Q SCI 482 Assignment 2 (100 points)

Q1: Are these Douglas firs about 100 years old? (50 points)

Douglas firs (*Pseudotsuga menziesii*) play a prominent role in the forests of the Pacific Northwest. Most patches were logged in the late 1800s or early 1900s, resulting in large Douglas firs in this region generally being secondary growth with an average age of about 100 years old. From a recent census of secondary growth trees of this age in the region, it is known that the population mean diameter is 88.5 cm and population standard deviation (SD) is 20.8 cm. A property owner in the Cascade mountains has approached you, asking if the trees on his property are typical of secondary growth. If they are larger than average, the property will be more valuable; if they were logged more recently and are smaller, the property will be worth less. They have measurements from their property of the diameters of Douglas firs in the file "`DouglasFirs.csv`". For hypothesis testing (parts c-e), assume that the sample has the same standard deviation as the population.

a. Read in the data, extract the appropriate column containing the diameters, and make sure that the data you have read in seem reasonable before proceeding. [5 points]

b. Find the sample mean, sample standard deviation, and sample size. [5 points]

c. Assuming that the <u>known population SD</u> applies to the sample of trees, calculate the standard error, the 90% confidence interval, 95% confidence interval, and 99% confidence interval for the sample. [15 points]

d. Conduct a hypothesis test that calculates the probability that the sample comes from a population of 100-year-old trees. Specify your null and alternative hypotheses, outline the test, note whether the test is one-tailed or two-tailed, calculate the test statistic, and report the resulting p-value. [15 points]

e. If the significance level, α, is 0.05, would you reject the null hypothesis? What if the chosen significance level is 0.005? Given the test results, are the trees on the property owner's land about 100 years old, or are they significantly older or younger? [10 points]

Q2: Are my trees 100 years old? (20 points)

The trees in my neighborhood may have been logged several times since old growth forest covered Snohomish County. The first major logging event was about 100 years ago. I went out into my garden and measured the diameter of the eight Douglas firs on my property (data in "`MyDouglasFirs.csv`"). <u>Assume now that you do not know the population standard deviation of diameters for secondary growth trees,</u> nor is my sample size large enough that the standard deviation of my sample can be assumed to be the same as the standard deviation of the population.

Conduct a statistical test, using a similar sequence of steps to question 1, to test whether the sample of trees on my property could be secondary growth, if secondary growth trees (the result of a single logging event) have a mean diameter of 88.5 cm. Explain your choice of statistical test, outline the steps involved, highlight your assumptions, report the 95% confidence interval from this sample, test statistic, p-value, and whether you accepted or rejected the null hypothesis, and finally answer the question above.

Q3: Is my (Trevor Branch's) scale accurate? (30 points)

I have been weighing myself regularly to examine the effects of exercise and diet changes, but the measurements appear to fluctuate greatly from day to day, far in excess of the normal 1.1 lb (pound) expected variation in adult mass. The scale manufacturer promises that repeated measurements of a standard mass have a standard deviation of $\sigma = 0.2$ lb. I repeatedly weighed a heavy bag of sand on various surfaces, removing the batteries and recalibrating between measurements, discovering no variation when weighed on a flat, hard surface. But measurements of my own mass varied by up to 0.8 lb, presumably because people move a bit more than bags of sand. These values are contained in the file `"scalereadings.csv"`. A randomly chosen number has been subtracted from these measurements to avoid personal embarrassment.

Important note: the statistical tables in Zar, and the corresponding equations, assume the probability of χ^2 value *greater than* a calculated value, but `pchisq()` and `qchisq()` return the probability of being *less than* a calculated value (i.e., the CDF) unless you specify `lower.tail=FALSE`.

a. Calculate the 95% CI for the population SD, from the sample of data. [5 points]

b. Conduct a statistical hypothesis test to detect whether the SD of measurements is equal to the SD promised by the manufacturer, for $\alpha = 0.05$. Outline the null and alternative hypotheses, the choice of test and whether it is one-tailed or two-tailed, and report the sample SD, the value of the test statistic, the *p*-value, and whether the null hypothesis is accepted or rejected. Does the scale conform to the manufacturer's promises of accuracy? [25 points]

R functions and constants that might be useful for this assignment

```
c()       #concatenate values together into a vector, e.g. c(0,3,5)
dim()     #number of rows and number of columns in a data frame
head()    #print out the first few lines of a data frame or matrix, head(dataframe)
hist()    #creates a histogram from a vector of values
length() #the number of items in a vector
mean()    #returns the mean of a set of values in a vector or one data frame column
min()     #find the minimum of two or more values, need not be in a vector or data
frame
ncol()    #the number of columns in a data frame or matrix
nrow()    #the number of rows in a data frame or matrix
pchisq() #the CDF up to a chi-square value. Important note, that this is the area
           to the left, but Zar tables and writeup reports the area to the right
           of a chi-square value, unless lower.tail=F.
pnorm()   #the cumulative distribution function up to value Z for given mean and sd
pt()      #the CDF up to a value t for given mean and sd, from t-distribution
qchisq() #given CDF probability and d.f., returns the chi-square value. Important
           note, that this is the area to the left, but Zar tables and writeup
           reports the area to the right of a chi-square, unless lower.tail=F.
qnorm()   #given CDF probability, returns the corresponding Z value
qt()      #given CDF probability and d.f., returns the value t from t-distribution
read.csv() #read in a CSV file, e.g. xdata <- read.csv(file="values.csv")
round()   #round to a number of decimal places, e.g. round(3.14159, 3)
sd()      #returns the standard deviation of values in a vector, e.g. sd(vector)
sqrt()    #the square root of a number (or vector), e.g. sqrt(10)
```

Other useful operators and constants

```
?  #get help on a function, e.g. ?seq or ?":" (quotes needed for operators)
#  #everything after a # on a line is not read by R, use for comments
[] #for extracting parts of a data frame, e.g. X[,10] for the 10th column
"" #double quotation marks are needed around text values ("" not “”)
TRUE  #the Boolean value true
FALSE #the Boolean value false
```