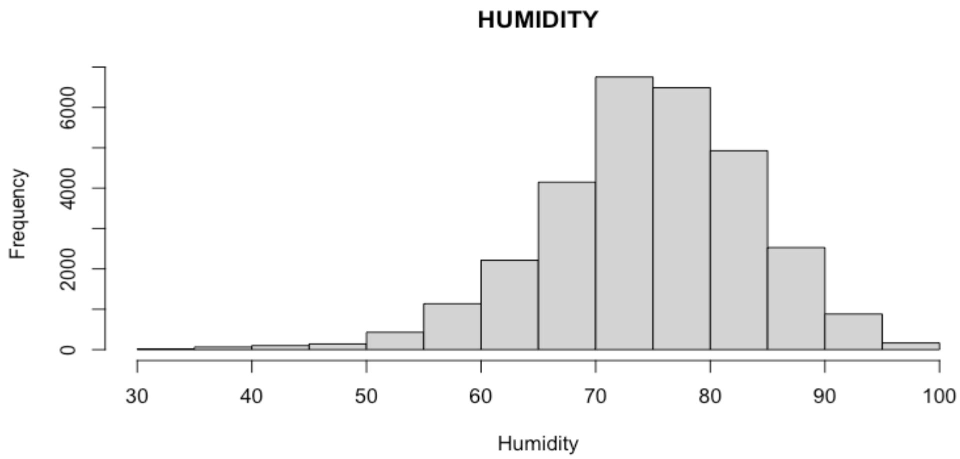
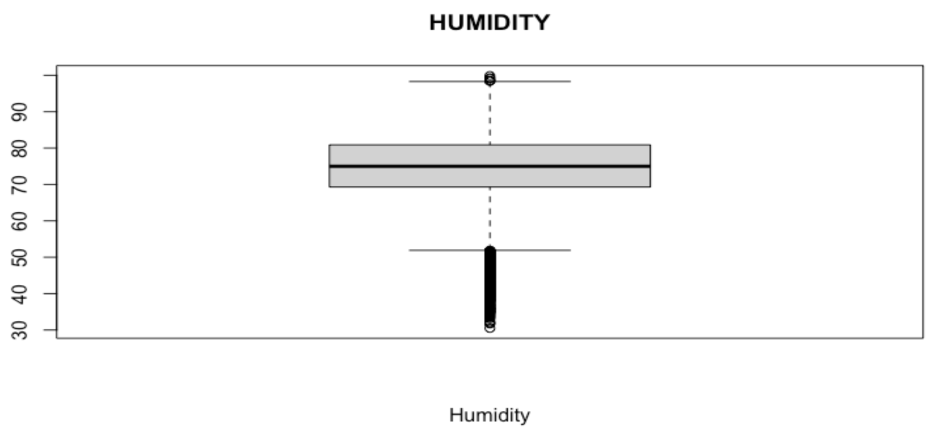
SIT743 Bayesian Learning and Graphical Models
Assignment-1

Said Abdullahi
sabdullahi@deakin.edu.au
221377847

Question 1.1

Draw a histogram and boxplot for the *'Humidity'* variable.





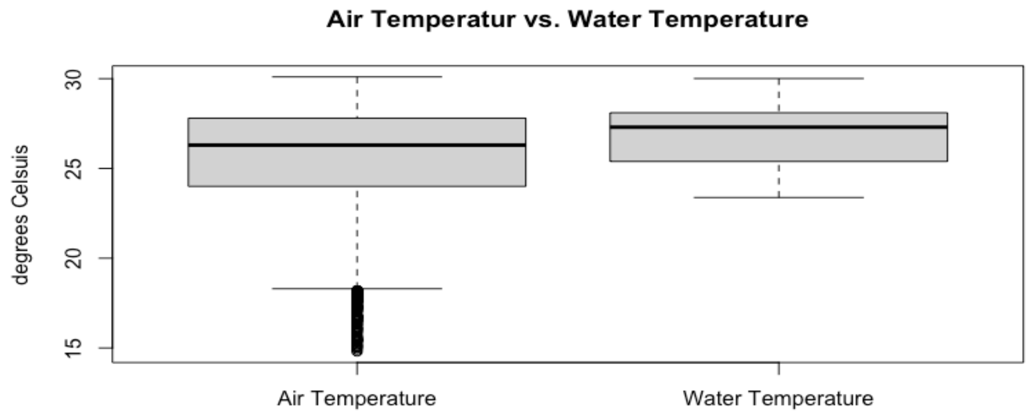
Min	1st Qu.	Median	Mean	3rd Qu.	Max
30.50	69.30	75.00	74.66	80.90	99.90

The histogram and boxplot of the distribution of humidity in percentages show that most of the data are between 70-80 with a range of 69.4. Our distribution is unimodal which negatively skewed. There are outliers with most at the lower end of the distribution. The median is 75 and a mean of 74.66, this indicates that the skewness is negatively with outliers most at the lower end. The interquartile range is 80.90 – 69.30 = 11.6.

Question 1.2: Which summary statistics would you choose to summarize the center and the spread for the 'Humidity' variable? Why?

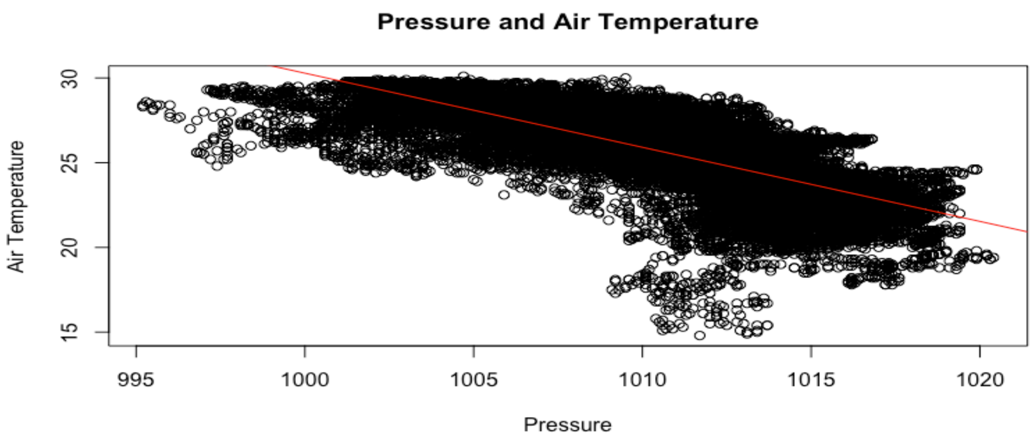
The median is a better measure of centre, since the distribution of humidity has outliers that are mostly at the lower end and negatively skewed. The median is more robust when there are outliers and data is skewed compared to the mean. While median is more robust when that the data is negatively skewed the interquartile is a robust measure for spread when the data has outliers. This is because IQR measure the range between the 1st Quartile (25%) and 3rd Quartile (75%) this excludes the outliers.

Question 1.3: Draw a parallel box plot for variables 'AirTemperature' and 'WaterTemperature'. Compare and comment on the results.



The two boxplots overlap therefore there are some similarities between the two. The medians are close. The air temperature is more variable (greater variance) while water temperature has smaller range which indicates that their data points are close to the centre values. Air temperature has many outliers while water temperature there is no outliers. Air temperature is more skewed than the water temperature.

Question 1.4: Draw a scatterplot of '*Pressure'* (as x) and '*AirTemperature*' (as y). Name the axes.



The association between air temperature and pressure is negative linear relationship. The correlation between air temperature and pressure is r = -0.733, which indicates a negative and moderate linear relationship. Correlation does not mean causation. The slope is -0.44 which states that 1 point increase in pressure lead to a decline of 0.44 point of air temperature. The regression equation is: air temperature = 468.08 – 0.44. The coefficient of determination is 53.80, which indicates that 53.80% of the variation of the air temperature is explained by the variations in the Pressure.

Question 1.5

a) 3-way table for WSB, PreB, and WaterTB

			WSB		
		PreB	**High**	**Low**	**Total**
	High	**High**	152	192	344
		Low	3065	5283	8348
WaterTB	**Low**	**High**	2881	2446	5327
		Low	600	596	1196
	Moderate	**High**	1149	2067	3216
		Low	4023	7546	11569
		Total	11870	18130	30000

b) Use the above obtained cross table to answer the following questions. Show all the steps/workings clearly. Consider that a record (row) is selected from the data at random

i. what is the probability that the *WSB* is low?

P(WSB=LOW) = 18130/30000 = 0.604

ii. what is the probability that the *WaterTB* is moderate given that the *WSB* is high?

P(WaterTB=moderate|WSB=high) = (1149 + 4023)/ 11870 = 0.436

iii. what is the probability that the *WaterTB* is low given that the *WSB* is high and the *PreB* is low?

P(WaterTB=low|WSB=high and PreB=low) = 600/(3065+600+4023) = 600/7688 = 0.078

iv. Are low *WaterTB* and high *PreB* mutually exclusive? Explain

P(WaterTB=low|PreB=high) = (2881 + 2446) / (344+5327+3216) = 5327/8887 = 0.599

P(WaterTB=low|PreB=high) = 0.599 ≠ 0

Explain: No, its disjoint because the P(WaterTB=low|PreB=high) = 0.599 ≠ 0.

v. Are high *WSB* and low *WaterTB* independent events? Explain.

P(WSB=high|WaterTb=low) = (2881+600) /(5327+1196) = 3481/6523 = 0.534

P(WSB=high) = 11870/30000 = 0.396

They are not equal so we can consider it as not independent because the difference is greater than the sampling errors.

Question 2.1

a) State two differences between frequentist way and the Bayesian way of estimating a parameter.

1. **How parameter is treated:** The first difference between frequentist and Bayesian is that the parameter θ is fixed parameter and tries to find the point estimate of the parameter with a certain level of confidence while the Bayesian way of estimating parameter θ is considered to be random variable with a probability distribution.

2. **Interpretation of uncertainty and probability:** The frequentist probability is considered objective. It is the long run frequency of an event occurring in a repeated trail. In contrast Bayesian considers probability is a measure of uncertainty based on prior knowledge and current data. Fornacon-Wood et al., 2021 states that "[t]he fundamental difference between these 2 schools is their interpretation of uncer- tainty and probability: the frequentist approach assigns probabilities to data, not to hypotheses, whereas the Bayesian approach assigns probabilities to hypotheses. Furthermore, Bayesian models incorporate prior knowl- edge into the analysis, updating hypotheses probabilities as more data become available." (p.1). In other word the probability is objective for frequentist and subjective for Bayesian.

Source: SIT743 Week 3 lecture notes

Source: Fornacon-Wood, I., Mistry, H., Johnson-Hart, C., Faivre-Finn, C., O'Connor, J. P. B. & Price, G. J. (2021*). Understanding the Differences Between Bayesian and Frequentist Statistics*. Journal of Thoracic Oncology, 16(5), 749-753.

b) How the uncertainty (or variance/error) in an (parameter) estimate is computed using the frequentist approach?

In the frequentist approach the bootstrap method is used to obtain the error bars. It takes a multiple Data of size N (obtaining by sampling with replacement) then we estimate the parameter this is done multiple times. Then the statistical accuracy of parameter estimate can be evaluated by looking at the variability of predictions between different bootstrap datasets. Note that estimated parameter is done using Maximum Likelihood Method. We can compute the standard deviation or standard error of the estimate obtained through bootstrap datasets.

Source: SIT743 Week 3 lecture notes

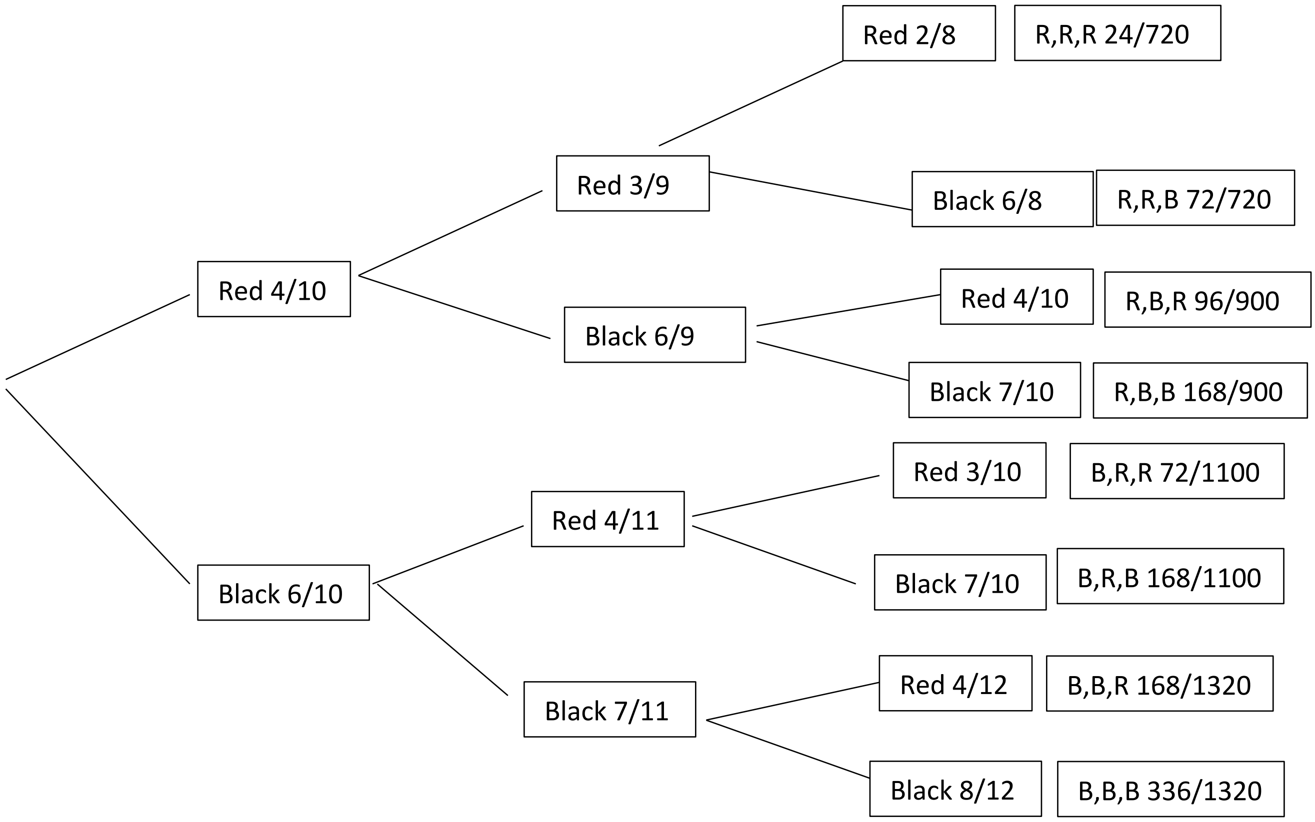
c) Why are conjugate priors useful in Bayesian statistics? Give an example of a Conjugate pair.

Conjugate priors are useful because it allows us to simply get the posterior distribution. Especially if the prior and the likelihood belong to the same conjugate pair which mean that the posterior will also be the same conjugate family. For example, Donovan & Mickey 2019 give us a problem using beta-binomial conjugate. This is the White House Problem. A beta prior distribution plus a binomial data gives a beta posterior distribution. This always to get the parameter in simple way.

Source: Book: *Bayesian Statistics for Beginners: a step-by-step approach, Therese Donovan and Ruth M. Mickey, Oxford University Press, 2019.*
https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198841296.001.0001/oso-9780198841296

Question 2.2

a) **Draw tree diagram**



b) What is the probability that John ended up keeping two black marbles with him? (i.e., *what is the probability that John has chosen two black marbles from the basket over the three trials*?)

RBB, BRB, and BBR

P(two black marble) = P(Red, Black, Black) + P(Black, Red, Black) + P(Black, Black, Red) = 168/900 + 168/1100 + 168/1320 = 0.187 + 0.153 + 0.127 = 0.467

c) What is the probability that only the third one john has chosen is the red marble?

P(Black, Black, Red) = 168/1320 = 0.127

d) What is the probability that john gets at least one red marble?

P(one red marbles) = 1 – P(no red marbles) = 1 – P(Black, Black, Black) = 1 – 336/1320

= 1- 0.255 = 0.745

e) Given that his 3^{rd} selection was a black marble, what is the probability that his 2^{nd} selection was a black marble?

P(2^{nd} trail = Black | 3^{rd} trail=Black) = P(Red, Black, Black) + P(Black, Black, Black)

= 168/900 + 336/1320 = 0.187 + 0.255 = 0.442

Question 3

Question 3.1.

$$X_i \sim Exp(\theta) = p(x_i|\theta) = \theta e^{-(xi\theta)}$$

a) Show that the expression for the likelihood, p(X|θ), of N printers (X= {x_1, x_2 … x_n} can be given by the below equation (show the steps clearly to obtain this).

The joint distribution of n trials will be multiplication of each individual distribution of each trail.

$$P(D|\theta) = P(x_1|\theta) = P(x_1|\theta) \text{ x } P(x_2|\theta) \text{ x } \dots \text{ x } P(x_n|\theta)$$

$$= \Pi^n_{i=1} P(x_i|\theta) = \Pi^n_{i=1} \theta e^{-(xi\theta)}$$

$$= \theta e^{-(x1\theta)} \text{ x } \theta e^{-(x2\theta)} \text{ x } \dots \theta e^{-(xn\theta)}$$

$= \theta^n e^{-(x1+x2+x3\dots xn)\theta}$ we used the exponential rule a^n= a x a x a x a

$$= \theta^n e^{-(w\theta)} \text{ , where } W = \Sigma^n_{i=1} x_i = x_1 + x_2 + x_3 \dots x_n$$

$$\text{Likelihood, } P(X|\theta) = \theta^n e^{-(w\theta)} \text{ , where } W = \Sigma^n_{i=1} x_i$$

b) Find a simplified expression for the log-likelihood function L(θ) = ln (p(X|θ))

First, we take the log of the likelihood function Ln ($\theta^n e^{-(w\theta)}$)

= Nlnθ - Wθ

Log rules:

$Loga^x$ = xloga

c) Show that the Maximum likelihood Estimate (θ) of the parameter θ is given by:

$$\theta = \frac{N}{W}$$

then we differentiate,

$$\frac{L(\theta)}{(\theta)} = \frac{d}{d\theta}\,(N\ln\theta - W\theta) = 0$$

$$= \frac{N}{\theta} - W = 0$$

d) **From the past history, the company has found out that the lifetimes of eight of the 3D printers were {7, 5, 8, 12, 10, 8, 9, 8}. Find the Maximum likelihood Estimate of the parameter $\boldsymbol{\theta}$ given this data? Hence, find *on average, how long a certain 3D printer last?***

On average, how long a certain 3D printer last is:

$$W = \Sigma^{n}_{i=1} x_i = 7 + 5 + 8 + 12 + 10 + 8 + 9 + 8$$

$$= 67$$

$$N = 8$$

$$\theta = \frac{N}{W} = \frac{8}{67} = 0.12$$

e) **Hence, find the probability that a 3D printer lasts for more than 8 years. (Hint: you may use cumulative distribution function (CDF) of exponential distribution. The cdf of an exponential distribution is given by F(t) = 1 – $e^{-(t\theta)}$**

F(t) = 1- $e^{-(8 \times 0.12)}$= 0.6321, this the cumulative distribution function

The probability that a 3D printer goes for more than 8 years is

=1- 0.6321 = 0.3679 or 36.79%

Question 3.2: The company consulted a 3D printer manufacturing company to learn more about the lifetime of the printers. Through discussions with the engineers of that company, it has found out that the pattern of the parameter (θ) follows a Gamma distribution, G*(a,b)*, as given below, with hyperparameters a = 0.1 and *b*= 0.2.

G*(a,b)* = Z $b^{a}e^{(a-1)}e^{-(b\theta)}$, where Z is a constant

a) Posterior is proportional to likelihood times prior

$$P(\theta|D) \propto P(D|\theta) \text{ x } P(\theta)$$

$$= \theta^{n}e^{-(w\theta)} \text{ x Z } b^{a}e^{(a-1)}e^{-(b\theta)}$$

$$= \text{Z } b^{a}\theta^{(n+a-1)} \text{ x } \theta e^{-(w\theta)+-(b\theta)}$$

$$= Z\, b^a \theta^{(n+a-1)} \times \theta e^{-\theta(b+w)},\ \text{ where } W = \Sigma^n_{i=1} x_i$$

$$= Z\, b^a \theta^{(a`-1)} \times \theta e^{-\theta(b`)},\ \text{where Z is a constant (we remove it)}$$

$$P(\theta|X, a, b) = \text{Gamma}(a+n, b+w)$$

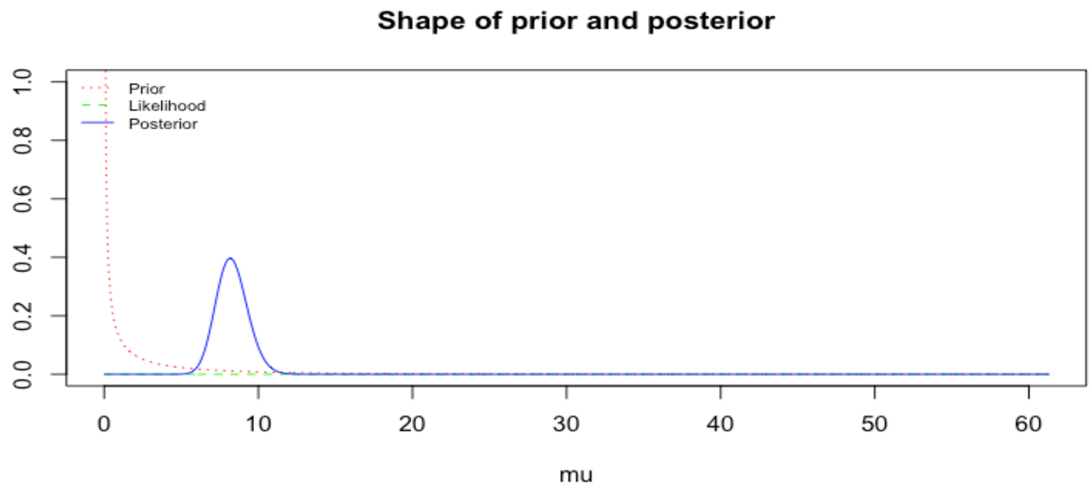
b) a=0.1 and b=0.2. data is {7,5,8,12,10,8,9,8}. N=8

find the values a` and b`

$$a` = a + n = 0.1 + 8 = 8.1, \quad b` = b + w = 0.2 + 67 = 67.2$$

$$\theta = 8.1/67.2 = 0.12$$

c)



Poisson is the inverse of exponential distribution. The graph is a Poisson with a gamma prior. We can see that the mu is around 8 or 9. Inverse_to_get_exponential_mu <- 1/mean(results) = 1/8.296296 = 0.121. Prior gamma + Possion data = posterior gamma.

Question 4: Bayesian inference for Gaussians (unknown mean and known variance)

a) Poster distribution of θ with observation X={$x_1 = 0.5$}, m=0, $\tau^2=2^2$, and $\sigma^2=1^2$.

Prior is a Gaussian with P(θ) ~N(m, τ^2). Likelihood is a gaussian with P(X/θ)~N(θ, σ^2). Posterior is a Gaussian with P(θ|X)~ N(μ_n, σ^2_n) where, μ_n is the mean of the posterior and σ^2_n is the variance of the posterior.

$$\mu_n = \sigma_n^2 \left(\frac{N\bar{x}}{\sigma_n^2} + \frac{m}{\tau 2} \right)$$

$$\frac{1}{\sigma_n^2} = \frac{N}{\sigma^2} + \frac{1}{\tau^2} \ \text{ where, } \ \bar{x} = \frac{1}{N} \Sigma_{i=1}^{N} x_i$$

Source: SIT743 Week 4 Prac Task2

b) Find the mean and standard deviation for the posterior distribution for θ for each of the following values of n:

i. n=10

ii. n=100

Finding mean and standard deviation for n=10

$$n=10,\ m=0,\ \tau^2=4,\ \sigma^2=1,\ \bar{x} = x_1 = 0.5$$

$$\frac{1}{\sigma_n^2} = \frac{10}{1} + \frac{1}{4} = \frac{41}{4} \Rightarrow \sigma_n^2 = \frac{4}{41}$$

$$\sigma = \sqrt{\frac{4}{41}} = 0.31$$

$$\mu_n = \frac{4}{41}\left(\frac{10 \times 0.5}{1} + \frac{0}{4}\right) = 0.09756(5 + 0) = 0.4878$$

Finding mean and standard deviation for n=10

$$n=100,\ m=0,\ \tau^2=4,\ \sigma^2=1,\ \bar{x} = x_1 = 0.5$$

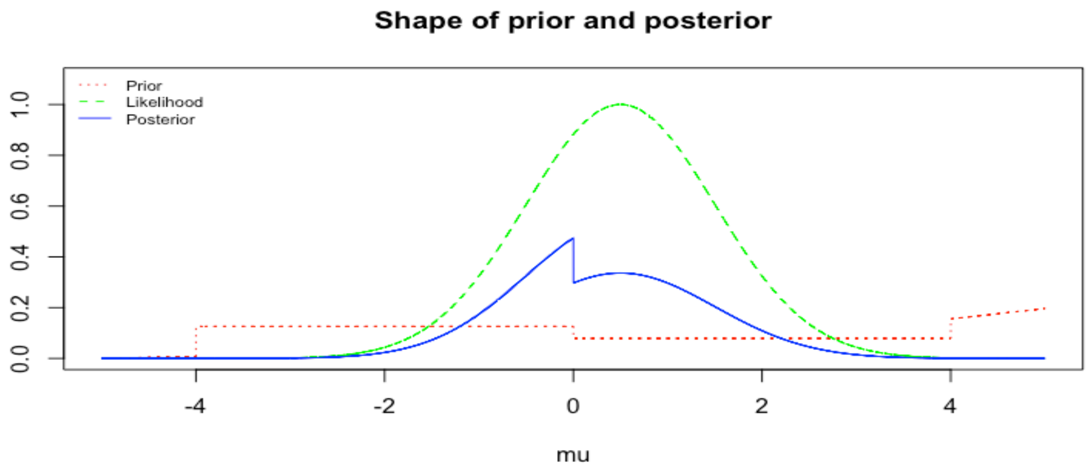
$$\frac{1}{\sigma_n^2} = \frac{100}{1} + \frac{1}{4} = \frac{401}{4} \Rightarrow \sigma_n^2 = \frac{4}{401}$$

$$\sigma = \sqrt{\frac{4}{401}} = 0.099875$$

$$\mu_n = \frac{4}{401}\left(\frac{100 \times 0.5}{1} + \frac{0}{4}\right) = 0.00997506(50 + 0) = 0.49875$$

The variance at n=10 is 0.09756 and at n=100 is 0.00997506. This is showing that the variance declines tenfold when we increase n by tenfold and this in turn increase the mean slightly.

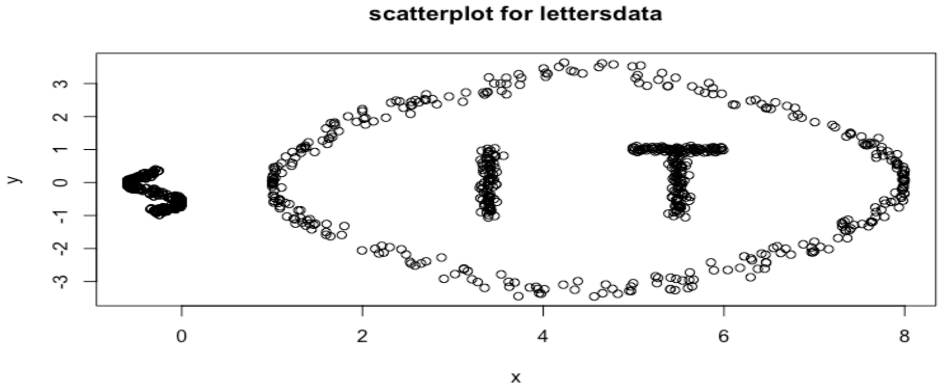
d) Write a R program to implement this prior, and compute the posterior distribution considering n = 1. Using R program find the *posterior mean estimate* of θ and the *posterior standard deviation*. Sketch, on a single coordinate axis, the obtained prior, likelihood and posterior distributions.



We can see that the posterior is affect by the prior and exhibiting a normal/gaussian distribution with a mean of estimate of 0.3228789 and posterior standard deviation of 1.029051.

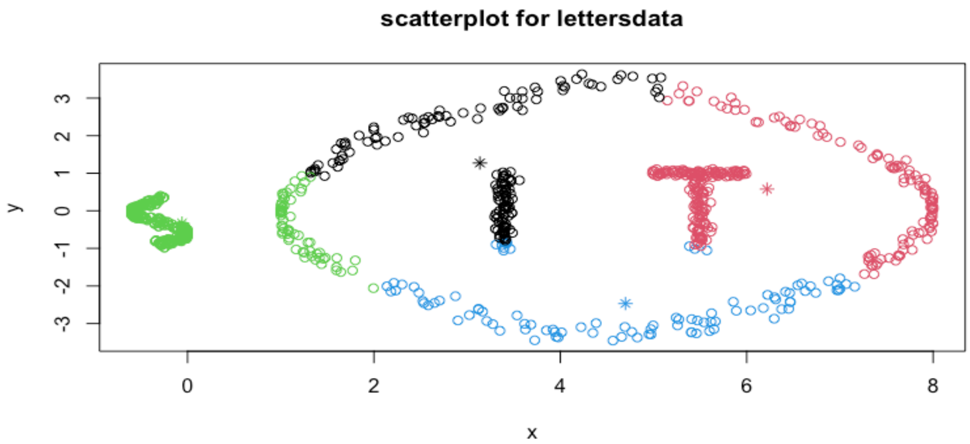
Question 5: K-Means clustering

a) Scatterplot



There seems to be 4 classes/clusters in data as seen in the above scatterplot. Maybe the S is one cluster the I and T their own cluster and the circle its own cluster. If I would group this dataset I would use 4 clusters/classes.

b) Use the above number of classes as the k value and perform the k-means clustering on that data. Show the results using a scatterplot (show the different clusters with different colours). Comment on the clusters obtained.



We used four clusters for the letter dataset. Each letter is its own cluster and they are taking some of the circle with them. With the blue cluster taking the bottom of the letter I and T. This is due the cluster centres located close to the letter and the bottom of the circle. The third cluster has the least data with 90. Between_SS/total_SS is 84.4% which is good but it could be better.

K-means clustering with 4 clusters of sizes 277, 400, 90, 159

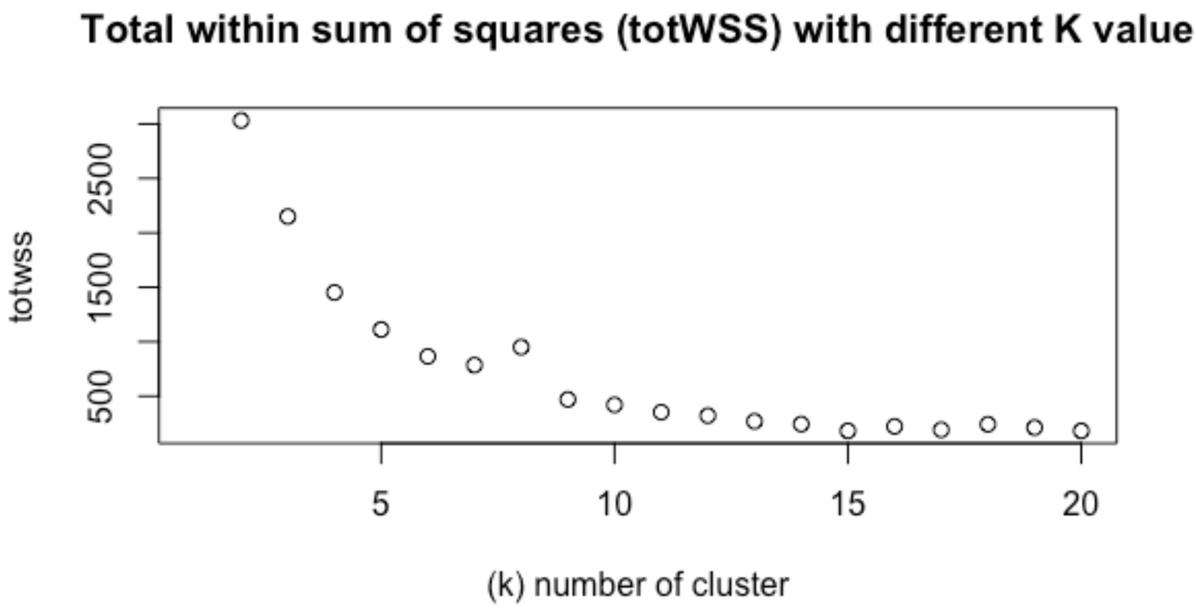
Cluster means: (points)

```
          x          y
1  6.21510650  0.5730457
2 -0.05926076 -0.3254617
3  4.77839610 -2.5916661
4  3.14651195  1.1917342
```

Within cluster sum of squares by cluster:
[1] 580.1895 222.2355 238.9843 406.4542
 (between_SS / total_SS = 84.4 %)

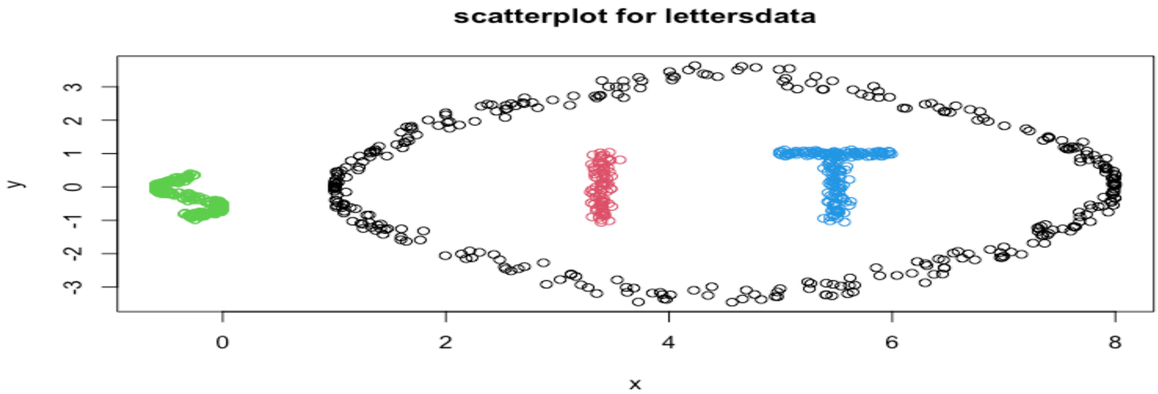
Spectral clustering would be better.

c) Vary the number of clusters (k value) from 2 to 20 in increments of 1 and perform the k-means clustering for the above data. Record the *total within sum of squares (TOTWSS)* value for each k, and plot a graph of TOTWSS verses k. Explain how you can use this graph to find the correct number of classes/clusters in the data.



This graph sometimes will tell us the optimal clusters to use. However, with our graph it is hard to tell how many clusters is optimal. The elbow is not clearly defined. Maybe 6 clusters. Total Within Sum of Squares graph indicates the total sum of squared distances between each data point and the centroid (point of each cluster) of its given cluster for different values of clusters.

Question 5.2: Spectral Clustering



K-means clustering with 4 clusters of sizes 336, 85, 336, 169

Cluster means:

```
          [,1]       [,2]          [,3]
1  0.000000e+00 0.00000000 -5.445296e-02
2  0.000000e+00 0.00000000  1.162738e-18
3 -1.317336e-19 0.05433495 -2.712276e-20
4 -7.669770e-02 0.00000000  1.079942e-19
```

Within cluster sum of squares by cluster:

```
[1] 3.718090e-03 5.400557e-31 8.031751e-03 5.851187e-03
 (between_SS / total_SS =  99.2 %)
```

Visually there seems to be an improvement with spectral clustering. Each letter is its own cluster and the circle is its own cluster. There is also an improvement with the between_SS/total_SS with the kmean being 84.4 and spectral clustering being 99.2% which is a significant improvement.

Question 6

a)

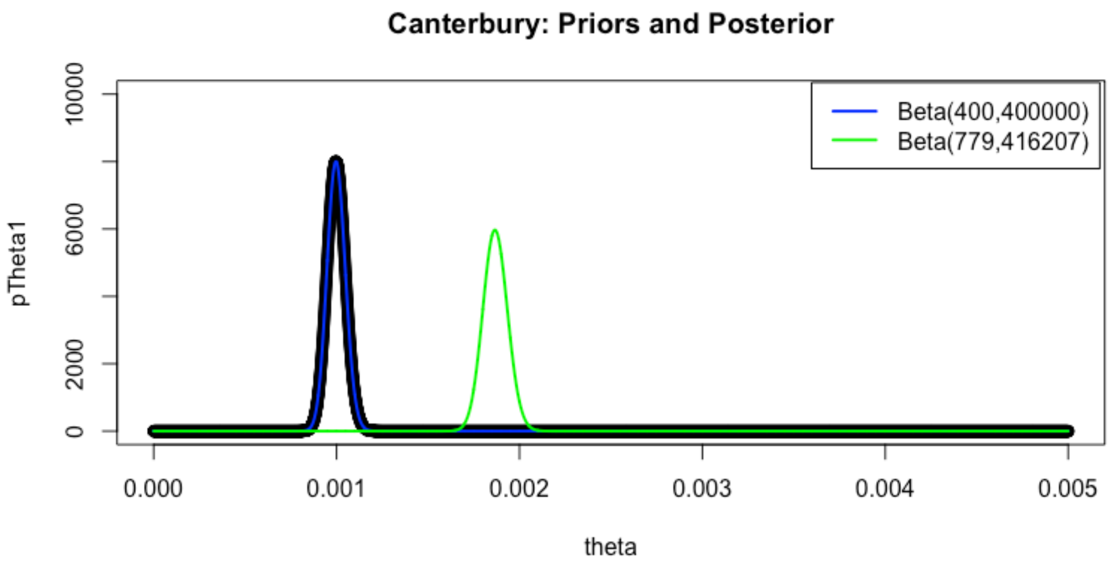
i. What parameters are estimated in this work?

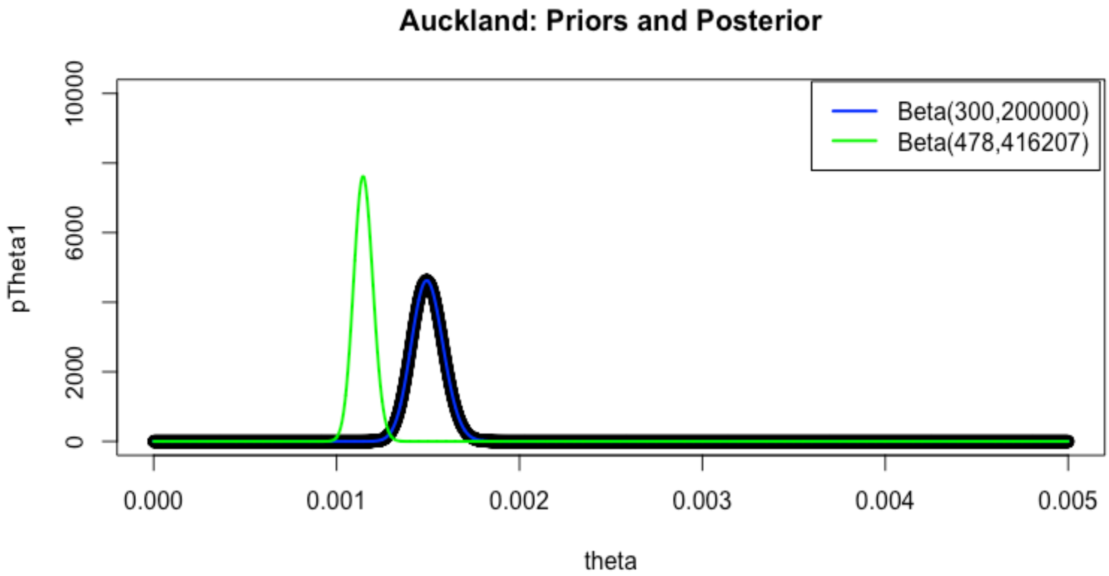
The parameter being estimated is the Posterior Beta(a +y, b+N-y) and the Posterior median at 95% credible interval.

ii.

Table 1. Bayesian analysis of Canterbury and Auckland COVID-19 data sets - 7 April 2023

Study	Cities	Deceased (y)	Total Cases (N)	Prior Beta(a, b)	Posterior Beta($a + y$, $b + N - y$)	Posterior median (95% credible interval) (%)
	Canterbury	379	297169	Beta(400, 400,000)	Beta(779, 696790)	0.11 (0.10 - 0.12)
	Auckland	178	216385	Beta(300, 200,000)	Beta(478, 416207)	0.11 (0.10 - 0.13)





We can see that with Canterbury the prior theta is 0.001 and the posterior theta slightly more. However for Auckland the prior is close to 0.001 and the posterior is greater between 0.001 and 0.002.

b)

i. ***What is MCMC (Markov Chain Monte Carlo) sampling, and how MCMC sampling-based method can be used for Bayesian posterior estimation?***

When Bayesian conjugates cannot be used to solve our problem, we can use MCMC. Remember Bayesian conjugates are special cases. The MCMC build the posterior from scratch. According to Donovan & Mickey, 2019, "[t]he main idea behind MCMC for Bayesian inference is that you “build” the posterior distribution by drawing samples from it”. And this is done by using Bayes’ Theorem. We rewrite it as posterior ∝ likelihood * prior this is done by holding the denominator of the Bayes’ Theorem constant. The algorithm Metropolis is used for the MCMC analysis with 8 steps of operations.

1. Select a plausible value (parameter) from the posterior distribution
2. We calculate the posterior density of the observing data (x) under our chosen value in step 1. Plug this into the likelihood and prior distribution equation with the chosen hyper parameter.

$$\text{Posterior} \propto \text{likelihood} * \text{prior}$$

3. Next we choose a value randomly from a symmetrical distribution. Now we a new random chosen value.
4. Now we calculate the posterior for the new random chosen value.

$$\text{Posterior} \propto \text{likelihood} * \text{prior}$$

5. The next step we throw one away. We use the Metropolis algorithm:

$$P_{move} = \min\left(P(\theta_{new}|\text{data}) / P(\theta_{old}|\text{data}),\ 1\right)$$

For example:

$$P_{move} = \min\left(0.0113 / 0.0161,\ 1\right)$$

$$P_{move} = \min\left(0.7019,\ 1\right) = 0.7019$$

Please see (Donovan & Mickey, 2019) for detail on above example.

6. In the next step we choose a random number from a uniform distribution and if the number is less than the probability of moving (0.7019) we accept the new value (parameter)
7. We repeat the steps above hundredths or thousandths of times.
8. We summarize the accepted values (parameter) in the form of a distribution with summary statistics. And this is posterior distribution!

Note: The above question is based on the book - *Bayesian Statistics for Beginners: a step-by-step approach, Therese Donovan and Ruth M. Mickey, Oxford University Press, 2019. https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198841296.001.0001/oso-9780198841296]*

ii. Write Report on chosen journal or article

Title: "Did noise pollution really improve during COVID-19? Evidence from Taiwan."

Authors: Rezzy Eko Caraka, Yusra Yusra, Toni Toharudin, Rung-Ching Chen, Mohammad Basyuni, Vilzati Juned, Prana Ugiana Gio, and Bens Pardamean.

Introduction

Covid-19 was first reported in Wuhan, China in December 2019. The virus quickly spread around the world leading to large outbreaks. World Health Organisation declared a global health emergency shortly after on 20 January 2020. The Covid-19 pandemic had significant impact on public health and governments around the world where implementing policies to counter the adverse effects of Covid-19. However, according Caraka et al., 2021 states that noise pollution improve in Taiwan during the COVID-19 pandemic. This is due to the government implementing lockdown consequently "most people stayed at home and worked from home, which probably reduced the noise pollution". (Caraka et. al., 2021, p1)

Dataset

This papers dataset are from the Local Environmental Protection Bureaus, the EPA, Executive Yuan, Taiwan, from 2015 to the 4^{th} quarter of 2020. (Caraka et al., 2021, p4) According to Caraka et. al., 2021) "The dataset consists of the cases of over-standard noise per time frame, number of noise petitions, number of industry petitions, number of motorcycles, number of cars and density of vehicles." (p.4)

Method

Caraka et al., 2021 uses "various statistical methods following odds ratios, Wilcoxon and Fisher's tests and Bayesian Markov chain Monte Carlo (MCMC) with various comparisons of prior selection." (p.1) Bayesian method is used in order to obtain expectations and make predictions. However, due to complexity in empirically calculating the expectations using analytical methods the paper uses Markov Chain Monte Carlo (MCMC). (Caraka et al., 2021, p4)

Results

The study found that noise pollution has declined during the COVID-19 pandemic in Taiwan. The Fisher's tests is significant at $\alpha = 5\%$. Household activities did not change with people still going supermarkets to buy groceries. According to Caraka et al., 2021 this because "in Taiwan there was no lockdown nor large-scale restriction, then the number of cases of over-standard noise per time frame before and during COVID-19 should be the same".p(5) This is "[i]n line with this, with the *p*-value of both Wilcoxon's and Fisher's tests, there is a significant difference in the values of motorcycles, cars and vehicle density both before and during the COVID-19 period." (Caraka et al., 2021, p. 5) In orther words the reduction in noise pollution may be due to changes in transportation rather than changes in household activities. The measures noise pollution using Bayesian MCMC using 10 different priors. Caraka et al., 2021 found that [t] he Bayesian model that provides high accuracy is Bayesian-MCMC-AIC prior with R^2 84.70%". (p.8)

Unfortunately, this paper does not state what software is used for evaluations. However, we can assume that they used statistical software for data analysis. The benefit of this paper is that uses data from reliable sources which increases robustness of the results. However the limitation is that the study is based on Taiwan which means it may not be generalisable to other countries.

Conclusion:

This study found that noise pollution declined in Taiwan during the COVID-19 pandemic. This was mainly due to people driving less and staying home.

Reference:

1. *Bayesian Statistics for Beginners: a step-by-step approach, Therese Donovan and Ruth M. Mickey, Oxford University Press, 2019. https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198841296.001.0001/oso-9780198841296*
2. Caraka, R. E., Yusra, Y., Toharudin, T., Chen, R.-C., Basyuni, M., Juned, V., Gio, P.U., & Pardamean, B. (2021). Did noise pollution really improve during COVID-19? Evidence from Taiwan. Science of the Total Environment , 764, 142787.