COLUMBIA UNIVERSITY — Prof. Marco Avella
Statistical Inference and Modeling — TA: Ian Kinsella
Graded Homework 1 — Fall 2019

You are free to present your solutions to the exercises below in *groups of at most three*. This homework is due on September 26th at the begining of that day's class. Good luck!

Exercise 1

Assume the life times of the lightbulbs in a building can be modeled by an exponential distribution (i.e. let D be a random variable representing the lifetime of a lamp, then the probability density function of D is $f(d;\lambda) = \lambda e^{-\lambda d}$ for $d > 0$). We would like to construct point and interval estimates of the quantiles of the survival time.

1. Compute the p^{th} population quantile (denoted $Q_D(p)$) of this model.
2. Find the MLE of $Q_D(p)$ given iid samples $D_1, D_2, \cdots, D_n$.
3. Give an approximate confidence interval of $Q_D(p)$ based on its MLE.
4. Show that $\lambda\overline{D_n}$ is an exact pivot and use it to contruct an exact confindence interval of the median (i.e. $Q_D(.5)$).

Exercise 2

Given the sample $X_1, \ldots, X_n$ issued from a Poisson distribution with parameter λ, we would like to estimate λ.

1. Show that the Method of Moments estimator based on the second moment equation $E(X^2) = \mu_2$ is
$$\hat{\lambda}_M = \frac{-1 + \sqrt{1 + \frac{4}{n}\sum_{i=1}^n X_i^2}}{2}$$
2. Find the asymptotic distribution of $\hat{\lambda}_M$.
Hints: Use the delta method *or* take $\sqrt{n}(\hat{\lambda}_M - \lambda)$, and multiply and divide it by $\{(1+2\lambda) + \sqrt{1+4Z}\}$, where $Z = 1/n\sum_{i=1}^n X_i^2$. Note that $\text{Var}(X_i^2) = 4\lambda^3 + 6\lambda^2 + \lambda$.
3. Compare the asymptotic efficiency of $\hat{\lambda}_M$ with the methods of moments estimator obtained using the first moment.

4. Give approximate confidence intervals using the asymptotic distribution of the above estimators. What can you conclude?

Exercise 3

Let $R_1, \ldots, R_n \overset{iid}{\sim} N(\mu, \sigma^2)$ be observations representing earnings of shares in a portfolio, where μ and σ^2 are unknown parameters. We would like to find an estimator of $\gamma = \mathbb{E}[R_1^3]$.

1. Write γ as a function of μ and σ^2.
2. Consider the estimator $\hat{\gamma} = (\frac{1}{n}\sum_{i=1}^n R_i)^3$
   (a) Derive the bias of $\hat{\gamma}$.
   (b) Is $\hat{\gamma}$ consistent? Support your answer with a mathematical argument.
3. Based on your work in the preceeding question, propose an unbiased estimator of μ^3 based on $(\frac{1}{n}\sum_{i=1}^n R_i)^3$ and show that it is unbiased.
4. Consider the estimator $\tilde{\gamma} = \frac{1}{n}\sum_{i=1}^n R_i^3$
   (a) Derive the bias of $\tilde{\gamma}$.
   (b) Is $\tilde{\gamma}$ consistent? Support your answer with a mathematical argument.
5. Using any of the preceeding unbiased estimators of γ, derive the minimum variance unbiased estimator of γ using the Rao-Blackwell Theorem.

Exercise 4

The dataset `kidney.txt` consists of measurements on 157 healthy volunteers (potential kidney donors). These data originated in the nephrology laboratory of Dr. Brian Myers, Stanford University. The two variables are the `age` of the volunteers in years and a composite measure "`tot`" of overall function. The goal of this exercise is to use straight line linear regression framework to analyze these data. Provide a minimum amount of `R` output to justify your answers.

1. Visualize the data and discuss the pertinence of fitting a straight line to this data set.
2. Which of the two variables would you interpret as your response variable?
3. What sign do you expect your two parameters to have? Justify this intuition and interpret the meaning of it these data.

4. How do you interpret the meaning of the parameters α and β if you assumed that your observations (y_i, x_i), $i = 1, \ldots, n$ were generated from the following two linear models

   (a) $Y_i = \alpha + \beta X_i + \varepsilon_i$, $i = 1, \ldots, n$,
   (b) $Y_i = \alpha + \beta(X_i - \overline{X_n}) + \varepsilon_i$, $i = 1, \ldots, n$,

   where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. noise variables with $\mathbb{E}[\varepsilon_1] = 0$ and $\text{var}(\varepsilon_1) = \sigma^2$.

5. Compute the least squares estimator of (α, β). Give your estimated values, and comment them. Are they statistically significant?

6. If you do not want to assume a linear model as in point 4, how do you interpret the least squares estimates that you gave in the previous question?

7. Give a prediction for the overall kidney function `tot` for a 100 years old person. Is this a good predictor given the data at hand?

8. Plot the residuals of your least squares fit. Does it seem reasonable to assume that errors ε_i are i.i.d.?

9. Give an exact 95% confidence interval for β assuming that the noise terms are i.i.d. normal. Compare it with a 95% asymptotic confidence interval that does not assume that the errors are normal. Discuss briefly their relative merits.

10. Generate 1000 bootstrap samples and use them to compute a 95% bootstrap confidence interval for β. Plot the bootstrap distribution that you obtained and compare your bootstrap confidence interval with the two obtained in point 9.

11. Compute the correlation between `age` and `tot` denoted by $\hat{\rho} = corr(y, x)$. Repeat this n times by leaving out one observation (y_i, x_i) and denote the resulting "leave-one-out" correlation by $\hat{\rho}_{(i)}$. Examine the differences $\hat{\rho}_{(i)} - \hat{\rho}$ and assess whether there are any observations (y_i, x_i) being particularly influencial in the analysis.

Exercise 5 (Optional bonus question)

Let's revisit the data set of scientific discoveries studied in Simonton (1979) [1]. You will check the pertinence of the statistical model proposed by the author and study an alternative approach. Print the code and output of your analysis when answering the questions below.

[1] D. K. Simonton (1979), "Independent Discovery in Science and Technology: A closer look at the Poisson distribution", *Social Studies of Science*, Vol. 8, No. 4, pp. 521–532.

1. Give a one paragraph summary of the goals of the paper and briefly describe the main statistical model used by the author. Does it seem like a reasonable choice to you? (Note: you don't need to read the whole paper to answer this question, just the beginning should suffice)

2. As an alternative approach, consider using the truncated Poisson as the frequency function for the number of simultaneuous scientific discoveries Y

$$\mathbb{P}(Y = k) = \frac{e^{-\mu}\mu^k}{k!}\frac{1}{1 - e^{-\mu} - \mu e^{-\mu}},\ k \geq 2.$$

   What advantages would an alternative statistical analysis which fits this model provide?

3. Compute the expectation and the variace of Y.

4. Write down the log likelihood and plot it for the data presented in Table 1.

5. Compute numerically the MLE of μ for the data presented in Table 1. Describe the algorithm you chose, print your results, and submit your code with this assignment.

6. Give the asymptotic distribution of $\hat{\mu}_{ML}$.

7. Give a 0.95 asymptotic confidence interval for μ.

8. In the paper Simonton appears to estimate the intensity parameter μ via an ad-hoc technique. First, a handful of reasonable values μ are chosen as candidates. Next, a sample size n_μ is selected for each of candidates such that observed counts in each bin closely matched the expected value under the model (this is done since data is not observed for $Y = 1$ and $Y = 0$, hence the "true" sample size n is not observed). Finally, for each pair (μ, n_μ) a χ^2 "goodness of fit" statistic is computed using the observed data. The μ which provided the best fit according to this statistic is selected as the estimate. Does this seem like a reasonable approach to fitting the desired model? What can you say about it mathematically? (e.g. can you attempt to answer questions about the estimator's bias, consistency, or variance?)

9. Comment on the estimate obtained by Simonton's technique as compared to the results obtained in the preceeding parts of this problem.