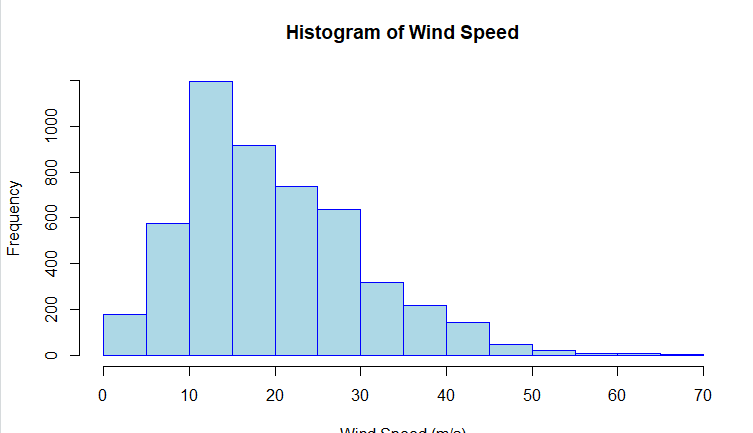
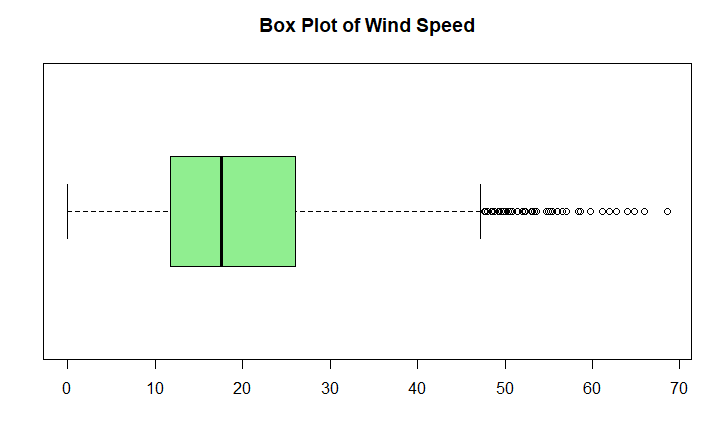
**1) Use the sampled data (“my.data”) to answer the following questions.**

**1.1) Draw a histogram and a box plot for the ‘Wind speed’ variable. Provide a five number summary for the Wind speed values. Use these to comment about the distribution of the Wind speed variable.**

****

****

**Five Number Sumamry**

****

**1.2) Which summary statistics would you choose to summarize the center and the spread for the ‘Wind speed’ variable? Why?**

**Central Tendency: Median**

The median is preferred measure for central bias in this context because it is less sensitive to outliers and skewed data than mean. Given that wind speed can have a wide range of values as indicated by significant difference between minimum and maximum in dataset.

**Spread: Interquartile Range (IQR)**

The IQR is chosen to summarize spread because it focuses on middle 50% of data and calculated as difference between the third quartile (Q3) and the first quartile (Q1. This measure of spread is particularly robust in the presence of outliers or a skewed distribution, which is common in environmental data like wind speed.

**Interpretation**

Given provided five-number summary for **Wind speed** variable, I would select median to summarize center of distribution because it accurately reflects midpoint of data and unaffected by potential outliers or extreme values. For spread of data I would choose Interquartile Range (IQR) because it offers a measure of the middle 50% of data .

**1.3) Draw a scatterplot of ‘Temperature’ (as x) and ‘Humidity‘ (as y). Name the axes. Fit a linear regression model to the above two variables, and plot the (regression) line on the same scatter plot. Write down the linear regression equation. Compute the correlation coefficient and the coefficient of Determination. Explain what these results reveal.**

**Linear Regression Equation**

Equation is Humidity = 101.712 - 2.244769 \* Temperature. This means that for every one degree Celsius increase in temperature humidity is expected to decrease by about 2.245 percentage points on average. The intercept (101.712) represents estimated humidity when temperature is 0 degrees Celsius.

**Correlation Coefficient**

The correlation coefficient of -0.6445942 suggests a moderate to strong negative linear relationship between temperature and humidity. As temperature increases humidity tends to decrease.

**Coefficient of Determination (R²)**

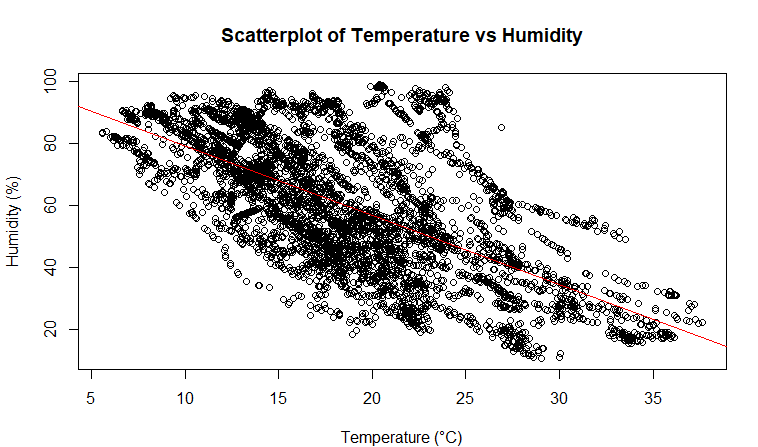
An R² of 0.4155017 indicates that approximately 41.55% of variability in humidity can be explained by linear relationship with temperature. While this is a large portion it also tells that there are other factors affecting humidity since over 58% of the variability is not showing for by temperature alone.

**Scatterplot Interpretation:**

The scatterplot with regression line provides a visual representation of this relationship. The downward slope of line confirms negative relationship. And spread of data points around line shows variability that is not explained by temperature alone, which aligns with the R² value.

A diagram of a temperature

Description automatically generated

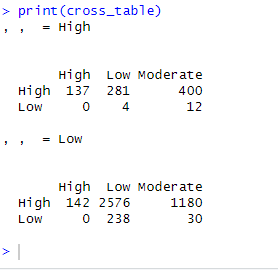


**Explanation of Results:**

The results reveal that while there is a clear association between temperature and humidity and relationship is not perfect and other variables may also play an important role in determining humidity levels. This kind of analysis is essential in climatology and environmental science, where multiple factors usually influence variables of interest. The results would be useful for predicting humidity levels based on temperature but should be used with understanding that other factors may also need to be considered for more accurate predictions.

**1.4) Create three new variables, as defined below:**

**a) Write R program to construct a cross table (cross tabulation) using the above three new variables (WSB, TB and IrrB). Show the obtained cross table.**



**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 IrrB is High? [1 Marks] ii) what is the probability that the TB is high given that the WSB is low? [2 Marks] iii) what is the probability that the IrrB is low given that the TB is moderate and the WSB is low? [2 Marks] iv) Are high TB and low IrrB independent events? Explain. [2 marks] v) Are low TB and high WSB mutually exclusive? Explain**

**i) Probability that the IrrB is High**

The calculated probability is 0.1668 which means there's a 16.68% chance that Irradiance Bin (IrrB) is High in randomly selected record from data.

**ii) Probability that the TB is high given that the WSB is low**

The probability is 0 indicating that there are no records with a high Temperature Bin (TB) when Wind Speed Bin (WSB) is low.

**iii) Probability that the IrrB is low given that the TB is moderate and the WSB is low**

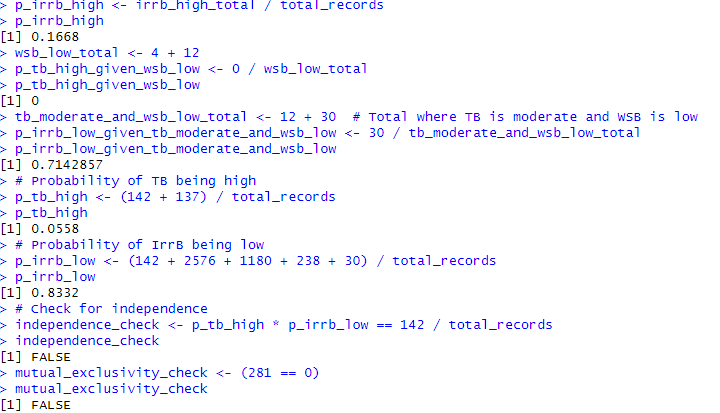
The probability is 0.7142857 meaning that there's a 71.43% chance that Irradiance Bin (IrrB) is Low when Temperature Bin (TB) is moderate and Wind Speed Bin (WSB) is low.

**iv) Are high TB and low IrrB independent events?**

The result is FALSE meaning that high TB and low IrrB are not independent events. The probability of them occurring together is not equal to product of their individual probabilities and which implies there's a relationship between two events.

**v) Are low TB and high WSB mutually exclusive?**

The result is FALSE. Low TB and high WSB are not mutually exclusive events because there are records where both occur simultaneously.



**Q2)**

**2.1)**

**a) How the uncertainty (or variance/error) in an (parameter) estimate is computed using the Bayesian approach?**

In Bayesian approach uncertainty in parameter estimate is expressed through posterior distribution. This distribution incorporates both data through the likelihood and prior beliefs through the prior distribution, The variance or standard error of an estimate can be directly computed from posterior distribution. For given parameter θ uncertainty is variance of posterior distribution P(θ∣data).

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

In frequentist approach uncertainty in parameter estimate is often computed using concept of confidence intervals or standard errors. This is based on sampling distribution of estimator. The standard error of an estimate is standard deviation of its sampling distribution which is often estimated from data.

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

Conjugate priors are useful in Bayesian statistics because they simplify computation of posterior distribution. When prior and likelihood are conjugate pairs posterior distribution is in same family as prior distribution which allows for analytical solutions. This is particularly advantageous for computational efficiency.

**Example of a Conjugate Pair:** The Beta distribution is a conjugate prior for Binomial likelihood. If likelihood of observing data is Binomial and we use a Beta prior for probability of success p and posterior distribution of p will also be Beta distribution.

**2.2) A box contains five red apples and seven green apples. Anne performed three selections (trials) from the box in a sequence. In each selection (trial), if she picks a red apple, she returns three red apples to the box (i.e., she puts two more ‘additional’ red apples to the box. Note that the ‘additional’ red apples are coming from a separate stockpile that Anne has). If she picks a green apple, she returns it back in the box. At the end of three selections, compute the following probabilities. Show all the steps/workings clearly.**

P(R)= 5/12

P(G)= 7/12

**​**

**a) Draw a tree diagram and mark all the probabilities on the branches.**

**b) What is the probability that only the third selection was the green apple?**

This is probability of picking a red apple twice followed by a green apple. Let's calculate it using adjusted probabilities after each red apple is picked:

P(R)×P(R∣ after 1st red)×P(G∣ after 2nd red)

First draw:

P(R)= 12/5

​

Second draw after adding two reds:

P(R)=7/14

​

Third draw after adding two more reds:

P(G)= 7/16

​

The combined probability for R-R-G sequence is

(5/12) × (7/14) × (7/16)

​

**c) What is the probability that Anne has chosen only two green apples from the box over the three trials?**

This is sum of probabilities of sequences RGG, GRG, and GGR

**d) What is the probability that she selects at least one green apple?**

This is complement of probability of selecting no green apples (all red). 1−P(all red apples)

**e) Given that her 3rd selection was a green apple, what is the probability that her 1 st selection was a red apple?**

P(RRG)= (5/12) × (7/14) × (7/16)

P(GRG)= (7/12) × (5/12) × (7/14)

P(RGG)= (5/12) × (7/14) × (7/14)

P(1st pick is R∣3rd pick is G)= P(RRG)+P(RGG) / P(RRG)+P(GRG)+P(RGG) = 0.6164383561643836

**Q3) Frequentist and Bayesian estimations**

**3.1)**

​**a) Show that the joint distribution of dwell times of 𝑁 web pages can be given by the below equation (show the steps clearly to obtain this).**

Given a set of dwell times X={X1, X2, ..., XN} for N web pages and assuming that these times are independently and identically distributed. according to an exponential distribution with rate parameter A probability density function (pdf) for exponential distribution for asingle observation ; is given by:

ƒ(xi|λ) = λe^(-λxi)

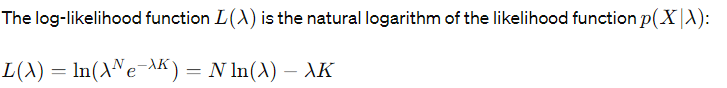
for xi >= 0 and λ > 0.

Since the dwell times are i.i.d., the joint pdf of X given is the product of the individual pdfs:

**A math equations and formulas

Description automatically generated with medium confidence**

**b)Find a simplified expression for the log-likelihood function 𝐿(𝝀𝝀) = 𝐥𝐥𝐥𝐥 (𝑝(𝑿𝑿|𝝀𝝀)).**

****

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

**A white background with black text

Description automatically generated**

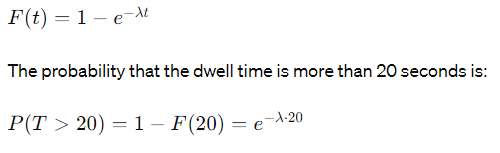
**d)From the past history, the software company has found out that the dwell times of six of the web pages were {10, 12, 15, 20, 15, 20}. Find the Maximum likelihood Estimate of the parameter 𝜆 given this data, i.e., the estimate of the average dwell time?**

Given data {10,12,15,20,15,20}{10,12,15,20,15,20}, we can calculate *λ MLE*​.

The Maximum Likelihood Estimate (MLE) of parameter *λ* given data is approximately 0.06520.0652, which represents rate parameter of exponential distribution for dwell times.

**e) Hence, find the probability that the dwell time of a web page lasts for more than 20 seconds.**

Given MLE of *λ* we can use cumulative distribution function (CDF) of exponential distribution to find probability that dwell time of a web page lasts for more than 20 seconds. The CDF is given by.

****

Using this estimate probability that dwell time of a web page lasts for more than 20 seconds is approximately 0.2710.271, indicating there's about a 27.1% chance a page's dwell time exceeds 20 seconds.

**3.2) SMkt Ltd has now consulted another retail business firm to obtain some additional information about its past experience on the use of web sites for its business. The retail business firm has mentioned that the average dwell time (𝜆) follows a pattern that can be described using an Inverse-Gamma distribution, IGam (a,b), as given below, with hyperparameters 𝑎 = 0.5 and 𝑏 = 10.**

**a)Posterior Distribution**

To derive expression for posterior distribution we'll use Bayes theorem where posterior distribution is proportional to likelihood times prior distribution.

Given:

Prior distribution: Inverse-Gamma IGam(a, b)

Hyperparameters: a = 0.5, b = 10

Likelihood function for given data is typically an exponential or gamma distribution but exact form depends on the data likelihood you are given.

The general form of posterior for Inverse-Gamma distribution when data has a gamma likelihood (which is a conjugate prior for the gamma likelihood) is:

IGam(a', b')

where:

a' = a + n/2

b' = b+Σ(xi)/2

n is the number of observations and î; represents the data points.

The posterior distribution will also be an Inverse-Gamma distribution where a' and b' are updated hyperparameters after observing data.

**b)**

Finding a' and b by Using observed dwell times; = {10, 12, 15, 20, 15, 20}, we calculate a' and b':

n = 6 (number of observations)

E=10+12+15+20+15+20 = 92

a' = 0.5+6/2 = 3.5

b' = 10 +92/2 = 56

The MAP estimate of A for the Inverse-Gamma distribution is the mode which is given by:

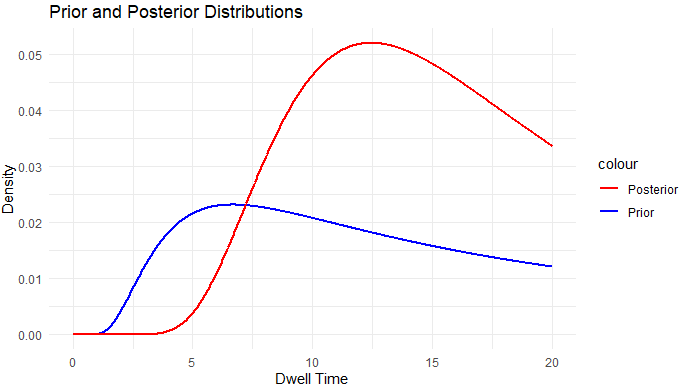
MAP(λ) = b'/a'+1

MAP(λ)= 56 / 3.5+1

MAP(λ)= 56 / 4.5

MAP(λ)= 12.444

**c)**



**Q4) Bayesian inference for Gaussians (unknown mean and known variance)**

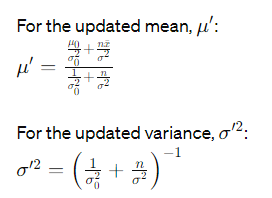
**a) Write an expression for the posterior distribution for 𝜃𝜃 in terms of 𝑛. (Do not derive the formulae)**

• Prior distribution: Normal with mean μ = 150 cm and standard deviation σ0 = 20 cm.

• Likelihood of observing the data: Normal with unknown mean and known standard deviation σ = 10 cm.

• Sample mean = 160 cm and sample size n.

The posterior distribution of 0 given data is also Normal. The mean (u') and variance (σ12) of the posterior distribution can be calculated as follows without derivation:



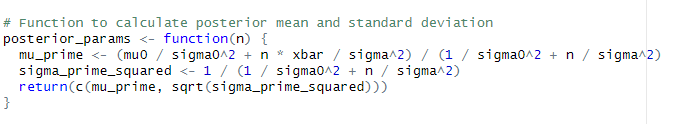
**b) Find the mean and the standard deviation of the posterior distribution for each of the following values of n: i. n=10 ii. n=100 Compare and comment on the posterior variances obtained.**

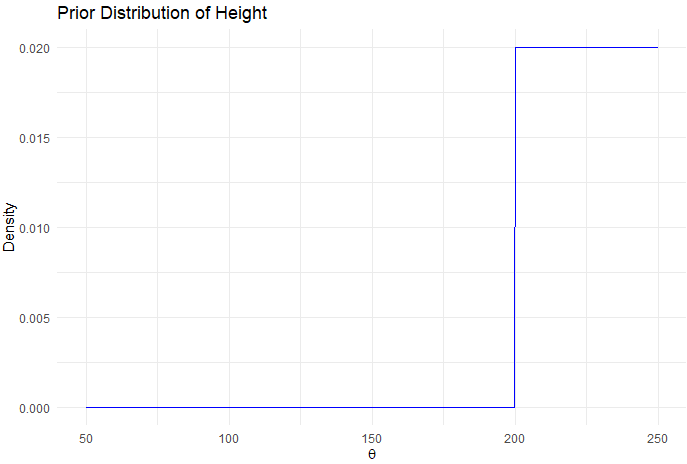
****

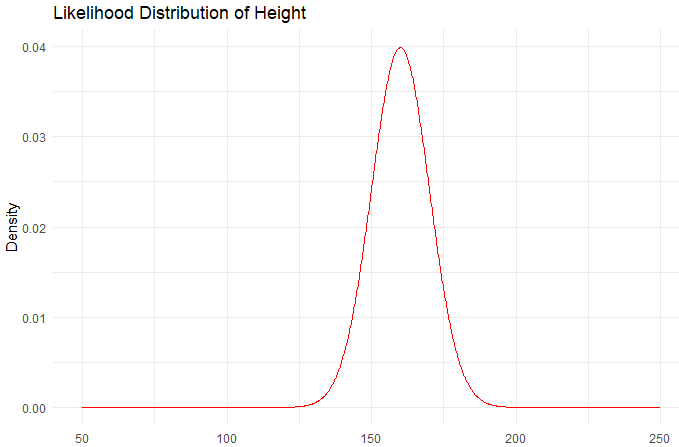
**c) Assume that the prior distribution of the height is changed, and now the prior is distributed as defined below, over the range between 50 and 250 cm:**

* **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 𝜃𝜃 and the posterior standard deviation.**
* **Sketch the obtained prior, likelihood and posterior distributions in separate plots.**

****

****

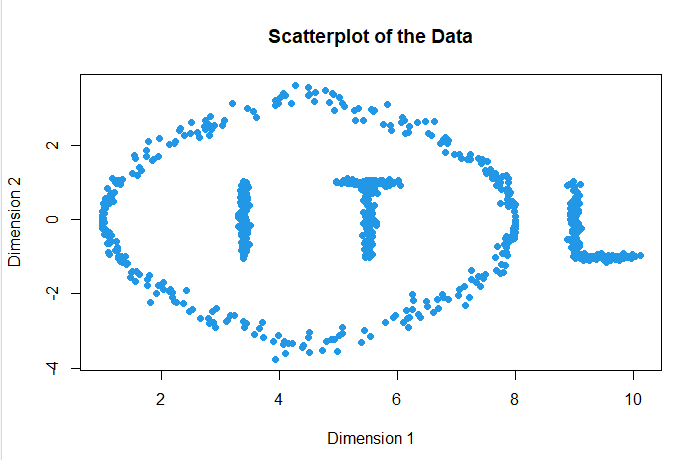
****

****

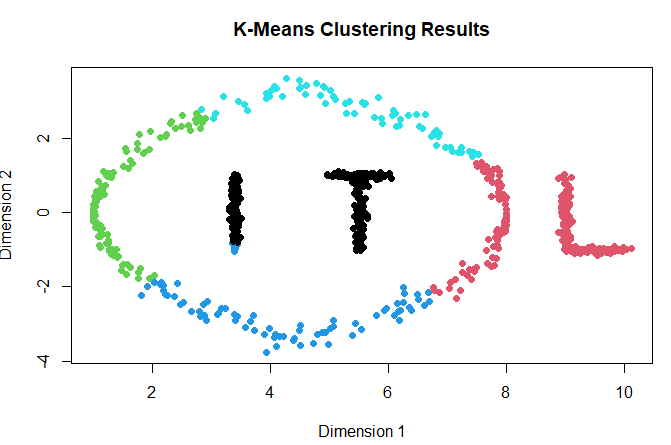
**Q5) Clustering:**

**5.1) K-Means clustering: Use the data file “lettersdata.txt” provided for thisquestion. Load the file “lettersdata.txt” using the following**

**a) Draw a scatterplot for the data and state the number of classes/clusters that can be found in the data (by visual examination of the scatter plot)**

****

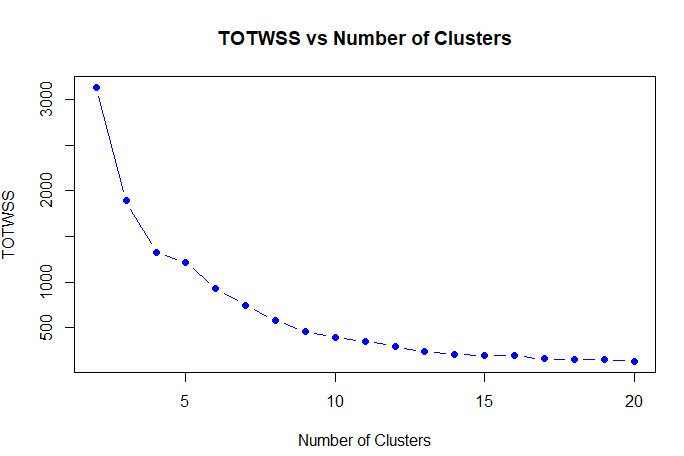
**b) Use the above number of classes as the k value and perform the k-means clustering on that data. Show the results using a scatterplot (show the different clusters with different colours). Comment on the clusters obtained.**

****

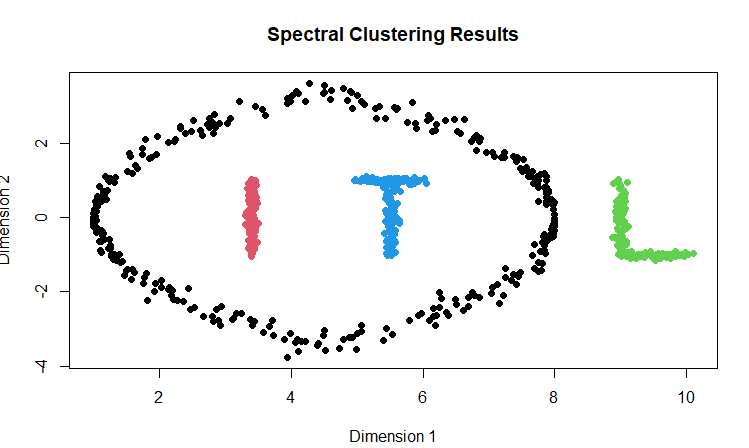
**c)Vary the number of clusters (k value) from 2 to 20 in increments of 1 and perform the k-means clustering for the above data. Record the total within sum of squares (TOTWSS) value for each k, and plot a graph of TOTWSS verses k. Explain how you can use this graph to find the correct number of classes/clusters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

**A close-up of a white background

Description automatically generated**

****

**5.2) Spectral Clustering: Use the same dataset (zz) and perform a spectral clustering (use the number of clusters/centers as 4). Show the results on a scatter plot (with colour coding). Compare these clusters with the clusters obtained using the k-means above and comment on the results. [3 Marks]**

****

**Interpretations:**

Compared to k-means spectral clustering tends to work well with non-convex clusters and can group points based on connectivity or neighborhood relationship rather than just distance from cluster center.

In terms of interpretability clusters formed by spectral clustering might be more meaningful especially if true groups are known to be non-linearly separable. This can be observed in output you provided where clusters appear to capture pattern that may not be linearly separable scenario where spectral clustering excels compared to k-means.

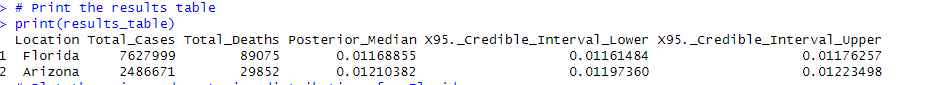
**Question 6:**

**a)**

**1)**Prevalence of Infection Fatality: The unknown prevalence of death in individuals infected with SARS-CoV-2 virus, denoted as ℎ.

Prevalence of Asymptomatic Cases: The unknown prevalence of asymptomatic status in people who were infected by SARS-CoV-2 virus.

**2)**

****

**A graph with red lines

Description automatically generated**

**A graph with red lines

Description automatically generated**

**b)**

**1)What is MCMC (Markov Chain Monte Carlo) Sampling?**

MCMC sampling refers to class of algorithms used to sample from probability distributions based on constructing Markov chain that has desired distribution as its equilibrium distribution. It especially useful in Bayesian statistics where posterior distribution is often complex and multidimensional and making direct sampling challenging.

The essence of MCMC is to build Markov chain with stationary distribution that matches posterior distribution we want to estimate. By simulating chain over many iterations and we can collect samples that are approximately from posterior distribution.

How MCMC Sampling-Based Method Can Be Used for Bayesian Posterior Estimation?

In Bayesian analysis posterior distribution combines prior belief about a parameter with likelihood of observed data. MCMC methods like Metropolis Hastings algorithm and Gibbs sampling enable approximation of posterior distribution when its intractable analytically.

Through MCMC we can draw samples from posterior without having to know full normalization constant of Bayes' theorem which is often the intractable part. Over time chain of samples represents posterior distribution from which we can calculate statistics like mean, median, mode and credible intervals.

**2) Report**

**Problem Being Solved:**

The paper addresses classification of malignant breast tumors in Western Nigeria by using MCMC convergence in Bayesian logistic regression. It aims to profile patients with benign and malignant breast tumors focusing on prognostic factors and accuracy of MCMC convergence in predictions.

Objective:

The main objective is to provide a detailed profile of patients living with both benign and malignant breast tumors. The study not only seeks to identify prognostic factors that could affect outcomes but also aims to improve predictive accuracy of these factors using Bayesian methods. By focusing on MCMC convergence researchers ensure that statistical models they are using provide reliable and robust estimates of parameters in question.

**Methodology:**

To achieve this authors employ Bayesian binary logistic regression. Bayesian methods are particularly suitable for this kind of analysis because they allow incorporation of prior knowledge or lack thereof into the model and they provide probabilistic interpretations of model parameters. This is opposed to classical statistical methods which often offer point estimates without a clear measure of the associated uncertainty.

The logistic regression model is used to predict probability of a tumor being malignant based on various covariates. The Markov Chain Monte Carlo MCMC algorithm is crucial here and it used to sample from posterior distribution of these parameters since an analytical solution is not feasible due to complexity of model.

**Datasets Used:**

The study utilized hospital based records to identify prognostic factors for malignant breast cancer among women in Western Nigeria. The records consisted of 237 patients and 20 variables including socio-demographic information and cancer-specific details.

**Bayesian Method for Modelling/Solving:**

The Bayesian binary logistic regression was applied treating model parameters as random variables and utilizing prior distributions to reflect prior knowledge or assumptions. The MCMC algorithm was used to sample from posterior distribution of model parameters.

The study started with non-informative priors to avoid influencing posterior distribution.

Diagnostics tests such as Heidelberger-Welch were performed to ensure MCMC algorithm's convergence.

Posterior distributions were derived and parameters were estimated providing more accurate reflection of uncertainty in parameter values compared to classical methods.

**Pros and Cons:**

The paper illustrates advantage of Bayesian analysis in providing precise estimates for modeling malignant breast cancer. It can incorporate prior knowledge and offers probabilistic measure of parameter uncertainty. However Bayesian methods can be complex and computationally intensive often requiring specialized software and a deeper understanding of statistical modeling.

**Implications:**

The findings highlight the need for increased awareness and targeted health interventions, particularly among certain demographics such as educated women who were shown to be at higher risk. The paper underscores potential of Bayesian methods to produce more accurate models in medical research and can serve as an advocacy tool for more focused healthcare policies and screening programs.

**Experiment and Results:**

R software was used for classical statistical analysis.

WinBUGS software was used for Bayesian analysis.

After MCMC burn-in posterior means and medians of the coefficients indicated significance for variables such as education level and age.

MCMC diagnostics showed convergence suggesting robustness in Bayesian estimates.

**Conclusions:**

The study concluded that Bayesian analysis provides more precise and robust estimates for modeling malignant breast cancer. Efforts are needed to raise awareness and advocate for reducing prevalence of malignant breast lesions particularly among women.

This paper serves as excellent example of how MCMC sampling and Bayesian methods can be applied in medical research to provide more nuanced insights than traditional statistical methods. The utilization of MCMC allows for a comprehensive exploration of parameter space, taking into account prior knowledge and likelihood of observed data and to arrive at a more informed posterior distribution.

**Reference:**

Ogunsakin RE, Siaka L. Bayesian Inference on Malignant Breast Cancer in Nigeria: A Diagnosis of MCMC Convergence. *Asian Pac J Cancer Prev*. 2017;18(10):2709-2716. Published 2017 Oct 26. doi:10.22034/APJCP.2017.18.10.2709