MATH2406 Applied Analytics

Assessment 1: Report

Thaddeus Lee, S3933533

Install and load the packages you need to produce the report here:

```
library(dplyr) # Useful for data manipulation
library(ggplot2) # Useful for building data visualisations
library(knitr) # Useful for creating nice tables
library(magrittr)
library(LearningStats)
library(car)
library(MASS)
library(VGAM)
library(pwr)
library(tinytex)
```

Question 1

a) Generate 100 data values at random from a uniform distribution on the unit interval from zero to one.
Referring to code below, we use *runif* function to randomly create 100 observations of continuous values between zero and one. The data is assigned to **data_values**

b) Produce a histogram of your data. Discuss its shape. Compare and contrast your histogram with the uniform density. What conclusions can you draw from this comparison?
In a uniform distribution, all outcomes between the set min and max are equally likely. Compared with the uniform density, the histogram generated in the code below, is relatively flat overall, but of the 100 samples that were randomly generated, we can observe lower frequencies in the 0.2 range, and higher frequencies around 0.6.

c) Calculate the mean and standard deviation of the data. Compare these two statistics with the mean and standard deviation of your uniform distribution. Discuss.
The sample mean calculated is 0.553. This is close to the population or parameter mean of 0.5.
The sample standard deviation calculated is 0.268, this is close to the population standard deviation of 0.2887.

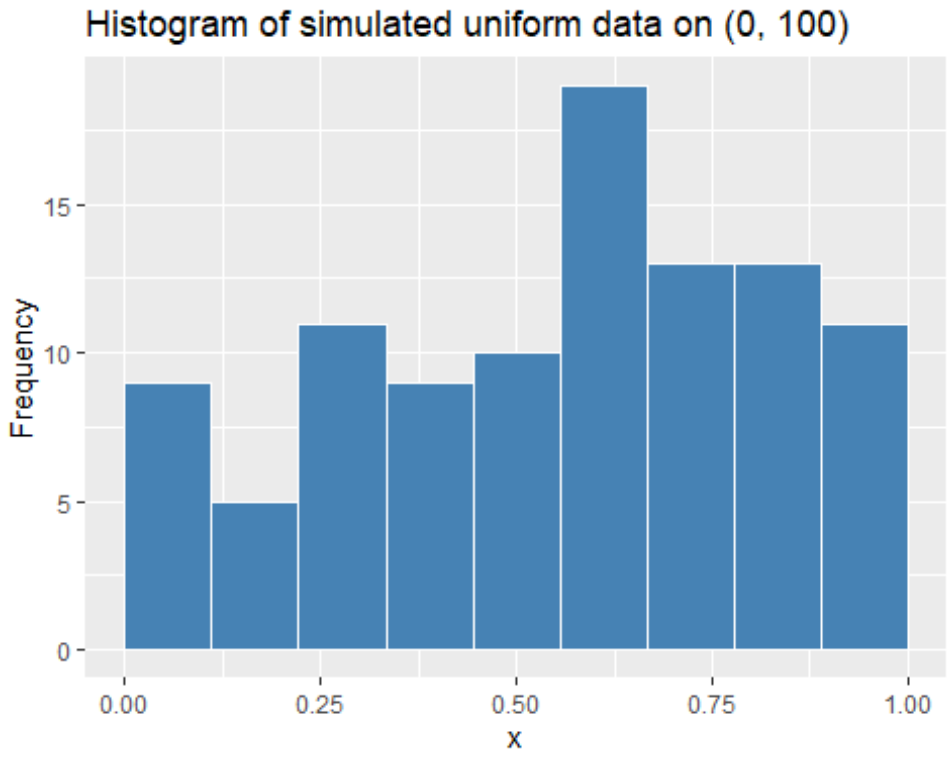
d) Estimate the following three cumulative probabilities based on your sample data and compare them with the corresponding cumulative probabilities of your uniform distribution:
i. Pr(X ≤ 0.5)

The cumulative probability from our generated data is 0.505, which is close to the value of o.5 from the uniform distribution.

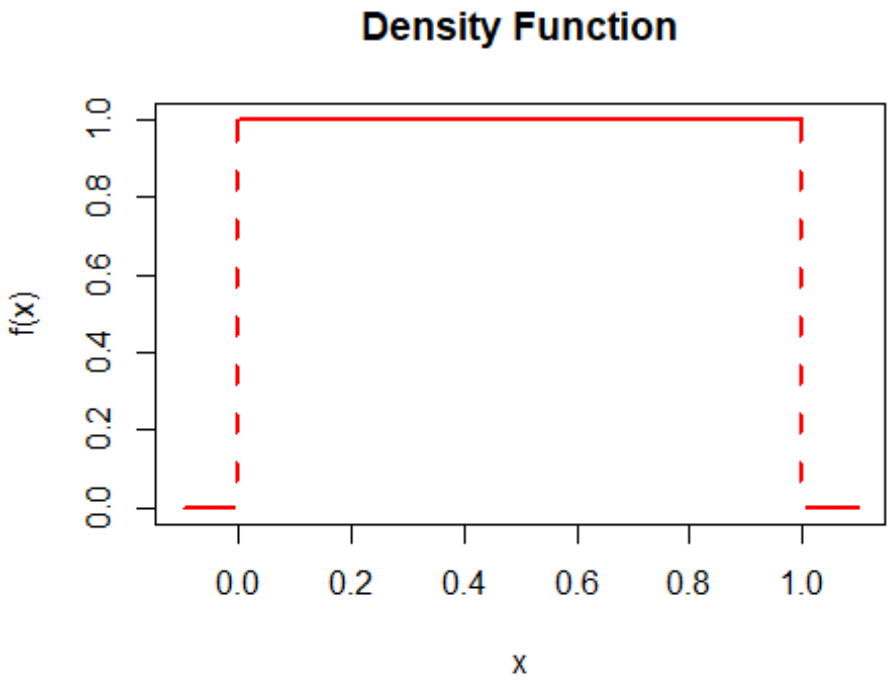
ii. Pr(0.3 ≤ X ≤ 0.5)
The cumulative probability calculated from our generated data is 0.203, which is close to the 0.2 calculated from the uniform distribution.

iii. Pr(X ≤ 0.7 | X > 0.25). The cumulative probability calculated from our generated data is 0.61, which is close to the 0.6 calculated from our uniform distribution.

```
##############################################################################
###
# Question 1
##############################################################################
###
# a) Generate 100 data values at random
seed <- 3933533
set.seed(seed)

a <- 0       #min value
b <- 1       #max value
n <- 100     #sample size

data_values <- runif(n, min = a, max = b)
head(data_values, 10)

##  [1] 0.35076953 0.08431648 0.44524975 0.64411892 0.39894326 0.93250161
##  [7] 0.60536972 0.27479247 0.04394548 0.30109492

##############################################################################
###
# b) Plotting Histogram
ggplot() +
  geom_histogram(aes(data_values), boundary = b, bins = 10, color = 'white',
fill = 'steelblue') +
  scale_x_continuous(limits = c(a, b)) +
  labs(x = "x", y = 'Frequency', title = "Histogram of simulated uniform data
on (0, 100)")
```




```
## Density Plot
plotUnif(min = a, max = b, type = "den", col = "red")
```



```r
################################################################################
###
# c) Calculating Mean and Standard Deviation.
distribution_mean <- (a + b)/2
distribution_variance <- ((b-a)^2)/12
distribution_sd <- sqrt(distribution_variance)

sample_mean <- mean(data_values)
round(sample_mean, 3)
```

```
## [1] 0.553
```

```r
sample_standard_deviation <- sd(data_values)
round(sample_standard_deviation, 3)
```

```
## [1] 0.268
```

```r
################################################################################
###
# d)
sample_min <- min(data_values)
sample_max <- max(data_values)


#i.    Pr(X ≤ 0.5)
# From distribution
x1 <- 0.5
prob1 <- punif(x1, min = a, max = b)

# From sample
sample_prob1 <- punif(x1, min = sample_min, max = sample_max)
round(sample_prob1, 3)
```

```
## [1] 0.505
```

```r
#ii.   Pr(0.3 ≤ X ≤ 0.5)
# From Distribution
x2 <- 0.3
x3 <- 0.5

prob2 <- punif(x2, min = a, max = b, lower.tail = TRUE)
prob3 <- punif(x3, min = a, max = b, lower.tail = TRUE)

ans2 <- prob3 - prob2
round(ans2, 3)
```

```
## [1] 0.2
```

```r
# From sample
sample_prob2 <- punif(x2, min = sample_min, max = sample_max, lower.tail =
TRUE)
```

```r
sample_prob3 <- punif(x3, min = sample_min, max = sample_max, lower.tail =
TRUE)

sample_ans2 <- sample_prob3 - sample_prob2
round(sample_ans2, 3)
```

```
## [1] 0.203
```

```r
#iii.  Pr(X ≤ 0.7 | X > 0.25).
# From Population
x4 <- 0.7
x5 <- 0.25


probA <- punif(x4, min = a, max = b, lower.tail = TRUE)
probB <- punif(x5, min = a, max = b, lower.tail = TRUE)
prob_b <- punif(x5, min = a, max = b, lower.tail = FALSE)

ans3 <- (probA - probB)/prob_b
round(ans3, 3)
```

```
## [1] 0.6
```

```r
# From Sample
sample_probA <- punif(x4, min = sample_min, max = sample_max, lower.tail =
TRUE)
sample_probB <- punif(x5, min = sample_min, max = sample_max, lower.tail =
TRUE)
sample_prob_b <- punif(x5, min = sample_min, max = sample_max, lower.tail =
FALSE)

sample_ans3 <- (sample_probA - sample_probB)/sample_prob_b
round(sample_ans3, 3)
```

```
## [1] 0.61
```

Question 2

An experiment is conducted on the effectiveness of a flue vaccine with a group of 200 people. People are equally split into to groups:
Group A:- receive the flu jab. Group B:- receive a placebo.

At the end of winter all participants are asked to disclose whether they contracted the flu.

It can be assumed that the number in each group who contract the flu will follow a Binomial distribution. Participants in group A are expected to have a 10% chance of contracting the flu whereas, for those in group B, the chance of catching the flu is 30%.

a) Generate synthetic data for the number who 'contract the flu' in each group, and complete the above table.

Table is completed in **Flu_data** data frame. with results from synthetic data as follows: Group A: 10 contracted flu, 90 did not get flu. Group B: 31 contracted flu, 69 did not get flu.

b)
i. What is your estimated probability that a person receiving the placebo will not contract the flu during the winter season? From the data generated, probability is 69%.

ii. Derive a 95% confidence interval for the proportion of people receiving the placebo who do not contract the flu in that winter season. Give a non-technical explanation of your result.
From the code located in section Question 2b(ii) below, we estimate with a 95% level of confidence, that the proportion of people receiving the placebo who do not get the flu in winter is between 0.599 and 0.781.

The result is arrived at point estimate is created for the proportion of people who are not receiving the jab and are not infected with the flu. The standard of error of the point estimate is obtained and used to calculate the margin of error for our estimate.

iii. If 40% of the wider adult population receive the flu vaccine, what percentage would you anticipate contracting the flu that year?
As calculated below in section question 2b(iii), the percentage of people we expect contracting the flu is 22.6%

iv. What is your estimate of the probability that a person who is vaccinated against the flu for ten years running, will contract the flu on at least three of those years?
From code under Question 2b (iv) below, the probability is estimated to be %0.07

```
##############################################################################
###
# Question 2
##############################################################################
###
# a) Generate synthetic data for the number who 'contract the flu' in each
group, and complete the above table.

# Group A:-
seed <- 2933533
set.seed(seed)

p_a <- 0.10 # probability of catching flu
x_a <- 100
n_a <- 1


Group_A <- rbinom(x_a, n_a, p_a)
head(Group_A, 10) # look at first 10 sampled values
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0
```

```
# Group B:-
seed <- 2933533
set.seed(seed)

p_b <- 0.30 # probability of catching flu
x_b <- 100
n_b <- 1


Group_B <- rbinom(x_b, n_b, p_b)
head(Group_B, 10) # look at first 10 sampled values
```

```
##  [1] 1 0 0 0 0 0 0 0 0 0
```

```
Test <- cbind(Group_A, Group_B)
Test <- as.data.frame(Test)


# Totalling frequencies for Group A and B
A_No_Flu <- sum(Test$Group_A == 0)
A_Got_Flu <- sum(Test$Group_A == 1)

B_No_Flu <- sum(Test$Group_B == 0)
B_Got_Flu <- sum(Test$Group_B == 1)


#Creating Flu dataframe

Flu_Data <-
  data.frame(Condition = c("Contracted the Flu", "Did not contract the Flu",
"Total"),
                       Group_A = c(A_Got_Flu,A_No_Flu, x_a),
                       Group_B = c(B_Got_Flu, B_No_Flu, x_b))

Flu_Data
```

```
##                   Condition Group_A Group_B
## 1       Contracted the Flu      10      31
## 2 Did not contract the Flu      90      69
## 3                     Total     100     100
```

```
################################################################################
###
# Question 2 - b
################################################################################
###
# i) Probability person receiving placebo will not get flu
```

```r
ans_bi <- B_No_Flu/x_b
ans_bi
```

```
## [1] 0.69
```

```r
##############################################################################
###
# ii)

#Calculations
est_p <- ans_bi   # sample proportion
sample_var <- est_p*(1-est_p) # sample variance
s_error    <- sqrt(sample_var/x_b)   # standard error
conf_level <- 0.95
alpha <- 1-conf_level
z_value <- qnorm(alpha/2, mean = 0, sd = 1, lower.tail = FALSE)
m_error <- z_value*s_error

# Confidence limits
LCL <- est_p-m_error # Lower
UCL <- est_p+m_error # Upper

round(LCL, 3)
```

```
## [1] 0.599
```

```r
round(UCL, 3)
```

```
## [1] 0.781
```

```r
##############################################################################
###
# iii)

numberA_get_flu <- A_Got_Flu
vacc_p <- numberA_get_flu/x_a


numberB_get_flu <- B_Got_Flu
unvacc_p <- B_Got_Flu/x_b

#Expected number of people vaccinated getting flu:-
vaccinated_proportion <- 0.4
expected_vacc_flu <- vacc_p*vaccinated_proportion

#expected number of unvaccinated getting flu:-
unvaccinated_prop <- 1-vaccinated_proportion
expected_novaccs_flu <- unvacc_p*unvaccinated_prop
```

```r
# Expected portion catching flu
expected_prop_flu <- expected_vacc_flu+expected_novaccs_flu
##############################################################################
#
# iii) Alternate solution

n <- 100 # size of random sample
prop_unvacc <- 1 - 0.4 # incidence of smoking in pop'n of British physicians
nbr_unvacc <- seq(0,n) # create seqeuence vector
nbr_vacc <- 100 - nbr_unvacc
prob_comb <- dbinom(nbr_unvacc, n, prop_unvacc)

# Expected number of unvaccinated with flu
exp_nbr_unvacc_flu <- nbr_unvacc*unvacc_p


# Expected number of vaccinated with flu
exp_nbr_vacc_flu <- nbr_vacc*vacc_p

# total expected flu per combination
exp_flu_per_comb <- exp_nbr_unvacc_flu + exp_nbr_vacc_flu
ans_2biii <- crossprod(prob_comb,exp_flu_per_comb) # equivalent to sumprod in
Excel

round(ans_2biii, digits = 3)
```

```
##      [,1]
## [1,] 22.6
```

```r
##############################################################################
###
# iv)

#P(x > 2) as we are testing that they will get infected at least 3 times

x_2biv <- 2
years <- 10

number_get_flu <- A_Got_Flu
prob_a_flu <- number_get_flu/x_a

ans_2biv <- pbinom(x_2biv, years, prob_a_flu, lower.tail = FALSE)
round(ans_2biv, 3)
```

```
## [1] 0.07
```

Question 3

This question explores the distribution of the sample mean when the underlying variable has an exponential distribution with mean 10.

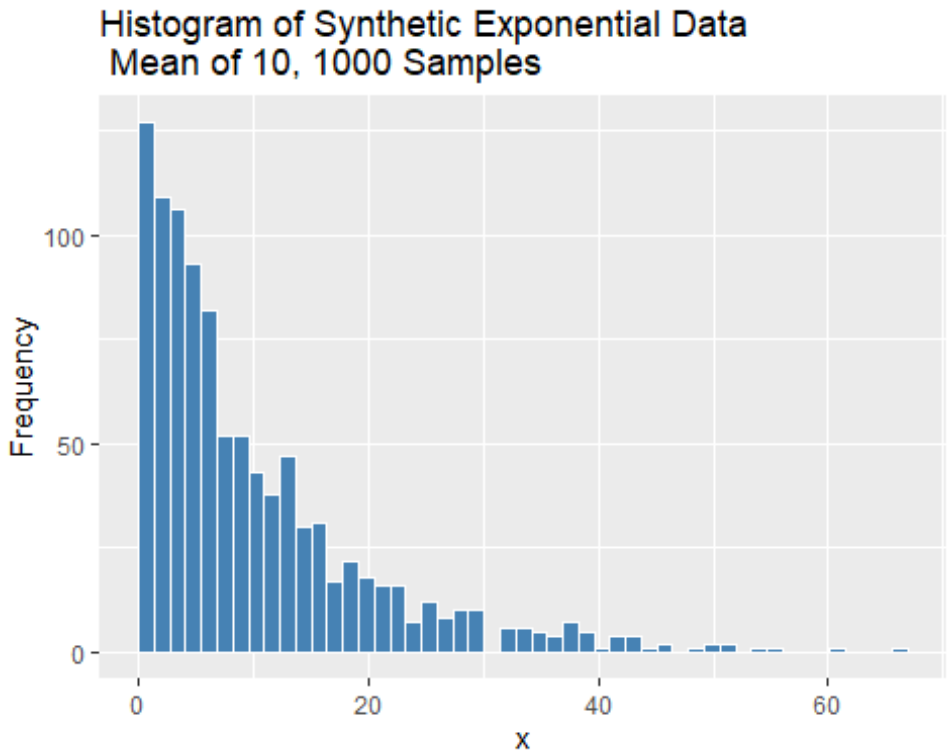
a) Generate 1000 observation from the given distribution.
i. Compare the sample estimates for the mean and standard deviation with the population mean and standard deviation for the exponential distribution. Discuss.
The sample mean calculated in from our generated data is 10.067, which is close to our population mean of 10. The sample standard deviation calculated from our generated data is 10.196 which is close to out population standard deviation of 10.

The differences between the generated values compared to the distribution differ slightly due to the randomness in our synthetic data.

ii. Compare the histogram of the sample data with the density function of the exponential distribution. Discuss.
Comparing the histogram of our sample data in **data_q3** with the density of the exponential distribution, we can see that they are very close to each other. Given that the sample mean and standard deviation calculated and provided in question 3ai is very close to the population mean and standard deviation, the similarities between the histogram and density plot are in keeping with these values being very close.

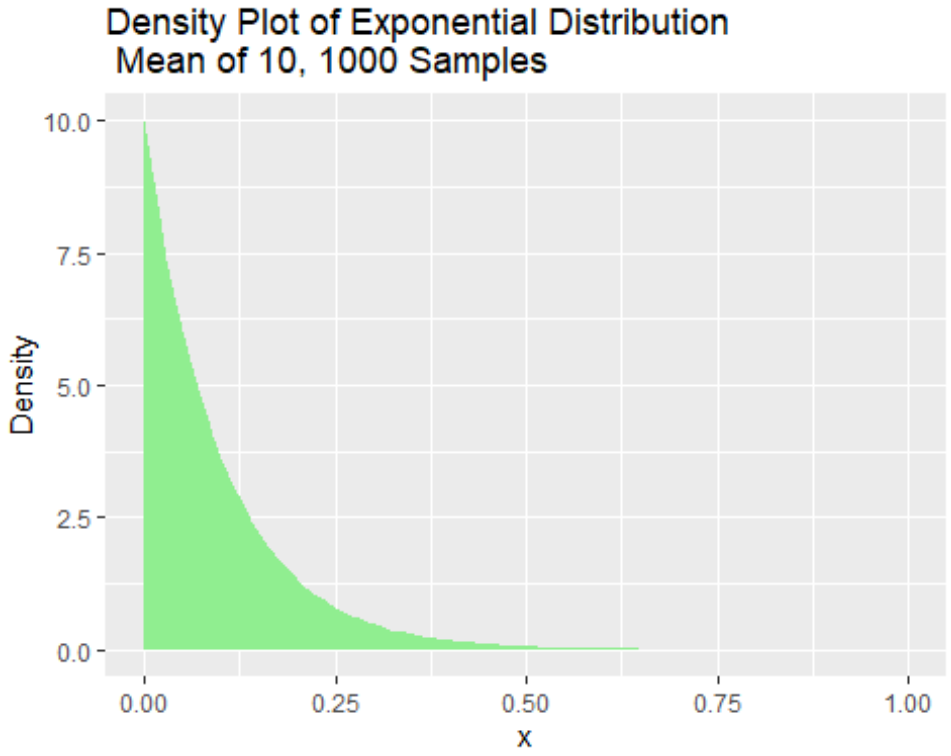
```
################################################################################
###
# 3a)
################################################################################
###
# Suppose X has an exponential distribution with lambda = 1/10

lambda <- 1/10

# Generate 1000 random observations on X

# Set Seed
random_number <- 2933533 # this can be any number
set.seed(random_number)

n3a <- 1000 # number of simulated values

data_q3 <- rexp(n3a, lambda)
head(data_q3,10) # look at first 10 sampled values

##  [1]  5.4758393 13.5066799 14.6812206  1.5684873 29.7954097 35.1787606
##  [7] 42.2145443 11.3536779  0.1179311  4.9222968

################################################################################
###

#### i)
#mean

sample_mean_q3 <- mean(data_q3)
round(sample_mean_q3, 3)
```

```
## [1] 10.067

# Standard Deviation
sample_sd_q3 <- sd(data_q3)
round(sample_sd_q3, 3)

## [1] 10.196

# ii)
# Histogram
q3_min <- 0
q3_max <- max(data_q3)

GG_data_q3 <- ggplot() +
  geom_histogram(aes(data_q3), boundary = q3_max, bins = 50, color = 'white',
fill = 'steelblue') +
  scale_x_continuous(limits = c(q3_min, q3_max)) +
  labs(x = "x", y = 'Frequency', title = "Histogram of Synthetic Exponential
Data\n Mean of 10, 1000 Samples")
GG_data_q3
```




```
# Density plot

Density_q3plot <- ggplot(data = data.frame(x = c(0, 1)), aes(x)) +
  stat_function(fun = dexp,
                geom = "area",
                fill = "light green",
```

```
                n = n,
                args = list(rate = 10)) +
  labs(x = "x", y = 'Density', title = "Density Plot of Exponential
Distribution\n Mean of 10, 1000 Samples")

Density_q3plot
```



Question 3 - continued

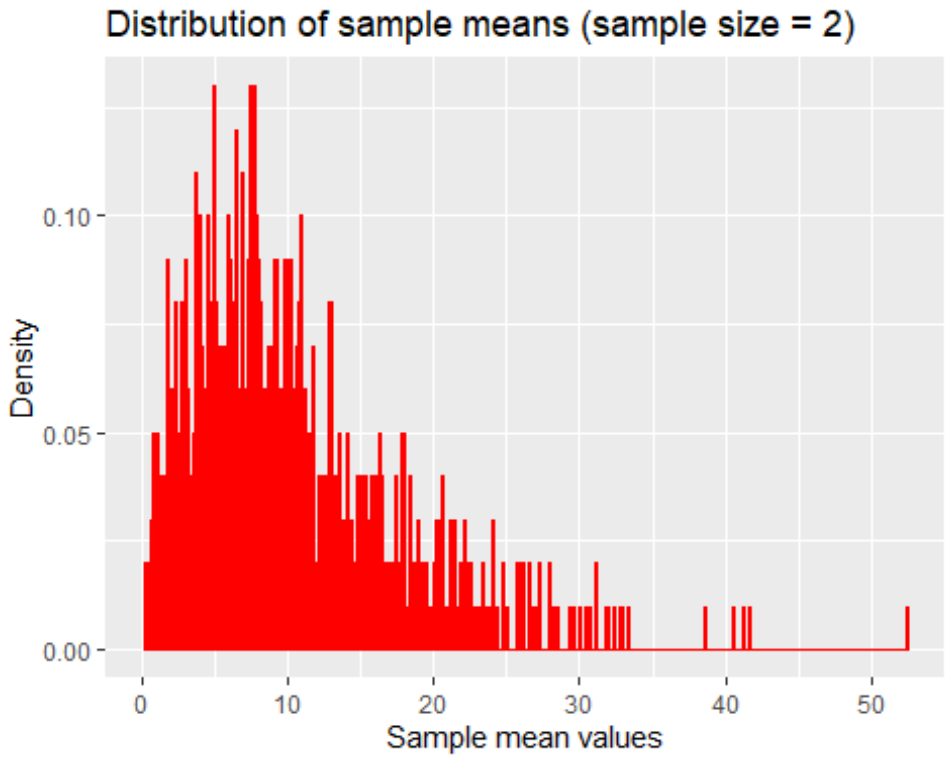
b) Generate 1000 sample means of sample size 2 where the observations are drawn at random from the exponential distribution.
i. Display a histogram of the sample means and discuss its characteristics.
From 1000 sample means with a sample size 2 drawn from an exponential distribution we can plot the histogram below. We can observe that there is a very slight tendency towards a normal distribution.

This is part of the Central Limit Theorem (CLT) which states that when sampling from a population with mean (μ) and standard deviation (σ), the sampling distribution of the sample mean will tend to a normal distribution with mean (μ) and standard deviation ($\sigma/\sqrt{n}$) as the sample size increases.

c) Generate 1000 sample means of sample size 30 where the observations are drawn at random from the exponential distribution.
i. Display a histogram of the sample means and discuss its characteristics.
From 1000 sample means with a sample size 30 drawn from an exponential distribution we can plot the histogram below. We can observe that the plot of 1000 sample means

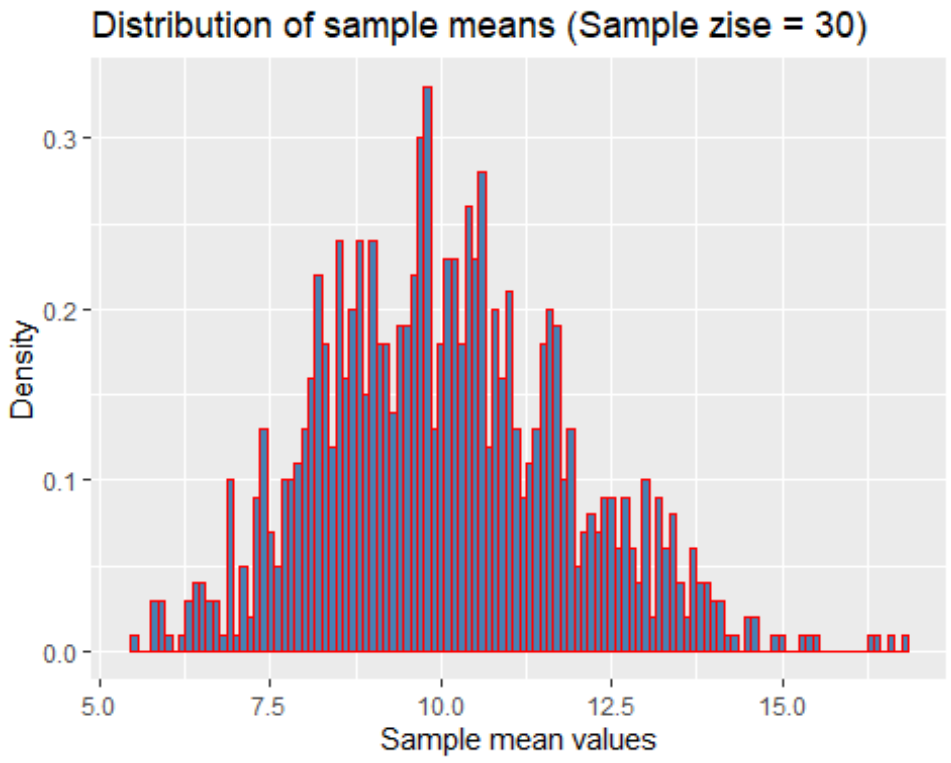
resembles a uniform distribution given the larger sample size of 30. This is consistant with the Central Limit Theorem (CLT).

```
###############################################################################
###
# 3b)
###############################################################################
###


n3b <- 2 # sample size
rows_3b <- 1000 # samples

data_q3b <- rexp(n3b*rows_3b, lambda)
head(data_q3b,10) # look at first 10 sampled values

##  [1]  5.7792819  2.4528646  8.8410829  5.0730442 10.3752675  1.5409576
##  [7]  2.7816476  0.9965296  0.4378116 13.0122377

# reshape vector into matrix with 1000 rows
m3b <- matrix(data_q3b, rows_3b)
x.bars_3b <- apply(m3b, 1, mean) # 1 returns a vector of row means


# Calculates means of sample means
avg.x.bars_3b <- format(round(mean(x.bars_3b), 3), nsmall = 3)
cat(paste('Mean of sampling distribution (sample size = 2) ',avg.x.bars_3b))

## Mean of sampling distribution (sample size = 2)  9.990

#Calculates Standard Deviation of sample means.
sd.x.bars_3b <-format(round(sd(x.bars_3b), 3), nsmall = 3)
cat(paste('Standard deviation of sampling distribution: ',sd.x.bars_3b))

## Standard deviation of sampling distribution:  6.967

x.bars_3b <- as.data.frame(x.bars_3b) # required for ggplot
head(x.bars_3b$x.bars_3b,10) # look at first 10 sampled values

##  [1] 5.6587726 3.2039119 8.1345760 9.8593396 9.6908631 2.8543654 5.4297300
##  [8] 0.8009811 3.7411600 6.9999561

Plot_GG_data_q3b <-ggplot(data = x.bars_3b) +
  geom_histogram(mapping = aes(x = x.bars_3b, y=..density..),
                 color = 'red',fill="steelblue", binwidth = 1/10) +
  labs(title = paste("Distribution of sample means (sample size = 2)"),
       x ="Sample mean values", y = "Density")
Plot_GG_data_q3b
```




```
###############################################################################
###
# 3c)
###############################################################################
###


n3c <- 30 # sample size
rows_3c <- 1000 # samples

data_q3c <- rexp(n3c*rows_3c, lambda)
head(data_q3c,10) # look at first 10 sampled values

##  [1]  0.9441471 20.3723737 16.1345835  0.6233299  1.0560436  7.1456850
##  [7]  4.1257369  8.2887600  4.9186411  0.8162019

# reshape vector into matrix with 1000 rows
m3c <- matrix(data_q3c, rows_3c)
x.bars_3c <- apply(m3c, 1, mean) # 1 returns a vector of row means


# Calculates means of sample means
avg.x.bars_3c <- format(round(mean(x.bars_3c), 3), nsmall = 3)
cat(paste('Mean of sampling distribution: ',avg.x.bars_3c))

## Mean of sampling distribution:  10.065
```

```r
#Calculates Standard Deviation of sample means.
sd.x.bars_3c <-format(round(sd(x.bars_3c), 3), nsmall = 3)
cat(paste('Standard deviation of sampling distribution: ',sd.x.bars_3b))

## Standard deviation of sampling distribution:  6.967

x.bars_3c <- as.data.frame(x.bars_3c) # required for ggplot
head(x.bars_3c$x.bars_3c,10) # look at first 10 sampled values

##  [1]  9.052568 11.160521  9.891790 13.391250  9.660742  9.957862  9.942457
##  [8]  8.206482  8.297210  8.609383

Plot_GG_data_q3c <- ggplot(data = x.bars_3c) +
  geom_histogram(mapping = aes(x = x.bars_3c, y=..density..),
                 color = 'red',fill="steelblue", binwidth = 1/10) +
  labs(title = paste("Distribution of sample means (Sample zise = 30)"),
       x ="Sample mean values", y = "Density")
Plot_GG_data_q3c
```



Question 4

a) In preparation for your meeting with management:
i. Conduct a one-tailed test of the difference between the means, using a 5% significance level. What can you conclude from the test?

From the market research conducted using a sample of 20 randomly chosen people, the average ratings of attractiveness for the current and new packaging were 7.5 and 8.2 out of 10 respectively. The standard deviation of the difference calculated is 1.9.

As the population standard deviation is unknown and the sample size is less than 30, a right-tailed t-test is conducted with a 5% significance level.

Our null hypothesis is that there is no difference between the means of the current and new packaging.

As the analysis is based on the difference of means with the same underlying population we first calculate D which is calculated below as:

$$D = x_2 - x_1$$

from there, our t-test is calculated: $T = \frac{\overline{D}}{(s_D/\sqrt{n})}$

Degrees of freedom is calculated as $df = n - 1$ where n is our sample size of 20.

The resulting t-test gives us a p-value of 0.058. Given that the resulting p-value is greater than our significance level of 5%, we fail to reject our null hypothesis.

ii. Conduct any further analysis that you feel is necessary.

While there is a difference of means between current and new packaging, the result of the t-test tells us that this is insignificant. It's possible that the test conducted resulted in a type ii error (β), where we have incorrectly failed to reject the null hypothesis and thereby miss the benefits of increased sales that the dairy producer may enjoy as a result from updating their packaging.

An analysis of the effect size of the test conducted above provides a result of 0.368. Using Jacob Cohens measure of effect size, where there is an inverse relationship between the effect size d and sample size, the result of 0.368 is small.

Therefore, in order to conduct a test that avoids a type ii (β) error, market research would need to be conducted using a larger sample size. Assuming, the effect size is equal and a desired power of 0.8, we can determine the appropriate sample size for a test which will avoid the type ii error. Using the **pwr.t.test** function in R gives us a result of 47.

b) Write up your advice to management, together with key supporting points.

A hypothesis test was conducted using the results of the market research and found that there is insufficient evidence in the difference of the attractiveness between the current packaging and new packaging. The estimated effect size of the test was found to be small. In order to appropriately measure the difference of attractiveness of the new packaging, a larger sample size is needed.

```
################################################################################
###
## Question 4
```

```r
################################################################################
###
# a) Conduct a one-tailed test of the difference between the means
n_participants <- 20
mean_old <- 7.5
mean_new <- 8.2

sd_package <- 1.9

d_bar <- mean_new - mean_old


# t_score
t_score <- d_bar/(sd_package/(sqrt(n_participants)))

# Degrees of freedom
df <- n_participants -1

# Significance level - alpha
alpha <- 0.05

# Right-tail test
p_value <- 1 - pt(t_score, df)
round(p_value, 3)
```

```
## [1] 0.058
```

```r
#result
result_4 <- p_value < alpha
result_4
```

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

```r
################################################################################
###
# Question 4 - b
################################################################################
###

#Estimate of effect
d_hat <- (mean_new - mean_old)/sd_package
round(d_hat, 3)
```

```
## [1] 0.368
```

```r
#referring to Jacob Cohen's measure of effect, the value is quite small...

################################################################################
###
# power test to determine sample size for effectiveness.
```

```r
#pwr.2p.test

# desired detectable effect
d <-0.5

# desired power
power <- 0.8

#desired significance level
sig.level <-0.05


#power test - type is "paired" as samples are dependent.
pwr.t.test(d = 0.368, power = 0.8, sig.level = 0.05,
           type = "paired", alternative = "greater")

##
##      Paired t test power calculation
##
##               n = 47.0361
##               d = 0.368
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
##
## NOTE: n is number of *pairs*
```

Question 5

a) Generate 10 pairs of observations from a bivariate Normal distribution with the following parameter values: $(\mu_{1)}, \mu_2, \sigma_1, \sigma_2, \rho)$ **= (50, 55, 10, 10, 0.8)**

i. Decide on an appropriate hypothesis test of the difference between the two means based only on these 10 observations (i.e. as though you are unaware of the parameter values). Explain how you arrived at that decision.

Though the parameters are unknown. We do know that the underlying distribution is a bivariate normal distribution with a small sample of 10 pairs. Given that the sample size is small ($n \leq 30$) and the population standard deviation (σ) is unknown, the use of a **Two-Sided t-test** to test differences of means would be an appropriate test to perform.

ii. Conduct the test using a 5% significance level, and write up your findings. Conclude this part of the question by reflecting on the ability of the test to reach a correct conclusion.

The t-test is conducted with a 5% significance level and the following hypotheses: $H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$

The result of t-test returns a p-value of 0.1229. As this result is greater than 0.05, we fail to reject the null hypothesis that $\mu_0 = \mu_1$.

b) Generate 30 pairs of observations from a bivariate Normal distribution with the following parameter values: $(\mu_{1)}, \mu_2, \sigma_1, \sigma_2, \rho)$ **= (50, 55, 10, 10, 0.8)**

i. Decide on an appropriate hypothesis test of the difference between the two means based only on these 30 observations (i.e. as though you are unaware of the parameter values). Explain how you arrived at that decision.
Though the parameters are unknown. We do know that the underlying distribution is a bivariate normal distribution with a sample of 30 pairs. Given that the sample size is ($n \leq 30$) and the population standard deviation (σ) is unknown, the use of a **Two-Sided t-test**, as used in Question 5a, to test differences of means would be an appropriate test to perform.

ii. Conduct the test using a 5% significance level, and write up your findings. Conclude this part of the question by reflecting on the ability of the test to reach a correct conclusion.

The t-test is conducted with a 5% significance level and the following hypotheses: $H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$

The result of t-test returns a p-value = 8.239e-05 As this result is lesser than 0.05, we reject the null hypothesis that $\mu_0 = \mu_1$, and assume that they are different.

The ability of the test to reach the correct conclusions are affected by the factors affecting the power of the test:

- significance level
- sample size
- effect size
- standard deviation.

In our case, the test conducted on the distribution where sample $n = 30$, the higher sample size, with other factors remaining equal, will mean a smaller standard of error, and a higher likelihood to reject the null hypothesis.

c) Compare your answers to parts (a) and (b). What can you conclude?
Though we assumed that the parameters of the distributions generated in questions 5a and 5b were unknown for the purposes of our hypothesis tests, they were generated using the the same parameters.

The differing results between 5a and 5b stress the importance of the factors which affect the power of the test:

- significance level
- sample size
- effect size
- standard deviation.

The higher sample size means a reduced standard of error and thus a higher likelihood to reject the null hypothesis.

The results highlight the importance of a high sample size in hypothesis testing.

```
##############################################################################
###
## Question 5 - a
##############################################################################
###
seed <- 2933533
set.seed(seed)

mean_X <- 50 # mean of X
mean_Y <- 55 # mean of Y

sd_X <- 10 # standard deviation of X
sd_Y <- 10 # standard deviation of X

cor_XY <- 0.8 # correlation of X and Y

mu <- c(mean_X,mean_Y) # vector of means
cov_XY <- sd_X*sd_Y*cor_XY # rearranging the correlation formula
sigma <- matrix(c(sd_X^2, cov_XY, cov_XY, sd_Y^2), nrow = 2) # variance-
covariance matrix

# Generate observations

random_number <- 23933533 # this can be any number
set.seed(random_number)

n5a <- 10 # number of simulated values
Q5a_Data <- mvrnorm(n5a, mu, sigma)
colnames(Q5a_Data) <- c("X", "Y")

Q5a_Data

##              X        Y
##  [1,] 51.30465 49.37548
##  [2,] 39.34527 50.42266
##  [3,] 42.94387 54.97885
##  [4,] 53.99379 52.68075
##  [5,] 50.46263 52.92391
##  [6,] 53.15481 57.92838
##  [7,] 59.47273 65.06521
##  [8,] 49.21299 42.09455
##  [9,] 52.80954 55.37687
## [10,] 40.89308 43.94911
```

```
#############################################################################
##
###  Convert Q5a_Data into data frame

Q5a_Data_DF <- as.data.frame(Q5a_Data)


#############################################################################
###
# t-test
t.test(Q5a_Data_DF$X,Q5a_Data_DF$Y, paired = TRUE)

##
##  Paired t-test
##
## data:  Q5a_Data_DF$X and Q5a_Data_DF$Y
## t = -1.7023, df = 9, p-value = 0.1229
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.266568  1.026084
## sample estimates:
## mean of the differences
##               -3.120242

#############################################################################
###
# Question 5b
#############################################################################
###


random_number <- 23933533 # this can be any number
set.seed(random_number)

n5b <- 30 # number of simulated values
Q5b_Data <- mvrnorm(n5b, mu, sigma)
colnames(Q5b_Data) <- c("X", "Y")

Q5b_Data

##              X        Y
##  [1,] 42.46860 58.21152
##  [2,] 43.92693 45.84101
##  [3,] 48.04134 49.88138
##  [4,] 51.81838 54.85615
##  [5,] 46.81776 56.56877
##  [6,] 48.56436 62.51883
##  [7,] 60.89525 63.64269
##  [8,] 34.68061 56.62693
##  [9,] 57.77829 50.40812
```

```
## [10,] 42.28011 42.56208
## [11,] 39.37961 44.83288
## [12,] 62.93714 60.29502
## [13,] 57.26367 68.84129
## [14,] 43.94805 42.11283
## [15,] 46.16755 51.21629
## [16,] 48.27664 56.04408
## [17,] 54.70411 52.07333
## [18,] 35.98793 32.65673
## [19,] 45.04399 52.65799
## [20,] 46.20263 52.96549
## [21,] 44.01426 49.63985
## [22,] 53.24990 57.63230
## [23,] 50.15245 46.26600
## [24,] 33.66583 44.48021
## [25,] 58.59344 64.60588
## [26,] 53.80033 68.19383
## [27,] 42.06108 51.17474
## [28,] 35.62919 37.95754
## [29,] 48.97778 64.92469
## [30,] 44.12628 52.56610
```

```
##############################################################################
##
### Convert Q5b_Data into data frame

Q5b_Data_DF <- as.data.frame(Q5b_Data)
```

```
##############################################################################
###
# t-test
t.test(Q5b_Data_DF$X,Q5b_Data_DF$Y, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Q5b_Data_DF$X and Q5b_Data_DF$Y
## t = -4.5756, df = 29, p-value = 8.239e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.238247 -3.148490
## sample estimates:
## mean of the differences
##               -5.693369
```

Question 6

A supermarket chain conducts an experiment to compare the sales effectiveness of three promotional scenarios for its home brand chocolate blocks:

1. Include home brand chocolate blocks in its weekly catalogue of specials 2. Promote home brand chocolate blocks with an end of isle display 3. Include home brand chocolate blocks in its weekly catalogue of specials, and at the same time promote the chocolate blocks with an end of isle display.

Each scenario is trialed in 6 different stores comparable in size and sales (18 stores in total).

a) Generate six monthly sales (recorded in thousands of units) for each scenario, according to the following:

- **All sales are normally distributed with a standard deviation of 10.**
- **Scenario 1 sales have a mean of 50.**
- **Scenario 2 sales have a mean of 60.**
- **Scenario 3 sales have a mean of 70.**

i. Select an appropriate test

For our scenario, we have a single factor which we will refer to as **Scenario**, with 3 levels:

- **Scenario1**
- **Scenario2**
- **Scenario3**

Though we do not know the parameters values of our data, we can assume the following:

- **A random sample from each group's population**
- **Groups are independent of each other**
- **Observations within groups are normally distributed**
- **Group variances are the same. Confirmed using Levene's test for homogeneity).**
- **Appropriate treatment of any outliers**

Thus, the One-Way ANOVA test appropriate for testing the differences between means.

ii - Results - One-Way Anova with 5% significance level
Before the we conduct the One-Way Anova test, we ensure that the assumptions are met:

Random data is generated separately for each scenario below using the **rnorm** function. There are 6 stores and 6 monthly sales, and therefore each scenario has 36 observations each.

The resulting observation within each group or scenario are normally distributed. This is also supported by the Shapiro-Wilks test.

Using Levene's Test returns p-value of 0.6809, thus we can assume homogeneity.

A boxplot of the sales data generated, reveals there are 3 outliers found in **scenario 2.**

We conduct a One-Way ANOVA with the following hypotheses: H_0: $\mu_1 = \mu_2$ H_a: At lease one mean is different.

The One-Way Anova is conducted twice. Once with the outliers present in scenario 2 and once without.

Results:-
ANOVA with outliers:- p-value = <2e-16 **ANOVA without outliers:-** p-value = <2e-16

In both scenarios, we reject the null hypothesis and can assume there is a significant difference between the means.

Post-Hoc Test using Tuckeys As the ANOVA test we conducted has found a significant difference between each scenario. A pot-Hoc test using Tuckeys is used for further analysis in determining where the differences between each scenario lie.

As the presence of outliers had no impact on the outcome of the ANOVA test, we will look at the post-hoc results without the outliers.

The marketing **1. Include home brand chocolate blocks in its weekly catalogue of specials 2. Promote home brand chocolate blocks with an end of isle display 3. Include home brand chocolate blocks in its weekly catalogue of specials, and at the same time promote the chocolate blocks with an end of isle display.**

Between **Scenario 2 vs Scenario 1**, the results do not support a significant difference, as p-value = 0.111, between promoting the home brand chocalate with an end of isle display and promoting them in the weekly catalogue.

Between **Scenario 3 vs Scenario1** the results support a significant difference, as p-value = 0.000, between inclusion of home brand chocalate in weekly catalogue specials and end of isle displays vs only including them in weekly catalogue specials. There is a mean sales difference of 20,916 units (3 decimal places).

Between **Scenario 3 vs Scenario 2** the results support a significant difference, as p-value = 0.000, between inclusion of home brand chocalate in weekly catalogue specials and end of isle displays vs only promoting them with an end of isle display. There is a mean sales difference of 25,272 units (3 decimal places).

```
##############################################################################
###
# Generating Sample Data for each marketing scenario.
##############################################################################
###
#############                Scenario 1                ###############
seed <- 3933533
set.seed(seed)
```

```r
scenario1_mean <- 50
scenario_sd <- 10

Six_monthly_sales <- 6

Obs <- 36 # Number of observations. From 6 stores per scenario * 6 monthly sales

scenario1_data <- rnorm(Obs, scenario1_mean, scenario_sd)



Stores_Scenario1 <- as.factor(c(rep("Store_1A", 6),
                        rep("Store_1B", 6),
                        rep("Store_1C", 6),
                        rep("Store_1D", 6),
                        rep("Store_1E", 6),
                        rep("Store_1F", 6)))

CandyShop1_DF <- data.frame(scenario1_data, Stores_Scenario1)

# Create new column labelled "Market Scenario"
CandyShop1_DF %<>% mutate(Market_Scenario = "Scenario_1")
str(CandyShop1_DF)
```

```
## 'data.frame':    36 obs. of  3 variables:
##  $ scenario1_data  : num  46.2 48.6 47.4 52.7 32.9 ...
##  $ Stores_Scenario1: Factor w/ 6 levels "Store_1A","Store_1B",..: 1 1 1 1
1 1 2 2 2 2 ...
##  $ Market_Scenario : chr  "Scenario_1" "Scenario_1" "Scenario_1"
"Scenario_1" ...
```

```r
###############            Scenario 2           ###############
scenario2_mean <- 60
scenario_sd <- 10
Obs <- 36 # number of simulated values

scenario2_data <- rnorm(Obs, scenario1_mean, scenario_sd)



Stores_Scenario2 <- as.factor(c(rep("Store_2A", 6),
                        rep("Store_2B", 6),
                        rep("Store_2C", 6),
                        rep("Store_2D", 6),
                        rep("Store_2E", 6),
                        rep("Store_2F", 6)))
```

```r
CandyShop2_DF <- data.frame(scenario2_data, Stores_Scenario2)


# Create new column labelled "Market Scenario"
CandyShop2_DF %<>% mutate(Market_Scenario = "Scenario_2")
str(CandyShop2_DF)
```

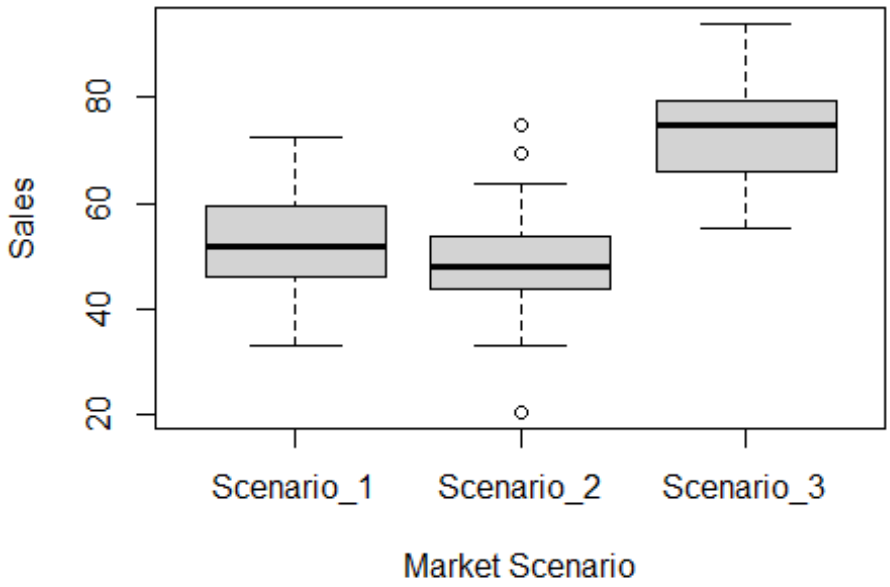
```
## 'data.frame':    36 obs. of  3 variables:
##  $ scenario2_data  : num  55 38.2 50.8 55.1 46.5 ...
##  $ Stores_Scenario2: Factor w/ 6 levels "Store_2A","Store_2B",..: 1 1 1 1
1 1 2 2 2 2 ...
##  $ Market_Scenario : chr  "Scenario_2" "Scenario_2" "Scenario_2"
"Scenario_2" ...
```

```r
################           Scenario 3           ##############
scenario3_mean <- 70
scenario_sd <- 10
Obs <- 36 # number of simulated values

scenario3_data <- rnorm(Obs, scenario3_mean, scenario_sd)



Stores_Scenario3 <- as.factor(c(rep("Store_3A", 6),
                        rep("Store_3B", 6),
                        rep("Store_3C", 6),
                        rep("Store_3D", 6),
                        rep("Store_3E", 6),
                        rep("Store_3F", 6)))

CandyShop3_DF <- data.frame(scenario3_data, Stores_Scenario3)

# Create new column labelled "Market Scenario"
CandyShop3_DF %<>% mutate(Market_Scenario = "Scenario_3")
str(CandyShop3_DF)
```

```
## 'data.frame':    36 obs. of  3 variables:
##  $ scenario3_data  : num  81.1 77.1 66.8 55.2 58 ...
##  $ Stores_Scenario3: Factor w/ 6 levels "Store_3A","Store_3B",..: 1 1 1 1
1 1 2 2 2 2 ...
##  $ Market_Scenario : chr  "Scenario_3" "Scenario_3" "Scenario_3"
"Scenario_3" ...
```

```r
################################################################################
###

# Renaming columns

#Scenario 1
CandyShop1_DF <- rename(CandyShop1_DF, "Sales" = "scenario1_data")
```

```r
CandyShop1_DF <- rename(CandyShop1_DF, "Store" = "Stores_Scenario1")

#Scenario 2
CandyShop2_DF <- rename(CandyShop2_DF, "Sales" = "scenario2_data")
CandyShop2_DF <- rename(CandyShop2_DF, "Store" = "Stores_Scenario2")

# Scenario 3
CandyShop3_DF <- rename(CandyShop3_DF, "Sales" = "scenario3_data")
CandyShop3_DF <- rename(CandyShop3_DF, "Store" = "Stores_Scenario3")

##############################################################################
###
# Combine Data into new Data Frame

Chocolate_DF <- rbind(CandyShop1_DF, CandyShop2_DF, CandyShop3_DF)
Chocolate_DF$Market_Scenario <- as.factor(Chocolate_DF$Market_Scenario)


##############################################################################
###
# Box plot of Dataframe for overview

boxplot1 <- boxplot(Chocolate_DF$Sales ~ Chocolate_DF$Market_Scenario,
                    Main = "Sales by Market Scenario", ylab = "Sales",
                    xlab = "Market Scenario")
```



```
#############################################################################
###
##  Shapiro Test for Normality

# Test Scenario 1:-
shapiro.test(CandyShop1_DF$Sales)

##
##  Shapiro-Wilk normality test
##
## data:  CandyShop1_DF$Sales
## W = 0.98735, p-value = 0.9471

# Test Scenario 2:-
shapiro.test(CandyShop2_DF$Sales)

##
##  Shapiro-Wilk normality test
##
## data:  CandyShop2_DF$Sales
## W = 0.9601, p-value = 0.2165

# Test Scenario 3:-
shapiro.test(CandyShop3_DF$Sales)

##
##  Shapiro-Wilk normality test
##
## data:  CandyShop3_DF$Sales
## W = 0.97311, p-value = 0.5165

#############################################################################
###
## Levene Test

leveneTest(Sales ~ Market_Scenario, data = Chocolate_DF, center = mean)

## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value Pr(>F)
## group   2  0.3857 0.6809
##       105

#############################################################################
###
# ANOVA Test - With Outliers
results1 <- aov(Sales ~ Market_Scenario, data = Chocolate_DF)
summary(results1)

##                  Df Sum Sq Mean Sq F value Pr(>F)
## Market_Scenario   2  12746    6373   66.59 <2e-16 ***
## Residuals       105  10049      96
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

################################################################################
###
## Removed outliers from scenario 2
outliers <- boxplot1$out
#outliers <- boxplot(Chocolate_DF$Sales, plot = FALSE)$out
Chocolate_DF2_Clean <- Chocolate_DF
Chocolate_DF2_Clean <- Chocolate_DF2_Clean[-which(Chocolate_DF2_Clean$Sales
%in% outliers),]


################################################################################
##
# ANOVA Test - Without Outliers
results1_clean <- aov(Sales ~ Market_Scenario, data = Chocolate_DF2_Clean)
summary(results1_clean)

##                  Df Sum Sq Mean Sq F value Pr(>F)
## Market_Scenario   2  12840    6420   80.55 <2e-16 ***
## Residuals       102   8130      80
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

################################################################################
###
# Post-hoc


TukeyHSD(results1)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Sales ~ Market_Scenario, data = Chocolate_DF)
##
## $Market_Scenario
##                             diff      lwr       upr     p adj
## Scenario_2-Scenario_1 -3.78846 -9.27027  1.693351 0.2323046
## Scenario_3-Scenario_1 20.91582 15.43401 26.397633 0.0000000
## Scenario_3-Scenario_2 24.70428 19.22247 30.186093 0.0000000

TukeyHSD(results1_clean)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Sales ~ Market_Scenario, data = Chocolate_DF2_Clean)
##
## $Market_Scenario
```

```
##                               diff       lwr        upr     p adj
## Scenario_2-Scenario_1 -4.355982 -9.473337  0.7613736 0.1114946
## Scenario_3-Scenario_1 20.915822 15.910950 25.9206949 0.0000000
## Scenario_3-Scenario_2 25.271804 20.154449 30.3891598 0.0000000

#############################################################################
###

# ii) 5% significance
#############################################################################
###

#with outliers
pf(5.00, df1 = 2, df2 = 105,lower.tail = FALSE)

## [1] 0.008429495

# Without Outliers
pf(5.00, df1 = 2, df2 = 102,lower.tail = FALSE)

## [1] 0.008481831
```

b) Repeat part (a) under the changed assumption the standard deviation of sales is 5.
We repeat the steps of question 6a with the only change that the standard deviation is 5. The results of our ANOVA tests are as follows:

Results:-
ANOVA with outliers:- p-value = <2e-16 **ANOVA without outliers:-** p-value = <2e-16

As with Question 6a, our p-values are very small, and we can assume to reject the null hypothesis and assume a significant difference between the means.

Post-Hoc Test using Tuckeys

As the ANOVA test we conducted has found a significant difference between each scenario. A pot-Hoc test using Tuckeys is used for further analysis in determining where the differences between each scenario lie.

Again, as the presence of outliers had no impact on the outcome of the ANOVA test, we will look at the post-hoc results without the outliers present in the data set.

Unlike the results of Question 6a, Tuckeys has found a significant difference in all combinations.

Between **Scenario 2 vs Scenario 1**, the results support a significant difference, as p-value = 0.000 between promoting the home brand chocalate with an end of isle display and promoting them in the weekly catalogue. There is a mean sales difference of 7822 units. (3 decimal places).

Between **Scenario 3 vs Scenario1** the results support a significant difference, as p-value = 0.000, between inclusion of home brand chocalate in weekly catalogue specials and end of

isle displays vs only including them in weekly catalogue specials. There is a mean sales difference of 20,458 units (3 decimal places).

Between **Scenario 3 vs Scenario 2** the results support a significant difference, as p-value = 0.000, between inclusion of home brand chocalate in weekly catalogue specials and end of isle displays vs only promoting them with an end of isle display. There is a mean sales difference of 12,636 units (3 decimal places).

c) Compare your answers to parts (a) and (b). What can you conclude?

In question 6a and 6b, we were able to reject the null hypothesis and assume that the differences of means between scenarios were different.

However, in 6a, our post-hoc test did not support a significant difference in **Scenario 2 vs Scenario 1**, where as the difference was supported in the other 2 combinations.

In 6b, our post-hoc test supported a significant difference in all 3 scenarios. The difference of results between 6a and 6b can be attributed to the factors which affect the power of the test. In 6b, all parameters remained the same from 6a apart from the standard deviation.

The smaller Standard Deviation reduced the standard error of the test statistic, in turn reducing the p-value, and increasing the power of the test.

```
#############################################################################
###
#  Question 6 - b
#############################################################################
###

##############                Scenario 1                ###############
seed <- 3933533
set.seed(seed)


scenario_1b_mean <- 50
scenario_b_sd <- 5



Obs <- 36 # Number of observations. From 6 stores per scenario * 6 monthly
sales

scenario_1b_data <- rnorm(Obs, scenario_1b_mean, scenario_b_sd)



Stores_Scenario_1b <- as.factor(c(rep("Store_1A", 6),
                          rep("Store_1B", 6),
                          rep("Store_1C", 6),
                          rep("Store_1D", 6),
```

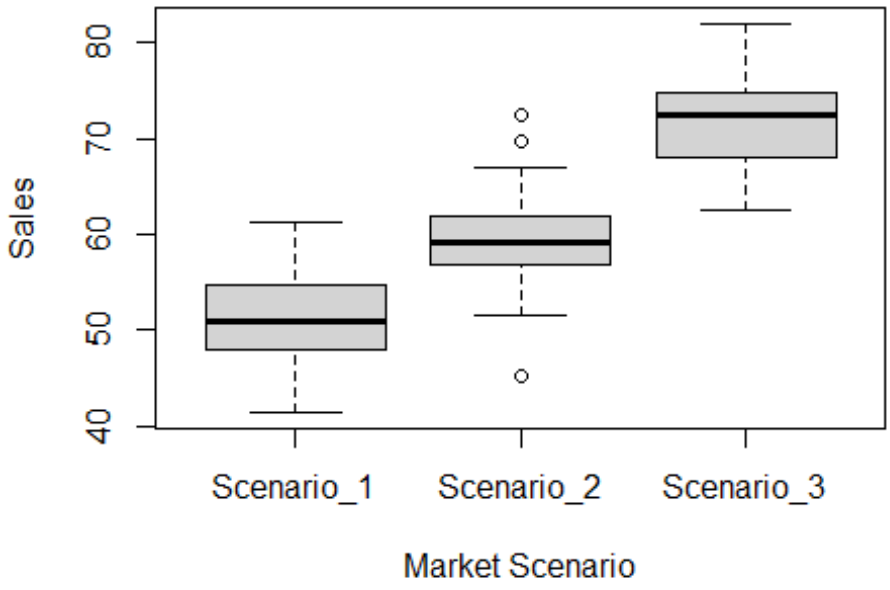
```
                        rep("Store_1E", 6),
                        rep("Store_1F", 6)))

CandyShop1b_DF <- data.frame(scenario_1b_data, Stores_Scenario_1b)

# Create new column labelled "Market Scenario"
CandyShop1b_DF %<>% mutate(Market_Scenario = "Scenario_1")
str(CandyShop1b_DF)

## 'data.frame':    36 obs. of  3 variables:
##  $ scenario_1b_data  : num  48.1 49.3 48.7 51.3 41.5 ...
##  $ Stores_Scenario_1b: Factor w/ 6 levels "Store_1A","Store_1B",..: 1 1 1
1 1 1 2 2 2 2 ...
##  $ Market_Scenario   : chr  "Scenario_1" "Scenario_1" "Scenario_1"
"Scenario_1" ...

###############             Scenario 2           ###############
scenario_2b_mean <- 60
scenario_b_sd <- 5

Obs <- 36 # number of simulated values

scenario_2b_data <- rnorm(Obs, scenario_2b_mean, scenario_b_sd)



Stores_Scenario_2b <- as.factor(c(rep("Store_2A", 6),
                        rep("Store_2B", 6),
                        rep("Store_2C", 6),
                        rep("Store_2D", 6),
                        rep("Store_2E", 6),
                        rep("Store_2F", 6)))

CandyShop2b_DF <- data.frame(scenario_2b_data, Stores_Scenario_2b)


# Create new column labelled "Market Scenario"
CandyShop2b_DF %<>% mutate(Market_Scenario = "Scenario_2")
str(CandyShop2b_DF)

## 'data.frame':    36 obs. of  3 variables:
##  $ scenario_2b_data  : num  62.5 54.1 60.4 62.6 58.2 ...
##  $ Stores_Scenario_2b: Factor w/ 6 levels "Store_2A","Store_2B",..: 1 1 1
1 1 1 2 2 2 2 ...
##  $ Market_Scenario   : chr  "Scenario_2" "Scenario_2" "Scenario_2"
"Scenario_2" ...

################             Scenario 3           ###############
scenario_3b_mean <- 70
scenario_b_sd <- 5
```

```r
Obs <- 36 # number of simulated values

scenario_3b_data <- rnorm(Obs, scenario_3b_mean, scenario_b_sd)



Stores_Scenario_3b <- as.factor(c(rep("Store_3A", 6),
                          rep("Store_3B", 6),
                          rep("Store_3C", 6),
                          rep("Store_3D", 6),
                          rep("Store_3E", 6),
                          rep("Store_3F", 6)))

CandyShop3b_DF <- data.frame(scenario_3b_data, Stores_Scenario_3b)

# Create new column labelled "Market Scenario"
CandyShop3b_DF %<>% mutate(Market_Scenario = "Scenario_3")
str(CandyShop3b_DF)
```

```
## 'data.frame':    36 obs. of  3 variables:
##  $ scenario_3b_data  : num  75.5 73.6 68.4 62.6 64 ...
##  $ Stores_Scenario_3b: Factor w/ 6 levels "Store_3A","Store_3B",..: 1 1 1
1 1 1 2 2 2 2 ...
##  $ Market_Scenario   : chr  "Scenario_3" "Scenario_3" "Scenario_3"
"Scenario_3" ...
```

```r
##############################################################################
###

# Renaming columns

#Scenario 1
CandyShop1b_DF <- rename(CandyShop1b_DF, "Sales" = "scenario_1b_data")
CandyShop1b_DF <- rename(CandyShop1b_DF, "Store" = "Stores_Scenario_1b")

#Scenario 2
CandyShop2b_DF <- rename(CandyShop2b_DF, "Sales" = "scenario_2b_data")
CandyShop2b_DF <- rename(CandyShop2b_DF, "Store" = "Stores_Scenario_2b")

# Scenario 3
CandyShop3b_DF <- rename(CandyShop3b_DF, "Sales" = "scenario_3b_data")
CandyShop3b_DF <- rename(CandyShop3b_DF, "Store" = "Stores_Scenario_3b")

##############################################################################
###
# Combine Data into new Data Frame

Chocolate_B_DF <- rbind(CandyShop1b_DF, CandyShop2b_DF, CandyShop3b_DF)
Chocolate_B_DF$Market_Scenario <- as.factor(Chocolate_B_DF$Market_Scenario)
```

```
#############################################################################
###
# Box plot of Dataframe for overview

boxplot2 <- boxplot(Chocolate_B_DF$Sales ~ Chocolate_DF$Market_Scenario,
                    Main = "Sales by Market Scenario", ylab = "Sales",
                    xlab = "Market Scenario")
```




```
#############################################################################
###
##  Shapiro Test for Normality

# Test Scenario 1:-
shapiro.test(CandyShop1b_DF$Sales)

##
##  Shapiro-Wilk normality test
##
## data:  CandyShop1b_DF$Sales
## W = 0.98735, p-value = 0.9471

# Test Scenario 2:-
shapiro.test(CandyShop2b_DF$Sales)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CandyShop2b_DF$Sales
## W = 0.9601, p-value = 0.2165

# Test Scenario 3:-
shapiro.test(CandyShop3b_DF$Sales)

##
##  Shapiro-Wilk normality test
##
## data:  CandyShop3b_DF$Sales
## W = 0.97311, p-value = 0.5165

################################################################################
###
## Levene Test

leveneTest(Sales ~ Market_Scenario, data = Chocolate_B_DF, center = mean)

## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value Pr(>F)
## group   2  0.3857 0.6809
##       105

################################################################################
###
# ANOVA Test
resultsb <- aov(Sales ~ Market_Scenario, data = Chocolate_B_DF)
summary(resultsb)

##                  Df Sum Sq Mean Sq F value Pr(>F)
## Market_Scenario   2   7642    3821   159.7 <2e-16 ***
## Residuals       105   2512      24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

################################################################################
###
## Removed outliers from scenario 2
outliers_b <- boxplot2$out
#outliers <- boxplot(Chocolate_DF$Sales, plot = FALSE)$out
Chocolate_B_Clean <- Chocolate_B_DF
Chocolate_B_Clean <- Chocolate_B_Clean[-which(Chocolate_B_Clean$Sales %in%
outliers_b),]



resultsb_clean <- aov(Sales ~ Market_Scenario, data = Chocolate_B_Clean)
summary(resultsb_clean)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## Market_Scenario    2   7665    3832   192.3 <2e-16 ***
## Residuals        102   2032      20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
###############################################################################
###
# Post-hoc Test - Tuckeys


TukeyHSD(resultsb)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Sales ~ Market_Scenario, data = Chocolate_B_DF)
##
## $Market_Scenario
##                             diff       lwr      upr p adj
## Scenario_2-Scenario_1  8.10577  5.364865 10.84668     0
## Scenario_3-Scenario_1 20.45791 17.717006 23.19882     0
## Scenario_3-Scenario_2 12.35214  9.611236 15.09305     0
```

```
TukeyHSD(resultsb_clean)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Sales ~ Market_Scenario, data = Chocolate_B_Clean)
##
## $Market_Scenario
##                               diff       lwr      upr p adj
## Scenario_2-Scenario_1  7.822009  5.263331 10.38069     0
## Scenario_3-Scenario_1 20.457911 17.955475 22.96035     0
## Scenario_3-Scenario_2 12.635902 10.077224 15.19458     0
```


datasciencemadesimple.com 2022, *Aggregate() Function in R*, Data Science Made Simple, viewed 20 May 2022, https://www.datasciencemadesimple.com/aggregate-function-in-r/

LAERD Statistics, 2022, *Testing for Normality using SPSS Statistics*, LAERD Statistics, viewed 21 May 2022, https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics.php

Cikaric, D, 2021, *20 Must-Know Energy Drink Statistics and Facts for 2022* Med Alert Help, viewed 2021 May 2002, https://medalerthelp.org/blog/energy-drink-statistics/

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.37.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Stefan Milton Bache and Hadley Wickham (2022). magrittr: A Forward-Pipe Operator for R. R package version 2.0.2. https://CRAN.R-project.org/package=magrittr

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Yihui Xie (2022). tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents. R package version 0.38.

Yihui Xie (2019) TinyTeX: A lightweight, cross-platform, and easy-to-maintain LaTeX distribution based on TeX Live. TUGboat 40 (1): 30–32. https://tug.org/TUGboat/Contents/contents40-1.html

María Isabel Borrajo-García, Mercedes Conde-Amboage and Alejandra López-Pérez (2021). LearningStats: Elemental Descriptive and Inferential Statistics. R package version 0.1.0. https://CRAN.R-project.org/package=LearningStats

John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Thomas W. Yee (2015). Vector Generalized Linear and Additive Models: With an Implementation in R. New York, USA: Springer.

Thomas W. Yee and C. J. Wild (1996). Vector Generalized Additive Models. Journal of Royal Statistical Society, Series B, 58(3), 481-493.

Stephane Champely (2020). pwr: Basic Functions for Power Analysis. R package version 1.3-0. https://CRAN.R-project.org/package=pwr