BACS HW7
Student ID: 110065508

Due: 2022/04/03

Question 1. First, read the data with the following code:

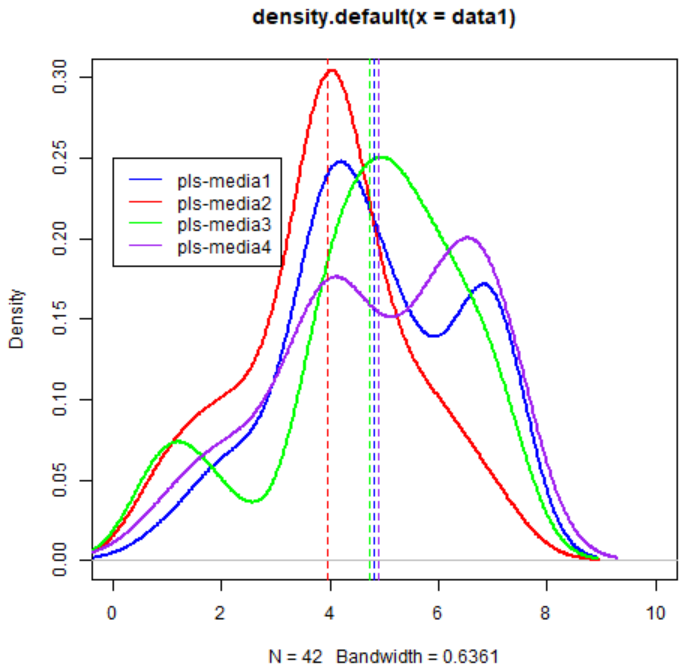
```r
library(reshape2)
library(tidyverse)

# Question 1
data1 <- read_csv('pls-media1.csv')$INTEND.0 # read the data
data2 <- read_csv('pls-media2.csv')$INTEND.0
data3 <- read_csv('pls-media3.csv')$INTEND.0
data4 <- read_csv('pls-media4.csv')$INTEND.0
```

(a) Using the `mean` function directly,

```r
# Question 1 (a)
data1_mean <- mean(data1)
data2_mean <- mean(data2)
data3_mean <- mean(data3)
data4_mean <- mean(data4)
```

(b) The density plot and the vertical lines (marking the means) with same colors are from the same type of data.

```r
# Question 1 (b)
png(filename = "1b.png")
plot(density(data1), col="blue", lwd=2, xlim=c(0, 10), ylim=c(0,0.3))
lines(density(data2), col="red", lwd=2)
lines(density(data3), col="green", lwd=2)
lines(density(data4), col="purple", lwd=2)
abline(v=data1_mean, lty="dashed", col="blue")
abline(v=data2_mean, lty="dashed", col="red")
abline(v=data3_mean, lty="dashed", col="green")
abline(v=data4_mean, lty="dashed", col="purple")
legend(0, 0.25,
       legend = c("pls-media1", "pls-media2", "pls-media3", "pls-media4"),
       col = c("blue","red","green","purple"),
       lty = 1:1, cex = 1)
dev.off()
```



(c) Yes, the mean of data from `pls-media2.csv` (red lines) looks like having a different mean. More specifically, the mean of the other types are close. ■

Question 2. (a) Let the four types of data have means $\mu_1, ..., \mu_4$. The null and alternative hypotheses of ANOVA are H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; H_1: $\mu_1 \neq \mu_2$, $\mu_1 \neq \mu_3$, $\mu_1 \neq \mu_4$, $\mu_2 \neq \mu_3$, $\mu_2 \neq \mu_4$, and $\mu_3 \neq \mu_4$.

(b) I conduct the one-way ANOVA by the following hand-carved code:

```
# Question 2 (b)
# Compute MSTR
data <- list(data1, data2, data3, data4)
sstr <- sum(sapply(data, length)*(sapply(data, mean) - mean(sapply(data, mean)))^2)
df_mstr <- 4-1
mstr <- sstr/df_mstr
# Compute MSE
sse <- sum((sapply(data, length)-1)*sapply(data, var))
df_mse <- sum(sapply(data, length)) - 4
mse <- sse/df_mse
# Compute F-value and p-value
f_value <- mstr/mse # F-value
qf(p=0.95, df1=df_mstr, df2=df_mse) # The cutoff value of F-value
p_value <- pf(f_value, df_mstr, df_mse, lower.tail=FALSE) # The p-value
```

(i) MSTR= 7.53239, MSE= 2.869151, and $F = 2.625303$.

(ii) The p-value of F is 0.05230686. Since $\alpha = 0.05$, The F-value is not significant. Hence we cannot reject null hypothesis H_0.

(c) To use the built-in function in `R`, we need to apply the `melt` function first to transform the 4 types of data into a dataframe.

```
# Question 2 (c)
names(data) <- c("pls-media1", "pls-media2", "pls-media3", "pls-media4")
data_aov <- melt(data,
                 id.vars = NULL,
                 variable.name = "type",
                 value.name = "intend")
anova_model <- aov(data_aov$intend ~ factor(data_aov$L1))
summary(anova_model)
```

It gives $F = 2.617$. The p-value of F is 0.0529. Since $\alpha = 0.05$, The F-value is not significant. Hence we cannot reject null hypothesis H_0.

(d) To conduct a post-hoc Tukey test,

```r
# Question 2 (d)
TukeyHSD(anova_model, conf.level = 0.05)
```

Maybe in this case, a screenshot of results is more appropriate. From the results, we may conclude that type 2 data (from `pls-media2.csv`) have a different mean from other types.

	diff	lwr	upr	p adj
pls-media2-pls-media1	-0.86215539	-1.06562977	-0.6586810	0.1085727
pls-media3-pls-media1	-0.08452381	-0.28530983	0.1162622	0.9959223
pls-media4-pls-media1	0.08178054	-0.11218249	0.2757436	0.9959032
pls-media3-pls-media2	0.77763158	0.57175512	0.9835080	0.1825044
pls-media4-pls-media2	0.94393593	0.74470805	1.1431638	0.0573229
pls-media4-pls-media3	0.16630435	-0.03017708	0.3627858	0.9687417

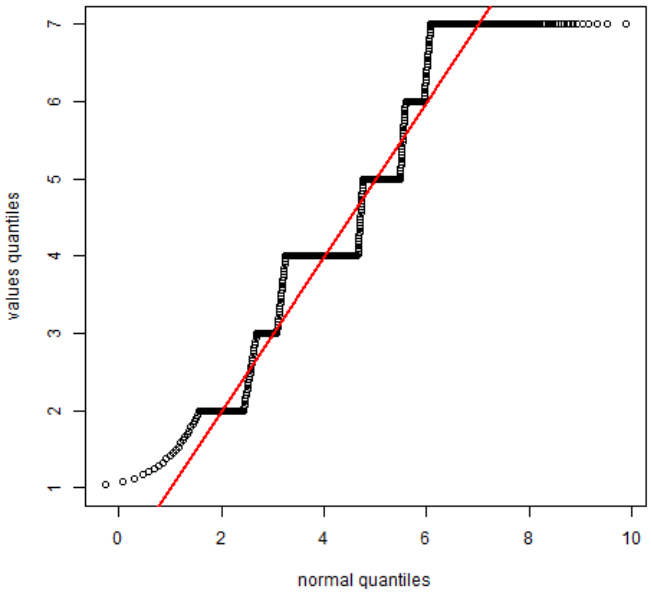
(e) Let's review the three conditions of ANOVA[1]:

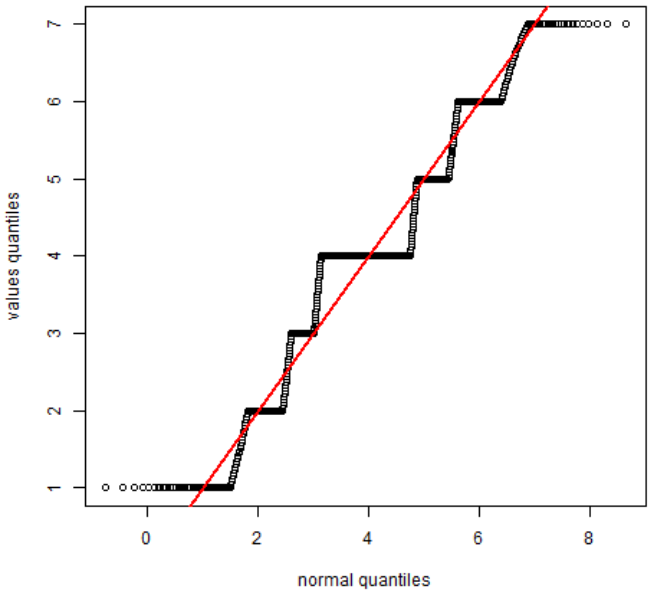
(i) **Each treatment/population's response variable is normally distributed**

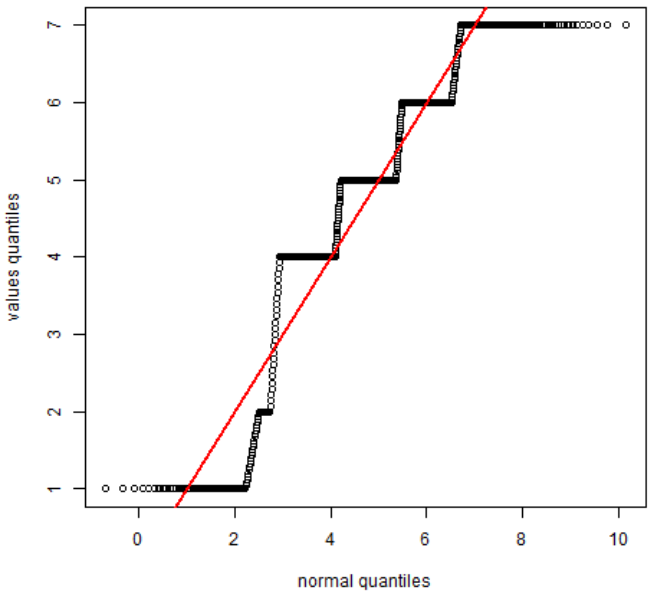
The Q-Q plot based method we used in last homework[2] helps.

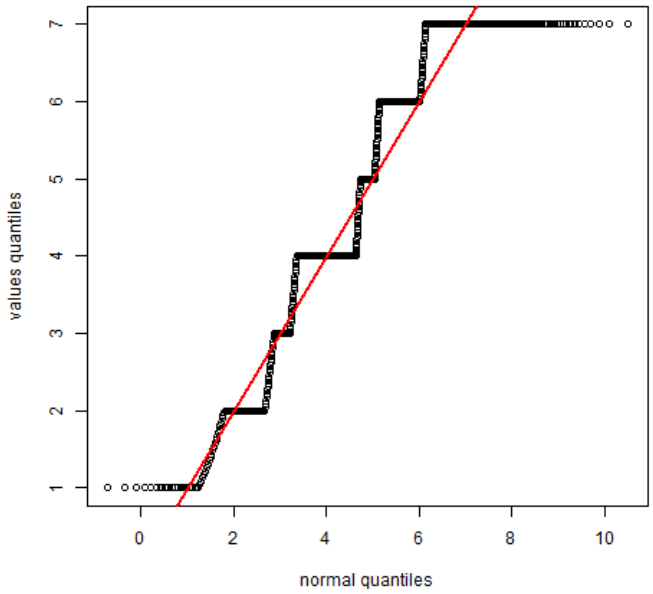
```r
norm_qq_plot <- function(values)
{
  probs1000 <- seq(0, 1, 0.001)
  q_vals <- quantile(values, probs=probs1000)
  q_norm <- qnorm(probs1000, mean=mean(values), sd=sd(values))
  plot(q_norm, q_vals, xlab="normal quantiles", ylab="values quantiles")
  abline(a=0, b=1, col="red", lwd=2)
}

png(filename = "2e-type1.png")
norm_qq_plot(data1)
dev.off()

png(filename = "2e-type2.png")
norm_qq_plot(data2)
dev.off()

png(filename = "2e-type3.png")
norm_qq_plot(data3)
dev.off()

png(filename = "2e-type4.png")
norm_qq_plot(data4)
dev.off()
```

Note that it is really unlikely that those data being normally distributed since they are discrete, whose values take ranges from 1 to 7. So we may predict that the Q-Q plot may look like step functions.


values quantiles
normal quantiles
0
2
4
6
8
10
1
2
3
4
5
6
7


values quantiles
normal quantiles
0
2
4
6
8
1
2
3
4
5
6
7


values quantiles
normal quantiles
0
2
4
6
8
10
1
2
3
4
5
6
7



(ii) **The variance (s^2) of the response variables is the same for all treatments/populations**

```
# Question 2 (e)
data1_sd <- sd(data1)
data2_sd <- sd(data2)
data3_sd <- sd(data3)
data4_sd <- sd(data4)
data_sd <- sd(data_aov$intend)
```

Compute the s^2 using `sd` functions, I found that the s^2 for all four types are 1.641506, 1.52364, 1.753933, 1.816324, and s^2 among all data is 1.718571. We may run a F-test (or something similar) to test whether they are significantly different.

(iii) **The observations are independent: the response variables are not related between groups**

This is hard to test, I assume (with a faith of belief that) this is true.

■

Question 3. (a) Let $\overline{R}_i$ be the mean of sum of ranks, $i = 1, 2, 3, 4$. The null and alternative hypotheses are H_0: $\overline{R}_1 = \overline{R}_2 = \overline{R}_3 = \overline{R}_4$; H_1: $\overline{R}_i \neq \overline{R}_j$, $\forall i \neq j$, $i, j = 1, 2, 3, 4$

(b) I compute Kruskal Wallis H by the following hand-carved code:

```
# Question 3 (b)
intend_rank <- rank(data_aov$intend)
data_aov$rank <- intend_rank # add as a new column to data_aov
group_ranks <- split(data_aov, data_aov$L1)
rank_sum_1 <- sum(group_ranks$`pls-media1`$rank)
rank_sum_2 <- sum(group_ranks$`pls-media2`$rank)
rank_sum_3 <- sum(group_ranks$`pls-media3`$rank)
rank_sum_4 <- sum(group_ranks$`pls-media4`$rank)
R <- c(rank_sum_1^2/length(data1),
       rank_sum_2^2/length(data2),
       rank_sum_3^2/length(data3),
       rank_sum_4^2/length(data4))
H = (12/(length(data)*(length(data)+1))) * sum(R) - 3*(length(data)+1)
kw_p <- 1 - pchisq(H, df=4-1)
```

(i) The code gives $H = 706140.1$.

(ii) The p-value is 0, we can reject the null hypothesis H_0. Hence The rank sum of all types are all different.

(c) The built in function of Kruskal Wallis test:

```
# Question 3 (c)
kruskal.test(intend~L1, data=data_aov)
```

It gives p-value= 0.03166. Since this $p < 0.05$, we can reject the null hypothesis H_0 and having the same conclusion of the hypotheses as in (b).

(d) The built in function of the post-hoc Dunn test:

```
# Question 3 (d)
dunnTest(intend~L1, data=data_aov, method = "bonferroni")
```

	Comparison	Z	P.unadj	P.adj
1	pls-media1 - pls-media2	2.30087819	0.021398517	0.12839110
2	pls-media1 - pls-media3	-0.09233644	0.926430736	1.00000000
3	pls-media2 - pls-media3	-2.36408588	0.018074622	0.10844773
4	pls-media1 - pls-media4	-0.31452459	0.753122646	1.00000000
5	pls-media2 - pls-media4	-2.65613380	0.007904225	0.04742535
6	pls-media3 - pls-media4	-0.21613379	0.828883460	1.00000000

We may conclude that type 2 data (from `pls-media2.csv`) have a different rank sum from other types.

■

References

[1] Handouts of week 7.

[2] A Q-Q Plot Dissection Kit, https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html