STAT5101 Foundations of Data Science Assignment 2

Yiu Chung WONG 1155017920

1. The dean of a business school wishes to form an executive committee of 5 from among the 40 tenured faculty members at the school. The selection is to be random, and at the school there are 8 tenured faculty members in accounting.

(a) What is the probability that the committee will contain at least 1 of the accounting faculty members?

```
N <- 40 #Total number of faculty members
m <- 8 #Total number of accounting faculty members
k <- 5 #Number of draws
x <- 1 #Number of accounting faculty members selected
phyper(q = x-1, m = m, n = N-m, k = k, lower.tail = FALSE)
```

```
## [1] 0.6939612
```

(b) Prof. Chan is a tenured faculty member at the business school, what is the probability that Prof. Chan will be selected as a committee member.

```
5/40
```

```
## [1] 0.125
```

(c) If Prof. Chan is a tenured faculty member in accounting at the business school, what is the probability that Prof. Chan will be selected as a committee member.

```
5/40
```

```
## [1] 0.125
```

(d) According to a survey research, the probability a professor catches a cold during winter is 0.2. Assume 10 professors are randomly selected. Consider the random variable defined by the number of professors, among the 10 professors, that catch a cold during winter. Propose an appropriate distribution for the random variable. Based on the distribution, calculate the probability that at least 2 professors catch a cold.

- Random Variable: number of professors catching cold.
- Binomial Distribution.

```
N = 10 #Total number of professors selected.
p = 0.2 #Probability of a professor catches a cold during winter
x = 2 #Number of professor catching a cold
pbinom(q = x-1, size = N, prob = p, lower.tail = FALSE)
```

```
## [1] 0.6241904
```

2. Assume that the flaws along a magnetic tape follow a Poisson distribution with a mean of 0.1 flaw per meter. Let Y denote the distance between two successive flaws.

(a) What is the mean of Y?

```r
1/0.1
```

```
## [1] 10
```

(b) What is the probability that there are no flaws in 5 consecutive meters of tape?

```r
dpois(x = 0, lambda = 5 * 0.1)
```

```
## [1] 0.6065307
```

3. Cinema advertising is increasing. According to a survey research, the probability a viewer will remember a cinema advertisement is 0.74.

(a) Suppose that 10 viewers of a cinema advertisement are randomly sampled. Consider the random variable defined by the number of viewers who recall the advertisement. What assumptions must be made in order to assume that this random variable is distributed as a binomial random variable?

- The number of observations n is fixed.
- Knowing one viewers's answer to recalling the advertisement gives no information to the answer of another viewer.
- Probability of viewers recalling the advertisement remains constant.
- The viewers either recall or don't recall. There is no third outcome.
- Sampling with replacement: we might interview the same viewer more than once to ensure each trial is independent to each other and the probability of success remains constant.

Assuming that the number of viewers who recall the cinema advertisement is a binomial random variable,

(b) what is the probability of at least two viewers who recall the advertisement?

```r
N <- 10
p <- 0.74
x <- 2
pbinom(q = x-1, size = N, prob = p, lower.tail = FALSE)
```

```
## [1] 0.9999584
```

(c) what are the mean and standard deviation of this distribution?

```r
c <- c(N*p, sqrt(N*p*(1-p)))
names(c) <- c("mean", "sd")
c
```

```
##     mean       sd
## 7.400000 1.387083
```

4. The time between arrivals of customers at a bank during the noon to 1 P.M. hour has a uniform distribution over an interval from 0 to 120 seconds. What is the probability that the time between the arrival of two customers will be

####(a) between 10 and 30 seconds?

```r
punif(q = 30, min = 0, max = 120) - punif(q = 10, min = 0, max = 120)
```

```
## [1] 0.1666667
```

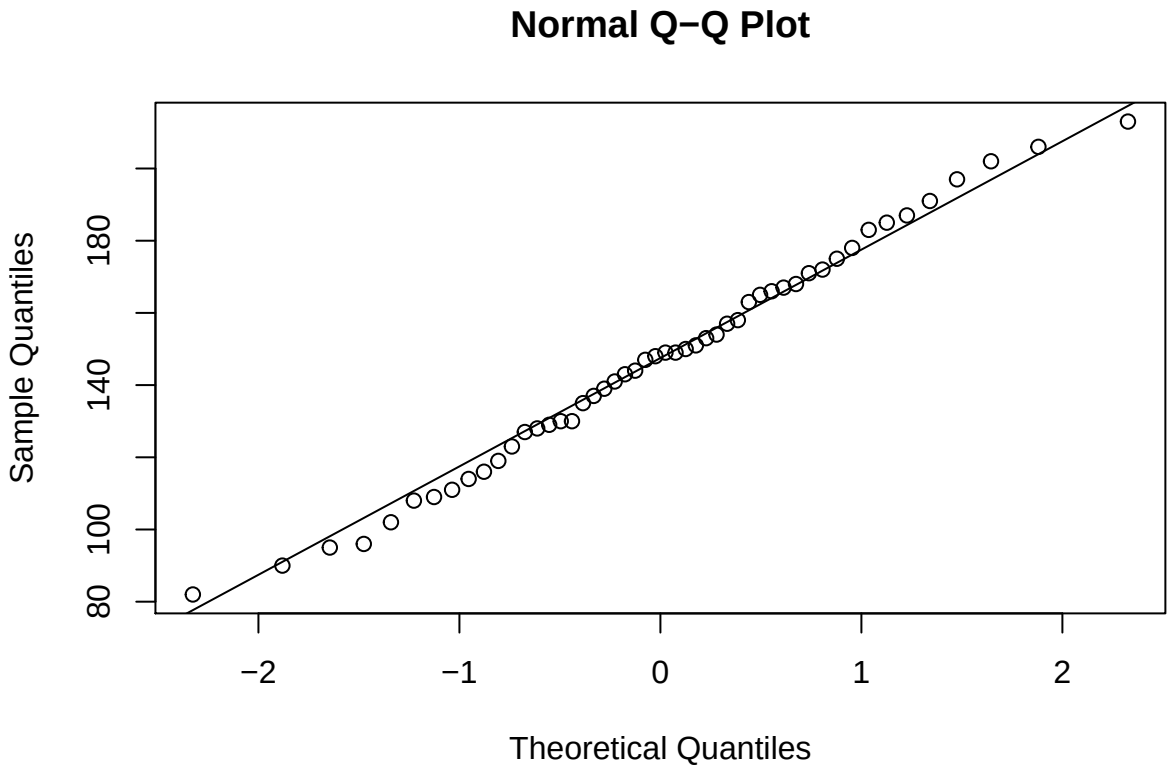
####(b) What is the expected value and the standard deviation of the time between arrivals?

```r
q4b <- c(
        (120 + 0)/2,
        sqrt((120 - 0)^2 * 1/12)
        )
names(q4b) <- c("Epected Value a", "Standard Deviation")
q4b
```

```
##    Epected Value a Standard Deviation
##           60.00000           34.64102
```

5. The data *UTILITY.xls* represent the electricity cost in dollars during the month of July 2002 fora random sample of 50 two-bedroom apartments in a large city. Decide whether the data appear to be approximately normally distributed by constructing a normal probability plot.

```r
UTILITY <- readxl::read_xls("UTILITY.xls")
qqnorm(UTILITY$`Utility Charge`)
qqline(UTILITY$`Utility Charge`)
```



- The electricity cost seems to follow a normal distribution reasonably well, except in the tails which deviates from normality very lightly. Overall, the sample quantile seems to match the theoretical quantile pretty closely.

6. You are trying to set up a portfolio that consists of a corporate bond fund and a common stock fund. The following information about the annual return (per $1,000) of each of these investments under different economic conditions is avaliable along with the probability that each of these economic conditions will occur.

```
probability <- c(0.1, 0.2, 0.3, 0.25, 0.15)
corporate_bonds <- c(-40, 60, 80, 105, 100)
common_stocks <- c(-120, -30, 115, 170, 230)

investments <- data.frame(probability, corporate_bonds, common_stocks)

w <- 0.4
corporate_bonds_expected <- sum(apply(investments, 1, function(x) x[1]*x[2]))
common_stocks_expected <- sum(apply(investments, 1, function(x) x[1]*x[3]))
corporate_bonds_var <- sum(apply(investments, 1, function(x) x[1]*(x[2]-corporate_bonds_expected)^2))
common_stocks_var <- sum(apply(investments, 1, function(x) x[1]*(x[3]-common_stocks_expected)^2))
corporate_common_covar <- sum(apply(investments, 1, function(x){
        (x[2]-corporate_bonds_expected) * (x[3]-common_stocks_expected) * x[1]
        }))

expected_return <- w * corporate_bonds_expected + (1-w) * common_stocks_expected
risk <- sqrt(
        w^2 * corporate_bonds_var +
        (1-w)^2 * common_stocks_var +
        2*w*(1-w)*corporate_common_covar
)
portfolio  <- c(expected_return, risk)
names(portfolio) <- c("expected return (per $1,000)", "risk (per $1,000)")
portfolio
```

```
## expected return (per $1,000)            risk (per $1,000)
##                     85.40000                     80.53161
```

7. A set of final examination grades in an introductory statistics course was found to be normally distributed with a mean of 73 and a standard deviation of 8.

(a) What percentage of students scored between 65 and 89?

```
a <- pnorm(q = 89, mean = 73, sd = 8) - pnorm(q = 65, mean = 73, sd = 8)
a <- percent(a)
a
```

```
## [1] "81.9%"
```

(b) Only 5% of the students taking the test scored higher than what grade?

```
b <- qnorm(p = .05, mean = 73, sd = 8, lower.tail = FALSE)
b
```

```
## [1] 86.15883
```

(c) If the professor grades on a curve (gives As to the top 10% of the class regardless of the score), are you better off with a grade of 81 on this exam or a grade of 68 on a different exam

where the mean is 62 and the standard deviation is 3? Show your answer statistically and explain.

```
exam1 <- pnorm(q = 81, mean = 73, sd = 8, lower.tail = FALSE)
exam2 <- pnorm(q = 68, mean = 62, sd = 3, lower.tail = FALSE)
q7c <- c(exam1, exam2)
names(q7c) <- c("exam1", "exam2")
q7c
```

```
##      exam1      exam2
## 0.15865525 0.02275013
```

- The latter exam. With a grade of 68, you will be in the top 2.28%; whereas in the first exam, a grade of 81 can only get you in the top 15.9%