Homeworks for James' Stats classes

Zhifei Yu

2023-7-14

Contents

Welcome!

Howdy! This is a collection of 4 assignments from James Scott's probability and Statistics courses, written by Zhifei Yu a.k.a Kevin.

1 Homework 1

1.1 Problem 1 Playlists revisited

1.1.1 Part A

	0	1
0	0.925	0.912
1	0.075	0.088

Column:

0: never plays Daft Punk

1: plays Daft Punk

Row:

0: never plays David Bowie

1: plays David Bowie

1.1.2 Part B

To check out if 2 events are independent, we can use the definition: If A and B are independent, then P(A|B) = P(A|~B) = P(A) To make it clear, “plays Pink Floyd” is considered as event B, “plays Johnny cash” is event A.

0	1
0.94	0.06

0: never plays Johnny Cash

1: plays Johnny Cash

	0	1
0	0.945	0.895
1	0.055	0.105

Column:

0: never plays Johnny Cash

1: plays Johnny Cash

Row:

0: never plays Pink Floyd

1: plays Pink Floyd

So, in this case P(A) = 6%, P(A|B) = 10.5%, P(A|not B) = 5.5%; clearly, they are not equal. Therefore, they are not independent and seem to have positive relationship.

Or we can check it by if P(B) = P(B|A) = P(B|~A)

0	1
0.895	0.105

0: never plays Pink Floyd

1: plays Pink Floyd

	0	1
0	0.9	0.817
1	0.1	0.183

Column:

0: never plays Pink Floyd

1: plays Pink Floyd

Row:

0: never plays Johnny Cash

1: plays Johnny Cash

Clearly, P(B) = 10.5%, P(B|A) = 18.3%, and P(B|~A) = 10%, so they are not close to each other.

1.2 Problem 2 Super Bowl ads

1.2.1 Part A

FALSE	TRUE
0.7	0.3

True: should be danger

False: not danger

Which returns the results that P(danger = TRUE) = 30%

	FALSE	TRUE
FALSE	0.88	0.61
TRUE	0.12	0.39

Column:

True: should be danger

False: not danger

Row:

True: should be funny

False: not funny

From the table, we can know that:

P(danger = TRUE | funny = TRUE) = 39%

P(danger = TRUE | funny = FALSE) = 12%

Undoubtedly, from this statistics, humor and danger are absolutely not independent because P(danger) P(danger|funny) P(danger|not funny)

It seems that humor are indeed more or less a indication of danger for this ads, because under the condition that ads are funny, the probability of danger seems to be higher than unconditional probability and under the another condition that ads are not funny, the probability of it shows way much lower than unconditional probability.

1.2.2 Part B

FALSE	TRUE
0.63	0.37

True: with animals False: without animals

Which returns the results that P(animals = TRUE) = 37%

	FALSE	TRUE
FALSE	0.63	0.62
TRUE	0.37	0.38

Column:

True: with animals

False: without animals

Row:

True: has sex contents

False: not have sex contents

From the table, we can know that:

P(animals = TRUE | use_sex = TRUE) = 38%

P(animals = TRUE | use_sex = FALSE) = 37%

From the probability tables and unconditional probability, I think animals and use_sex are statistically independent.My argument is that the unconditional probability of animals seems to be very close to the conditional probabilities on both conditions that using sex and not using, which, from definition, shows this 2 events are independent.

1.2.3 Part C

FALSE	TRUE
0.71	0.29

True: with celebrities

False: without celebrities

Which returns the results that P(celebrity = TRUE) = 29%

	FALSE	TRUE
FALSE	0.71	0.71

	FALSE	TRUE
TRUE	0.29	0.29

Column:

True: with celebrities

False: without celebrities

Row:

True: has patriotic contents

False: not have patriotic contents

From the table, we can know that:

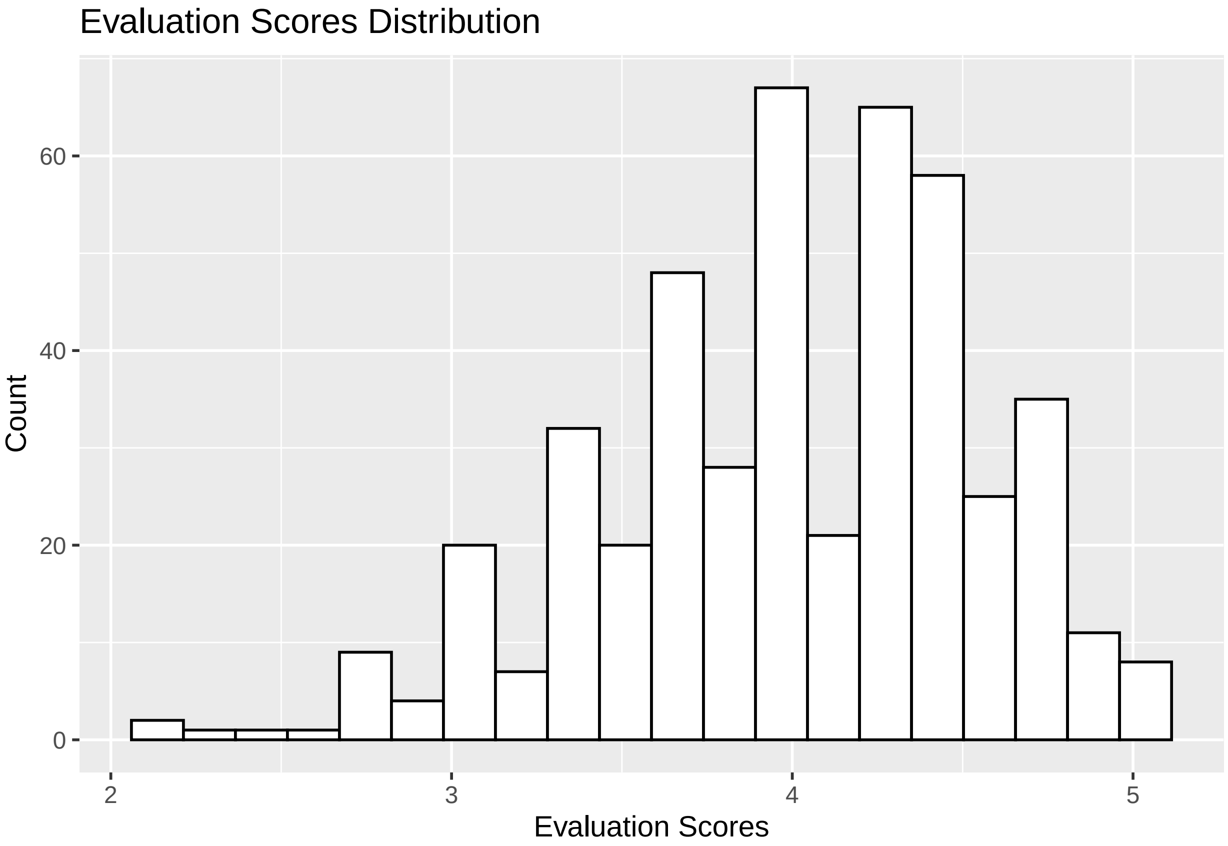
P(celebrity = TRUE | patriotic = TRUE) = 29%

P(celebrity = TRUE | patriotic = FALSE) = 29%

Similar with Part B, in this part, the unconditional probability of celebrity is nearly equal to the 2 conditional probabilities of both showing patriotic contents and not showing this. Thus, they are independent on the basis of this data.

1.3 Problem 3 Beauty, or not, in the classroom

1.3.1 Part A



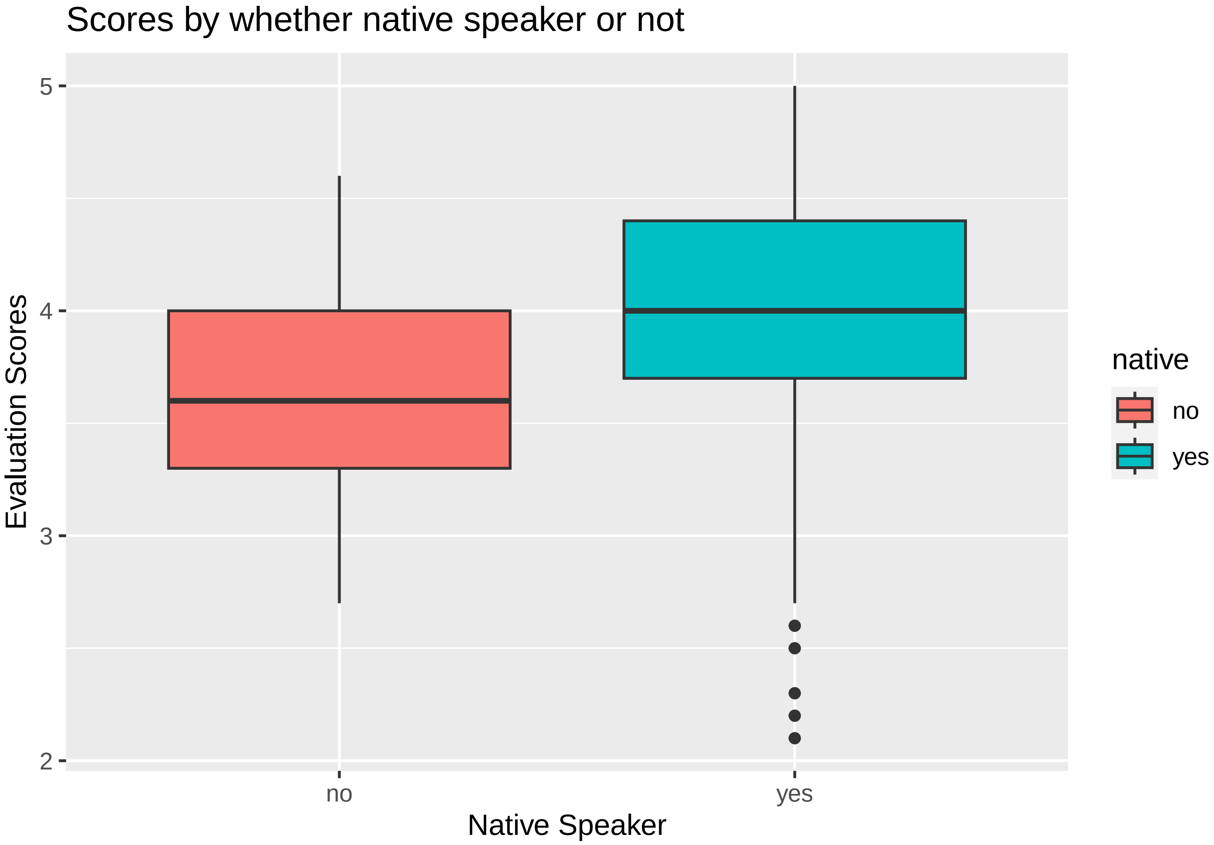
Above is the histogram plot that shows course evaluation scores of all professors.

X axis shows the evaluation scores which professors have gained.

Y axis shows tha counting of each score.

We can see around 4 is where the most scores are sited. So I guess that most UT's professor are pretty good so that they can get good evaluations scores from students.

1.3.2 Part B

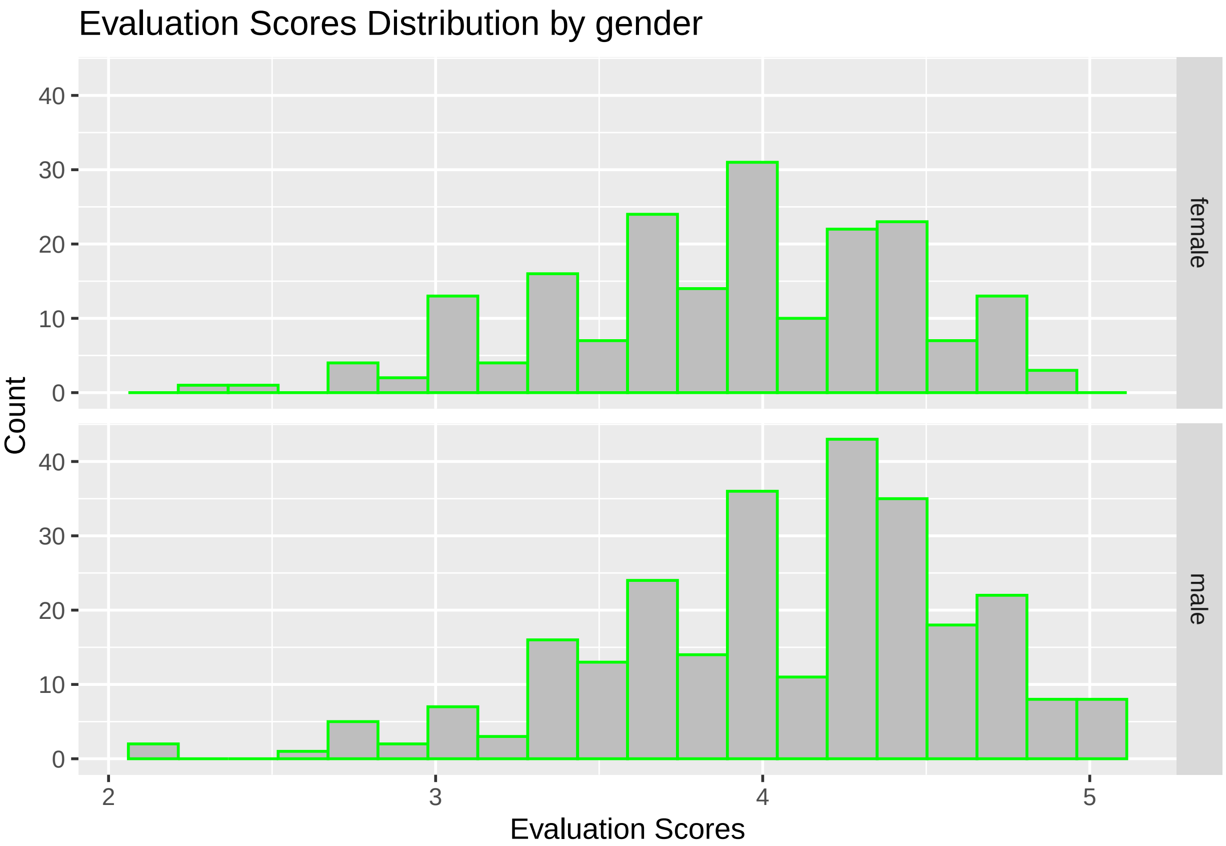


The left red box represents the non-native English speaker.

The right blue box stands for native speaker. Y axis means the evaluation scores.

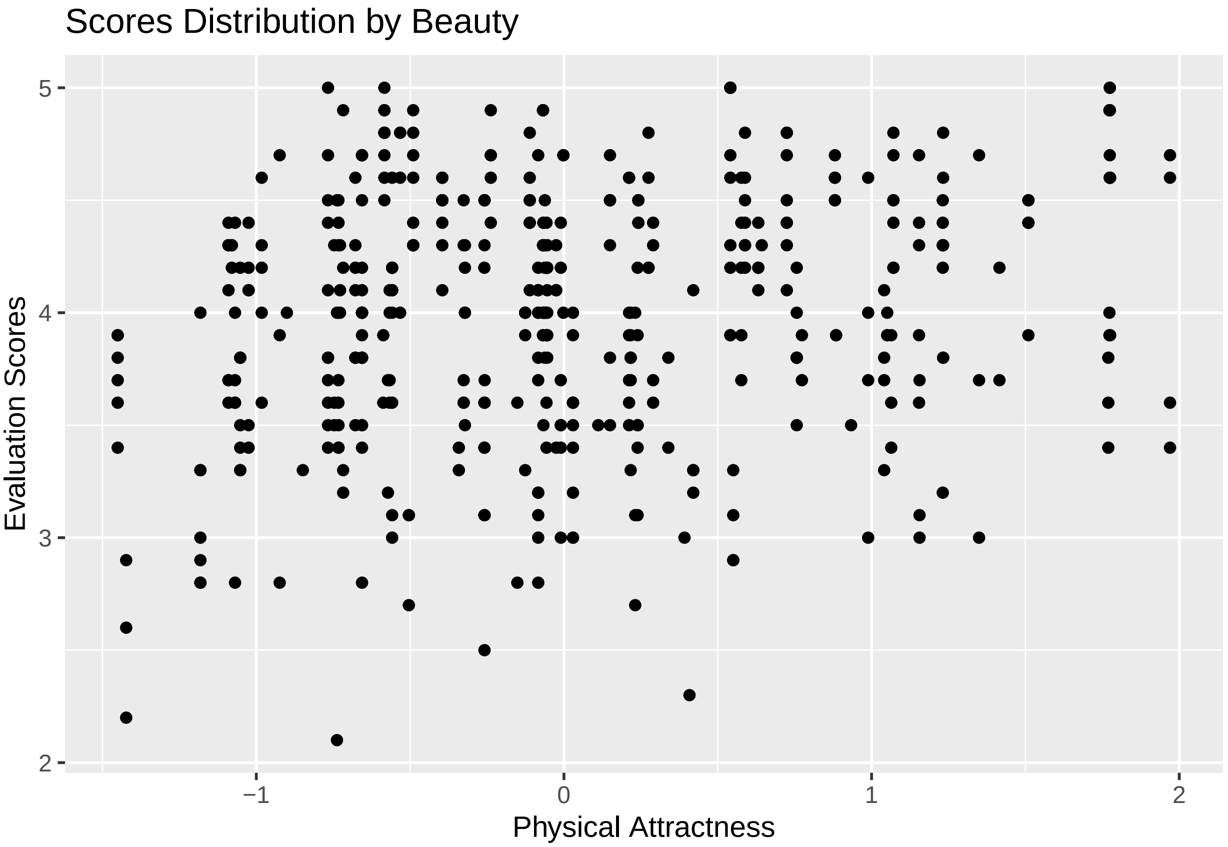
So, from this boxplot, we can conclude that native speakers generally get better score than non-native speaker.

1.3.3 Part C



From this plot, we can see that male professors are more focused around 4, whereas female professors are more spread-out.

1.3.4 Part D



X axis stands for the score of beauty.

I think that the physical attraction basically has slightly positive relationship with evaluation scores but cannot be a predictive indicator of evaluation.

1.4 Problem 4 SAT scores for UT students

Scores

Mean

Std

IQR

quan5

quan25

median

quan75

quan95

1

SAT-V

595.05

83.77

110.00

460.00

540.00

590.00

650.00

730.00

2

SAT-Q

619.98

83.08

120.00

480.00

560.00

620.00

680.00

760.00

3

GPA

3.21

0.48

0.72

2.36

2.87

3.25

3.59

3.92

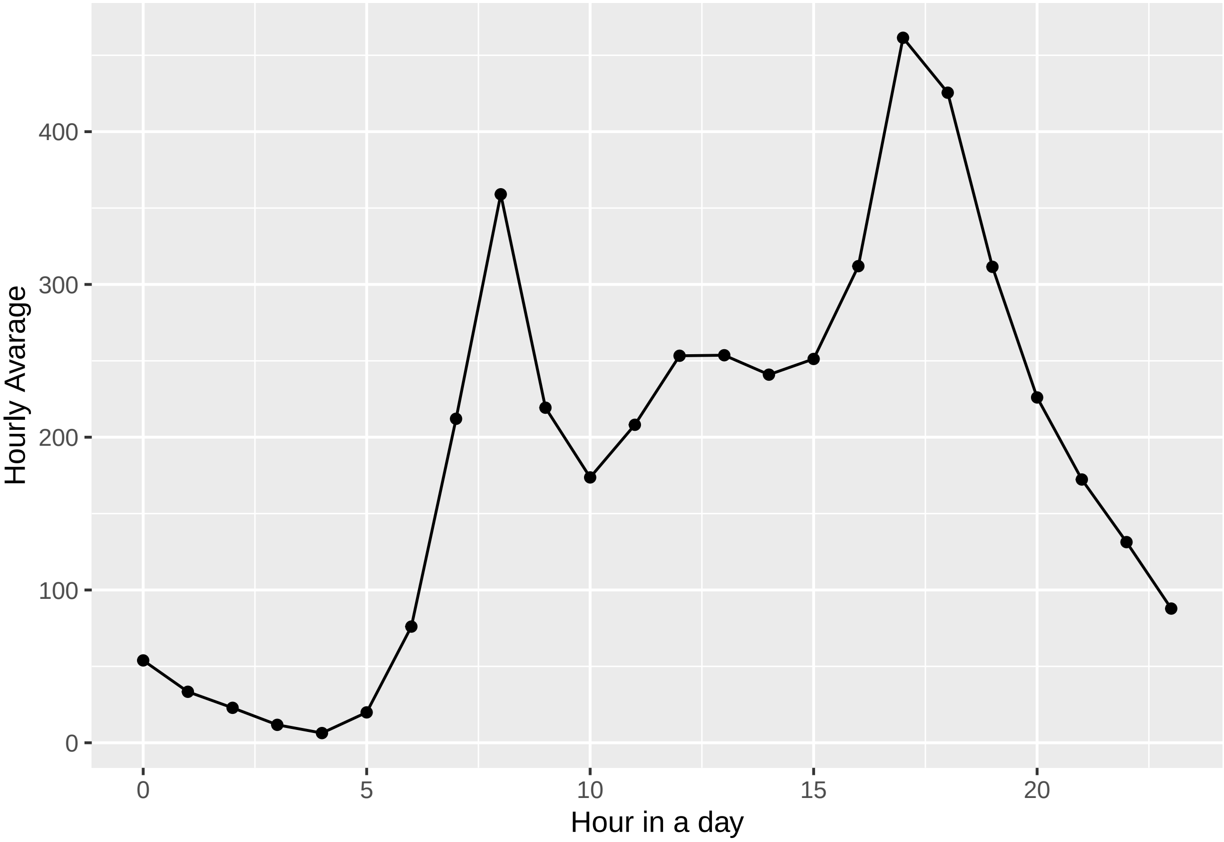
SAT-V means SAT verbal scores and SAT-Q means SAT quantitative score, while GPA means accumulative grade points.

Mean is the average of each score, std means standard deviation, IOR is inter-quantile range.

Quan5 is 5th percentile and so on so forth.

1.5 Problem 5 bike sharing

1.5.1 Plot A

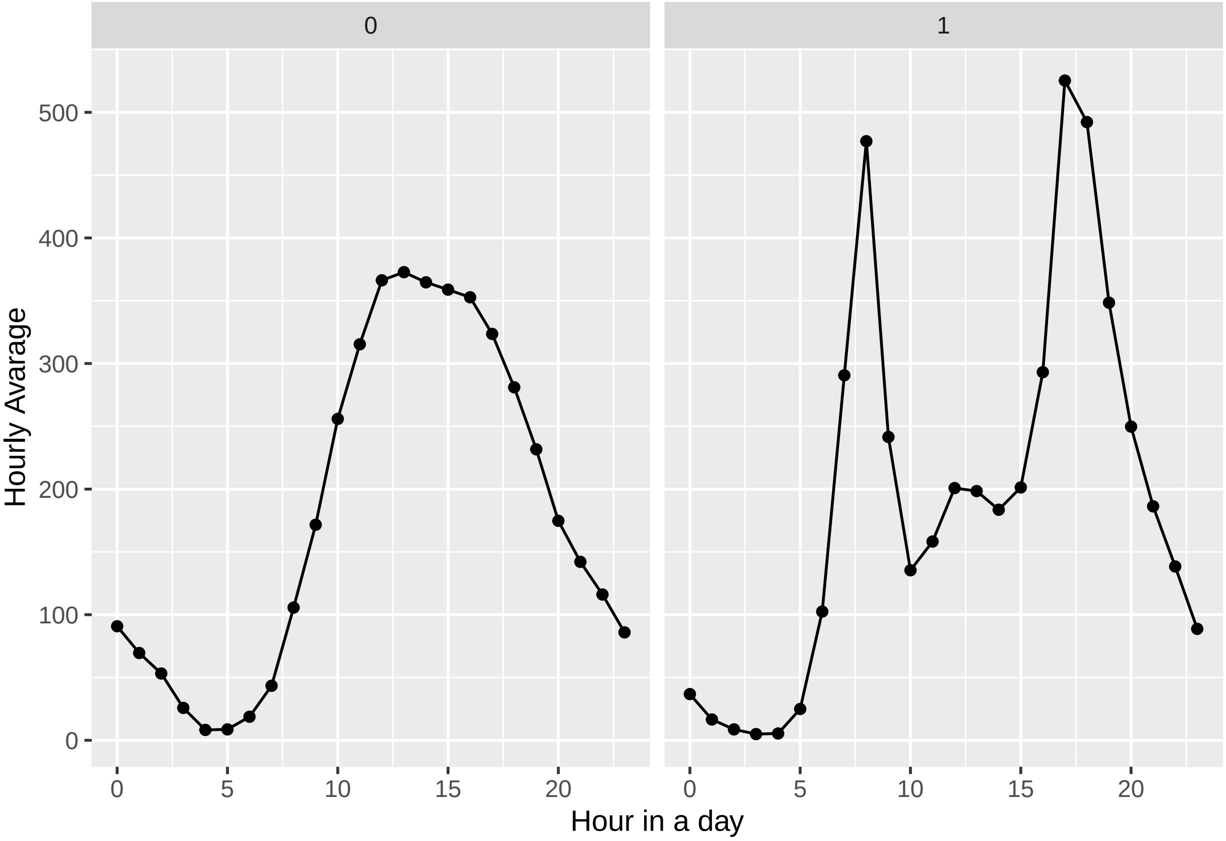


In this plot, x axis stands for the 24 hours in a single day. 0 is midnight and 10 is 10a.m., so on so forth.

Y is the average ridership of each hour throughout all days in this data.

We can see that the average ridership around morning rush hour and afternoon rush hour are 2 peaks. Also, it remains fairly high during daytime but swiftly decreases in the evening.

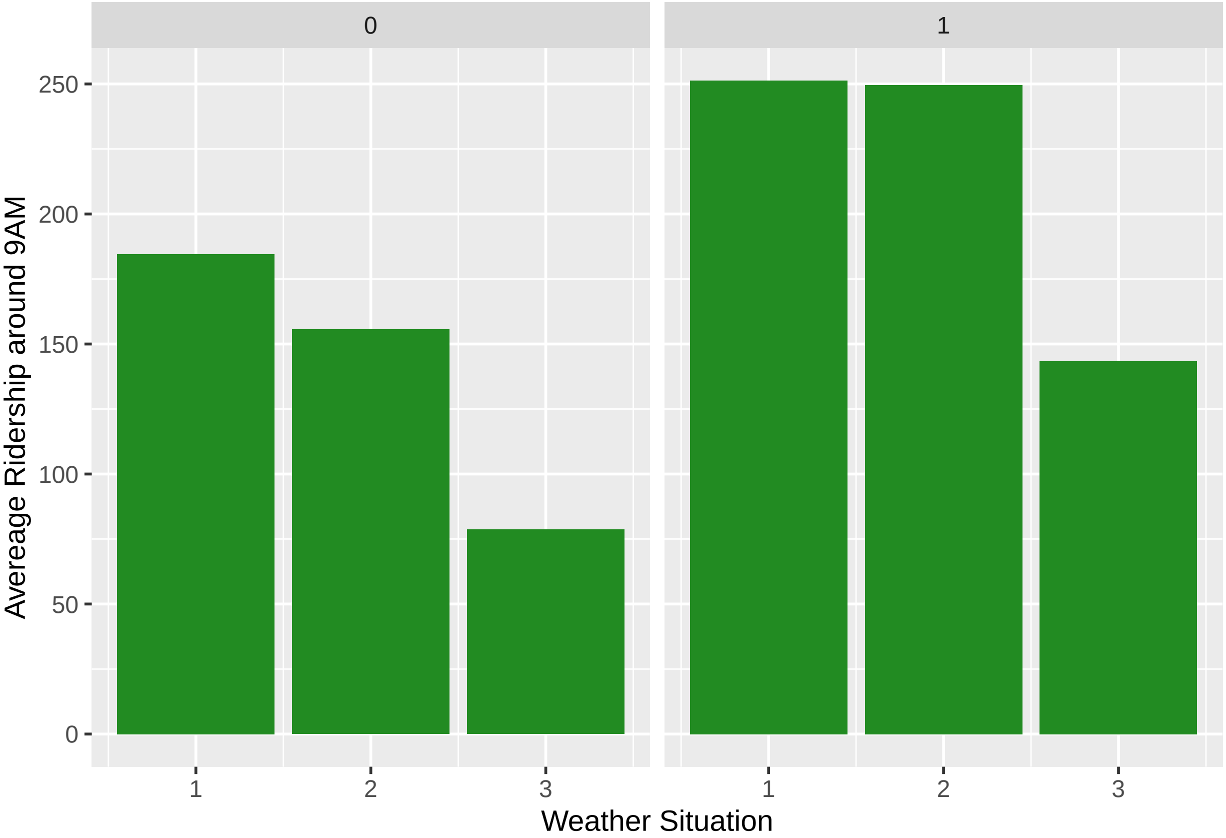
1.5.2 Plot B



X and Y axis basically are the same as the plot A.

0 means weekends and holidays, while 1 means workdays. Workdays' pattern is pretty close to plot A and it complies with common sense. But non-holidays' pattern are quite different from Plot days, people tends to use bikes around noon till afternoon. I guess that people are likely to hang out during this time.

1.5.3 Plot C



0 and 1 has the same meaning as Part B.

X axis means weather situation where 1 means sunny day, 2 means cloudy and misty day, and 3 means light snowy and light rainy day.

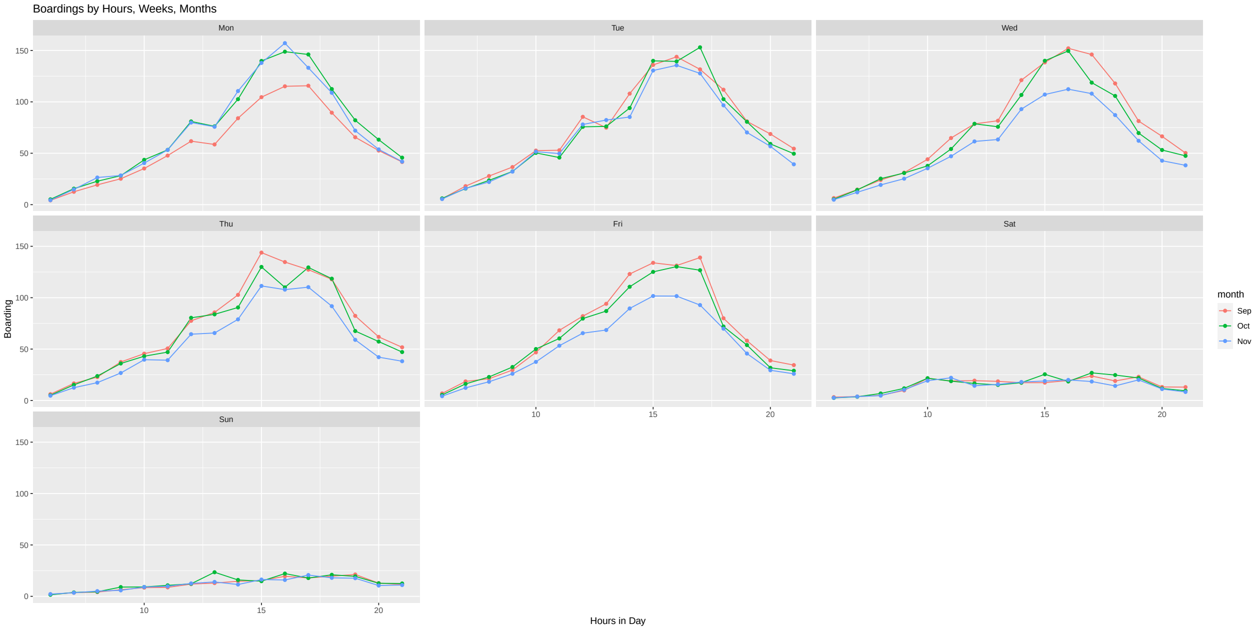
Y axis represents the avarage ridership around 9a.m.

So, in non-workdays, when the weather becomes cloudy, a few people may not go out riding bike because they might think that there will be potential rains. However, in workdays, almost all people will not change their original plan, which is getting a bike, just because of potential rains. They have to work!

2 Homework 2

2.1 Problem 1 Capital Metro UT Ridership

2.1.1 Plot 1



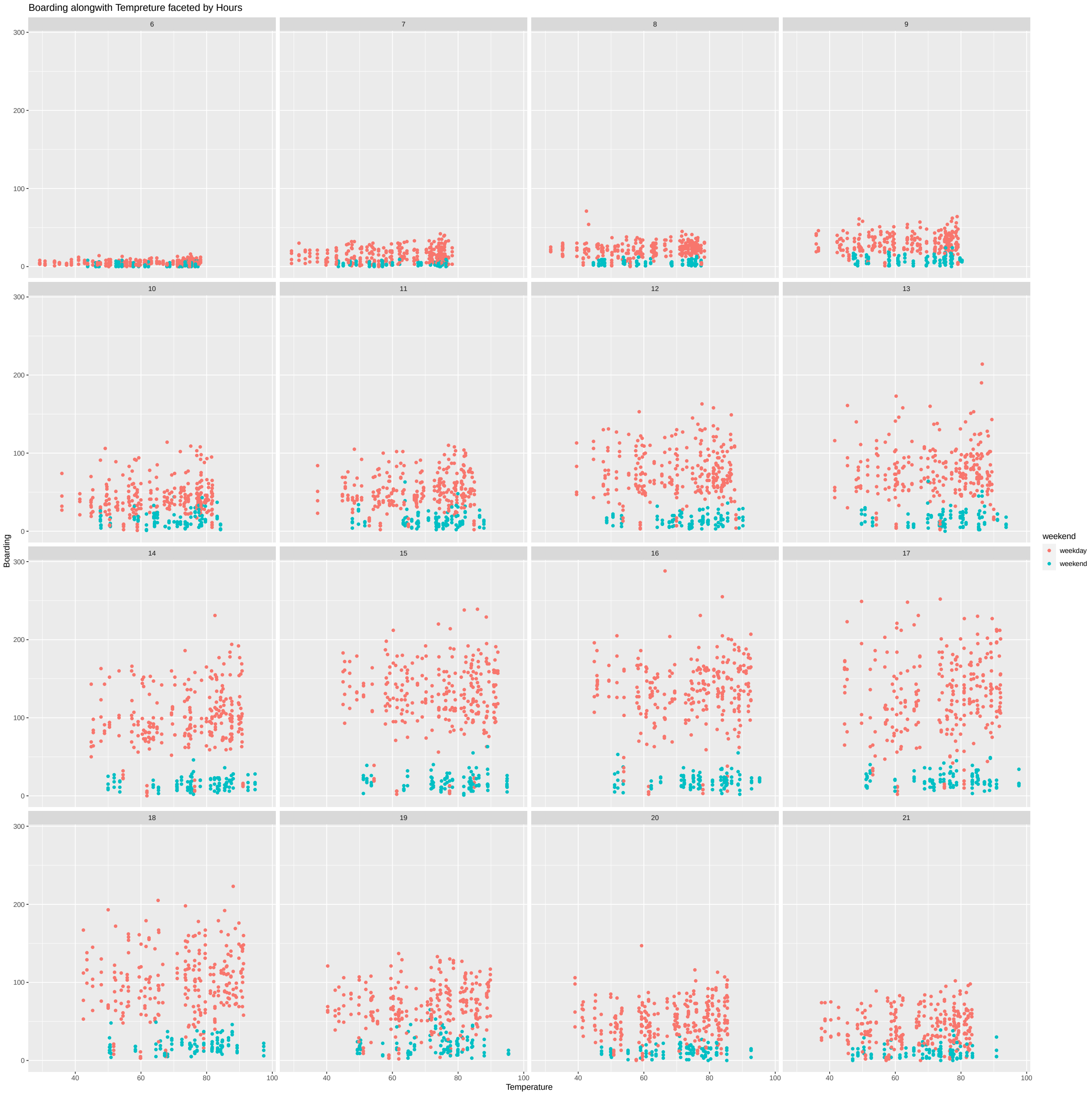
Caption: Above is the the combination of line plots above, each plot stands for a day in week. x-axis means the hours throughout a day from 6 a.m. to 9 p.m. The y-axis means the average boarding. We have 3 lines in each plot, representing 3 months from September to November. Orange line shows September, green line shows October, and blue line shows November.

Answer: From the combination of plots above, I found that the peak boarding from Monday to Friday are quite similar because these five days are workdays; students, teachers, and staffs are taking regular routines in these days. But, on weekend, the boarding curves become so flat because of no class and no work at school.

My reason for that the average boarding on Monday in September is lower than the other 2 months is September has a lot of new students coming. Most new students just know new friends and maybe hanging out with new friends on Sunday night. So, on Monday, they probably could not get up early enough to catch up a bus. That is why the average boarding on Monday is quite less than the other 2 months.

The average boarding on Weds/Thurs/Fri in November looks lower. I guess it is related to the Halloween Holiday. As we know, in Halloween, we have a short break and most people definitely do not come to school. Thus, undoubtedly, the average boarding should be lower.

2.1.2 Plot 2



Caption: This combination of scatter plots shows the relationship between boarding number and temperature. The x-axis is temperature of each section and the y-axis is the boarding amount during each 15min section. Also each faceted plot stands for each hour from 6 a.m. to 9 p.m. e.g. the first plot has a "6" above, which means 6 a.m. and the last plot has a 21 above which means 9 p.m.

Answer: Actually, I did not see any trend alongside with temperature. In each hour of days, the points distribution seems to be very chaotic and ruleless. Therefore, I dare to say temperature has no solid effect on the number of UT student boarding the bus.

2.2 Problem 2 Wrangling the Billboard Top 100

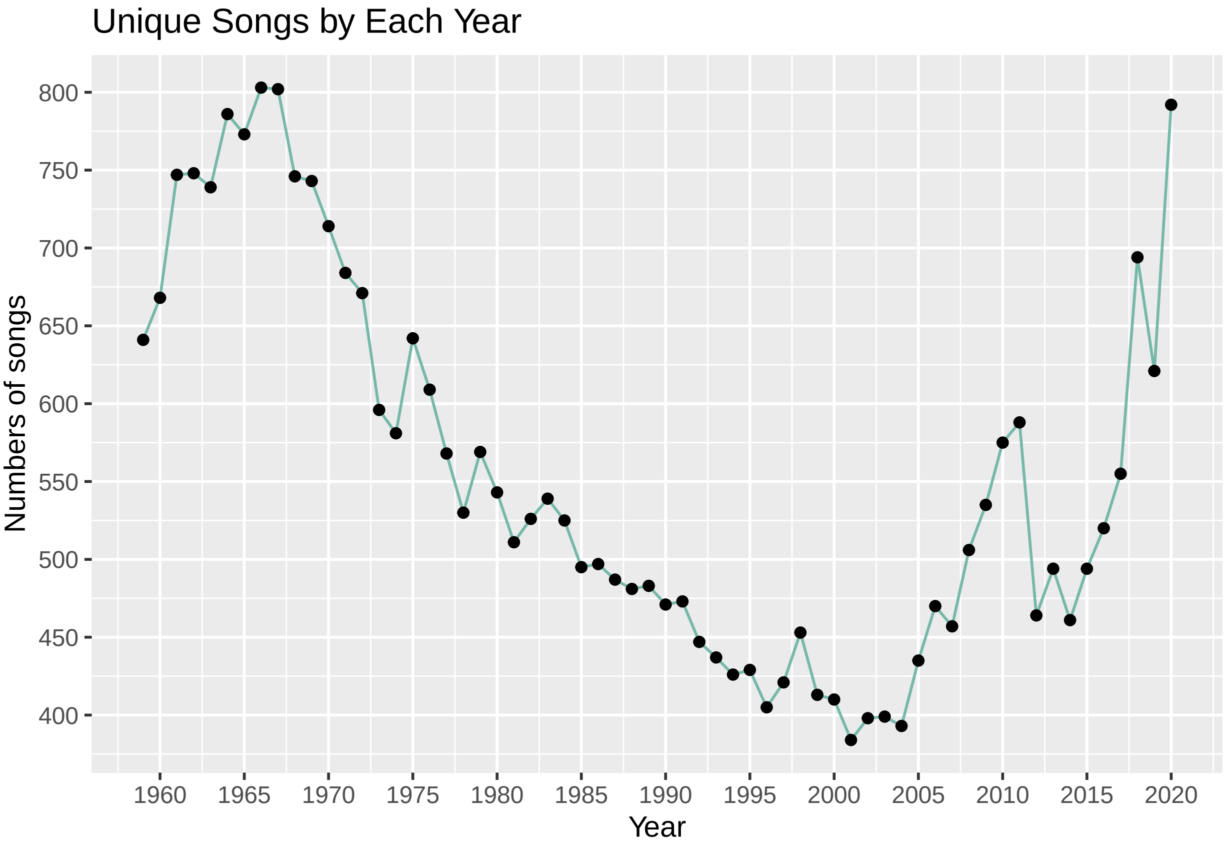
2.2.1 Part A

performer	song	weeks
Imagine Dragons	Radioactive	87
AWOLNATION	Sail	79
Jason Mraz	I'm Yours	76
The Weeknd	Blinding Lights	76
LeAnn Rimes	How Do I Live	69
LMFAO Featuring Lauren Bennett & GoonRock	Party Rock Anthem	68
OneRepublic	Counting Stars	68
Adele	Rolling In The Deep	65
Jewel	Foolish Games/You Were Meant For Me	65
Carrie Underwood	Before He Cheats	64

Caption: This is the table for top 10 songs spending most weeks on the Billboard Top 100 from 1958 to 2021.

The first column shows the performer name of these songs, the second column lists the name of these songs, and the third colume stands for the total number of weeks for each song.

2.2.2 Part B

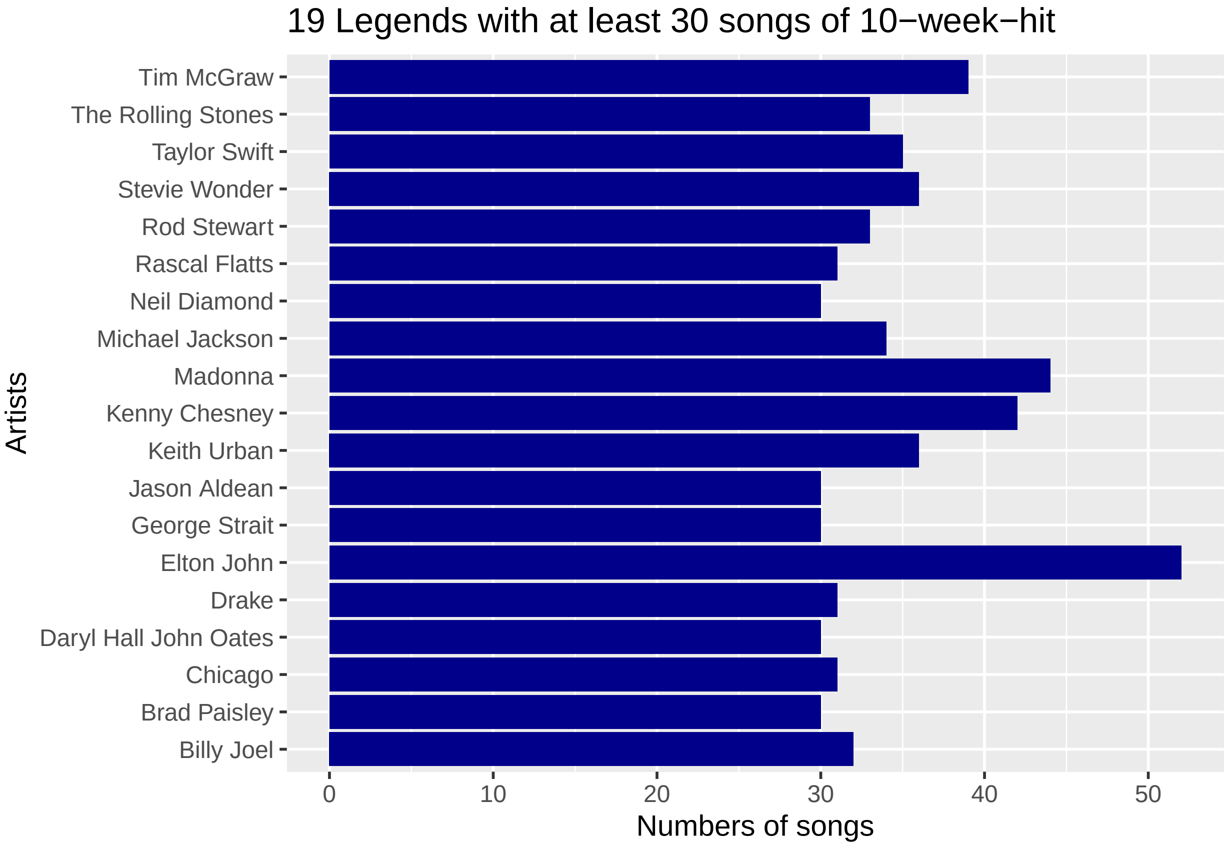


Caption: This is a line graph which shows the total amount of unique songs appearing in the Billboard Top 100 in each year from 1959 to 2020. x-axis represents the year and y-axis stands for the number of the songs.

Comment: I found that around the middle of 1960s, there was a peak which a lot of songs appearing in the Top List. I guess that one of potential reasons is that in these years, the hippie movement arose and gradually dominated the United States. Alongside with this movement, a lot of genres of music, like Rock&Roll, R&B,

was becoming more and more popular. So during this time, songs were pretty diverse. Another rising trend seems to happen after 2010, maybe just because of globalization, songs from other cultures, such as K-pop, Latin, are becoming more and more public.

2.2.3 Part 3



Caption: This is a bar plot showing 19 artists with equal or over 30 songs at least 10 weeks on the Billboard. The x-axis is the the number songs of each artist and the y-axis shows the name of each artist.

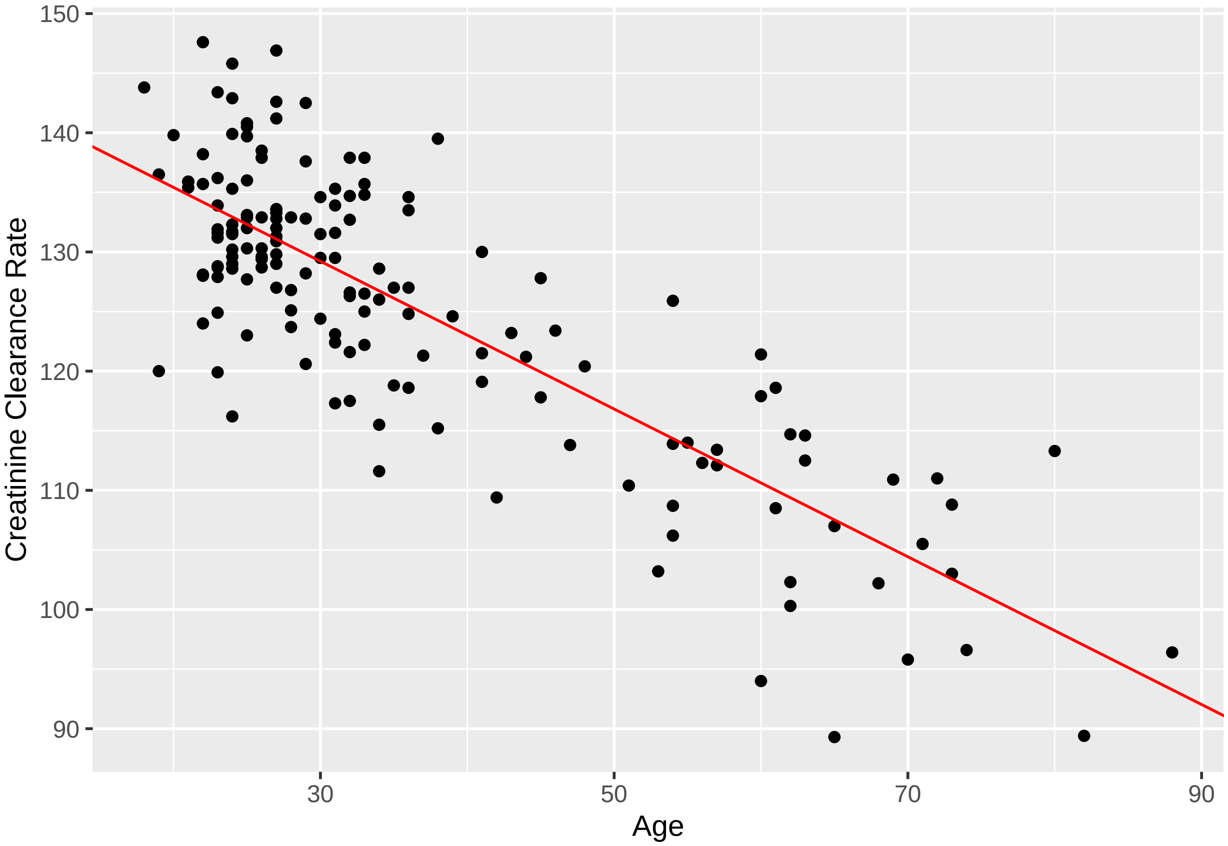
Sir Elton John has over 50 songs appearing more than 10 weeks in Billboard, which is a great achievement!

2.3 Problem 3 Regression Practice

Below is the intercept and estimated coefficient of age:

```
## (Intercept)         age
## 147.8129158  -0.6198159
```

Below shows the scatter plot of age(x-axis) and creatinine clearance rate(y-axis) with a reference line with the estimators above:



2.3.1 Part A

Let me put *age = 55* i.e *x = 55* into the equation *y(creatinine) = 147.81 - 0.62x*. We can get:

```
## [1] 113.71
```

So, we expect that a 55-year-old person has 113.71 ml/min creatinine clearance rate.

2.3.2 Part B

The coefficient of age told us that when we grow up 1 year, we expect to decrease 0.62 ml/min of creatinine clearance rate. i.e. -0.62 ml/min per year.

2.3.3 Part C

Expected 40-year-old creatinine clearance rate:

```
## [1] 123.01
```

Real rate: 135 ml/min

Expected 60-year-old creatinine clearance rate:

```
## [1] 110.61
```

Real rate: 112 ml/min

So, we can know the difference between real creatinine clearance rate for the 40-year-old is *135-123.01 = 11.99* and the difference between real creatinine clearance rate for the 60-year-old is *112 - 110.61 = 1.39*. So from these two difference 11.99 is bigger than 1.39, we can say the 40-year old is healthier.

Let us look another comparison, the comparative rate which defines as *difference/expected value*:

For the 40:

```
## [1] 0.097
```

For the 60:

```
## [1] 0.013
```

So, the rate of the 40 is almost 9 times of the rate of the 60, which means the 40 is comparatively healthier than the expected value for 40.

In conclusion, the 40-year-old is healthier than the 60-year-old in terms of creatinine clearance rate.

2.4 Problem 4 Probability Practice

2.4.1 Part A

Let event A be buying 3 cars and none of them is lemon.

So, ~A should be buying 3 cars and at least one of them is lemon.

P(~A) = 1 - P(A) =

```
## [1] 0.719
```

The probability of getting at least one lemon should be 72%.

2.4.2 Part B

Since each time of throwing a dice is independent, we can treat throwing a dice twice as independent events.

For the probability of getting the sum of 2 number is odd, we have 2 scenarios: First, first time is a even and second time is a odd; the other is converse. So the probability of this event should be *P(odd) * P(even) + P(even) * P(odd)*, which is

```
## [1] 0.5
```

For the probability of the sum of 2 number is less than 7, we can go through case by case. If we get 1 at first time, we can get 1,2,3,4,5 at the second time, the probability of it should be 1/6 * 5/6 = 5/36. If we get 2 at first time, we can get 1,2,3,4 at the second time, the P of it should be 1/6 * 4/6 = 4/36. If getting 3 at the first time, the P should be 1/6 * 3/6 = 3/36. If 4, then 1/6 * 2/6 = 2/36. If 5, then 1/6 * 1/6 = 1/36. If 6, no chance to get less than 7, so the P is 0. Then we add up these probabilities, we can get

```
## [1] 0.417
```

So the chance to get the sum of 2 numbers less than 7 is about 42%

In this cas, we can go through case by case again. If we get 1 at first time, we can get 2,4 at the second time, 2 cases here. If we get 2 at first time, we can get 1,3 at the second time, 2 cases hear. If getting 3 at the first time, we can get 2 at second time, then 1 case. If 4, then get 1 at second time, still 1 case. If 5, then 0. Then we add up these case, we can get 6 case.

The total case is 6 * 6=36, and the chance of getting odd number is 0.5, so the odd case is 0.5 * 36 = 18.

Thus, the P(<7|the number is odd) = 6/18 = 1/3

If they are independent, then P(A|B) = P(A), but here, P(<7) = 0.42 and P(<7|odd) = 0.33. Therefore, they are not independent apparently.

2.4.3 Part C

Information we already have:

P(RC) = 0.3

P(TC) = 0.7

P(yes|RC) = 0.5

P(no|RC) = 0.5

P(yes) = 0.65

P(no) = 0.35

What we need to get: P(yes|TC)

P(yes) = P(RC) * P(yes|RC) + P(TC) * P(yes|TC), which is 0.65 = 0.3 * 0.5 + 0.7 * x

Solve the equation, we get

```
## [1] 0.714
```

Thus, there are 71% Truthful visitors saying yes in the survey.

2.4.4 Part D

Given information:

P(positive|disease) = 0.993

P(negative|non-disease) = 0.999

P(disease) = 0.000025

P(non-disease) = 1 - 0.000025 = 0.999975

What we need to get: P(disease|positive) by Bayes Rule

P(positive|non-disease) = 1 - P(negative|non-disease) = 0.0001

P(positive) = P(positive|disease) * P(disease) + P(positive|non-disease) * P(non-disease) =

```
## [1] 0.0001248225
```

P(disease|positive) = P(positive|disease) * P(disease) / P(positive) =

```
## [1] 0.1988784
```

So, there is 19.9% chance someone have the disease when testing positive.

2.4.5 Part E

Given Information:

P(R/A) = 0.99

P(R/~A) = 0.10

P(A) = 0.05

P(~A) = 0.95

Need to show: P(A/R)

P(R) = P(A) * P(R/A) + P(~A) * P(R/~A) =

```
## [1] 0.1445
```

P(A/R) = P(R/A) * P(A) / P(R) =

```
## [1] 0.3425606
```

So, the probability of that an aircraft is present given it is registered should be around 34.26%

2.5 Problem 5 Modeling soccer games with th Poisson distribution

In order to answer the question what are the estimated probabilities of win/lose/draw results between 2 teams, we need to use the Probability Mass Function of Poisson Distribution to calculate the probabilities of all possible goals for both teams.

The PMF of Poisson distribution takes the following form:

$$P(X = x) = (\lambda^x/x!)e^-\lambda$$

means the expected goals for each team and event X means how many goals scored in a game.

Approach: First, I need to make an assumption that the goals scored by each team is an independent event. So, when a probability of a game score(like 2-1) should be the probability of first team scoring 2 goals times the probability of second team scoring 1 goal.

Find the (expected goal by a team) by the formula:

The team's attach power * The rival's defense weakness * Home/away boost, in which:

Attach power = The total goals for a team in a season/the average goals of all teams in a season

Defense weakness = The total goals lost by a team in a season/the average goals lost of all teams in a season

Home boost = the average goals for all home teams in a match

Away boost = the average goals for all away teams in a match

After we get , we need to use PMF of Poisson to calculate a probabilty of an outcome. For example:

Set team A's is 1.5 and team B's is 1.7, then the goal for A is Xa and the goal for B is Xb.

Since Xa and Xb are independent, we have

$$P(Xa = 2, Xb = 1) = (1.5^2/2!)e^-1.5 * (1.7^1/1!)e^-1.7$$

With just a little bit more work, we can use calculate the probability of all possible outcomes up to a rip-roaring (but very unlikely) 7-7 draw by 'tibble', 'expand', 'mutate' functions in R.

Then, we have a table for all probabilities of games from 0-0 to 7-7 (64 outcomes). Then by 'filter' function, we can calculate the probability of each scenario(win/lose/draw).

Next, Let me calculate the home/away boost, and seasonal average goals for further use

The average number of goals scored i.e. conceded by teams is (all the goals/20 teams):

```
## [1] 53.6
```

The home boost is

```
## [1] 1.57
```

The away boost is

```
## [1] 1.25
```

2.5.1 Question 1

Liverpool(home) vs Tottenhan(away)

The Attack Strength for Liverpool(home) sould be:

```
## [1] 1.66
```

The Defense Weakness for Liverpool(home) should be:

```
## [1] 0.41
```

The Attack Strength for Tottenhan(away) should be:

```
## [1] 1.25
```

The Defense Weakness for Tottenhan(away) should be:

```
## [1] 0.73
```

The for Liverpool(home) should be:

```
## [1] 1.89
```

The for Tottenhan(away) should be:

```
## [1] 0.64
```

Now, we have the s for both Liverpool and Tottenhan.

Now, show head 5 rows in a probabilities table for all possible game scores up to 6-6:

Liverpool

Tottenhan

prob

1

0

0

0.079

2

0

1

0.051

3

0

2

0.016

4

0

3

0.003

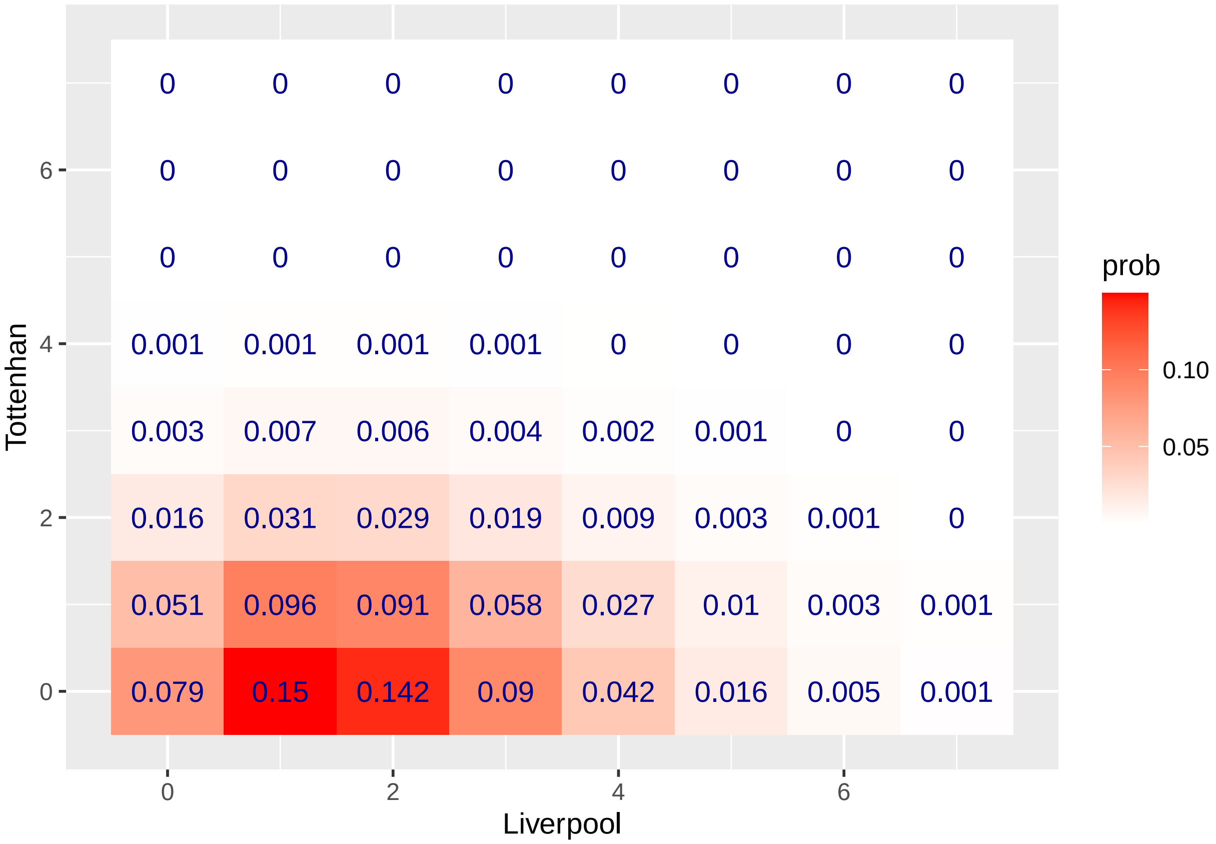
5

0

4

0.001

Then visualize it by heatmap:



Then calculate the probabilities of 3 scenarios:

Liverpool win:

```
## # A tibble: 1 x 1
##   `sum(prob)`
##         <dbl>
## 1       0.672
```

Tottenhan win:

```
## # A tibble: 1 x 1
##   `sum(prob)`
##         <dbl>
## 1       0.118
```

Draw:

```
## # A tibble: 1 x 1
##   `sum(prob)`
##         <dbl>
## 1       0.209
```

In conclusion, the chance of Liverpool winning is around 67.2%, the chance of Tottenhan winning is 11.8%, and the chance of draw is 20.9%

2.5.2 Question 2

Manchester City(home) vs Arsenal(away)

The Attack Strength for Manchester City(home) sould be:

```
## [1] 1.77
```

The Defense Weakness for Manchester City(home) should be:

```
## [1] 0.43
```

The Attack Strength for Arsenal(away) should be:

```
## [1] 1.36
```

The Defense Weakness for Arsenal(away) should be:

```
## [1] 0.95
```

The   for Manchester City(home) should be:

```
## [1] 2.65
```

The   for Arsenal(away) should be:

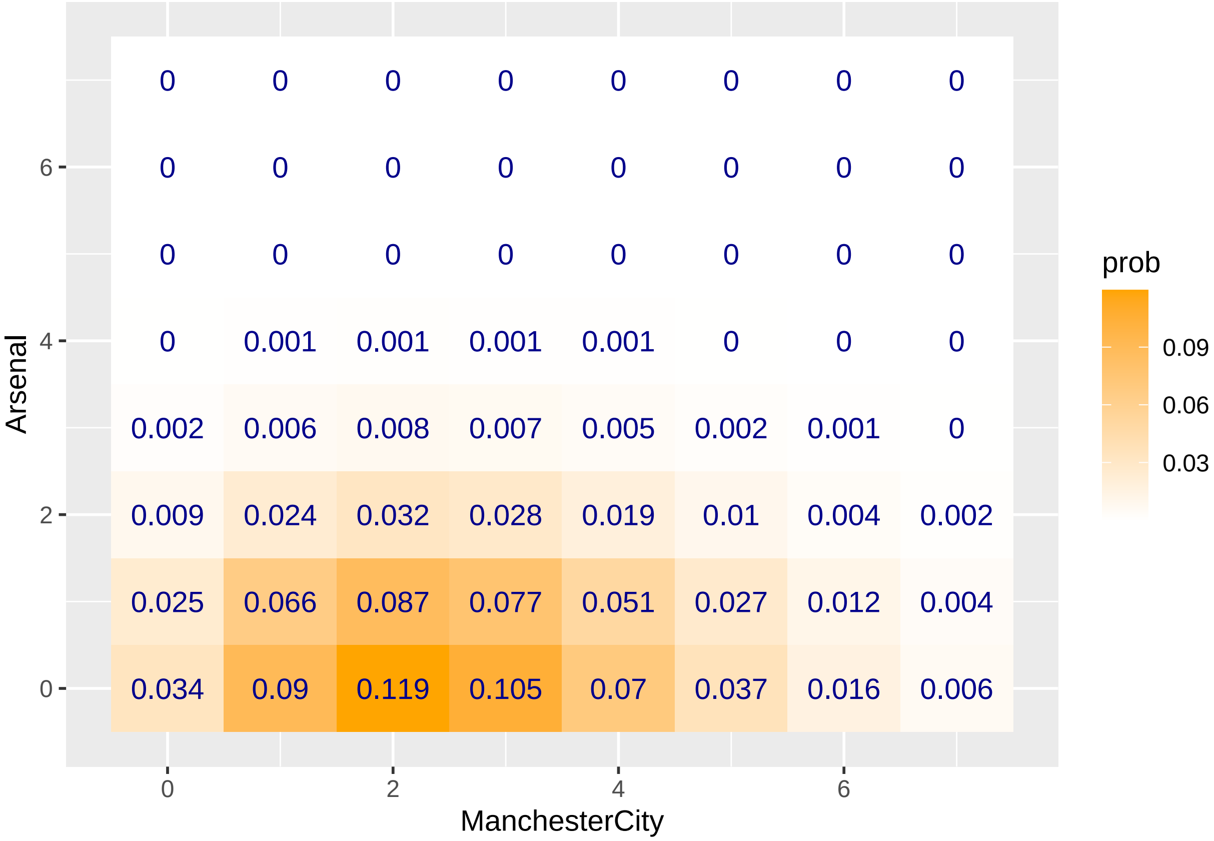
```
## [1] 0.73
```

Now, we have the s for both Liverpool and Tottenhan.

Now, show head 5 rows in a probabilities table for all possible game scores up to 7-7:

	ManchesterCity	Arsenal	prob
1	0	0	0.034
2	0	1	0.025
3	0	2	0.009
4	0	3	0.002
5	0	4	0.000

Then visualize it by heatmap:



Then calculate the probabilities of 3 scenarios:

ManchesterCity win:

```
## # A tibble: 1 x 1
##   `sum(prob)`
##         <dbl>
## 1       0.775
```

Arsenal win:

```
## # A tibble: 1 x 1
##   `sum(prob)`
##         <dbl>
## 1      0.0793
```

Draw:

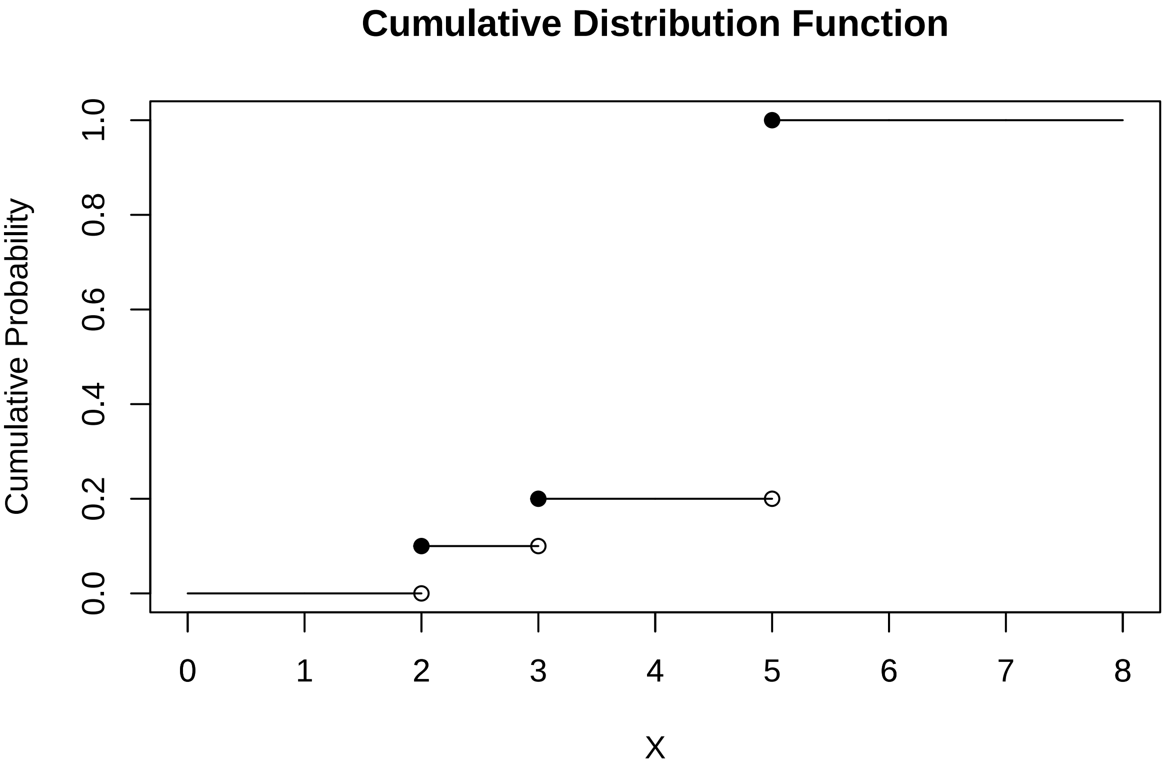
```
## # A tibble: 1 x 1
##   `sum(prob)`
##         <dbl>
## 1       0.140
```

In conclusion, the chance of ManchesterCity winning is around 77.5%, the chance of Arsenal winning is 7.9%, and the chance of draw is 14%

3 Homework 3

3.1 Problem 1 CDFs and PDFs

3.1.1 Part A



The CDF's plot shows above. X is the event, and Y-axis shows the cumulative probability.

We can conclude that:

$$P(2 < X \leq 4.5) = P(X = 3) = 0.1$$

$$P(2 \leq X < 4.5) = P(X = 2, 3) = 0.1 + 0.1 = 0.2.$$

3.1.2 Part B

3.1.2.1 (i)

$$P(X^2 \leq 0.25) = P(X \leq \sqrt{0.25}) = P(X \leq 0.5)$$

Since it follows a uniform distribution, then by the formula

$$P(X \leq 0.5) = \frac{0.5 - 0}{1 - 0} = 0.5$$

3.1.2.2 (ii) We have to make clear that 0 < a < 1, otherwise it will fail to get a probability.

$$P(X^2 \leq a) = P(X \leq \sqrt{a}) = \frac{\sqrt{a} - 0}{1 - 0} = \sqrt{a}$$

3.1.2.3 (iii) From part (ii), we know that

$$P(Y \leq y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = \frac{\sqrt{y} - 0}{1 - 0} = \sqrt{y}$$

So the CDF of Y is:

$$F(y) = \sqrt{y}$$

And then take first-order derivatives to PDF:

$$f_y(y) = \frac{d(\sqrt{y})}{dy} = \frac{1}{2(\sqrt{y})}$$

3.1.2.4 (iiii) Compute E(Y):

$$E(Y) = \int_0^1 yf(y)dy = \int_0^1 \frac{y}{2(\sqrt{y})}dy = \int_0^1 \frac{\sqrt{y}}{2}dy = \frac{1}{3} * (1)^{\frac{3}{2}} - 0 = \frac{1}{3}$$

In order to compute var(Y), we need to calculate E(Y^2) first.
Compute E(Y^2):

$$E(Y^2) = \int_0^1 y^2 f(y)dy = \int_0^1 \frac{y^2}{2(\sqrt{y})}dy = \int_0^1 \frac{\sqrt{y^3}}{2}dy = \frac{1}{5} * (1)^{\frac{5}{2}} - 0 = \frac{1}{5}$$

Thus, the var(Y) = E(Y^2) - E(Y)^2 as follow:

$$var(Y) = E(Y^2) - E(Y)^2 = \frac{1}{5} - (\frac{1}{3})^2 = \frac{4}{45}$$

3.2 Problem 2 Practice with Expected Value

3.2.1 Part A

We are trying to get E(X):

$$E(X) = E(Z_1^2 + Z_2^2 ... + Z_d^2) = E(Z_1^2) + E(Z_2^2) + ... + E(Z_d^2)$$

Since Z_1 to Z_d are all follows the same distribution, we have:

$$E(Z_1^2) = E(Z_2^2) = ... = E(Z_d^2)$$

Thus, as long as we get E(Z_i^2), we are done. So the E(Z_i^2):

$$E(Z_i^2) = Var(Z_i) + E(Z_i)^2 = 1 + 0^2 = 1$$

The we can get E(X):

$$E(X) = E(Z_1^2) + E(Z_2^2) + ... + E(Z_d^2) = \sum_{i=1}^{d} 1 = d$$

3.2.2 Part B

I do not agree with that, because average velocity does not take into account the time spent at different velocities during the entire journey.

Therefore, if we want to figure out the average time, we should calculate the time for walking and biking separately, and then take weighted average of them.

Thus, the time for walking should be:

$$T(walking) = Distance/Velocity(walking) = 2(miles)/5(miles/hour) = 0.4hour$$

And the time for biking should be:

$$T(biking) = Distance/Velocity(biking) = 2(miles)/10(miles/hour) = 0.2hour$$

And the add in the probabilities of two situations, we have:

$$T = w(walking)*T(walking)+w(biking)*T(biking) = 0.4*0.4+0.6*0.2 = 0.16+0.12 = 0.28hour = 16.8minutes$$

Thus, Markov takes average 0.28 hour, or 16.8 minutes, to get to the class.

3.3 Problem 3 Inverse CDF

We can calculate the P(X<x) first:

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(F(F^{-1}(U)) \leq F(x)) = P(U \leq F(x)) = F(x)$$

Then once we get the CDF, we can get the PDF of it.

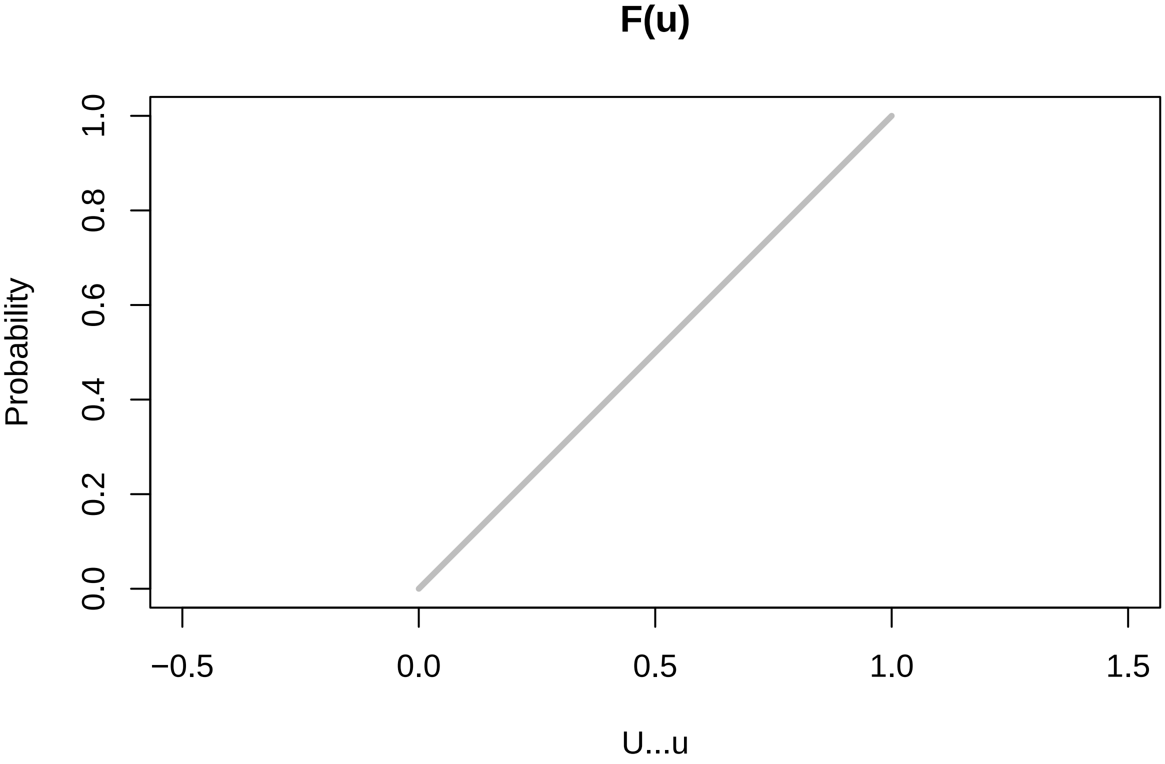
$$f(x) = F'(x)$$

Let's look at the detailed case here (0 u 1):

$$F(u) = P(U \leq u) = \frac{u-0}{1-0} = u$$

The plot as followed:

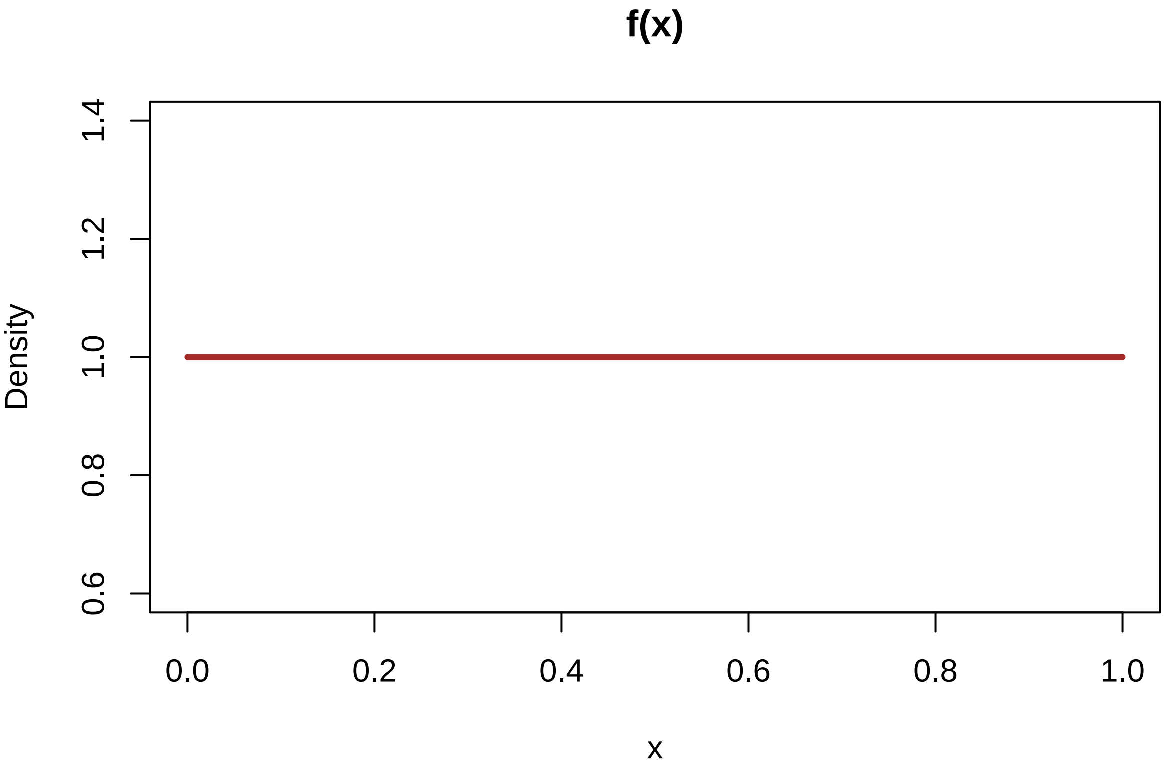
```
## Warning in title(...): conversion failure on 'U u' in 'mbcsToSbcs': dot
## substituted for <e2>

## Warning in title(...): conversion failure on 'U u' in 'mbcsToSbcs': dot
## substituted for <89>

## Warning in title(...): conversion failure on 'U u' in 'mbcsToSbcs': dot
## substituted for <a4>
```



Since X = F^-1(U), we can see from the picture X and U are identical and all a uniform distribution on [0,1]. Thus, we can conclude that the pdf of f(x) equals to of f(u) as followed (0 x 1):

$$f(x) = \frac{1}{1-0} = 1$$

The plot shows as followed:



3.4 Problem 4 Simulation

3.4.1 Part A

The expected value for pn shows below:

$$E(\hat{p_N}) = E(\frac{X_N}{N}) = \frac{E(X_N)}{N} = \frac{NP}{N} = P$$

The standard deviation for pn shows below:

$$sd(\hat{p_N}) = \sqrt{(var(\hat{p_N}))} = \sqrt{var(\frac{X_N}{N})}_= \sqrt{\frac{var(X_N)}{N^2}} = \sqrt{\frac{NP(1-P)}{N^2}} = \sqrt{\frac{P-P^2}{N}}$$

3.4.2 Part B

If we put N=5, P=0.5 in formula E(P_N) and sd(P_N), we can get

Probability for N=5:

```
## [1] 0.5
```

Standard deviation for N=5:

```
## [1] 0.2236
```

Then we use randomly simulate 1000 times, we can get Monte Carlo Mean:

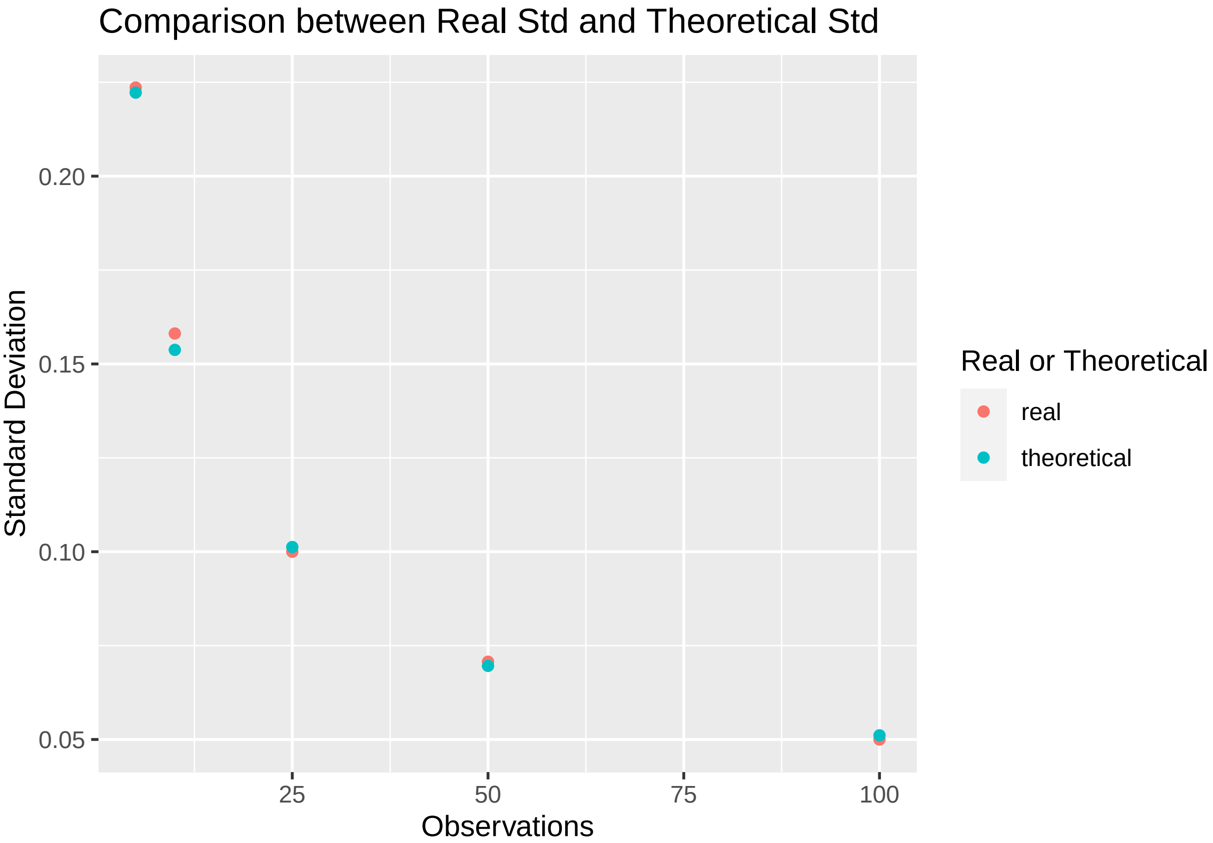
```
## [1] 0.5038
```

Monte Carlo Standard Deviation:

```
## [1] 0.2222
```

From the results above, we can find Monte Carlo Mean is very close to theoretical mean, and of cause, Monte Carlo standard deviation basically agrees with the theoretical standard deviation.

Repeat the same process for p10, p25, p50, p100, and make the plot as required below:



Caption: X-axis stands for the observation times in an experiment, noted as N; Y-axis represents the value of standard deviation. Besides, orange points mean the real value from experiments, and blue points are the theoretical value calculated be the formula in Part (A).

Comments: From the plot above, we can see that, as the observations(N) of experiment increase more and more, the standard deviation become less and less, which matches the standard deviation formula in part (A).

In addition to that, I can find that, as the observations(N) of each experiment increase, the real standard deviations are more close to the theoretical standard deviations.It complies large number theorem.

3.5 Problem 5 More PDF/CDF Practice

The CDF of X_i(i=1,2,…,N) as given:

$$F_{X_i}(x) = P(X_i \leq x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}$$

Since max{X_i(i=1,2,…,N)}) < y is equivalent to X_i < y for any i, then we have:

$$F_{Y_N}(y) = P(Y_N < y) = P(max(X_1, X_2 ... X_N) < y) = P(X_1 < y, X_2 < y, X_3 < y ... X_N < y)$$

And then because X itself is independent, then we have:

$P(X_1 < y, X_2 < y, X_3 < y ... X_N < y) = P(X_1 < y) * P(X_2 < y) * ... * P(X_N < y) = F_{X_1}(y) * F_{X_2}(y) * ... * F_{X_N}(y) = (1-e^{-\lambda y})^N$

Then, differentiate the CDF of Y_N, we can get the PDF of Y_N:

$$f_{Y_N} = \frac{dF_{Y_N}(y)}{dy} = \frac{d}{dy}(1 - e^{\lambda y})^N = N(1 - e^{\lambda y})^{N-1} * (-e^{-\lambda y}) * -\lambda = N\lambda e^{-\lambda y}(1 - e^{-\lambda y})^{N-1}$$

4 Homework 4

4.1 Problem 1 NBC Pilot Survey

4.1.1 Part A

Question: Does either “Living with Ed” or “My Name is Earl” make audiences happier than the other one?

Approach: I will use t-test to check if the mean of “happy” scores for “Living with Ed” and “My Name is Earl” have any difference. Thus, in this case, the null hypothesis is the mean difference to be 0, which indicates that they do not have any difference in terms of their means of happy “score”.

Results:

```
##
##  Welch Two Sample t-test
##
## data:  ed$Q1_Happy and earl$Q1_Happy
## t = 1.1676, df = 162.57, p-value = 0.2447
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1030341  0.4011371
## sample estimates:
## mean of x mean of y
##  3.926829  3.777778
```

Conclusion: From the t test summary above, we can see 0 is in the 95% confidence interval. Thus, we fail to reject the null hypothesis here. Therefore, I conclude that there is not a significant evidence showing either “Living with Ed” or “My Name is Earl” is more funny than the other one.

4.1.2 Part B

Question: Is either “The Biggest Loser” or “The Apprentice: Los Angeles” more annoying than the other one?

Approach: Likewise, I am going to use t-test to check if they have any difference in terms of the mean of “annoyed” scores. Therefore, the null hypothesis is the mean difference to be 0, which suggests that they have no difference on the mean of “annoyed” scores.

Results:

```
##
##  Welch Two Sample t-test
##
## data:  loser$Q1_Annoyed and la$Q1_Annoyed
## t = -2.1032, df = 300.66, p-value = 0.03628
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.52455614 -0.01743792
## sample estimates:
## mean of x mean of y
##  2.036232  2.307229
```

Conclusion: From the t test summary above, we can see the 95% confident interval doesn't contain 0, which means our null hypothesis has to be rejected, so they really have difference in annoying level. To be more specific, The "Apprentice: Los Angeles" is more annoying than "The Biggest Loser", because it has higher average "annoyed" score.

4.1.3 Part C

Question: As we expect, what proportion of American TV audiences would think "Dancing with the Stars" is confusing?

Approach: I am going to use proportion test with null hypothesis as P(not confusing) = P(confusing).

Results:

```
##
##  1-sample proportions test with continuity correction
##
## data:  $  [with success = TRUE]stars  [with success = TRUE]confusing  [with success = TRUE]
## X-squared = 127.65, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.04453431 0.12893254
## sample estimates:
##          p
## 0.07734807
```

Conclusion: From the proportions test summary above, we can see the null hypothesis fails to hold. Also, the 95% confident interval of actual proportion is from around 4.5% to 12.9%. In other words, The proportion of American audiences we expect to be befuddled by "Dancing with the Stars" falls into range between 4.5% to 12.9% with 95% confidence level.

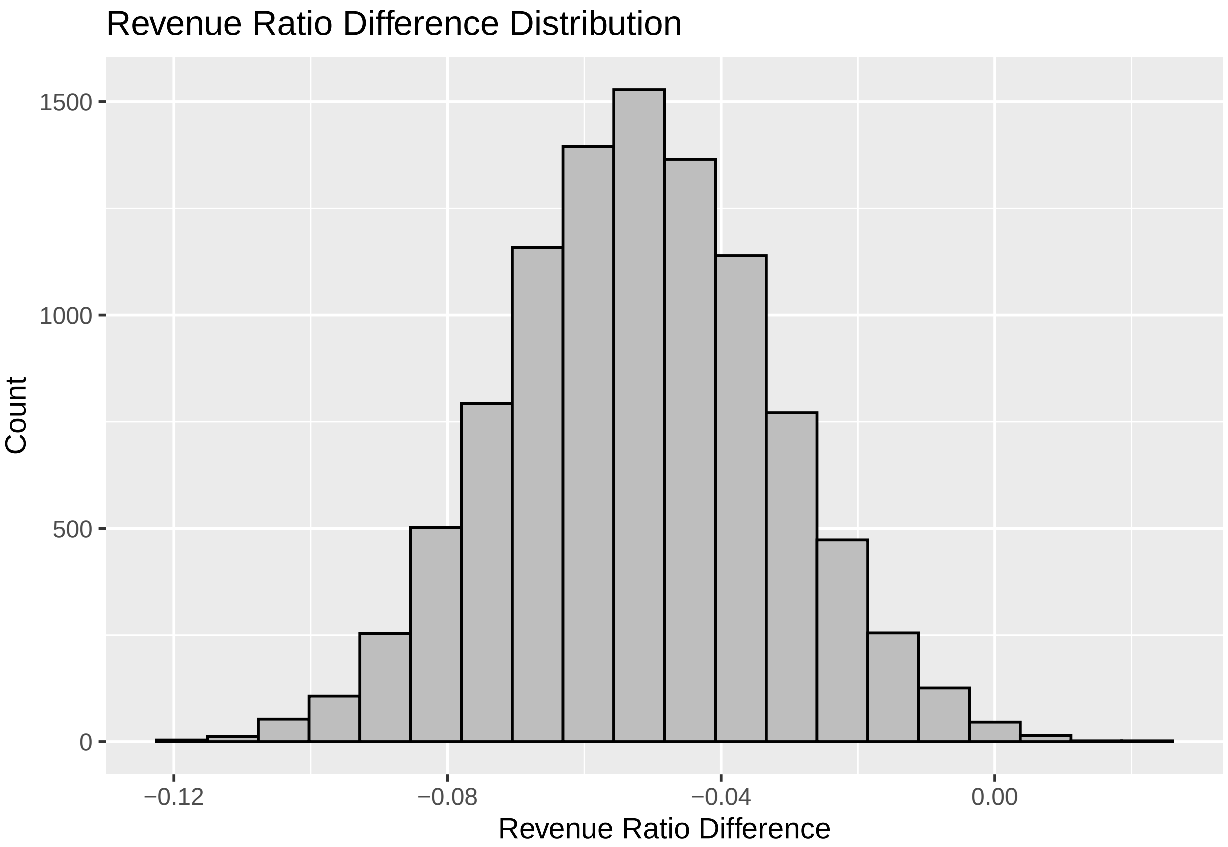
4.2 Problem 2 Ebay

Question: Does the difference of revenue ratios between treatment group and control group indicate the paid search advertising on Google generates more revenue for EBay?

Approach: I am going to bootstrap 10000 times of mean of revenue ratios for both treatment group and control group, and then compute their differences. Once I have 10000 times of differences, I can create a 95% confidence interval of all differences to see whether 0 falls in to it or not. If 0 is in the interval, I can conclude that no significant evidence shows that paid search advertising boosts the revenue for EBay. If the interval falls into positive side, than I can say that paid search advertising diminish the revenue. If the interval falls into negative side, than I can say that paid search advertising boosts the revenue.

Results:

Let us see the histogram of the 10000 times bootstrap of the revenue ratio difference between treatmentgroup and control group:



```
##     name       lower       upper level     method    estimate
## 1 result -0.08978001 -0.01338304  0.95 percentile -0.05228145
```

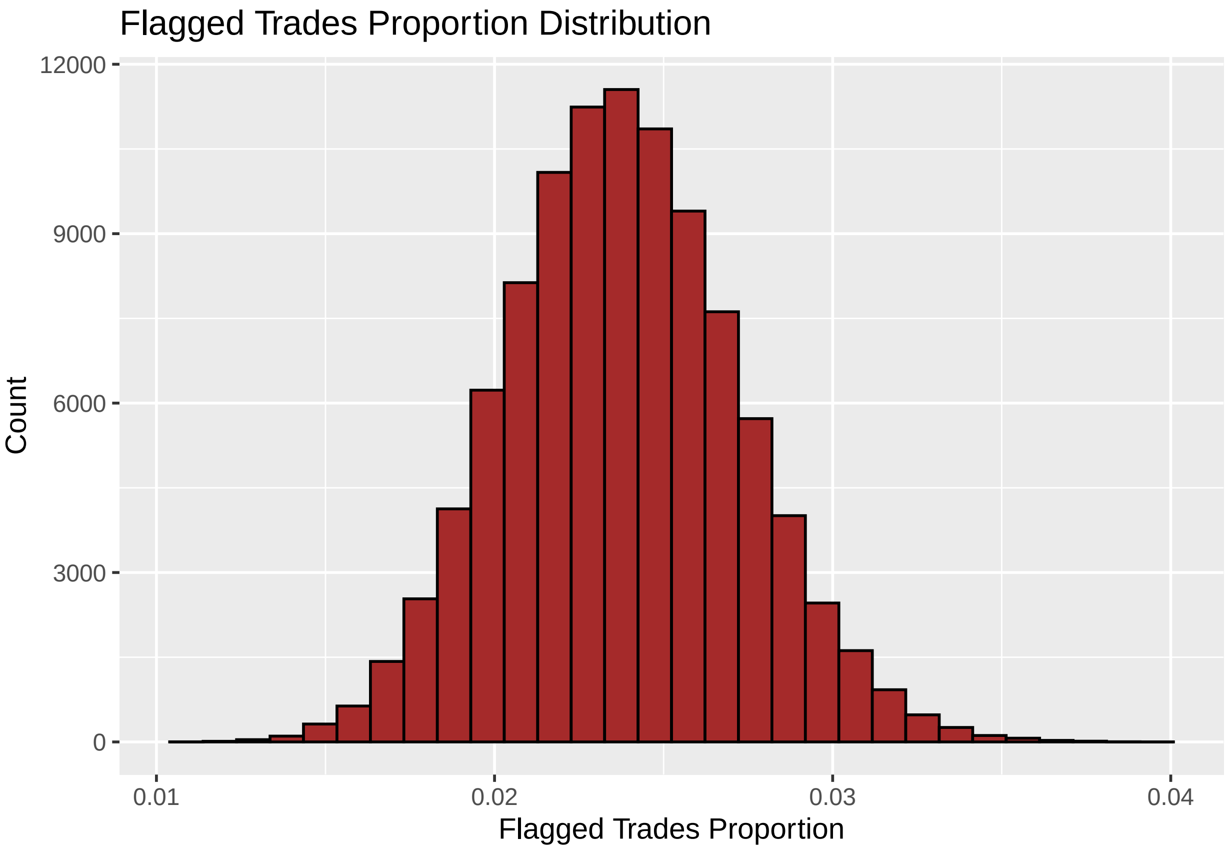
Conclusion: From the summary above, we can see the 95% confidence interval falls below 0, which means with 95% confident level, the real difference between the revenue ratio mean of treatment group and control group is in the range of around -9% and -1%. Thus, I conclude that paid search advertising does boost the revenue.

4.3 Problem 3 Iron Bank

In the case, I set the null hypothesis as that the proportion of flagged trades from Iron Bank is the same as 2.4%. In other words, they were clean under this algorithm.

My test statistic will be the proportion of flagged trades out of 2021 in each simulation.

Let us get into the simulation part. First, the histogram below shows distribution of the proportion in 100000 simulations:



Next, let us look at the P-value:

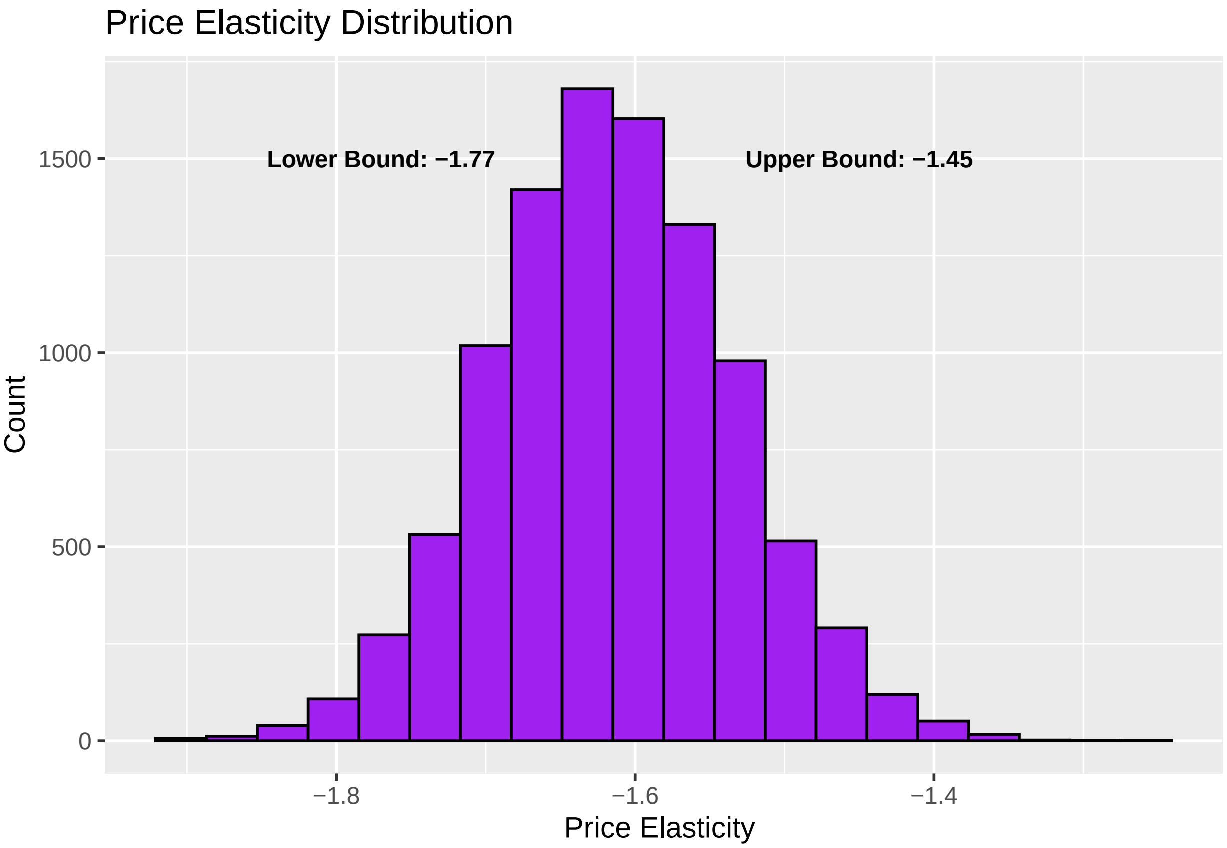
```
## [1] 0.000224
```

As we see, the P value is around 0.02%, which means given that the null hypothesis is true, the chance of observing the number of flagged trades to be equal to or larger than 70 should be around 0.02%. It is very close to 0, so the null hypothesis fails to hold so that we claim that Iron Bank violated the "Inside Trading" laws.

One_sentence conclusion: If the P value is larger than 2.5%, then the null hypothesis will look plausible to me.

Defensive statements: My presumptive confidence level is 95% and since this is one-side test, we should ignore the probability on the lower extreme side. Therefore, 95% confidence interval + 2.5%(the chance of the other extreme side) = 97.5%. Thus, if the P value here is larger than 2.5%, then it would be in 95% confidence interval. Hence, the null hypothesis fails to be rejected. Therefore, I would claim that there is not a significant evidence showing that the proportion of flagged trades from Iron Bank is not 2.4%.

4.4 Problem 4 Milk, Demend, Revisited



Caption: This figure shows 10000 times bootstrap of linear regression of demand on the price. X-axis represents the price elasticity of demand and the y-axis stands for count for the values of the price elasticity of demand. In this graph, 'lower Bound' means the lower bound of 95% confidence interval and 'Upper Bound' means the upper bound of 95% confidence interval. Thus, the whole 95% confidence interval is from 'lower Bound' to 'Upper Bound'.

4.5 Problem 5 Standard-error Calculations

4.5.1 Part A

4.5.1.1 i First, we have

$$E(\hat{p}-\hat{q}) = E(\hat{p}) - E(\hat{q})$$

Then, we have

$$E(\hat{p}) = E(\overline{X_N}) = E(\frac{\sum_{i=1}^{N} X_i}{N}) = \frac{Np}{N} = p$$

Also, we have

$$E(\hat{q}) = E(\overline{Y_M}) = E(\frac{\sum_{i=1}^{M} Y_i}{M}) = \frac{Mq}{M} = q$$

Thus, we have

$$E(\hat{p}-\hat{q}) = E(\hat{p}) - E(\hat{q}) = p - q$$

4.5.1.2 ii As we know, the variance of a random variable X following Bernoulli distribution should be:

$$Var(X) = p(1-p)$$

$$Var(\hat{p}) = Var(\overline{X_N}) = Var(\frac{X_1 + X_2 + ... + X_N}{N}) = \frac{1}{N^2}\sum_{i=1}^{N} Var(X_i) = \frac{1}{N^2} * Np(1-p) = \frac{p(1-p)}{N}$$

Thus, the standard deviation should be

$$Std(\hat{p}) = \sqrt{\frac{p(1-p)}{N}}$$

4.5.1.3 iii Following the same logic, we have

$$Var(\hat{q}) = \frac{q(1-q)}{M}$$

Thus, we have

$$Var(\hat{\triangle}) = Var((\hat{p}-\hat{q}) = Var(\hat{p}) - Var(\hat{q}) = \frac{p(1-p)}{N} - \frac{q(1-q)}{M}$$

Therefore, we have

$$Std(\hat{\triangle}) = \sqrt{\frac{p(1-p)}{N} - \frac{q(1-q)}{M}}$$

4.5.2 Part B

The expected value of this estimator should be

$$E(\hat{\triangle}) = E(\overline{X_N} - \overline{Y_M}) = E(\overline{X_N}) - E(\overline{Y_M}) = \mu_X - \mu_Y$$

The standard error of this estimator should be

$$Std(\hat{\triangle}) = \sqrt{Var(\hat{\triangle})} = \sqrt{Var(\overline{X_N} - \overline{Y_M})} = \sqrt{Var(\overline{X_N}) - Var(\overline{Y_M})} = \sqrt{\frac{1}{N^2}\sum_{i=1}^{N} Var(X_i) + \frac{1}{M^2}\sum_{i=1}^{M} Var(Y_i)} = \sqrt{\frac{N}{N^2}{\sigma_X}^2 +$$