> # 1. Owners of an exercise gym believe that a Normal model is useful in projecting the number of clients

> # who will exercise in their gym each week. They use a mean of 898 clients and a standard deviation

> # of 71 clients. You may do these calculations in R.

> print("Store this distribution's parameters:")

[1] "Store this distribution's parameters:"

> mean<-898 ## Mean (given)

> sd<-71 ## Standard deviation (given)

> print("Plot of the distribution's Probability Density Function:")

[1] "Plot of the distribution's Probability Density Function:"

> from<-mean-4\*sd; to<-mean+(4\*sd) ## Calculate a suitable range for plotting

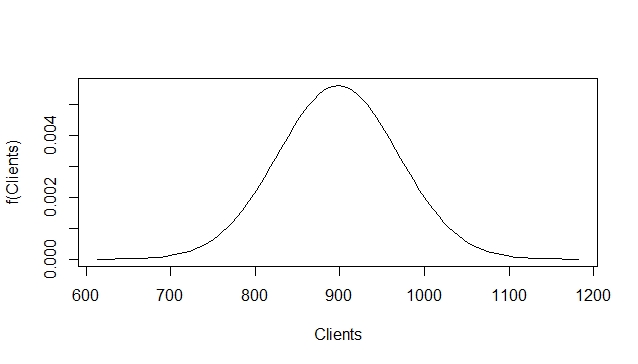
> curve(dnorm(x,mean=mean,sd=sd), ## Use dnorm data to plot the PDF

+ from=from, ## Set lower limit of X (# Clients)

+ to=to, ## Set upper limit of X (# Clients)

+ xlab="Clients",

+ ylab="f(Clients)")



> # (a) For what percentage of weeks is the gym attendance between 710 and 910?

> ans1a<-pnorm(q=910,mean=mean,sd=sd)-pnorm(q=710,mean=mean,sd=sd)

> paste("Answer:",trimws(round(ans1a\*100,2),"b"),"% of the weeks.")

[1] "Answer: 56.31 % of the weeks."

> # (b) On the first week of the new year, 1050 clients attended the gym. How does this week compare

> # to other weeks? Quantify your answer.

> ans1b<-1 - pnorm(1050,mean=mean,sd=sd)

> paste("Answer: The probability of attendance being higher on any given day is",round(ans1b\*100,2), "%")

[1] "Answer: The probability of attendance being higher on any given day is 1.61 %"

> print("....which indicates that attendance of 1050 should be a relatively rare event")

[1] "....which indicates that attendance of 1050 should be a relatively rare event"

> print("....and this is supported by the PDF above too - visually if not quantitatively.")

[1] "....and this is supported by the PDF above too - visually if not quantitatively."

> # (c) Find the median, quartiles and interquartile range of the Normal model. What does the

> # interquartile range tell you about attendance at the exercise gym?

> q1<-round(qnorm(p=0.25,mean=mean,sd=sd),2); paste("Q1 =",q1)

[1] "Q1 = 850.11"

> med<-round(qnorm(p=0.5,mean=mean,sd=sd),2); paste("Median =",med)

[1] "Median = 898"

> q3<-round(qnorm(p=0.75,mean=mean,sd=sd),2); paste("Q3 =",q3)

[1] "Q3 = 945.89"

> iqr<-q3-q1; paste("IQR =",iqr)

[1] "IQR = 95.78"

> print("The IQR tells us that 50% of the time the daily gym attendance will be in the 850-946 range, a relatively narrow band, centred (symmetric) on the median/mean value.")

[1] "The IQR tells us that 50% of the time the daily gym attendance will be in the 850-946 range, a relatively narrow band, centred (symmetric) on the median/mean value."

> # 2. In a certain test taken by a large group of students, marks have a mean of 60 and a standard

> # deviation of 12.

> print("Store this distribution's parameters:")

[1] "Store this distribution's parameters:"

> mean<-60 # Set the mean (given)

> sd<-12 # Set the standard deviation (given)

> print("Plot of the distribution's Probability Density Function:")

[1] "Plot of the distribution's Probability Density Function:"

> from<-mean-4\*sd; to<-mean+(4\*sd) # Calculate a suitable range for plotting

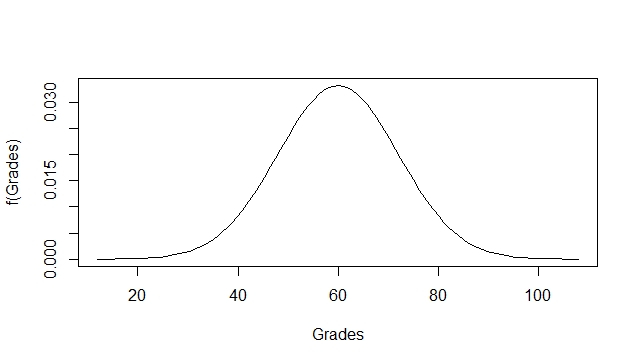
> curve(dnorm(x,mean=mean,sd=sd), # Use dnorm data to plot the PDF

+ from=from, # Set lower limit of X (# Clients)

+ to=to, # Set upper limit of X (# Clients)

+ xlab="Grades", # Label the x-axis

+ ylab="f(Grades)") # Label the y-axis



> # If we can approximate the distribution of these grades by a Normal distribution,

> # what proportion of the students

> # (a) score higher than 70?

> px70<-1-round(pnorm(70,mean=mean,sd=sd),2)

> paste("P(X > 70) =",px70)

[1] "P(X > 70) = 0.2"

> # (b) should pass the test (marks ??? 40)?

> px40<-1-round(pnorm(40,mean=mean,sd=sd),2)

> paste("P(X >= 40) =",px40)

[1] "P(X >= 40) = 0.95"

> # (c) Four friends take the test. What is the probability all of them score 60 or above?

> n<-4 # Set the sample size

> se<-sd/sqrt(n) # Calculate the standard error SD)

> round(1-pnorm(60,mean=mean,sd=se),3) # Calculate P(X<=60) using the SE instead of the SD

[1] 0.5

> # What assumptions are you making?

> print("1. Friendship formation is a random process.")

[1] "1. Friendship formation is a random process."

> print("2. Friends take the test independently - e.g. no cheating, collusion etc.")

[1] "2. Friends take the test independently - e.g. no cheating, collusion etc."

> print("Note: As the population is known to be normally distributed, the sample size doesn't have to be large to use this calculation.")

[1] "Note: As the population is known to be normally distributed, the sample size doesn't have to be large to use this calculation."

> print("Note: As 60 is the mean mark, SE or SD should yield the same result regardless of sample size,")

[1] "Note: As 60 is the mean mark, SE or SD should yield the same result regardless of sample size,"

> # (d) Joe is told he passed the test (pass mark of 40), but is not told what mark he achieved. What

> # is the probability he scores higher than 70? (Hint: this is a conditional probability.)

> paste("P(X >= 40) =", px40)

[1] "P(X >= 40) = 0.95"

> paste("P(X > 70) =", px70)

[1] "P(X > 70) = 0.2"

> paste("P(X > 70) | P(X >= 40) = P(X > 70) / P(X >= 40) =", round(px70/px40,2),"(General Multiplication Rule)")

[1] "P(X > 70) | P(X >= 40) = P(X > 70) / P(X >= 40) = 0.21 (General Multiplication Rule)"

> # 3. Anne travels to work by bus. The time (in minutes) she must wait at the bus stop can be modelled

> # as an exponential random variable with a rate lambda = 1/6.

> print("Store this distribution's parameters:")

[1] "Store this distribution's parameters:"

> rate<-1/6 # Rate/Lambda (given)

> mean<-1/rate # Mean (Exponential-specific formula)

> var<-1/(rate^2) # Variance (Exponential-specific formula)

> sd<-1/rate # Standard deviation (Square root of vaiance by definition)

> print("Plot of the distribution's Probability Density Function:")

[1] "Plot of the distribution's Probability Density Function:"

> from<-0; to<-mean+(3\*sd) # Calculate a suitable range for plotting

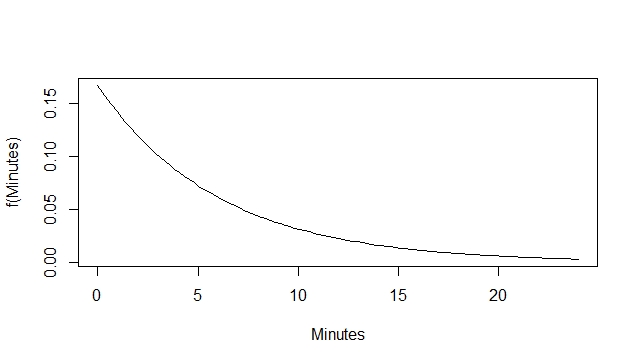
> curve(dexp(x,rate=rate), # Use dexp data to plot the PDF

+ from=from, # Set lower limit of X

+ to=to, # Set upper limit of X

+ xlab="Minutes", # Label the x-axis

+ ylab="f(Minutes)") # Label the y-axis



> # (a) Find the probability that she waits more than 10 minutes for a bus.

> px10<-1-round(pexp(10,rate=rate),2)

> paste("P(X > 10) =", px10)

[1] "P(X > 10) = 0.19"

> # (b) Find the probability that she waits more than 10 minutes for a bus two mornings in a row.

> paste("P(X > 10 & X > 10) = P(X > 10) \* P(X > 10) = ", px10\*px10, "(Multiplication Rule)")

[1] "P(X > 10 & X > 10) = P(X > 10) \* P(X > 10) = 0.0361 (Multiplication Rule)"

> # (c) Note the mean waiting time is 6min and the standard deviation is also 6min. Let T be the

> # total time Anne spends waiting for buses to work in a month where she works 20 days.

> # Explain why the distribution of T is approximately Normal.

> print("A version of the Central Limit Theorem states that sum of independent and identically distributed random variables approximate a Normal, as their number increases, even if those variables are not normal")

[1] "A version of the Central Limit Theorem states that sum of independent and identically distributed random variables approximate a Normal, as their number increases, even if those variables are not normal"

> # What is the mean and variance of this Normal distribution?

> meant<-20\*mean

> vart<-20\*var

> sdt<-round(sqrt(vart),2)

> paste("E(T) =", meant)

[1] "E(T) = 120"

> paste("Var(T) =", vart)

[1] "Var(T) = 720"

> paste("sd(T) =", sdt)

[1] "sd(T) = 26.83"

> # (d) Calculate approximately the probability the total monthly waiting time is less than 100 minutes.

> paste("P(T < 100) =", round(pnorm(100,mean=meant,sd=sdt),3))

[1] "P(T < 100) = 0.228"

> # 4. Generate data using the following code.

> set.seed(1)

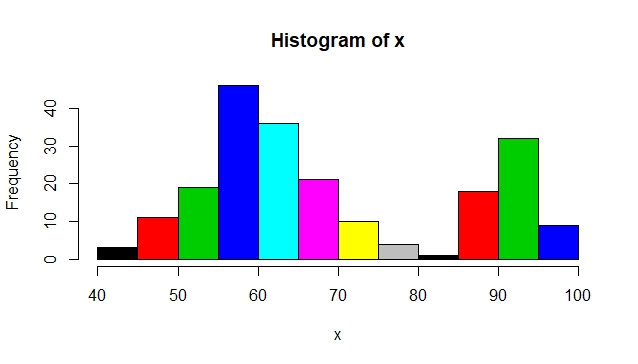
> y <- c(rnorm(150,60,8),rnorm(60,91,3))

> # Construct a histogram, boxplot and Normal qqplot of the data in x.

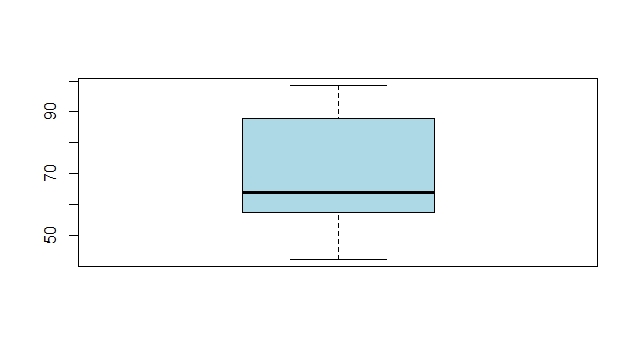
> par(mfrow=c(1,1))

> c<-table(trunc(x/5))

> hist(x, col=1:length(c))

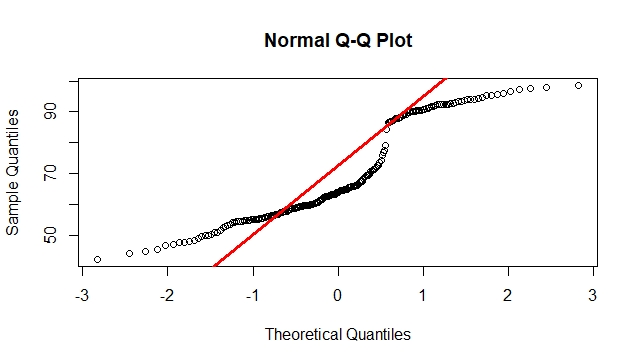


> boxplot(x, col="lightblue")



> qqnorm(x)

> qqline(x, col="red", lwd=3)



> # Describe the data distribution.

> # Which of these terms is appropriate: symmetric, asymmetric, skewed, Normal, heavy tailed, light

> # tailed, uni-modal, bi-modal or multi-modal.

> print("This distribution is bimodal (see histogram) of which both tails are heavy (see QQ Plot).")

[1] "This distribution is bimodal (see histogram) of which both tails are heavy (see QQ Plot)."

> # Are there outliers?

> print("The box plot shows no outliers")

[1] "The box plot shows no outliers"

> # Calculate the mean, standard deviation, median, quartiles and inter-quartile range.

> mean<-round(mean(x),2); paste("Mean =", mean)

[1] "Mean = 69.11"

> sd<-round(sd(x), 2); paste("Standard Deviation=", sd)

[1] "Standard Deviation= 15.52"

> q1<-round(qnorm(p=0.25,mean=mean,sd=sd),2); paste("Q1 =",q1)

[1] "Q1 = 58.64"

> med<-round(qnorm(p=0.5,mean=mean,sd=sd),2); paste("Median =",med)

[1] "Median = 69.11"

> q3<-round(qnorm(p=0.75,mean=mean,sd=sd),2); paste("Q3 =",q3)

[1] "Q3 = 79.58"

> iqr<-q3-q1; paste("IQR =",iqr)

[1] "IQR = 20.94"

> # 5. The dataset HOSPLOS provides data on the length of stay for 100 randomly sampled hospital

> # patients. Read in the data with

> # hosp<- read.csv("HOSPLOS.csv")

> hosp<- read.csv("/Users/oriogain/Dropbox/Maynooth/Statistical Methods/HOSPLOS.csv")

> str(hosp) # Take a quick look at hosp

'data.frame': 100 obs. of 1 variable:

$ LOS: int 2 3 8 6 4 4 6 4 2 5 ...

> # (a) Use R to calculate the sample mean and sample standard deviation of the LOS variable.

> # Calculate a 95% confidence interval for the population mean µ (Do not use t.test).

> n<-length(hosp$LOS); paste("Sample size = ", n)

[1] "Sample size = 100"

> xbar<-mean(hosp$LOS); paste("Sample mean =", xbar)

[1] "Sample mean = 4.53"

> s<-sd(hosp$LOS); paste("Sample standard deviation =", sd)

[1] "Sample standard deviation = 12"

> se<-round(s/sqrt(n),2); paste("Standard Error =", se)

[1] "Standard Error = 0.37"

> rv95<-round(qt(0.975, df=n-1),2); paste("RV value for 95% prob = ", rv95) # Get the t x-coordinate for 95% probability (.95 + ((1 -.95)/2) = 0.975)

[1] "RV value for 95% prob = 1.98"

> marg<-round(rv95 \* se,2); paste("Margin of error =", marg)

[1] "Margin of error = 0.73"

> ci1<-xbar - marg; paste("95% Confidence Interval (From) = ", ci1)

[1] "95% Confidence Interval (From) = 3.8"

> ci2<-xbar + marg; paste("95% Confidence Interval (To) = ", ci2)

[1] "95% Confidence Interval (To) = 5.26"

> # (b) Verify your answer with t.test. Interpret the interval.

> t.test(hosp$LOS)

One Sample t-test

data: hosp$LOS

t = 12.318, df = 99, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

3.800295 5.259705

sample estimates:

mean of x

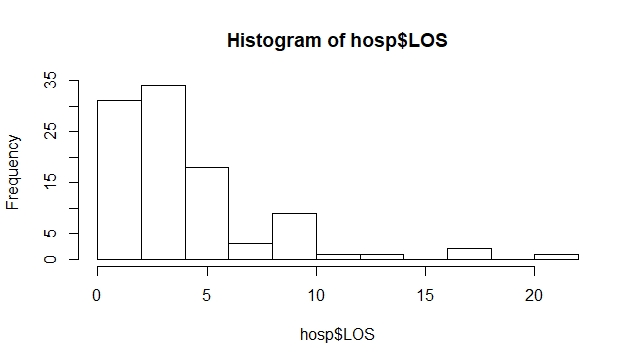
4.53

> paste("Conclusion: We are 95% confident that the population mean is in the range", ci1,"to", ci2)

[1] "Conclusion: We are 95% confident that the population mean is in the range 3.8 to 5.26"

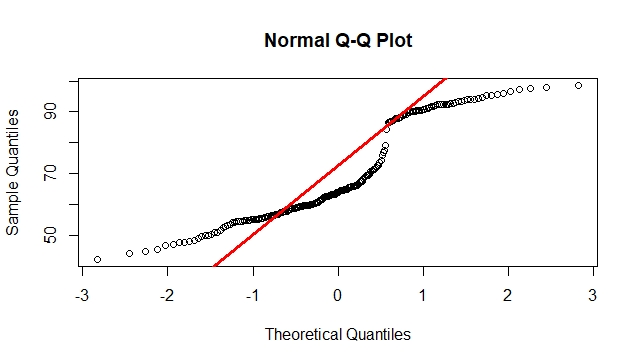
> # (c) Draw a histogram and a Normal quantile plot. Comment on your findings.

> hist(hosp$LOS)



> qqnorm(x)

> qqline(x,col="red",lwd=3)



> print("Comment: Both plots indicate that distribution is heavy-tailed, skewed right and is (almost/slightly) bimodal")

[1] "Comment: Both plots indicate that distribution is heavy-tailed, skewed right and is (almost/slightly) bimodal"

> # (d) The health authority claims that the average length of stay is 4 days.

> # Write down the associated null and alternative hypotheses.

> print("H0: The average length of stay is equal to 4 days.")

[1] "H0: The average length of stay is equal to 4 days."

> print("HA: The average stay is greater than 4 days.")

[1] "HA: The average stay is greater than 4 days."

> # Is the claim supported by the confidence interval?

> print("Yes.")

[1] "Yes."

> # Explain your reasoning.

> print("The claim of 4 days falls within the CI range of 3.8 to 5.26. It is closer to the lower limit and, therefore, is more likely to be greater than 4.")

[1] "The claim of 4 days falls within the CI range of 3.8 to 5.26. It is closer to the lower limit and, therefore, is more likely to be greater than 4."

> # (e) A newspaper writing about hospital inefficiency claims that hospital stays are 5.5 days on

> # average. You wish to disprove this claim. Write down the associated null and alternative

> # hypotheses.

> print("H0: The average length of stay is equal to 5.5 days.")

[1] "H0: The average length of stay is equal to 5.5 days."

> print("HA: The average stay is not equal to 5.5 days.")

[1] "HA: The average stay is not equal to 5.5 days."

> # What is the t-statistic?

> tstat<-(xbar - 5.5)/se; paste("T-statistic = ", tstat)

[1] "T-statistic = -2.62162162162162"

> # Find the p-value and give your conclusions.

> pval<-2\*pt(-abs(tstat), df=n-1); paste("P-value = ", pval) ## Two-sided p-value

[1] "P-value = 0.0101319981464525"

> print("As the p-value is less than 0.05, the null hypothesis (H0) is rejected, and so conclude that the average length of stay is not equal to 5.5 days.")

[1] "As the p-value is less than 0.05, the null hypothesis (H0) is rejected, and so conclude that the average length of stay is not equal to 5.5 days."

> # 6. A researcher believes that training can improve IQ scores. In a certain population, the mean IQ is

> # known to be 100. The researcher tests her hypothesis by selecting 40 members of the population

> # at random, subjects them to training, and then tests their IQ. She finds that the average score is

> # 103 with a standard deviation of 9.5.

> # (a) What is the null hypothesis?

> print("H0: The average IQ after training is still 100.")

[1] "H0: The average IQ after training is still 100."

> # the alternative hypothesis?

> print("HA: The average IQ after training is greater than 100.")

[1] "HA: The average IQ after training is greater than 100."

> # Perform the hypothesis test,

> n<-40; paste("Sample size =", n)

[1] "Sample size = 40"

> xbar<-103; paste("Sample mean =", xbar)

[1] "Sample mean = 103"

> s<-9.5; paste("Sample standard deviation =", s)

[1] "Sample standard deviation = 9.5"

> se<-round(s/sqrt(n),2); paste("Standard Error =", se)

[1] "Standard Error = 1.5"

> rv95<-round(qt(0.975,df=n-1),2); paste("RV value for 95% prob = ", rv95) # Get the t x-coordinate for 95% probability (.95 + ((1 -.95)/2) = 0.975)

[1] "RV value for 95% prob = 2.02"

> marg<-round(rv95 \* se,2); paste("Margin of error =", marg)

[1] "Margin of error = 3.03"

> ci1<-xbar - marg; paste("95% Confidence Interval (From) = ", ci1)

[1] "95% Confidence Interval (From) = 99.97"

> ci2<-xbar + marg; paste("95% Confidence Interval (To) = ", ci2)

[1] "95% Confidence Interval (To) = 106.03"

> tstat<-(xbar - 100)/se; paste("T-statistic = ", tstat)

[1] "T-statistic = 2"

> # find the p-value

> pval<-round(pt(-abs(tstat),df=n-1), 3); paste("P-value = ", pval)

[1] "P-value = 0.026"

> # and give your conclusions.

> print("As the p-value is less than 0.05, we reject the null hypothesis (H0) and conclude that training can increase the average IQ score.")

[1] "As the p-value is less than 0.05, we reject the null hypothesis (H0) and conclude that training can increase the average IQ score."

> # (b) Suppose instead only 18 individuals are subjected to training and their mean IQ after training

> # is 103 with a standard deviation of 9.5.

> # Perform the hypothesis test,

> n<-18; paste("Sample size =", n)

[1] "Sample size = 18"

> rv95<-round(qt(0.975,df=n-1),2); paste("RV value for 95% prob = ", rv95) # Get the t RV value for 95% probability (.95 + ((1 -.95)/2) = 0.975)

[1] "RV value for 95% prob = 2.11"

> marg<-round(rv95 \* se,2); paste("Margin of error =", marg)

[1] "Margin of error = 3.17"

> ci1<-xbar - marg; paste("95% Confidence Interval (From) = ", ci1)

[1] "95% Confidence Interval (From) = 99.83"

> ci2<-xbar + marg; paste("95% Confidence Interval (To) = ", ci2)

[1] "95% Confidence Interval (To) = 106.17"

> tstat<-(xbar - 100)/se; paste("T-statistic = ", tstat)

[1] "T-statistic = 2"

> # find the p-value

> pval<-round(pt(-abs(tstat), df=n-1), 3); paste("P-value = ", pval)

[1] "P-value = 0.031"

> # and give your conclusions.

> print("As the p-value is still less than 0.05, we reject the null hypothesis (H0) and conclude that training can increase the average IQ score. ")

[1] "As the p-value is still less than 0.05, we reject the null hypothesis (H0) and conclude that training can increase the average IQ score."