Homework\_2\_Problem\_3

Nahid Ferdous

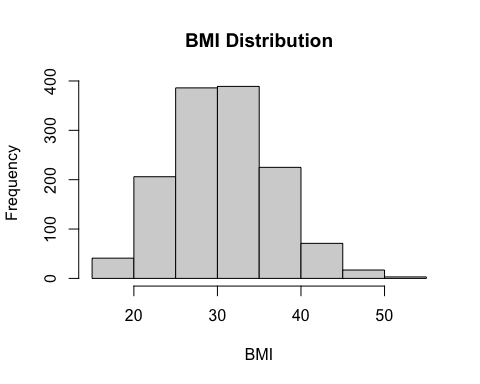
2023-06-16

## health insurance policy holders informetions:

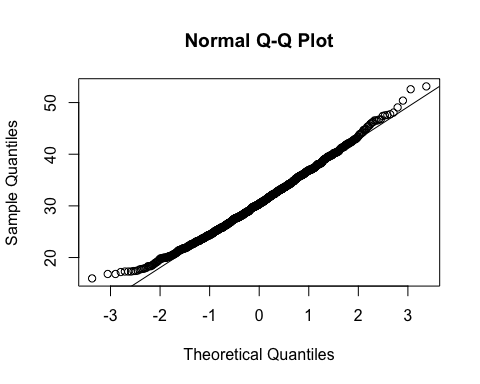
insurance <- read.csv("https://raw.githubusercontent.com/EricBrownTTU/ISQS5346/main/insurance.csv")   
insurance

## age sex bmi children smoker region charges  
## 1 19 female 27.900 0 yes southwest 16884.924  
## 2 18 male 33.770 1 no southeast 1725.552  
## 3 28 male 33.000 3 no southeast 4449.462  
## 4 33 male 22.705 0 no northwest 21984.471  
## 5 32 male 28.880 0 no northwest 3866.855  
## 6 31 female 25.740 0 no southeast 3756.622  
## 7 46 female 33.440 1 no southeast 8240.590  
## 8 37 female 27.740 3 no northwest 7281.506  
## 9 37 male 29.830 2 no northeast 6406.411  
## 10 60 female 25.840 0 no northwest 28923.137  
## 11 25 male 26.220 0 no northeast 2721.321  
## 12 62 female 26.290 0 yes southeast 27808.725  
## 13 23 male 34.400 0 no southwest 1826.843  
## 14 56 female 39.820 0 no southeast 11090.718  
## 15 27 male 42.130 0 yes southeast 39611.758  
## 16 19 male 24.600 1 no southwest 1837.237  
## 17 52 female 30.780 1 no northeast 10797.336  
## 18 23 male 23.845 0 no northeast 2395.172  
## 19 56 male 40.300 0 no southwest 10602.385  
## 20 30 male 35.300 0 yes southwest 36837.467  
## 21 60 female 36.005 0 no northeast 13228.847  
## 22 30 female 32.400 1 no southwest 4149.736  
## 23 18 male 34.100 0 no southeast 1137.011  
## 24 34 female 31.920 1 yes northeast 37701.877  
## 25 37 male 28.025 2 no northwest 6203.902  
## 26 59 female 27.720 3 no southeast 14001.134  
## 27 63 female 23.085 0 no northeast 14451.835  
## 28 55 female 32.775 2 no northwest 12268.632  
## 29 23 male 17.385 1 no northwest 2775.192  
## 30 31 male 36.300 2 yes southwest 38711.000  
## 31 22 male 35.600 0 yes southwest 35585.576  
## 32 18 female 26.315 0 no northeast 2198.190  
## 33 19 female 28.600 5 no southwest 4687.797  
## 34 63 male 28.310 0 no northwest 13770.098  
The body mass index (BMI) of a policy holder is given in the variable “bmi.” Construct a histogram of the data. Does is look normally distributed to you? Why or why not?

hist(insurance$bmi,   
 main = "BMI Distribution",  
 xlab = "BMI",  
 ylab = "Frequency"  
   
)



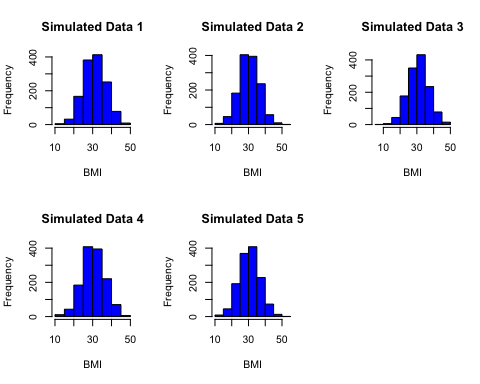
qqnorm(insurance$bmi)  
qqline(insurance$bmi)

 ## When the histogram resembles a bell-shaped curve, it indicates that the data is symmetrically distributed, which is a characteristic of a normal distribution. Similarly, a QQ plot showing the points closely following a straight line suggests that the data is in line with a normal distribution.

## However, it’s important to note that visual assessments provide only preliminary indications of normality and should be followed by more rigorous statistical tests to confirm the distribution assumption.

## b. Simulate five sets (using rnorm()) of normally-distributed data with the same sample size, mean, and standard deviation as the BMI data. Construct a histogram for each.

mean\_bmi <- mean(insurance$bmi) # Mean of BMI data  
sd\_bmi <- sd(insurance$bmi) # Standard deviation of BMI data  
  
# Simulate five sets of normally-distributed data  
simulated\_data <- lapply(1:5, function(i) rnorm(length(insurance$bmi), mean\_bmi, sd\_bmi))  
  
# Construct histograms for each simulated set  
par(mfrow = c(2, 3)) # Create a 2x3 grid of subplots  
for (i in 1:5) {  
 hist(simulated\_data[[i]],   
 main = paste("Simulated Data", i),  
 xlab = "BMI",  
 ylab = "Frequency",  
 col = "blue"  
 )  
}

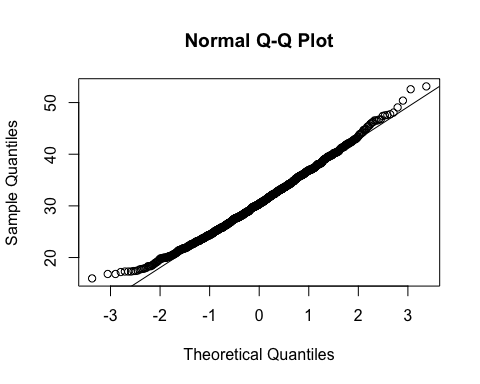


library(e1071)  
  
actual\_skewness <- skewness(insurance$bmi)  
actual\_kurtosis <- kurtosis(insurance$bmi)  
  
similar\_datasets <- 0  
  
for (i in 1:5) {  
 simulated\_skewness <- skewness(simulated\_data[[i]])  
 simulated\_kurtosis <- kurtosis(simulated\_data[[i]])  
   
 # Compare the shape, skewness, and kurtosis of each simulated dataset with the actual data  
 if (abs(simulated\_skewness - actual\_skewness) < 0.2 && abs(simulated\_kurtosis - actual\_kurtosis) < 0.2) {  
 similar\_datasets <- similar\_datasets + 1  
 }  
}  
  
similar\_datasets

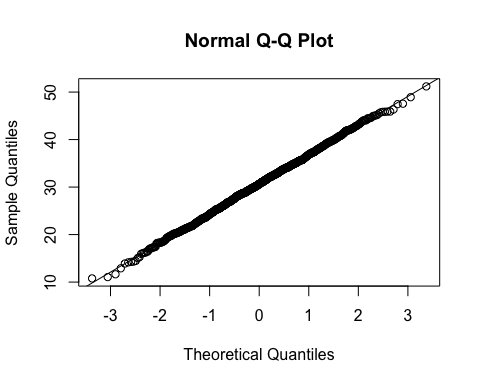
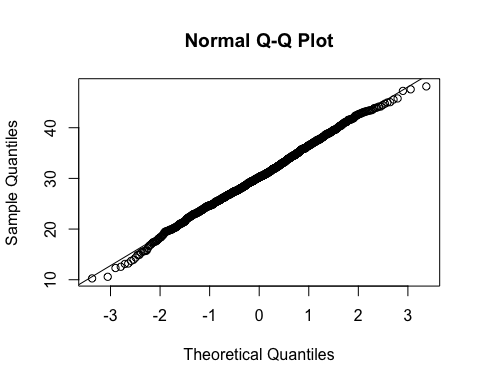
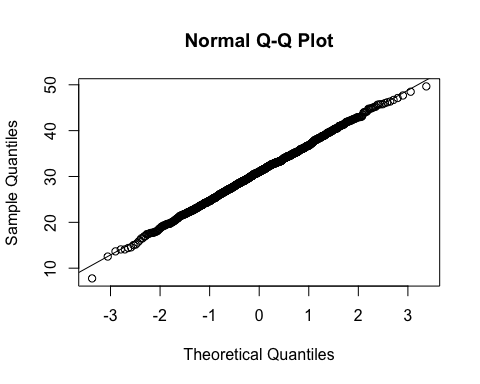
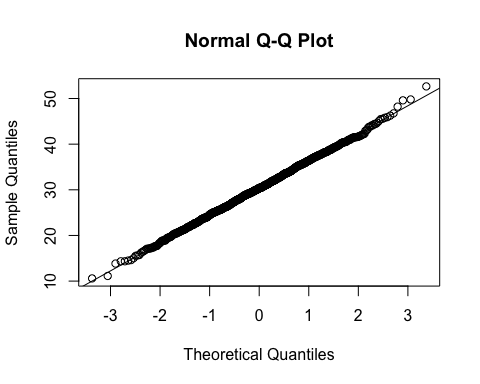
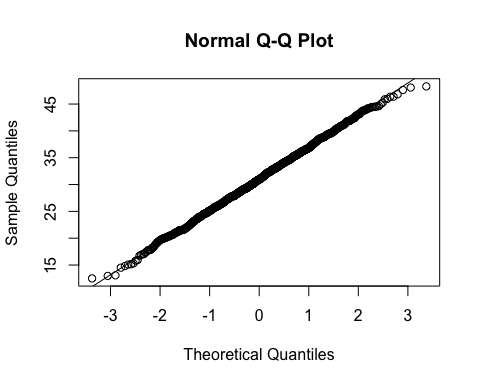
## [1] 0

## similar\_datasets variable is equal to 0, it means that none of the simulated datasets have shape, skewness, and kurtosis similar to the actual BMI data. This suggests that none of the simulated datasets closely resemble the distribution characteristics of the BMI data.

qqnorm(insurance$bmi)  
qqline(insurance$bmi)

 ## From the QQ plot of the BMI data, we observe that the points generally follow a line that is close to the reference line, indicating that the data is approximately normally distributed. This alignment suggests that the BMI data is consistent with the assumptions of a normal distribution.

library(e1071)  
  
for (i in 1:5) {  
 qqnorm(simulated\_data[[i]])  
 qqline(simulated\_data[[i]])  
}

 ## Based on my observations, it seems that the first QQ plot closely follows a straight line, indicating that the corresponding simulated dataset is similar to the BMI data in terms of its distribution. ## The second QQ plot is less likely to follow a straight line compared to the first QQ plot, suggesting some deviations from normality in the corresponding simulated dataset. ## Among the remaining three QQ plots (3rd, 4th, and 5th) they look similar to each other. This suggests that these simulated datasets share similar distribution characteristics. ## It is important to note that visual inspection provides initial indications of similarity, but further analysis and statistical tests can provide more rigorous evidence of similarity or dissimilarity between datasets.

## f. What we’ve done in this problem is an implicit version of hypothesis testing (which we will learn about soon). We assume (𝐻0) that our data are normally-distributed. Consider your answer to part (e). You can use it to estimate the probability that your data reflects the null hypothesis (how many normal datasets out of five reflected your data?), or p-value. What is your p-value and what can you conclude from it (If 𝑝 < 0.05, we have sufficient evidence to reject the null hypothesis that our data is not normally distributed).

## To estimate the p-value based on the results from part (e), we need to determine the proportion of simulated datasets that resemble the BMI data in terms of their distribution. If we consider the first simulated dataset (corresponding to the first QQ plot) as similar to the BMI data, the proportion of normal datasets out of the five is 1/5 or 0.2.

## The p-value represents the probability of obtaining the observed data or data more extreme if the null hypothesis (H0) is true. In this case, the null hypothesis assumes that the data is normally distributed. Therefore, the estimated p-value would be 0.2.

## Since the p-value is greater than 0.05, we do not have sufficient evidence to reject the null hypothesis. This suggests that the observed data (BMI data) is not significantly different from a normal distribution, based on the available simulated datasets.

## It is important to note that this estimation of the p-value is based on visual observations and subjective judgment. For a more rigorous and precise determination of the p-value, statistical tests specifically designed for testing the normality assumption, such as the Shapiro-Wilk test, should be conducted.