Advices About Flight Travelling

YICHUN ZHANG - 1006033187

April 16, 2021

Abstract

The study aims to give some advice for travellers when they are planning to travel by airplane. Data chosen for this study is downloaded from openintro and records all the flights departing from New York City in 2013. We will navigate through the data from a statistical perspective and use methods including maximum likelihood estimation, credible bayesian interval, confidence interval, hypothesis test, the goodness of fit and linear regression model separately to study for our interested questions. The key results for this study are:

- Air time is positively correlated to the distance between origin and destination.
- World flight mean (average) air time equals our sample mean air time.
- The maximum possibility of flight delayed happens when the mean equals to 1072.333.
- Flights departed between 12:01-24:00 is more than flights departed between 0:00-12:00.
- The proportion of flight delayed equals the proportion of flight not delayed.
- There is a 95% probability that the average of number of flight delay in month is between 1070.206 and 1074.284.

From these findings, we concluded that passengers should plan for air time according to their travelling distance. Passengers, for this stage, can belive that we are 90% confident that the world average flight delay time is between 149.4283 and 151.3788. We also found that The maximum possibility of flight delay happens when the mean equals 1072.333, and flights departed between 12:01-24:00 are more than flights departed between 0:00-12:00. Lastly, The proportion of flight delayed equals the proportion of flight not delayed and there is 95% probability that the average number of flight delays in a month is between 1070.206 and 1074.284.

Introduction

In the past year, many people have to put off their travel plans because of a global pandemic. Nowadays, with vaccines widely accepted and used, people have become more and more optimistic about the re-connection of the world. Many may have been getting more and more excited about their planned trip. Hence, this study will analyze the data called NYCflights, a collection of information for flights departing from New York City (NYC). This study is essential for those who are planning for a flight trip, especially those who plan to travel from NYC. Advice for preparing for the trip regarding what time to leave when travelling. Specifically, this study will highlight flight time (time on-air), delay time and flight number distribution at different times of a day. Variable air time describes how long the flight will be on the air in minutes. Arr_delay is the delay time of flight. A negative value means the flight is left early. Dep_time is the departure time of the flight. In the data summary section, we will use tables and graphs to display some important aspects visually. The following sections will study our questions and problem with statistical methods, like linear regression, goodness of fit, maximum likelihood model, etc. The statistical principle will be stated in the method part of the section, and the results will be stated in the result section. Needed mathematical justification of statistical method are displayed in appendix.

Terminologies

The common terminologies used in the reports are:

1. Mean: the average of the number set. In statistics, it measures the central tendency of a probability distribution.
2. Standard deviation: The Standard Deviation measures how spread numbers are. It is usually represented by the symbol σ.
3. Normal distribution: In probability theory, a normal distribution is a type of continuous probability distribution for a real-valued random variable.
4. Sample: the collection of part of population data.
5. Population: the set of every case that we want to study. In other words, it is the entirety of relevant data.
6. (True) parameter: Refers to the characteristics of a given population.

Hypothesis

Other relevant concepts related to this study will be specified in the following sections when needed. Hypothesis made for this study are:

- Air time is positively correlated to the distance between origin and destination.
- World flight mean (average) air time equals to our sample mean air time.
- Flights departed between 12:01-24:00 is more than flights departed between 0:00-12:00.
- The proportion of flight delayed equals to the proportion of flight not delayed.

Data

NYCflights data is taken from openintro, and is downloaded directly from its R package. It includes 32,735 rows (total flights) and 16 variables in total. To plan for a trip from the point of passengers, we are most interested in: dep_delay, air_time, and dep_time as mentioned above. There are no missing variables in this data set. Therefore, there is no filter-the-missing step. For the convenience of the following steps, we mutated a new variable called delay. If the flight is delayed (positive number in arr_delay), it is recorded as TRUE. Otherwise, it is stored as FALSE. In this section, by using the data summary techniques with R, we wish to get a brief answer of the following questions:
1. How long should we expect to wait for flight delay? 2. How long should we expect to stay on air? 3. What time flights are departed most frequent in a day?

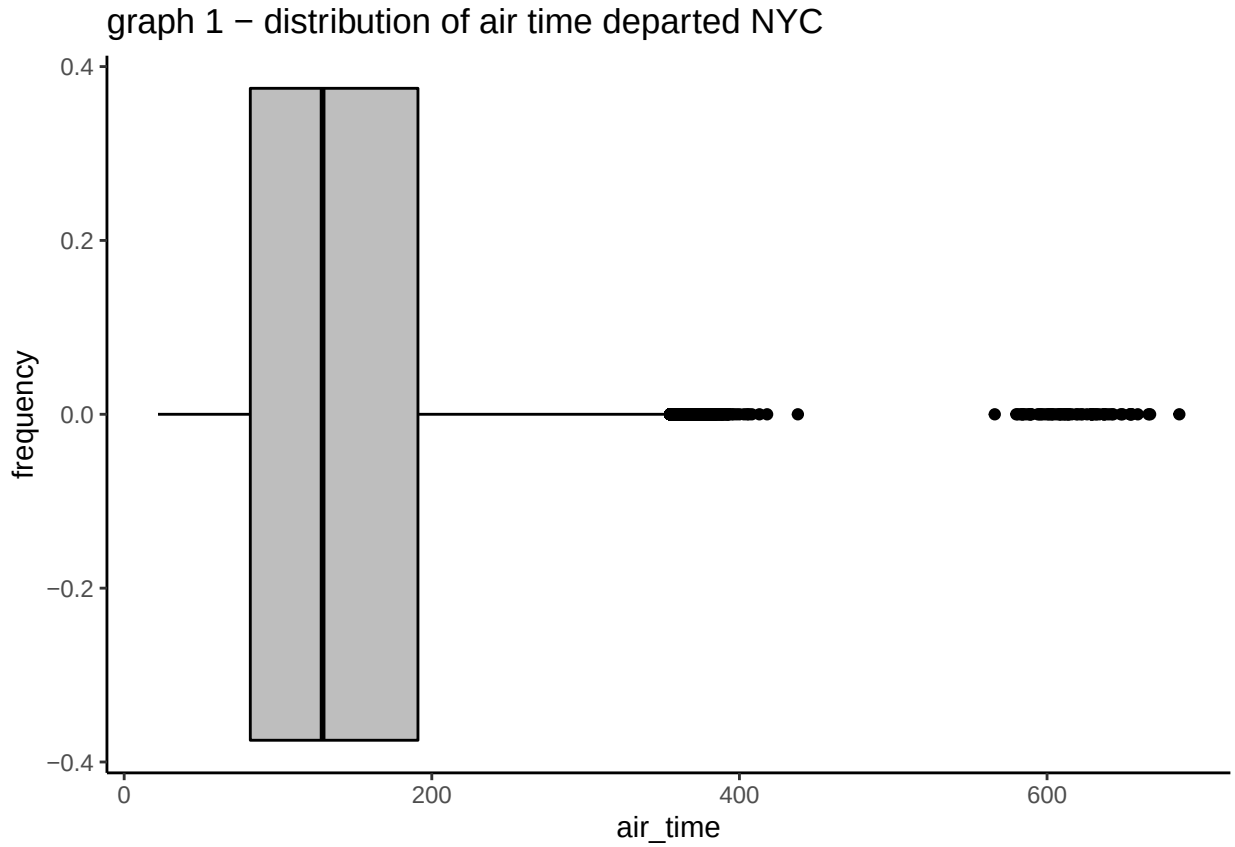
The key variables of interest for this study is:
- air time: time the flight spend on air, in minutes

- dep_delay/delay: time of the flight delayed, in minutes
- dep_time: time of departure with numerical expression, eg. 13:00 is recorded as 1300, 18:57 is recorded as 1857.
- distance: distance of travelling (from origin to destination).

Table1	Air-time mean	Air-time mean Standard Deviation	delay_time mean	delay_time Standard Deviation
Count	150.4419	93.52766	12.70515	40.40743

The mean (average) of flight time (spend on air) is 150.4419 minutes, and the mean (average) delay time

of all flights is 12.70515. The standard deviation of the air time is 93.52766, and the standard deviation of delay time is 40.40743. This means we should expect that the air time of flights is around 150 minutes, but it is not quite centred around the mean, so there is a lot of variance of air time. Similarly, we should expect flights to delay for around 12 minutes, but the variance of delay time happens.



This boxplot (graph #1) is centred at around 120. The right tail of the box plot is extended. This means that there are few outstandingly long flights in our record. The 25th percentile locates at around 80 minutes, and the 75th percentile of the distribution locates at around 200. This means half of the flights have air time between 80 minutes and 200 minutes. This is reasonable because domestic short-mid flights are more than long.

All analysis for this report was programmed using `R version 4.0.4`.

Methods

In following following sections, we will study our problems with statistical knowledge. First, I will give a brief explain about each method.

Linear Regression

Simple linear regression is a linear model for describing relationship between two varaibles. The methematical process is plotting two-dimensional sample points into a cartesian coordinate system, and find a line that best fit the relationship between x and y. In the simple linear regression model, with a bivariate data set of (x_i, y_i), we assume x_i is nonrandome while y_i is random for predicting. Values stored in each variable can be considered as samples. Samples of X are $X_1, X_2, X_3, X_4.....X_n$, and samples of Y are $Y_1, Y_2, Y_3, Y_4.....Y_n$. When given a non-random x, we can use our linear regression model to predict the corresponding y. Assumptions when using linear regression model are the following. First assumption is that the response variable, which is air

time of all flights among our sample, is normally distributed. This assumption is roughly satisfied in this case, because there are some outliers. Secondly, our sample is independent. Finally, expectation of $U_i = 0$ ($E(U_i) = 0$) and variance of $U_i = \sigma^2$ ($V(U_i) = \sigma^2$)in this case. The parameters we are interested in, in this qeuestion, is α and β.

We also assume that $y_1, y_2....y_n$ are realizatoins of random variables Y1, Y2, ...Yn. $X_1...X_n$ and $y_1...y_n$ satisfies a linear relation:

$$Y_i = \alpha + \beta x_i + U_i$$

for $i = 1, 2...n$ α is the y-intercept in this linear relationship; β is the slope (change of y divide by change of x). U_i can be understood simply as the difference of real Y_i and the estimated Y_i of each sample we calculated using this relationship. Mathematically, α measures where the line intercpet y-axis, and β measures the slope of the line. More practically, α is the value of y when x equals the zero, and β is the change of y when x increase by 1. In this certain dataset we chose, α do have a specific meaning, β means the how much does occupancy of shelters increases by 1.

Confidence Interval

For this study, I will do confidence interval with bootstrap, and use confidence interval 90%. Our result should give us a range that we can be 90% confident about our true parameter. The confidence interval is a possible set of true parameter values. More precisely, it gives an approximation of a series of values that is likely to have an unknown true parameter. A confidence level is a percentage that is correlated with each confidence interval. This percentage represents how confident we are that the results will capture the true population parameter, relying on the bond's luck together with your random sample. As I wish to account for 90% of the possible results, It means that 1-α is 90%. Bootstrap sampling distribution follows a normal distribution, and a 90% means the middle 90%. In this study,we will use Bootstrap. Bootstrap is a statistical method that uses random sampling with replacement to estimate the sampling distribution about a given population. These repeated experiments are called resamples. The empirical bootstrap is one type of bootstrap, which samples from an estimator's sampling distribution without specifying the data distribution. Since we don't know our original data distribution, we'll use empirical bootstrap in this study. Moreover, the z-distribution assumes that we know the population standard deviation; however, the t-distribution is only based on the sample standard deviation. In our analysis, we will use t distribution because we do not know the true parameter. Also, since we don't know the true parameter's value, we'll use a sampling estimator to approximate it. Nevertheless, since each sample is unique (chosen at random), each estimator is unique. In this case, we can use the confidence interval to determine how often an interval would include the true parameter. The assumption for using this method is: sample is independent, and the parameter of interest is: mean air time of flights.

Maximum Likelihood Estimator

Maximum likelihood estimation (MLE) is used to estimate the parameter of probability distribution. The parameter that determinies the model distribution is unknown, MLE can give it a estimation. The parameter values are found such that they maximise the likelihood that the process described by the model produced the data that were actually observed. Estimator in MLE is a equation or function with the unknown parameter, and the estimate the result from plugging parameter into the estimator. We assume that sample is independent, and the number of delay time follows a poisson distribution with mean being λ. Given that some parameter λ characterizes the distribution of random sample $x_1, ...x_n$. In this case, the possibility for sample to be true depends on λ, and we want to find λ that maxmizes the likelihood of correctness. In other words, we aim to find a $lambda_0$ such that $\lambda_0 = \lambda$ such that true sample distribution equals to sample distribution. λ is assumed to be constant when we use MLE. We are interested in the true (world) number of flight delay during a year (See mathematical justification of in appendix section).

Hypothesis Test

A statistical hypothesis test is a statistical inference method. There is a set of two hypothesis to make before the study, alternative hypothesis (H_0) and null hypothesis (H_a). Null hypothsis is a claim about a population parameter that is assumed to be true until it is declared false. Hypothesis test can specifiy which outcome of study can reject H_0 and support H_a, by using significance test. Significance test is done according to a threshold probability—the significance level—the data would be unlikely to occur if the null hypothesis were true. Rejection of H_0 happens when $p \leq \alpha$ or the test statistics falls into rejection region. It is shown with very small p-value. When p-value is less than 0.001, there is very strong evidence to reject H_0 and support H_a. We use 2 types of error to distinguish between the null hypothesis and the alternative hypothesis, type 1 and type 2 error. The first type of error occurs when the null hypothesis is wrongly rejected. The second type of error occurs when the null hypothesis is wrongly not rejected. Our assumptions for doing this test is that sample independent. The paramter of doing a hypothesis test can be mean, median, proportion and so on.

Goodness of Fit Test

Assume equation shows the likelihood for this sample to happen when given parameter θ is $L(\theta) = f(x_1, x_2...x_n|\theta)$, then $L(\hat{\theta}) = f(x_1, x_2...x_n|\hat{\theta})$ shows the likelihood of sample happening when parameter changes to $\hat{\theta}$. Similarly, $L(\theta_0) = f(x_1, x_2...x_n|\theta_0)$ is the likelihood of sample happening when parameter changes to θ_0. In likelihood ratio test, θ_0 is a claim, means assuming that $\theta = \theta_0$. The likelihood ratio would be

$$\Lambda(\theta_0) = \frac{L(\theta_0)|X1, X_2, ...X_n)}{L(\hat{\theta}_0)|X1, X_2, ...X_n)}$$

. Our assumptions for Goodness of fit is that: sample is independent and flight delay distribution follow bernuilli distribution with the proability of delay being p_1, not delay being p_2.

Bayesian Credible Interval

Bayesian inference is another statistical inference method. It is used to find a estimated range of paramter of our interest. It is normally used to update the probabilty, when more information are given. We assume that the number of flight delay in a month follows a poisson distribution with mean being λ. We assume that λ is random, and its distribution is assumed to be $Exp(\beta = 1072.333)$. We are interested in finding the 95% credible inrerval of λ, which is the number of flight delay happen in month.

If the likelihood of parameter happening is $f(x_1, ...x_n|\theta)$, then by the law of total probability, these ingredients specify a joint distribution for $(x_1, ...\theta)$: $g(\theta)f(x_1, ...x_n|\theta)$. The marginal distribution for the data would be: $m(x_1, ...x_n) = \int g(\theta)f(x_1, ...x_n|\theta)$Bayesian inference computes the posterior probability according to Bayes' theorem:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) * P(B|A)}{P(B)}$$

, and posterior distribution of θ is a conditional distribution of θ, when s is given. The definition of posterior distribution is:

$$Posterior\, density = \frac{Joint\, density\, of\, (x_1, ...x_n, \lambda)}{Marginal\, distribution\, of\, data}$$

Results

The result we get, for each method we used, are explained below.

Linear regression.

How long should we expect to stay in the airplane? My assumption is that flying time of the flight depends on the distance for the destination. The longer the distance is, the longer the flight time will be. We will use

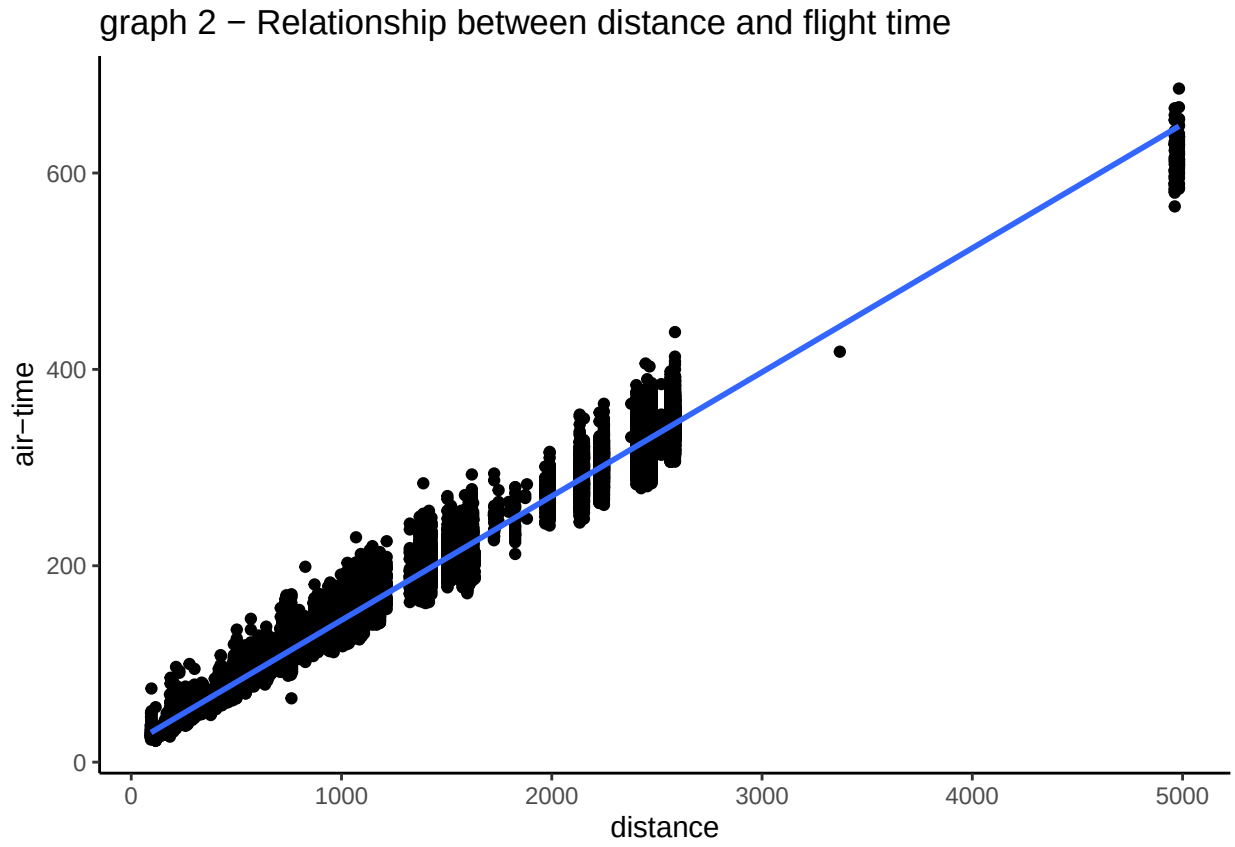
our sample data to make a linear regression to investigate the relationship betwwen distance and air time of flights.

The actual output is:

Table2	Estimate std.	error	t-value	Pr(>
(intercept)	-1.230e+02	1.044e+00	-117.8	<2e-16
air_time	7.772e+00	5.892e-03	1319.1	<2e-16

$\hat{\beta} = 7.77178$. This means for every unit increase in distance, the average of air time is expected to increase by 7.77178. The p-value for hypothesis $H_0 : \beta = 0$ vs $H_a : \beta \neq 0$ is 0, so we have strong evidence to show that there is (no if pvalue is large) linear relationship between gained and weight. This means that air time is strongly correlated to distance of travelling. Therefore, when planning for a trip, passengers should consider distance of travelling to leave time for spending on air.

Visually, this plot can show the relationship between air time and distance on a x-y coradination. See graph #2:



The x-axis of the plot is the distance, and the y-axis is the air time.As we can see, points on the graph are located near the line, which means the correlation is quite strong. Also there is a gap in distance, from 2700 to 5000. This fact makes sence with geographical knowledge. Distance over 2500 may go over one of the four oceans. There is no major countries on the ocean, so flights will not stop in the middle.

Confidence Interval

How long is the flight delay should we include when we make a plan? In this study, I will use our data to find out the true (world-wide) flight delay time using confidence interval.

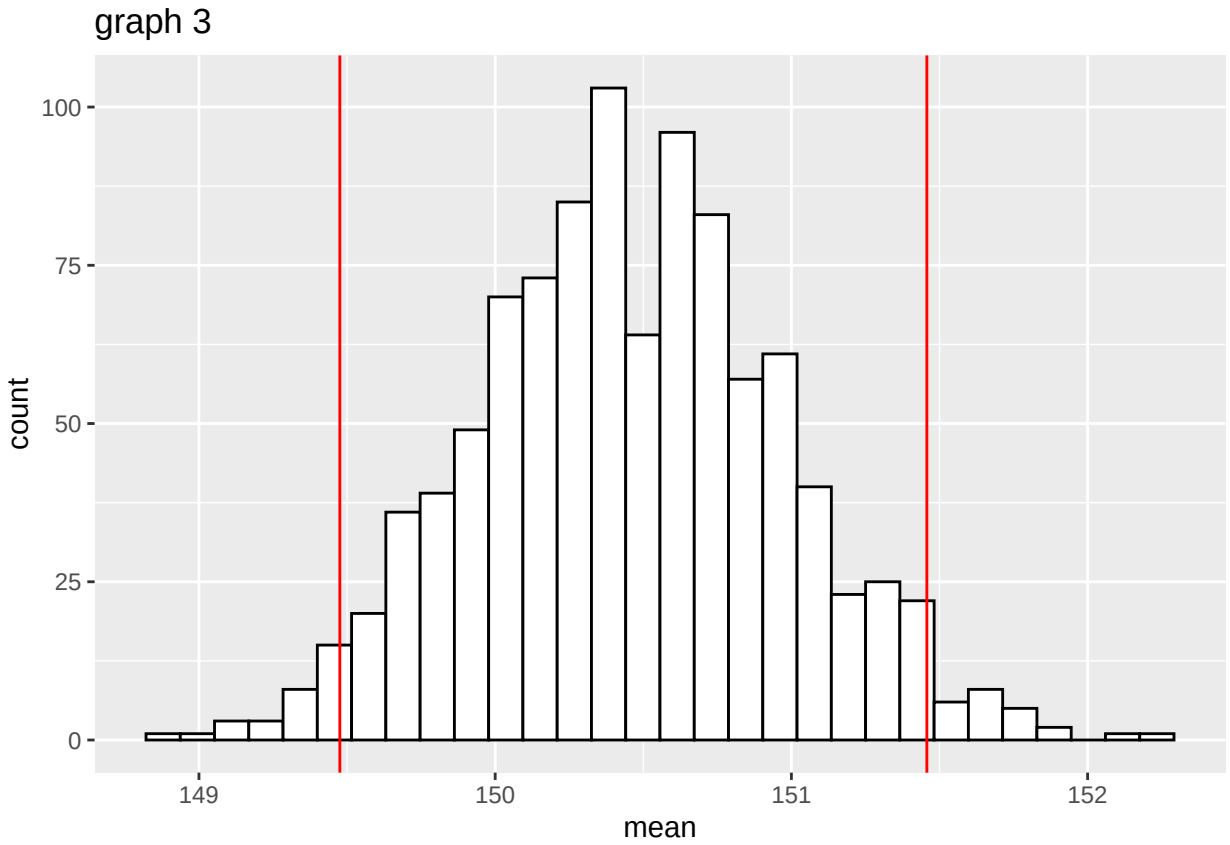


Table3	5%	95%
	149.6279	151.3354

The sample size of this bootstrap is: the original size of sample the all the values in our study. We resampled with replacement for 1000 times, and sum up all the results. We get a distribution of the sampling result. The distribution is shown in a histogram. The distribution is often normal distributed, since we are just sampled randomly with replacement using our original data. Given that our prefered confidence level is 90%, we can find the confidence interval. Values between the 2 red lines are in the 95% interval. We are 90% confidence to show that the true (population) average air time of is between 149.6279 and 151.3354. This result makes sense. The sample mean 150.4419 is bounded in this range and the number at 5% percentaile and 95% percentile is close to our sample mean. We can imagine this two numbers being the two-sided tails for the distribution. This means passengers, for current stage, can believe that there is a 90% chance that the world flight time average is between 149.6279 and 151.3354.

Maximum Likelihood Estimator

Flight delay is one of the problems that most passenger will consider when travelling with plane. Delay of flight can disturb the trip and make the experience less enjoyable. In this section, we use MLE to answer the question: When does the prossibiltity of flights being delayed is maxmized?

If $X_1, X_2, ...X_n$ are continuous random variables with joint pdf (probability distribution function) $f(x_1, ...x_n|\theta)$,

where θ is the parameter of interest. Then, for a given vector of observations $(x_1, x_2, ...x_n)$, we have the likelihood of θ being $L(\theta) = f(x_1, ...x_n|\theta)$. Parameter of interest, in this case, is λ, and our function is the function of poission distribution.

Given that $X_1, ..., X_n \overset{iid}{\sim} Poission(\lambda)$

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_1!}$$

. We then use the derivitive to find the λ that maxmizes L. This process is done with R code (Mathematical process in Appendix), and the output is:

Table4	mean
	1072.333

The estimated true (population) number of flight delay in a month is 1072.333. This is when the paramter (mean frequency of delay/$lambda$) is maxmized. This means that it is most likely, in a month, that 1072.333 flights will be delayed. This result seems normal and reasonable.

Hypothesis Test

Dep_time is the depature time of flights in local time zone. When dep_time is smaller than 1200 in our data, the flight is departutred during 0:00-12:00. In the contrast, when dep_time is larger than 1200, the flight is departured during 12:00-24:00. It makes sense that more flights are departed between 18:00-24:00 compared to 0:00-6:00, because 0:00-6:00 mid-night time. Since there are more flights departed in the second half of day, the mean of flight departure time should be more biased to night-time. Therefore, the hypothesis of this section is: the mean of flight departure time is higher than 1200.

In this study, we are interested in the average (mean) of dep_time. Our hypotheses are:

$H_0 : \mu = 1200$
$H_a : \mu > 1200$, where μ represents the mean of dep_time.

```
## # A tibble: 1 x 3
##    mean     n    sd
##   <dbl> <int> <dbl>
## 1 1349. 32735  489.
```

Table5	mean	n
	1349.266	32735

Test statistic is obtained with the formula:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Then, we use this formula to get p-value: $p - value = P[T \geq t]$ or $P[Z \geq z]$

Table6	p-value
	0.0003135572

Since p-value is 0.0003135572, which is below the threshold, we have strong evidence to reject H_0. This result

support our original hypothesis before testing, which is H_a in our testing process. This means that there are evidence supports that flight departure time is biased toward night time (12:00-24:00). Therefore, when travelling, it is encouraged to choose early flights if it is needed to avoid crowd.

Goodness of Fit Test

Our hypothesis is : $H_0 : p_1 = 0.5$ and $p_2 = 0.5$.

Table7	ratio	p-value
result	0.9842849	0.8587325

The output shows we have no evidece to against the null hypothesis: $H_0 : p_1 = 0.5$ and $p_2 = 0.5$. In other word, we fail to find evidence for rejecting that delay happens half of the time. Though we cannot provide true delay proportion, passengers can think delay time will happen half of time, and use this fact as a guidance for planning.

Bayesian Credible Interval

How many flight delays happen in one month for flights departed form NYC? In this section, we want to find a range for the true true number of flight delay in a month.

The number of delay in each month is recorded as the following:

Table8	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
delay_freq	1136	992	1146	1281	1047	1252	1302	1154	660	956	1015	1521

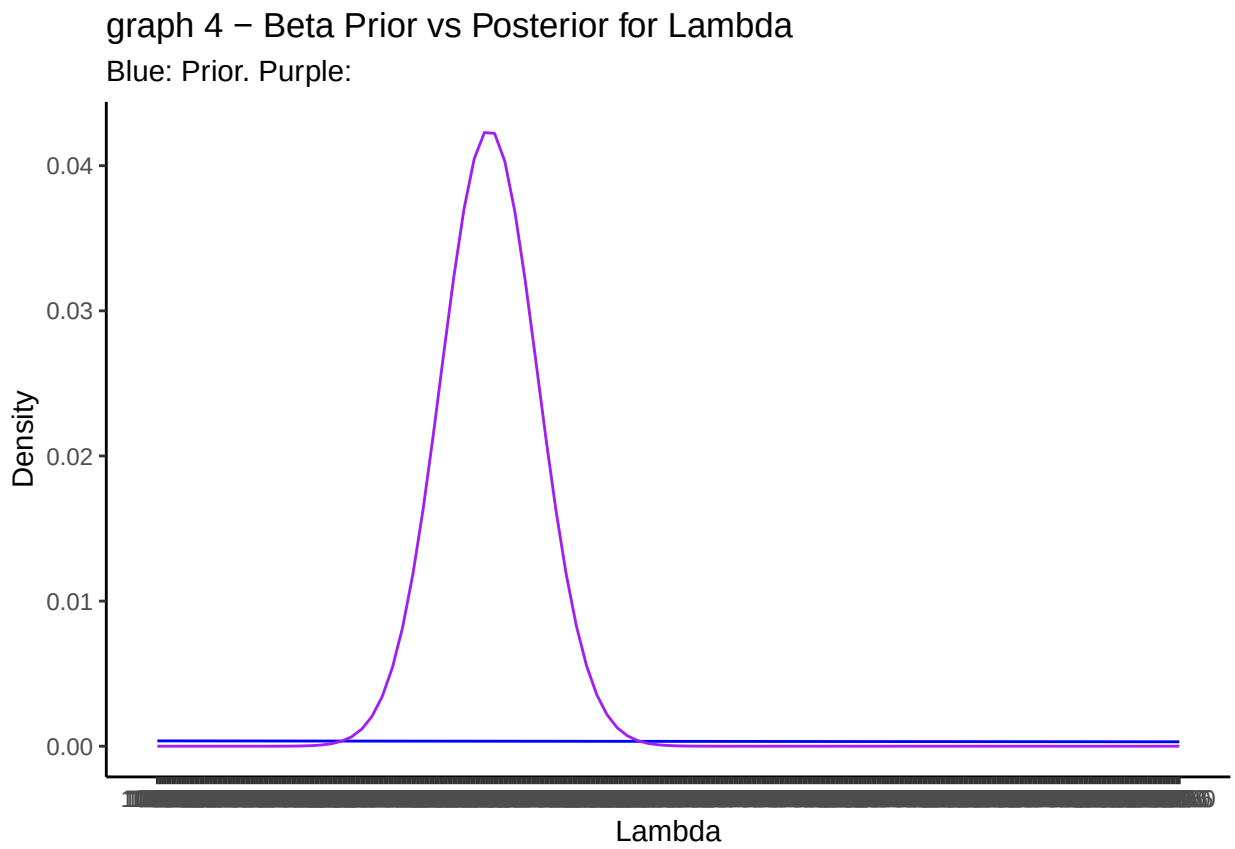


Table9	0.025	0.0975
	1070.206	1074.284

The blue line of the graph is prior and the purple line of the graph is distribution of λ. The purple line shows a smooth and normal flow of distribution. The part where purple line is above blue line are the range of 95%.

By using `R`, we see there is 95% probability that the parameter λ (the average of Number of hospital visits during pregnancy) is between 1070.206 and 1074.284. This means our estimated range for the true paramter (frequency of flight delay in a month) is between 1070.206 and 1074.284 with our data and assumption. This result seems normal and sense-making when we compare this result with the number of delay in each month in the above graph.

All analysis for this report was programmed using `R version 4.0.4`.

Conclusions

This study explored air-time, delay time and frequency of delay for flights, from a statistical perspective. We used maximum likelihood estimation, bayesian credible interval, confidence interval, hypothesis test, goodness of fit and linear regression model to study our problems. Our key problems are about the air time and delay time of flights. From this study result, passengers can get some adivces about planning for a flight trip. Most results are align with our hypothesis. Key results we found are: - Air time is positively correlated to the distance between origin and destination.
- World flight mean (average) air time equals to our sample mean air time.
- The maximum possibility of flight delayed is happens when mean equals to 1072.333.
- Flights departed between 12:01-24:00 is more than flights departed between 0:00-12:00. - The proportion of

flight delayed equals to the proportion of flight not delayed.
- There is 95% probability that the average of number of flight delay in month is between 1070.206 and 1074.284.

Big picture

Overall, this data is just one year of the world database. Studying this dataset can give us some sense of the big picture of all Canada flight, while we do also need to expand and elaborate on our study to obtain a more scientific result. The work we've done can be useful for future studies. For example, the bootstrap method we used this time can be when we need to estimate a parameter, while we do not have the whole population we need as a sample. Using bootstrap, though this cannot give a certain number, we can find a range of where the parameter can be at.

Weaknesses

The study is made based only on fundamental statistical knowledge. We made lots of assumptions about sample and distributions in this study, which are not neccessarily true. When we are with more knowledge more methods, accuracy of our study can increase. More sophisticated models may be applied for this data better than ours.

Next Steps

In the future, possible improvements includes: 1. Accomplish this study with more models that can study our questions better. 2. Elaborate more on methods and results to see we can find more interesting facts about our study.

Discussion

This report focus on one data set from openintro, and study it from different aspects using statistical methods. We mainly focus on air time, delay, and departure time of the data. We summarized data with tables and graphs. Methods section introduces most statistical methods we are using. They are: linear regression model, maxmimum likelihood estimation, hypothesis test, confidence interval, goodness of fit and bayesian credible interval. By using R, we tested our hypotheis and solved our questions. At last, the mathematical justification of MLE and Bayesian credible interval is also included as a reference.

Bibiography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)
4. "Homepage." City of Toronto Open Data Portal, open.toronto.ca/.
5. Alison Gibbs, Alex Stringer (2021) *Chapter 6: introduction to Bayesian inference.* Probability, Statistics, and Data Analysis.[https://awstringer1.github.io/sta238-book/section-introduction-to-bayesian-inference.html]
6. Brownlee, Jason. "A Gentle Introduction to the Bootstrap Method." Machine Learning Mastery, 25 May 2018, machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/.
7. Richard Arnold Johnson; Dean W. Wichern (2007). Applied Multivariate Statistical Analysis. Pearson Prentice Hall. ISBN 978-0-13-187715-3. Retrieved 10 August 2012.

8. Mark Galassi, Jim Davies, James Theiler, Brian Gough, Gerard Jungman, Michael Booth, and Fabrice Rossi. GNU Scientific Library - Reference manual, Version 1.15, 2011. Sec. 21.7 Weighted Samples
9. The World Question Center 2006: The Sample Mean, Bart Kosko
10. Richard. "Sample Mean and Covariance." Wikipedia, 2015, en.wikipedia.org/wiki/Sample_mean_and_covariance.
11. Brownlee, Jason. "A Gentle Introduction to the Bootstrap Method." Machine Learning Mastery, 25 May 2018, machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/.

Appendix:

Equation of calculating mean:

$$\bar{x_n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Equation of calculating standard deviation:

$$s_n^2 = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i^2 - n\bar{x}^2)}{n-1}}$$

MLE

We assume that sample is independent, and the number of delay time follows a poisson distribution with mean being λ, since are interested in the frequency of delay happening within an interval of time.

The likelihood function is

$$L(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

Now, for a easiler computation, we compute the log likelihood function $l(\lambda)$:

$$l(\lambda) = log(L(\lambda)) = log(\frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}) = log(e^{-n\lambda} * \lambda^{\sum_{i=1}^{n} x_i}) - log(\prod_{i=1}^{n} x_i!) = log(e^{-n\lambda}) + log(\lambda^{\sum_{i=1}^{n} x_i}) - log(\prod_{i=1}^{n} x_i!) = -n\lambda + \sum$$

We set derivative being 0 to find the maxmimum:

$$l'(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i - 0 = 0$$

This means :

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i = n => \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Therefore, the estimator for Poission(λ) will happen when $\hat{\lambda} = \bar{x}$

Now we use second derivative to test it:

$$l''(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i.$$

Since $n > 0$ and $\lambda^2 > 0$, $l''(\lambda) < 0$.

Bayesian model

We assume that the number of flight delay in a month follows a poisson distribution with mean being λ, since flight delay time is a frequency within a time interval.

Assume the posterior distribution of λ follows gamma distribution, $\pi(\lambda|x_1,...,x_n) \propto Gamma(n\bar{x}+1,(n+\frac{1}{\beta})^{-1})$

Suppose $X_1,...,X_n \overset{iid}{\sim} Poisson(\lambda)$, $f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ and $\lambda \sim Exponential(\beta)$, $f(x) = \frac{1}{\beta}e^{-\frac{\lambda}{\beta}}$ The likelihood function of λ is $L(\lambda) = f(x_1,x_2,...,x_n|\lambda)$

$$L(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda_i^x}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} = \frac{e^{-n\lambda}\lambda^{n\bar{x}}}{\prod_{i=1}^n x_i!}$$

According to the Posterior density function (Chapter 6, Probability, Statistics, and Data Analysis):

$$Posterior\,density = \frac{Joint\,density\,of\,(x_1,...x_n,\lambda)}{Marginal\,distribution\,of\,data}$$

$\pi(\lambda|x_1,...,x_n) = \frac{f(\lambda)f(x_1,...x_n|\lambda)}{\int f(\lambda)f(x_1,...x_n|\lambda)d\lambda}$

Marginal distribution of data $\int f(\lambda)f(x_1,...x_n|\lambda)d\lambda$ is the normalizing constant (Chapter 6, Probability, Statistics, and Data Analysis).

then $\pi(\lambda|x_1,...,x_n) \propto f(\lambda)f(x_1,...x_n|\lambda)$

$\pi(\lambda|x_1,...,x_n) \propto \frac{e^{-n\lambda}\lambda^{n\bar{x}}}{\prod_{i=1}^n x_i!} * \frac{1}{\beta}e^{-\frac{\lambda}{\beta}}$

$\pi(\lambda|x_1,...,x_n) \propto \frac{1}{\beta\prod_{i=1}^n x_i!}e^{-\lambda(n+\frac{1}{\beta})}\lambda^{n\bar{x}}$

Given x_i is knowen, $\prod_{i=1}^n x_i!$ is a constant, we can simply:

$\pi(\lambda|x_1,...,x_n) \propto \frac{1}{\beta}e^{-\lambda(n+\frac{1}{\beta})}\lambda^{n\bar{x}}$

Gamma distribution: $f(\lambda) = \frac{\lambda^{\alpha^*-1}e^{-\frac{\lambda}{\beta^*}}}{\Gamma(\alpha^*)\beta^{*\alpha^*}}$

Supposing $\alpha^* = n\bar{x}+1$ and $\beta^* = (n+\frac{1}{\beta})^{-1}$

$\frac{\lambda^{\alpha^*-1}e^{-\frac{\lambda}{\beta^*}}}{\Gamma(\alpha^*)\beta^{*\alpha^*}} = \frac{(n+\frac{1}{\beta})^{n\bar{x}+1}}{(n\bar{x})!}e^{-\lambda(n+\frac{1}{\beta})}\lambda^{n\bar{x}}$

Thus, posterior distribution of λ follows gamma distribution, $\pi(\lambda|x_1,...,x_n) \propto Gamma(n\bar{x}+1,(n+\frac{1}{\beta})^{-1})$

In conclusion, the posterior distribution of λ follows gamma distribution, $\pi(\lambda|x_1,...,x_n) \propto Gamma(n\bar{x}+1,(n+\frac{1}{\beta})^{-1})$. In function $\pi(\lambda|x_1,...,x_n) \propto \frac{1}{\beta}e^{-\lambda(n+\frac{1}{\beta})}\lambda^{n\bar{x}}$, n is the sample size, $\bar{x}$ is the sample mean.