ST447 Data Analysis and Statistical Methods (Individual Project)

Candidate Number: 22996

30-11-2023

```r
XYZprofile(ID)
```

```
## The profile of XYZ:
## - Age:  19
## - Gender:  Female
## - Home address:  Hendon (London)
```

Introduction

The above output suggests that XYZ can take his driving test from an either of the test centers at Wood Green (close to LSE) or at Hendon (close to his home). The Problem statement is to suggest the best test center with statistical backing based on data rather than gut feeling. Further, XYZ - without statistical knowledge should be able to understand and replicate this.

Firstly, The initial data cleaning was performed in Excel. Where the data of 19 yaer old females was inserted from these locations with sheet names being Hendon and Wood_Green. Aditionally, I have also utilised the existing sheet summary (Sheet3) covering yearly data of all ages and locations. Owing to the real life knowledge, Data from 2018 was taken into account which is further elaborated as we dwelve further.

Secondly, I have performed the Mann Whitney U Test to test if there is a statistical significant difference in the average pass rate. Further, I have trained two different models on the outcomes based on these two datasets rather than existing pass rates. Further, I ustilised the MSE metric to test the accuracy of my models. Finally, To conduct the Analysis the following assumptions have been made.

- The Data collected from the actual population is random, Each sample is independent ofthe other, Say Location Hendon has no effect on the Wood Green and vice versa. Similarly, each Age group and year are independent and has no influence on other ages and years respectively.
- To Fit Logistic regression model, A Linear relationship and Homoscedasticity in the data is assumed.

```r
#Loading all of the necessary package using lapply.
lapply(c("readxl", "ggplot2", "dplyr", "knitr", "ggpubr", "moments", "gridExtra")
       , require, character.only = TRUE)
```

```r
# Lists all sheet names in the Excel file
sheet_names <- excel_sheets('Book2_copy.xlsx')

# Loops through each of the sheets and create a data frame for each, named after the sheet.
for(sheet in sheet_names) {assign(sheet, read_excel('Book2_copy.xlsx', sheet = sheet))}
Hendon$Location = "Hendon"
Wood_Green$Location= "Wood_Green"
tail(Hendon,2)
```

```
## # A tibble: 2 x 5
##    year Female_Conducted Female_Passes Female_Pass_rate Location
##   <dbl>            <dbl>         <dbl>            <dbl> <chr>
## 1  2010              389           170             43.7 Hendon
## 2  2009              321           137             42.7 Hendon
```

```
tail(Wood_Green,2)
```

```
## # A tibble: 2 x 5
##    year Female_Conducted Female_Passes Female_Pass_rate Location
##   <dbl>            <dbl>         <dbl>            <dbl> <chr>
## 1  2009                7             2             28.6 Wood_Green
## 2  2008              372           123             33.1 Wood_Green
```

```
tail(Sheet3,2)
```

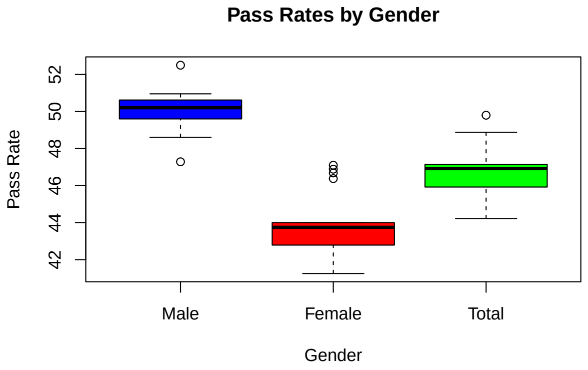
```
## # A tibble: 2 x 10
##    year Male_Conducted Male_Passes Male_Pass_rate Female_Conducted Female_Passes
##   <dbl>          <dbl>       <dbl>          <dbl>            <dbl>         <dbl>
## 1  2022         886928      444828           50.2           801673        371806
## 2  2023         229948      115452           50.2           203511         95420
## # i 4 more variables: Female_Pass_rate <dbl>, Total_Conducted <dbl>,
## #   Total_Passes <dbl>, Total_Pass_rate <dbl>
```

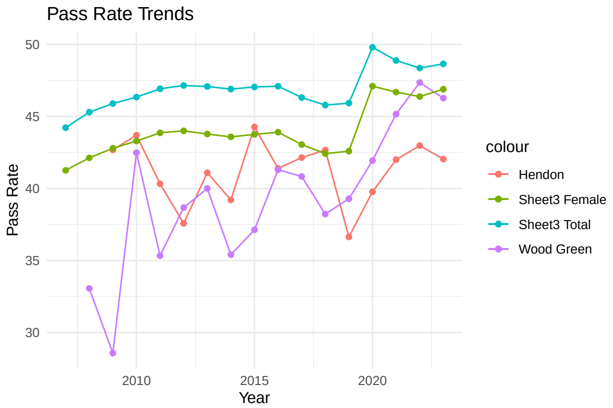
EDA

An attempt has been made to understand the existing data through exploratory data analysis. Fig-1 (Box plots) contains the Male, Female and Total pass rates. Fig-2 (Multiple Line Graph) contains the trends demonstrated by the 19 year olds from both the locations, Female and total pass rates from sheet3.

```
#Plots a box plot from sheet3 (Summary data sheet with all ages and locations)
boxplot(Sheet3$Male_Pass_rate, Sheet3$Female_Pass_rate, Sheet3$Total_Pass_rate,
        names = c("Male", "Female", "Total"), col = c("blue", "red", "green"),
        main = "Pass Rates by Gender", ylab = "Pass Rate", xlab = "Gender")

# Create an empty plot for Multiple line graph
plot<- ggplot()+theme_minimal()+ labs(title="Pass Rate Trends", x ="Year", y ="Pass Rate")

# Add Hendon data to the plot
plot<- plot+ geom_line(data=Hendon,aes(x=year, y=Female_Pass_rate, color ="Hendon")) +
         geom_point(data = Hendon, aes(x = year, y = Female_Pass_rate, color = "Hendon"))

# Add Wood_Green data to the plot
plot<- plot+geom_line(data=Wood_Green, aes(x=year,y=Female_Pass_rate,color="Wood Green"))+
  geom_point(data=Wood_Green, aes(x=year, y=Female_Pass_rate, color = "Wood Green"))

# Add Sheet3_Total data to the plot
plot<- plot+geom_line(data=Sheet3, aes(x=year, y=`Total_Pass_rate`,color="Sheet3 Total"))+
  geom_point(data = Sheet3, aes(x = year, y = `Total_Pass_rate`, color = "Sheet3 Total"))

# Add Sheet3_Female data to the plot
```

```r
plot<- plot+geom_line(data=Sheet3,aes(x=year,y=`Female_Pass_rate`,color="Sheet3 Female"))+
  geom_point(data = Sheet3, aes(x = year, y = `Female_Pass_rate`,color = "Sheet3 Female"))
print(plot) # Display the plot
```





From the above EDA, It is clearly evident that the average female pass rate is considerably below than the Average Male and overall pass rates. The data also exhibit potential outliers (which will be dealt further). Also, The Line graph shows some important trends, with both lines related to sheet3 exhibit similar trend. Further,The data from wood green exhibits an upward trend whereas the Hendon has a stable trend.

Mann Whitney U Test

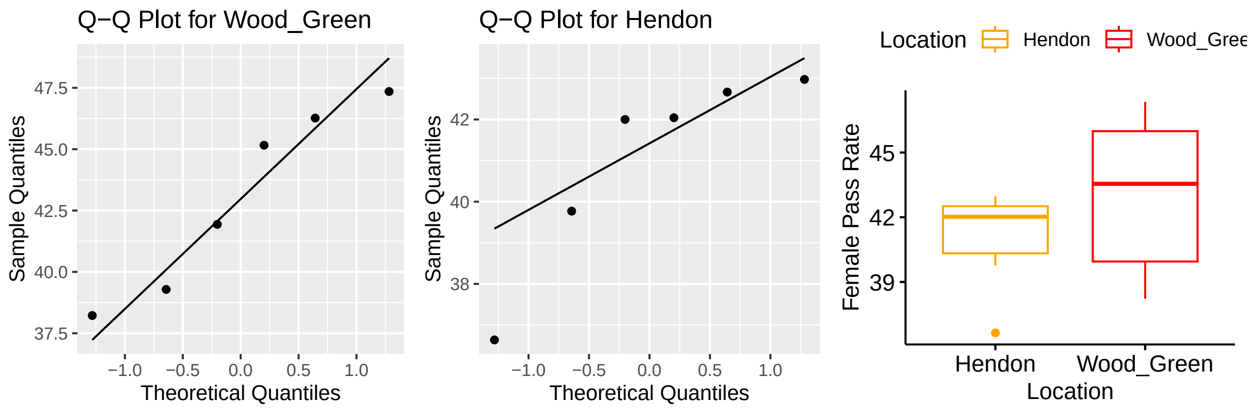
To further test the true difference between the samples, two-sided Mann Whitney U Test has been conducted. It is ideal as compared to the two sampled t test and other tests as the sample data doesn't follow the normal distribution (as seen from the qq plots below) and the sample size is small. Further, The From the Box plot and the summary statistics below it seems like the means of both the distributions are left skewed and have means are close. Which satisfies the essential **requirements** of this test.

Note: As mentioned in the above section, there are some outliers evident from the boxplot as well as the line graph (observations of Wood Green before 2010), which could be due to random occurrence or some clerical mistakes, as the centers back in 2010 are not as digitalised as of now. Further, there has been changes in the driving tests from 4th December 2017. So, In view of mitigating the effect of the driving test change and outliers, I'm conducting further analysis on data after 2018.

```r
# Filtering out the data from 2018 to till date from both locations
Hendon <- Hendon %>% filter(year >= 2018)
Wood_Green <- Wood_Green %>% filter(year >= 2018)
combined_data <- rbind(Hendon, Wood_Green)

cat(paste("Skewness, and Kurtosis for Hendon's Female Pass Rate:",
  "\nSkewness: ", skewness(Hendon$Female_Pass_rate),
  "\nKurtosis: ", kurtosis(Hendon$Female_Pass_rate),
  "\nMean: ", mean(Hendon$Female_Pass_rate),
  "\nMedian: ", median(Hendon$Female_Pass_rate),
  "\n\nSkewness, and Kurtosis for Wood_Green's Female Pass Rate:",
  "\nSkewness: ", skewness(Wood_Green$Female_Pass_rate),
  "\nKurtosis: ", kurtosis(Wood_Green$Female_Pass_rate),
  "\nMean: ", mean(Wood_Green$Female_Pass_rate),
  "\nMedian: ", median(Wood_Green$Female_Pass_rate),sep = ""))
```

```
## Skewness, and Kurtosis for Hendon's Female Pass Rate:
## Skewness: -1.10781635377438
## Kurtosis: 2.75161015006904
## Mean: 41.0154457666111
## Median: 42.021377672209
##
## Skewness, and Kurtosis for Wood_Green's Female Pass Rate:
## Skewness: -0.169093962507012
## Kurtosis: 1.4023173890012
## Mean: 43.0373934658977
## Median: 43.5483870967742
```

```
# Creating a Q-Q plot for Wood_Green
qq_wood_green <- ggplot() + stat_qq(aes(sample = Wood_Green$Female_Pass_rate)) +
  stat_qq_line(aes(sample = Wood_Green$Female_Pass_rate)) +
  labs(title = "Q-Q Plot for Wood_Green", x = "Theoretical Quantiles", y = "Sample Quantiles")

# Creating a Q-Q plot for Hendon
qq_hendon <- ggplot() + stat_qq(aes(sample = Hendon$Female_Pass_rate)) +
  stat_qq_line(aes(sample = Hendon$Female_Pass_rate)) +
  labs(title = "Q-Q Plot for Hendon", x = "Theoretical Quantiles", y = "Sample Quantiles")

# Creating a boxplot
boxplot_combined <- ggboxplot(combined_data, x = "Location", y = "Female_Pass_rate",
      color = "Location", palette = c("#FFA500", "#FF0000")) +
      labs(y = "Female Pass Rate", x = "Location")

# Arrange the plots side by side
grid.arrange(qq_wood_green, qq_hendon, boxplot_combined, ncol = 3)
```



```
mw_test <- wilcox.test(Female_Pass_rate~ Location, data = combined_data,exact = FALSE)
mw_test
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Female_Pass_rate by Location
## W = 14, p-value = 0.5752
## alternative hypothesis: true location shift is not equal to 0
```

Since, with the given test statistic W = 14. We have the respective p-value as 0.5752 which is greater than 0.05. So, We fail to reject the null hypothesis. Which means we do not have sufficient evidence to prove against the null hypothesis that there is no difference in median pass rates between both the locations and any difference observed is due to random chance. Again This could be the case of Type-II error, Due to the nature of the data and small sample sizes, The power to detect a difference between groups can still be influenced by sample sizes and effect size.

So, Even If we try to Predict the expected pass rates using the above data with percentages, it might not yield better outcomes due to limited data and the nature of percentage where 2% pass rate without context can mean 2 people passed the test of 100 test takers as well as 4 people passed the test with 200 test takers. So, Rather than taking the pass percentages to predict taking individual outcomes enhances the actual data.

Data Wrangling

As Discusses above regarding the downside of taking percentages, I'm Further conducting the analysis on the percentages can have potential downside of hiding the actual information by restricting to per 100. So, I chose to backtrack and re-populate the actual observations where each test pass is represented by 1 and fail by 0. It helped to increase the available data to train model on, from 6 in each case to 2183 and 1545 respectively (as we can see from the output of the tail below.)

```
# Defining a function to process each dataframe
process_dataframe <- function(df, transformed_data) {
  for(i in 1:nrow(df)) {
    passes <- rep(1, df$Female_Passes[i])  # Generate successes (1s) and failures (0s)
    fails <- rep(0, df$Female_Conducted[i] - df$Female_Passes[i])
    outcomes <- c(passes, fails)  # Combine the outcomes
    # Create a data frame for this year and Bind the year data to the transformed data
    year_data <- data.frame(year = rep(df$year[i], length(outcomes)), outcome = outcomes)
    transformed_data <- rbind(transformed_data, year_data)}
  return(transformed_data)}

# Initializing empty data frames to store the transformed data with required column dtypes
transformed_data1_Hendon <- data.frame(year = integer(), outcome = integer())
transformed_data2_Wood_Green <- data.frame(year = integer(), outcome = integer())

# storing the transformed data
transformed_data1_Hendon <- process_dataframe(Hendon, transformed_data1_Hendon)
transformed_data2_Wood_Green <- process_dataframe(Wood_Green,transformed_data2_Wood_Green)

tail(transformed_data1_Hendon,2)  # Displays the last two rows
```

```
##      year outcome
## 2182 2018       0
## 2183 2018       0
```

```
tail(transformed_data2_Wood_Green,2)  # Displays the last two rows
```

```
##      year outcome
## 1544 2018       0
## 1545 2018       0
```

Logistic Regression

There are multiple methods to predict the pass-rate ranging from classical linear regression to the state of the art ML models like neural networks. But picking Logistic regression seems to be a sensible decision compared to other models, Given that we are supposed to predict binary outcomes and other methods might result in unbounded and negative probablities as output. Due to the fact that logistic regression handles this issue and bounds the predicted probabilities between 0 and 1 by employing the following **sigmoid function**:

$$P(\text{Pass} = 1|\text{Year}) = \frac{\exp(\beta_0 + \beta_1 \times \text{Year})}{1 + \exp(\beta_0 + \beta_1 \times \text{Year})} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times \text{Year})}}$$

Here, β_0 (intercept) and β_1 (slope of the year) are the parameters obtained by training the model. Which represents the log odds baseline of passing and the effect of year as an independent variable on these odds. Further, The utilization of the following **Log Loss function** serves as an effective way of finding loss between the predicted probabilities $\hat{p}_i$ and the actual binary outcomes y_i.

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)]$$

Further, In most of the cases logistic regression offers similar results or better results, despite being more than 65 year old method, when there is a linear relationship. Further High interpretability, less computational requirements makes it an ideal pick. However, we don't have access to the real life variables such as the temperatures, roads and other demographics. Since, we have more data to train from the Data Wrangling. Two different models were trained rather than one inorder to get more tailored insights.

```r
# Training  two different logistic regression models for Hendon and Wood_Green
m1_Hendon <- glm(outcome ~ ., family = "binomial", transformed_data1_Hendon)
m2_Wood_Green <- glm(outcome ~ ., family = "binomial", transformed_data2_Wood_Green)

summary(m1_Hendon) # Summary of model1_Hendon
```

```
##
## Call:
## glm(formula = outcome ~ ., family = "binomial", data = transformed_data1_Hendon)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -40.35337   47.94824  -0.842    0.400
## year          0.01979    0.02373   0.834    0.404
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2950.7  on 2182  degrees of freedom
## Residual deviance: 2950.0  on 2181  degrees of freedom
## AIC: 2954
##
## Number of Fisher Scoring iterations: 4
```

```r
summary(m2_Wood_Green) # Summary of model2_Wood_Green
```

```
##
## Call:
```

```
## glm(formula = outcome ~ ., family = "binomial", data = transformed_data2_Wood_Green)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -155.87705   58.60978  -2.660  0.00782 **
## year           0.07700    0.02901   2.655  0.00793 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2112.4  on 1544  degrees of freedom
## Residual deviance: 2105.3  on 1543  degrees of freedom
## AIC: 2109.3
##
## Number of Fisher Scoring iterations: 4
```

```r
# Making predictions on the existing data to find MSE
predictions1_Hendon<-predict(m1_Hendon, newdata=transformed_data1_Hendon ,type='response')
predictions2_Wood_Green<-predict(m2_Wood_Green,newdata=transformed_data2_Wood_Green,type='response')

# Evaluate the models using Mean Squared Error (MSE)
mse1_Hendon <- mean((predictions1_Hendon - transformed_data1_Hendon$outcome)^2)
mse2_Wood_Green <- mean((predictions2_Wood_Green - transformed_data2_Wood_Green$outcome)^2)
#MSE Values for both Locations
cat(paste("MSE of Hendon:", mse1_Hendon, ";", "MSE of Wood_Green:", mse2_Wood_Green))
```

```
## MSE of Hendon: 0.241317770867793 ; MSE of Wood_Green: 0.24412889712663
```

From the above results after being trained again on the seen data the MSE scores of both models with Hendon being 0.241 and Wood Green being 0.244 are pretty close. However, The Summaries tells a different story. So, It clearly implies that year variable is not signifcant enough to predict the outcome in Hendon.

```r
# Making predictions for both locations using the for 2023,2024,205
new_data <- data.frame(year = c(2023, 2024, 2025))
new_predictions1_Hendon <- predict(m1_Hendon, newdata = new_data, type = 'response')
new_predictions2_Wood_Green <- predict(m2_Wood_Green, newdata =new_data, type = 'response')

# Displays the predicted pass rates for both locations for years 2023, 2024, 2025
cat("The XYZ's expected passing rate at the nearest test centre to his/her home (Hendon):\n"
, paste(new_predictions1_Hendon, collapse = ", "),"\n
The XYZ's expected passing rate at the nearest test centre to the LSE (Wood Green):\n"
, paste(new_predictions2_Wood_Green, collapse = ","),"\n")
```

```
## The XYZ's expected passing rate at the nearest test centre to his/her home (Hendon):
##  0.420122095016548, 0.424950309485684, 0.429792885407411
##
## The XYZ's expected passing rate at the nearest test centre to the LSE (Wood Green):
##  0.47596322809719,0.495196048457133,0.514443096553275
```

The expected passing rate for both locations for years 2023, 2024, 2025. It turns out that the predictions in case of the Hendon model is pretty much same over the years, partly due to the stable trend that we have observed in the line graph earlier and partly due to the poor significance of the independent year variable. However, the Wood Green predictions are going on an increasing trend just like the line graph.

Conclusion

Of these two locations, where should XYZ take the test? Is there any evidence to (statistically) support this suggestion? The results from the Mann Whitney U Test states there is no significant difference amoung either of the test centers. As Discussed this could be the case of **Type-II error**. As we can see from the the above output, The Expected passing rates from the logistic regresssion model in year 2023 are 42.01% and 47.59% for Hendon and Wood Green respectively. However, From the summary of the Hendon model, The year variable proved to be ineffective to predict.

Based on the given data and this statistical Analysis, I would recommend **"Wood Green"** as

- The Pass rate is already at its peak (as seen in initial line graph) and continue to rise at 2% each year.
- The Year variable proved to be more siginificant in case of Wood Green compared to the Hendon.

In my opinion going to Wood Green will make more sense. Even if the the Mann Whitney U Test is true, Going to the Wood Green is same as the other (**Optimal**) or going by the gut feeling without any solid proof (**Negative**). This is due to the fact that going there in the worst case is same as going to Hendon (**Optimal**). However It can turn out to be the best decision (**Positive**) due to above two Findings. But if he goes to Hendon, It can either turn out **Optimal** or much more **Negative** as there is no statistical Proof.

Note: It seems ideal to postpone test due to rise in pass rate year by year, But no prediction is 100% accurate and future always holds uncertainty. So, It is okay to take the test next year but It is advisable not to postpone too long, As the trend is already at its peak at Wood_Green.

Strengths and Weakness:

- Strengths: Logistic regression always returns Probability between 0 and 1. Multiple methods, More personalized model for each location and Linear regression is known for High interpretability and less computational requirements. Further, Taking the driving change rules into account can possibly serve effectively as the person is going to take the test on these new rules. Finally, EDA also served effectively in obtaining a lot of information on how to structure the approach.

- Weakness: Firstly, Taking the data of two locations and two models might leave some generalized patterns in the data. Secondly, The Hypothesis testing is conducted on the actual percentages. However, To provide more data to the model, The model was trained on the outcomes. Thirdly, Some other ML and non-parametric models like KNN, SVM can fit model with any linear shape assumption. Finally, just like mentioned in the instructions **"It is widely believed that the driving test routes around some centres are probably more difficult than others (e.g. there are far less bus lanes, roundabouts and cyclists in rural areas than in London)."**, Obtaining more information regarding the reviews of the people who took test, Type of car used for the test, Weather conditions can further help to make a better model and provides more context.

References

- https://www.statology.org/mann-whitney-u-test-r/
- https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/

Appendix: The use of generative AI tools

I have utilized ChatGPT to get R markdown code to conserve space, Which displays the plots side by side in EDA and Mann Whitney U Test sections pdf output. However, The actual code to make the plots was done on my own. Further, I confirm it was not used in enhancing the text or explanation.