STAT430 Homework #3: Due Friday, February 18, 2022.

Name: KEY

Submit this homework as a pdf file to Canvas. I have created boxes after each problem for you to write you solution.

Note that when you 'Knit' the R Markdown script, a pdf is automatically created in the same folder as the .Rmd file. Thus, add your solutions to the R Markdown file, Knit it, and submit the created pdf. If you have questions about this process, ask a fellow student, come to office hours, or set up a meeting with me.

Question 1

My dog Peter snorts in the deep phase of his sleep, from midnight until 3:00am, at the rate of 12 snorts per hour. I am also in the deep phase of my sleep during this period, and he will wake me up if he shorts at least 25 times. We are interested in the probability (on any specific night) that such a situation occurs.

1. Provide the exact formula for the probability and compute the value. See the `ppois` function which is used similar to `pnorm` introduced above.

$S \sim$ Poisson(36). We want to know the probability $P(S \geq 25)$. The exact formula is

$$P(S \geq 25) = 1 - P(S < 25) = \sum_{i=1}^{24} \frac{36^i e^{-36}}{i!}$$

```
1-ppois(24, 36) #ppois inclusive of quantile value
```

```
## [1] 0.9775541
```

2. Using the CLT, explain how you could approximate the probability (i.e. what are the Y_is, their expectation, and variance). Give an approximation to the probability in part 1.

$S = \sum_i^{36} Y_i$, where $Y_1, \ldots, Y_{36} \overset{\text{iid}}{\sim}$ Poisson(1). Using CLT, we have

$$\frac{\bar{Y} - 1}{1/\sqrt{36}} = \frac{\sum_{i=1}^{36} Y_i - 36}{\sqrt{36}} = \frac{S - 36}{6} \overset{\text{approx}}{\sim} N(0, 1)$$

So,

$$P(S \geq 25) = P\left(\frac{S - 36}{6} \geq \frac{25 - 36}{6}\right) = P\left(\frac{S - 36}{6} \geq -1.83\right) \approx 1 - \Phi(-1.83).$$

And using R to find this, we have

```
1-pnorm(-1.83, 0, 1)
```

```
## [1] 0.966375
```

3. We can also approximate the probability using simulation. For example, suppose we were interested in the probability that a Poisson(10) random variable is greater than 15. We can simulate Poisson random variables in `R` using the function `rpois`. Take a look at the `help` for `rpois`:

```
help(rpois)
```

Notice that `rpois` takes two arguments. The first, labeled `n`, is the number of *realizations* (simulated random values) you want and `lambda` is the mean of the Poisson (the model parameter). To estimate the desired probability, we will simulate many, many random variables Poission(10) random variables. Here I generate 20 Poisson random variables with `lambda=10` and store them in a vector `Y`.

```
set.seed(4321)
Y <- rpois(20, 10)
Y
```

```
##  [1]  8  9 12 12  9 15  9 10 13  7 10 12  7  6 16 11  9  9  9 11
```

Notice that to make the results reproducible, I set the random seed generated to 4321 To empirically estimate the probability a Poisson(10) random variable is greater than 15, we look at the fraction of our simulated values that are greater than 15.

```
mean(Y>15)
```

```
## [1] 0.05
```

Note that we should use far more than 20 realizations to estimate a probability using simulation. A reasonable value would have been 1000 for example. Write your own code (or modify that above) to use simulation to estimate the probability that the snorts wake me up.

SAMPLE SOLUTION

```
# draw samples from distribution
pois_sample <- rpois(1000, 36)
# look at proportion of sample that is >= 25
sum(pois_sample >= 25)/length(pois_sample)
```

```
## [1] 0.97
```

Question 2

Suppose $Y_1, ..., Y_{40}$ denote a random sample of measurements on the proportion of weekends that the dogs of the Statistics Department (the dogs are not themselves employed by CSU—their owners are) go on hikes. Let each Y_i have probability density function given by

$$f(y) = 3y^2 \qquad 0 \leq y \leq 1.$$

I estimate that my dog Peter goes on a hike on 70% of weekends. Peter complains that all his friends go on hikes more often than he gets to. I am interested in whether the average Statistics dog goes hiking on more weekends than Peter. Approximate $P(\overline{Y} > 0.7)$.

By the CLT

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} N(0,1).$$

First find

$$\mu = E(Y_i) = \int_0^1 3y^3 dy = \frac{3}{4}y^4\Big|_0^1 = .75$$

and

$$\sigma^2(Y_i) = E(Y_i^2) - (E(Y_i))^2 = -(.75)^2 + \int_0^1 3y^4 dy = -.5625 + \frac{3}{5}y^5\Big|_0^1 = .0375$$

with $\sigma(Y_i) = .194$ Now, applying the CLT, we can find

$$P(\bar{Y} > 0.7) = P\left(\frac{\bar{Y} - 0.75}{0.194/\sqrt{40}} > \frac{0.7 - 0.75}{0.194/\sqrt{40}}\right) = P\left(\frac{\bar{Y} - 0.75}{0.194/\sqrt{40}} > -1.63\right) \approx 1 - \Phi(-1.63).$$

Using R, we can find this:

```
1-pnorm(-1.63, 0, 1)
```

```
## [1] 0.9484493
```

So the probability that the average CSU dog goes on a hike for more than 70% of weekends is approximately 94.8%.

Question 3

Researchers analyzed the 600 most popular songs since 2012 (as ranked on Billboard's Hot 100), and found that a small group of just 12 songwriters was responsible for 21% of these songs. Suppose you sample n = 35 songs randomly, with replacement, from the 600 most popular songs.

a) What is the exact probability that your 35 songs will include at least 11 songs from the small group of songwriters?

Let X be the random variable representing the number of songs from the small group of songwriters that you select. Then $X \sim$ Binomial(35, .21). The exact probability that your selection of 35 will include at least 11 songs from the group is

$$P(X \geq 11) = 1 - P(X < 11) = 1 - \sum_{i=1}^{10} \binom{35}{i} .21^i (1 - .21)^{35-i}.$$

We can use R to compute this value

```
1- pbinom(10, 35, .21) #pbinom is inclusive of quantile
```

```
## [1] 0.0991568
```

b) What is the approximate probability, using the CLT?

Since X is a binomial(35,.21) distributed random variable, it is the sum of $Y_1, \ldots, Y_{35} \overset{\text{iid}}{\sim}$ bernoulli(.21) trials. So we can use the Central Limit Theorem (with a continuity correction) to approximate:

$$P(X \geq 11) \approx 1 - P\left(\frac{X - 35(.21)}{\sqrt{35(.21)(1-.21)}} < \frac{10.5 - 35(.21)}{\sqrt{35(.21)(1-.21)}}\right) = 1 - P\left(\frac{X - 7.35}{\sqrt{5.8065}} < 1.307\right) \approx 1 - \Phi(1.51).$$

Using R to compute this:

```
1 - pnorm(1.307, 0, 1)
```

```
## [1] 0.09560636
```

Question 4

Suppose $\hat{\theta}$ is an estimator for a parameter θ and $E[\hat{\theta}] = a\theta + b$ for some non-zero constants a and b.

a) In terms of a, b, and θ, what is Bias($\hat{\theta}$)?

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta = a\theta + b - \theta = \theta(a-1) + b$$

b) Define a new estimator $\hat{\theta}_2$ which is a function of $\hat{\theta}$ and is an unbiased estimator for θ.

We can define a new parameter $\hat{\theta}^*$ such that $E(\hat{\theta}^*) = \theta$ with

$$\hat{\theta}^* = \frac{\hat{\theta} - b}{a}.$$

We can show that this is unbiased:

$$E(\hat{\theta}^*) = E\left(\frac{\hat{\theta} - b}{a}\right) = \frac{a\theta + b - b}{a} = \theta.$$

Question 5

Let $Y_1, ..., Y_n$ denote a random sample of size n from a population whose density is given by

$$f(y) = \frac{\alpha y^{\alpha-1}}{\theta^\alpha} \qquad 0 \le y \le \theta$$

where $\alpha > 0$ is a known, fixed value but θ is unknown. Consider the estimator $\hat{\theta} = \max(Y_1, ..., Y_n)$.

a) Show that $\hat{\theta}$ is a biased estimator of θ. Hint: Find the distribution of $\hat{\theta}$ by following the example from class. Start by expressing the c.d.f. of $\hat{\theta}$ as a function of the Ys.

To compute the Bias of the estimator, we need the distribution of $\hat{\Theta} = \max_i(X_i)$. We can compute this using the CDF method (unless you already have the formula for pdfs of order statistics, then use that.)

$$\begin{aligned} F_{\hat{\Theta}}(t) &= P(\max_i(Y_i) \le t) \\ &= P(Y_1 \le t, \ldots, Y_n \le t) \\ &= P(Y_1 \le t) \ldots P(Y_n \le t) \\ &= P(Y_1 \le t)^n && (Y_i \text{ are iid}) \\ &= F_Y(t)^n \end{aligned}$$

Then the PDF of $\hat{\Theta}$ is

$$\begin{aligned} f_{\hat{\Theta}}(t) &= \frac{d}{dt}\left[F_Y(t)^n\right] \\ &= nF_Y(t)^{n-1} f_Y(t) \\ &= n\left(\int_0^t \frac{\alpha y^{\alpha-1}}{\theta^\alpha} dy\right)^{n-1}\left(\frac{\alpha t^{\alpha-1}}{\theta^\alpha}\right) \\ &= n\left(\frac{t^\alpha}{\theta^\alpha}\right)^{n-1}\left(\frac{\alpha t^{\alpha-1}}{\theta^\alpha}\right) \\ &= n\left(\frac{\alpha t^{n\alpha-1}}{\theta^{n\alpha}}\right) && 0 \le t \le \theta \end{aligned}$$

Now we can compute the expectation of $\hat{\Theta}$.

$$\begin{aligned} E(\hat{\Theta}) &= \int_0^\theta n\left(\frac{\alpha t^{n\alpha}}{\theta^{n\alpha}}\right) dt \\ &= n\left(\frac{\alpha t^{n\alpha+1}}{(n\alpha+1)\theta^{n\alpha}}\right)\Big|_0^\theta \\ &= \frac{n\alpha\theta^{n\alpha+1}}{(n\alpha+1)\theta^{n\alpha}} \\ &= \frac{n\alpha\theta}{(n\alpha+1)} \end{aligned}$$

b) Define an *unbiased* estimator of θ.

Since the expectation of $\hat{\Theta}$ is a constant multiple of θ, we can define an unbiased estimator by simply multiplying by the reciprocal of the constant:

$$\hat{\Theta}^* = \frac{n\alpha+1}{n\alpha}\hat{\Theta}$$

c) Compute the MSE of the estimator you defined in part (b).

The MSE of an estimator is defined as

$$\text{MSE}(\hat{\Theta}) = E((\hat{\Theta} - \theta)^2).$$

If the estimator is unbiased, then this is simply the variance. So

$$\begin{aligned}
MSE(\hat{\Theta}a^*) &= E((\hat{\Theta}^* - \theta)^2) \\
&= \text{Var}(\hat{\Theta}^*) \\
&= \text{Var}\left(\frac{n\alpha+1}{n\alpha}\hat{\Theta}\right) \\
&= \left(\frac{n\alpha+1}{n\alpha}\right)^2 (E(\hat{\Theta}^2) - E(\hat{\Theta})^2) \\
&= \left(\frac{n\alpha+1}{n\alpha}\right)^2 \left[\int_0^\theta n\left(\frac{\alpha t^{n\alpha+1}}{\theta^{n\alpha}}\right) dt - \left(\frac{n\alpha\theta}{(n\alpha+1)}\right)^2\right] \\
&= \left(\frac{n\alpha+1}{n\alpha}\right)^2 \left(\frac{n\alpha\theta^2}{n\alpha+2} - \left(\frac{n\alpha\theta}{n\alpha+1}\right)^2\right) \\
&= \frac{(n\alpha+1)^2\theta^2}{n\alpha(n\alpha+2)} - \theta^2 \\
&= \frac{\theta^2}{n\alpha(n\alpha+2)}
\end{aligned}$$

Question 6

For each of the following, state whether the claim is true or false. If false, explain why and provide a counter example if appropriate.

a) There is a unique unbiased estimator $\hat{\theta}$ for each parameter θ.

False. You can have multiple estimators for a single parameter and they can be constructed to be unbiased.

b) If $\hat{\theta}$ is an unbiased estimator for θ, then $(\hat{\theta})^2$ is an unbiased estimator for θ^2.

False. Let $\hat{\theta}$ be an unbiased estimator of the parameter θ. Then $E(\hat{\theta}) = \theta$. However, we have

$$\begin{aligned} E(\hat{\theta}^2) &= \text{Var}(\hat{\theta}) + (E(\hat{\theta}))^2 \\ &= \text{Var}(\hat{\theta}) + \theta^2 \end{aligned}$$

Unless $\text{Var}(\hat{\theta}) = 0$ (i.e. $\hat{\theta}$ is a constant), $\hat{\theta}^2$ is a biased estimator of θ^2.

c) Let $Y_1, ..., Y_n$ be i.i.d. random variables such that $E[Y_i] = \mu$ and $\text{Var}[Y_i] = \sigma^2 < \infty$. By the CLT,

$$\frac{Y_1 - \mu}{\sigma} \to_d N(0, 1).$$

False. This quantity does not depend on n, so it converges to the distribution

$$F_{\frac{Y_1-\mu}{\sigma}} = P(Y_1 \leq \sigma y + \mu),$$

And since the distribution of Y_i was not specified, we do not know that this distribution is $N(0, 1)$.

d) An estimator is a function of sample observations and parameters.

True-ish and False-ish. An estimator can not be a function of the parameters it is intended to estimate. It CAN be a function of values of other estimators or known parameters.