Stat414/614; Fall 2024; Worksheet 06; 20 Points; NAME:

This worksheet, is based on the Module06 material, but also draws on your knowledge from previous modules, especially Module05. Main event here is the introduction of a super useful and widely used technique called "Bootstrapping" to estimate standard errors (and in general, sampling distributions) of estimators, without making any distributional assumptions on the population. After the simulation technique you learned in Module03, bootstrapping is perhaps the most important skill you are picking up in this course. Happy bootstrapping!

Launch RStudio. Install the package EnvStats. In the `Files,Plots,...` pane, enter `EnvStats` in the search box and find the main help page of this package. Enter `EPA` in the search box and find all data sets available in this package. Find the tccb data set, `attach(EPA.94b.tccb.df)` it and using `names(EPA.94b.tccb.df)` find the names of the variables in it.

Subset the data corresponding to the Reference area and the clean up area as follows.

```
library(EnvStats)
attach(EPA.94b.tccb.df)
TcCB.Ref <- TcCB[Area=="Reference"]
TcCB.Cleanup <- TcCB[Area=="Cleanup"]
detach(EPA.94b.tccb.df)
```

(Note that once we run the above segment, `TcCB.Ref` and `tcCB.Cleanup` will be available through out the current R session. This method has the advantage of not having to re-run attach() functions again and again, which often leads to the annoying `masking` problem.)

1. Determine if we can comfortably assume a theoretical model for each of the above variables.
   (a) Plot the histogram of the data, and overlay a log-normal curve.
   (b) Use the function `gofTest()` of `EnvStats` to test the log-normality assumption of the data for each variable.
   (c) State your conclusions.
2. Apply the log transformation to each of the above variable and create

   ```
   ln.TcCB.Ref <- log(TcCB.Ref)
   ln.TcCB.Cleanup <- log(TcCB.Cleanup)
   ```

   (a) Use the appropriately modified version of the Bootstrapping code provided to obtain estimates of the mean [$=\mu = E(log(X))$] for each variable and corresponding bootstrap standard errors.
   (b) Compare the above SE's to the corresponding $\frac{s}{\sqrt{n}}$, and comment on how close they are.