# STAT 425 - Statistical Design and Analysis of Experiments

# Assignment 2

**Name:** Jana Osea
**Student ID:** 30016679

1. The phenomenon of road rage has received much media attention in recent years. Is a driver's propensity to engage in road rage related to his or her income? Researchers in Mississippi State University attempted to answer this question by conducting a survey of a representative sample of 1,000 U.S. adult driver1 . Based on how often each driver engaged in certain road rage behaviors (ie., making obscene gestures at, tailgating, and thinking about physically hurting another driver), a road rage score was assigned. (Higher scores indicate a greater pattern of road rage behavior.) The drivers were also grouped by annual income: under $30,000, between $30,000 and $60,000, and over $60,000. The data was subjected to an analysis, producing the given statistics in the document.

   (a) Find the value of $\overline{X}$. (Hint: Are the sample sizes the same?)
   **Solution:** The overall average can be calculated by the following where $w_k$ is the sample weight of the $k-th$ treatment.

$$\begin{aligned}\overline{X} &= \sum_{k=1}^{3} w_k \overline{X}_k \\ &= \frac{379}{1038} \times 4.6 + \frac{392}{1038} \times 5.08 + \frac{267}{1038} \times 5.15 \\ &= 4.9227\end{aligned}$$

   (b) Provide the ANOVA table.
   **Solution:** The values can be obtained by the following:

$$SST = 8417.447$$

$$\begin{aligned}SSB &= \sum_{k=1}^{3} n_k \overline{X}_k^2 - n\overline{X}^2 \\ &= \left(379 \times 4.6^2 + 392 \times 5.08^2 + 267 \times 5.515^2\right) - 1038 \times 4.9227^2 \\ &= 62.961\end{aligned}$$

$$\begin{aligned}SSE &= SST - SSB \\ &= 8417.447 - 62.961 \\ &= 8354.486\end{aligned}$$

$$\begin{aligned} df_{between} &= k - 1 \\ &= 3 - 1 \\ &= 2 \end{aligned}$$

$$\begin{aligned} df_{within} &= n - k \\ &= 1038 - 3 \\ &= 1035 \end{aligned}$$

$$\begin{aligned} df_{total} &= n - 1 \\ &= 1038 - 1 \end{aligned}$$

$$\begin{aligned} MSB &= \frac{SSB}{df_{between}} \\ &= \frac{62.961}{2} \\ &= 31.481 \end{aligned}$$

$$\begin{aligned} MSW &= \frac{SSW}{df_{within}} \\ &= \frac{8354.486}{1035} \\ &= 8.072 \end{aligned}$$

$$\begin{aligned} F_{obs} &= \frac{\frac{SSB}{df_{between}}}{\frac{SSW}{df_{within}}} \\ &= \frac{\frac{62.961}{2}}{\frac{8354.486}{1035}} \\ &= \frac{31.481}{8.072} \end{aligned}$$

$$\begin{aligned} P - value &= P(F_{2,1035} > F_{obs}) \\ &= 0.0205 \end{aligned}$$

| Source | DF | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Between | 2 | 62.961 | 31.481 | 3.900 | 0.0205 |
| Within | 1035 | 8354.486 | 8.072 | | |
| Total | 1037 | 8417.447 | | | |

(c) What are the conditions under which an ANOVA analysis is conducted? State the two conditions in the context of the data collected.

**Solution:** There are 2 conditions that must be met.

i. Normality: The response variable $X_{ij}$ is Normally distributed. As well, the residual terms $e_{ij}$ are independent random variables that are Normally distributed with a mean of 0, or $E(e_{ij}) = 0$

ii. Homoscedasticity: The variance in the response variable (road range values) are the same across all groups (income groups).

(d) Assuming the conditions stated in (c) hold true, test the hypothesis that there is no difference in the road rage score across the three income classes. What is your decision? (Use $\alpha = 0.05$).

**Solution:** We wish to test the following hypothesis:

$$H_0 : \mu_{<30000} = \mu_{30000-60000} = \mu_{>60000}$$
$$H_a : \text{at least one } \mu \text{ is not the same for the } k = 3 \text{ populations}$$

Using the ANOVA table above, we get that

$$F_{obs} = 3.9$$
$$P - value = P(F_{2,1035} > F_{obs}) = 0.0205$$

Based on this data, the $P - value = P(F_{2,1035} > F_{obs}) = 0.0205$ which is less than 0.05. One can conclude that the three populations are not equal with respect to the response variable (road range values). The mean value of the response variable is different for at least one of these $k = 3$ populations.

(e) Interpret the meaning of the P-value in (d) in the context of the data.
**Solution:** Based on this data, the mean the road range value of at least one income group is not the same; there exists a 0.0205 probability of another experiment producing stronger statistical evidence against the null hypothesis than the current evidence.

(f) If the null hypothesis in (d) is not rejected, construct a 95% confidence interval for $\mu$ - the mean road rage score of all individuals. IF the null hypothesis in (d) is rejected, construct a 95% confidence interval estimates for (i) $\mu_{<30}$, (ii) $\mu_{30-60}$ and (iii) $\mu_{>60}$. Ensure you interpret the meaning of these intervals in the context of the data.
**Solution**: Null hypothesis is rejected and so we will be using the following formula where $k$ is the different income groups.

$$\overline{X_k} \pm t_{0.025,n_k-1}\frac{MSE}{\sqrt{n_k}}$$

Using R to perform the calculation, we obtain the following confidence intervals

```
4.6 + c(-1,1)*abs(qt(0.05/2, 379-1))*(sqrt(8.072)/sqrt(379))
```

```
## [1] 4.313046 4.886954
```

```
5.08 + c(-1,1)*abs(qt(0.05/2, 392-1))*(sqrt(8.072)/sqrt(392))
```

```
## [1] 4.797875 5.362125
```

```
5.15 + c(-1,1)*abs(qt(0.05/2, 267-1))*(sqrt(8.072)/sqrt(267))
```

```
## [1] 4.807655 5.492345
```

i. Based on this data, the 95% confidence interval for $\mu_{<30}$ is between 4.313046 and 4.886954. This means the if we perform this experiment many times, 95% of the road range averages calculated will be captured by this interval.

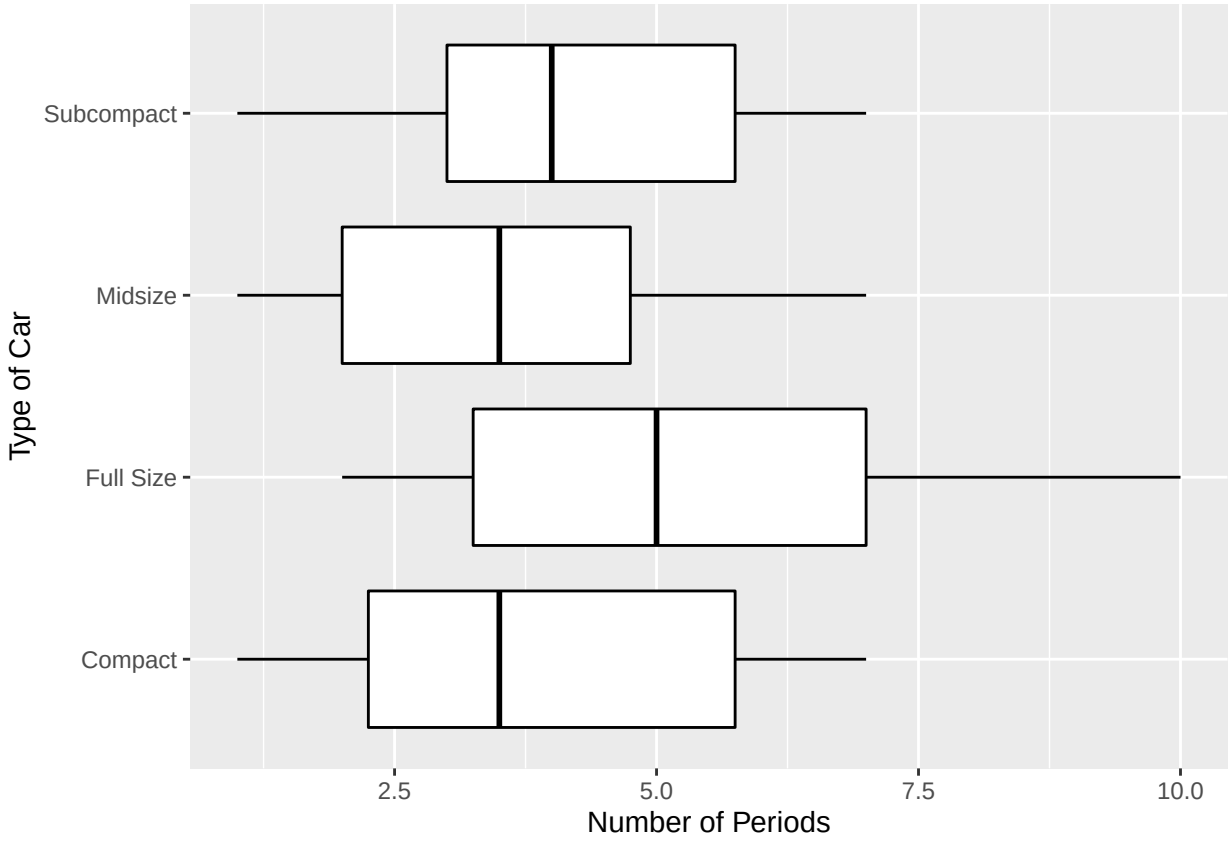
ii. Based on this data, the 95% confidence interval for $\mu_{30-60}$ is between 4.797875 and 5.362125. This means the if we perform this experiment many times, 95% of the road range averages calculated will be captured by this interval.

iii. Based on this data, the 95% confidence interval for $\mu_{>60}$ is between 4.807655 and 5.492345. This means the if we perform this experiment many times, 95% of the road range averages calculated will be captured by this interval.

2. Refer to the data in Chapter 3, end of chapter question 3.15.

(a) Create boxplots of these data with ggplot. What do these boxplots tell you about the association between the number of weeks a car is rented and type of car?

**Solution:** Based on our data, the boxplots below seem to show that the variances between each sample are the same because the inter quartile ranges look around the same size.




(b) Compute SST, SSB, and SSW and summarize your findings in the form of an ANOVA table.

**Solution:** Using the following R code, we can produce parts of the ANOVA table.

```
summary(aov(period~type, data=p2))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## type         3  16.67   5.558    1.11  0.358
## Residuals   36 180.30   5.008
```

Completing the table using the fact the $df_{Total} = n - 1$ and $SST = SSW + SSB$

| Source | DF | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Type | 3 | 16.67 | 5.558 | 1.11 | 0.358 |
| Residuals | 36 | 180.30 | 5.008 | | |
| Total | 39 | 196.97 | | | |

(c) From (b): Does the data indicate that the type of car rented affects the length of the rental contract? State the appropriate statistical hypothesis, provide the value of the test statistic, and compute/provide the P-value. What can you infer from these data? Carry out the test at $\alpha = 0.05$. In addition to computing its value, ensure you interpret the meaning of the P-value.

**Solution:** We are testing the following test of hypothesis:

$$H_0 : \mu_{Subcompact} = \mu_{Midsize} = \mu_{FullSize} = \mu_{Compact}$$
$$H_a : \text{at least one } \mu \text{ is not the same for the } k = 4 \text{ populations}$$

From the ANOVA table above we get the following results

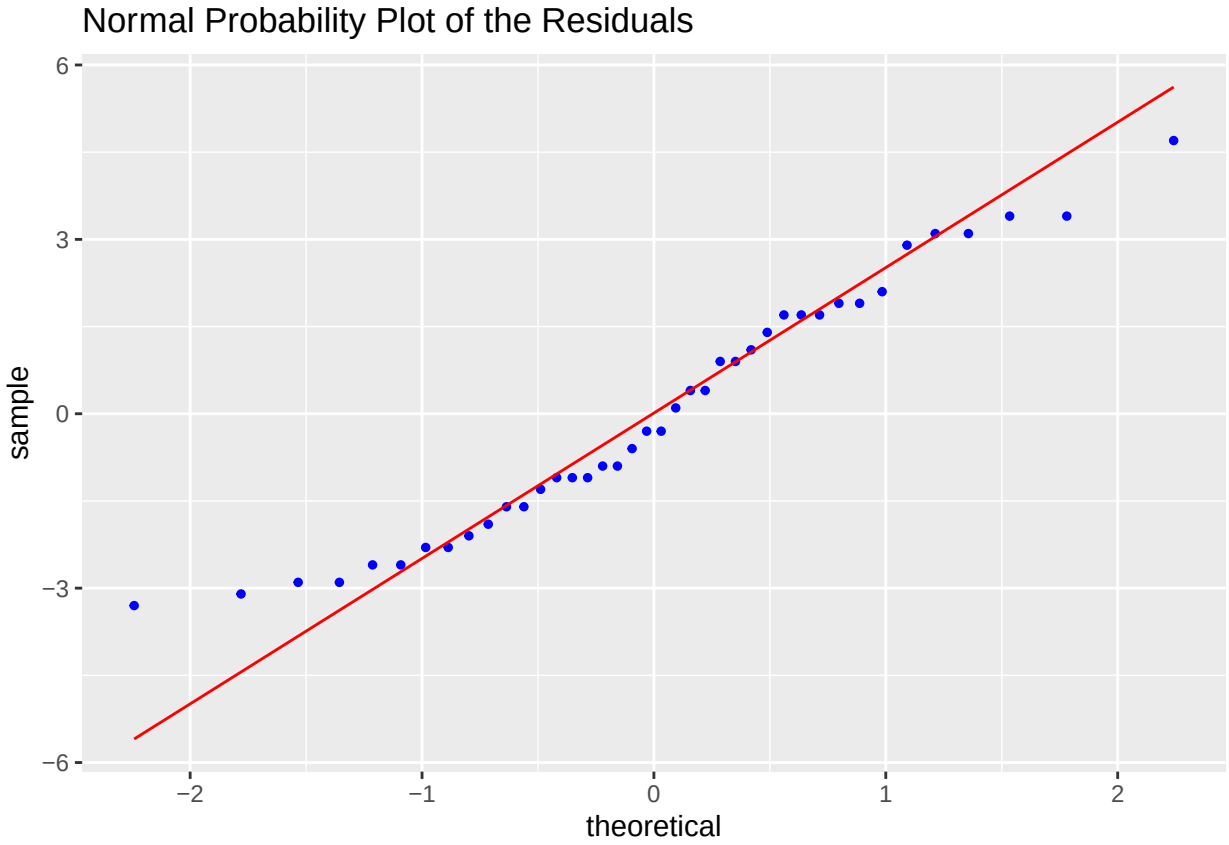
$$F_{obs} = 1.11$$
$$P - value = P(F_{3,26} > F_{obs}) = 0.358$$

Based on this data, the $P - value = P(F_{3,26} > F_{obs}) = 0.358$ which is greater than 0.05. One can conclude that the four populations are equal with respect to the response variable (rental periods). The mean value of the response variable is the same for all these $k = 4$ populations. There exists a 0.358 probability of another experiment producing stronger statistical evidence against the null hypothesis.

(d) Analyze the residuals from this experiment. Are the one-way model assumptions satisfied? Explain.

**Solution:** There are 2 conditions that must be met as outline in 1(c).

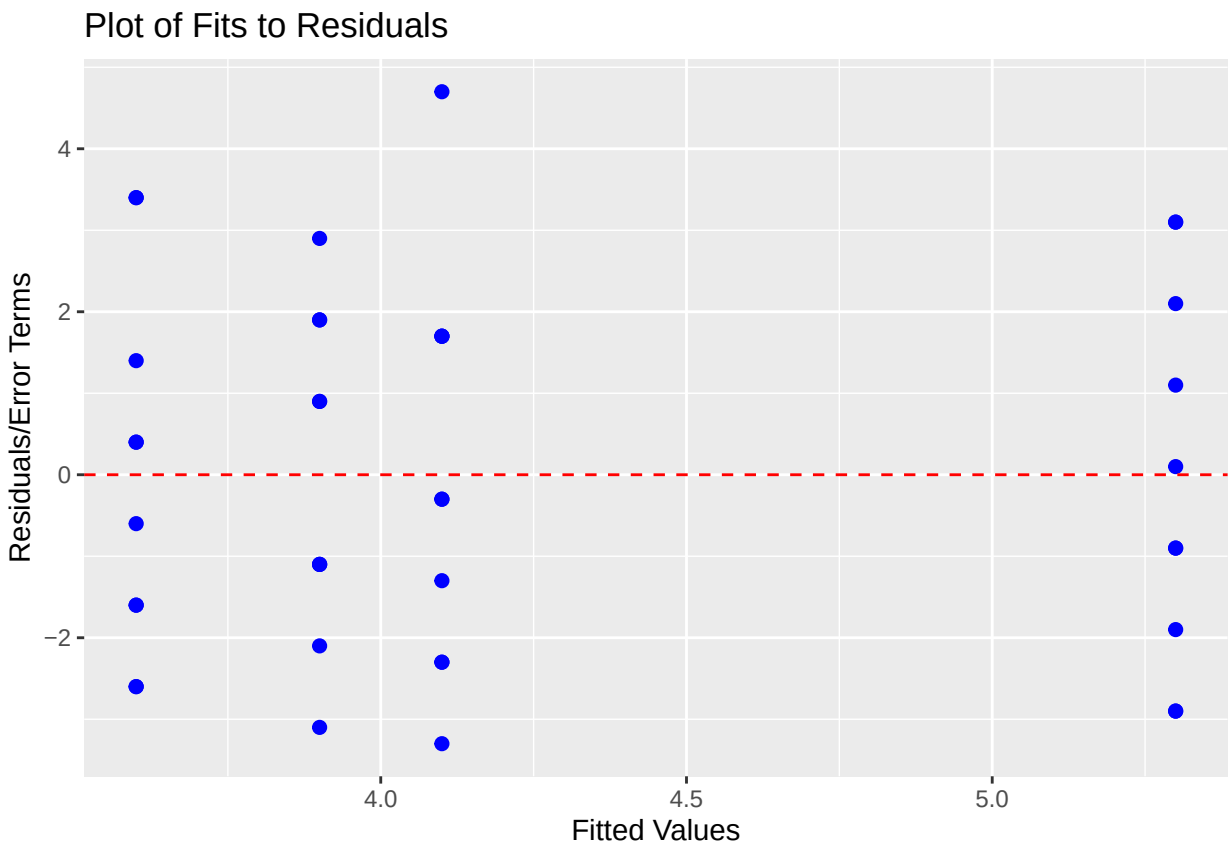
1. Normal distribution of residuals




```
##
##  Shapiro-Wilk normality test
##
## data:  p2$ei_terms
## W = 0.95761, p-value = 0.1386
```

From the qq plot above, the residuals seem to be generally normally distributed. However, the tails seem to deviate too much. In order to check if the residuals are indeed are normally distributed, we check the Shapiro-Wilk test. We get that the p-value=0.1386 > 0.05.

2. Homoscedasticity




The residual plot shows no distinct pattern. As well, the points are balanced around the horizontal line equal to 0. Thus, the variance between the groups are indeed true.

(e) Consider your comments from part (a): Carry out the appropriate test to that will determine if the variation in these data is the same for the four different types of cars

**Solution:** In order to check if the variances are the same for all populations, we perform an Levene's test by the one-way ANOVA of $|X_{ij} - \tilde{X}_i|$ for $i =$ Subcompact,Midisize, Full size, and Compact. We have the following hypothesis:

$$H_0 : \sigma_{Subcompact} = \sigma_{Midsize} = \sigma_{FullSize} = \sigma_{Compact} = \sigma_{Common}$$
$$H_a : \text{at least one } \sigma \text{ is different for all } k = 4 \text{ populations}$$

```
a <- favstats(period~type, data=p2)$median
p2$median <- c(rep(a[4],10), rep(a[1],10), rep(a[3],10), rep(a[2],10))
p2$absdiff <- abs(p2$period - p2$median)
summary(aov(absdiff~type, data=p2))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## type         3   0.28  0.0917   0.067  0.977
## Residuals   36  49.50  1.3750
```

From the above, we get that

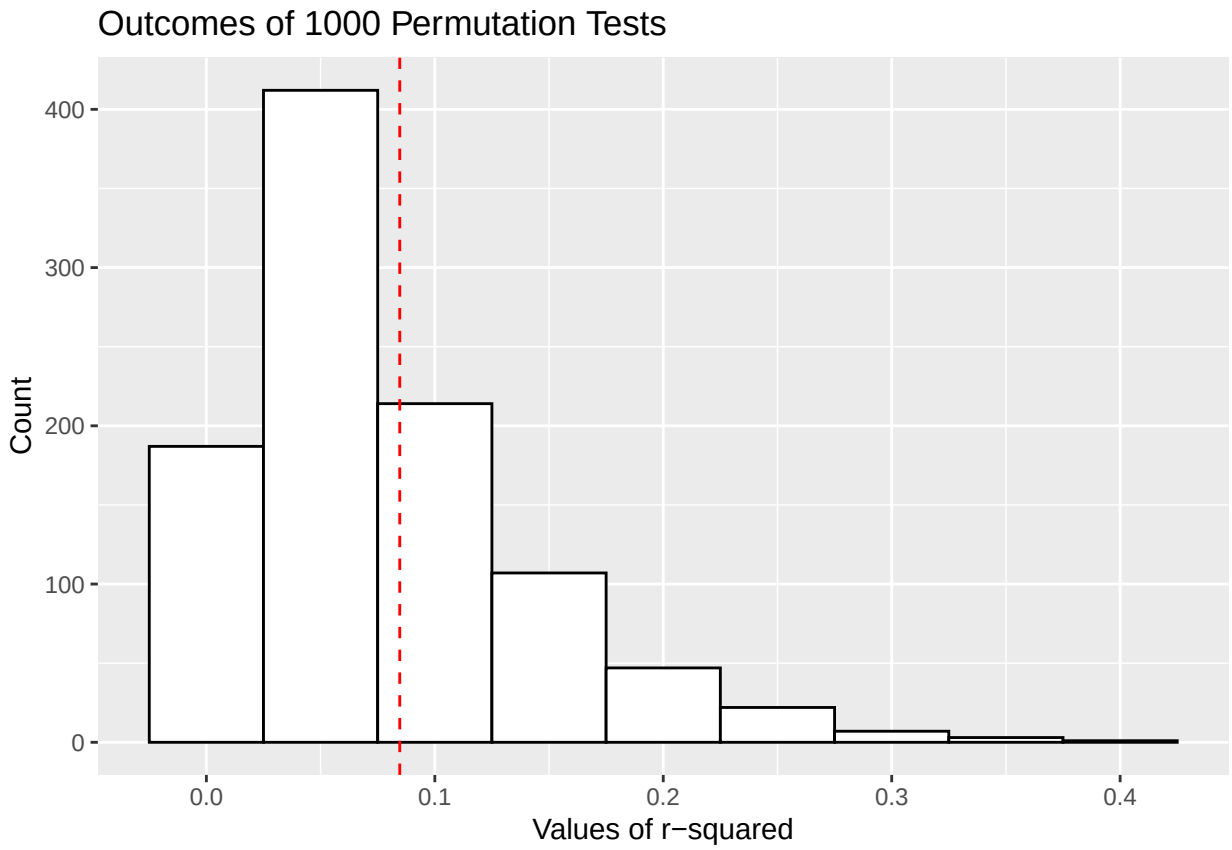
$$F_{obs} = 0.067$$
$$P - value = P(F_{3,36} > F_{obs}) = 0.977$$

Based on the data, the $P - value = P(F_{3,36} > F_{obs}) = 0.977$ which is greater than 0.05 and so we do not reject the null hypothesis. The standard deviation in the period of rental for each type of car is the same; there exists a 0.977 probability of another experiment producing stronger statistical evidence against the null hypothesis.

(f) Using 1000 iterations or replications, carry out a condition-free statistical test to these data. In your summary of your result(s), ensure you indicate (i) the observed result and (ii) your empirical P-value. (iii) Are you results consistent with your findings in part (c)?

**Solution:** We produce the following histogram:

```
set.seed(1)
demopermtest1000.df = do(1000) * rsquared(lm(period ~ shuffle(type), data=p2))
obsrsquared = rsquared(lm(period~type, data=p2))
ggplot(data=demopermtest1000.df, aes(x = rsquared)) + geom_histogram(col="black", fill="white", binw
```




(i) Using the do function from the Mosaic package, we perform 1000 iterations of the test to these data. Based on the data, we will super impose the $r^2 = 0.08465541$ value to the histogram plot. We observe that our observed $r^2$ is quite commonly observed.

(ii) The empirical p-value can be calculated through the following:

```
howmany = sum(demopermtest1000.df$rsquared > obsrsquared)
emp_P_value = (howmany/1000)
emp_P_value
```

```
## [1] 0.349
```

From the above, we get that the empirical p-value = 0.349 This means that 349 tests out of 1000 of permutation tests produced a value of $r^2$ that exceeded 0.0847.

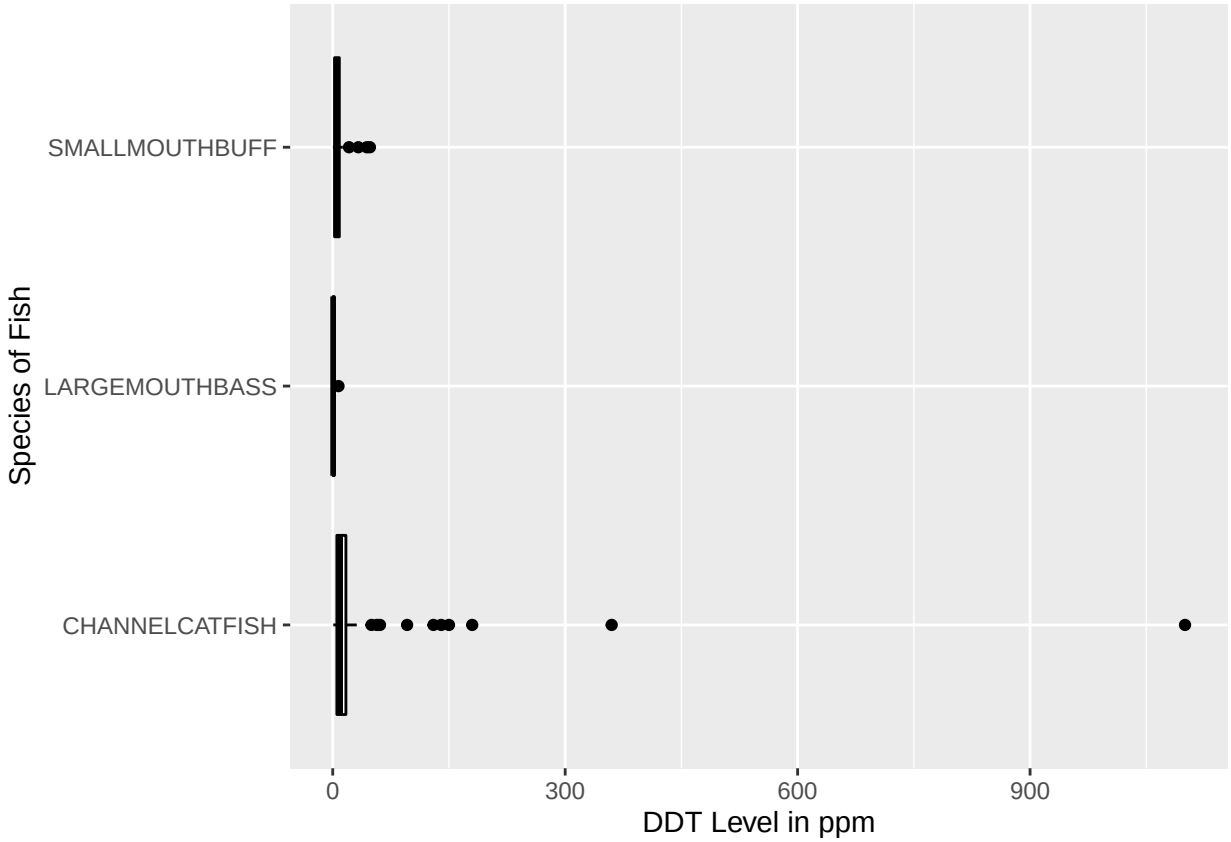
(iii) Finally, this result is consisten with the results from (c). The p-value from (c) is $p - value_c = 0.358$ and the p-value from (f) is $p - value_f = 0.384$. These two values are close to each other. As well, the $r^2$ value from the data is quite common from the histogram above.

3. The U.S. Army Corps of Engineers collected data on the DDT levels of three different species of fish: channel catfish, largemouth bass, and smallmouth buffalofish. The DDT levels were measured in PPM (parts per million) on 144 captured fish. The data is found in the FISHDDT.csv file in the directory:

   (a) Provide a boxplot and commentary of the DDT levels comparing the distribution of DDT levels amongst the three different types of fish. If necessary, provide the test statistic and P-value of the statistical test that compares the variation in the DDT levels amongst the three different species of fish.

   **Solution:**

```
p3 <- read.csv("p3.csv")
p3 <- p3[order(p3$SPECIES),]
ggplot(data=p3, aes(x = SPECIES, y = DDT)) + geom_boxplot(col="black", fill="white") + xlab("Species
```




From the boxplot above, it is difficult to make any conclusion concerning the variance between the populations to be the same or not. Hence, we perform a Levene's test using a one-way ANOVA on the $|X_{ij} - \tilde{X}_i|$ for $i =$ Channel Cat Fish, Small Mouth Buff, and Large Mouth Buff. We have the following hypothesis where $X_{ij}$ is the DDT levels in ppm for the specified fish and $\tilde{X}_i$ is the median for that species of fish:

$$H_0 : \sigma_{ChannelCatFish} = \sigma_{SmallMouthBuff} = \sigma_{LargeMouthBuff} = \sigma_{Common}$$
$$H_a : \text{at least one } \sigma \text{ is different for all } k = 3 \text{ species}$$

```
a <- favstats(DDT~SPECIES, data=p3)$median
p3$median <- c(rep(a[1],96), rep(a[2],12), rep(a[3],36))
p3$absdiff <- abs(p3$DDT - p3$median)
summary(aov(absdiff~SPECIES, data=p3))
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## SPECIES       2   17417    8709   0.917  0.402
## Residuals   141 1338770    9495
```

From the above, we get that

$$F_{obs} = 0.917$$
$$P - value = P(F_{2,141} > F_{obs}) = 0.402$$

Based on this data, the $P-value = P(F_{2,141} > F_{obs}) = 0.402$ which is greater than 0.05 and so we do not reject the null hypothesis. The standard deviation in the DDT levels of each species of fish is the same; there exists a 0.402 probability of another experiment producing stronger statistical evidence against the null hypothesis.

(b) Does this data indicate that one species of fish has higher DDT levels than any other species, on average? Use $\alpha = 0.05$.

**Solution:** We wish to perform an ANOVA test. For the sake of the of procedure, we will assume that the residual normality condition has been met. And so, we test the following hypothesis

$$H_0 : \mu_{ChannelCatFish} = \mu_{SmallMouthBuff} = \mu_{LargeMouthBuff}$$
$$H_a : \text{at least one of the k=3 species is not the same}$$

Using R to perform the test:

```
summary(aov(DDT~SPECIES, data=p3))
```

```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## SPECIES        2   23454   11727   1.215    0.3
## Residuals    141 1360549    9649
```

From which, we get

$$F_{obs} = 1.22$$
$$P - value = P(F_{2,141} > F_{obs}) = 0.3$$

Based on this data, the $P-value = P(F_{2,141} > F_{obs}) = 0.3$ which is greater than 0.05 and so we do not reject the null hypothesis. One can conclude that all the species are the same in terms of the mean DDT levels in ppm. There exists a 0.3 probability of another experiment producing stronger statistical evidence against the null hypothesis.

(c) If the null hypothesis in (b) is not rejected, construct a 95% confidence interval for $\mu$ - the DDT level in the fish. If the null hypothesis in (b) is rejected, construct a 95% confidence interval estimates for
(i)$\mu_{ChannelCatFish}$ (ii) $\mu_{SmallMouthBuff}$ (iii) $\mu_{LargeMouthBuff}$

**Solution:** Frome (b) we do not reject the null hypothesis. Hence, we use the following formula for the 95% confidence interval of the mean DDT levels:

$$\overline{X} \pm t_{0.025,n-1}\sqrt{\frac{MSE}{n}}$$

Using, R to compute the intervals we get:

```
mse <- 9649
mean(p3$DDT)+ c(-1,1)*abs(qt(0.05/2, 144-1)*(sqrt(mse/144)))
```

```
## [1]  8.174239 40.535761
```

Based on this data, the 95% confidence interval for $\mu$ is between 8.1742 and 40.5358. This means the if we perform this experiment many times, 95% of the DDT levels in ppm for this species calculated will be captured by this interval.

(d) With an experiment error rate of 0.05, carry out the appropriate multiple-comparison method to that will identify which species has the highest DDT level.

**Solution:** We will perform a Bonferroni multiple comparison method with an error rate=0.05 using the R code below:

```r
PostHocTest(aov(DDT ~ SPECIES, data=p3), method="bonferroni", conf.level=0.95)
```

```
##
##   Posthoc multiple comparisons of means : Bonferroni
##     95% family-wise confidence level
##
## $SPECIES
##                                      diff     lwr.ci   upr.ci   pval
## LARGEMOUTHBASS-CHANNELCATFISH -31.919375 -104.79207 40.95332 0.8712
## SMALLMOUTHBUFF-CHANNELCATFISH -25.137708  -71.65124 21.37582 0.5776
## SMALLMOUTHBUFF-LARGEMOUTHBASS   6.781667  -72.55208 86.11541 1.0000
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will summarize the results in a table below:

| Lower Bound | $\mu_i - \mu_j$ | Upper Bound | Finding |
|---|---|---|---|
| $-104.79$ | $\mu_{LargeMouthBass} - \mu_{ChannelCatFish}$ | 40.95 | $\mu_{LargeMouthBass} = \mu_{ChannelCatFish}$ |
| $-71.65$ | $\mu_{SmallMouthBuff} - \mu_{ChannelCatFIsh}$ | 21.38 | $\mu_{SmallMouthBuff} = \mu_{ChannelCatFIsh}$ |
| $-72.55$ | $\mu_{SmallMoutBuff} - \mu_{LargeMouthBass}$ | 86.12 | $\mu_{SmallMoutBuff} = \mu_{LargeMouthBass}$ |

From the summary above, we see can conclude that:

$$\mu_{LargeMouthBass} = \mu_{ChannelCatFish} = \mu_{SmallMouthBuff}$$

Based on the data, there is no specific specific species with the highest DDT levels in ppm. This is consistent with our result in (b).

4. Can a secondary task, like a word association task, improve your performance when you are fatigued? Data from the a study appearing in Human Factors resulted from researchers using a driving simulation experiment. Each of n = 40 students was randomly assigned to one of four groups, then each student was to simulate driving long-distance (the same distance) in a driving simulator. Students assigned to Group 1 performed a verbal task continuously (continuous verbal). Students assigned to a Group 2 performed the verbal task during only at the end of their drive (end verbal); students assigned to Group 3 did not perform the task at all (no verbal communication), and students assigned to Group 4 listed to a certain radio program while driving the long-distance (radio show condition).

   At the end of the trial, each student was asked to recall billboards they they saw along the way. The response variable measured was the percentage of billboards recalled by each student-driver. These data can be found in the FATIGUE.csv file found in the same data file directory as indicated in Question 3. Read these data into R/R Studio and answer the following questions:

   (a) Suppose the response variable in this case is Normally distributed, and the standard deviation in the percentage of billboards recalled is the same for the four groups, can you conclude from these data that there is no difference in the mean percentage recall for student-drivers in the four groups?
   **Solution:** Assuming the condtions are met. We perform an ANOVA test with the following hypothesis:

$$H_0 : \mu_{continuousverbal} = \mu_{EndVerbal} = \mu_{NoVerbal} = \mu_{Radio}$$
$$H_a : \text{at least one } \mu \text{ is not the same for the } k = 4 \text{ groups}$$

   Using R, we get that:

```
p4 <- read.csv("p4.csv")
summary(aov(RECALL~GROUP, data=p4))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## GROUP        3   5922  1973.9   5.388 0.00362 **
## Residuals   36  13189   366.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   From the ANOVA table above we get the following results:

$$F_{obs} = 5.388$$
$$P - value = P(F_{3,36} > F_{obs}) = 0.00362$$

   Based on this data, the $P - value = P(F_{3,36} > F_{obs}) = 0.00362$ which is less than 0.05 and so we reject the null hypothesis. One can conclude that the three populations are equal with respect to the response variable (percent billboards recalled). The mean value of the percent billboards recalled is the different for at least one of these $k = 4$ groups There exists a 0.00362 probability of another experiment producing stronger statistical evidence against the null hypothesis.

   (b) Do these data indicate that the distribution of the percentage of billboards recalled is the same for the four groups? Carry out the appropriate statistical test, showing all relevant work. What can you infer?
   **Solution:** In order to test for the similar distributions, we perform a Kruskal-Wallis test with the following hypothesis:

$H_0$ : The distribution of the percent billboards recalled are the same for all k=4 groups
$H_a$ : Not all k=4 groups are distributated the same

   Using R to perform the test:

```
kruskal.test(RECALL ~ GROUP, data=p4)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  RECALL by GROUP
## Kruskal-Wallis chi-squared = 12.846, df = 3, p-value = 0.004983
```

From which, we get

$$KW_{obs} = 12.846$$
$$P - value = P(\chi^2_2 > KW_{obs}) = 0.004983$$

Based on this data, the $P - value = P(\chi^2_2 > KW_{obs}) = 0.004983$ which is less than 0.05 and so we reject the null hypothesis. One can conclude that at least one group is not distributed the same in terms of the percent billboards recalled. There exists a 0.004983 probability of another experiment producing stronger statistical evidence against the null hypothesis.

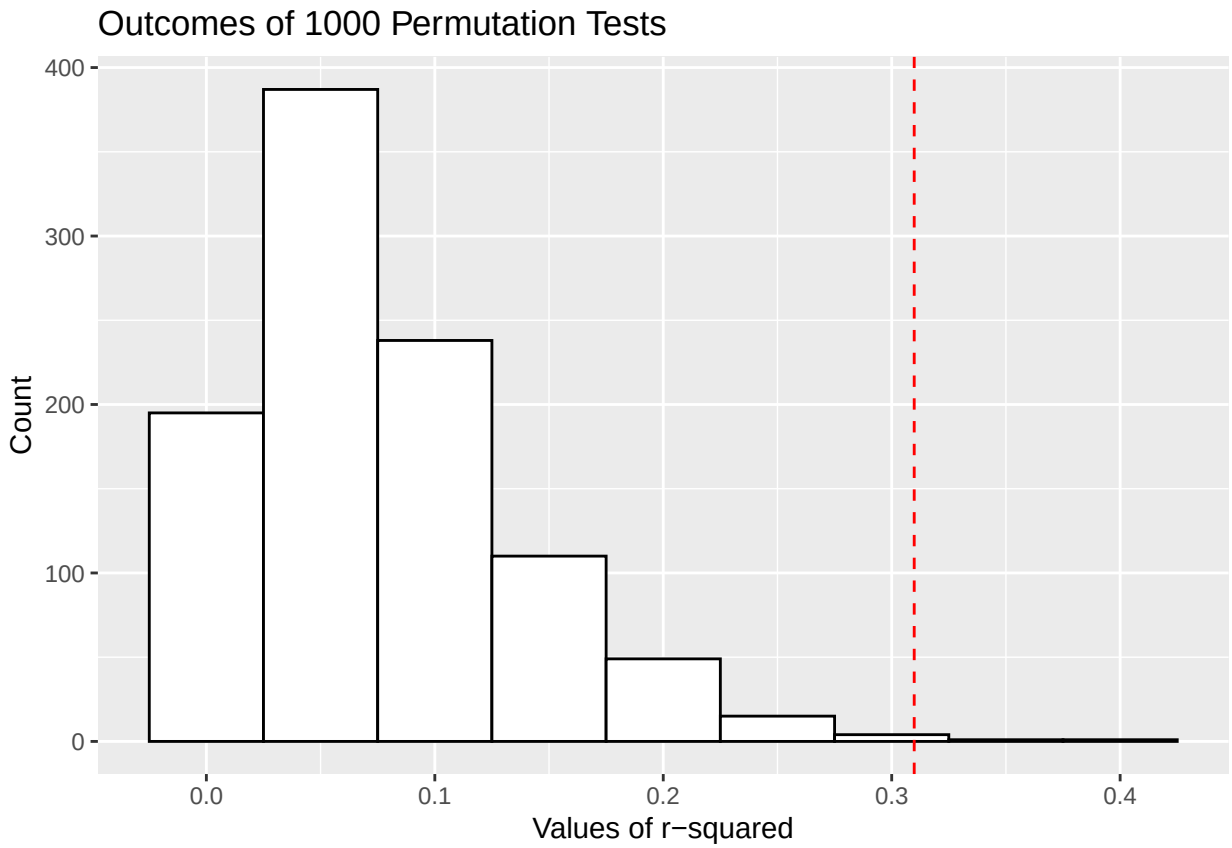
(c) Similar to Question 2(f), conduct a permutation test. Use 1000 iterations/replications. Ensure you provide (i) your code and (ii) your empirical P-value. What can you infer?

**Solution:** We produce the following histogram:

```
set.seed(1)
demopermtest1000.df = do(1000) * rsquared(lm(RECALL ~ shuffle(GROUP), data=p4))
obsrsquared = rsquared(lm(RECALL~GROUP, data=p4))
obsrsquared
```

```
## [1] 0.3098566
```

```
ggplot(data=demopermtest1000.df, aes(x = rsquared)) + geom_histogram(col="black", fill="white", binw
```




(i) Using the do function from the Mosaic package, we perform 1000 iterations of the test to these data. Based on the data, we will super impose the $r^2 = 0.3098566$ value to the histogram plot. We observe that our observed $r^2$ is not that commonly observed.

(ii) The empirical p-value can be calculated through the following:

```
howmany = sum(demopermtest1000.df$rsquared > obsrsquared)
emp_P_value = (howmany/1000)
emp_P_value
```

```
## [1] 0.002
```

From the above, we get that the empirical p-value = 0.002. This means that only 2 tests out of 1000 of permutation tests produced a value of $r^2$ that exceeded 0.3098566. This result is consistent with the results from (a). The p-value from (a) is $p - value_a = 0.004683$ and the p-value from (c) is $p - value_c = 0.002$. These two values are close to each other. As well, the $r^2$ value from the data is not that common from the histogram above.

(d) Return to part (a): Given what you know about these data, carry out the appropriate multiplecomparison method to identify which treatment results in the "best" proportion of billboards remembered. Summarize your results.

**Solution:** We will perforam a Tukey's Honestly Significant difference multiple comparison test:

```
PostHocTest(aov(RECALL ~ GROUP, data=p4), method="hsd", conf.level=0.95, ordered = TRUE)
```

```
##
##   Posthoc multiple comparisons of means : Tukey HSD
##     95% family-wise confidence level
##     factor levels have been ordered
##
## $GROUP
##                    diff     lwr.ci   upr.ci   pval
## Radio-ContVerb     14.7  -8.354157 37.75416 0.3298
## LateVerb-ContVerb 29.1   6.045843 52.15416 0.0086 **
## NoVerb-ContVerb   29.6   6.545843 52.65416 0.0074 **
## LateVerb-Radio     14.4  -8.654157 37.45416 0.3476
## NoVerb-Radio       14.9  -8.154157 37.95416 0.3183
## NoVerb-LateVerb     0.5 -22.554157 23.55416 0.9999
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can summarize the results using the table below:

| Lower Bound | $\mu_i - \mu_j$ | Upper Bound | Finding |
|---|---|---|---|
| $-8.35$ | $\mu_{Radio} - \mu_{ContVerb}$ | 37.75 | $\mu_{Radio} = \mu_{ContVerb}$ |
| 6.05 | $\mu_{LateVerb} - \mu_{ContVerb}$ | 52.15 | $\mu_{LateVerb} > \mu_{ContVerb}$ |
| 6.54 | $\mu_{NoVerb} - \mu_{ContVerb}$ | 52.65 | $\mu_{NoVerb} > \mu_{ContVerb}$ |
| $-8.65$ | $\mu_{LateVerb} - \mu_{Radio}$ | 37.45 | $\mu_{LateVerb} = \mu_{Radio}$ |
| $-8.15$ | $\mu_{Noverb} - \mu_{Radio}$ | 37.95 | $\mu_{Noverb} = \mu_{Radio}$ |
| $-22.55$ | $\mu_{NoVerb} - \mu_{LateVerb}$ | 23.55 | $\mu_{NoVerb} = \mu_{LateVerb}$ |

From the summary above, we see can conclude that:

$$\mu_{NoVerb} = \mu_{LateVerb} > \mu_{ContVerb} = \mu_{Radio};$$
$$\mu_{Radio} = \mu_{LateVerb};$$
$$\mu_{Radio} = \mu_{NoVerb}$$

Based on the data and the table above derived from the Bonferroni multiple comparison method, we get that no verbal communication and late verbal communication are the treatment methods that provide the "best" proportion of billboards recalled. It is hard to determine which method is better than the other since the mean difference is statistically the 0.

5. Do television shows with violence and sex impair memory for commercials? Investigators publishing in the Journal of Applied Psychology carried out an experiment where 324 subjects were randomly assigned to one of three viewer groups. One group watched a television show with violent content code (or a V-rating); the second group viewed a show with a six content code (or a S-rating), the third group watched a television show that was rated as "general". Nine commercials were embedded into each television show. After viewing the program, each subject was scored on their recall of the brand names from the nine commercials, with scores ranging from 0 to 9. The data can be found in the TVADS.csv. Does the data suggest that TV shows with violent content and sexual content impair memory for commercials? For Question 5, I want a one-page report/summary. This summary should include:

    1. Visualizations of these data, along with appropriate commentary associated with each visualization. For example, why are you creating boxplots? What do the boxplots suggest?
    2. An application of two different statistical methods, each with their associated statistical hypotheses, computed values of test statistic and P-value.
    3. Should the null hypothesis be rejected, a deeper analysis into "why" the null hypothesis is rejected. That is, if the data do suggest that violent content and sexual content in television shows do impair memory for commercials, is the effect on the response variable the same for these two types of content? Do persons have better memory of commercials embedded in television shows with violent content or sexual content, or is their memory worse?

    **Solution:** Report found on next page.

# Do Television Shows With Violence and Sex Impair Memory for Commercials?

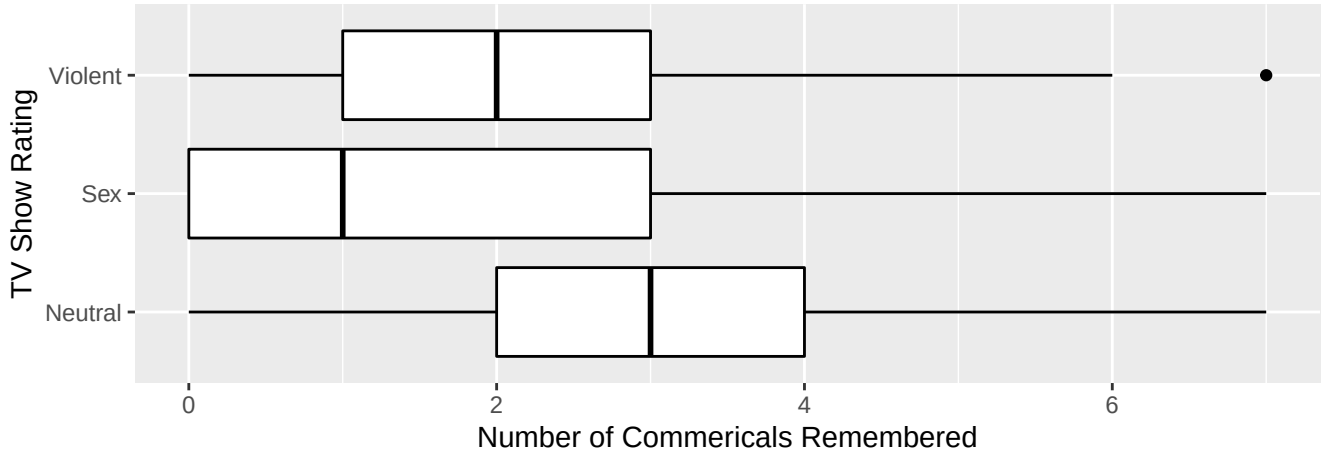
By: Jana Osea

## Introduction

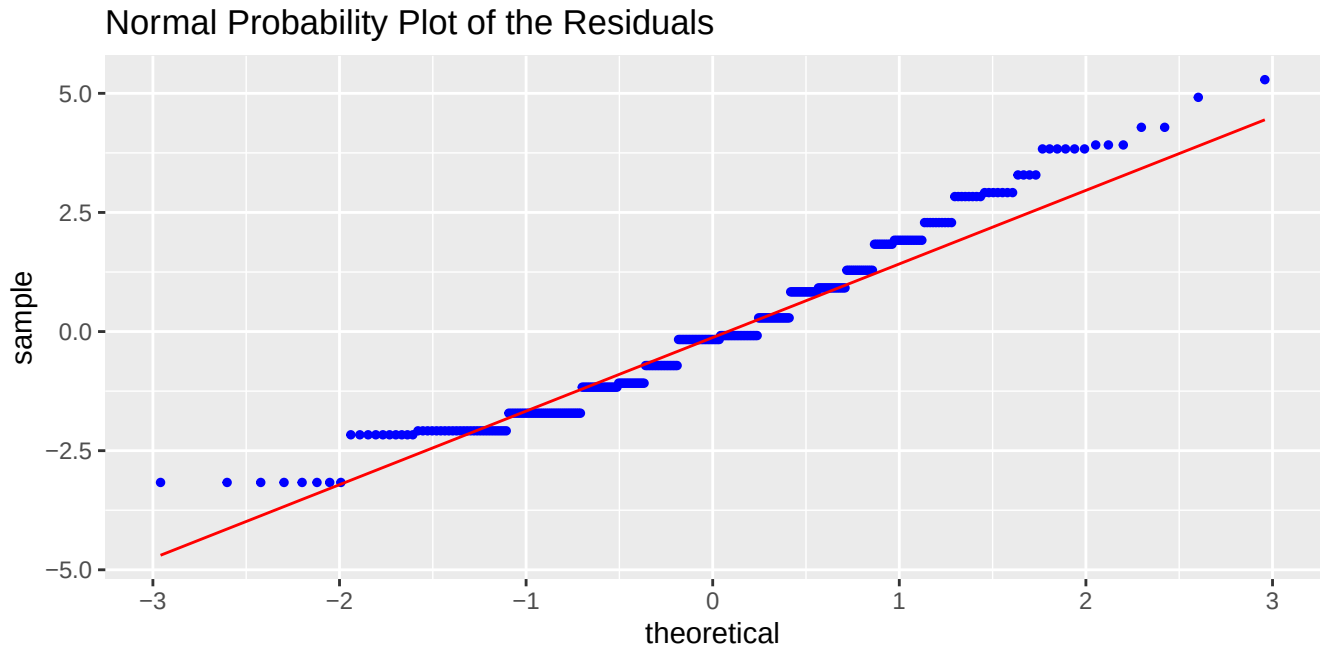
Invetigators from the *Journal of Applied Psychology* wanted to test whether TV shows with violent and sexual content impair memory for commercials. The study design description can be found on the last page of the assignment sheet. Our goal is to test whether there is a difference between all the different groups and if there exists a difference, we want to test what is the treatement increase memory scores.

## Conditions

1. Equal Variance: In order to check for equal variance among each group, we will plot a boxplot as shown below. Since the interquartile ranges seem to be of the same size, we can conlcude that the condition of homoscedasticity has been met.




2. Normality: Next, we must check the assumption of normality of the residuals



```
##
##  Shapiro-Wilk normality test
##
## data:  p5$residuals
## W = 0.96009, p-value = 9.761e-08
```

From the results of the Shapiro-Wilk test ($p - value \approx 0$) and the qq plot above, it is clear that the data residuals do not follow an approximately normal distribution. Hence, we will use a non-parametric method to analyze the data instead.

# Analysis

We will perform two tests of hypothesis:

1. First will perform a Kruskal Wallis test with the following hypothesis:

   $H_0$ : The distribution of the commericals recall score are the same for all k=3 groups
   $H_a$ : Not all k=3 groups are distributated the same

   Using R we get the following:

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SCORE by GROUP
## Kruskal-Wallis chi-squared = 37.153, df = 2, p-value = 8.556e-09
```

   From which, we get the following test statistic and p-value

$$KW_{obs} = 37.153$$
$$P - value = P(\chi_2^2 > KW_{obs}) = 8.556e - 09$$

   Based on this data, the $P - value = P(\chi_2^2 > KW_{obs}) = 8.556e - 09$ which is less than 0.05 and so we reject the null hypothesis. One can conclude that at least one group is not distributed the same in terms of the commercials recalled score. There exists a approximately 0 probability of another experiment producing stronger statistical evidence against the null hypothesis.

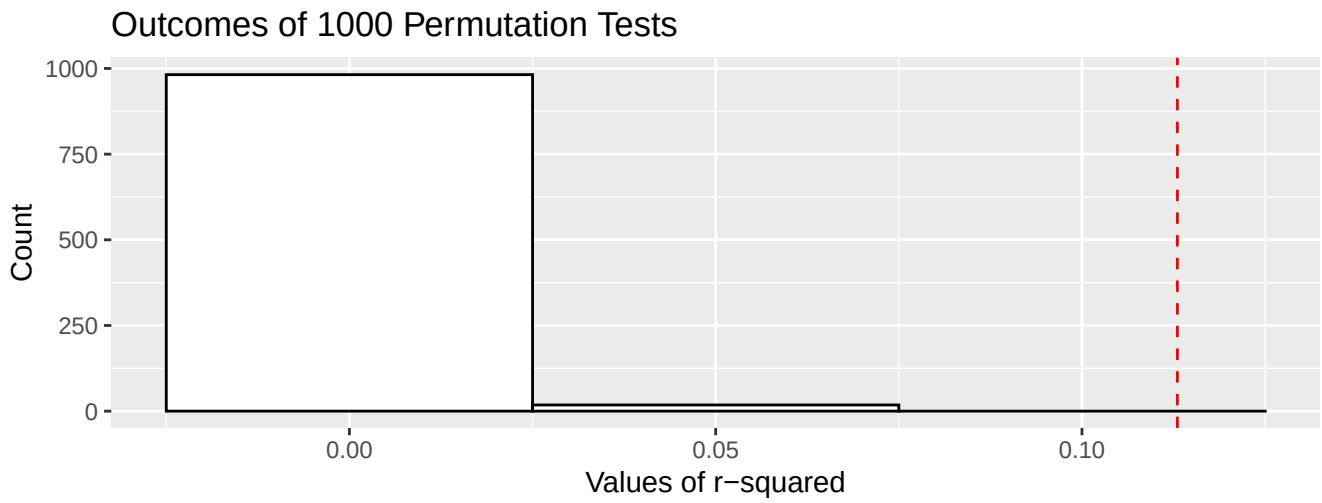
2. Next, we will perform a permutation test with the following hypothesis:

   $H_0 : \mu_{Violent} = \mu_{Sex} = \mu_{Neutral}$
   $H_a$ : at least one $\mu$ is not the same for the $k = 3$ treatment groups

   Using R, we will perform a permutation test:

```r
set.seed(1)
demopermtest1000.df = do(1000) * rsquared(lm(SCORE ~ shuffle(GROUP), data=p5))
obsrsquared = rsquared(lm(SCORE~GROUP, data=p5))
howmany = sum(demopermtest1000.df$rsquared > obsrsquared)
emp_P_value = (howmany/1000)
c(emp_P_value, obsrsquared)
```

```
## [1] 0.0000000 0.1130235
```




From which we get that

$$r^2_{observed} = 0.1130235$$
$$P - value_{empirical} = 0$$

From the above, we get that the empirical p-value = 0. This means that only approximately 0 tests out of 1000 of permutation tests produced a value of $r^2$ that exceeded 0.1130235. This is consistent with the results from the Kruskal-Wallis test above. Based on the data, one can conclude that there is a difference in the mean esponse variable (commercials recalled score) between the violent, sexual, and neutral TV show groups.

## Further Analysis

Since we rejected the global hypothesis that these groups are the same. We are not sure if we should treat the Neutral tv program group as control, so we will perform a Tukey's Honest Significant Different method using the R code below:

```
TukeyHSD(SCORE ~ GROUP, ordered=T, conf.level=0.95, data=p5)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##     factor levels have been ordered
##
## Fit: aov(formula = x)
##
## $GROUP
##                       diff        lwr       upr     p adj
## Violent-Sex     0.3703704 -0.1858756 0.9266163 0.2611283
## Neutral-Sex     1.4537037  0.8974578 2.0099496 0.0000000
## Neutral-Violent 1.0833333  0.5270874 1.6395793 0.0000193
```

From which, we can derive the following table

| Lower Bound | $\mu_i - \mu_j$ | Upper Bound | Finding |
|---|---|---|---|
| $-0.1858756$ | $\mu_{Violent} - \mu_{Sex}$ | 0.9266163 | $\mu_{Violent} = \mu_{Sex}$ |
| 0.8974578 | $\mu_{Neutral} - \mu_{Sex}$ | 2.0099496 | $\mu_{Neutral} > \mu_{Sex}$ |
| 0.5270874 | $\mu_{Neutral} - \mu_{Violent}$ | 1.6395793 | $\mu_{Neutral} > \mu_{Violent}$ |

Based on the data, we can conclude the following:

$$\mu_{Neutral} > \mu_{Violent} = \mu_{Sex}$$

# Conclusion

Based on the data along with all the analysis, we conclude that there is a difference between the mean commercials recall score of the different TV groups with 95% confidence. Furthermore, the group with neutral TV ratings were able to retain the most commercial recall score compared to TV ratings with sex or violence. Interestingly, there is no statistical difference between violent and sexual TV watchers. To conclude, this brings light the fact that the content of TV shows have an impact on our memory and that perhaps TV viewers should be more aware of their TV watching endeavors.