COLUMBIA UNIVERSITY — Prof. Marco Avella
Statistical Inference and Modeling — TA: Casey Bradshaw, Leon Fernandes
Graded Homework 1 — Spring 2020

You are free to present your solutions to the exercises below in *groups of at most three*. This homework is due on February 13 at the begining of that day's class. Good luck!

Exercise 1

Assume the life times of the neon lamps 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. Given iid samples $D_1, D_2, \cdots, D_n$, find a method of moments-based estimator of $Q_D(p)$ defined as a function of the first empirical moment.
3. Give an approximate confidence interval of $Q_D(p)$ based on the method of moments estimator from part (2).
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

The Poisson model assumes that the population mean equals the population variance. However, in practice it is common to encounter count data where the observed variance is significantly larger than mean. This phenomenon is called overdispersion. The goal of this exercise is to construct an asymptotic test of overdispersion, which you will propose and evaluate. You are free to present the results in either tables or plots, but limit yourselves to at most 2 pages of output including some lines of code. In your simulations, you will generate samples corresponding to the following two models:

(a) $X_1, \ldots, X_{50} \overset{iid}{\sim} Poisson(5)$

(b) $X_1, \ldots, X_{50} \overset{iid}{\sim} Poisson(\theta)$, where θ is a gamma random variable with $\mathbb{E}[\theta] = 5$ and $\text{var}[\theta] = 10$.

The first model will serve to verify that the test provides the desired confidence level. The second will serve to check the power of the test.
Let's start with some theoretical considerations.

1. Given $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$, find $\mathbb{E}[\overline{X}^2]$.

2. Show that if $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$, we have that $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2)$ is an unbiased estimator of λ.

3. With $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$, define new variables $Y_i = (X_i - \lambda)^2 - X_i$. Find $\mathbb{E}[Y_i]$ and $var(Y_i)$.

4. With s^2 as defined above, rewrite the quantity $s^2 - \overline{X}$ in terms of $Y_1, \ldots, Y_n$, $\overline{X}$, and λ.

5. Characterize the asymptotic distributions of $\sqrt{n}(s^2 - \overline{X})$ and $\sqrt{\frac{n}{2}}\frac{(s^2-\overline{X})}{\overline{X}}$

6. Based on the above result, propose a test of the null hypothesis $H_0 : \mathbb{E}[\overline{X}] = \mathbb{E}[s^2]$ at significance level α.

7. Illustrate the finite sample performance of your test by simulating 500 samples according to models (a) and (b). Briefly comment on your results.

8. The data presented in Table 1 below are daily numbers of deaths of women, with brain vessel diseases as cause of death. The data is for the year 1989 in West Berlin. Does your test detect some overdispersion for this data set?

Table 1: Female deaths by brain vessel disease in West Berlin, 1989

Deaths per day	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Frequency	1	4	15	31	39	55	54	49	47	31	16	9	8	4	3

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 preceding 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 preceding unbiased estimators of γ, derive the minimum variance unbiased estimator of γ using the Rao-Blackwell Theorem.

Exercise 4

The dataset `liver.csv` consists of measurements of measurements of the per capita liquor consumption and cirrhosis mortality rate for 46 different geographical regions. The two variables are `liquor` consumption per capita (ounces) and `cirrhosis` mortality rate, i.e. deaths per 100,000 people. Cirrhosis is condition in which the liver does not function properly due to long-term damage.
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 $\mathrm{var}(\varepsilon_1) = \sigma^2$.

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 cirrhosis mortality rate, `cirrhosis`, for a region with per capita liquor consumption of 180 ounces. 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 `liquor` and `cirrhosis` 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 influential 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 other and study an alternative approach. Print the code and output of your analysis when answering the questions below. The code and output should not exceed two pages.

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?

[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.

3. Compute the expectation and the variance 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 those candidates such that observed counts in each bin closely match 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.