**Harrison Jansma**

**Stats for Data Science**

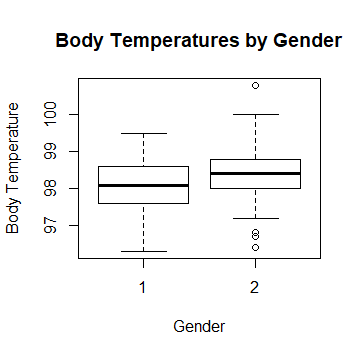
**Mini Project 5**

**4/18/2019**

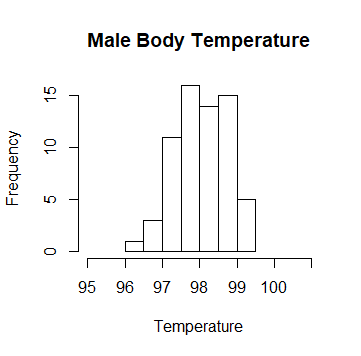
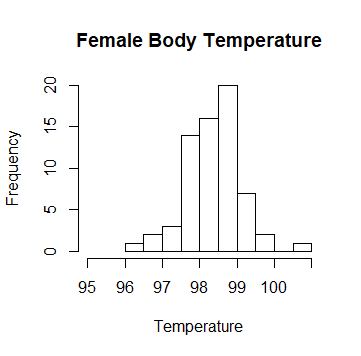
**1. Consider the data stored in bodytemp-heartrate.csv on eLearning, containing measurements of body temperature and heart rate for 65 male (gender = 1) and 65 female (gender = 2) subjects.**

**(a) Do males and females differ in mean body temperature? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.**

Looking at the dataset of 130 observations of body temperatures for men and women, we can see that the mean temperature of the 65 observed men and 65 observed women is 98.10 and 98.39 respectively. We can see this difference in the following figure.



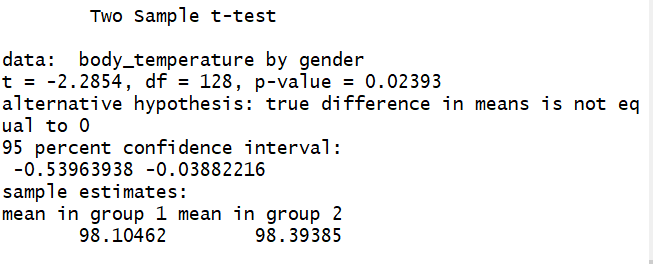
The below histograms show that both distributions of body temperature for men and women are normally distributed. We also see a few potential outliers in the female dataset. In particular, the outlier that falls close to one hundred may be affecting our estimate of the mean. We will see in the next section, if this perceived difference between men and women is due to chance.



It is a reasonable assumption that both distributions random, independent, and normally distributed with equal variances, we can perform a two-sample t-test to determine whether the difference between the two means is statistically significant (at a 95% confidence level).

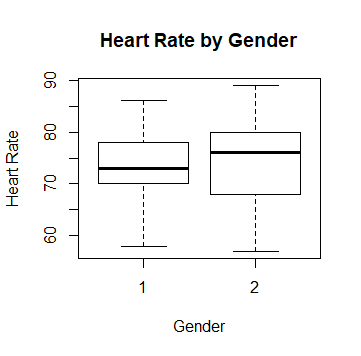
The null hypothesis for this test being the two means are equal, and the alternative is that they are not equal (two-sided).

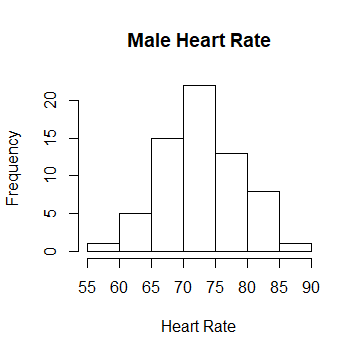
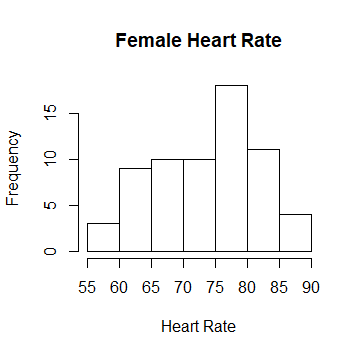
Upon performing the t-test we get the below results. With a p-value of 0.024 the difference between body temperatures of men and women is statistically significant. We reject the null hypothesis and conclude that men and women do have different body temperatures.



**(b) Do males and females differ in mean heart rate? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.**

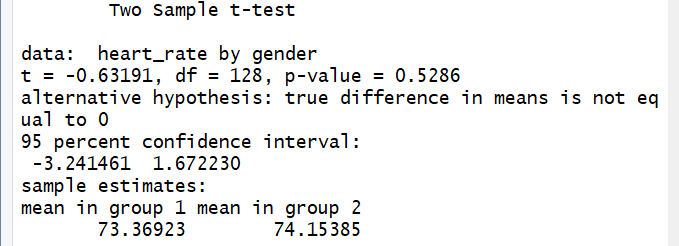
From the same dataset we can also compare the heart rates of 130 men and women. From this data we see that the mean heart rate for men and women is 73.37 and 74.15 respectively. We can see this difference in the below barplot.





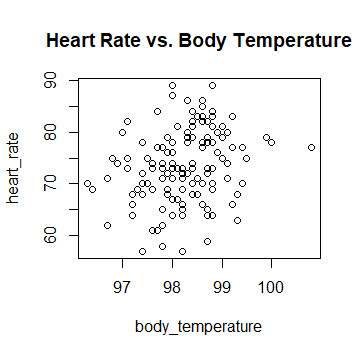
It is a reasonable assumption that both distributions random, independent, and normally distributed with equal variances, we can perform a two-sample t-test to determine whether the difference between the two means is statistically significant (at a 95% confidence level). The null being that the two means are equal, and the alternative is that they are not equal (two-sided).

Upon computing the results of the test, we reach the following conclusion. Due to the high p-value (0.5286) we fail to reject the null hypothesis. There isn’t enough evidence to conclude that there is a difference between the heart rates of men and women.



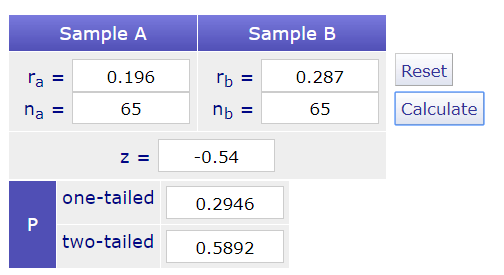
**(c) Is there a linear relationship between body temperature and heart rate? Does this relationship depend on gender? Answer these questions by performing an appropriate analysis of the data, including an exploratory analysis.**

From our dataset I compute the linear correlation coefficient between heart rate and body temperature as 0.254, which notes a weak positive correlation. Upon visualizing the dataset in the below plot I fail to see any clear upward trend in the data.



But perhaps this relationship can be seen more clearly if we separate the sample by gender. The correlation between heart rate and body temperature for men and women are 0.196 and 0.287 respectively.

To test whether this difference in correlation between heat rate and body temperature for men and women is statistically significant, we will need to convert the correlation values into z-scores with Fischer’s r-to-z transformation. I used an online calculator at (<http://vassarstats.net/rdiff.html>) and found that the p-value for the difference between the two correlation coefficients is 0.589. This shows that there is not enough evidence to conclude that the correlation coefficients for heart rate and body temperature differ by gender.



**2. The goal of this exercise to see how large n should be for the large-sample and the (parametric) bootstrap percentile method confidence intervals for the mean of an exponential population to be accurate. To be specific, let X1, . . . , Xn represent a random sample from an exponential (λ) distribution. Note that this distribution is skewed and its mean is µ = 1/λ. We can construct two confidence intervals for µ — one the large-sample z-interval (interval 1) and the other a (parametric) bootstrap percentile method interval (interval 2). We would like to investigate their accuracy, i.e., how close their estimated coverage probabilities are to the assumed nominal level of confidence, for various combinations of (n, λ). This investigation will focus on 1 − α = 0.95, λ = 0.01, 0.1, 1, 10 and n = 5, 10, 30, 100. Thus, we have a total of 4 ∗ 4 = 16 combinations of (n, λ) to investigate.**

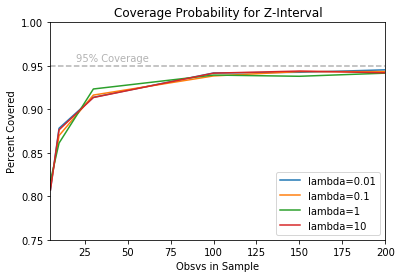
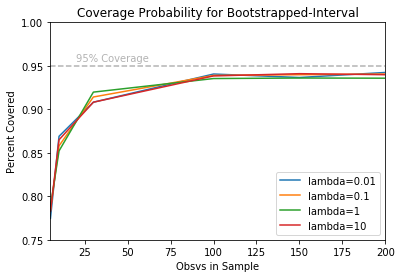
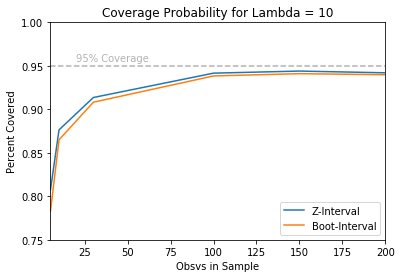
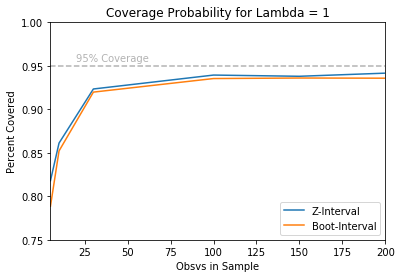
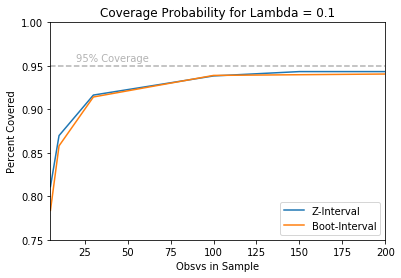
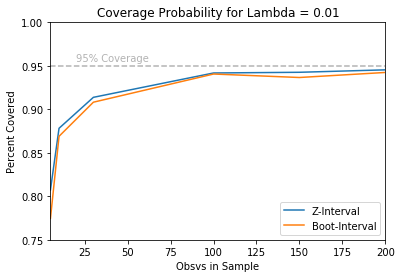
**(a) For a given setting, compute Monte Carlo estimates of coverage probabilities of the two intervals by simulating appropriate data, using them to construct the two confidence intervals, and repeating the process 5000 times.**

We first simulate 5 observations drawn from an exponential (λ = 0.01) distribution. We will then utilize R to generate 5000 large-sample z-confidence intervals and 5000 bootstrapped percentile confidence intervals from this sample. All intervals will be at the 95% confidence level. The bootstrap samples will contain 500 resamplings of the simulated data.

By performing the above computation, we estimate coverage probability for z-intervals and bootstrapped percentile intervals at 0.8132 and 0.7820 respectively.

**(b) Repeat (a) for the remaining combinations of (n, λ). Present an appropriate summary of the results.**

To better answer the question in part c, I simulated two additional coverage percentages at n=150 and n=200. The results of my analysis are displayed below.



**(c) Interpret all the results. Be sure to answer the following questions: In case of the large-sample interval, how large n is needed for the interval to be accurate? Likewise, in case of the bootstrap interval, how large n is needed for the interval to be accurate? Do these answers depend on λ? Can we say that one method is more accurate than the other? Which interval would you recommend? Provide justification for all your conclusions.**

It seems that both methods of interval estimation have reasonable accuracy if sample size is greater than one hundred. We see in the first four plots above that the coverage percentage of both intervals have diminishing returns as n increases. That “elbow point” where coverage stops decreasing lies somewhere in between n=30 and n=100. However, to stay on the safe side it is best to get more data if resources allow.

For the most part, this answer does not depend on lambda. We see in the overlaid plots above that both bootstrapped and z-intervals tend to have consistent coverage probabilities for each value of n despite changes in lambda.

Regardless of the value of n or lambda, the z-interval tended to be more accurate than the bootstrapped percentile interval. However, this difference might be influenced by the number of resamplings in each bootstrap sample. During this analysis I increased bootstrap sample size from 100 to 500 and saw substantial increases in coverage probability for every combination of n and lambda.

I would recommend the Z-interval for its accuracy and computational efficiency if the desired estimated parameter suits its assumptions. It was easy in this example since we estimate means (known to have normal sampling dist. w/ large n). However, for more complex estimations where standard error might not have a simple formula, bootstrap percentile intervals tend to be your only option.

**(d) Do your conclusions in (c) depend on the specific values of λ that were fixed in advance? Explain.**

No, the coverage probability of each interval was consistent despite changes in lambda. There is no reasonable trend that would justify that larger lambda results in larger/smaller coverage.

library(boot)

library(ggplot2)

#############################

#MINI PROJECT 5

#QUESTION 1

#BODY TEMPERATURE

MyData <- read.csv(file="C:/Users/Harrison/Desktop/bodytemp-heartrate.csv", header=TRUE, sep=",")

attach(MyData)

#seperate the data

men = MyData[1:65,]

women = MyData[66:130,]

mean(men$body\_temperature)

mean(women$body\_temperature)

#visuals

boxplot(body\_temperature~gender,data=MyData, main="Body Temperatures by Gender",

xlab="Gender", ylab="Body Temperature")

hist(men$body\_temperature, main="Male Body Temperature",xlab="Temperature",xlim=c(95,101))

hist(women$body\_temperature, main = "Female Body Temperature", xlab="Temperature",xlim=c(95,101))

t.test(body\_temperature~gender,data = MyData,var.equal=TRUE)

###########

#HEART RATE

mean(men$heart\_rate)

mean(women$heart\_rate)

#visuals

boxplot(heart\_rate~gender,data=MyData, main="Heart Rate by Gender",

xlab="Gender", ylab="Heart Rate")

hist(men$heart\_rate, main="Male Heart Rate",xlab="Heart Rate")

hist(women$heart\_rate, main = "Female Heart Rate", xlab="Heart Rate",xlim=c(55,90))

t.test(heart\_rate~gender,data = MyData,var.equal=TRUE)

###########

#HEART RATE BODY TEMP CORR

cor(heart\_rate,body\_temperature)

plot(heart\_rate~body\_temperature,MyData, main="Heart Rate vs. Body Temperature")

cor(men$heart\_rate,men$body\_temperature)

cor(women$heart\_rate,women$body\_temperature)

#########################################

#QUESTION 3

ns=c(5, 10, 30, 100,150,200)

lambdas=c(0.01,0.1,1,10)

#Get Conf Intervals for n/lambda

getCIs =function(n=5,lambda=0.01){

#Generates a sample of size n from exp(lambda)

# then computes Z-interval and bootstrap CIs

samp = rexp(n=n, rate=lambda)

serror = sd(samp)

Zci = c(mean(samp)-1.96\*(serror/(sqrt(n))),mean(samp)+1.96\*(serror/(sqrt(n))))

meany = function(data, indices){

return(mean(data[indices]))

}

res = boot(data=samp, statistic = meany, R=500)

Bci = quantile(res$t, probs = c(0.025,0.975))

return(c(Zci,Bci))

}

#Support function. Returns TRUE if x is in [left,right]

between=function(x,left,right){

if(x<left){return(FALSE)}

if(x>right){return(FALSE)}

else{return(TRUE)}

}

#Returns percent of 5000 intervals that covered true param

Coverage\_Percent = function(n,lambda){

coverage\_percent = c(0,0)

for(i in 1:5000){

res = getCIs(n=n,lambda = lambda)

if(between(1/lambda,res[1],res[2])){

coverage\_percent[1]=1+coverage\_percent[1]

}

if(between(1/lambda,res[3],res[4])){

coverage\_percent[2]=1+coverage\_percent[2]

}

}

return(coverage\_percent/5000)

}

#Placeholder to insert probs into

z\_est = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24)

b\_est = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24)

#Loop over all values of n and lambda

for(i in 1:6){

for(j in 1:4){

print(ns[i])

print(lambdas[j])

print("Z-Interval Coverage Bootstrap Coverage")

cv= Coverage\_Percent(n[i],lambda[j])

print(cv)

z\_est[(4\*(i-1))+j]=cv[1]

b\_est[(4\*(i-1))+j]=cv[2]

}

}

n = rep(ns,times=6)

lambda =rep(lambdas,times=6)

counts = data.frame(lambda,z\_est, b\_est)

#Export data to generate plots w/ Matplotlib

write.csv(counts, file = "C:/Users/Harrison/Desktop/counts.csv")