93

Mini-Project 1

Yaretsy Castro

2024-03-29

Introduction:

This survey investigates if there is an app that Smith students are more likely to use depending on the time of day. A student was asked in the survey to rate three apps that they are most likely to use. The apps are social media, streaming app, and email. This project focuses on streaming app.

The null hypotheses is that there is no difference of likeliness of email usage between different times of day. There are three times 8:00 am, 12:00 pm, and 4:00 pm.

-2 Your hypothesis? Ha What did you expect to happen?

Method:

To collect data for the experiment, a survey was given to a group of students in Smith College. The survey was conducted using Qualtrix. The survey randomly assigned each student with a time of day. During the survey, a screenshot with the time of day was given which was also bright blue to signify that the time is during the day. To make sure that was communicated, the time of day was explicitly shown in the question. The student was asked to rate three types of apps, social media, email and streaming app, of likeliness of being opened.

The factor of interest is time. Time has three levels, 8:00 am, 12:00 pm, and 4:00 pm. The experimental unit is a student at Smith. The response variable is rate. Rate is in a scale of 1-10. We will focus on the usage of email.

Writing needs some work/ Fleshing out. -1

Results:

As a reminder our null hypotheses is that there is no difference of likeliness of email usage between different times of day. There are three times 8:00 am, 12:00 pm, and 4:00 pm. ?

First, we will look at what the data looks visually and statistically. We will create a boxplot to visualize the distributions of rate by time. Statistically, we will create a table that has descriptive statistics such as number of observations per time, mean, and standard error.

Secondly, we will conduct an ANOVA test that will give us an F-statistic and p-value. This will tell us if there is evidence to reject or fail to reject the null hypothesis.

Thirdly, we will check the ANOVA conditions. This will tell us if we can rely in our ANOVA inference.

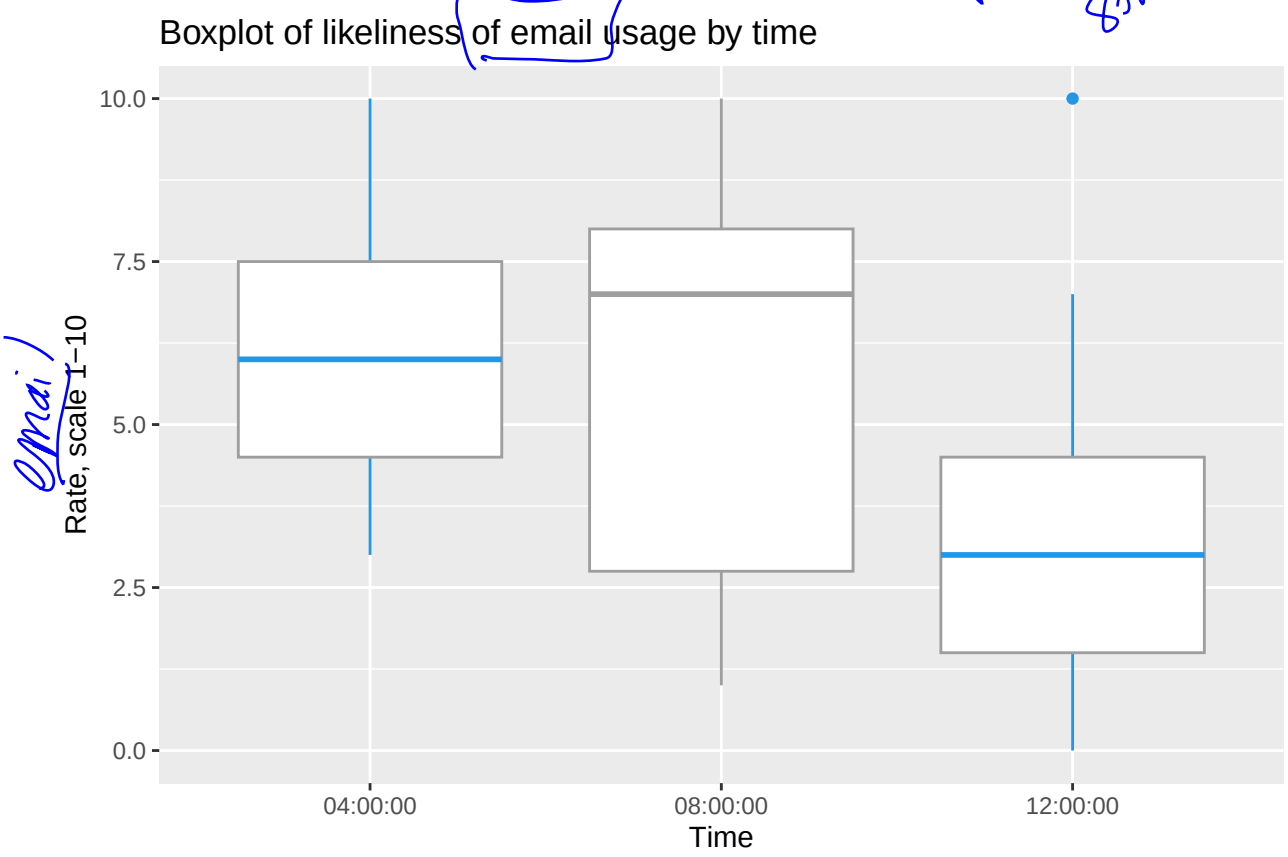
Next, we calculate the effect difference and confidence intervals. This will tells us how much the difference is between groups.

Lastly we will check how much of the variation is explained by time.

Visually

If we look at the graph below, we see that the rate means are different in each time. We can also see that for the most part the distributions are normal. 8:00 am has a skewness at the bottom. 12:00 pm has one outlier in the in at the top.

```
#boxplot visual of email
vis1 <- ggplot(ds, aes(x = as.factor(Time), y = Rate, color = Time)) +
  geom_boxplot() +
  labs(title = "Boxplot of likeliness of email usage by time",
       x = "Time",
       y = "Rate, scale 1-10")
vis1
```



Statistically

If we look at the table we can see that for the most part the experiment is balanced, with 8:00 am having one more observation. When we look at the means, we see that the means are different from each other, 12:00 pm, mean of 3.45, with the most difference from the others.

```
#descriptive stats
stats <- ds%>%
  group_by(Time)%>%
  summarise(count = n(),
            mean = mean(Rate),
```

```
            standard_deviation = sd(Rate))
stats
```

```
## # A tibble: 3 x 4
##   Time    count  mean standard_deviation
##   <time> <int> <dbl>              <dbl>
## 1 04:00     11  6.27               2.20
## 2 08:00     12  5.58               3.09
## 3 12:00     11  3.45               2.94
```

ANOVA

From the ANOVA table below we can see that we have a F-statistic, 5.73, higher than 1 with a p-value, 0.02, lower than 0.05. This means that we have evidence that there is a there are statistically significant differences between times and reject the null hypotheses.

```
#anova model
mod1 <- lm(Rate ~ Time, data = ds)
aov1 <- anova(mod1)
aov1
```

```
## Analysis of Variance Table
##
## Response: Rate
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## Time       1  43.682  43.682  5.7323 0.02269 *
## Residuals 32 243.848   7.620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA Conditions

To check the ANOVA conditions we look at the acronym SINZ. As we go through the conditions, we cans ee that all conditions are met and we can move unto what are the effect sizes and confidence intervals.

Same Standard Deviations To check this this condition, we look at the descriptive statistics again. We check if the highest standard deviation is less than two times than the lowest standard deviation. As we look below, we can see that this condition passes as the lowest sd is 2.20 and the highest is 3.09.

Independence This experiment was conducted at random, so each observation is separate from one another, therefore the independence assumption passes.

```
stats
```

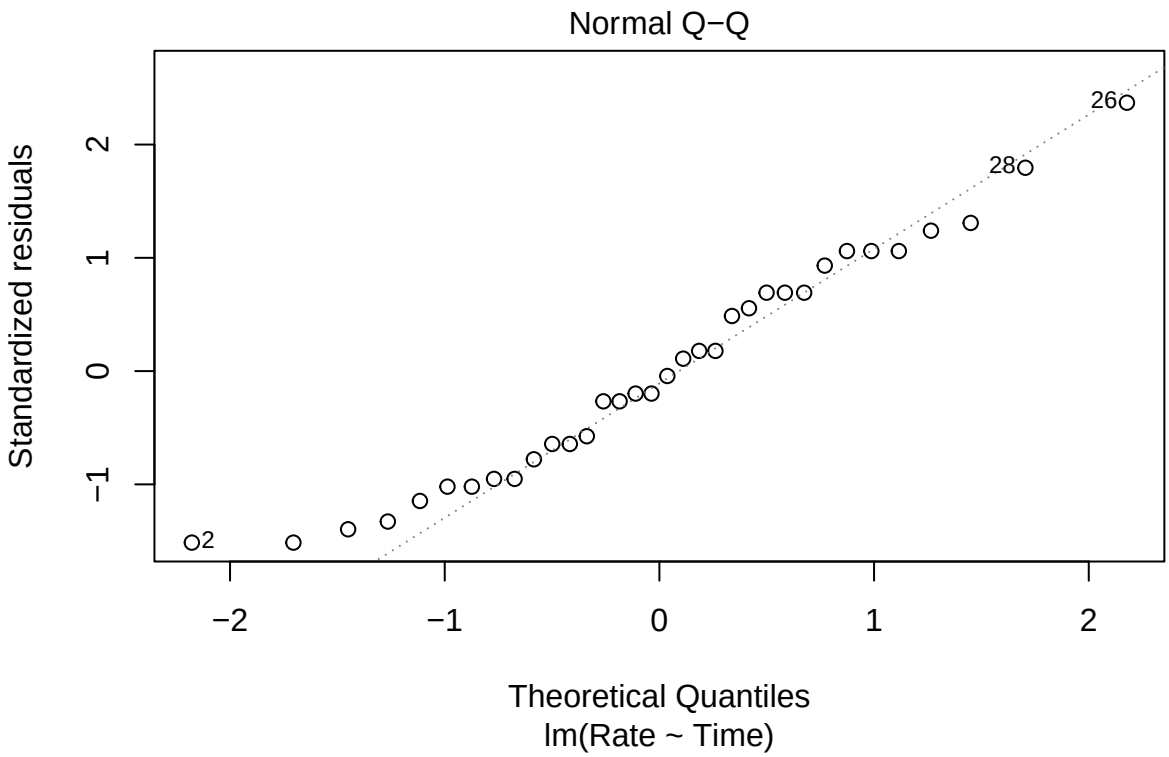
```
## # A tibble: 3 x 4
##   Time    count  mean standard_deviation
##   <time> <int> <dbl>              <dbl>
## 1 04:00     11  6.27               2.20
## 2 08:00     12  5.58               3.09
## 3 12:00     11  3.45               2.94
```

```r
3.09/2.20
```

```
## [1] 1.404545
```

Normality To check this assumption we create a qq-plot that will tells us the distribution of residuals. As we see below the model is mostly normal. Therefore, the model passes the normality condition.

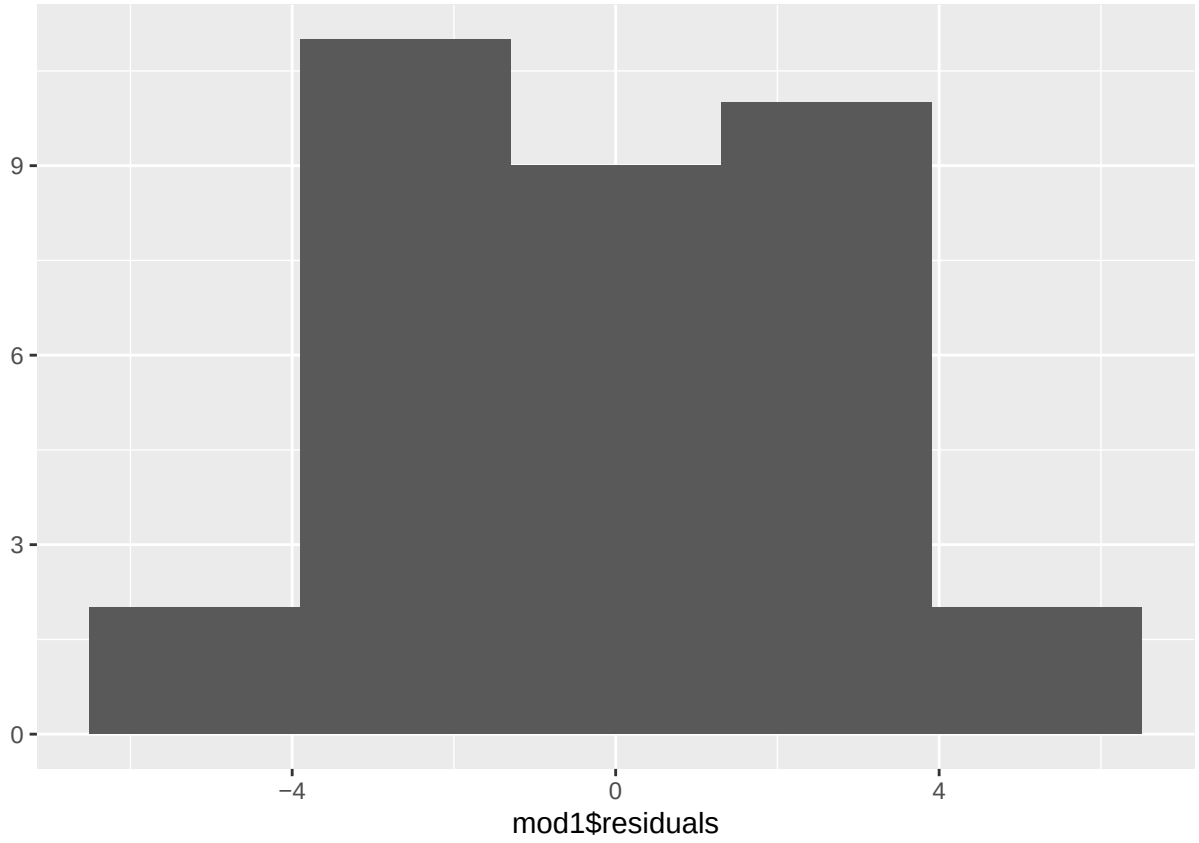
```r
plot(mod1, which = 2)
```



Zero Mean Residuals For this assumption, we create a histogram of the residuals. As we can see below, the residuals are fairly centered at zero.

```r
qplot(mod1$residuals,bins = 5)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Confidence Intervals

We check confidence intervals to check if the group differences are statistically significant different. Two groups are significantly different if the 95% interval does not contain 0. If we look below we can see that the **only one** that is statistically different is between times 12:00 pm and 4:00 pm. There is 95% confidence that the difference of usage rate of email during 12:00 pm and 4:00 pm is between -5.22 and -0.42.

The other two comparisons have 0 in their intervals. Interval between 8:00 am and 12:00 pm: (-0.22, 4.48). Interval between 8:00 am and 4:00 pm: (-3.04, 1.66).

```
MSE = 7.620 #from our ANOVA source table
df_E = 32 #from our ANOVA source table
t <- qt(.975, df_E) #for 95% CI

count_8 <- 12
count_12_4 <- 11

mean_eight <- 5.583333
mean_twelve <- 3.454545
mean_four <- 6.272727
```

```
#Confidence interval for 8:00 am and 12:00 pm
UL_8_12 <- (mean_eight-mean_twelve) + t*sqrt(MSE)*sqrt(1/count_8+1/count_12_4) #upper limit
LL_8_12 <- (mean_eight-mean_twelve) - t*sqrt(MSE)*sqrt(1/count_8+1/count_12_4) #lower limit

LL_8_12
```

```
## [1] -0.2183102
```

```
UL_8_12
```

```
## [1] 4.475886
```

```
#Confidence interval for 8:00 am and 4:00 pm

UL_8_4 <- (mean_eight-mean_four) + t*sqrt(MSE)*sqrt(1/count_8+1/count_12_4) #upper limit
LL_8_4 <- (mean_eight-mean_four) - t*sqrt(MSE)*sqrt(1/count_8+1/count_12_4) #lower limit

LL_8_4
```

```
## [1] -3.036492
```

```
UL_8_4
```

```
## [1] 1.657704
```

```
#Confidence interval for 12:00 pm and 4:00 pm

UL_12_4 <- (mean_twelve-mean_four) + t*sqrt(MSE)*sqrt(1/count_12_4+1/count_12_4) #upper limit
LL_12_4 <- (mean_twelve-mean_four) - t*sqrt(MSE)*sqrt(1/count_12_4+1/count_12_4) #lower limit

LL_12_4
```

```
## [1] -5.215761
```

```
UL_12_4
```

```
## [1] -0.4206028
```

Effect Sizes

As we saw above, there is only a significant difference between 12:00 pm and 4:00 pm. We want to find out how big that difference is, so we calculate it below. To calculate the difference we subtract the means and divide by the square root of MSE.(Mean Square Error)

As we can see below the difference between 12:00 pm and 4:00 pm is -1.02092 times the size of the typical within-group deviations in usage rate of email.

```
#difference between 12:00 pm and 4:00 pm
D = (mean_twelve-mean_four)/sqrt(MSE)

D
```

```
## [1] -1.02092
```

R-squared for One-Way ANOVA

Now we check how much of the variation is explained by time. We calculate this by dividing the Sum Sq of Time divided by the sum of Sum Sq of Time and Residuals.

As we we see below the R^2 is 0.15. This means that 15% of the variation of usage rates for email is explained by time.

```
aov1
```

```
## Analysis of Variance Table
##
## Response: Rate
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## Time       1  43.682  43.682  5.7323 0.02269 *
## Residuals 32 243.848   7.620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
r2 <- 43.682/(43.682+243.848)
r2
```

```
## [1] 0.1519215
```

Conclusion:

Through this experiment, I was attempting to find out if there is a difference of likelyness of app usage depending on time of day. I wanted to focus specifically in the likeliness of email usage. Likeliness is measure using a rating scale of 1-10.

Limitations:

This model was unable to prove that there is a significant difference between 8:00 am & 12:00 pm and 8:00 am & 4:00 pm. The model was also only to explain 15% of variation between rate usage of email.

This data also has certain limitations. It only contains 11 to 12 observations per time. **I would also like to point out that for the variable 4:00 pm, originally the scale was 1-26, due to a typo in the survey.** I re-scaled the data for 4:00 pm from 1-26 to a scale of 1-10.

Findings

There is no evidence to support that there is a difference in likeliness of email usage due to time. There is a possibility that there is a difference between 12:00 pm and 4:00 pm, however this is still not conclusive as the data was not in the correct scale during the collection of data. It also important to note that 15% of the variation of the data is explained by time, which is a small percentage.