for $n < N$)?

LOTUS

56. Ⓢ For $X \sim \text{Pois}(\lambda)$, find $E(X!)$ (the average factorial of X), if it is finite.
57. For $X \sim \text{Pois}(\lambda)$, find $E(2^X)$, if it is finite.
58. For $X \sim \text{Geom}(p)$, find $E(2^X)$ (if it is finite) and $E(2^{-X})$ (if it is finite). For each, make sure to clearly state what the values of p are for which it is finite.
59. Ⓢ Let $X \sim \text{Geom}(p)$ and let t be a constant. Find $E(e^{tX})$, as a function of t (this is known as the *moment generating function*; we will see in Chapter 6 how this function is useful).
60. Ⓢ The number of fish in a certain lake is a $\text{Pois}(\lambda)$ random variable. Worried that there might be no fish at all, a statistician adds one fish to the lake. Let Y be the resulting number of fish (so Y is 1 plus a $\text{Pois}(\lambda)$ random variable).

    (a) Find $E(Y^2)$.

    (b) Find $E(1/Y)$.
61. Ⓢ Let X be a $\text{Pois}(\lambda)$ random variable, where λ is fixed but unknown. Let $\theta = e^{-3\lambda}$, and suppose that we are interested in estimating θ based on the data. Since X is what we observe, our estimator is a function of X, call it $g(X)$. The *bias* of the estimator $g(X)$ is defined to be $E(g(X)) - \theta$, i.e., how far off the estimate is on average; the estimator is *unbiased* if its bias is 0.

    (a) For estimating λ, the r.v. X itself is an unbiased estimator. Compute the bias of the estimator $T = e^{-3X}$. Is it unbiased for estimating θ?

    (b) Show that $g(X) = (-2)^X$ is an unbiased estimator for θ. (In fact, it turns out to be the only unbiased estimator for θ.)

    (c) Explain intuitively why $g(X)$ is a silly choice for estimating θ, despite (b), and show how to improve it by finding an estimator $h(X)$ for θ that is always at least as good as $g(X)$ and sometimes strictly better than $g(X)$. That is,

    $$|h(X) - \theta| \leq |g(X) - \theta|,$$

    with the inequality sometimes strict.

Poisson approximation

62. Ⓢ Law school courses often have assigned seating to facilitate the Socratic method. Suppose that there are 100 first-year law students, and each takes the same two courses: Torts and Contracts. Both are held in the same lecture hall (which has 100 seats), and the seating is uniformly random and independent for the two courses.

    (a) Find the probability that no one has the same seat for both courses (exactly; you should leave your answer as a sum).

    (b) Find a simple but accurate approximation to the probability that no one has the same seat for both courses.

    (c) Find a simple but accurate approximation to the probability that at least two students have the same seat for both courses.