R code (Section 2.2)

# Using R to plot an empirical distribution function for a set of data:

# Example from class:

female.heights <- c(69.3,66.3,62.6,62.9,67.4)

# Three different styles of edf plot:

plot(ecdf(female.heights))

plot(ecdf(female.heights), verticals=TRUE)

plot(ecdf(female.heights), verticals=TRUE, do.points=FALSE)

# An example with a simulated data set with LOTS of observations:

simul.height.data <- rnorm(n=500, mean=65, sd=2.5)

plot(ecdf(simul.height.data))

plot(ecdf(simul.height.data), verticals=TRUE)

plot(ecdf(simul.height.data), verticals=TRUE, do.points=FALSE)

# Note how close the estimator of the cdf comes to the true cdf when the sample size is very large:

gridpts <- seq(55,75,by=0.1)

lines(gridpts,pnorm(gridpts,mean=65,sd=2.5),col='red') # true cdf superimposed in red

### Bootstrap example:

# function to calculate the 85th percentile of a sample vector:

perc85 <- function(input.vec){

output <- quantile(input.vec, prob=0.85)

return(output)

}

bmi.samp <- c(21.8,36.6,22.0,24.4,22.2,20.0,19.2,21.6,27.2,28.9,19.4,28.1,18.6,26.6,20.6,26.7,26.5,25.3,29.6,24.7)

my.n <- length(bmi.samp)

#Defining the number of resamples:

my.m <- 1000

# Setting up the matrix to hold bootstrap-sample values

setup.data.matrix <- matrix(bmi.samp, nrow=my.m, ncol=my.n, byrow=T)

# carrying out the sampling (with replacement):

bootstrap.data.matrix <- apply(setup.data.matrix, 1, sample, size=my.n, replace=TRUE)

# Transposing to get back to same dimensions as setup.data.matrix

bootstrap.data.matrix <- t(bootstrap.data.matrix)

# Calculating the sample mean for each of the bootstrap samples

my.85.percs <- apply(bootstrap.data.matrix, 1, perc85)

# standard error of this statistic:

sd(my.85.percs)

# 95% Bootstrap interval estimates for population 85th percentile:

lower.upper.CI <- quantile(my.85.percs, probs=c(0.025, 0.975))

print(paste("95% bootstrap interval for 85th percentile: ", round(lower.upper.CI,2) ))