**On the passing rate of UK driving test centres**

Candidate number: 22137

The profile of XYZ is a 25-year-old male which has two possible locations to take the driving test: Wood Green London and Worcester.

## Data preparation

The data available identifies the pass rate for each test centre, gender, age (between 17 to 25) and the year of the test. I used the data relevant for the test centres already mentioned grouped in the following way in an Excel file:

| Centre | Age | Year | Gender | Total | Pass | NoPass |
|---|---|---|---|---|---|---|
| Worcester | 17 | 2022 | Male | 1325 | 779 | 546 |
| Worcester | 18 | 2022 | Male | 830 | 433 | 397 |
| Worcester | 19 | 2022 | Male | 440 | 236 | 204 |
| Worcester | 20 | 2022 | Male | 282 | 138 | 144 |
| Worcester | 21 | 2022 | Male | 197 | 95 | 102 |

**Table 1.** example of the five first rows of the data used.

Centre accounts for the testing centre, Age represent how old was the person tested, the third column is the year when the test was conducted, Gender represents whether the person is male or female, Total is the total number of tests, Pass is the number of test passed and NoPass represents the number of tests not passed. There are 16 data points available for each year (2007 to 2022), for each age (17 to 25), for each gender (male and female) and for each centre (Wood Green and Worcester) summing to a total of 576 data points. Some boxplots are shown below.

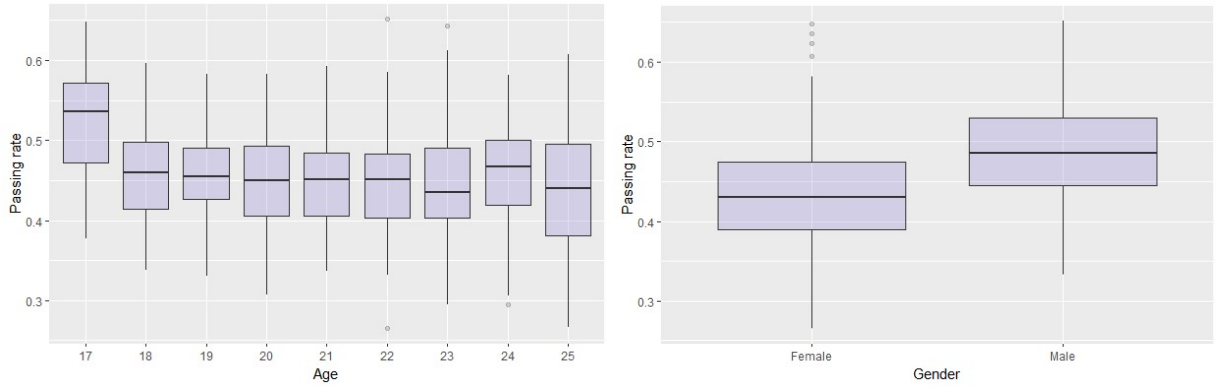


**Figure 1a.** Boxplot for Passing rate against Age. **Figure 1c.** Boxplot for Passing rate against Gender.

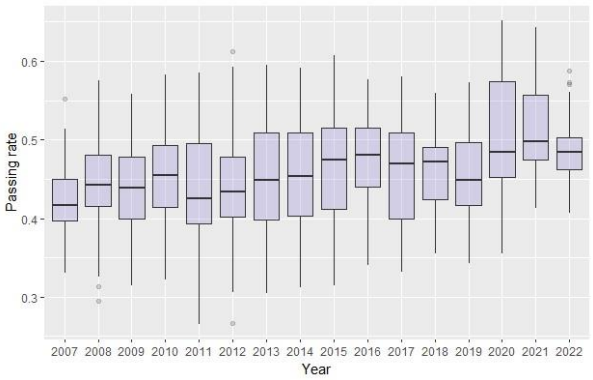



**Figure 1c.** Boxplot for Passing rate against year. **Figure 1d.** Boxplot for Passing rate against Centre.

The above boxplots in figure 1 help us identify trends in the data, for example one can infer that the passing rate in Worcester is higher than in Woodgreen, also that for males is higher than females and that it has been increasing throughout the last years. By the other hand, for the age it does not seem to change in a significant way. These boxplots help us visualise, but to make valid conclusions we need to use statistical methods.

## Logistic regression approach

As a first approach I consider a logistic regression to predict the pass rate that XYZ will have at each of the test centres considered (Worcester and Wood Green). For this approach I want to regress the probability of passing the test by having gender, age and centre as independent variables. Therefore, I first need to make sure that these variables are significant to regress the dependent variable. For this, I first build the deviance table to determine the following hypothesis for every coefficient:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

I am only interested in determining if each independent variable is able to regress the probability of passing the test. The results are shown below:

| Variable | Df | Deviance | Residual df | Resid. dev | p-value |
|---|---|---|---|---|---|
| Null | 1 | - | 575 | 3049.8 | - |
| Gender | 1 | 432.95 | 574 | 2616.9 | 0.00 |
| Age | 1 | 648.02 | 573 | 1968.9 | 0.00 |

| | | | | | |
|---|---|---|---|---|---|
| Centre | 1 | 672.54 | 572 | 1296.3 | 0.00 |

**Table 2.** Analysis of deviance.

Table 2 shows the results of the analysis of deviance, but a simple way to understand the relevant results of this table is to consider the p-values for every variable and one can notice that these p-values are always lower than 0.05, therefore I can conclude with enough statistical evidence that every coefficient is different to 0 (I reject the null hypothesis under a 0.05 significance level).

Now that I know the variables are significant and there is no need to remove any, I continue by fitting the logistic regression model in RStudio.

| Coefficient | Estimate | Std. Error | z- value | p-value |
|---|---|---|---|---|
| Intercept | 0.27 | 0.047 | 5.699 | 0.00 |
| Gender: Male | 0.21 | 0.010 | 20.033 | 0.00 |
| Age | -0.03 | 0.002 | -14.737 | 0.00 |
| Centre: Worcester | 0.29 | 0.011 | 25.903 | 0.00 |

**Table 3.** Summary of the logistic regression model.

The above table shows the results and, as expected, every variable is significant in the model, hence they have predictive power to the pass rate. The estimation of coefficients with its standard errors are also shown. These tell us that being a male and performing the test at Worcester increase the passing rate because the coefficients are positive. However, the older the individual is, the lower it is the passing rate because the coefficient is negative. Just by this analysis I can start inferring about the outcome of the question of interest: given the estimation of the coefficients, XYZ will have a better chance to pass the test in Worcester rather than in Wood Green.

It is important to notice that, since every coefficient in the model is significant, then the model is globally significant as well and has predictive power to the pass rate.

Now that I have the estimates of the coefficients, I can make predictions given the characteristics of XYZ. The logistic model will return a number between 0 and 1 that can be interpreted as the pass rate (or probability) given the value that each variable takes. For both centres, I set the age to be 25 and the gender as Male and I proceed by comparing the probability given by the model by setting the dummy variable “Centre: Worcester” to 1 and then to 0, meaning that in the first case I predict the expected passing rate in Worcester and in the

second case I predict the expected passing rate in Wood Green. The results are shown as follows:

| Centre | Expected passing rate | Standard error | Lower confidence interval | Upper confidence interval |
|---|---|---|---|---|
| Worcester | 0.489 | 0.004 | 0.481 | 0.497 |
| Wood Green | 0.416 | 0.003 | 0.409 | 0.423 |

**Table 4.** Results by logistic regression model.

And a friendlier visualisation for the confidence interval:

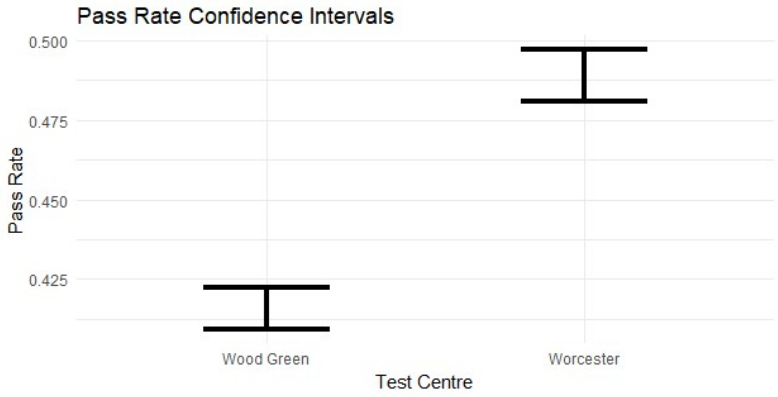


**Figure 2.** Confidence interval for the passing rate of each centre for 25-year-old males.

As can be seen in table 4 and figure 2, the passing rate for Worcester is higher than in Wood Green, therefore it is advisable for XYZ to take the test in the Worcester centre based on the statistical analysis given by the logistic regression model. It is also worth noting that the confidence intervals do not overlap, meaning that they are significantly different.

## Wald Test approach

Taking into account the results in table 2 stating that age, gender and centre are significant and one can differentiate between them, I will filter the data related to the profile of XYZ, i.e., I will only use the data for 25-year-old males to calculate the expected passing rate of Worcester and Wood Green test centres.

Also, Assuming asymptotic normality for the unbiased estimators of the means and to have stronger statistical evidence, I want to confirm my previous results via the Wald test by testing the following hypotheses for the means difference:

$$H_0: \mu_{Worcester} - \mu_{Wood\ Green} = 0$$

$$H_1: \mu_{Worcester} - \mu_{Wood\ Green} \geq 0$$

The null hypothesis states that the expected passing rate for both centres are equal, whereas the alternative hypothesis states that the expected passing rate is greater in the Worcester test centre. Now I construct the test statistic:

$$W = \frac{\bar{X}_{Worcester} - \bar{X}_{Wood\ Green}}{\widehat{se}(\bar{X}_{Worcester} - \bar{X}_{Wood\ Green})} = \frac{\bar{X}_{Worcester} - \bar{X}_{Wood\ Green}}{\sqrt{\frac{S^2_{Worcester}}{n_{Worcester}} + \frac{S^2_{Wood\ Green}}{n_{Wood\ Green}}}} = 2.63$$

With this test statistic and assuming that $z_\alpha = 1.645$ with a p-value of 0.00 I can reject the null hypothesis and I conclude that with a 0.05 level of significance the expected passing rate of the Worcester centre is higher than the expected passing rate of the Wood Green centre.

## Conclusion

To sum up, based on both the logistic regression model and the Wald test, it is advisable for XYZ to take the test in Worcester since XYZ has a higher expected passing rate in that centre. This recommendation is based and supported by two different and valid statistical methods. Regarding the Wald test, a significance level of $0.05$ was used, i.e., there is only a $5\%$ probability that the expected passing rate in Worcester is not greater than in Wood Green, this is a very small probability. Furthermore, the Wald test provides a satisfactory answer when one have data with random fluctuations.

One important strength of both approaches is the simplicity of both. Also, I am using confidence intervals to conclude accordingly and in both cases I get the same answer. Therefore, the advise to XYZ is confirmed by two different statistical methods.

One limitation is the small number of datapoints used for the Wald test. Since I filtered the data for 25-year-old males there were only 16 datapoints for each centre and it may not perfectly follow the asymptotic normality as the number of datapoints should be more. Another option different to Wald test is to perform the $t$-test with the student's $t$-distribution, but when the student's $t$-distribution has a large number of degrees of freedom (that is datapoints) it converges to the normal distribution used in the Wald test, so it might not be a big change. It would be definitely better to have more data, as for example using a monthly or even daily dataset.

Another limitation is that logistic regression is sensible to outliers and within our dataset, we have one that can be considered an outlier which is 2020 when the COVID pandemic occurred and there was less traffic than usual.

### RStudio code

```
current_working_dir <- dirname(rstudioapi::getActiveDocumentContext()$path)
setwd(current_working_dir)
rm(list=ls())
library('readxl')
library('lmtest')

library('ggplot2')

library('rmarkdown')
library('DescTools')

ID = 202338054
source("XYZprofile.r")
XYZprofile(ID)

## The profile of XYZ:
## - Age:  25
## - Gender:  Male
## - Home address:  Worcester

df <- read_excel('data.xlsx', sheet = 1)

# Define the type of the variables
df$Age = as.numeric(df$Age)
df$Centre = as.factor(df$Centre)
df$Gender = as.factor(df$Gender)
### For boxplots
ggplot(df, aes(x=as.factor(Age), y=Pass/Total)) +
  geom_boxplot(fill="slateblue", alpha=0.2) +
  xlab("Age") + ylab('Passing rate')

########## Logistic regression ##########

# Formula for the model with the independent variables to use
mod.form = "cbind(Pass,NoPass) ~ Gender + Age + Centre"
# Fitting the logistic regression
glm.out = glm(mod.form, family=binomial(logit), data=df)
# Analysis of deviance table
anova(glm.out, test="Chisq")

# Summary of the results of the model
summary(glm.out)
```

```r
# Creates a dataframe with the desired values of the variables
d1 <- data.frame(Centre = 'Worcester', Age = 25, Gender = 'Male')
# Makes a prediction for the logistic regression and calculates standard erro
r
preds1 = predict(glm.out, d1 , type="response", se.fit = TRUE)
# Confidence interval for the passing rate
critval <- 1.96
upr1 <- preds1$fit + (critval * preds1$se.fit)
lwr1 <- preds1$fit - (critval * preds1$se.fit)
fit1 <- preds1$fit
fit1
## 0.4890129

# Creates a dataframe with the desired values of the variables
d2 <- data.frame(Centre = 'Wood Green', Age = 25, Gender = 'Male')
# Makes a prediction for the logistic regression and calculates standard erro
r
preds2 = predict(glm.out, d2 , type="response", se.fit = TRUE)
# Confidence interval for the passing rate
critval <- 1.96 ## approx 95% CI
upr2 <- preds2$fit + (critval * preds2$se.fit)
lwr2 <- preds2$fit - (critval * preds2$se.fit)
fit2 <- preds2$fit
fit2
## 0.416001

########## Confidence interval plot ##########
df2 <- data.frame(
  TestCentre = c("Worcester", "Wood Green"),
  Mean = c(fit1, fit2),
  Lower = c(lwr1, lwr2),
  Upper = c(upr1, upr2))
# Create a ggplot object
p <- ggplot(df2, aes(x = TestCentre, y = Mean, fill = TestCentre)) +
  # Plot the error bars with a more visually attractive appearance
  geom_errorbar(aes(ymin = Lower, ymax = Upper), position = position_dodge(0.
9), width = 0.4, color = "black", size = 1.5) +
  # Customize the plot appearance
  theme_minimal() +
  labs(title = "Pass Rate Confidence Intervals",
       x = "Test Centre",
       y = "Pass Rate")

########## Wald test ##########

# Retrieves the data for 25-year-old males in Wood Green and Worcester, respe
ctively
passed_woodgreen = df$Pass[df$Age == 25 & df$Gender == 'Male' & df$Centre ==
'Wood Green']
passed_worcester = df$Pass[df$Age == 25 & df$Gender == 'Male' & df$Centre ==
```

```
'Worcester']
Tot_woodgreen = df$Total[df$Age == 25 & df$Gender == 'Male' & df$Centre == 'W
ood Green']
Tot_worcester = df$Total[df$Age == 25 & df$Gender == 'Male' & df$Centre == 'W
orcester']
# Calculates passing rate
rate_woodgreen = passed_woodgreen/Tot_woodgreen
rate_worcester = passed_worcester/Tot_worcester
# Calculates mean
X_woodgreen = mean(rate_woodgreen)
X_worcester = mean(rate_worcester)
# Calculates variance
Var_woodgreen = Var(rate_woodgreen)
Var_worcester = Var(rate_worcester)
# Number of observations
n_woodgreen = length(rate_woodgreen)
n_worcester = length(rate_worcester)
# Calculates standard error of the difference of means
se = sqrt(Var_woodgreen/n_woodgreen + Var_worcester/n_worcester)
# Wald test statistic
W = (X_worcester - X_woodgreen)/se
W

## [1] 2.633432)
```