CFRM 461: Probability and Statistics for Computational Finance Homework 1

Instructions: Please round your answers to 2 decimal places unless stated otherwise. Since R codes in this homework are simple, please paste them directly in your write-up. There is NO need to upload separate R files for this homework.

1. A Seattle hospital is concerned about the rising number of low birth weight of babies. They then collected data on the mother's age, the mother's pre-pregnancy weight, the level of pre-natal care (none, minimal, adequate), and whether the mother used drugs during pregnancy (cigarettes, alcohol, etc).

   Note: Please identify the type (categorical or quantitative) for the four variables in the study: "age", "pre-pregnancy weight", "level of prenatal care", and "drug use".

   Mothers Age: Quantitative
   Pre-pregnancy weight: Quantitative
   Pre-natal care: Categorical
   Drugs: Categorical

2. The results of a survey of automobiles parked in student and staff lots at a large university are listed in the table below.

	Student	Staff
North American	91	90
European	31	16
Asian	68	54

Table 1: Two-Way Table

   We name the two variables in Table 5 by "Driver" and "Origin" (both are categorical variables). The "Driver" variable has two categories: "Student" and "Staff", while the "Origin" variable has three categories: "North American", "European", and "Asian".

   Hints: A distribution gives the probabilities (relative frequency) for all possible outcomes. Please report the probabilities in percentage with two decimal places accuracy (e.g., 12.34%).

	Student	Staff	Total
North American	91	90	181
European	31	16	47
Asian	68	54	122
Total	190	160	350

Table 2: Contingency Table

(a) What is the marginal distribution of "Origin"?

Origin	Frequency	Relative Frequency (%)
North American	181	51.72
European	47	13.43
Asian	122	34.86

Table 3: Marginal Distribution

(b) What is the marginal distribution of "Driver"?

Driver	Student	Staff
Frequency	190	160
Relative Frequency (%)	54.29	45.71

Table 4: Marginal Distribution

(c) What is the conditional distribution of "Origin" for "Staff"?

Origin	Frequency	Relative Frequency (%)
North American	90	56.25
European	16	10
Asian	54	33.75

Table 5: Marginal Distribution

(d) What is the probability (relative frequency) that the driver of a car is a student given that he/she is a North American?

$$\frac{91}{181} \implies 50.27\%$$

3. In this `R` programming question, please paste the codes in your write-up, or alternatively, you may use R-markdown to prepare your submission.

(a) Create a 3×2 matrix, named by `M`, to record all the entries in Table 1, i.e.,

$$M = \begin{pmatrix} 91 & 90 \\ 31 & 16 \\ 68 & 54 \end{pmatrix}$$

To receive full credits, you need to provide at least two different methods.

```
mat <- matrix(c(91,90,31,16,68,54),3,2,TRUE)
```

Output :

```
91  90
31  16
68  54
```

```
> x <- c(91,31,68)
> y <- c(90,16,54)
> cbind(x,y)
```

Output :

```
91  90
31  16
68  54
```

(b) Find the row sums and the column sums of matrix `M`.

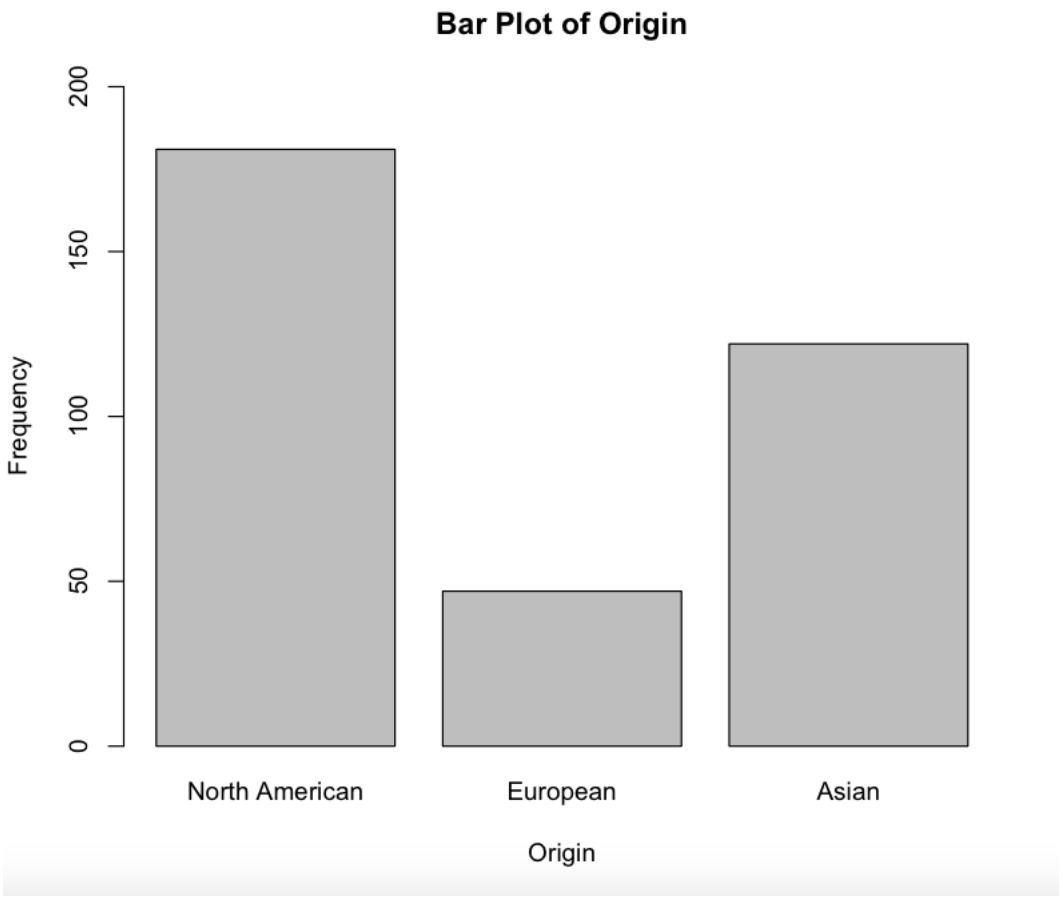
```
rowSums(mat)
colSums(mat)
```

Output : 181 47 122
Output : 190 160

(c) Create a bar plot for the variable "Origin". In the plot, please (1) name the title of the plot as "Bar Plot of Origin", (2) name all three categories (North American, European, Asian), (3) use frequency (count) as the y-axis and choose [0, 200] as the range of the y-axis.

In the `barplot()` command, please specify `main`, `names.arg`, `ylim` as above. Please use `width = 1` when drawing the barplot. Please read the help document on `barplot` to understand the meanings of those arguments.

```
origin <- c(181,47,122)
barplot(origin,main = "Bar Plot of Origin", names.arg =
c("North American","European","Asian"),ylim = c(0,200),ylab =
"Frequency",xlab = "Origin", width = 1)
```



4. The prices of walking shoes in $ are listed below:

$$90, 70, 70, 70, 75, 70, 65, 68, 60, 74, 70, 95, 75, 68, 85, 65$$

(a) We use the build-in function `hist()` of `R` to construct a histogram for the given data set through the following steps:

(a) Create a vector to record all the observations

(b) Create a vector for the breakpoints between histogram cells: use 60, 65, 70, 75, 80, 85, 90, 95, 100, i.e., the first bin is [60, 65], the second is (65, 70], etc.

(c) Use frequency as the y-axis, and pick the range of [0, 10] for the y-axis

(d) Label the x-axis by "Price", and pick the range of [60, 100] for the x-axis

(e) Name the title of the histogram as "Histogram of Price"

(f) Choose `TRUE` for plot option

Please paste the codes and the final histogram in your write-up.

Hints: You need to specify `breaks`, `freq`, `main`, `xlim`, `ylim`, `xlab`, `plot` in the `hist()` command to obtain the desired histogram.

```
counts <- c(90, 70, 70, 70, 75, 70, 65, 68, 60, 74, 70, 95,
75,68, 85, 65)
brk <- c(60,65,70,75,80,85,90,95,100)
```

(b) Identify the shape of the distribution of the shoes price, i.e., identify three features of a distribution: mode (unimodal, bimodal, etc), symmetry (symmetric, skewed to the left/right), and outliers (yes or no).

It is a unimodal distribution that is skewed to the right and contains outliers.

(c) Compute the mean and the median of the given data set, and compare them to confirm your conclusion on the symmetry feature of the distribution in (b).

```
> mean(counts)
[1] 73.125
> median(counts)
[1] 70
```

Clearly the mean ¿ median so the distribution is positively skewed

(d) Calculate the range, the interquantile range (IQR), and the standard deviation of the given data set.

```
> range(counts)
[1] 60 95
counts <- sort(counts)
> IQR <- median(counts[9:16]) - median(counts[1:8])
> IQR
[1] 7
```

```
> sd(counts)
[1] 9.372833
```

Remark: You may use `R` to calculate Q4 (c) and (d), and if you choose to do so, please paste the codes in your write-up as well. But please be aware that in the exams you will need to compute those by hands.

5. (a) The volumes of soda in 1 litre cola bottle can be described by a Normal model with a mean of 0.95 L and a standard deviation of 0.04 L. What percentage of bottles can we expect to have a volume less than 0.94 L?

$$z = \frac{y - \bar{y}}{s}$$
$$\implies \frac{0.94 - 0.95}{0.04} = -0.25$$
$$\implies 0.4013$$

Therefore, 40.13 % of bottles will have a volume less than 0.94 L

(b) A bank's loan officer rates applicants for credit. The ratings can be described by a Normal model with a mean of 200 and a standard deviation of 50. If an applicant is randomly selected, what percentage can be expected to be between 200 and 275?

$$\frac{275 - 200}{50} = 1.5 \implies 0.9332$$

$$\implies 0.9332 - 0.5 = .4332$$

Therefore, 43.32% will be between 200 and 275

(c) The lengths of human pregnancies can be described by a Normal model with a mean of 268 days and a standard deviation of 15 days. What percentage can we expect for a pregnancy that will last at least 300 days?

$$\frac{300 - 268}{15} = 2.13 \implies 0.9834$$
$$\implies 0.0166$$

Therefore 1.66 % of the pregnancies will last at least 300 days.

Remarks: To receive full credits, you must provide formulas and use the Normal Distribution table (see the "Module" section in Canvas) to look for probabilities. Keep **4 decimal places** for the final results, e.g., 0.1234. If you use `R` or other software to solve Q5, only partial credits will be awarded!