The London School of Economics and Political Sciences

Department of Statistics

Data Analysis and Statistical Methods: Individual Project on Pass Rate Analysis

Candidate ID:

26990

Module Code: ST447

Submitted in partial fulfillment of the requirements for the MSc degree in Data Science

(Title Page + 8 Pages)

November 2023

1 Introduction

The objective of the data analysis outlined in this report is to provide a quantitative foundation for assisting the fictional student XYZ in making an informed decision about the optimal location to undertake their practical UK car test. XYZ's profile is summarized in Table 1.

Table 1: XYZ's Profile

Age	**Gender**	**Home Address**
19	Female	Worthing

We assume that XYZ wants to maximize her probability of passing the exam on the first try. XYZ's driving skills are exactly average, and the only factor we can alter to influence the probability of passing is the choice of the test center location. Based on the project details, as XYZ is a student at the London School of Economics and Political Science, she has the choice of taking the driving test near either the LSE or her home address. Therefore, the analysis compares the two test centers, **Worthing** (near her home address) and **Wood Green** (near the LSE).

This evaluation aims to guide XYZ's test center selection, considering her profile attributes, test attempt frequencies, and pass rates associated with each location recorded in UK government data.[1]

2 Data

The data, initially in *.ods* format, details annual outcomes for UK test centers. Converting it to *.xlsx* enables compatibility with R's `readxl` package, simplifying data reading. Looping through all tables in the file, we remove empty cells, standardize formats, and merge the yearly tables into a single data frame. The relevant R-code for the procedure is outlined below:

```r
library(readxl) # for reading xlsx data
library(dplyr) # common package for data processing
library(zoo) # for the na.locf() function
library(boot) # use the boot lib to estimate CIs (section 3.1)
# path to the raw data
path <- "dvsa1203.xls"
# define which sheets to read from the original file
sheet_names <- excel_sheets(path = path)[-1]
# create an empty data frame to store the data in
```

[1] https://www.gov.uk/government/statistical-data-sets/car-driving-test-data-by-test-centre.

```r
uk_df_wide <- data.frame()
# loop over all the tables in the original file
for (i in 1:length(sheet_names)) {# read and store table i in sheet i.
  # cleaning by skipping empty rows and headers and defining NA strings
  df_i <- read_xls(path, sheet = sheet_names[i], skip = 7,
    col_names = F, na = c("","-"))
  # only select non-empty columns
  df_i_relevant_cols <- df_i %>% select_if(~sum(!is.na(.)) > 0)
  # assign appropriate column names
  colnames(df_i_relevant_cols) <- c("Location", "Age",
                      "M_conducted","M_passes","M_pass_rate",
                      "F_conducted","F_passes","F_pass_rate",
                      "T_conducted","T_passes","T_pass_rate")
  # fill in some empty cells and remove unnecessary rows
  # set the correct data format for columns
  df_i_clean <- df_i_relevant_cols %>%
    mutate(Location = na.locf(Location) ) %>%
    filter(!is.na(Age) & Age != "Total") %>%
    mutate(Date = sheet_names[i]) %>%
    mutate(Date = as.factor(Date)) %>%
    mutate(Location = as.factor(Location)) %>%
    mutate(Age = as.factor(Age)) %>%
    select(Date, everything())
  # append the data for sheet i to the data frame
  uk_df_wide <- rbind(uk_df_wide, df_i_clean)}
# reformat df for data analysis by creating a column for gender
uk_m_df <- uk_df_wide %>% # select male results
  rename(Passrate = M_pass_rate, Conducted = M_conducted) %>%
  select( -F_conducted, -F_passes, -F_pass_rate,
          -T_conducted, -T_passes, -T_pass_rate, -M_passes) %>%
  mutate(Gender = "M")
uk_f_df <- uk_df_wide %>% # select female results
  rename(Passrate = F_pass_rate, Conducted = F_conducted) %>%
  select( -F_passes, -T_conducted, -T_passes, -T_pass_rate,
          -M_conducted, -M_passes, -M_pass_rate) %>%
  mutate(Gender = "F")
# stack the df for male and female to get the full data again
uk_df <- rbind(uk_m_df, uk_f_df)
uk_df <- uk_df %>%
  arrange(Location, Date, Age) %>%
  mutate(Location = ifelse(Location == "Wood Green",
  "Wood Green (London)", as.character(Location))) # Fix naming change
```

Code Listing 1: Reading, Cleaning, and Structuring the Data

In the resulting data frame, six columns remain: *Date*, *Location*, *Age*, *Conducted*, *Passrate*, *Gender*. Table 2 visually demonstrates the structure of the resulting data frame, featuring the top three entries based on the pass rate. Figure 1 (left) reveals correlations between relevant variables in the data set. Upon further data exploration, it's apparent that the recorded mean pass rates[2] vary significantly across different locations, as shown in Figure 1 (right).

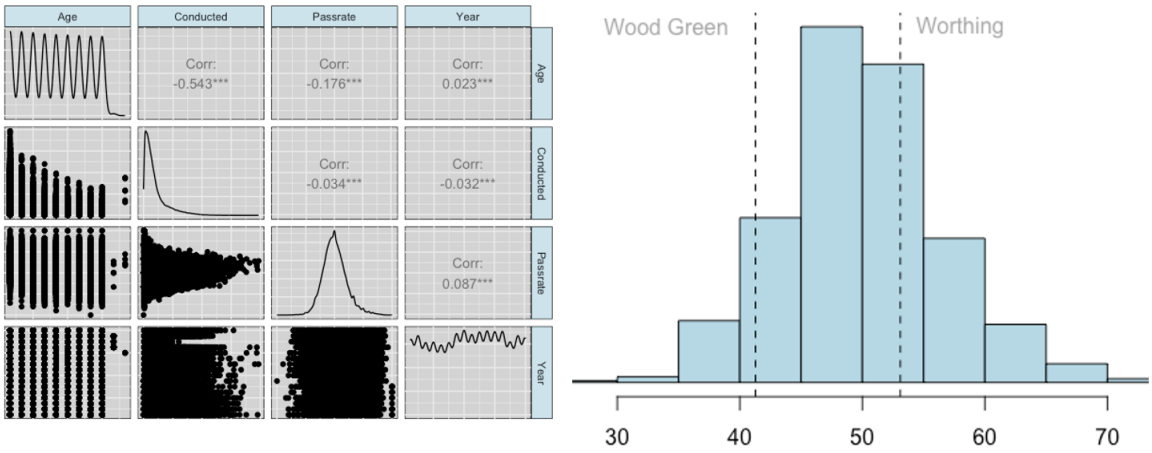
[2]Mean pass rate across all observations recorded for each location.

Table 2: Structure of the Clean Data Frame

Date	Location	Age	Conducted	Passrate	Gender
2011-12	Alness	24	10	100	M
2012-13	Callander	22	11	100	M
2012-13	Cardigan	24	7	100	M

Our main interest lies in the test centers where XYZ can take the driving test, highlighting their mean pass rates in Figure 1 (right). Clearly, the considerable difference between these two pass rates strongly indicates Worthing as the optimal choice for XYZ's test location. The next section applies several statistical methods to explore this difference further and determine its statistical significance to ultimately find the test center that maximizes XYZ's probability of passing the test.

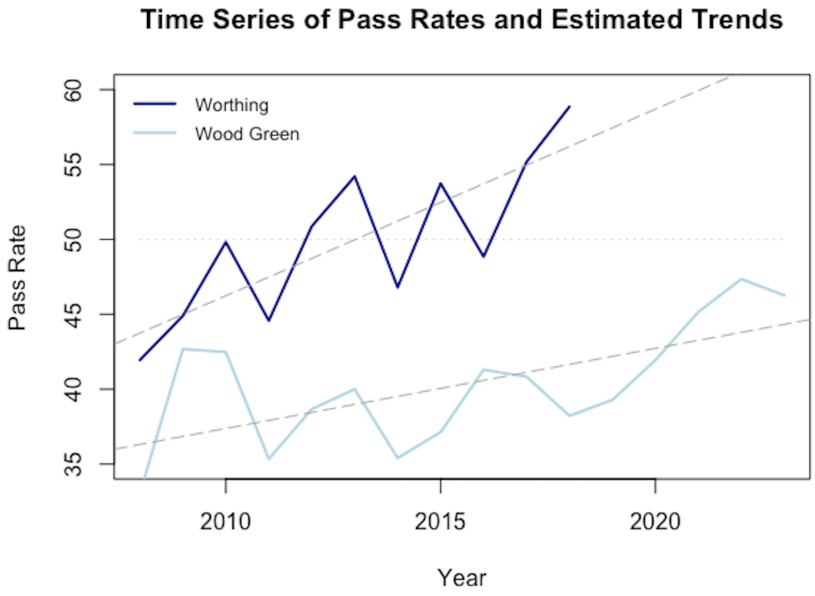
Figure 1: Correlation Plots (left) and Distribution of Pass Rates (right)



3 Methods and Results

To get a better understanding of how the pass rates between the two test centers compare, we look at the time series of their pass rates for XYZ's characteristics. Note the data range discrepancy: Worthing's series spans 2008 to 2018, while Wood Green's covers 2008 to 2023, posing challenges for comparison and prediction. Nonetheless, the difference illustrated in Figure 1 persists: Worthing consistently shows a higher pass rate, as is illustrated in Figure 2. In the upcoming subsection, we aim to employ statistical methods to further validate and provide robust support for the visual findings observed in Figure 2.

Figure 2: Comparing Pass Rates across Time and Test Centers



3.1 Pass Rate Estimation with Logistic Regression

To estimate the effect of the test center location on the pass rate, we fit the pass rate on *Age*, *Gender*, the *Year*,[3] and the *Location* in a data set that contains only data for the two relevant locations. Besides assessing location impact on pass rates, we will later use the model to predict each location's expected pass rate, if XYZ were to take the test in 2024. The expected pass rate, constrained within the 0 to 1 range, can be interpreted as a probability, warranting the use of a logistic regression model for prediction. Such a model is useful when decision-making requires an understanding of uncertainty. The approach gains further credence when considering the foundation of the data recorded in *Passrates*. Individual test-taker data is absent, and the *Passrates* variable aggregates results based on subgroups - year, gender, location, and age. However, each individual test can be viewed as a Bernoulli trial with binary outcomes and unknown p ($X_i \sim \text{Bernoulli}(p)$), implying that the average pass rate within each subgroup is a sample mean of N such random variables, $\hat{p} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Given the assumption of the central limit theorem (trial independence), we can also ascertain the distribution, expected value, and variance of this sample mean $\hat{p} \dot{\sim} \text{Normal}(\hat{p}, \frac{\hat{p}(1-\hat{p})}{N})$. Owing to the aggregated data structure, merging information from *Passrates* and *Conducted* is necessary to generate a binary dependent variable for the logistic model. Fortunately, this process is straightforward in R:

[3]To allow for prediction, the originally categoric *Date* variable is transformed into a numeric *Year* variable.

```r
# create a new data frames for the locations of interest
# create a numeric "Year" variable from the "Date"
# select only variables needed for the model estimation
relevant_locations <- uk_df %>%
  filter(Location=="Wood Green (London)" | Location=="Worthing" ) %>%
  mutate(Year = as.numeric(paste("20",substr(Date,6,7), sep = ""))) %>%
  select(Year, Age, Passrate, Gender, Location, Conducted) %>%
  mutate(Passrate = Passrate/100 ) # scale Passrate
# estimate a logistic model by creating a binary dependent variable
gl_mod = glm(cbind(Passrate*Conducted,Conducted-Passrate*Conducted)~.,
             family=binomial(logit), data = relevant_locations)
summary(gl_mod) # show the model output
```

Code Listing 2: Fitting a Logistic Model

Table 3 gives more insight into the specific model used and the estimated coefficients. In the context of this analysis, the most important coefficient is for the location of *Worthing*. The coefficient of 0.4582 implies that when changing the test location from Wood Green to Worthing, the log-odds of the outcome (passing or failing the test) change by a factor of $e^{0.4582} = 1.5812$. Judging from the p-value, the coefficient is statistically highly significant.

Table 3: Output for Logistic Regression Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-42.6441	3.0544	-13.96	0.0000
Year	0.0211	0.0015	13.89	0.0000
Age18	-0.2362	0.0203	-11.62	0.0000
Age19	-0.2349	0.0224	-10.48	0.0000
...	...	...	...	...
Age25	-0.3335	0.0280	-11.89	0.0000
GenderM	0.2735	0.0123	22.26	0.0000
Worthing	**0.4582**	0.0139	32.98	0.0000

Using the logistic model, we finally predict the expected pass rates linked to XYZ's attributes, considering the effect of the two locations and assuming that she takes the test in 2024. Predicting pass rates using the model is straightforward, but establishing confidence intervals for these predictions poses more difficult because of the logistic model's nature. However, to assess the significance of differences between estimated rates, confidence intervals are essential. One possible approach involves estimating confidence intervals for predictions through a bootstrapping method[4]. With the code below we can predict XYZ's pass rate for each location and estimate 95%-confidence intervals for the results.

[4]Bootstrapping is a resampling technique that involves creating multiple datasets by randomly sampling with replacement from the original data to estimate statistics or uncertainties.

```r
set.seed(8) # set a seed for reproducibility
# Create new df with the profile of XYZ for both locations
XYZ_profile <- data.frame(Year = 2024, Age = as.factor(19),
  Gender = "F",Location = c("Wood Green (London)","Worthing"))
gl_predictions <- c() # vector to store results (pred,ci_lb,ci_ub)
for (i in 1:2) { # fit and predict for worthing, wood green
  gl_boot_pred <- function(data, samp) {#def logit function for boot()
    gl_m<-glm(cbind(Passrate*Conducted,Conducted-Passrate*Conducted)~.,
              family=binomial(logit), data = relevant_locations[samp,])
    predict(gl_m, newdata = XYZ_profile[i,], type = "response")}
  # use the boot() function to implement the algorithm directly
  boot_results <- boot(data = relevant_locations,
              statistic = gl_boot_pred, R = 1000)
  # use boot.ci to estimate a CI for the boot results above
  boot_ci <- boot.ci(boot_results, conf = 0.95, type = "basic")
  gl_predictions <- c(gl_predictions,boot_results$statistic(),
              boot_ci$basic[4:5])}# store the two rows in matrix
gl_pred_results <- matrix(gl_predictions, nrow = 2, byrow = T)
```

Code Listing 3: Predicting Pass Rates and Estimating Confidence Intervals

Table 4's summarized data on predicted pass rates and their associated confidence intervals validate the initial observation: XYZ has better odds of passing the test in Worthing. The quality of these results, however, depends on the validity of the model in this setting. The model's potential limitations in fit might arise due to the sample size being relatively small or a breach in the assumption of independent samples. Such a violation of independence might occure due to the observed time trend in the data, as illustrated in Figure 2.

Table 4: Predicted Pass Rates for XYZ's Profile and Bootstrap-CIs

Location	**Prediction**	**95%-CI**
Worthing	**54.96%**	[53.49, 56.53]
Wood Green	**43.56%**	[42.38 , 44.74]

3.2 Directly Predicting Mean Difference Significance

To verify our results, instead of predicting pass rates for each location and computing confidence intervals, we can directly estimate whether Worthing's pass rate will be significantly higher in 2024 or not. By leveraging the central limit theorem, as outlined in the previous subsection, we can compute expected values and variances for the sample means. This facilitates Two-Sample Welch t-tests, enabling the significance assessment of mean differences among specific subgroups across test centers (e.g., comparing 23-year-old males' pass rates in 2008 across different locations). For the overlap of the two test centers' time series, we test the significance of the mean differences between all corresponding subgroups. Formally, we test:

$$H_0 : \mu_1 - \mu_2 = 0, \qquad H_1 : \mu_1 - \mu_2 \neq 0$$

With the test statistic and degrees of freedom:

$$\text{t-statistic} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}, \qquad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

The code snippet below describes the procedure in more detail:

```
# Function to calculate t-tests (Welch two-sample)
significant_mean_diff <- function(passrate, conducted) {
  n1 <- conducted[1]
  n2 <- conducted[2]
  mean_x <- passrate[1]/100
  mean_y <- passrate[2]/100
  var_x <- (mean_x*(1-mean_x)) / n1
  var_y <- (mean_y*(1-mean_y)) / n2
  df <- ((var_x / n1 + var_y / n2)^2) / # degrees of freedom
    ((var_x^2) / (n1^2 * (n1 - 1)) + (var_y^2) / (n2^2 * (n2 - 1)))
  t_stat <- (mean_x - mean_y) / sqrt((var_x / n1) + (var_y / n2))
  p_value <- 1 - pt(t_stat, df) # p-value
  return(p_value<0.01)}
# select relevant data and compute significance of mean difference
binom_data <- uk_df %>%
  filter(Location=="Wood Green (London)" | Location=="Worthing" ) %>%
  mutate(Year = as.numeric(paste("20",substr(Date,6,7), sep = ""))))%>%
  filter(Year <= 2018) %>% select(-Date) %>%
  arrange(Age,Year, Gender, desc(Location) ) %>%
  group_by(Age,Year, Gender) %>%
  mutate(SignificantMeanDiff=significant_mean_diff(Passrate,Conducted))
# evaluate results for XYZs profile in Worthing
sup_worthing_xyz <- binom_data %>%
  filter(Location == "Worthing", Age == "19", Gender == "F",
         SignificantMeanDiff==T)
```

Code Listing 4: Computing t-Tests on Sample Mean Differences across Subgroups

Running the code reveals, that the mean pass rate for XYZ's profile is significantly higher in Worthing for all years on record (2008-2018). Since XYZ is taking the test in 2024, our objective is to predict whether this will also be the case for that year. Given that this is again a problem of binary classification, we achieve this by fitting a logistic regression model on *SignificantMeanDiff*. The resulting predicted probability is 100%.[5] Therefore, we anticipate that XYZ's expected probability of passing the test will remain significantly higher in Worthing, also for the year 2024. The following code outlines the details:

[5]For comparison: If XYZ were 25 years old, this value would drop to 96.99%.

```r
worthing_data <- binom_data %>% #select Worthing
    filter(Location=="Worthing" )  %>%
    select(SignificantMeanDiff,Age,Gender,Year)
# Fit a logistic model to classify significance of the difference
gl_mod <- glm(SignificantMeanDiff ~ Age + Gender + Year,
    data = worthing_data, family = binomial)
# Create new df with the profile of XYZ for the logistic model
xyz_worthing_2024 <- data.frame(Year = 2024, Age = as.factor(19),
    Gender = "F")
# Predicted prob of the PR being significantly higher in Worthing
pred_worthing <- predict(gl_mod,
    newdata = xyz_worthing_2024, type= "response")
cat(pred_worthing)
```

Code Listing 5: Predicting the Significance of the Mean Difference in 2024

While producing a direct and unambiguous answer to the question at hand, one notable limitation of the previously described approach is its dependency on a logistic model trained on a highly imbalanced dataset: Worthing predominantly exhibits a considerably higher mean pass rate, leading to a scarcity of negative labels within the sample.

3.3 Simple Two-Sample Hypothesis Test for Pass Rate Difference

A simpler method of analysis is to calculate the mean pass rate for a test taker with XYZ's profile across the years at each location and conduct a single Welch Two-Sample t-test on the difference. The formal test setup parallels the one outlined in the previous subsection with $\bar{X}_1 = 49.98$, $\bar{X}_2 = 40.32$, $n_1 = 11$, $n_2 = 10$, ${s_1}^2 = 5.19$, ${s_2}^2 = 3.98$, and $df = 17.912$. In this case, the test can be implemented using R's `t.test()`. The computation indicates a p-value below the 5% level, allowing us to reject the null hypothesis of no significant mean difference. This aligns with the conclusions drawn from other methods. While this approach is simple, the upward trend observed in the two time series it relies upon, along with the difference in their endpoint year (Figure 2), might again impair the validity and predictive power of its results.

4 Conclusions

The analysis outlined above indicates that XYZ should take the practical driving test at the test center in Worthing. All tests suggest Worthing as the preferable choice. While acknowledging analysis limitations, including data range discrepancies, the influence of time trends, and other potential independence violations, the consistency of results from various methodologies strongly favors Worthing. However, access to individual-level test data would significantly enhance the depth of analysis possible.