Name: ___________________________________

Computing in Statistics, STAT:5400
Midterm 2, Fall 2016

You must work in the Linux environment. Submit your answers in the ICON drop box as an .Rnw file and .pdf file produced using Sweave, with your name as author. If you can't get your .Rnw file to compile, submit it anyway and include your R output in a separate text file.

If you arrive at 3:30, you must upload your solutions by 5:20.

If you arrive at 4:30, you must upload your solutions by 6:20.

In your document, have a named section for each problem, and, where needed, a numbered list of answers to multipart questions. You don't have to type any other text except where needed to answer a question.

1 Parametric bootstrap

This problem is based on the variable `indoor` from the `BaP.txt` dataset provided under `Datasets` on the course web page.

1. Read the data into R. Assume that the values of `indoor` are a random sample from a population. Obtain a point estimate of the median of the population. Show your R code and output.

```
> BaP <- read.table("http://homepage.divms.uiowa.edu/~kcowles/Datasets/BaP.txt",
+       header = TRUE)
> indoorMed <- median(BaP$indoor)
> print(indoorMed)

[1] 50

>
```

2. Assume that the population distribution of `indoor` is exponential. Use the `boot` function in the `boot` package to carry out a parametric bootstrap to assess the bias and standard error of the point estimate that you obtained in the previous problem. Use 3000 bootstrap datasets. Use an R function to time how long execution of the `boot` function takes. Show your R code and output.

```
> myRanGen <- function( dat, parms )
+ {
+    n <- length(dat)
+    rexp( n, parms$lambda)
+ }
> library(boot)
> system.time(bootout <- boot( BaP$indoor, median, 3000, sim = "parametric", ran.gen =
+          myRanGen, mle = list( lambda = 1 / mean(BaP$indoor) ) ))

   user  system elapsed
   0.07    0.00    0.07

> unbiased <- 2 * bootout$t0 - mean(bootout$t)
> print(bootout)

PARAMETRIC BOOTSTRAP


Call:
boot(data = BaP$indoor, statistic = median, R = 3000, sim = "parametric",
    ran.gen = myRanGen, mle = list(lambda = 1/mean(BaP$indoor)))


Bootstrap Statistics :
    original  bias    std. error
t1*       50 6.63701    20.67438

> print(paste("Unbiased estimate: ", unbiased))

[1] "Unbiased estimate:  43.3629902107471"

>
```

2 Simulation study

A student reasons as follows. Approximately 95% of the area under a normal curve lies between $\mu - 2\sigma$ and $\mu + 2\sigma$. Therefore, one should be able to estimate a population standard deviation using a random sample from the population by calculating the empirical .025 and .975 quantiles of the sample data and computing:

$$\frac{.975 quantile - .025 quantile}{4}$$

Conduct a simulation study to evaluate the *bias* in the student's estimator when

- the population distribution is exponential with rate parameter 0.25 and
- the sample size $n = 200$.

Choose your number of replicate datasets large enough that the standard error of your estimate of bias is no larger than 0.01. (Note that the true variance of an exponential random variable with rate parameter λ is $\frac{1}{\lambda^2}$).

```
> mystat <- function(v)
+ {
+     qs <- quantile(v, c(0.975, 0.025) )
+     (qs[1] - qs[2]) / 4.0
+ }
> truesd <- 1.0 / 0.25
> S <- 1000
> n <- 200
> myfunc <- function(S,n)
+ {
+     set.seed(72)
+
+     mydat <- matrix(rexp( S * n, rate = 0.25), nrow = S)
+
+     myests <- apply(mydat, 1, mystat)
+
+     biases <- myests - truesd
+
+     list( estbias = mean(biases), stderr = sd(biases) / sqrt(S))
+ }
> myfunc(S=S, n=n)

$estbias
[1] -0.4435608

$stderr
[1] 0.01406721

>
```

The standard error is too large by a factor of 1.4. I will increase S by a factor of 1.4^2.

```
> S <- ceiling(1.4^2 * S)
> myfunc(S=S, n = n)

$estbias
[1] -0.4388089

$stderr
[1] 0.009289604

>
```

3 Relational database structures

A university wishes to store information on its classrooms. Here is a flat file structure that could represent the data.

```
Building abbreviation (example:  SH)
Building name  (example:  Schaeffer Hall)
Street address
Year building constructed
Building has central air conditioning?
Room number
Number of seats
Has blackboard?
Has whiteboard?
Has computer with projector?
Has overhead projector?
```

Put these data into a relational structure that is in third normal form. In a verbatim environment, just list all the tables that you would need, and the fields in each. Indicate all primary and foreign keys.

```
Buildings table
---------------
Building abbreviation (primary key)
Building name
Street address
Year built
Central air?

Rooms table
-----------

Building abbreviation   \      First 2 fields form primary key
Room number             /      Building abbreviation also is foreign key
Number of seats
blackboard?
whiteboard?
computer with projector?
overhead projector?
```