Data Analysis and Statistical Methods
Frans Simanjuntak - S3038971

1. **Problem 1**

A. For this question I will use X^2-test (Chi Square test)

Given:

$\alpha = 0.05$

H_0 = The probability of exceptionally lucky

H_1 = The probability of not lucky

Critical value: $X^2_{1-\alpha,df}$

Rejection Region: Reject H_0 if $X^2 > X^2_{1-\alpha,df}$

B. On paper test

Given:

$\alpha = 0.05$

k = 5 (category)

df = k-1 = 5-1 = 4

$X^2_{0.95,4}$ = **9.488**

Following the formula:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - E_i)^2}{E_i}$$

We can calculate the X^2

Correct Number	**Freq (x)**	**Probability(p)**	**n*pi**	**x - (n*pi)^2**	**e(x) = x - (n*pi)^2 / (n*pi)**
0	71	0.401	74.185	10.144225	0.136742266
1	74	0.424	78.44	19.7136	0.2513207547
2	31	0.152	28.12	8.2944	0.2949644381
3	8	0.022	4.07	15.4449	3.794815725
4	1	0.001	0.185	0.664225	3.590405405
Total (n)	185			$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - E_i)^2}{E_i}$	8.068248589

From the calculation we obtain the value of X^2 = **8.068248589**

Since $X^2 \leq X^2_{1-\alpha}$ **therefore we can not reject the null hypothesis**.

C. Test in R

R-Code

```
library(tidyverse)

frequency <- c(71,74,31,8,1)
probability <- c(0.401,0.424,0.152,0.022,0.001)

n <- sum(frequency, na.rm=TRUE)


expected_value <- ((frequency - n*probability)^2)/(n*probability)
expected_value

#Test the chisquare
chisq.test(frequency, p=probability)

#Get the chisquare value from the table
qchisq(0.95,4)
```

Results:

```
        Chi-squared test for given probabilities

data:  frequency
X-squared = 8.0682, df = 4, p-value = 0.08911

Warning message:
In chisq.test(frequency, p = probability) :
  Chi-squared approximation may be incorrect
>
> #Get the chisquare value from the table
> qchisq(0.95,4)
[1] 9.487729
```

The value of X^2 is 8.0682, the degree of freedom is 4, the p-value is 0.08911 and the critical value of X^2 is 9.487729. Since the p value > 0.05 therefore **we can not reject the null hypothesis**.

2. **Problem 2**

Given :
dataset = data-HW_2.csv
Category of number of years of education (k) = 0-5 yrs, 6-11 yrs, 12+ yrs
k = 3
Category of number of times the person has been unemployed (l) = 0, 1,2
l= 3

A. *In R, make the data into table.*

R-Code:

```
library(tidyverse)
library(dplyr)

setwd("X:/My Desktop/Statistic/assignment 4")
#load the dataset into data frame
df_hw <- read.csv("Data_HW_2.csv", header = TRUE)

#convert data frame into table
dt = table(df_hw$education, df_hw$jobless)

#rename the column names and row names
colnames(dt) <- c("unemployment_0","unemployment_1", "unemployment_2")
rownames(dt) <- c("education_0_to_5","education_12_and_over", "education_6_to_11")

dt
```

The Result:

	unemployment_0	unemployment_1	unemployment_2
education_0_to_5	4	2	6
education_12_and_over	61	39	16
education_6_to_11	78	27	15

B. ***By hand: test if "education" and "unemployment" are independent at the 95% confidence level.***

Given:

$\alpha = 0.05$

H_0 = The variables are independent

H_1 = The variables are dependent

df = (k-1)*(l-1) = 4

Critical value: $X^2_{1-\alpha,df}$ = **9.488**

Rejection Region: Reject H_0 if $X^2 > X^2_{1-\alpha,df}$

Given the **Observed Value**

		Unemployment		
Education	0	1	2	Total
0-5 years	4	2	6	**12**
12+years	61	39	16	**116**
6-11 years	78	27	15	**120**
Total	**143**	**68**	**37**	**248**

Step 1. Convert to fraction of probabilities

		Unemployment		
Education				Total
0-5 years				**0.04838709677**
12+years				**0.4677419355**
6-11 years				**0.4838709677**
Total	**0.5766129032**	**0.2741935484**	**0.1491935484**	

Step 2. Fill in the table assuming independence

		Unemployment		
Education				Total
0-5 years	0.02790062435	0.01326742976	0.007219042664	**0.04838709677**
12+years	0.2697060354	0.128251821	0.06978407908	**0.4677419355**
6-11 years	0.2790062435	0.1326742976	0.07219042664	**0.4838709677**
Total	**0.5766129032**	**0.2741935484**	**0.1491935484**	**1**

Step 3. Multiply with n total (248) again in order to get the expected Value

		Unemployment		
Education				Total
0-5 years	6.919354839	3.290322581	1.790322581	**12**
12+years	66.88709677	31.80645161	17.30645161	**116**
6-11 years	69.19354839	32.90322581	17.90322581	**120**
Total	**143**	**68**	**37**	

Step 4. Calculate Test Statistic

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

Observed Value

	Unemployment			
Education	0	1	2	**Total**
0-5 years	4	2	6	**12**
12+years	61	39	16	**116**
6-11 years	78	27	15	**120**
Total	**143**	**68**	**37**	**248**

Expected Value

Education	**Unemployment**			**Total**
0-5 years	6.919354839	3.290322581	1.790322581	**12**
12+years	66.88709677	31.80645161	17.30645161	**116**
6-11 years	69.19354839	32.90322581	17.90322581	**120**
Total	**143**	**68**	**37**	

X^2 = 1.231709151+ 0.5060088552 + 9.898430689 + 0.5181553708 + 1.626938428 + 0.09862309473 + 1.120821114 + 1.059108159 + 0.470793374

X^2 = 16.53058824

Since $X^2 > X^2_{1-\alpha}$ therefore **we must reject null hypothesis**. It's most likely the variables are dependent event.

C. ***In R: What commands do you use in R to do the test?***

R-Code

```
library(tidyverse)
library(dplyr)

setwd("X:/My Desktop/Statistic/assignment 4")
#load the dataset into data frame
df_hw <- read.csv("Data_HW_2.csv", header = TRUE)

#convert data frame into table
dt = table(df_hw$education, df_hw$jobless)

#rename the column names and row names
colnames(dt) <- c("unemployment_0","unemployment_1", "unemployment_2")
rownames(dt) <- c("education_0_to_5","education_12_and_over", "education_6_to_11")

dt

# calculate sums over the rows and columns
education_level = rowSums(dt) / sum(dt)
education_level

# calculate sums over the columns and columns
unemployment_level = colSums(dt) / sum(dt)
unemployment_level

#calculate the probabilities
probabilities = education_level %*% t(unemployment_level)
# add back in the row names and column names
colnames(probabilities) = c("unemployment_0","unemployment_1", "unemployment_2")
rownames(probabilities) = c("education_0_to_5","education_12_and_over",
"education_6_to_11")
probabilities

#calculate expected matrix
expected_matrix = probabilities * sum(dt)
expected_matrix

#perform chisquare test
chisq.test(dt, correct=FALSE)
```

```
#Get the chisquare value from the table
qchisq(0.95,4)
```

The result:

```
        Pearson's Chi-squared test

data:  dt
X-squared = 16.531, df = 4, p-value = 0.002384

Warning message:
In chisq.test(dt, correct = FALSE) :
  Chi-squared approximation may be incorrect
>
> #Get the chisquare value from the table
> qchisq(0.95,4)
[1] 9.487729
```

Since the p-value < 0.05 therefore we must reject the null hypothesis.

What would be the conclusions from the test, if you only tested the 2 higher education levels (omit 0-5 years)

R-Code

```
library(tidyverse)
library(dplyr)

setwd("X:/My Desktop/Statistic/assignment 4")
#load the dataset into data frame
df_hw <- read.csv("Data_HW_2.csv", header = TRUE)

#convert data frame into table
dt <- table(df_hw$education, df_hw$jobless)
#rename the column names and row names
colnames(dt) <- c("unemployment_0","unemployment_1", "unemployment_2")
rownames(dt) <- c("education_0_to_5","education_12_and_over", "education_6_to_11")
dt
#remove the first row [education_0_to_5] from the table
dt <- dt[-1,]

# calculate sums over the rows and columns
education_level = rowSums(dt) / sum(dt)
education_level
```

```
# calculate sums over the columns and columns
unemployment_level = colSums(dt) / sum(dt)
unemployment_level

#calculate the probabilities
probabilities = education_level %*% t(unemployment_level)
# add back in the row names and column names
colnames(probabilities) = c("unemployment_0","unemployment_1", "unemployment_2")
rownames(probabilities) = c("education_12_and_over", "education_6_to_11")
probabilities

#calculate expected matrix
expected_matrix = probabilities * sum(dt)
expected_matrix

#perform chisquare test
chisq.test(dt, correct=FALSE)

#Get the chisquare value from the table
qchisq(0.95,2)
```

The Result:

```
        Pearson's Chi-squared test

data:  dt
X-squared = 4.2266, df = 2, p-value = 0.1208

>
> #Get the chisquare value from the table
> qchisq(0.95,2)
[1] 5.991465
```

The conclusion

- We can't reject the null hypothesis since the p-value > 0.05 therefore we can say that the variables are likely to be independent.

What is your interpretation of the outcome?

- The X-squared is the X^2 test
- The df is the degree of freedom
- The p-value is the probability of observing X^2 test, assuming the null hypothesis is true

From these two test, we can draw a conclusion that the more we add category to our variables, the more the variables become dependent and vice versa.

3. **Problem 3**

   Given

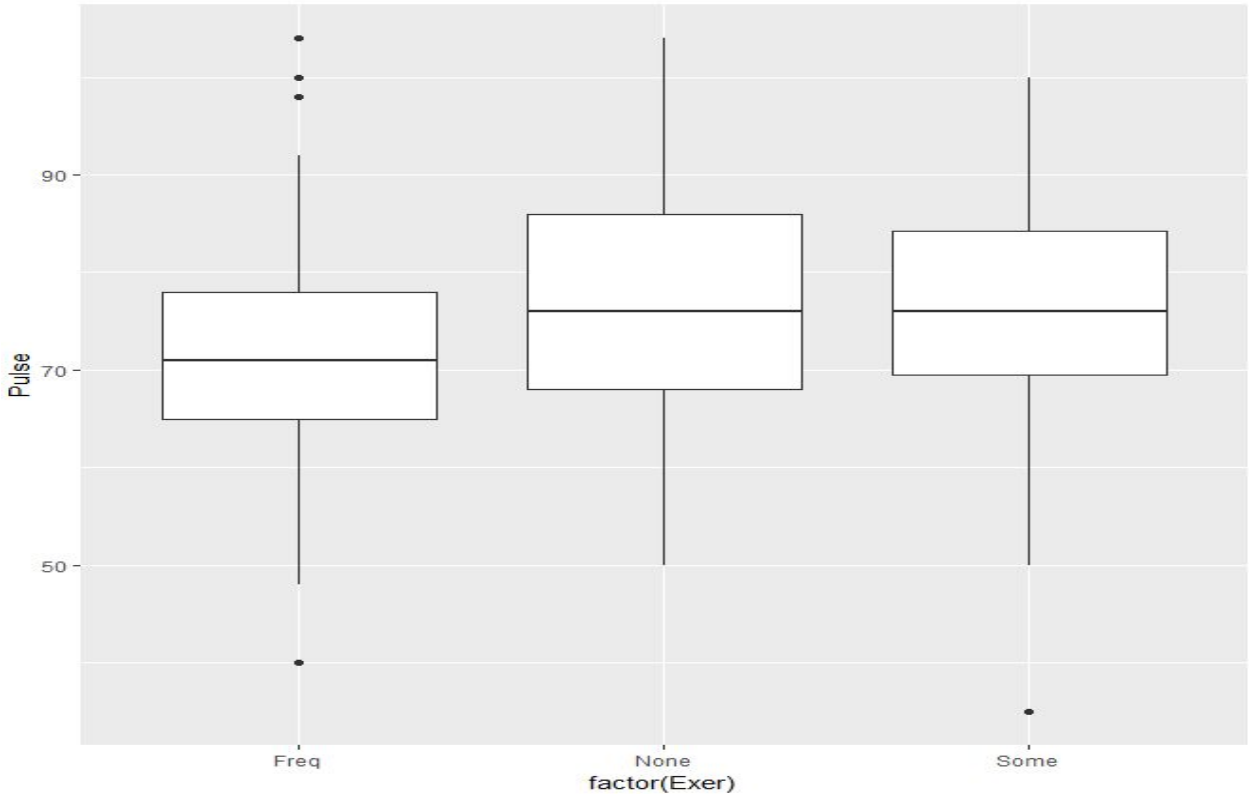
   Dataset = survey.csv

A. Make a box plot of pulse grouped by exercise behavior

R-Code

```
library(dplyr)
library(tidyverse)
library(ggplot2)

setwd("X:/My Desktop/Statistic/assignment 4")
#load the dataset into data frame
df_students <- read.csv("survey(2).csv", header = TRUE)

#Select Exercise and Pulse and Remove row if pulse not given
df <- df_students %>% select(Exer, Pulse) %>% filter(Pulse != "NA")
count(df)
#plot the pulse grouped by exercise berhaviour
ggplot(df, aes(x=factor(Exer), y=Pulse)) + geom_boxplot()
```

The Result



B. Do a hypothesis test if the means for each group are different at the 95% confidence level.

Which test do you choose?

- I will choose ***One way ANOVA test*** because we want to determine whether there are any statistically significant difference between the means of three or more independent (unrelated) groups.
- The one way anova test compares the means between the groups we are interested in and determines whether any of those means are statistically significantly different from each other.

Hypothesis Testing:

H_0 : $\mu 1 = \mu 2 = \mu 3$

H_1 : Not all μ are equal

$\alpha = 0.05$

Rejection Region: Reject H_0 if p-value <= α

Assume : Equal Variance

Test in R:

```
library(dplyr)
library(tidyverse)
library(ggplot2)

setwd("X:/My Desktop/Statistic/assignment 4")
#load the dataset into data frame
df_students <- read.csv("survey(2).csv", header = TRUE)

#Select Exercise and Pulse and Remove row if pulse not given
df <- df_students %>% select(Exer, Pulse) %>% filter(Pulse != "NA")
count(df)
#plot the pulse grouped by exercise behaviour
ggplot(df, aes(x=factor(Exer), y=Pulse)) + geom_boxplot()

# perform the 1-way ANOVA test
oneway.test(Pulse~Exer, df, var.equal = TRUE)
```

Results:

```
        One-way analysis of means

data:  Pulse and Exer
F = 3.3783, num df = 2, denom df = 189, p-value = 0.03618
```

Conclusion:
Since p-value <= 0.05 therefore we reject H_0

C. Explain each item of the output you get from R

- **F** is the value of one way anova test (3.3783)
- **num df** is the degree of freedom (2)
- **denom df** is the number of degree of freedom that the estimate of variance used in the numerator is based on (189)
- **p-value** is the probability of observing anova test, assuming the null hypothesis is true (0.03618)

How they are calculated?

- **k** = the number of groups
- **N** = the total number of subjects in the experiment

- **Degree of freedom**
  The variation due to the interaction between the samples is denoted SS(B) for Sum of Squares Between groups. If the sample means are close to each other this will be small. There are k samples involved with one data value for each sample (the sample mean), so there are k-1 degrees of freedom.
  df = k-1

- **Denom Degree of Freedom**
  The variation due to differences within individual samples, denoted SS(W) for Sum of Squares Within groups. Each sample is considered independently, no interaction between samples is involved. The degrees of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size:
  denom df = N - k.

- **F**
  The F test statistic is found by dividing the between group variance by the within group variance. The degrees of freedom for the numerator are the degrees of freedom for the

between group (k-1) and the degrees of freedom for the denominator are the degrees of freedom for the within group (N-k).

$$F = \frac{s_b^2}{s_w^2}$$

- **P-value**

  P-value is the probability of observing anova test, assuming the null hypothesis is true. It can only be obtained once we have degree of freedom of numerator, denominator of degree of freedom, and F statistic result. We can use R to calculate this.

4. Problem 4

Given:

v = 4.0 ± 0.2 m/s

t = 0.60 ± 0.06 s

$y = v*t - ½ * g * t^2$

a. Calculate y and the uncertainty of y. Assume g = 9.80 m/s² (no uncertainty in g)

Let:

p = v*t = 4.0 *0.60 = 2.4 m

q = ½ * g *t² = ½ *9.80*0.60² = 1.764 m

Then :

y = p - q

y = 2.4 - 1.764

y = 0.636 m

The uncertainty of y or δy are calculated as follows:

First lets calculate the uncertainty of p and q.

- The uncertainty of p or δp can be obtained using **multiplication rule**

$$\frac{\delta p}{|p|} = \sqrt{(\frac{\delta v}{v})^2 + (\frac{\delta t}{t})^2}$$

$$\frac{\delta p}{|p|} = \sqrt{(\frac{0.2}{4.0})^2 + (\frac{0.06}{0.60})^2}$$

$$\frac{\delta p}{|p|} = 0.1118$$

$$\delta p = |p| * 0.1118$$

$$\delta p = 2.4 * 0.1118$$

$$\delta p = 0.26832$$

$$\delta p = \mathbf{0.27}$$

- The uncertainty of q or (δq) can be obtained using the rules of **uncertainty in a power.**

$Q = t^n$

$\frac{\delta q}{|q|} = |n| \frac{\delta t}{|t|}$

$\frac{\delta q}{|q|} = 2 \left(\frac{0.06}{0.6} \right)$

$\frac{\delta q}{|q|} = 2 * 0.10$

$\frac{\delta q}{|q|} = 0.20$

$\delta q = |q| * 0.20$

$\delta q = 1.764 * 0.20$

$\delta q = 0.3528$

δq = **0.35 m**

Since the formula is y = p - q therefore, we should calculate δy using subtraction rule.

$\delta y = \sqrt{(\delta p)^2 + (\delta q)^2}$

$= \sqrt{(0.27)^2 + (0.35)^2}$

= **0.4420 m**

b. Round the result and uncertainty to the significant digits.

Thus, the reported value of y is **0.636 ± 0.4420 m**.

The round up to the significant digits is : **0.6 ± 0.4 m.**

Explain what significant digits are.

The significant digits are those digits that are certain and the first uncertain digit. The number of significant figures is dependent upon the uncertainty of the measurement or process of establishing a given reported value.

In this case, the y value and its uncertainty value (δy) contain two significant digits, 0.6 and 0.4 respectively. These two significant digits are classified as absolute uncertainty. The absolute uncertainty is the number which, when combined with a reported value, gives the range of true values. The reported value is 0.6 m and the absolute uncertainty is 0.4 m therefore the range of true values is **0.2 m to 1.0 m**.

5. **Problem 5**

Given

$$x = \Gamma(u) = \int_0^\infty t^{u-1} e^{-t} dt$$

A. Calculate x and s_x for u = 0.8 ± 0.1 (1s) using a Monte Carlo simulation with 10000 repetitions.

R-Code

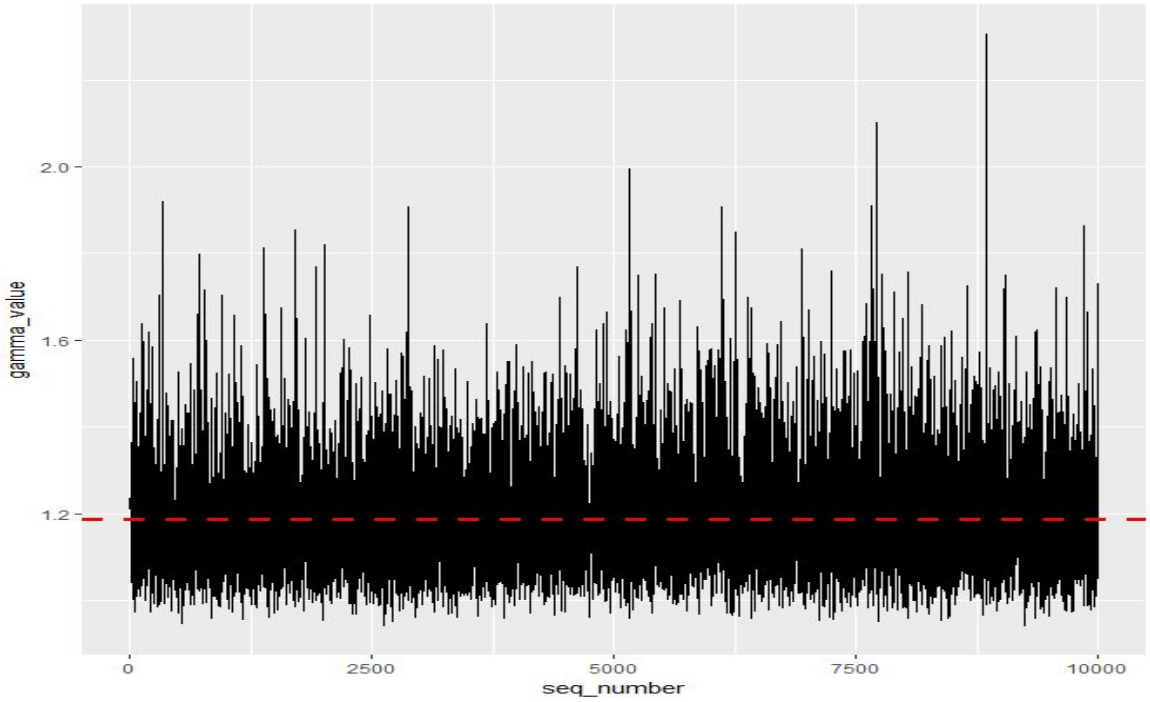
```
library(tidyverse)
library(ggplot2)
library(dplyr)

n <- 10000
u <- rnorm(n, mean = 0.8, sd = 0.1)
x<- gamma(u)

seq_number = seq (1, n, 1)

df <- data.frame(seq_number = seq_number, gamma_value = x)
df

#mean
x_bar <- mean(x)
x_bar

#median
x_median <- median(x)
x_median

#standard deviation
x_sd <- sd(x)
x_sd


#plot df
ggplot()+
  geom_line(data=df, aes(x=seq_number, y= gamma_value))+
  geom_hline(yintercept = x_bar, linetype="dashed", lwd= 1.2, color="red")+
  geom_hline(yintercept = x_median, linetype="dashed", lwd= 1.2, color="blue")
```

Compare the approaches of using the mean and the median for estimating x

The approach using mean:

```
> #mean
> x_bar <- mean(x)
> x_bar
[1] 1.185037
```

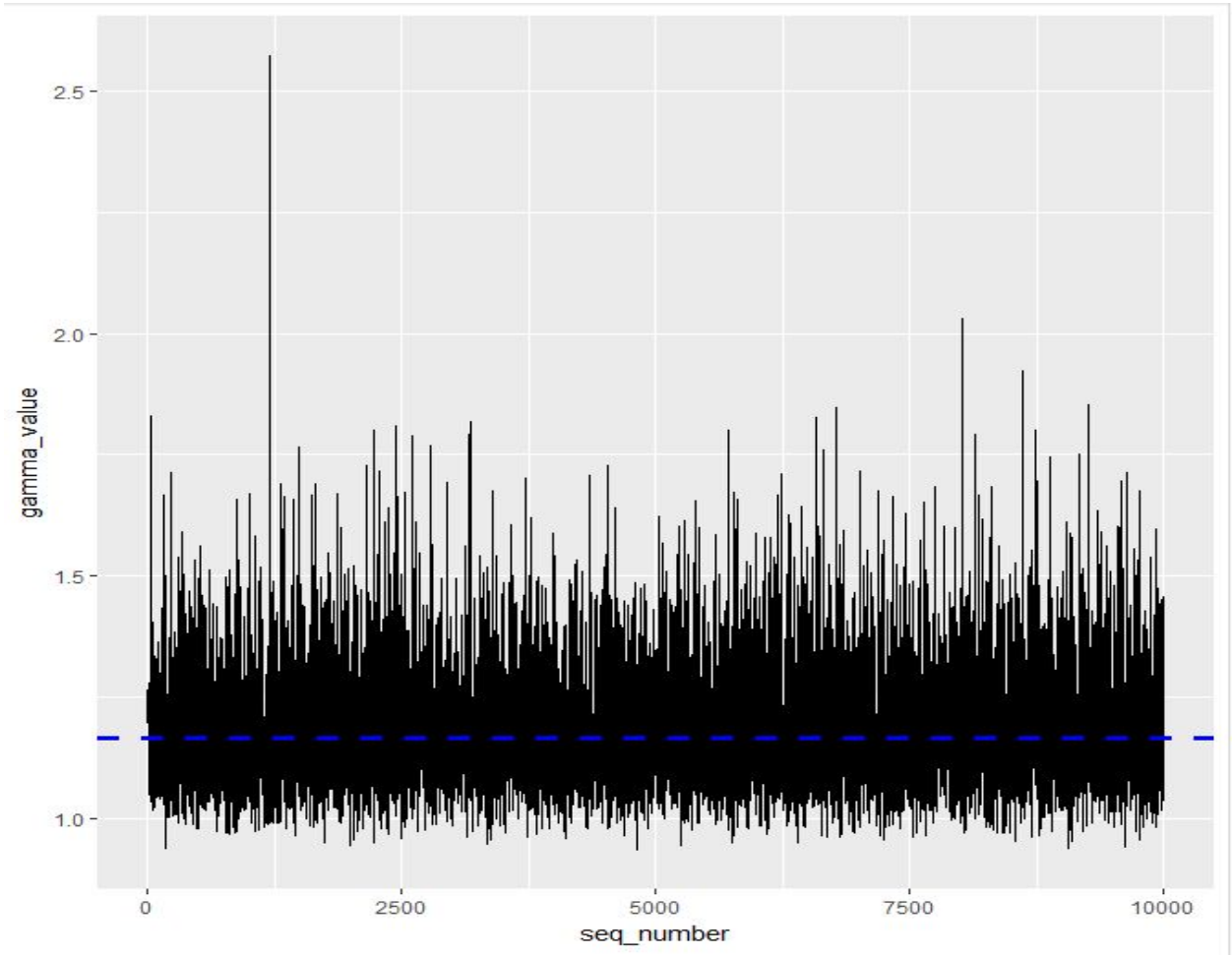
In the below plot, the mean value was indicated by the ***horizontal red line***.



The approach using median:

```
> #median
> x_median <- median(x)
> x_median
[1] 1.165116
```

In the below plot, the median value was indicated by the ***horizontal blue line***.



The standard deviation:

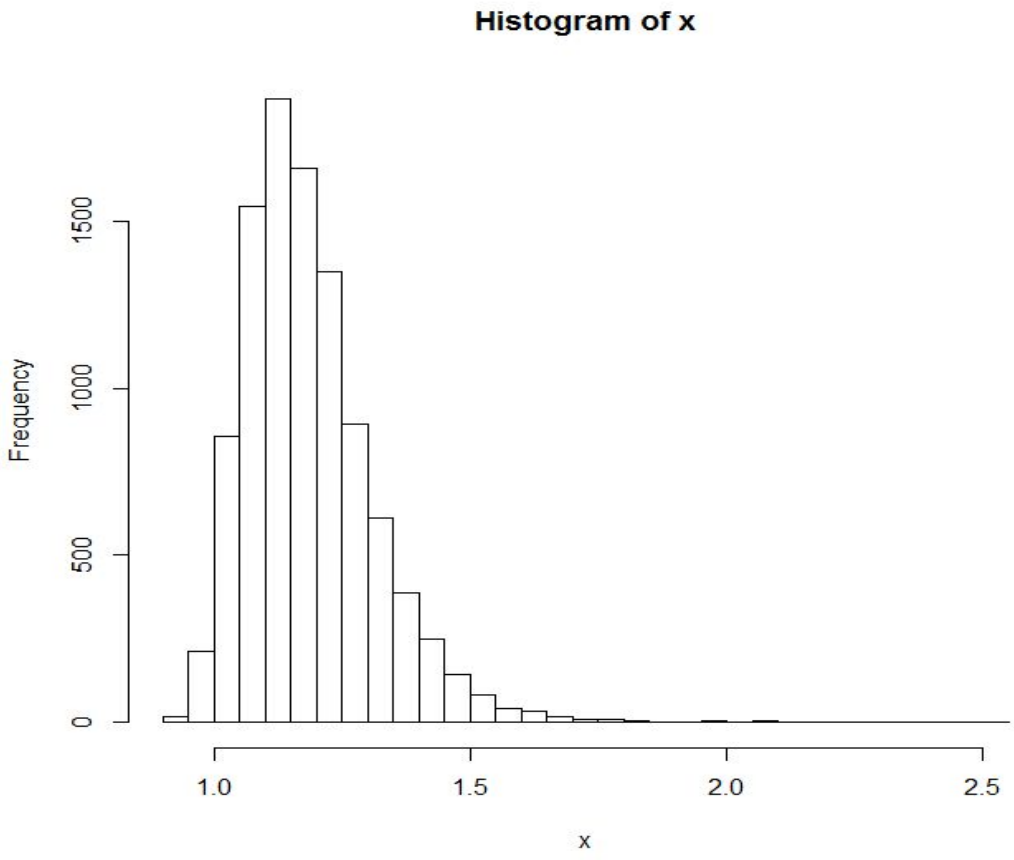
```
> x_sd
[1] 0.1240749
```

From both results, we can see that the mean value is 0.019709 higher above the median value.

B. Plot a histogram of x using 50 bars.

R-Code

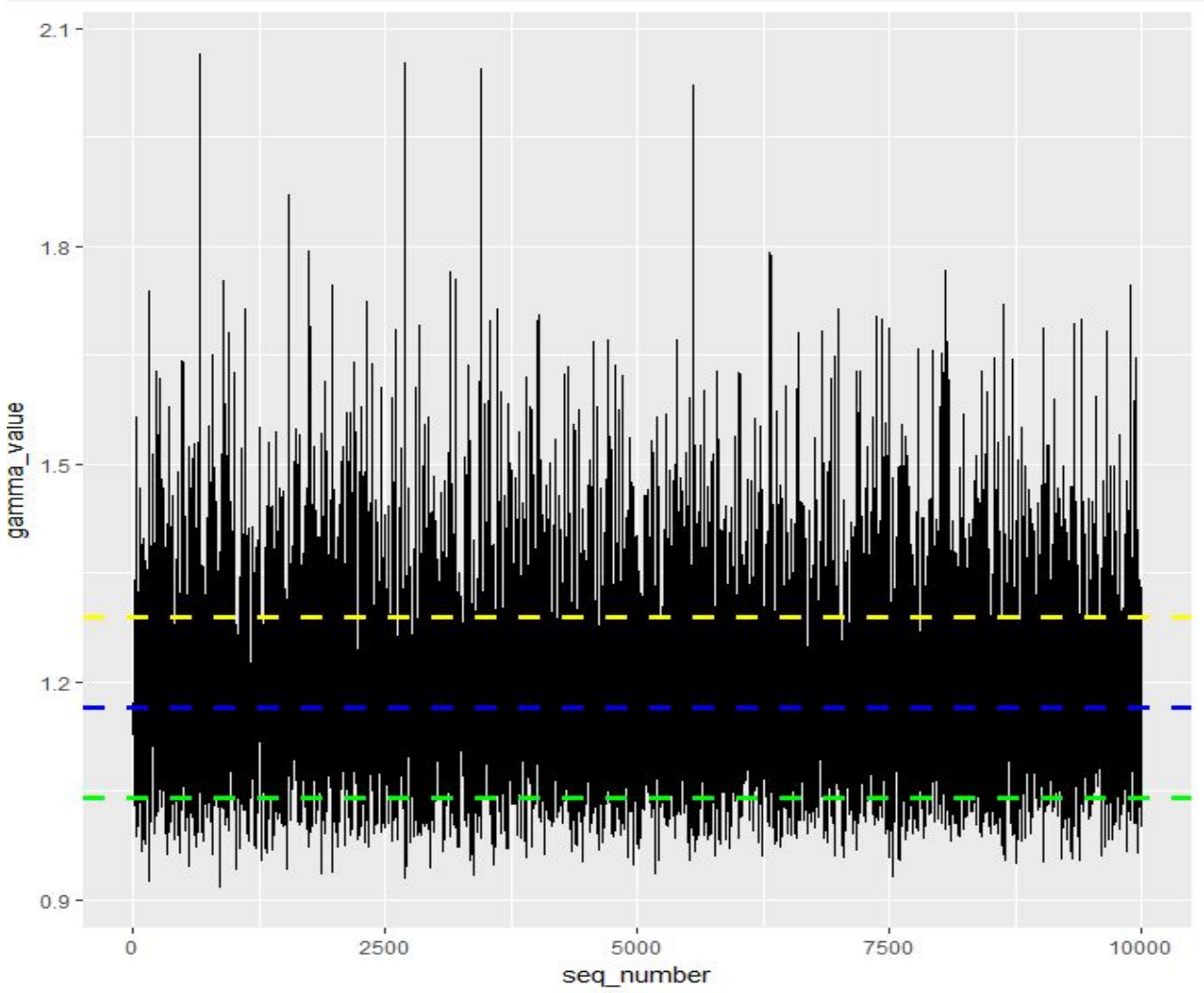
```
library(tidyverse)
library(ggplot2)
library(dplyr)

n <- 10000
u <- rnorm(n, mean = 0.8, sd = 0.1)
x<- gamma(u)

seq_number = seq (1, n, 1)

df <- data.frame(seq_number = seq_number, gamma_value = x)

#plot histogram using 50 bars
hist(x, breaks=50)
```



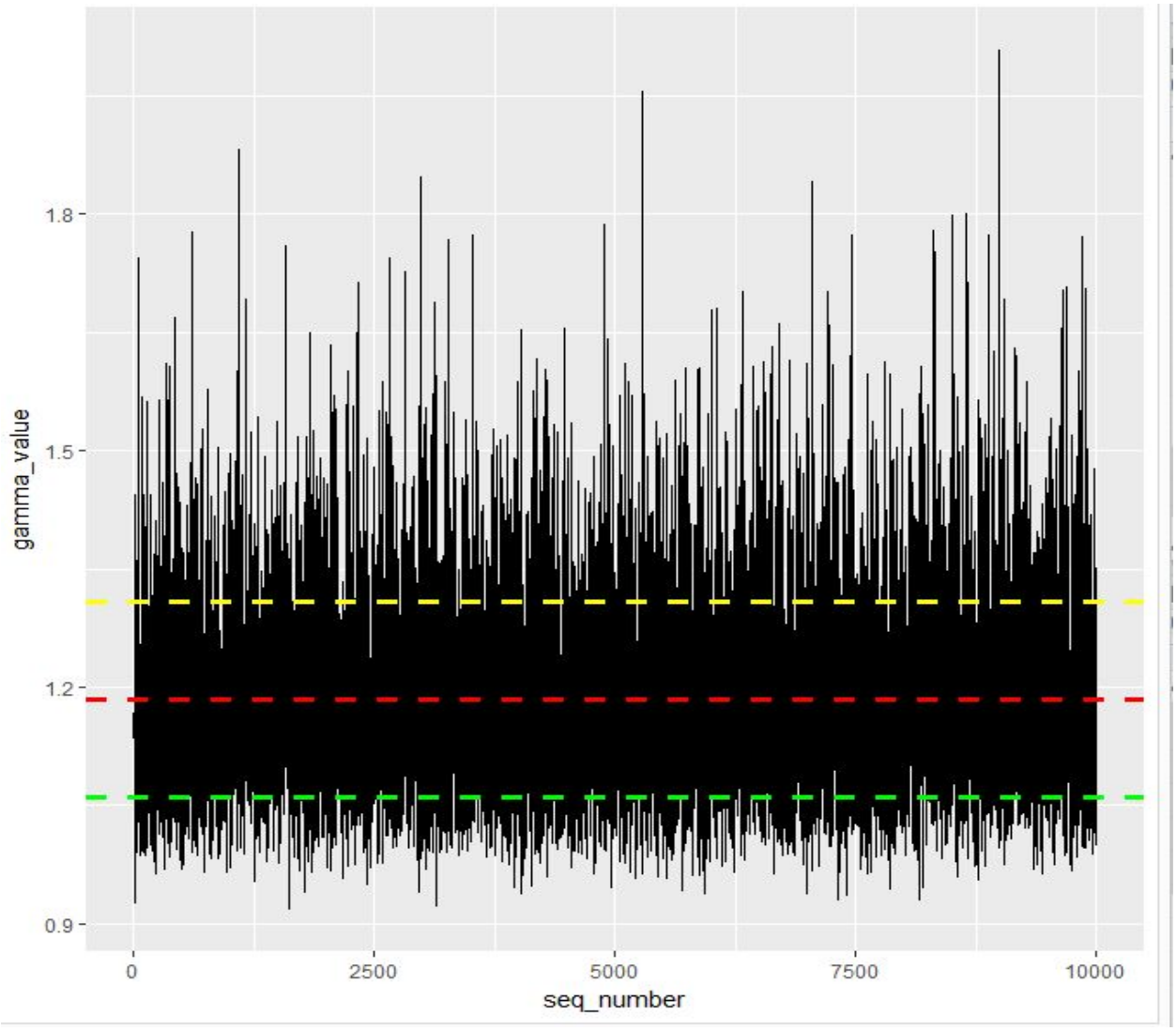
Discuss if the simple ± interval is appropriate and if you would choose the median or mean to represent x

Given
Mean = 1.185001
Median = 1.165292
Standard deviation = 0.1240749

If we we observe using **median** and apply ± interval (0.1240749), then the estimation range would be in between 1.0412171 and 1,2893669 as shown in below figure. The horizontal blue line indicates the median value, the yellow line indicates the value of media+standard deviation, and the green line indicates the value of media-standard deviation.



Conversely, if we observe using the mean value and apply ± interval (0.1240749), the value ranges would be in between 1.0609252 and 1.3090759 as shown in below figure. The horizontal red line indicates the median value, the yellow line indicates the value of media+standard deviation, and the green line indicates the value of media-standard deviation.



Based on our observation, it can be easily noticed that the ± interval is really helpful to determine which one is the best indicator to represent x. In my opinion, **mean value** is the best choice to represent x since after applying ± interval, the values are relatively close to the x values compared to the approach using median.