Assignment 1

Please show all the steps and keep the final answer 3-4 decimal places.

Q1 (1+3+1)

Suppose that the percentage of American drivers who multitask (e.g., talk on cell phones, eat a snack, or text at the same time they are driving) is approximately 80%. In a random sample of $n = 20$ drivers, let X equal the number of multitaskers. Review the Binomial distribution from extra notes, Unit 03.

a. How is X distributed?

ANSWER: X follows a binomial distribution. This is because, the sample size $n = 20$ is fixed and performed the exact way n times. There are also on two possible outcomes, an individual is a multitasker or they are not a multitasker. Each "trial" is independent, and the probability of being a multitasker is $p = 0.8$, while the probability of not being a multitasker is $q = 1 - p = 0.2$

b. What are the values of the mean, variance, and standard deviation of X.

ANSWER: We know for the binomial distribution, the mean $\mu = np$, where $n = 20$ and $p = 0.8$. Thus, $\mu = 20 * 0.8 = 16$

Similarly, the variance $\sigma^2 = np(1-p) = (20 * 0.8)(1 - 0.8) = 16 * 0.2 = 3.2$

And finally, the standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{3.2} = 1.7889$

c. What is the probability that there are more than three multitaskers in the sample?

ANSWER: $P(X > 3) = 1 - P(X \leq 3) = 1 - (P(X=0) + P(X=1) + P(X=2) + P(X=3)) = 1 - 0.0000 = 1$

Using R function 1 - pbinom(), we have found that the probability that there are more than three multitaskers in the sample is 1.

```
#insert your code to calcualte the probability in (c)
prob = 1 - pbinom(3, size=20, prob=0.8)
print(prob)
```

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

Q2 (1+2)

The mean and standard deviation of service times for customers at a post office are known to be 2.93 minutes and 1.79 minutes, respectively. Review extra notes, Unit 07 about the central limit theorem.

a. Can you calculate the probability that the average service time for the next two customers is less than 2.7 minutes? If so, calculate the probability. If not, explain why not.

ANSWER: The probability of the next two customers cannot be calculated, due to the central limit theorem needing a "sufficiently large" sample size n for the sample mean $\bar{X}$ to follow an approximate normal distribution.

b. What is the approximate probability that the total service time for the next 65 customers exceeds three hours? (Assume that the next 65 customers represent a simple random sample and that there is always a customer waiting in line.)

ANSWER: Time = 180 minutes so $\bar{X} = \frac{180}{65} = 2.769$

$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{2.769-2.93}{1.79/\sqrt{65}} = -0.7252$

```
#insert your code to calcualte the probability in (b)
prob = 1 - pnorm(-0.7252)
print(prob)
```

```
## [1] 0.7658353
```

Thus, $P(\bar{X} > 180) = 1 - P(\bar{X} \leq 180) = 0.7658$

Meaning that the total service time exceeding 3 hours for 65 customers has a 0.7658 probability.

Q3 (5)

We will classify emails as "Spam" or "Not Spam" based on word counts. The words considered are {buy, now, free}. For the training dataset, the summary of those words are

Class	buy	now	free	total
Spam	20	5	10	35
Not Spam	5	15	5	25

For a new email, the word counts are {buy: 1, now: 0, free: 2}. What class this email should be classfied as using the Naive Bayes?

ANSWER: Let $P(Y=1) = \frac{35}{60} = \pi_1$ represent the proportion of spam emails and $P(Y=0) = \frac{25}{60} = \pi_0$ represent the proportion of not spam emails.

Let $w = (1, 0, 2)$, representing the words $(buy, now, free)$ in our new email respectively.

$P(Y=1|w) = \frac{P(w|Y=1)P(Y=1)}{P(w|Y=1)P(Y=1)+P(w|Y=0)P(Y=0)} = \frac{\pi_1 P(w_1=1,w_2=0,w_3=2|Y=1)}{\pi_1 P(w|Y=1)+\pi_0 P(w|Y=0)}$

Also note that $P(w_1 = 1, w_2 = 0, w_3 = 2|Y = 1) = \prod_{i=1}^{3} P(w_i|Y = 1)$

Now, let's calculate the probability that this email is a spam

$P(Y=1|w) = \frac{(35/60)(20/35)(10/35)^2}{(35/60)(20/35)(10/35)^2+(25/60)(5/25)(5/25)^2} = 0.8909$

Then,

$P(Y=0|w) = 1 - P(Y=1|w) = 0.1091$

Since $P(Y=1|w) > P(Y=0|w)$, This message is a spam email.

Q4 (5)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the exponential distribution whose pdf is

$$f(x;\theta) = (1/\theta)e^{-x/\theta}, 0 < x < \infty, 0 < \theta < \infty.$$

a. Show that the maximum likelihood estimator of θ is $\bar{X}$.

ANSWER: So we have our PDF $f(x;\theta) = (1/\theta)e^{-x/\theta}$

Then, we can calculate the likelihood function $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n}(1/\theta)e^{-x/\theta} = \theta^{-n}e^{\frac{-\sum_{i=1}^{n} x_i}{\theta}}$

Now, we can calculate the log-likelihood function $l(\theta) = lnL(\theta) = ln(\theta^{-n}e^{\frac{-\sum_{i=1}^{n} x_i}{\theta}}) = -nln\theta - \frac{\sum_{i=1}^{n} x_i}{\theta}$

Now, we find the derivative of the log-likelihood function with respect to θ and set it equal to 0: $\frac{dl(\theta)}{d\theta} = \frac{-n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0$

$\frac{n}{\theta} = \frac{\sum_{i=1}^{n} x_i}{\theta^2}$

$\frac{n\theta^2}{\theta} = \sum_{i=1}^{n} x_i$

$n\theta = \sum_{i=1}^{n} x_i$

$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}$

Therefore, the maximum likelihood estimator for θ is $\bar{X}$

b. What is the maximum likelihood estimate of θ if a random sample of size 5 yielded the sample values 3.5, 8.1, 0.9, 4.4, and 0.5?

ANSWER: from the data, $\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{3.5+8.1+0.9+4.4+0.5}{5} = 3.48$

Therefore, the maximum likelihood estimate of θ is 3.48

Q5 (5)

Use the following CommuteAtlanta data and consider the commute time.

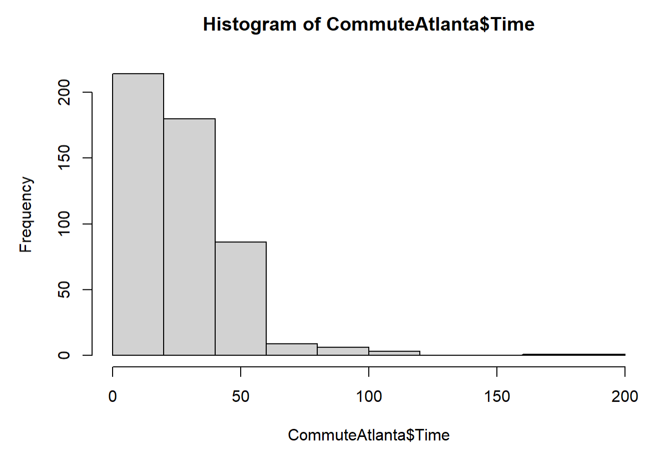
```
## Bootstrap confidence interval
library(Lock5Data)
library(boot)
```

```
## Warning: package 'boot' was built under R version 4.2.3
```

```
data(CommuteAtlanta)
str(CommuteAtlanta)
```

```
## 'data.frame':    500 obs. of  5 variables:
##  $ City    : Factor w/ 1 level "Atlanta": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age     : int  19 55 48 45 48 43 48 41 47 39 ...
##  $ Distance: int  10 45 12 4 15 33 15 4 25 1 ...
##  $ Time    : int  15 60 45 10 30 60 45 10 25 15 ...
##  $ Sex     : Factor w/ 2 levels "F","M": 2 2 2 1 1 2 2 1 2 1 ...
```

```
hist(CommuteAtlanta$Time)
```



(1). Using the boot() function, calculate the 95% bootstrap confidence interval for the median commute time.

```
#insert your code
my.median = function(x,indices){
  return(median(x[indices]))
}

time.boot = boot(CommuteAtlanta$Time, my.median, 10000)

boot.ci(time.boot)
```

```
## Warning in boot.ci(time.boot): bootstrap variances needed for studentized
## intervals
```

```
## Warning in norm.inter(t, adj.alpha): extreme order statistics used as endpoints
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = time.boot)
##
## Intervals :
## Level      Normal              Basic
## 95%   (21.41, 27.36 )   (20.00, 25.00 )
##
## Level     Percentile            BCa
## 95%   (25, 30 )   (20, 20 )
## Calculations and Intervals on Original Scale
## Warning : BCa Intervals used Extreme Quantiles
## Some BCa intervals may be unstable
```

(2). We create 1000 bootstrap samples of median commute time.

```
#Program the basic bootstrap by ourselves
#To construct the confidence interval for the mean commute time in Atlanta, we need to find the
#point estimate (sample median) from the original sample.
time.median = with(CommuteAtlanta, median(Time))
time.median
```

```
## [1] 25
```

```
# To find the standard error, we will create a huge matrix with 1000
# rows (one for each bootstrap sample) and 500 columns
# (one for each sampled value, to match the original sample size).
# We will then use apply() to apply mean() to each row of the matrix.
B = 1000
n = nrow(CommuteAtlanta)
boot.samples = matrix(sample(CommuteAtlanta$Time, size = B * n, replace = TRUE), B, n)
boot.statistics = apply(boot.samples, 1, median)
```

Calculate the 95% normal bootstrap condfidence interval and 95% percentile bootstrap confidence interval. Compare the results with the boot() function results

```
# insert your code
# 95% normal Bootstrap CI
boot.sd = sd(boot.statistics)
z = qnorm(0.975)
normal_ci = c(time.median-z*boot.sd, time.median+z*boot.sd)

# 95% percentile bootstrap CI
percentile_ci = quantile(boot.statistics, c(0.025,0.975))

# Display Results
cat("Normal CI: (", normal_ci[1], ",", normal_ci[2], ")\n")
```

```
## Normal CI: ( 22.07324 , 27.92676 )
```

```
cat("Percentile CI: (", percentile_ci[1], ",", percentile_ci[2], ")\n")
```

```
## Percentile CI: ( 25 , 30 )
```

```
# boot() function
cat("\nboot() CI's:\n")
```

```
##
## boot() CI's:
```

```
print(boot.ci(time.boot, type = c("norm", "perc")))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = time.boot, type = c("norm", "perc"))
##
## Intervals :
## Level      Normal             Percentile
## 95%   (21.41, 27.36 )   (25.00, 30.00 )
## Calculations and Intervals on Original Scale
```

Comparing with the boot() function results, we can see that the percentile intervals are exactly the same, while the normal intervals have slight differences. This slight difference may be due to the calculation of the standard error.

Q6 (3+2+2+2+1+1)

Suppose that X follows a Poisson distribution $Poi(\lambda)$ with probability mass function

$$f_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

The prior distribution of λ is a Gamma distribution $Gamma(\alpha, \beta)$ with density function

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda\beta}$$

(1). With data $x_1, \ldots, x_n$ from the Possion distribution $\mathcal{P}(\lambda)$, using the Bayesian approach, show that the posterior distribution of λ follow the Gamma distribution

$$\pi(\lambda|x_1, \ldots, x_n) \sim Gamma(\alpha + n\bar{x}, n + \beta)$$

ANSWER: Since $f_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, $f_\lambda(x_1, .., x_n) = \prod_{i=1}^{n} \frac{\lambda^x e^{-\lambda}}{x!} = \lambda^{\sum_{i=1}^{n} x_i}e^{-n\lambda}(\prod_{i=1}^{n} x_i!)^{-1}$

We know that $\pi(\lambda|x_1, .., x_n) = \pi(\lambda)f_\lambda(x_1, .., x_n)/f(x_1, .., x_n) = \frac{(\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda\beta})(\lambda^{\sum_{i=1}^{n} x_i}e^{-n\lambda}(\prod_{i=1}^{n} x_i!)^{-1})}{f(x_1,..,x_n)} = \frac{\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{a+n\bar{x}-1}e^{-\lambda(\beta+n)}(\prod_{i=1}^{n} x_i!)^{-1}}{f(x_1,..,x_n)}$

Now we can work on our denominator:
$f(x_1, .., x_n) = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{a+n\bar{x}-1}e^{-\lambda(\beta+n)}(\prod_{i=1}^{n} x_i!)^{-1}d\lambda = \frac{\beta^\alpha}{\Gamma(\alpha)}(\prod_{i=1}^{n} x_i!)^{-1}\int_0^\infty \lambda^{a+n\bar{x}-1}e^{-\lambda(\beta+n)}d\lambda$

To solve the integral, let $u = (\beta + n)\lambda, \lambda = \frac{u}{\beta+n}, d\lambda = \frac{1}{\beta+n}du$

Which we can substitute into our integral:
$f(x_1, .., x_n) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\prod_{i=1}^{n} x_i!)^{-1}(\beta+n)^{-(a+n\bar{x})}\int_0^\infty u^{a+n\bar{x}-1}e^{-u}du = \frac{\beta^\alpha}{\Gamma(\alpha)}(\prod_{i=1}^{n} x_i!)^{-1}(\beta+n)^{-(a+n\bar{x})}\Gamma(a+n\bar{x})$ (using the gamma function formula)

Now we can fill in our posterior distribution and simplify:
$\pi(\lambda|x_1, .., x_n) = \frac{\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha+n\bar{x}-1}e^{-\lambda(\beta+n)}(\prod_{i=1}^{n} x_i!)^{-1}}{\frac{\beta^\alpha}{\Gamma(\alpha)}(\prod_{i=1}^{n} x_i!)^{-1}(\beta+n)^{-(\alpha+n\bar{x})}\Gamma(\alpha+n\bar{x})} = \frac{(\beta+n)^{\alpha+n\bar{x}}}{\Gamma(\alpha+n\bar{x})}\lambda^{\alpha+n\bar{x}-1}e^{-\lambda(\beta+n)} \sim Gamma(\alpha + n\bar{x}, n + \beta)$

(2). What is the posterior mean and variance of λ?

ANSWER: The mean and variance of the gamma distribution is as follows

MEAN:$\mu = \frac{\alpha+n\bar{x}}{\beta+n}$ and $\sigma^2 = \frac{\alpha+n\bar{x}}{(\beta+n)^2}$

(3). Derive the MAP estimator of λ.

ANSWER: $\pi(\lambda|x_1, .., x_n) \propto \lambda^{\alpha+n\bar{x}-1}e^{-\lambda(\beta+n)}$ (since we can ignore "constants" with respect to λ)

Now, lets find log-likelihood:
$ln(\pi(\lambda|x_1, .., x_n)) = ln(\lambda^{\alpha+n\bar{x}-1}e^{-\lambda(\beta+n)}) = ln(\lambda^{\alpha+n\bar{x}-1}) + ln(e^{-\lambda(\beta+n)}) = (\alpha + n\bar{x} - 1)ln\lambda - \lambda(\beta + n)lne = (\alpha + n\bar{x} - 1)ln\lambda - \lambda(\beta$

Next, we find the derivative w.r.t λ and set it equal to 0: $\frac{d}{d\lambda}((\alpha + n\bar{x} - 1)ln\lambda - \lambda(\beta + n)) = 0$

$=> \frac{\alpha+n\bar{x}-1}{\lambda} - (\beta + n) = 0$

$=> \lambda = \frac{\alpha+n\bar{x}-1}{\beta+n} = \hat{\lambda}_{MAP}$

(4). Suppose that we record the number of a specific bacteria present in 20 water samples taken in the Mekong Delta (Vietnam) so that we have the following data at hand:

$$x_i = 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 2, 0, 0, 5, 2, 0, 0, 2, 0, 1$$

Suppose that data follows $Poi(\lambda)$ and the prior $\lambda \sim Gamma(\alpha = 2, \beta = 2)$, calculate the the posterior mean and the MAP estimator of λ.

ANSWER: $n = 20, \bar{x} = \frac{\sum_{i=1}^{20} x_i}{20} = \frac{17}{20} = 0.85, \alpha = 2, \beta = 2$

so, $\hat{\lambda}_{MAP} = \frac{2+17-1}{2+20} = 18/22 = 0.8182$

$\mu = \frac{2+17}{2+20} = 19/22 = 0.8636$

$$

(5). Plot the prior and posterior density function of λ.

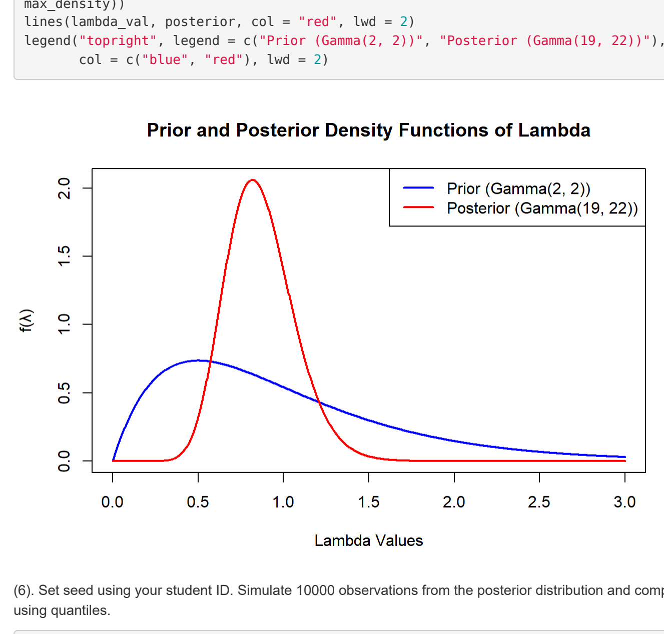
```
# insert your code
# lambda range (I chose an arbritrary 0-3)
lambda_val = seq(0,3,length.out=1000)

# prior density ~ Gamma(2,2)
prior = dgamma(lambda_val, shape=2, rate=2)

# posterior density ~ Gamma(2+17, 2+20) = Gamma(19,22)
posterior = dgamma(lambda_val, shape=19, rate=22)

max_density <- max(prior, posterior)

# plot densities
plot(lambda_val, prior, type = "l", col = "blue", lwd = 2,
     xlab = "Lambda Values", ylab = "f(λ)", main = "Prior and Posterior Density Functions of Lambda",ylim = c(0,
max_density))
lines(lambda_val, posterior, col = "red", lwd = 2)
legend("topright", legend = c("Prior (Gamma(2, 2))", "Posterior (Gamma(19, 22))"),
       col = c("blue", "red"), lwd = 2)
```



(6). Set seed using your student ID. Simulate 10000 observations from the posterior distribution and compute the 95% confidence interval of λ using quantiles.

```
# insert your code
set.seed(501112462)

# 10000 observations of posterior observation ~ Gamma(19,22)
posterior_observations = rgamma(10000, shape=19, rate=22)

# 95% confidence interval
ci = quantile(posterior_observations, probs=c(0.025,0.975))

print(ci)
```

```
##      2.5%     97.5%
## 0.5194879 1.3005391
```