Stats_project

December 5, 2024

1 Abstract

My advice to XYZ (21 year old woman from Reading), supported by statistical evidence, is to take the driving exam in Reading. However the data is not perfect. We do not know the skills/practice of people who take the exam in Reading and London. This unknown distribution of skills could affect both my recommendation on where to take the exam and the predicted passing rate for her.

[1]:
```
ID <- 202215485
source("XYZprofile.r")
XYZprofile(ID)
# With average driving skills.
```

```
The profile of XYZ:
- Age:  21
- Gender:  Female
- Home address:  Reading
```

[19]:
```
library(dplyr)
set.seed(100)
```

2 Exploratory data analysis:

I downloaded and cleaned the data using Python, this code is ommited due to space limitations. I then created the plot below which serves as exploratory data analysis.

[3]:
```
passing_rates_reading <- read.csv("passing_rates_reading.csv")
```

[4]:
```
passing_rates_reading[15:20,]
```

A data.frame: 6 × 7

	Location <chr>	Age <dbl>	Conducted <dbl>	Passes <dbl>	Pass.rate.... <dbl>	gender <chr>	year <chr>
15	Wood Green (London)	22	306	144	47.05882	men	2023-2
16	Wood Green (London)	23	323	151	46.74923	men	2023-2
17	Wood Green (London)	24	317	131	41.32492	men	2023-2
18	Wood Green (London)	25	292	128	43.83562	men	2023-2
19	Reading	17	309	188	60.84142	women	2023-2
20	Reading	18	296	162	54.72973	women	2023-2

I use this exploratory data analysis to inform which part of the sample I use in posterior analysis.

First, let us look at the trends across time - left column of the plot -, we can infer some important facts from this data (which is aggregated across different ages). The passing rates in London are lower for both men and women. There seems to be an upward trend in passing rates both in London and Reading over the sample years, however, in the post covid years there seems to be a negative trend, especially in Reading. Also notice that since Covid, there has been a big decrease in the total tests taken in Reading. Finally, males pass the exam at higher rates, especially in London.

Since there seems to be some uncertainty across time -both in passing rates and exams taken- I will only use the last year of available data. Since there are also large differences across gender, I will only be using women to test where XYZ -who is a 21 year old woman- should take the exam.

Now we can look at the differences across ages for the year 2023-24 (right column). We can immediately see that there are differences across ages, some may be due to noise/randomness while others may be due to concrete reasons. I would argue that people that take a driving test at 17 are very motivated and prepared. I do not think these people are equivalent to others who take the driving test later in their lives. On the other hand, I would argue that differences between 21 and 22 years olds are probably due to randomness.

To start, I will take the most restrictive approach possible, I will only compare women aged 21 that took the driving test in 2023-24. That will leave me with a small sample and from there I will relax my assumptions (including women of other ages).

3 Predicted passing rate.

[5]:
```
# This function preps the data for any subset I want. In all cases I only use␣
↪data from 2023-24

probs_for_test <- function(df,age=c(21)){
    #Works for any age grouping. Including a single age.

    p_rate_age_subset <- subset(df, Age %in% c(age))

    # I want to sum over age keeping gender year and location constant.

    p_rate_age_subset <- p_rate_age_subset %>% group_by(gender, year, Location)␣
↪%>%
      summarise(conducted=sum(
      Conducted), passes = sum(Passes),
            .groups="keep")

    #Only use 2023-24 data.

    p_rate_age_subset <- data.frame(subset(p_rate_age_subset, year=="2023-24" &␣
↪gender=="women"))

    rownames(p_rate_age_subset) <- p_rate_age_subset$Location

    #Put the elements in a nice vector for subsequent analysis.
```

```
    pass_r <- p_rate_age_subset["Reading", "passes"]
    tot_r <- p_rate_age_subset["Reading", "conducted"]
    pass_l <- p_rate_age_subset["Wood Green (London)", "passes"]
    tot_l <- p_rate_age_subset["Wood Green (London)", "conducted"]

    data_for_test <- c(pass_r,tot_r,pass_l,tot_l)

    return (data_for_test)
}
```

[6]:
```
cond_mean_boot <- function(params, iter){
    # The pre processing function will give the parameters in the following␣
 ↪order:
    pass_r <-params[1]
    total_r <- params[2]
    pass_l <- params[3]
    total_l <-params[4]

    estimate_r <- round(pass_r/total_r, digits = 2)
    estimate_l <- round(pass_l/total_l, digits = 2)

    return(c(estimate_r,estimate_l))

}
```

[7]:
```
param <- probs_for_test(passing_rates_reading,c(21))
cond_mean_boot(param,10000)
```

1. 0.5 2. 0.44

[8]:
```
param <- probs_for_test(passing_rates_reading,c(20,21,22))
cond_mean_boot(param,10000)
```

1. 0.49 2. 0.43

[9]:
```
param <- probs_for_test(passing_rates_reading,c(19,20,21,22,23))
cond_mean_boot(param,10000)
```

1. 0.48 2. 0.43

Predicted passing rate (women in 2023/24)	Women aged 21	Women aged 20-22	Women aged 19-23
Reading	0.5	0.49	0.48
London	0.44	0.43	0.43

To estimate XYZ's passing rate one can just divide the total number of people who pass over the total number of people who take the exam. The difficult part is to determine who one should count.

Following my exploratory analysis I only count women aged 21 for the first analysis and expand it to women aged 20-22 for the second one; arguing that women in this age range are very similar. In all cases I only use data from 2023-24.

In Reading, the expected probability of passing the exam seems to be roughly 50% for women. In London this expectation is a bit smaller, roughly 44% of women pass the test.

This analysis requires weak conditions (Law of Large Numbers) that are satisfied, however, it is important to know that this expected passing rate is for a group of women aged X taking the exam in Reading or London (for 2023-24). If we knew the distribution of skill/practice time amongst these women, I would be able to give a more precise prediction for XYZ. I come back to this in the Limitations section.

Technical note: I built a confidence interval for completeness using non-parametric bootstrap but I removed it as it was related to the following testing part.

4 Where to take the test? Are the passing rates for women higher in Reading?

Using the Permutation test I arrive at the conclusion that XYZ should take the exam in Reading. This test relies on few assumptions. I previously use the Wald test which gives similar results under different stronger assumptions.

4.1 Wald test:

This is a statistical test that can be derived under certain assumptions.

Issue 1: This is a test that requires you to have a very big sample so it might not work for my limited ones, especially in the case where I only include women aged 21.

Issue 2: This test requires an assumption on the distribution of the data that is not likely to hold. It basically requires that passing the exam is like flipping the same coin for everyone. But we know that people go in to the exam having practiced different amounts of time. Therefore, a different coin is being flipped for each individual. For those that practice more, the coin that is flipped is more likely to result in a pass.

Technical note: This can be derived using MLE under the assumption that the data comes from two independent Binomial distributions.

[10]:
```
wald_test <- function(params){
    pass_r <-params[1]
    size_r <- params[2]
    pass_l <- params[3]
    size_l <-params[4]

    prob_r <-pass_r/size_r
    prob_l <-pass_l/size_l

    # Assuming a Binomial distribution this is the standard error of the␣
 ↪estimator.
```

```
    se_mle <- sqrt( (prob_r)*(1-prob_r)/size_r + (prob_l)*(1-prob_l)/size_l )

    WALD <- (prob_r-prob_l)/se_mle

    return(WALD)
}
```

We are looking for values as close to 0 as possible. Those would allow us to say that it is better to take the exam in Reading.

[11]:
```
p_val_wald <-function(wald) {
    # Use a one sided test. Greater than. Reasonable assumption given that␣
 ↪Reading has always had a higher passing rate.
    p_val <-(1-pnorm(wald))
    return(round(p_val,digits=2))
}
```

[12]:
```
p_val_wald(wald_test(probs_for_test(passing_rates_reading,c(21))))
```

0.13

Under the most strict assumptions we have moderate evidence in favour of XYZ going to Reading. We can try to add more similar people, women aged 20 and 22 to our sample.

[13]:
```
p_val_wald(wald_test(probs_for_test(passing_rates_reading,c(20,21,22))))
```

0.03

This now shows strong evidence in favour of taking the exam in Reading. However, our assumptions may not be correct so this number could unfortunately be meaningless.

[14]:
```
p_val_wald(wald_test(probs_for_test(passing_rates_reading,c(19,20,21,22,23))))
```

0.03

4.2 Permutation test:

This is a very nice test that allows to compare two distributions or any statistic based on those distributions. It makes few assumptions, works for any distributions and does not require the sample to be very large. It is because of this lack of assumptions that I prefer to follow the recommendations from this test (although both test give very similar results). The results from this test lead me to recommend taking the exam in Reading.

[15]:
```
permutation_test <- function(params, iter){
    # The pre processing function will give the parameters in the following␣
 ↪order:
    pass_r <-params[1]
    total_r <- params[2]
    pass_l <- params[3]
    total_l <-params[4]
```

```
    # We generate a fake sample, for London and Reading respectively
    # They will have the same number of passes and fails observed in the data.
    sample_r <- c(numeric(pass_r)+1, numeric(total_r-pass_r))
    sample_l <- c(numeric(pass_l)+1, numeric(total_l-pass_l))

    full_sample <- c(sample_r, sample_l)

    original_statistic <- mean(sample_r) - mean(sample_l)

    larger_than_og <-c()

    for (i in 1:iter){
        # Sampled without replacement. -> Permutation
        # There may be repeated permutations but the probability is small so I␣
 ↪will not correct for it as it would be very expensive to check.
        permutation <- sample(full_sample)
        # The first elements will be "reading people"
        perm_r <- permutation[1:total_r]

        #These are "london people"
        perm_l <- permutation[(total_r+1):length(permutation)]

        # Compute the statistic with the permuted sample.
        perm_statistic <- mean(perm_r) - mean(perm_l)
        # Keep TRUE=1 and FALSE=0
        larger_than_og <- c(larger_than_og, perm_statistic > original_statistic)

    }

    p_val_perm <-mean(larger_than_og)
    return(round(p_val_perm,digits=2))
}
```

[16]:
```
params <- probs_for_test(passing_rates_reading,c(21))
permutation_test(params,10000)
```

0.11

We can see that even in the case where we only include women aged 21 we have moderately strong evidence in favour of going to Reading to take this exam.

[17]:
```
params <- probs_for_test(passing_rates_reading,c(20,21,22))
permutation_test(params,10000)
```

0.03

If we include women aged 20-22, we have very strong evidence in favour of going to Reading.

[18]:
```
params <- probs_for_test(passing_rates_reading,c(19,20,21,22,23))
permutation_test(params,10000)
```

0.03

What do these numbers mean? If the candidates repeated the test 100 times and the probabilities of passing were the same in both London and Reading, we would falsly recommend to go to Reading 100*z times. We are therefore looking for small numbers (z), 0.1 or 0.05.

We can see in the table below that even under the most stringent assumptions - to only consider women aged 21 as a valid comparison group - the permutation test gives moderate evidence towards discarding the baseline hypothesis - passing rates being the same in both locations -. For the assumption that women aged 21 are identical in all important aspects to women aged 20 and 22, both tests give strong evidence towards taking the exam in Reading.

Test p-values (women in 2023/24)	Women aged 21	Women aged 20-22	Women aged 19-23
Wald test	0.13	0.03	0.03
Permutation test	0.11	0.03	0.03

Technical note: I have used one sided alternative hypothesis to check where to take the exam. This is less conservative than the alternative hypothesis being $\mu_R \neq \mu_L$. For completeness I ran two sided tests, and while the p values roughly double, the message is similar, in favour of taking the exam in Reading.

5 Limitations

- Time series: I largely avoid making predictions for this present year (2024-25) -which would be the most relevant quantity for XYZ-, because that is beyond the scope of this course. But it does look like passing rates in Reading are going down. And, since post covid, the number of exams taken in Reading have been going down, so we might need to further research these temporal changes.
- Skill differences: My estimators aggregate/integrate over an unkown skill distribution that might vary across location. Practice/skill is a very important predictor on passing the exam. This might bias our results in two different ways. e.g.

  The exam is very difficult in London so people take more lessons to prepare than in Reading. Therefore, if we conditioned on practice time, the difference in passing rates between Reading and London would be greater. Passing in Reading would be much more likely than passing in London.

  On the other hand, practice lessons are very expensive in London and that may cause people to practice less. If we conditioned on practice time- compare people with the same practice time -, the difference in passing rates could be smaller or even negative. That might be because although driving in London is more difficult, there is a lot of traffic, which minimizes opportunities to make mistakes.

Note that both these scenarios are in agreement with the fact that XYZ reports she has average skill/practice time.