Q SCI 482 Assignment 1 (100 points)

Q1: standard normal in R (15 points)

The density of a normal distribution is $Y_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$.

For a standard normal, $\mu = 0$ and $\sigma = 1$, and this simplifies to $Y_i = \frac{1}{\sqrt{2\pi}} e^{-\frac{X_i^2}{2}}$.

1. Using elementary expressions in R, calculate the density for a standard normal when $X_i = 0$.

2. Using elementary expressions in R, calculate the density for a standard normal when $X_i = 2$.

3. Show the R code demonstrating how you can check your answers using `dnorm()`.

Q2: What proportion of sturgeon will be retained? (30 points)

Gulf sturgeon are a very large fish species, once a target of commercial fishing, but are listed as endangered. In the Suwannee River, Florida, sturgeon have the peculiar habit of leaping high above the river water, and since they are so large, they frequently injure and even kill boaters through collisions. Managers have tried to reduce human injuries and deaths by slowing boats down on this river. Imagine that in addition to boating speed limits, Florida Fish and Wildlife Conservation Commission considers a slot limit similar to that in the Columbia River, where only fish that are 44-50 inches in length may be kept, while both shorter and longer fish must be released. Length data (in millimeters) are available for 203 sturgeon caught in Florida.

1. Load the data in "`SturgeonData.csv`" into R and save it in a variable. Answer all questions in units of millimeters (mm).

2. What is the mean length of sturgeon?

3. What is the median length of sturgeon?

4. What is the variance of sturgeon lengths?

5. What is the standard deviation of sturgeon lengths?

6. What is the coefficient of variation of sturgeon lengths?

7. What is the Z-value for a sturgeon of length 2000 mm?

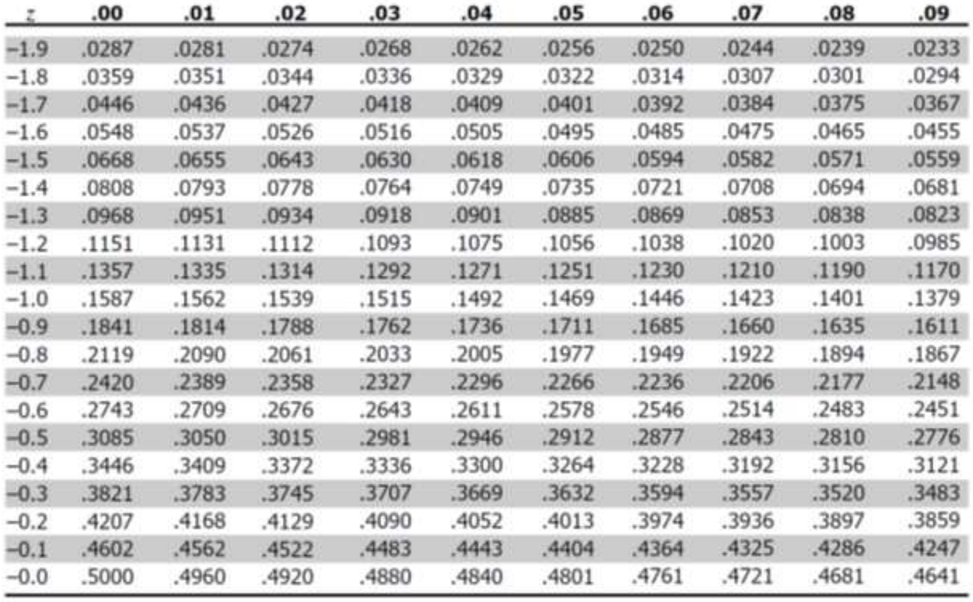
For parts 8-10, assume that sturgeon lengths follow a normal distribution with the parameters just calculated.
8. What proportion will be shorter than 1000 mm?

9. What proportion will be longer than 1500 mm?

10. What proportion will fall inside the slot limit of 44-50 inches (you'll need to convert to mm)?

Q3: statistical tables vs. R (10 points)

The statistical tables at the back of Zar's Biostatistical Analysis have largely been superseded by computer functions that can calculate exact values. For example, examine the portion of the table below for Z values, the CDF for the standard normal distribution. Remember that 95% CIs are 1.96 SDs or SEs from the mean: 95% of the area of a standard normal distribution is within -1.96 < Z < 1.96. Using R code and a choice of the pnorm/qnorm/rnorm/dnorm functions, recreate the values in the first row of the statistical table below for Z, rounded to the same number of decimal places (i.e. 0.0287, 0.0281, …, 0.0233). Note that there are several correct ways of doing this. If you want to find an elegant answer, you could use the `seq()` function and the knowledge that many R functions will accept vectors of values as inputs.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Q4: standard errors and the Central Limit Theorem (45 points)

High blood pressure (hypertension) is a long-term risk factor for many diseases, including heart disease, stroke, and kidney failure. However, because blood pressure is infrequently monitored, high blood pressure often goes undetected. Treatment for high blood pressure involves taking one of several types of medication, with side-effects that differ among the medication types. After two high blood pressure readings in a doctor's office, Patient A has been measuring blood pressure daily (data contained in "BPbefore.csv"), and started taking Medication One, while measuring blood pressure while medicated (data contained in "BPafter.csv"). Each blood pressure reading consists of two parts: the systolic and the diastolic, recorded in units of millimeters of mercury (abbreviated mmHg). A diagnosis of high blood pressure occurs when the systolic is over 140 or the diastolic is over 90, which is written as 140/90 mmHg.

In this example, we will be using standard errors and 95% confidence intervals to assess whether Patient A's mean blood pressure readings declined after taking Medication One, and what the implications are for Patient A.

1. Read in the data from "BPbefore.csv" and "BPafter.csv" into variables in R. Do some checks to make sure that there are no typos or missing data points. [5 points]

2. For the before and after treatment data, and for systolic and diastolic blood pressure, and assuming that the values are normally distributed, calculate the following values in R and write the answers in the data table below: number of data points (n), mean, standard deviation (SD), standard error of the mean (SE), and the lower and upper 95% confidence intervals (CIs) for the mean, i.e. based on the SE. Recall that 95% of the values in a standard normal distribution are -1.96 < Z < 1.96. [25 points]

Measurement						95% CI	95% CI
type	Treatment	n	mean	SD	SE	lower	upper
Systolic	Before meds						
Systolic	After meds						
Diastolic	Before meds						
Diastolic	After meds						

3. If there is no overlap in the 95% CIs of the means, it is safe to conclude that two means are significantly different. Given this, can you conclude that the medication lowered either or both of the systolic and diastolic blood pressure? [5 points, no R needed]

4. After taking the medication, does Patient A still have a medical diagnosis of high blood pressure? [5 points, no R needed]

5. Does Patient A need to take Medication One? [5 points, no R needed]

R functions and constants that might be useful for this assignment

```
c()       #concatenate values together into a vector, e.g. c(0,3,5)
dnorm()   #calculate the density (height) of a normal distribution
exp()     #exponential, e to the power of, e.g. exp(1) returns the constant e
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
median()  #returns the median of values in a vector
ncol()    #the number of columns in a data frame or matrix
nrow()    #the number of rows in a data frame or matrix
pi        #a constant, contains the value 3.141593, e.g. 2*pi*10^2
pnorm()   #the cumulative distribution function up to value q for given mean and sd
print()   #prints the value to the console
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)
seq()     #return a sequence of numbers, e.g. seq(from=1.5, to=2.3, by=0.1) or
          #   seq(from=0, to=1, length.out=51)
signif()  #round to specified number of significant digits, e.g. round(3876.463, 2)
sqrt()    #the square root of a number (or vector), e.g. sqrt(10)
var()     #returns the variance of values in a vector, e.g. var(vector)
```

Other operators and constants introduced in lab 1

```
apropos()#returns all functions that contain those letters, e.g. apropos("norm")
log()     #natural logarithm ln, e.g. log(exp(1)) returns 1
```

```
log10()  #base 10 logarithm, e.g. log(10) returns 1
max()    #returns the maximum value in a vector (or matrix or data frame)
rnorm()  #generate values randomly from a normal distribution, e.g. rnorm(100)
sum()    #add up all the numbers in a vector/array/data frame, e.g. sum(1:10)
vector() #create a vector
+  #add
-  #subtract
*  #multiply, e.g. 7*6  is 42
^  #to the power of, e.g. 10^3 is 1000
>  #greater than
>= #greater than or equal to
<  #less than
<= #less than or equal to
== #equal to
!= #not equal to
!  #not operator (turns TRUE into FALSE and FALSE into TRUE)
<- #assign the value or result on the right, to a variable on the left
[] #part of vector, matrix, or data frame, e.g. xvec[c(2,5)]
() #contains the parameters of a function
{} #contains a block of commands, used in if-then-else, functions
:  #generates sequences of numbers, e.g. 1:10
?  #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
"" #double quotation marks are needed around text values ("" not “”)
NA #special character for missing data: value exists but not available
NULL #special character for something that does not exist
TRUE #the Boolean value true
FALSE #the Boolean value false
```