Homework #7

Answer the following questions in a single R script called `hw07.R`. Answers must be given by R commands. You cannot simply look at the data set and answer the question via direct inspection. Use comments (#) to indicate which portion of your code answers which question. Be sure that you obtain the correct solutions to each question when you execute your script one line at a time from top to bottom.

Each question will be graded according to the following criteria:

- 0%: No attempts is made to answer the question.
- 25%: An attempt is made that, although unsuccessful, revealed some understanding of what the question was asking.
- 50%: Solution is incorrect, but with some modifications, could be corrected.
- 75%: Solution is incorrect, but easily resolved with minor modifications **OR** solution is correct, but obtained via convoluted reasoning or by avoiding standard approaches.
- 100%: Solution is correct and uses standard approaches.

#1) A random variable X has a **gamma distribution** if it has a probability density function:

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

where k and θ are parameters such that $k > 0$ and $\theta > 0$. We denote this by writing $X \sim \Gamma(k, \theta)$. We call k the **shape** and θ the **scale**.

(a) Create a single graphic with plots of the density functions with the following parameters and colors:

   (i) $k = 1$, $\theta = 1$, blue

   (ii) $k = 2$, $\theta = 1$, red

   (iii) $k = 3$, $\theta = 1$, green

   (iv) $k = 1$, $\theta = 2$, orange

   (v) $k = 2$, $\theta = 2$, purple

   (vi) $k = 3$, $\theta = 2$, black

   (Hint: In `dgamma`, k is given by `shape` and θ is given by `scale`.)

(b) The following code snippet takes 10,000 samples of size $n = 25$ from $X \sim \Gamma(2, 1)$.

```
n <- 25
tbl <- tibble(index=1:10000) %>% rowwise() %>%
  mutate(X = list(rgamma(n, shape=2, scale=1)))
```

For each sample, compute:

$$\hat{\theta} = \overline{x\ln(x)} - \overline{x} \cdot \overline{\ln(x)}$$

where

$$
\begin{aligned}
\overline{x} &= \frac{1}{n}\sum x_i, \\
\overline{\ln(x)} &= \frac{1}{n}\sum \ln(x_i), \\
\overline{x\ln(x)} &= \frac{1}{n}\sum x_i \cdot \ln(x_i).
\end{aligned}
$$

In R, $\ln(x)$ is given by `log(x)`.

(c) Plot a histogram for the values of $\hat{\theta}$. Then compute the mean of $\hat{\theta}$. Do you think $\hat{\theta}$ is an unbiased estimator? Be sure to justify your answer. Type your answer within a comment in your R script.

(d) For each sample, compute

$$\hat{k} = \frac{\overline{x}}{\hat{\theta}}.$$

Plot a histogram for the values of $\hat{k}$. Then compute the mean of $\hat{k}$. Do you think $\hat{k}$ is an unbiased estimator? Be sure to justify your answer. Type your answer within a comment in your R script.

(e) For each sample, compute

$$
\begin{aligned}
\tilde{\theta} &= \frac{n}{n-1} \cdot \hat{\theta} \\
\tilde{k} &= \hat{k} - \frac{1}{n}\left(3\hat{k} - \frac{2}{3}\cdot\frac{\hat{k}}{1+\hat{k}} - \frac{4}{5}\cdot\frac{\hat{k}}{(1+\hat{k})^2}\right).
\end{aligned}
$$

(f) Plot a histogram for the values of $\hat{\theta}$ and $\tilde{\theta}$. Make sure the histograms have different colors and set `alpha=0.5` for both. Then compute the mean of $\tilde{\theta}$. Do you think $\tilde{\theta}$ is an unbiased estimator? Be sure to justify your answer. Type your answer within a comment in your R script.

(g) Plot a histogram for the values of $\hat{k}$ and $\tilde{k}$. Make sure the histograms have different colors and set `alpha=0.5` for both. Then compute the mean of $\tilde{k}$. Do you think $\tilde{k}$ is an unbiased estimator? Be sure to justify your answer. Type your answer within a comment in your R script.

#2) Let X count the number of students who drop by my office on Friday, 3-4pm. We take $n = 40$ samples of X.

(a) Suppose $\mu = \mathrm{E}(X) = 4$ and $\sigma^2 = \mathrm{Var}(X) = 2$. Find $P(3.9 < \overline{X} < 4.1)$.

(b) Generate 10,000 samples of size $n = 40$ from $X \sim \mathcal{N}(4, 2)$. Compute the sample mean for each sample. Find the percentage of samples that were between 3.9 and 4.1.

(c) Find a two-sided confidence interval with $1 - \alpha = 0.99$ for μ.

(d) Generate 10,000 samples of size $n = 40$ from $X \sim \mathcal{N}(4, 2)$. Compute the sample mean for each sample. Find the percentage of samples that fall within the confidence interval you found in part (c).

(e) Find a lower one-sided confidence interval with $1 - \alpha = 0.95$ for μ.

(f) Find an upper one-sided confidence interval with $1-\alpha = 0.95$ for μ.

#3) Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. In $n = 20$ samples, we compute a sample variance of $s^2 = 5$. Find a $1-\alpha = 0.99$ two-sided confidence interval for σ^2.

#4) For this exercise we will look at survey data from Pew Research on American's understanding and sentiments toward religion in February in 2019. The archive `W44_Feb19.zip` contains a data file `ATP W44.sav` along with various documents. This data file can be loaded into R via the following code snippet:

```r
# install.packages("haven")
library(haven)
df <- read_sav('ATP W44.sav')
```

One of the questions from this survey asks respondents whether they have taken a course in world religions. Let X be a Bernoulli random variable where $X = 1$ if the respondent did take a course and $X = 0$ if the respondent did not. For a Bernoulli random variable, $\mu = \mathrm{E}(X)$ is identical to the probability of success θ. Since our sample size n for this survey is quite large, we can assume $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

(a) Compute the sample mean $\overline{x}$ and sample variance s^2 from the survey data with the assumption that $x_i = 1$ if the respondent did take a course in world religions and $x_i = 0$ if the respondent did not take a course in world religions (or failed to answer the question).

(b) Since X is a Bernoulli random variable, we know $\mathrm{Var}(X) \leq 0.5$ regardless of the probability of success θ (I showed this in class). Presuming this bound, compute a 99% confidence interval for μ the proportion of American's that have taken a course in world religions.

(c) Look at the reported percentages given to this question in `ATP W44 topline.pdf`. What might explain the disparity between our result and the one given in the report? Type your answer as a comment in your script. (Hint: Look through `ATP W44 methodology.pdf`.)