**Introduction to R Programming**

**Practical File**

**Q1) Write an R script to do the following:**

**a) simulate a sample of 100 random data points from a normal distribution with mean 100 and**

**standard deviation 5 and store the result in a vector.**

**b) visualize the vector created above using different plots.**

**c) test the hypothesis that the mean equals 100.**

**d) use wilcox test to test the hypothesis that mean equals 90.**

**Solution:**

sdata <- rnorm(n = 100, mean = 100, sd = 5)

hist(sdata, breaks = 100, main = "Histogram")

plot(density(sdata), main = "Density Plot")

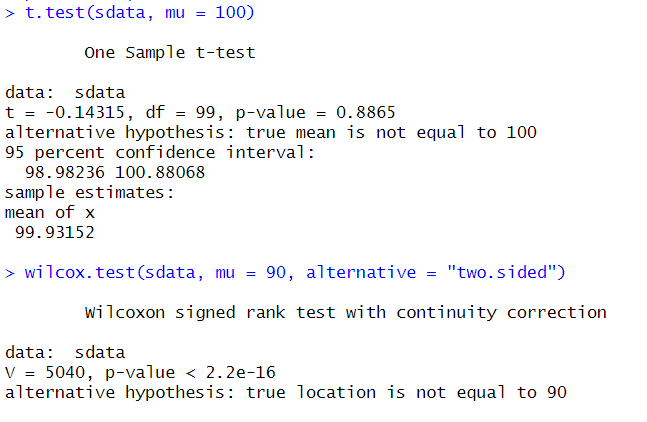
boxplot(sdata, main = "Box Plot")

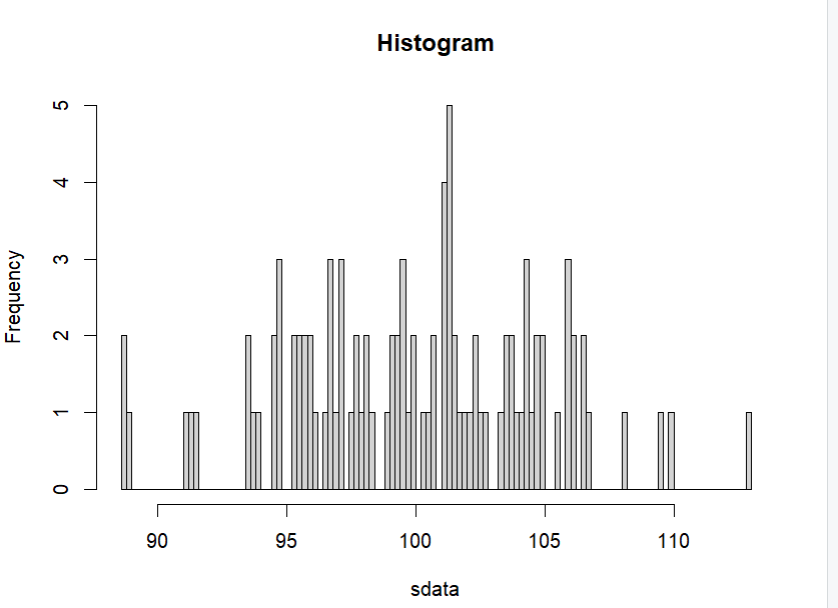
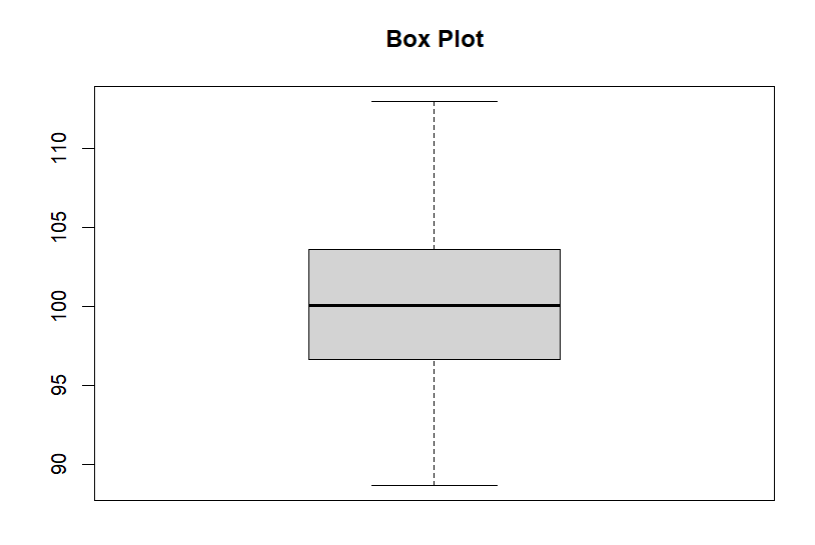
plot(sdata, main = "Scatter Plot")

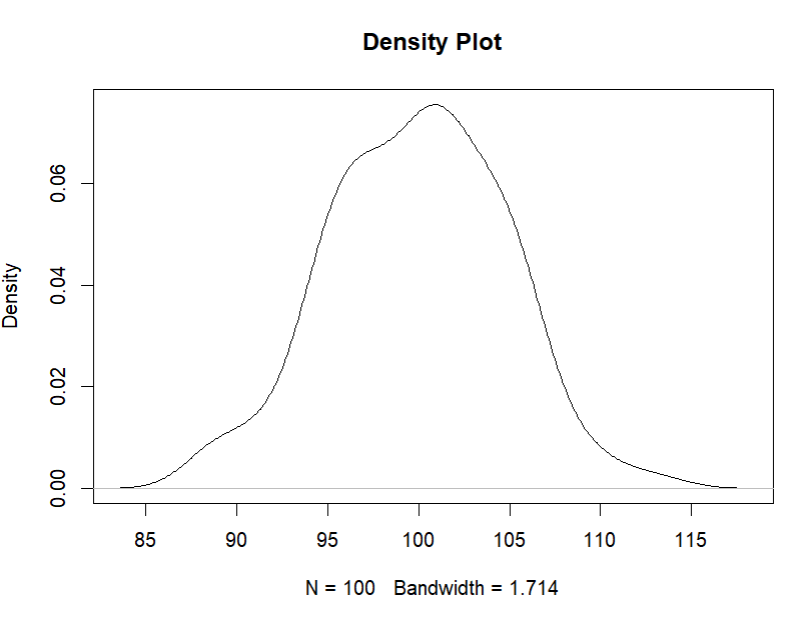
t.test(sdata, mu = 100)

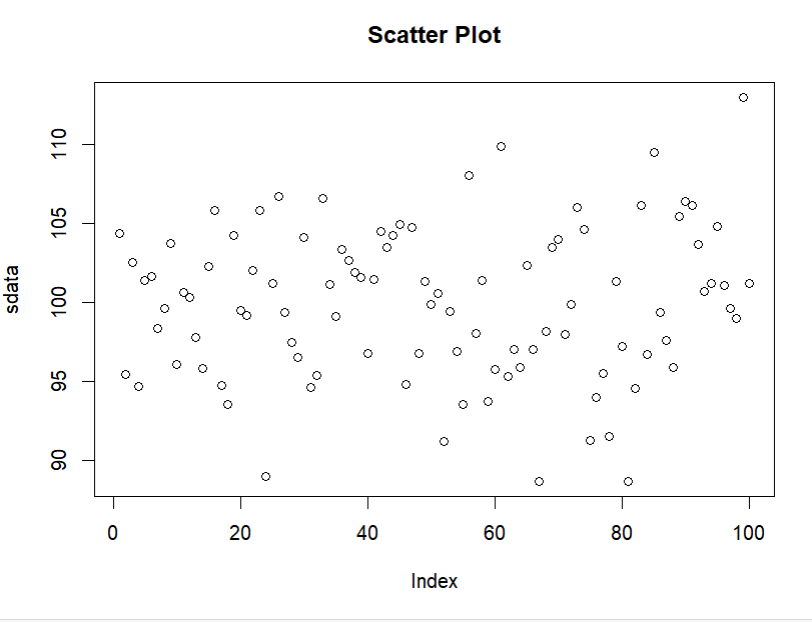
wilcox.test(sdata, mu = 90, alternative = "two.sided")

**Output:**









**Q** **2) Using the Algae data set from package DMwR to complete the following tasks.**

**a) create a graph that you find adequate to show the distribution of the values of algae a6.**

**b) show the distribution of the values of size 3.**

**c) check visually if oPO4 follows a normal distribution.**

**d) produce a graph that allows you to understand how the values of NO3 are distributed**

**e) across the sizes of river.**

**f) using a graph check if the distribution of algae a1 varies with the speed of the river.**

**g) visualize the relationship between the frequencies of algae a1 and a6. Give the**

**h) appropriate graph title, x-axis and y-axis title.**

**Solution:**

#a)

library(DMwR)

# Load the Algae dataset

data(algae)

# Plot the distribution of algae a6

hist(algae$a6, main = "Distribution of Algae a6", xlab = "Algae a6 Values", ylab = "Frequency")

# b) Plot the distribution of size 3

hist(algae$size3, main = "Distribution of Size 3", xlab = "Size 3 Values", ylab = "Frequency")

# c) Plot a histogram and a normal distribution curve

hist(algae$oPO4, main = "Distribution of oPO4", xlab = "oPO4 Values", ylab = "Frequency", prob = TRUE)

curve(dnorm(x, mean = mean(algae$oPO4), sd = sd(algae$oPO4)), add = TRUE, col = "blue")

#d)

# Install and load the 'ggplot2' package for more advanced plotting

install.packages("ggplot2")

library(ggplot2)

# Create a scatter plot to visualize the relationship between NO3 and size1

ggplot(algae, aes(x = size1, y = NO3)) + geom\_point() + labs(title = "Distribution of NO3 across River Sizes", x = "River Size", y = "NO3 Values")

# e) Create a boxplot to compare the distribution of a1 across speed categories

ggplot(algae, aes(x = speed, y = a1)) + geom\_boxplot() + labs(title = "Distribution of Algae a1 across River Speeds", x = "River Speed", y = "Algae a1 Values")

# f) Compute the frequencies of algae a1 and a6

a1\_freq <- table(algae$a1)

a6\_freq <- table(algae$a6)

# Create a bar plot to visualize the frequencies

barplot(rbind(a1\_freq, a6\_freq), beside = TRUE, legend.text = c("Algae a1", "Algae a6"),

main = "Frequency Relationship between Algae a1 and a6", xlab = "Algae", ylab = "Frequency")

**Output:**

**Q3) Read the file Coweeta.CSV and write an R script to do the following:**

**a) count the number of observations per species.**

**b) take a subset of the data including only those species with at least 10 observations.**

**c) make a scatter plot of biomass versus height, with the symbol colour varying by species, and use filled squares for the symbols. Also add a title to the plot, in italics.**

**d) log-transform biomass, and redraw the plot.**

**Solution:**

#a)

# Read the CSV file

data <- read.csv("Coweeta.CSV")

# Count the number of observations per species

species\_count <- table(data$species)

species\_count

# b) Filter the data to include species with at least 10 observations

filtered\_data <- subset(data, species\_count[species] >= 10)

#c) Make a scatter plot of biomass versus height

plot(filtered\_data$height, filtered\_data$biomass, pch = 15, col = filtered\_data$species,

xlab = "Height", ylab = "Biomass", main = expression(italic("Scatter Plot of Biomass vs Height")))

#d)

# Log-transform biomass

filtered\_data$log\_biomass <- log(filtered\_data$biomass)

# Redraw the scatter plot with log-transformed biomass

plot(filtered\_data$height, filtered\_data$log\_biomass, pch = 15, col = filtered\_data$species,

xlab = "Height", ylab = "Log-transformed Biomass", main = expression(italic("Scatter Plot of Log-transformed Biomass vs Height")))

**Output:**

**Q4) The built-in data set mammals contain data on body weight versus brain weight. Write R commands to:**

**a) Find the Pearson and Spearman correlation coefficients. Are they similar?**

**b) Plot the data using the plot command.**

**c) Plot the logarithm (log) of each variable and see if that makes a difference.**

**Solution:**

data(mammals)

cor(mammals$bodywt, mammals$brainwt, method = "pearson")

cor(mammals$bodywt, mammals$brainwt, method = "spearman")

plot(mammals$bodywt, mammals$brainwt,

xlab = "Body weight (kg)",

ylab = "Brain weight (g)",

main = "Body weight versus brain weight")

plot(log(mammals$bodywt), log(mammals$brainwt),

xlab = "Log(body weight)",

ylab = "Log(brain weight)",

main = "Log-transformed body weight versus brain weight")

**Output:**

**Q5) In the library MASS is a dataset UScereal which contains information about popular breakfast cereals. Attach the data set and use different kinds of plots to investigate the following relationships:**

**a) relationship between manufacturer and shelf**

**b) relationship between fat and vitamins**

**c) relationship between fat and shelf**

**d) relationship between carbohydrates and sugars**

**e) relationship between fibre and manufacturer**

**f) relationship between sodium and sugars**

**Solution:**

# Load the MASS library and the UScereal dataset

library(MASS)

data(UScereal)

# Create a plot of shelf vs. manufacturer

plot(UScereal$mfr, UScereal$shelf, xlab = "Manufacturer", ylab = "Shelf")

# Create a plot of fat vs. vitamins

plot(UScereal$vitamins, UScereal$fat,

xlab = "Vitamins", ylab = "Fats")

# Create a boxplot of fat vs. shelf

boxplot(UScereal$fat, UScereal$shelf, xlab = "Shelf", ylab = "Fat")

# Create a scatterplot of carbohydrates vs. sugars

plot(UScereal$carbo, UScereal$sugars, xlab = "Carbohydrates", ylab = "Sugars")

# Create a stacked barplot of fibre vs. manufacturer

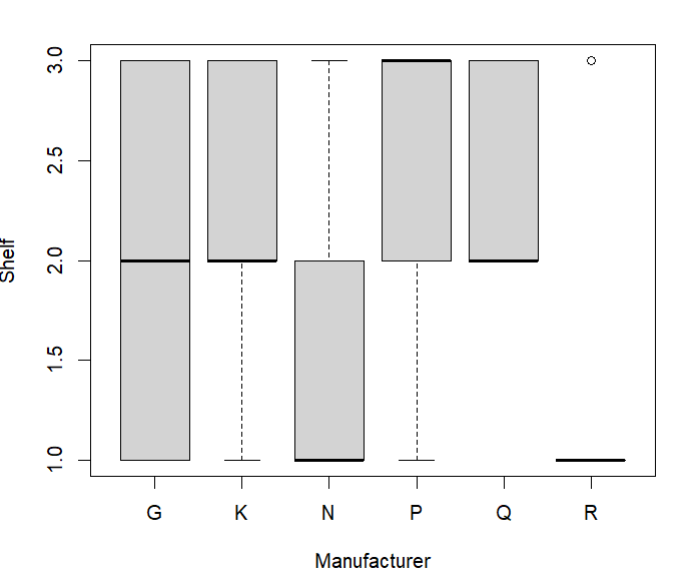
barplot(t(as.matrix(table(UScereal$fibre, UScereal$mfr))), col = rainbow(8),

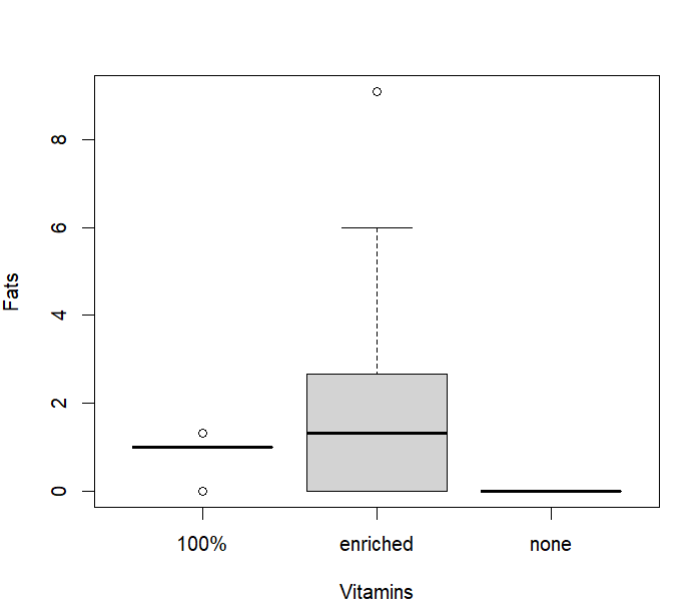
xlab = "Fiber", ylab = "Count")

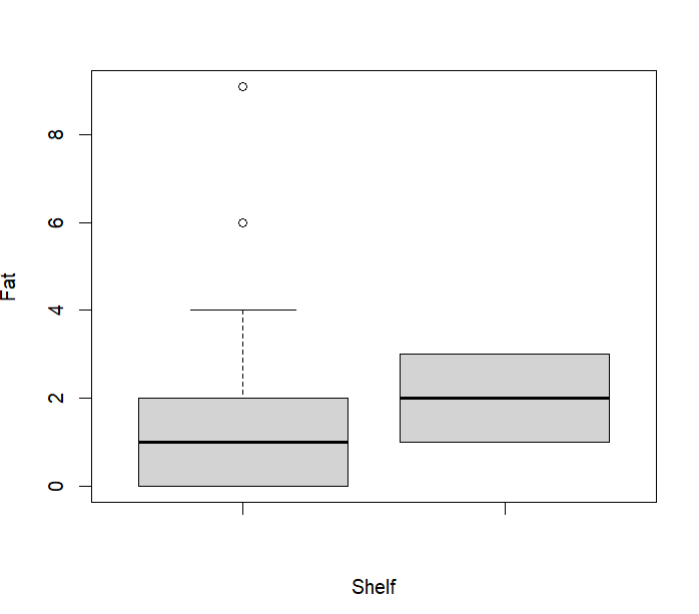
# Create a scatterplot of sodium vs. sugars

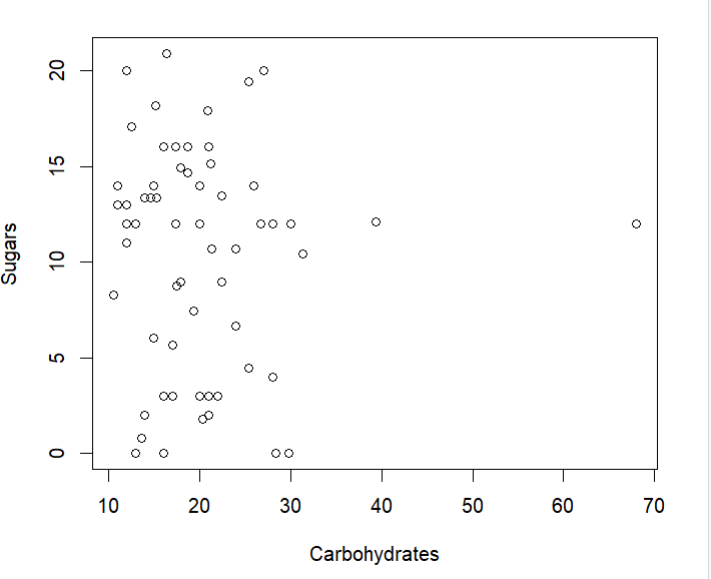
plot(UScereal$sodium, UScereal$sugars, xlab = "Sodium", ylab = "Sugars")

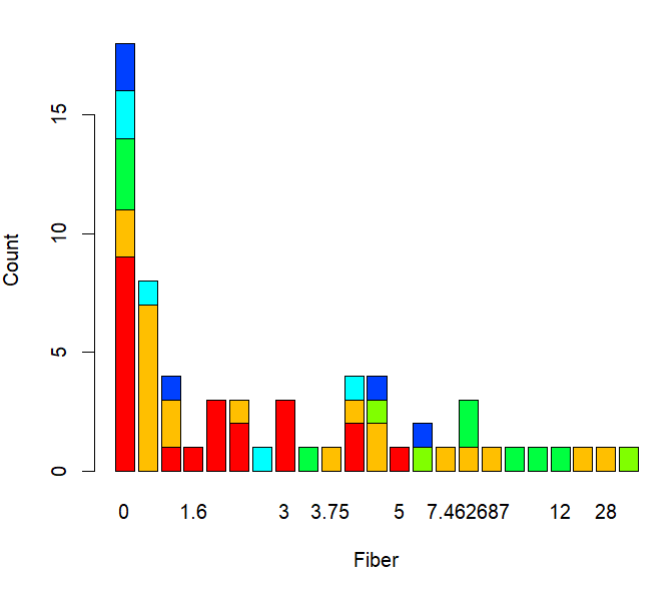
**Output:**

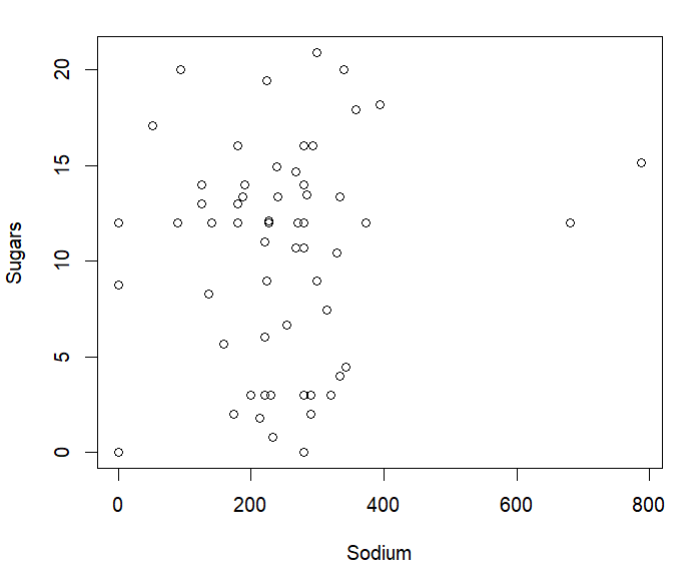












**Q6) Write R script to:**

**a) Do two simulations of a binomial number with n = 100 and p = 0.5.**

**b) Do you get the same results each time? What is different? What is similar?**

**c) Do a simulation of the normal two times. Once with n = 10, μ = 10 and σ = 10, the other with n = 10, μ = 100 and σ = 100.**

**d) How are they different? How are they similar?**

**e) Are both approximately normal?**

**Solutions:**

set.seed(0)

sim1 <- rbinom(n = 1, size = 100, prob = 0.5)

set.seed(1)

sim2 <- rbinom(n = 1, size = 100, prob = 0.5)

print(paste("Simulation 1: ", sim1))

print(paste("Simulation 2: ", sim2))

b) The results will differ each time the code is run. This is because the simulation is random, and the binomial distribution is a discrete distribution, so the values can only take on integer values between 0 and n. The two simulations will have a similar shape, but the exact values will differ.

c)

set.seed(1)

p <- rnorm(n = 10, mean = 10, sd = 10)

set.seed(0)

q <- rnorm(n = 10, mean = 100, sd = 100)

print(paste("Simulation 1: ", p))

print(paste("Simulation 2: ", q))

d) The two simulations will differ in their mean and standard deviation. The first simulation has a mean of 10 and a standard deviation of 10, while the second simulation has a mean of 100 and a standard deviation of 100. However, both simulations will be similar in their shape since they are both normal distributions.

e) Both simulations are approximately normal, as the central limit theorem states that the sum of many independent, identically distributed random variables will be approximately normal. In this case, each simulation is the sum of 10 independent, identically distributed random variables, so they will both be approximately normal.

**Output:**

