1. In lecture I drew normal probability plots using the Savona.csv data set. This data set has been included along with the exercises in order for you to complete this question.

   (a) Draw the normal probability plot that I demonstrated in lecture.
   Solutions:

   ```
   >savona=read.table("h:/Savona.csv",header=TRUE,sep=",")
   > SavRos=with(savona,ros(obs=Orthophos,censored=OthoCen,forwardT=NULL))
   > plot(SavRos)
   ```

   (b) Using the argument discussed in lecture, draw a lognormal probability plot of the data.
   Solutions:

   ```
   > logSavRos=with(savona,ros(obs=Orthophos,censored=OthoCen,forwardT="log"))
   > plot(logSavRos)
   ```

   (c) Sometimes it can be useful to compare two plots right next to each other. To do this, we can set graphing parameters in R using the function `par()`. If we type `par(mfrow=c(2,1))` Our graphics window will display 2 graphs, one on top of the other. Type the command `par(mfrow=c(2,1))`, and then type the commands to draw the normal and lognormal probability plots again to view them together.
   Solutions:

   ```
   >par(mfrow=c(2,1))
   plot(SavRos)
   plot(logSavRos)
   ```

   (d) Based on comparing the normal and lognormal plots right next to each other, would you choose to conduct the remainder of an analysis on the original scale, or on the log-scale?
   Solutions:
   Although there are a few small differences between the plots, it is hard to evaluate which one is 'best'. I would tend to choose the original scale for the analysis because it makes the interpretation at the end much simpler.

2. Being able to identify the *population* that you are sampling from is key to knowing what inferences can and cannot be legitimately drawn from a data set. Below are some sampling scenarios; see if you can identify the population that we can draw inference to based on descriptions of samples that were taken.

   (a) You are creating a stage height record for a river. You measure the river stage once a day, but on days where there is high precipitation or snow melt, the local bridge closes due to flooding so you cannot take an observation.
   Solutions:
   Our population here is the population of days when there is NOT high precipitation or snow melt. We can only draw inference back to the range of stage height values that we actually observe.

   (b) You are interested in contamination caused by a mining operation. What inferences can be drawn under the following sampling scenarios?

      i. You sample from the tailing pond only.
      Solutions:
      Our population here is water in the tailing pond. Nowhere else.

      ii. You sample from a lake near the tailing pond, and the tailing pond.
      Solutions:
      This makes it possible to broaden inference to the lake and the tailing pond.

      iii. You sample from the lake only.
      Solutions:
      Now inference is restricted to just the lake, we don't know anything about the tailing pond.

      iv. You sample from the lake, and groundwater sources near the tailing pond.
      Solutions:
      Inference now extends towards the lake and the groundwater sources, but not the tailing pond itself.

(c) You are looking at a report submitted as part of a permit for a mine plan. The particular focus is the adequacy of the reservoirs submitted for the mine plan. The report includes 60 years of precipitation data, but the reservoirs are supposed to be designed to withstand a 1 in 200 year flooding event. Is the precipitation record being used for reservoir estimation a sufficient sample to draw inference to our population of interest?

Solutions:
This is a tricky situation, and worthy of some discussion. The simplistic answer is that if you only have 60 years of data, you can draw inference to that 60 years of information. On the other hand there are methods for estimating the size of 1 in 200 year events based on shorter data sets. Whether the precipitation record being used extends to our population of interest (which includes 1 in 200 year flooding events) depends on the methods that are indicated in the report calculations.

(d) A permittee is applying to expand a landfill site this year. Local community members are concerned that landfill leachate might affect their nearby water sources. What inference can you draw about the impact of leachate in community water sources under the following circumstances?

i. Groundwater sampling occurred at the site in 2004 and 2005.
Solutions:
If sampling occurred only in 2004 and 2005, there is no valid way to extend inference all the way to 2009. Our population is water quality near the landfill in 2004 and 2005. It is not possible to draw inferences regarding the effect of leachate on local groundwater based on this population

ii. Groundwater sampling occurred between 2004 when the site opened, and the present within a 200m buffer zone around the landfill . The nearest residential well site is 150m from the landfill.
Solution:
Our population has now been redefined to groundwater sources within 200m of the landfill, in relevant years. If the nearest residential site is inside this zone, than we can make inferences about the effect of leachate at this well.

iii. Same as above, except the nearest residential well is 500m away from the landfill. Our population here is still groundwater within 200m of the landfill. Now that our residential site has moved further away, it is no longer part of the population that we can draw inference to. We cannot draw inference to the impact of leachate on this residential well.

3. Refer back to the Cadmium data set that was used in the third set of lectures and Exercise 3. We will be using it again here!
In lecture we discussed that there are 3 main steps to creating good summary statistics. Following these steps, conduct a summary data analysis of the Cadmium data set. Justify any transformations of the data that you take. If you cannot complete the analysis due to assumption violations, discuss this.
If you perform a data transformation, don't worry about transforming the summary data back to the original scale!

Solutions:
The first step is to assess the level of censoring to decide if it is less than 50%.

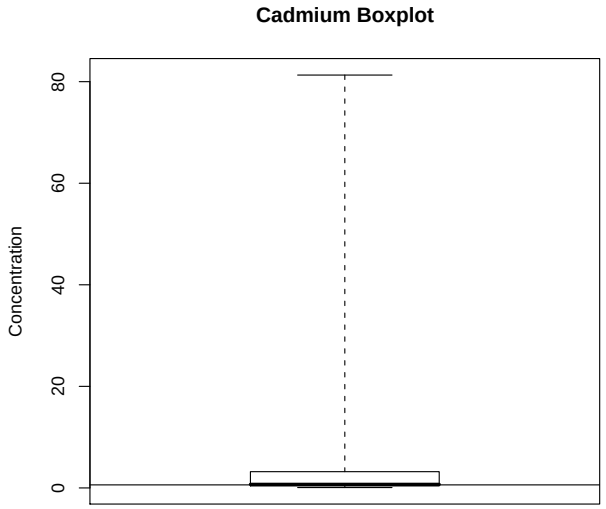
```
> with(Cadmium,censummary(obs=Cd,censored=CdCen))
all:
       n    n.cen  pct.cen      min      max
19.00000  4.00000 21.05263  0.20000 81.30000

limits:
  limit n uncen   pexceed
1   0.2 1     0 0.7142857
2   0.3 1     0 0.7142857
3   0.4 1     3 0.7142857
4   0.6 1    12 0.5714286
```

There is only 21% censoring of the data, and so this is suitable for further analysis.

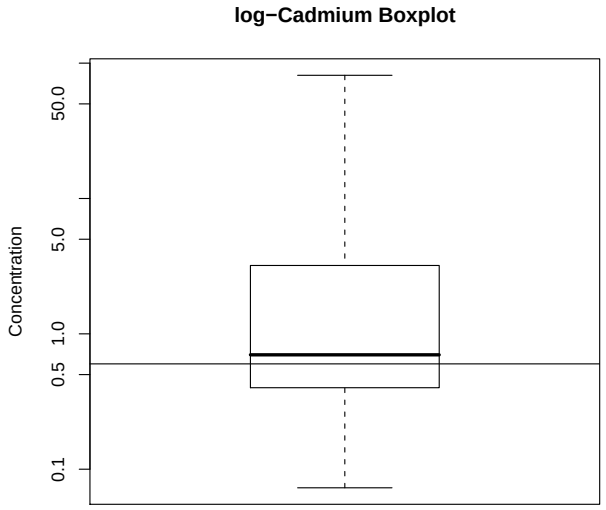
The second step is to look at a boxplot to see what the data looks like.

```
data(Cadmium)
with(Cadmium,cenboxplot(obs=Cd,cen=CdCen,main="Cadmium Boxplot",ylab="Concentration",log=FALSE))
```

This will result in the following plot being drawn.



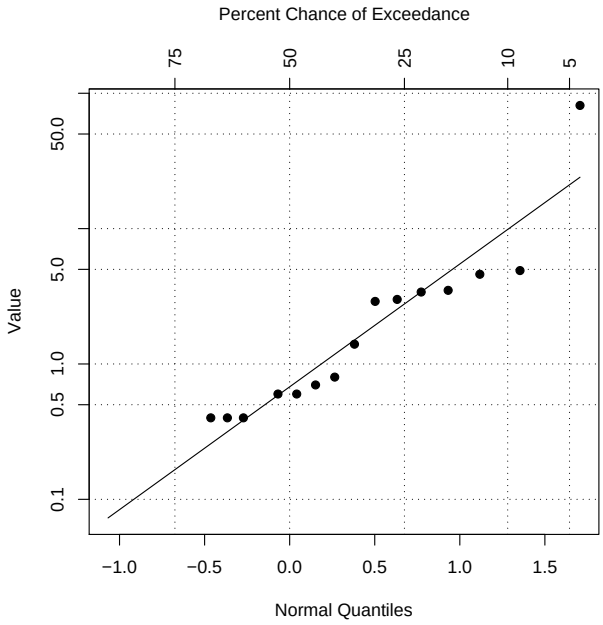
This shows evidence of right skewness, indicating that it might be better to take a log transformation and conduct analysis using the lognormal distribution. The boxplot on the log scale is shown below.



To confirm that the lognormal distribution is the more appropriate scale for analysis, we can construct a censored log probability plot based on the lognormal distribution.

```
CadmiumRos=with(Cadmium,ros(obs=Cd,censored=CdCen,forwardT="log"))
plot(CadmiumRos)
```

The lognormal probability plot shows no severe violations of the lognormal assumption, so we will conduct the rest of the analysis on this scale.



Now that appropriate evaluations of the data have taken place, we can calculate summary statistics.

```
> CadMLE
         n      n.cen     median       mean
19.0000000  4.0000000  0.9883846  4.0523779
        sd
16.1129838
> mean(CadMLE)
     mean        se   0.95LCL   0.95UCL
 4.052378  0.399717  1.276004 12.869680
> quantile(CadMLE,conf.int=TRUE)
  quantile       value    0.95LCL    0.95UCL
1     0.05  0.06236013 0.01574127  0.2470440
2     0.10  0.11480415 0.03473363  0.3794591
3     0.25  0.31830645 0.12443457  0.8142351
4     0.50  0.98838464 0.45152964  2.1635439
5     0.75  3.06906817 1.32723774  7.0968291
6     0.90  8.50931067 2.99801178 24.1521293
7     0.95 15.66552406 4.68418691 52.3908736
```

To learn how to transform these estimates back to the original scale, refer to your guidance document!!