ST447 Project

Candidate Number: 44051

2024-12-06

Meeting XYZ

Below, using the XYZprofile() function provided, we obtain and note the short profile of our friend XYZ. She is 22, female, and her closest driving test center is Burton on Trent. Lastly, given in the project instructions, we note the closest driving test center to LSE to be Wood Green (London).

```r
XYZprofile(ID)
```

```
## The profile of XYZ:
## - Age:  22
## - Gender:  Female
## - Home address:  Burton on Trent
```

```r
closest_XYZ <- "Burton on Trent"
closest_LSE <- "Wood Green (London)"
```

Reading-in and Procuring Data

Looking at the data we can see that from 2013-14 onwards, the test centre that follows Burton on Trent every year is Bury (Manchester), and similarly, the test centre that follows Wood Green (London) every year is Worcester. We can use this information to make new functions to retrieve the data most relevant to XYZ.

One thing to keep in mind is that the columns of the original data are broken up into *Male*, *Female*, and *Total*, and then subgroups of *Conducted*, *Passes*, and *Pass Rate*. Also note the lack of column names in their typical place in the original data set means they are reassigned names by R, and R typically outputs messages describing the reassignment, but I suppress the messages for those reassignments below.

```r
get_Burton_on_Trent <- function(year_or_sheet){

  # This function takes as input the year of interest from the .ods data file
  # between 2017-18 and 2023-24 and returns a data frame of the data from
  # Burton on Trent for all ages and sexes from that year

  read_in <- read_ods(path = "/Users/finbarrhodes/Documents/ST447/dvsa1203.ods", sheet = year_or_sheet)

  start_index <- which(read_in[,1] == closest_XYZ) + 1 # Finding where Burton on Trent appears in the d
  end_index <- which(read_in[,1] == "Bury (Manchester)") - 1 # Finding where Bury (Manchester) appears

  new_data <- read_in[(start_index):(end_index),2:11]
  names(new_data)[1:10] <- read_in[6,2:11] # Giving appropriate names to the new data's columns
  return(new_data)
}

get_Wood_Green <- function(year_or_sheet){

  # This function takes as input the year of interest from the .ods data file
  # between 2017-18 and 2023-24 and returns a data frame of the data from
  # Wood Green (London) for all ages and sexes from that year.

  read_in <- read_ods(path = "/Users/finbarrhodes/Documents/ST447/dvsa1203.ods", sheet = year_or_sheet)

  start_index <- which(read_in[,1] == closest_LSE)+1 # Finding where Wood Green (London) appears in the
  end_index <- which(read_in[,1] == "Worcester")-1 # Finding where Worcester appears in the data

  new_data <- read_in[(start_index):(end_index),2:11]
  names(new_data)[1:10] <- read_in[6,2:11] # Giving appropriate names to the new data's columns
  return(new_data)
}
```

Below I outline two data frames, one for Burton on Trent and one for Wood Green (London), and fill them with the appropriate data using the functions **get_Burton_on_Trent()** and **get_Wood_Green()** above as well as some manual retrieval due to inconsistencies in the original data.

```
BoT_rates
```

```
##       Year Female.Rate Age.Rate Joint.Rate
## 1  2023-24    55.13308 50.00000   48.51485
## 2  2022-23    51.62534 47.48603   48.00000
## 3  2021-22    49.13753 47.76119   48.57143
## 4  2020-21    50.88853 66.66667   73.68421
## 5  2019-20    49.19169 56.59341   53.53535
## 6  2018-19    48.88543 46.35417   43.11927
## 7  2017-18    55.17600 59.60000   55.80000
## 8  2016-17    50.27200 51.80000   47.10000
## 9  2015-16    47.63700 52.20000   40.70000
## 10 2014-15    47.28800 55.80000   48.00000
```

```
WG_rates
```

```
##       Year Female.Rate Age.Rate Joint.Rate
## 1  2023-24    44.51745 44.36860   41.42857
## 2  2022-23    45.41797 47.68683   47.47082
## 3  2021-22    45.07659 47.52475   47.47475
## 4  2020-21    44.93308 50.49505   45.28302
## 5  2019-20    39.91826 40.08621   39.46188
## 6  2018-19    39.06883 41.35338   35.49618
## 7  2017-18    38.06800 36.70000   33.20000
## 8  2016-17    40.22300 43.80000   40.30000
## 9  2015-16    37.29900 39.60000   34.80000
## 10 2014-15    36.90600 38.50000   38.70000
```

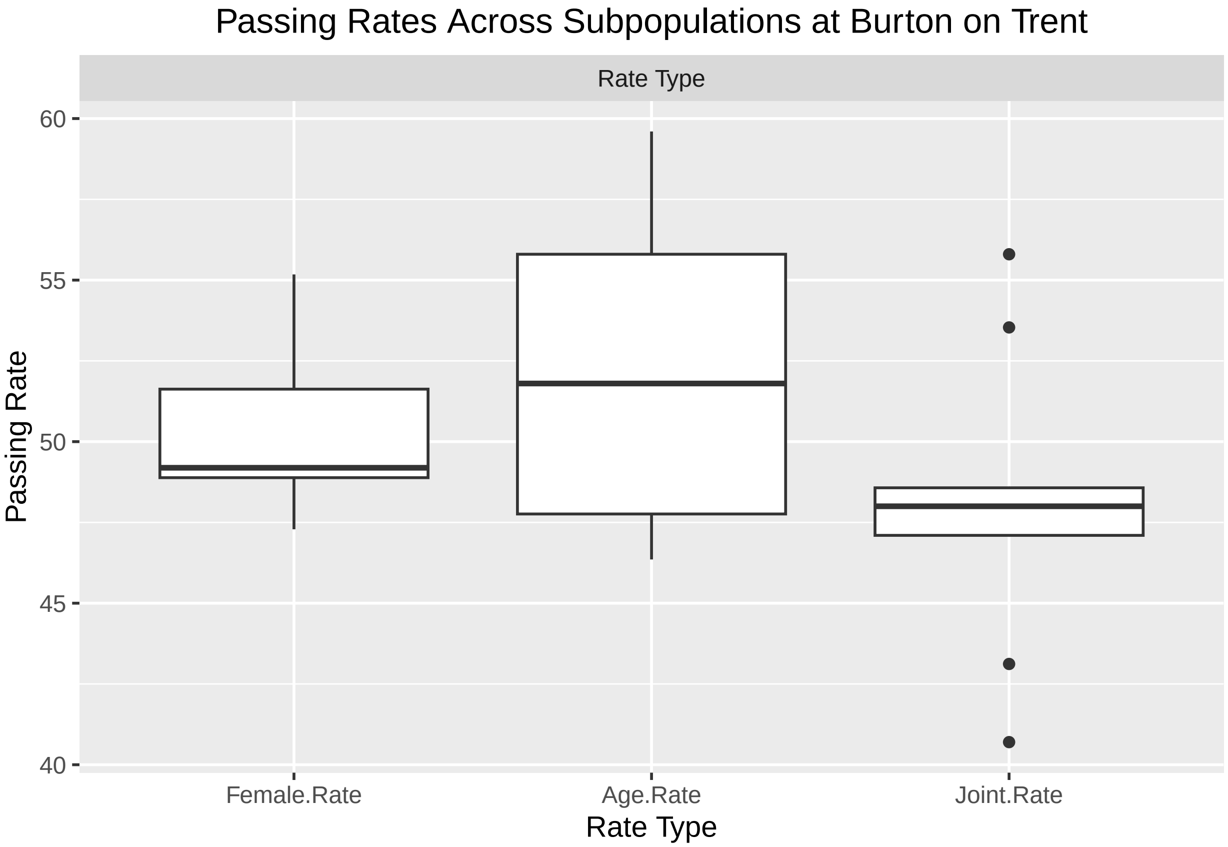
Looking at the raw .ods data we can notice that the driving test counts in 2020-21, the year of the COVID-19 pandemic, are far lower than the other years. This could potentially impact our estimates, as an implicit assumption of using the pass rates is that the counts underlying them are sufficiently large. The counts for the *Age Rate* and the *Joint Rate* are small (19 and 36 total tests respectively). In our analysis, we will be using the Permutation Test, a relatively flexible statistical test, that rests on an assumption that the data points are at the very least exchangeable. Removing the data from the year 2020-21 ensures this, and so we do that below.

```
BoT_rates <- BoT_rates[-which(BoT_rates$Year == "2020-21"),]

WG_rates <- WG_rates[-which(WG_rates$Year == "2020-21"),]
```

1. What is the expected passing rate at Burton on Trent?

In order to give an estimate of the expected passing rate at the Burton on Trent driving test center, I have gathered data relevant to the profile of XYZ given above, retrieving the marginal rates for female driving test-takers and test-takes the same age as XYZ, as well as the joint rate for test-takers her age and sex. Here we plot the rates to get a general idea of the data:



Now that we have our sample data most relevant to XYZ, we can provide XYZ with an expected pass rate of the sample data by taking an average over the past ten years.

```
## The expected passing rate over the past ten years for female test-takers at Burton on Trent is 50.48
```

We can take a different approach using the Age Rate, where instead of using XYZ's sex to better estimate her expected pass rate, we can use her age.

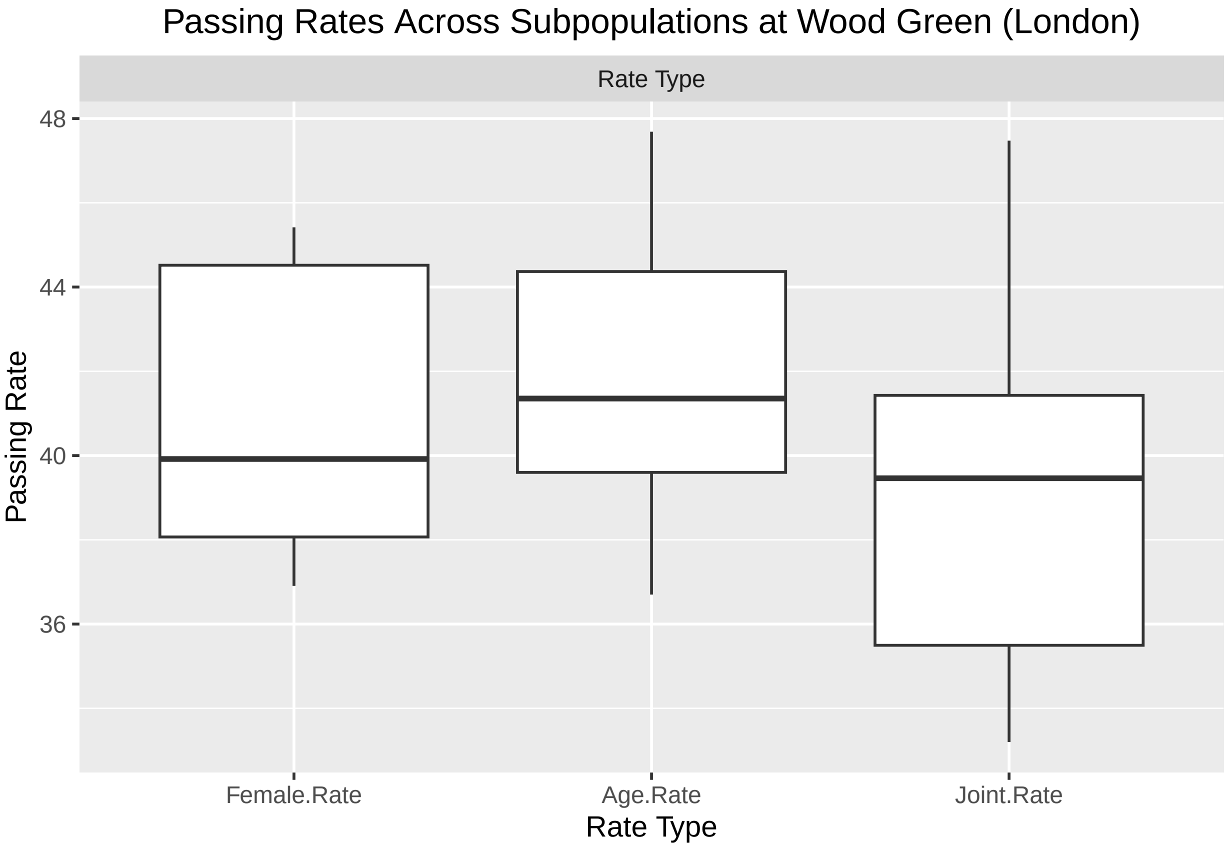
```
## The expected passing rate over the past ten years for 22 year-old test-takers at Burton on Trent is
```

Lastly, we can use a more specific lens to answer this question which would be to use the joint distribution of 22 year-old female test-takers.

```
## The expected passing rate over the past ten years for 22 year-old female test-takers at Burton on Tr
```

2. What is the expected passing rate on Wood Green (London)?

To answer question (2.), we take a similar approach to question (1.), but substituting the sample data from Burton on Trent for the sample data from Wood Green (London).



The sample mean for female test-takers from the Wood Green (London) driving center is

```
## The expected passing rate over the past ten years for female test-takers at Wood Green (London) is 4
```

Looking at the problem from a different angle, the sample mean from for 22 year-old test-takers at Wood Green (London) is

```
## The expected passing rate over the past ten years for female test-takers at Wood Green (London) is 4
```

Finally, a view of the joint distribution, taking a sample mean of 22 year-old female test-takers at the Wood Green center is

```
## The expected passing rate over the past ten years for female test-takers at Wood Green (London) is 3
```

I offer these six estimates across the two locations to provide a slightly more holistic estimate than just looking at one of these marginal distributions or the joint one on their own. It is worth noting that these estimates rest on a the assumption that there have been no changes in test administration or difficulty level over time. This is why it was important to investigate the rates in 2020-21, and why discarding them likely aids the soundness of our analyses.

3. Out of these two locations, where should XYZ take their test?

Due to the small sample sizes for each test center (n=9), it is not statistically sound to rely on asymptotic assumptions of the underlying distributions. In this situation, we can rely on the Permutation Test to judge which test center XYZ ought to go to to give her the best chances at passing her driving exam. The permutation test can be used to see if two samples originate from the same underlying distribution and it does not rely on any asymptotic assumptions or properties of the data. We can use it reasonably here and don't have to presume any underlying distribution of the driving test pass rates, we can just test if the rates across the two locations could feasibly come from the same distribution. More explicitly, our null hypothesis is that the data from the two test locations, when combined, form a single distribution equal to that of its parts, or equivalently, that the distributions from the two test locations are equal. If we find evidence to reject this null hypothesis, this suggests that they are not from the same distribution, and XYZ can, with reasonable confidence, go to the one with a higher average passing rate and believe it gives her a better chance at acing her driving exam.

Here we are implementing our Permutation Test across *Female* rates:

```r
T_obs <- abs(mean(BoT_rates$Female.Rate)-mean(WG_rates$Female.Rate))
z <- c(BoT_rates$Female.Rate,WG_rates$Female.Rate)
k <- 0
for(i in 1:10000) {
  permutation <- sample(z, length(z)) # permutation is a permutation of z
  t <- abs(mean(permutation[1:(length(z)/2)])-mean(permutation[((length(z)/2) + 1):length(z)]))
  if (t > T_obs){
    k <- k + 1
  }
}
cat("Permutation Test p-value:", k/10000)
```

```
## Permutation Test p-value: 0
```

Now similarly across the *Age* rates:

```r
T_obs <- abs(mean(BoT_rates$Age.Rate)-mean(WG_rates$Age.Rate))
z <- c(BoT_rates$Age.Rate,WG_rates$Age.Rate)
k <- 0
for(i in 1:10000) {
  permutation <- sample(z, length(z))
  t <- abs(mean(permutation[1:(length(z)/2)])-mean(permutation[((length(z)/2) + 1):length(z)]))
  if (t > T_obs){
    k <- k + 1
  }
}
cat("Permutation Test p-value:", k/10000)
```

```
## Permutation Test p-value: 2e-04
```

Finally, across the *Joint* rates:

```
T_obs <- abs(mean(BoT_rates$Joint.Rate)-mean(WG_rates$Joint.Rate))
z <- c(BoT_rates$Joint.Rate,WG_rates$Joint.Rate)
k <- 0
for(i in 1:10000) {
  permutation <- sample(z, length(z))
  t <- abs(mean(permutation[1:(length(z)/2)])-mean(permutation[((length(z)/2) + 1):length(z)]))
  if (t > T_obs){
    k <- k + 1
  }
}
cat("Permutation Test p-value:", k/10000, "\n")
```

```
## Permutation Test p-value: 0.002
```

Conclusion

In all three implementations of the Permutation Test, we find that the p-value is below a significance level of .01, giving us ample evidence to reject the null hypothesis in all three cases that the data from Burton on Trent and Wood Green (London) have equal distributions. We have shown above that on average, over the last nine years of reasonable test-taking and societal conditions, that the average passing rate at Burton on Trent is higher than that of Wood Green (London). Thus, I would suggest for XYZ to conduct her driving test in the location closest to her home, in Burton on Trent. It is worth noting that collecting more years of passing rate, or using the count data itself, or using a metric to determine test centers most similar to the Burton on Trent and/or Wood Green (e.g traffic data, or simply expanding to closest x centers near home/LSE) and using those test centers in analysis could improve our estimates and potentially our suggestion.