MA 677 Final Project

Shuoqi Huang

5/7/2020

Introduction

The goal of this project is to test the idea that a state that has more airports or enplanements has a significant difference in confirmed cases than a state that has a relatively low capacity. The reason why I choose to do this project is that large international airports perhaps are potentially helpful for spreading the Coronavirus. For example, Hartsfield - Jackson Atlanta International (ATL), Los Angeles International (LAX), and Chicago O'Hare International (ORD) are the three largest airports by checking the enplanements. Therefore, I decided to check this idea by doing Bayesian hypothesis testing.

Data

I get the state confirmed cases data from Havard dataverse (https:// dataverse.harvard.edu/dataverse/2019ncov). Moreover, I explore data of Public-use airports and the number of enplanements from the Federal Aviation Administration.

State	Number of Airports	State	Number of Airports	State	Number of Airports
AK	406	KY	60	NY	140
AL	92	LA	74	OH	170
AR	99	MA	40	OK	139
AZ	79	MD	47	OR	97
CA	254	ME	71	PA	130
CO	76	MI	231	RI	8
CT	23	MN	153	SC	68
DC	4	MO	129	SD	73
DE	11	MS	80	TN	80
FL	128	MT	128	TX	394
GA	109	NC	114	UT	46
HI	14	ND	89	VA	65
IA	121	NE	85	VT	16
ID	123	NH	25	WA	138
IL	110	NJ	45	WI	133
IN	115	NM	63	WV	36
KS	138	NV	49	WY	41

We can see that Alaska has the largest number of airports, but the latest confirmed cases are 369 on May 5th. Alabama has 8437 confirmed cases with only 92 airports. Therefore, I decided to divide states into two groups by checking the enplanements.

The FAA only provides enplanements by each airport in some states. Therefore, I calculate the total enplanements by states and divide them into two groups by setting the checkpoint enplanes equal to 10,000,000.

	ST	Capacity	Class		ST	Capacity	Class
1	CA	117968341	High	25	LA	7795240	Low
2	FL	91150567	High	26	KY	6844603	Low
3	TX	86428773	High	27	IN	5623344	Low
4	NY	54592785	High	28	WI	5475554	Low
5	GA	53712314	High	29	SC	5227513	Low
6	IL	51892058	High	30	AK	4985422	Low
7	CO	33377605	High	31	OK	3653023	Low
8	NC	31217947	High	32	CT	3317026	Low
9	VA	27857139	High	33	NM	2858986	Low
10	WA	26956708	High	34	NE	2709190	Low
11	NV	26067981	High	35	AL	2568618	Low
12	AZ	24762641	High	36	ID	2348824	Low
13	NJ	23770909	High	37	MT	2193321	Low
14	PA	21681560	High	38	RI	2151912	Low
15	MI	20345855	High	39	IA	2082586	Low
16	MA	20314174	High	40	AR	1934533	Low
17	MN	18795176	High	41	ME	1441318	Low
18	HI	17792964	High	42	ND	1095631	Low
19	MO	14128339	High	43	NH	1014421	Low
20	MD	13459863	High	44	KS	953132	Low
21	UT	12488751	High	45	MS	943091	Low
22	TN	12010620	High	46	SD	903958	Low
23	OR	11313244	High	47	VT	664433	Low
24	OH	10266738	High	48	WY	603422	Low
				49	WV	374922	Low

Method

Unlike the null hypothesis testing, the Bayesian hypothesis testing is really simple to understand. I set the following hypothesis:

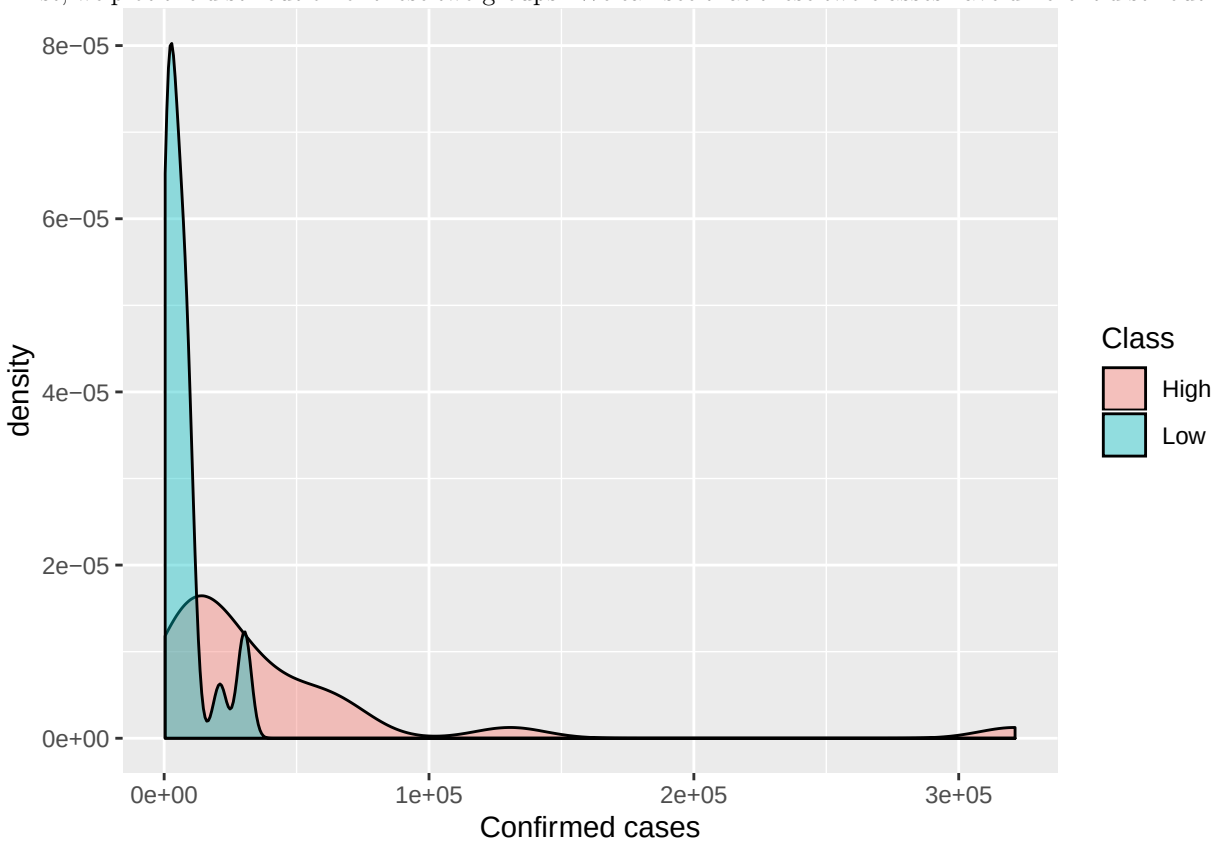
H_0 : two samples have no difference that implies transportation does not affect on confirmed cases

H_1 : two samples have differences that implies transportation affects on confirmed cases

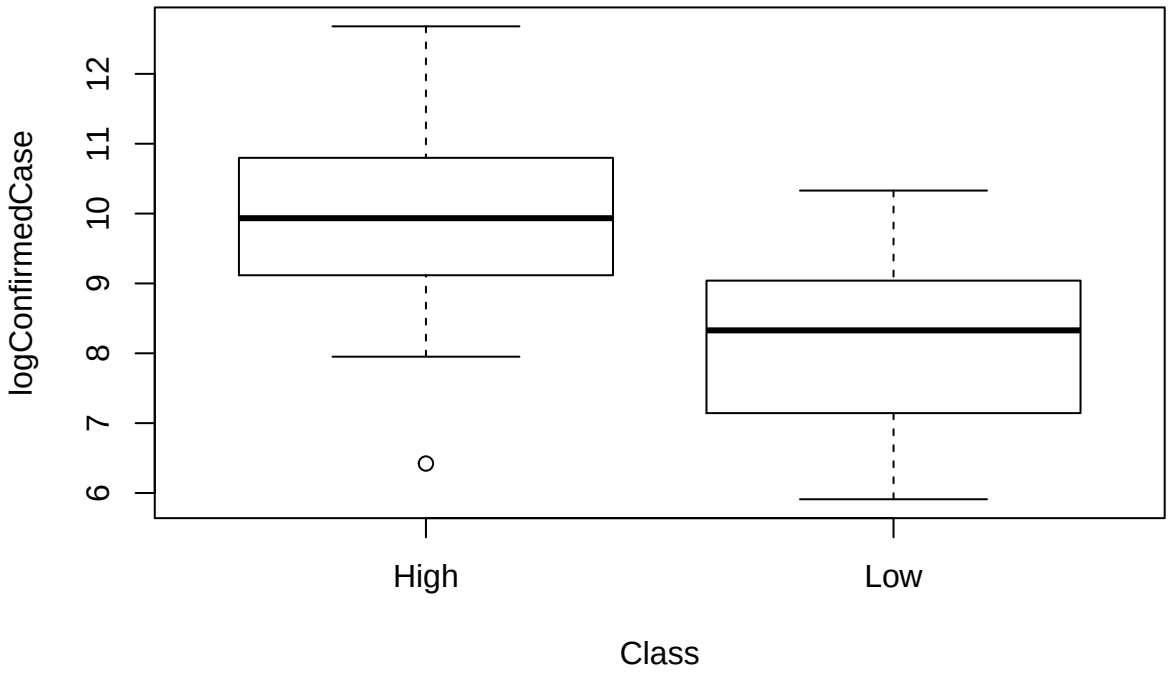
The Bayes factor plays a similar role to the p-value in the traditional hypothesis testing. I will use the Bayes factor to interpret my result.

Result

First, we plot the distribution of these two groups. We can see that these two classes have different distributions.



We can see from the previous plot that the data is right-skewed, so I perform a logarithmic transformation. Here is a summary of the data. The number of confirmed cases appears to be affected by transportation.



Before doing the Bayesian factor, I use a traditional t-test to show the result.

```
##
##  Two Sample t-test
##
## data:  logConfirmedCase by Class
## t = 4.611, df = 47, p-value = 3.092e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9504256 2.4216179
## sample estimates:
## mean in group High  mean in group Low
##           9.908112           8.222090
```

The p-value is really small so that I have strong evidence to reject the null hypothesis.

Let's now find the value of the Bayes factor.

```
## Bayes factor analysis
## --------------
## [1] Alt., r=0.707 : 610.2352 ±0%
##
## Against denominator:
##   Null, mu1-mu2 = 0
## ---
## Bayes factor type: BFindepSample, JZS
```

The Bayes factor is 610.2352.

Conclusion

I will use the Kass and Raftery (1995) table to interpret the result. According to Kass and Raftery (1995):

Bayes factor	Interpretation
1 - 3	Negligible evidence
3 - 20	Positive evidence
20 - 150	Strong evidence
>150	Very strong evidence

Therefore, the data are 610.2352 times as probable under the alternative as under the null hypothesis. There is very strong evidence that transportation affects the number of confirmed cases.