XYZ Driving Test - Candidate 23961

Introduction

As XYZ prepares to take the practical driving test in the UK, the choice of where to take the test can significantly impact the likelihood of success. In navigating this decision, XYZ has turned to data analysis to explore the passing rates at two potential test centers: one close to her home in Uxbridge, a little village in the suburbs of London, and another near the London School of Economics called Wood Green. In this report, we aim to provide a clear and accessible analysis for XYZ, who is a 17 years old girl, by evaluating the expected passing rates. By utilizing statistical methods, we hope to offer XYZ a data-driven perspective on which location might be more favorable based on historical outcomes of people with a similar profile to her. It's important to note that while statistical analysis can provide valuable insights, individual driving proficiency remains a crucial factor in the ultimate success of the driving test. The available dataset comprises information on individuals who have successfully passed the driving test and those who have undertaken the test, categorized by age and gender. A subset of the last 12 years has been chosen to be coherent with the two test centers given that the Uxbridge one has opened in 2011.

Creation and visualisation of the dataset

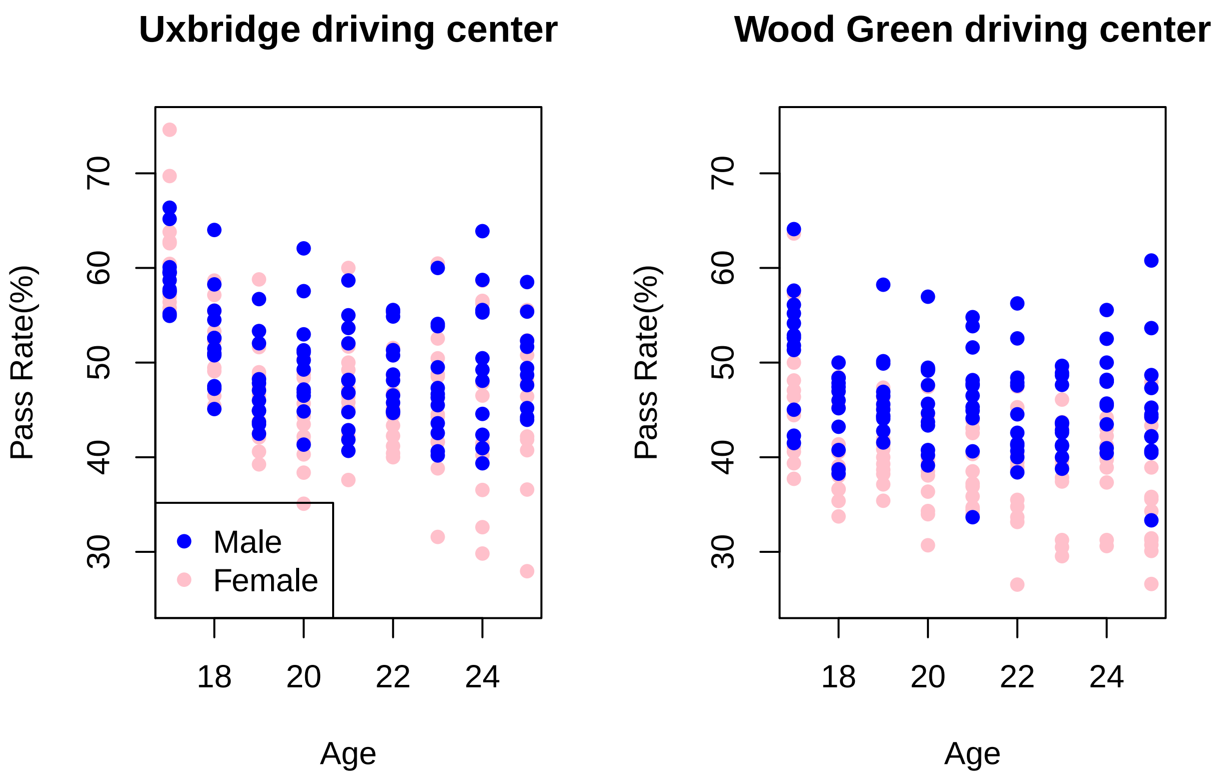
Twelve Excel files have been generated for each driving test center, spanning the years 2012 to 2023. Each file comprises counts of individuals who passed and those who conducted the test, categorized by age (ranging from 17 to 25) and gender. These files have been consolidated into a singular dataset, introducing a new variable labeled 'Year' to denote the respective year from which the data originates.

```r
library(readxl)
Uxbridge = data.frame()
for (i in 12:23){
  df = read_excel(paste('ux20',as.character(i),'.xlsx',sep=""))
  df$Year = as.numeric(paste('20',as.character(i),sep=""))
  Uxbridge = rbind(Uxbridge,df)
} # creation of Uxbridge database
WoodGreen = data.frame()
```

```
for (i in 12:23){
  df = read_excel(paste('wd20',as.character(i),'.xlsx',sep=""))
  df$Year = as.numeric(paste('20',as.character(i),sep=""))
  WoodGreen = rbind(WoodGreen,df)
} # creation of Wood Green database
Uxbridge$Age = as.numeric(Uxbridge$Age)
WoodGreen$Age = as.numeric(WoodGreen$Age) # Age is being transformed in a numeric variable
```

From the dataset contructed like this we can gather valuable insights and create informative plots. Dividing the count of the people that passed by the count of the people that conducted the test we will receive the pass rates. For every age there will be 12 pass rates per gender, one for each year taken into consideration.

```
par(mfrow=c(1,2))
plot(Uxbridge$Age[Uxbridge$Gender=='F'],Uxbridge$Passes[Uxbridge$Gender=='F']/Uxbridge$Conducte
points(Uxbridge$Age[Uxbridge$Gender=='M'],Uxbridge$Passes[Uxbridge$Gender=='M']/Uxbridge$Condu
legend('bottomleft',legend=c('Male','Female'),col = c('blue','pink'),pch=16)
plot(WoodGreen$Age[WoodGreen$Gender=='F'], WoodGreen$Passes[WoodGreen$Gender=='F']/WoodGreen$Co
points(WoodGreen$Age[WoodGreen$Gender=='M'],WoodGreen$Passes[WoodGreen$Gender=='M']/WoodGreen$C
```



Wald test for equality of the means of the pass rates

Dataset with 17 years old girls over all the 12 years

To be coherent with the profile of XYZ this first test will be a historical analysis of 17 years old girls for both driving centres on the last 12 years aggregated.

The variable U is considered to be a $Bin(n_u, \theta_u)$ where θ_u is the probability of passing in Uxbridge and n_u is the numbers of 17 years old girls who took a test there in the last 12 years. W, similarly, is a $Bin(n_w, \theta_w)$ where θ_w is the probability of passing in Wood Green and n_w is the numbers of 17 years old girls who took a test there in the last 12 years. For each person taking a driving test it will be assigned 1 if they pass or 0 if they fail.

The estimate for θ_u is going to be $\hat{\theta}_u = \bar{U}$, the sample mean of the test of 17 years old girls in the last 12 years. Being the sample size n_u quite big and the variance finite, we can say that $\bar{U}$, thanks to the Central Limit Theorem, is distributed as a $N(\theta_u, \frac{\theta_u(1-\theta_u)}{n_u})$ (and similarly for W).

```
Uxbridge_17F = Uxbridge[Uxbridge$Age == 17 & Uxbridge$Gender == 'F' ,]
WoodGreen_17F = WoodGreen[WoodGreen$Age == 17 & WoodGreen$Gender == 'F',]
# subset of data with a profile similar to XYZ for both driving centers

n_Ux = sum(Uxbridge_17F$Conducted) # number of observations for Uxbridge
n_WG = sum(WoodGreen_17F$Conducted) # number of observations for Wood Green

Ux_17F_mean = sum(Uxbridge_17F$Passes) / n_Ux
cat("Uxbridge average pass rate:" ,Ux_17F_mean)
```

```
## Uxbridge average pass rate: 0.5932203
```

```
WG_17F_mean = sum(WoodGreen_17F$Passes) / n_WG
cat("Wood Green average pass rate:" ,WG_17F_mean)
```

```
## Wood Green average pass rate: 0.4527972
```

By average, in the last 12 years taking a test in Wood Green had a lower pass rate ($\sim$ 45.28% while Uxbridge is at $\sim$ 59.32%) for a 17 years old girl. Is the difference big enough though? A Wald test for the equality of the mean can be used to see if this difference is statistically significant.

$\bar{U} - \bar{W} \sim N(\theta_u - \theta_w, \frac{S_u^2}{n_u} + \frac{S_w^2}{n_w})$
$H_0 : \theta_u - \theta_w = 0$

$H_1 : \theta_u - \theta_w > 0$

Compute the p-value: $P(\bar{U} - \bar{W} > 0) = 1 - \Phi\left(\frac{\bar{U}-\bar{W}}{\sqrt{\frac{S_u^2}{n_u}+\frac{S_w^2}{n_w}}}\right)$ under H_0

We don't know the actual variance S_u^2 and S_w^2 because θ_u and θ_w are unknown, but we can use an estimate. Being under the null hypothesis both distribution will have the same estimate and this will be:

$\hat{S}^2 = \hat{\theta}(1 - \hat{\theta})$ where $\hat{\theta} = \frac{\hat{\theta}_u n_u + \hat{\theta}_w n_w}{n_u + n_w}$

```
distr_mean = (Ux_17F_mean - WG_17F_mean)
# mean of the distribution in study
theta_hat = (Ux_17F_mean * n_Ux + WG_17F_mean * n_WG) / (n_Ux + n_WG)
# weighted average of the two sample means
var_hat = theta_hat * (1 - theta_hat)
distr_var = var_hat / n_WG + var_hat / n_Ux
# variance of the distribution in study

1 - pnorm(distr_mean/sqrt(distr_var)) # p-value
```

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

Under H_0 $P(\bar{U} - \bar{W} > 0) < 0.001$

The p-value is really small so we can reject the null hypothesis: the difference is statistically significant. With this Wald test it is proved with a very high confidence level that the expected pass rate in Uxbridge in the last 12 years has been higher than the one in Wood Green for a 17 years old girl.

Dataset with 17 years old girls for only 2023

Now the same test will be applied to only the data of 2023 under the assumption that the previous year is the most important to see if there is a difference between the pass rates of the two driving centres.

```
Uxbridge23_17F = Uxbridge[Uxbridge$Age == 17 & Uxbridge$Gender == 'F' & Uxbridge$Year == 2023,]
WoodGreen23_17F = WoodGreen[WoodGreen$Age == 17 & WoodGreen$Gender == 'F'& WoodGreen$Year == 20
# subset of data with a profile similar to XYZ in 2023
n_Ux23 = Uxbridge23_17F$Conducted # number of observations for Uxbridge
n_WG23 = WoodGreen23_17F$Conducted # number of observations for Wood Green
```

```
Ux23_17F_mean = Uxbridge23_17F$Passes / n_Ux23
cat("Uxbridge average pass rate in 2023:" ,Ux23_17F_mean)
```

```
## Uxbridge average pass rate in 2023: 0.6380952
```

```
WG23_17F_mean = WoodGreen23_17F$Passes / n_WG23
cat("Wood Green average pass rate in 2023:" ,WG23_17F_mean)
```

```
## Wood Green average pass rate in 2023: 0.5726496
```

```
distr_mean23 = (Ux23_17F_mean - WG23_17F_mean)
# mean of the distribution in study
theta_hat23 = (Ux23_17F_mean * n_Ux23 + WG23_17F_mean * n_WG23) / (n_Ux23 + n_WG23)
# weighted average of the two sample means
var_hat23 = theta_hat23 * (1 - theta_hat23)
distr_var23 = var_hat23 / n_WG23 + var_hat23 / n_Ux23
# variance of the distribution in study

1 - pnorm(distr_mean23/sqrt(distr_var23)) # p-value
```

```
## [1] 0.1218748
```

The p-value here is higher than 0.1, which doesn't strongly challenge the null hypothesis suggesting equal expected pass rates. This raises the question of significant changes in pass rates over the 12-year period. To explore this further, it might be necessary to conduct a more detailed analysis using logistic regression to determine the expected pass rates.

Logistic Regression to calculate the expected pass rate

Using the logistic regression, we can discern the significance of various variables in predicting the expected pass rate for a 17-year-old girl in 2024.

Uxbridge

The results from the ANOVA test within the logistic regression applied to the entire Uxbridge dataset indicate that the inclusion of 'Year' in the model is not statistically significant. This suggests that, hypothetically speaking, if an individual could choose the year for their driving test, it would not have a significant impact on the outcomes.

```r
Ux_mod.form = "cbind(Passes,Conducted-Passes) ~ Gender * Age"
Uxbridge_glm = glm(Ux_mod.form, family=binomial(logit), data=Uxbridge)
summary(Uxbridge_glm)
```

```
##
## Call:
## glm(formula = Ux_mod.form, family = binomial(logit), data = Uxbridge)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2784  -0.8942   0.0000   0.9788   4.2580
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.602379   0.132348  12.107  < 2e-16 ***
## GenderM     -0.513022   0.177190  -2.895 0.003788 **
## Age         -0.081749   0.006687 -12.225  < 2e-16 ***
## GenderM:Age  0.029592   0.008938   3.311 0.000931 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 678.79  on 209  degrees of freedom
## Residual deviance: 441.12  on 206  degrees of freedom
## AIC: 1559.6
##
## Number of Fisher Scoring iterations: 3
```

From the sign of the estimates of the coefficients we can say which variable has a positive effect on the odds of success and which has a negative. Being a male in Uxbridge gives a slight disadvantage: the odds of passing are $e^{-0.513022} = 0.5986836$ times the odds of passing as a girl. Being older gives you a disadvantage as well and, given the interaction, it’s lower if you’re a male.

```r
predict(Uxbridge_glm, newdata = data.frame(Age=c(17),Gender=c('F'),
                                           Year=c(2024)), type = "response")
```

```
##         1
## 0.5529631
```

This analysis tells us that the expected pass rate in this driving center for XYZ will be 55.29%.

Wood Green

In the Wood Green dataset 'Year' looks like being an important variable to take into consideration, as all the others. The interaction between the variables, instead, do not explain enough deviance to be added to the final model.

```
WG_mod.form = "cbind(Passes,Conducted-Passes) ~ Gender + Age + Year"
WoodGreen_glm = glm(WG_mod.form, family=binomial(logit), data=WoodGreen)
summary(WoodGreen_glm)
```

```
##
## Call:
## glm(formula = WG_mod.form, family = binomial(logit), data = WoodGreen)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9316  -0.7275   0.0741   0.8415   3.0248
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -68.012299   5.553855 -12.246  < 2e-16 ***
## GenderM       0.275570   0.019198  14.354  < 2e-16 ***
## Age          -0.020524   0.003931  -5.221 1.78e-07 ***
## Year          0.033706   0.002752  12.249  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 712.57  on 215  degrees of freedom
## Residual deviance: 283.34  on 212  degrees of freedom
## AIC: 1513.8
##
## Number of Fisher Scoring iterations: 3
```

Differently from Uxbridge, in Wood Green being a male is an advantage, but being older is always a disadvantage. From this output we can see that for every year passing by, the odds of passing the driving test increase by $e^{0.033706} = 1.03428$ times.

```
predict(WoodGreen_glm, newdata = data.frame(Age=c(17),Gender=c('F'),
                                            Year=c(2024)), type = "response")
```

```
##         1
## 0.4648094
```

This analysis tells us that the expected pass rate in this driving center for XYZ will be 46.48%.

Conclusions

In analyzing the historical pass rates for 17 years old girls at Uxbridge and Wood Green over the past 12 years, a notable disparity emerged, with Uxbridge exhibiting a statistically significant advantage. The subsequent Wald test reinforced this difference, affirming that Uxbridge's expected pass rates surpassed those of Wood Green. However, it's pivotal to recognize the assumptions underlying the Wald test. This statistical approach relies on the assumption that parameter estimates (such as pass rates) follow an asymptotically normal distribution. In the consolidated analysis of all 12 years, the substantial number of observations supports this assumption. However, in the 2023 analysis, where the data is more limited, this assumption may be somewhat tenuous due to the smaller sample size. Logistic regression models further indicated that while the year did not significantly impact pass rates at Uxbridge, it played a crucial role at Wood Green, reflecting an upward trend. The expected pass rate for XYZ in Uxbridge will be 55.29% while in Wood Green it will be 46.48%, making Uxbridge the designated choice for her. While the observed upward trend might imply that delaying the driving test in Wood Green could potentially enhance success rates, a closer examination reveals that the improvement is not substantial. Even if XYZ were to take the test three years from now, coinciding with her third year of bachelor's, the probability remains lower than that in Uxbridge.

```
predict(WoodGreen_glm, newdata = data.frame(Age=c(20),Gender=c('F'),
                                            Year=c(2027)), type = "response")
```

```
##        1
## 0.474659
```

It's crucial to recognize certain drawbacks in the logistic regression analysis. Historical data assumes a steady test environment, potentially overlooking recent changes in rules or test difficulty. Logistic regression models, while helpful, oversimplify the intricate factors affecting outcomes and might miss some essential details. The personal aspects such as driving skills and the subjective nature of test experiences are vital considerations for XYZ, emphasizing the need for a well-rounded approach in choosing a test center.