Consider the dataset SurveyFall2019.csv collected from the Getting To Know You Survey (see separate pdf file for description on the dataset). We will look at the distribution of the number of tattoos for male vs. female students in STAT3011 using R. Let X be the random variable representing the number of tattoos for female students and Y be the random variable for males. Use the following codes to load the dataset and show relevant R codes when addressing the questions below.

```
data <- read.csv("http://users.stat.umn.edu/~parky/SurveyFall2019.csv", header = T)
attach(data)
```

a. Examine the distribution for number of tattoos based on whether the student is from US.

```
> table(tattoos,from.US)
       from.US
tattoos International Prefer not to answer U.S.
     0             33                     2  255
     1              2                     0   31
     2              1                     0   12
     3              0                     0   10
     4              0                     0    1
     5              0                     0    4
     6              0                     0    2
     7              0                     0    3
     10             0                     0    2
```

b. Find the number of U.S. students in the dataset and construct the probability distribution.

```
> length(which(from.US=="U.S."))
[1] 320
> round(table(tattoos[from.US=="U.S."])/320, 3)

    0     1     2     3     4     5     6     7    10
0.797 0.097 0.038 0.031 0.003 0.012 0.006 0.009 0.006
```

c. Why does it not make sense to compute the mean number of tattoos that an international student has as $(0+1+2+3+4+5+6+7+10)/9$?

Since these probabilities are not all equal, one cannot simply compute the mean by adding up observed numbers and divide by total possible values of tattoos.

d. Find the probability that a randomly selected U.S. student has exactly 3 tattoos.

$P(X = 3) = 0.031$

e. Find the probability that a randomly selected U.S. student has over 6 tattoos.

$P(X > 6) = P(X = 7) + P(X = 10) = 0.015$

f. Find the 90th percentile of the distribution assuming it follows a normal distribution.

```
> qnorm(.9,mean=mean(tattoos),sd=sd(tattoos))
[1] 2.170215
```

g. Explore the distribution of exercise per day for male and female students with a side-by-side boxplot.

```
> boxplot(data$exercise~data$gender, main="Exercise by Gender", xlab="Gender",
ylab="Amount of exercise")
```

The boxplot looks similar between the two genders except for a few more extreme outliers for male students. The amount of exercise appears skewed to the right for male and female students since the upper whisker is slightly longer than the lower whisker. The median is also about the same for male and female students.

h. Calculate the probability that a randomly selected student works out greater than or equal to 2 hours assuming it follows a normal distribution.

```
> pnorm(2, mean=mean(data$exercise), sd=sd(data$exercise), lower=F)
[1] 0.3302893
```

i. Calculate the mean and standard deviation of the number of hours exercising for male and female students using mean(), sd() or tapply() commands, assuming it follows a normal distribution.

```
> tapply(exercise, gender, mean)
  Female     Male
1.304739 1.619048
> tapply(exercise, gender, sd)
  Female     Male
1.128252 1.475059
```

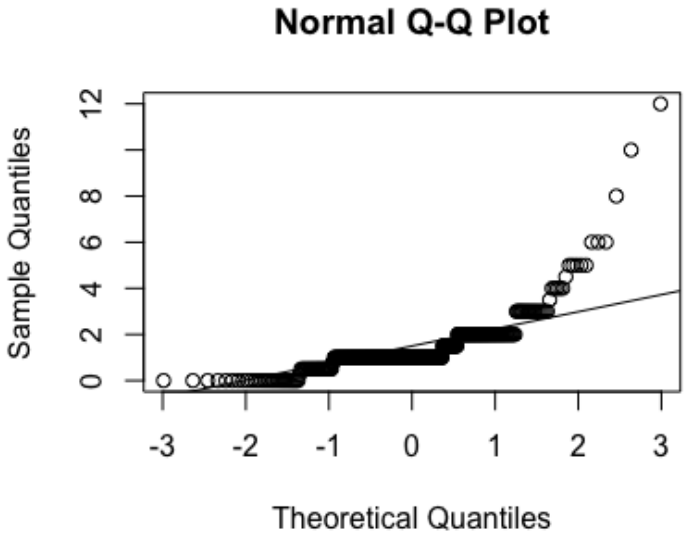
j. Calculate the probability of observing a female student working out more than 2 hours using the pnorm() command, assuming it follows a normal distribution.

$P(X > 2) = P(\frac{2-1.30}{1.13}) = P(Z > 0.62)$

```
> pnorm(0.62,lower=F)
[1] 0.2676289
```

k. Construct a Q-Q plot of exercise and assess whether the data is normal or not.

```
qqnorm(data$exercise)
qqline(data$exercise)
```



The distribution is heavy-tailed and the data is not normally distributed. A histogram of the physical exercise also confirms the Q-Q plot. The Q-Q plot here displays upward deviations from the straight line at the extremes indicating that the data is right-skewed.