HW3

Dukki Hong / A98058412

4/28/2019

Problem 1. Write a function chisq.normal.test(x,B = 2000) which takes in a vector x of numerical observations and tests whether this is a sample from a normal distribution based on B bootstrap replicates. Use the hist function to choose the bins automatically.

```r
#install.packages("MASS") #install and load package
require(MASS)
```

```
## Loading required package: MASS
```

```r
options(warn = -1) #ignore the warning
```

```r
chisq.normal.test <- function(x, B = 2000){
  n <- length(x)

  #Calculate for the MLE
  fit <- fitdistr(x, "normal")
  mean <- fit$est[1]
  sd <- fit$est[2]

  #Histogram to select the bins
  h <- hist(x, 30, col="skyblue", freq=F, main="histogram of data", xlab="Data")

  #Calculate the Expected counts
  p <- pnorm(h$breaks, mean = mean, sd = sd)
  E.C <- n*diff(p)
  O.C <- table(cut(x, h$breaks))
  test <- chisq.test(O.C, p = E.C, rescale.p = T)
  D <- as.numeric(test$stat)


  D.B <- numeric(B) #create an empty vector
  count <- 0
  for (i in 1:B){
    Boot.data <- rnorm(n, mean = mean, sd = sd) #bootstrapped data

    fit <- fitdistr(Boot.data, "normal")
    mean.boot <- fit$est[1] #find mean
    sd.boot <- fit$est[2] #find sd

    p <- pnorm(h$breaks, mean = mean.boot, sd = sd.boot)
    E.C <- n*diff(p)
    O.C <- table(cut(Boot.data, h$breaks))
    test <- chisq.test(O.C, p = E.C, rescale.p = T)
    D.B[i] <- as.numeric(test$stat)
    if(D.B[i] >= D){count <- count +1}
  }
  return((count+1)/(B+1))
}
```

Problem 2. Consider the following paper on eating disorders among women in some Portuguese-speaking countries (Brazil, Mozambique, and Portugal):https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0180125. The data are conveniently available for download:https://ndownloader.figshare.com/files/8846953. There are two sheets: One of them provides a description, while the other one contains the data themselves. We focus on the Weight variable. For each country, do the following, organizing the results in a compelling way.

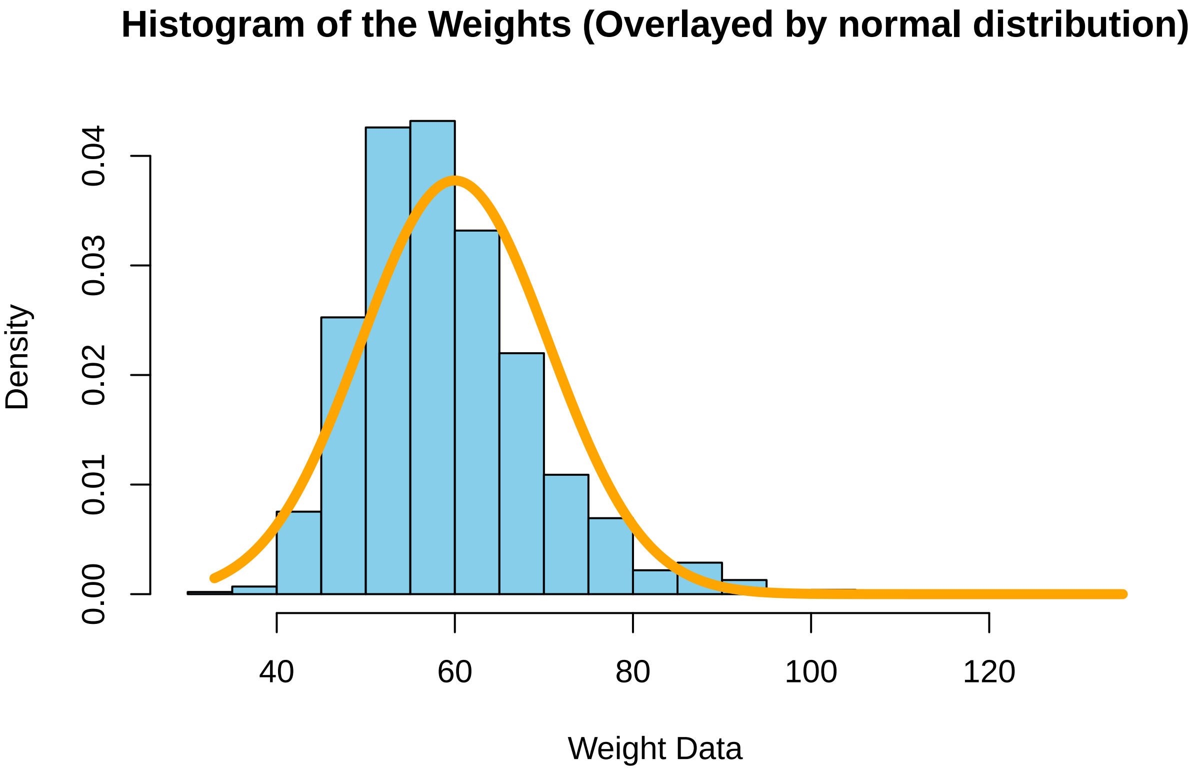
```
library(readxl) #load package
setwd("/Users/okiedukkki/Desktop/UCSD MATH/Math.185/hw/")
data <- read_excel("S1Dataset.xlsx", sheet=2)  #load data

w.data <- data$Weight #select the Weight column
w.data <- as.numeric(na.omit(w.data)) #remove all the rows with NA values
#is.na(w.data) #check if NA exists
#w.data
```

A. Plot a histogram and overlay the normal distribution fitted by maximum likelihood.

```
#Plot histogram
fit <- fitdistr(w.data, "normal")
mean <- fit$estimate[1]
sd <- fit$estimate[2]
cat('The minimum of the dataset is :', summary(w.data)[1],
    '. The maximum of the dataset is :', summary(w.data)[6])
```

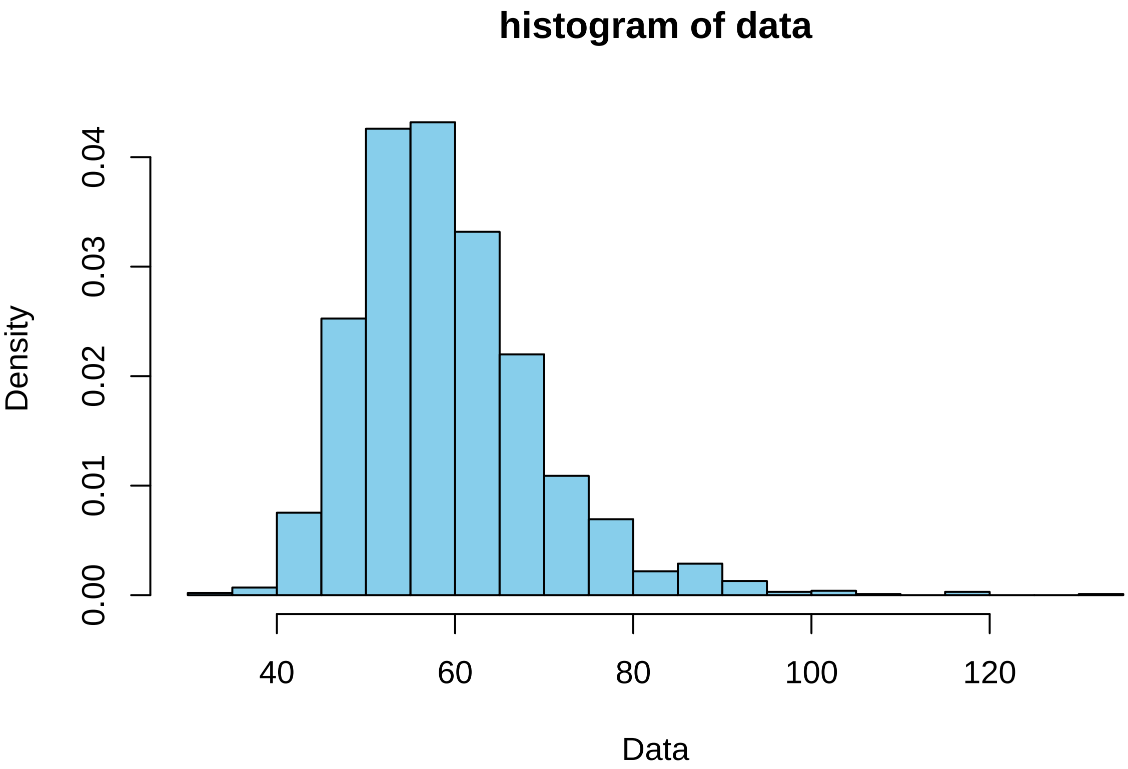
```
## The minimum of the dataset is : 33 . The maximum of the dataset is : 135
```

```
h <- hist(w.data, breaks= 30, col="skyblue", freq=F,
          main="Histogram of the Weights (Overlayed by normal distribution)", xlab="Weight Data")
x <- seq(33, 135, length = 1000)
lines(x, dnorm(x, mean = mean, sd = sd), col="orange", lwd=5)
```



B. Apply the function from Problem 1 to test for normality.

```
suppressWarnings(chisq.normal.test(w.data, B=2000))
```



```
## [1] 0.0004997501
```

Problem 3. With the same dataset as in Problem 2, and again focusing on the Weight variable, compare the countries pairwise using at least two different testing procedures. (Note that there are 3 countries, and therefore 3C(choose)2 = 3 pairs in total. Every time, state the null hypothesis you are testing, and name the testing procedure you are using to do that, specifying how the p-value is computed. Offer some brief comments.

```r
#Two different test procedures: Wilcoxen Rank Sum Test and Kolmogorov Smirnov Test

#Subsetting data by countries
Brazil <- data[which(data$Country == 1),]
Portugal <- data[which(data$Country == 2),]
Mozambique <- data[which(data$Country == 3),]

#Subset countries by their weights and remove NA's
Brazil <- Brazil$Weight
Brazil <- as.numeric(na.omit(Brazil))

Portugal <- Portugal$Weight
Portugal <- as.numeric(na.omit(Portugal))

Mozambique <- Mozambique$Weight
Mozambique <- as.numeric(na.omit(Mozambique))
```

Permutation Test: type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points.

H_0: X ~ Y vs H_1: X stochastically dominates Y

```r
#Permutation Test:
#H_0: X ~ Y vs H_1: X Stochastically dominates Y

permutation.test <- function(x,y,B){
  n_x <- length(x) #sample size for X
  n_y <- length(y) #sample size for Y
  test <- mean(x) - mean(y) #calculate for the difference in the two sample means
  z <- c(x,y)
  n <- length(z) #sum of the two sample sizes, X and Y

  count <- 0
  for( i in 1:B){
    temp <- sample(z, n, replace=F) #pick n samples without replacement from the total samples
    xtemp <- temp[1:n_x] #choose chosen samples from the 1st to the n_xth
    ytemp <- temp[(n_x+1):n] #choose chosen samples from the n_x+1st to the nth
    mean.temp <- mean(xtemp) - mean(ytemp) #take the difference of the two means
    if(mean.temp >= test){count <- count + 1} #check to see if the sample difference of mean is >= obse
  }
  return((count+1)/(B+1))
}

cat("The p-value is :", permutation.test(Brazil, Portugal, 10000), "< 0.05 = alpha.")
```

```
## The p-value is : 0.00189981 < 0.05 = alpha.
```

So we reject the null hypothesis. Hence X stochastically dominates Y. That is, weight data of Brazil stochastically dominates over the weight data of Portugal.

Wilcoxen Rank Sum Test

H_0: Distribution of the test statistic is same vs. H_1: Distribution of the test statistic is not same.

Given R_i, which is the rank of X_i, the Wilcoxen Rank Sum test rejects for a large value of the following: V = sum from i=1 to m of R_i.

```r
#Wilcoxen Rank Sum Test
wilcox.test(Brazil, Portugal, alternative="greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Brazil and Portugal
## W = 401190, p-value = 0.009522
## alternative hypothesis: true location shift is greater than 0
```

```r
cat("p-value is :", wilcox.test(Brazil, Portugal, alternative="greater")$p.value, "<0.05 = alpha.")
```

```
## p-value is : 0.009522257 <0.05 = alpha.
```

So we reject the null hypothesis. Hence the weight data of Brazil and the weight data of Portual does not have equal distribution.