ST447 Data Analysis and Statistical Methods

Individual Project: Practical car test centre decision

1st December 2023

1. Introduction

The aim of this project is to help our friend, XYZ, to decide where to take the practical car test. They must choose between the nearest test centre to their home and the nearest test centre to the London School of Economics (LSE).

To assist XYZ, we will employ data analysis tools and statistical methods to answer three questions:

1. What is XYZ's expected passing rate at the nearest test centre to their home?
2. What is XYZ's expected passing rate at the nearest test centre to the LSE?
3. Of these two locations, where should XYZ take the test? Is there any evidence to (statistically) support this suggestion?

The analysis will be presented in a manner accessible to XYZ, who may not have an extensive background in statistics. Through this report, we aim to provide not just answers to the above questions but also a transparent understanding of the data used, the methodology applied, and the reasoning behind our recommendations. Ultimately, our goal is to empower XYZ to make an informed decision about where to undertake the practical driving test.

2. Data and methodology

2.1 Dataset

The dataset at our disposal comprises information on tests pass rates, encompassing the age range of 17 to 25 years, gender, and spans the years from 2008 to 2023. For the purpose of this analysis, we isolated data relevant to the two test centres, Nelson and Wood Green (London). Rows containing missing values were removed.

2.2 Exploratory Analysis

A first analysis through descriptive statistics and charts has been made to have an introductory understanding of the data and to investigate any general trend about the pass rates across years. Afterwards, a visual investigation with bar charts has been conducted: we wanted to see how pass rates varied with respect to the gender and the age of exam-takers over the years.

2.3 Data Analysis and Statistical Methods

An initial approach to answer the questions of XYZ involved computing the overall pass rates for the two identified test centres, disregarding variations based on years, age, and gender. This initial step aimed to provide a broad understanding of the performance at each location.
A Wald test was then executed to assess whether the overall pass rates at the two test centres differed significantly. The Wald test scrutinizes the hypothesis that the pass rates are equal, computing a test statistic with which we are able to assess how confidently we can reject this hypothesis.

Moving beyond the broad comparison, Logistic regression was employed to quantify the impact of three key variables – the year of the exam, the age, and the gender of candidates – on the probability of passing the car practical test at a specific test centre. The logistic regression model not only quantified the relationship with these variables but also provided insights into the direction and magnitude of their impact.

Given these results, we wanted to further refine our analysis by calculating the expected pass rates, considering the gender of the exam-takers. This additional layer of granularity allowed for a more detailed understanding of how specific demographic characteristics contribute to variations in pass rates. Again, to ascertain the significance of the differences between the expected pass rates at the two test centres, we employed another Wald test.

2.4 Assumptions and Theoretical Notes

Throughout this process, it was assumed that data on people who pass the car practical test at a given test centre are random independent realisations of a Binomial distribution (n, p), where "n" is the number of exam-takers and "p" is the proportion of people passing the exam. A natural estimator for p is the sample pass rate. By the Central Limit Theorem (CLT) and the large sample size, the estimator of "p" is considered asymptotically distributed as a normal distribution with mean "p" (the real one) and variance of "p(1-p)/n". The hypothesis testing in the Wald test has as null hypothesis that the real difference between the average pass rate in the two test centres is 0. The alternative hypothesis is that the average pass rate in Nelson is higher than in Wood Green. To conduct this hypothesis testing, the t-test is performed using the sample variance, however, with a large sample this coincides with a z-test (normal distribution). We assume that the unknown population variance is the same for both test centres.

3. Results

3.1 Participant Profile

XYZ is a 23-year-old male.
ID = 202306391
source("XYZprofile.r")

XYZprofile(ID)

> XYZprofile(ID)

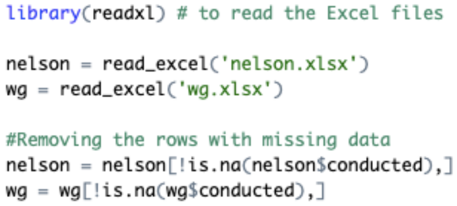
The profile of XYZ:

- Age: 23

- Gender: Male

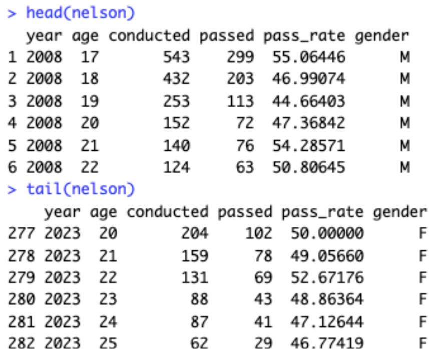
- Home address: Nelson

3.2 Exploratory Analysis

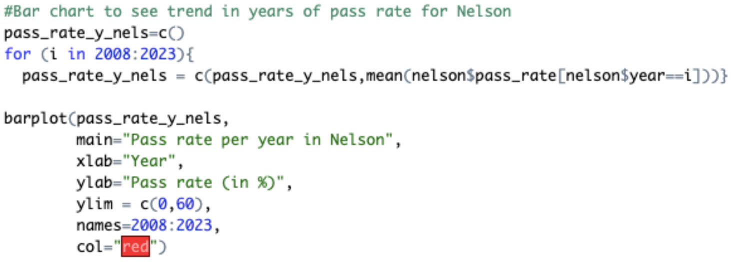


```
library(readxl) # to read the Excel files

nelson = read_excel('nelson.xlsx')
wg = read_excel('wg.xlsx')

#Removing the rows with missing data
nelson = nelson[!is.na(nelson$conducted),]
wg = wg[!is.na(wg$conducted),]
```

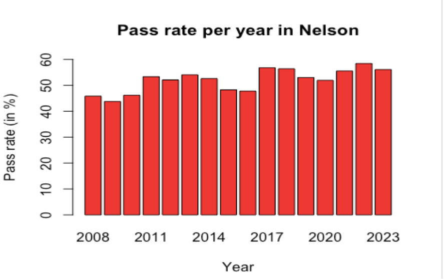
#To visualise the first and last lines of the nelson data

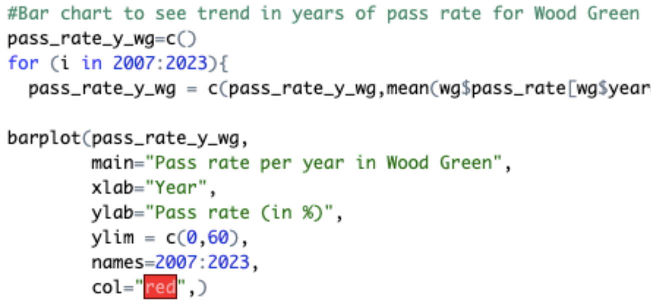


```
> head(nelson)
  year age conducted passed pass_rate gender
1 2008  17       543    299  55.06446      M
2 2008  18       432    203  46.99074      M
3 2008  19       253    113  44.66403      M
4 2008  20       152     72  47.36842      M
5 2008  21       140     76  54.28571      M
6 2008  22       124     63  50.80645      M
> tail(nelson)
    year age conducted passed pass_rate gender
277 2023  20       204    102  50.00000      F
278 2023  21       159     78  49.05660      F
279 2023  22       131     69  52.67176      F
280 2023  23        88     43  48.86364      F
281 2023  24        87     41  47.12644      F
282 2023  25        62     29  46.77419      F
```

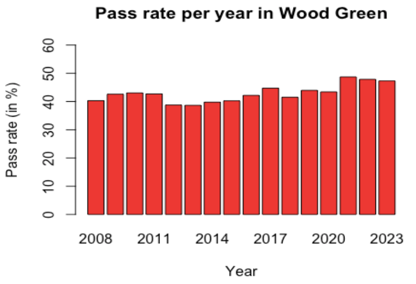


```
#Bar chart to see trend in years of pass rate for Nelson
pass_rate_y_nels=c()
for (i in 2008:2023){
  pass_rate_y_nels = c(pass_rate_y_nels,mean(nelson$pass_rate[nelson$year==i]))}

barplot(pass_rate_y_nels,
        main="Pass rate per year in Nelson",
        xlab="Year",
        ylab="Pass rate (in %)",
        ylim = c(0,60),
        names=2008:2023,
        col="red")
```

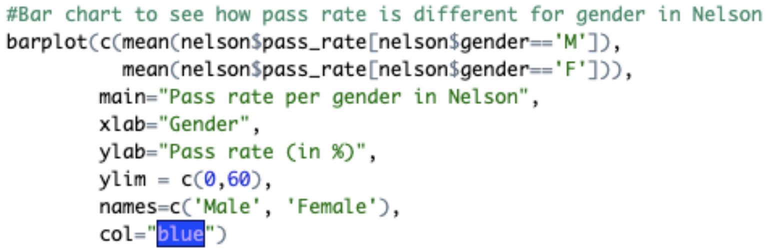




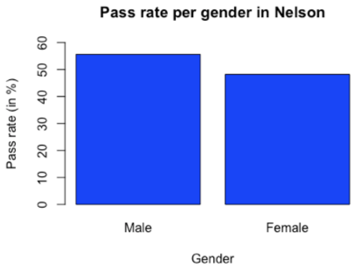
```
#Bar chart to see trend in years of pass rate for Wood Green
pass_rate_y_wg=c()
for (i in 2007:2023){
  pass_rate_y_wg = c(pass_rate_y_wg,mean(wg$pass_rate[wg$year==i]))}

barplot(pass_rate_y_wg,
          main="Pass rate per year in Wood Green",
          xlab="Year",
          ylab="Pass rate (in %)",
          ylim = c(0,60),
          names=2007:2023,
          col="red",)
```

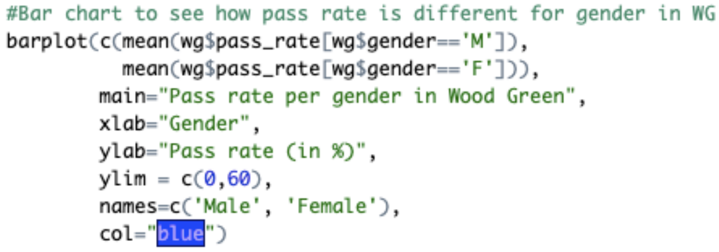


As we might observe for both test centres there is a weak tendency to have a higher pass rate year by year. We can observe that this trend is slightly more pronounced for the Nelson test centre.

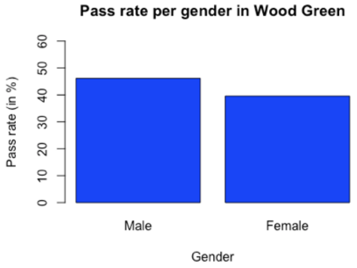


```
#Bar chart to see how pass rate is different for gender in Nelson
barplot(c(mean(nelson$pass_rate[nelson$gender=='M']),
          mean(nelson$pass_rate[nelson$gender=='F'])),
        main="Pass rate per gender in Nelson",
        xlab="Gender",
        ylab="Pass rate (in %)",
        ylim = c(0,60),
        names=c('Male', 'Female'),
        col="blue")
```

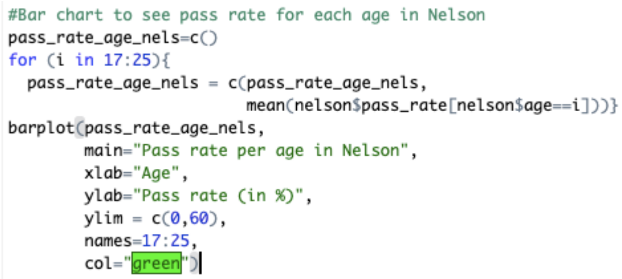




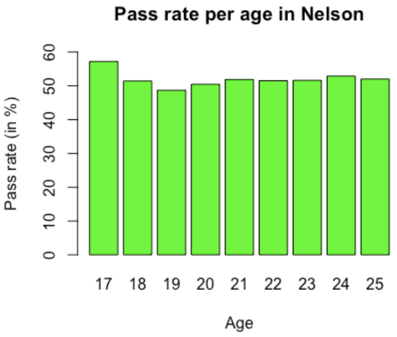
```
#Bar chart to see how pass rate is different for gender in WG
barplot(c(mean(wg$pass_rate[wg$gender=='M']),
          mean(wg$pass_rate[wg$gender=='F'])),
        main="Pass rate per gender in Wood Green",
        xlab="Gender",
        ylab="Pass rate (in %)",
        ylim = c(0,60),
        names=c('Male', 'Female'),
        col="blue")
```

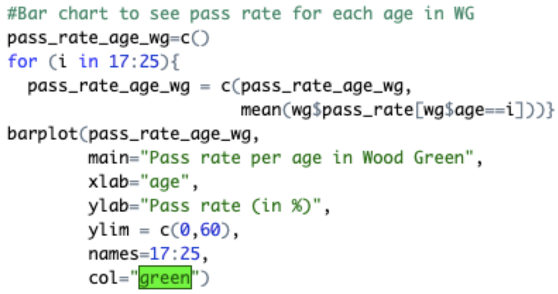


In this case, we notice that both test centres exhibit a notable difference between the pass rates for male and female individuals, with the former having a higher success rate.

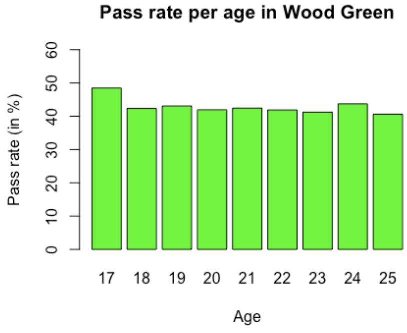


```
#Bar chart to see pass rate for each age in Nelson
pass_rate_age_nels=c()
for (i in 17:25){
  pass_rate_age_nels = c(pass_rate_age_nels,
                         mean(nelson$pass_rate[nelson$age==i]))}
barplot(pass_rate_age_nels,
        main="Pass rate per age in Nelson",
        xlab="Age",
        ylab="Pass rate (in %)",
        ylim = c(0,60),
        names=17:25,
        col="green")
```





```
#Bar chart to see pass rate for each age in WG
pass_rate_age_wg=c()
for (i in 17:25){
  pass_rate_age_wg = c(pass_rate_age_wg,
                       mean(wg$pass_rate[wg$age==i]))}
barplot(pass_rate_age_wg,
        main="Pass rate per age in Wood Green",
        xlab="age",
        ylab="Pass rate (in %)",
        ylim = c(0,60),
        names=17:25,
        col="green")
```

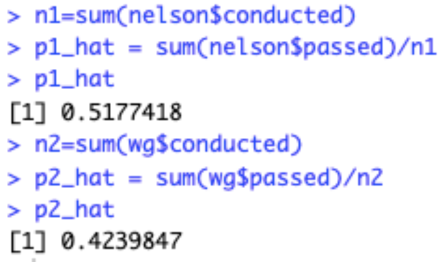


Looking at the pass rate dividing exam-takers by age, we observe a slight decrease in the rate for candidates older than 17 years old. However, for both centres, the pass rate for ages after 17 are almost equal.

With these preliminary results at hand, we can delve into the data analysis part, knowing that there exist these general trends: gender emerges as a significant factor influencing pass rates, while year and age exhibit more modest impacts. A subtle upward trajectory in pass rates is noticeable over successive years, and candidates aged 17 appear to moderately outperform their counterparts in other age groups.

3.3 Data Analysis and Statistical Methods

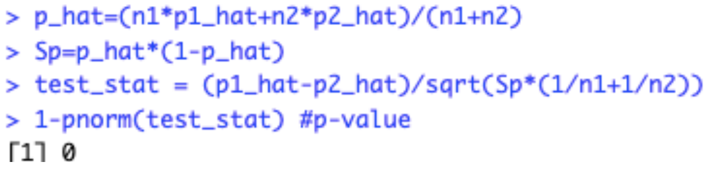
First, we compute the overall pass rate in Nelson and in Wood Green, ignoring demographic factors.



```
> n1=sum(nelson$conducted)
> p1_hat = sum(nelson$passed)/n1
> p1_hat
[1] 0.5177418
> n2=sum(wg$conducted)
> p2_hat = sum(wg$passed)/n2
> p2_hat
[1] 0.4239847
```

We notice that in Nelson the pass rate is higher by almost 0.1. Now, we are interested if this difference is statistically significant. Under the assumptions outlined in section 2.4, we run a Wald test. The null hypothesis to test is H0: $\hat{p}_1 - \hat{p}_2 = 0$, H1 is $\hat{p}_1 - \hat{p}_2 > 0$. Since the population variance is unknown, but it is assumed to be equal across the two centres, we create our test statistic using the pooled variance.

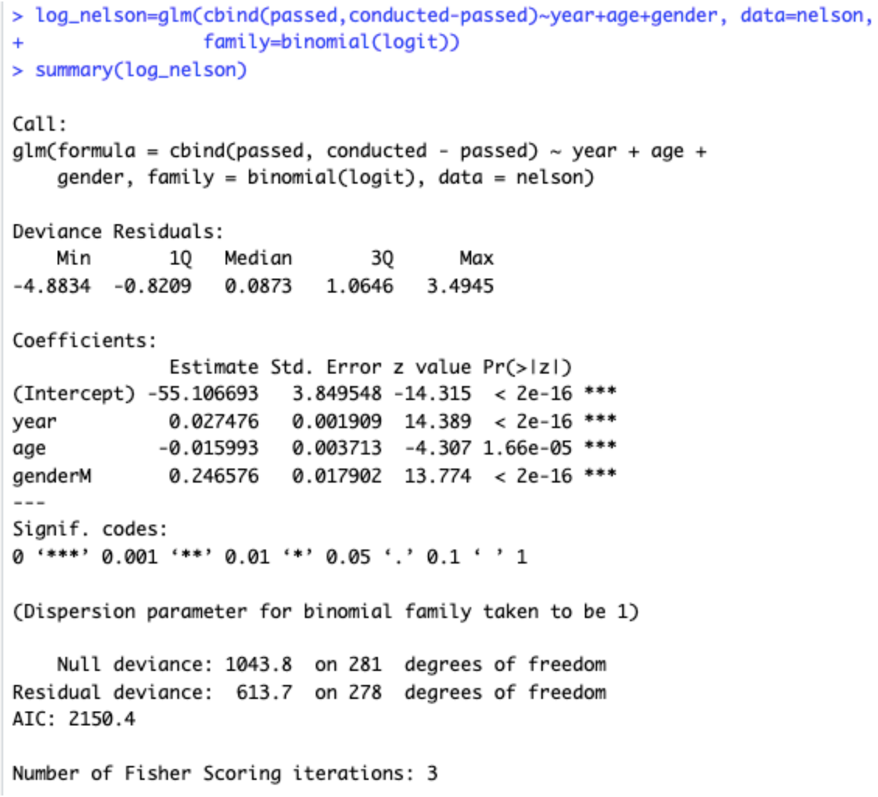
#Wald test



```
> p_hat=(n1*p1_hat+n2*p2_hat)/(n1+n2)
> Sp=p_hat*(1-p_hat)
> test_stat = (p1_hat-p2_hat)/sqrt(Sp*(1/n1+1/n2))
> 1-pnorm(test_stat) #p-value
[1] 0
```

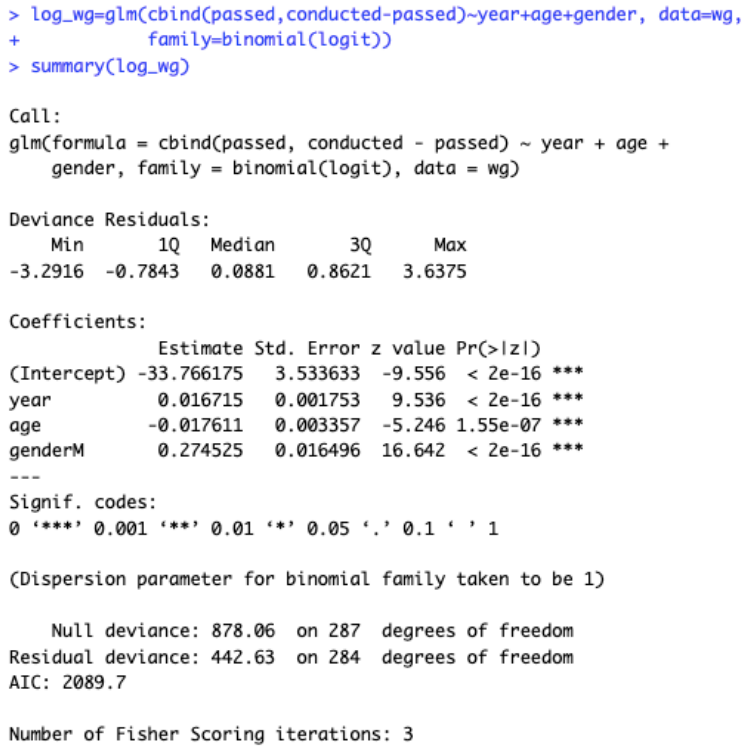
#We reject the null hypothesis with a confidence level of 99.9%. The two proportions are different.

To provide a more accurate recommendation to XYZ, we use a logistic regression to understand if age, gender and year, are significantly correlated to the probability to pass the practical car test.



```
> log_nelson=glm(cbind(passed,conducted-passed)~year+age+gender, data=nelson,
+               family=binomial(logit))
> summary(log_nelson)

Call:
glm(formula = cbind(passed, conducted - passed) ~ year + age +
    gender, family = binomial(logit), data = nelson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8834  -0.8209   0.0873   1.0646   3.4945

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -55.106693   3.849548 -14.315  < 2e-16 ***
year          0.027476   0.001909  14.389  < 2e-16 ***
age          -0.015993   0.003713  -4.307 1.66e-05 ***
genderM       0.246576   0.017902  13.774  < 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: 1043.8  on 281  degrees of freedom
Residual deviance:  613.7  on 278  degrees of freedom
AIC: 2150.4

Number of Fisher Scoring iterations: 3
```

#Logistic regression for Wood Green



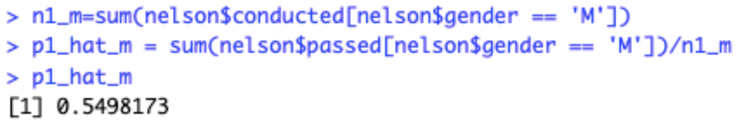
```
> log_wg=glm(cbind(passed,conducted-passed)~year+age+gender, data=wg,
+            family=binomial(logit))
> summary(log_wg)

Call:
glm(formula = cbind(passed, conducted - passed) ~ year + age +
    gender, family = binomial(logit), data = wg)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2916  -0.7843   0.0881   0.8621   3.6375

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -33.766175   3.533633  -9.556  < 2e-16 ***
year          0.016715   0.001753   9.536  < 2e-16 ***
age          -0.017611   0.003357  -5.246 1.55e-07 ***
genderM       0.274525   0.016496  16.642  < 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: 878.06  on 287  degrees of freedom
Residual deviance: 442.63  on 284  degrees of freedom
AIC: 2089.7

Number of Fisher Scoring iterations: 3
```

In these cases, the logistic regression has as response variable the success or failure of a driving test. R takes the total of successes and failures, and it adapts them to perform a logistic regression. As it may be seen from the p-value of the three variables to which we are interested, all the coefficients are strongly statistically significant, and we are therefore very confident that they are correlated to the pass rate of each test centre. The results confirm what we have observed in the descriptive analysis:

1. The coefficient on year is positive, indicating that there is a slight increase in the odds of passing the driving exam year after year, around 3% for Nelson centre and 2% for Wood Green.
2. The coefficient on age is negative, meaning that with an increase in age there is a decrease in the odds of passing the driving test. This decrease is around 2%.
3. The coefficient on gender is positive, reflecting that male candidates had more success in this test. The magnitude of this effect on the odds of passing the test is calculated by computing $e^{0.25} = 1.28$ and $e^{0.28} = 1.32$, therefore the odds of passing the test are 28% higher for Nelson and 32% for Wood Green. Odds are the ratio of probability of success and the probability of failure.

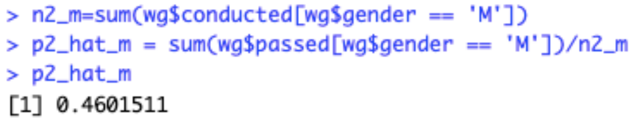
Since the effect of year and age are quite marginal on the pass rate, we are going to provide an expected pass rate, given the gender of the candidates.

#Expected pass rate given that the exam-taker is a male in Nelson



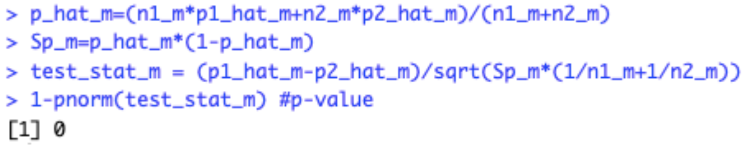
```
> n1_m=sum(nelson$conducted[nelson$gender == 'M'])
> p1_hat_m = sum(nelson$passed[nelson$gender == 'M'])/n1_m
> p1_hat_m
[1] 0.5498173
```

#Expected pass rate given that the exam-taker is a male in Wood green



```
> n2_m=sum(wg$conducted[wg$gender == 'M'])
> p2_hat_m = sum(wg$passed[wg$gender == 'M'])/n2_m
> p2_hat_m
[1] 0.4601511
```

#Wald test



```
> p_hat_m=(n1_m*p1_hat_m+n2_m*p2_hat_m)/(n1_m+n2_m)
> Sp_m=p_hat_m*(1-p_hat_m)
> test_stat_m = (p1_hat_m-p2_hat_m)/sqrt(Sp_m*(1/n1_m+1/n2_m))
> 1-pnorm(test_stat_m) #p-value
[1] 0
```

#We reject the null hypothesis with a confidence level of 99.9%. The two proportions are different.

3.4 Summary of results

The comprehensive analysis reveals intriguing patterns in driving test pass rates. Initially, an overall pass rate for Nelson was notably higher than for Wood Green, 52% for the former and 42% for the latter. Using a logistic regression, we understand that, while year and age have subtle effects on pass rates, gender emerges as a crucial determinant, with males exhibiting higher odds of success. Integrating gender into the analysis, **the expected pass rates for XYZ become 55% in Nelson and 46% in Wood Green**. Considering these findings, coupled with our initial overall expected rates that, despite being different, substantially confirm the same fact, **we confidently recommend XYZ to take the car driving test in Nelson, where the expected success rate is higher**.

4. Strengths and Weaknesses

Our analytical approach to determining the optimal test centre for XYZ exhibits several notable strengths, reinforcing the reliability and credibility of our recommendations. However, it is essential to acknowledge certain inherent limitations that warrant consideration.

4.1 Strengths

Our methodology is underpinned by statistical rigor. The Wald test and the Logistic regression allow us to have solid statistical claims. Another significant strength lies in the internal consistency of our results. The overall pass rate and the pass rate considering demographic factors align, instilling confidence in the reliability of our analysis. Furthermore, our analysis goes beyond mere numerical comparisons. By identifying a general tendency of increasing pass rates over the years, particularly in the recommended test centre, we contribute a nuanced understanding of temporal dynamics.

4.2 Weaknesses

An inherent limitation in our approach is the omission of a measure for candidate preparation. The success rate may be influenced by the candidates' level of readiness, and the absence of this consideration introduces a potential bias into our recommendations. In addition, our analysis does not encompass certain crucial aspects that could impact the decision-making process of XYZ. Variables such as waiting times for exams, general weather conditions, and traffic patterns are not considered. This oversight may introduce biases into our ultimate recommendation. Also, a potential correlation between success and failure at a particular location with external factors like weather and traffic conditions remains unexplored. Neglecting these correlations may result in an incomplete understanding of the dynamics influencing test outcomes.

5. Conclusion

Our statistical analysis has provided a robust foundation for recommending a suitable test centre based on available data and general demographic trends. The consistency in pass rates and the identification of temporal trends add confidence to our findings. However, the decision for XYZ is multi-faceted, and there are additional factors that should be considered for a well-informed choice.

Our approach is solid in its statistical foundation, offering a consistent and rigorous analysis of pass rates and their demographic variations. Despite these strengths, we must acknowledge the limitations inherent in the lack of candidate preparation measures and the exclusion of external factors, such as climate or traffic conditions. Therefore, we recommend that XYZ reflects on personal preferences and individual priorities. Factors such as comfort in driving conditions, preference for specific traffic and climate scenarios, should not be overlooked. XYZ may prioritize the setting in which they feel most at ease and capable of expressing their full driving potential.

In conclusion, keeping in mind our recommendation to take the practical car test in Nelson, based on the fact that this test centre has an expected pass rate of 55% compared to the 46% of Wood Green, we would advise XYZ to combine our insights with additional information gathered about their personal preferences, for instance, on traffic and weather conditions. The ultimate choice of the test centre should align with their comfort and confidence, ensuring a positive and successful experience. While statistics offer valuable guidance, personal preferences play a major role in optimizing the driving test experience.