EXERCISE SET 4

Exercise 1 (Bias and consistency)**:** Let

$$X \sim \text{Bernoulli}(q)$$

and $X_1, \ldots, X_N$ denote N samples of X. For each of the following estimators of q, (i) write down the standard error and (ii) state whether they are un-biased and/or consistent. (In each case, you can write down an exact formula for the standard error, so you do NOT need to use the CLT.)

(a)

$$\hat{q}_0 = \frac{1}{N}\sum_{i=1}^{N} X_i$$

(b)

$$\hat{q}_1 = \frac{Y}{N} + \frac{1}{\sqrt{N}}$$

(c)

$$\hat{q}_2 = \frac{1}{\lfloor N/2 \rfloor}\sum_{i=1}^{\lfloor N/2 \rfloor} X_i$$

The notation $\lfloor n \rfloor$ means the floor; that is, the largest integer less than n. For example, $\lfloor 101/2 \rfloor = \lfloor 50.5 \rfloor = 50$.

Exercise 2 (Estimator of mean in exponential model)**:** Let

$$T \sim \text{Exponential}(\lambda).$$

Recall that $E[T] = 1/\lambda$. We can estimate $E[T]$ via the sample average of measurements $T_1, \ldots, T_n$,

$$E[T] \approx \overline{T} = \frac{1}{n}\sum_{i=1}^{n} T_i.$$

This suggests that a natural way to estimate λ is by

$$\hat{\lambda} = \frac{1}{\overline{T}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} T_i}.$$

(a) The goal of the first part of this problem is to show, using simulations, that this is in-fact a biased estimator of λ, although the bias decreases with n. To achieve this, you should do the following:
- Make a list of 100 values of λ. You could use any range, but I picked between 0.2 and 2.
- For each value of λ,
  - simulate 10000 replicates of an experiment, where each replicate includes $n = 5$ values of T.
  - For each of these replicates, compute $\hat{\lambda}$ as defined above.
  - Then estimate the average $E[\hat{\lambda}]$ and save this value in a list.
- Make a plot of λ vs. $\left|E[\hat{\lambda}] - \lambda\right|$.

(b) (**optional – ungraded**) Consider the case $n = 2$. Prove that

$$E[\hat{\lambda}] = E\left[\frac{1}{\overline{T}}\right] \geq \lambda$$

This is a special case of Jenson's inequality.

Exercise 3 (Earnings data)**:** Consider the earnings data. This can be loaded with

```
> df = pd.read_csv("https://raw.githubusercontent.com/avehtari
> /ROS-Examples/master/Earnings/data/earnings.csv")
```

In this exercises, you will study the association between earnings and gender. In particular, you will explore how this depends on height. Later we will see there is a better way to answer this question by performing a regression with multiple predictors, but taking this more elementary approach will elucidate some key aspects of regression analysis.

(a) What do you expect the association between gender and earnings to be? Where do you expectations come from (news, intuition, other courses you've taken)?

(b) Using stats models, perform a linear regression on with gender (the column "male") as the predictor and earnings as the response variable. You can either use "earnk" or "earn", just keep track of the units. Then answer the questions
- Is there a statistically significant effect?
- Is the direction and size of the effect what you expected?

(c) Using stats models, perform a linear regression with height as the predictor and earnings as the response variable. Answer the same questions which are posed in part (a).

(d) You should have found there is an association between both gender and earnings, as well as height and earnings. A natural question is whether the association between height and earnings is simply a byproduct of the fact that men are taller on average. To answer this question, separate the data into males and females, then fit the linear regression model with height as a predictor separately for each group.
(e) Based on the results from the previous problem, what do you conclude? Is the association between height and earnings solely due to the association between gender and heights? Do you think it is partially due to the height?

Exercise 4 (Quiz practice)**:** Answer the questions on the practice quiz. You are welcome to provide some feedback on the quiz regarding difficultly level, clarity of problems, relevance to course material etc.

Exercise 5 (Statistical significance – **optional challenge**)**:** Show (using math OR simulations) that it is possible to conduct two experiments (let's use clinical trials as an example) so that $\Delta\hat{\mu}$ (using the same notation as my notes) is statistically significant for one experiment and not the other, yet the difference between $\Delta\hat{\mu}$ between the two experiments is not statistically significant. Here, by statistically significant I mean the p-value is < 0.05.