Name: Steve G. Mwangi
Date: November 10th, 2019
Assignment: TMATH 390 R Lab 6 Document.

Objectives

1. **Using simulated data to explore sampling distributions.**

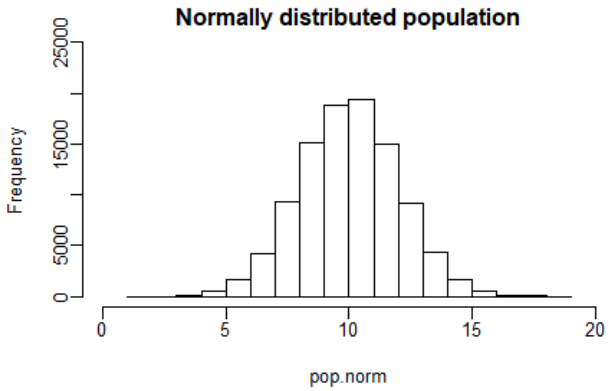
C1. (4) Submit your R script to Canvas.

C2 (2). Report the values for pop.norm.mu, pop.norm.med, pop.norm.var, and pop.norm.sigma

```
> ## First getting a clean slate on the environment by calling
> # rm to remove objects. The list is the character vector naming
> # objects to be removed.
> help(rm)
> rm(list=ls())
> # Each population will have 100000 individuals (N)
> # Available to be sampled from
> N = 100000
> # Setting the mean as 10, and sd as 2 for the population of N
> pop.norm=rnorm(N,mean=10,sd=2)
> # Let's record the population parameters
> # for this population
> pop.norm.mu=mean(pop.norm)
> pop.norm.med=median(pop.norm)
> pop.norm.var=var(pop.norm)
> pop.norm.sigma=sqrt(pop.norm.var)
> pop.norm.mu
[1] 10.00025
> pop.norm.med
[1] 10.00802
> pop.norm.var
[1] 3.997102
> pop.norm.sigma
[1] 1.999275
```

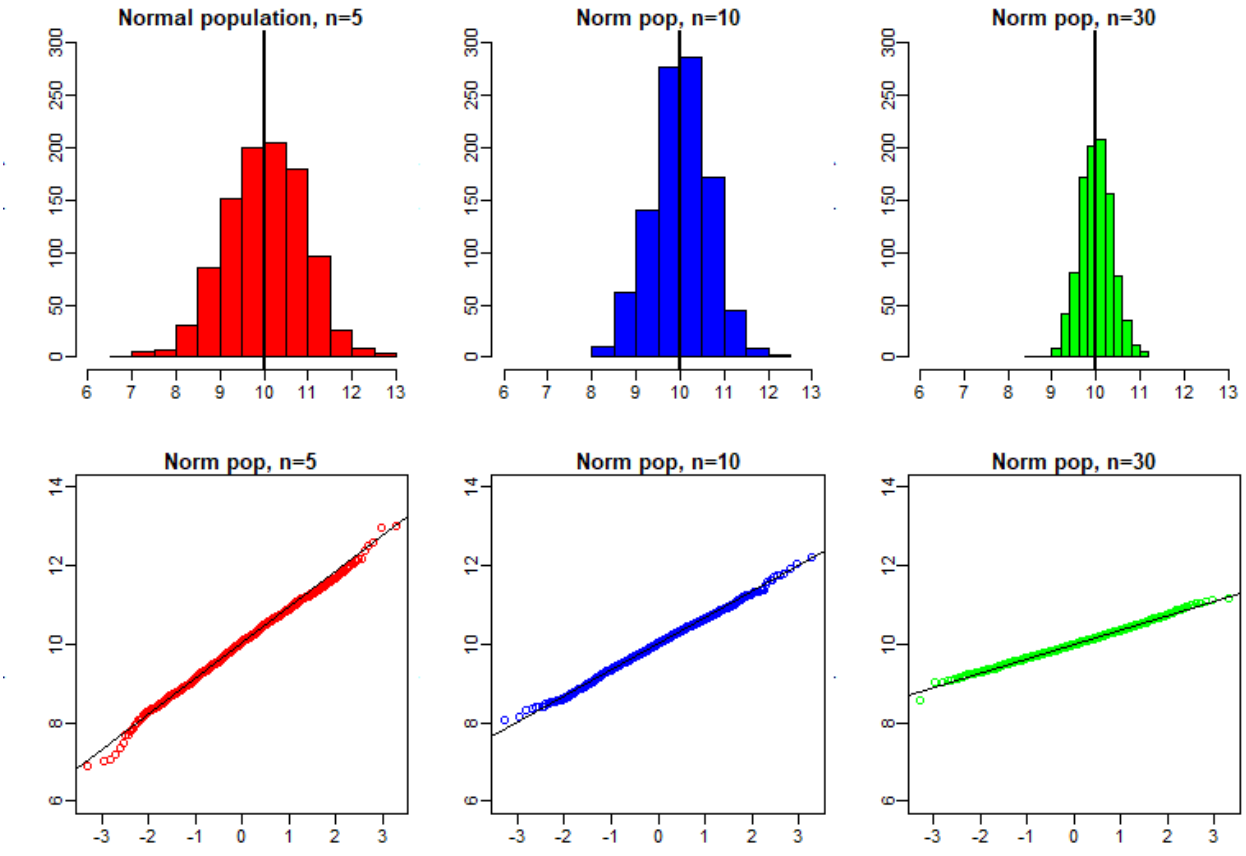
Now we will observe the distribution of individuals, to compare them to the distribution of sample means and medians for these populations. For your population of individuals draw a histogram of the distribution. Make sure to customize the xlim for each population because we will use this to compare the distributions to the sampling distributions.

C3 (2) Provide the graph showing the distribution of your population.

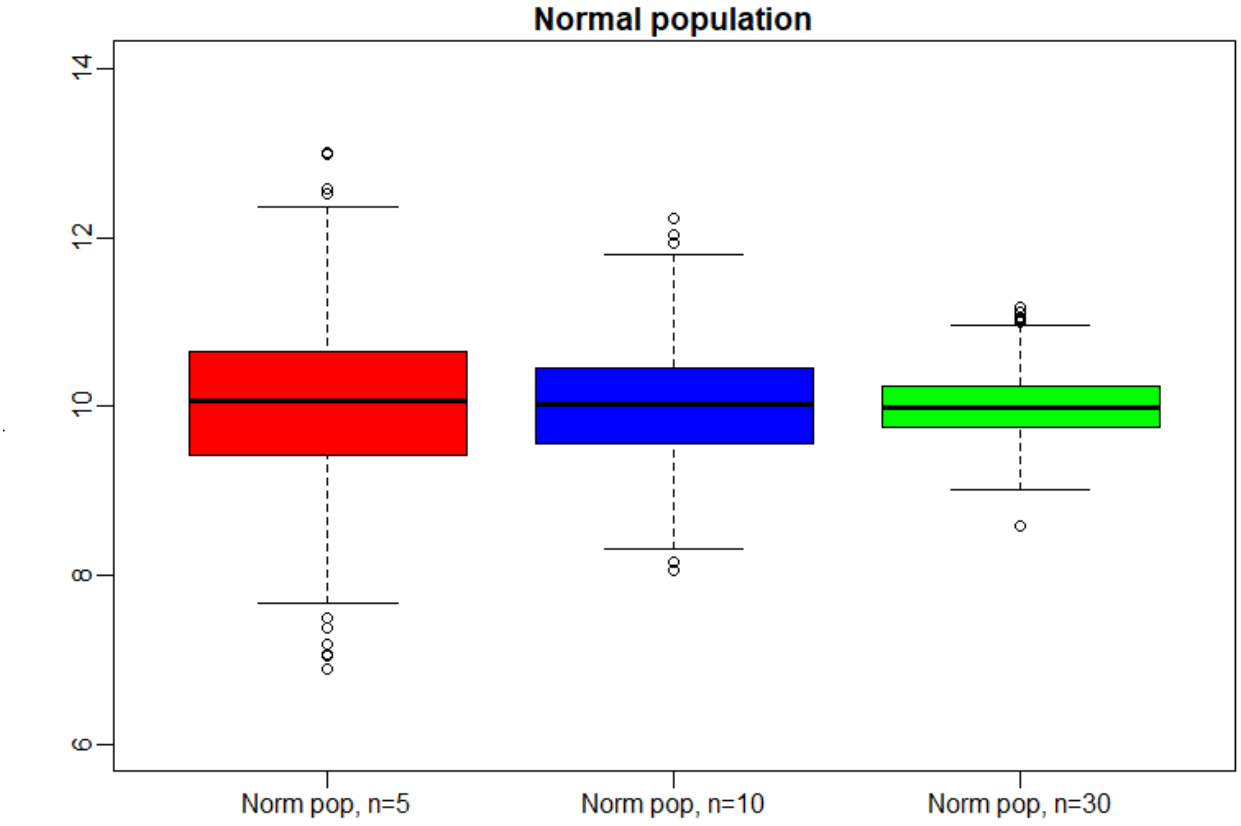


C4 (3) Provide the graph generated by the code above for the distribution of the sample mean.

```
> # NOTE: mfcol tells R to fill in the panels column by column rather than by row
> ### Size 5
> par(mfcol=c(2,3),mar=c(3.5,3.5,1.5,0.5),mgp=c(3.5,0.5,0))
> hist(samp.norm5$mean,xlab="Sample mean",main="Normal population, n=5", xlim=c(6,13),ylim = c(0,
300), col = "red")
> # draw a vertical line at the true value
> abline(v=pop.norm.mu,lwd=2)
> qqnorm(samp.norm5$mean,col = "red",ylim =c(6, 14), main="Norm pop, n=5")
> qqline(samp.norm5$mean)
> #-------------------------------------------------------------------------------------------------
----
> ### Now for samples of size 10
> hist(samp.norm10$mean,xlab="Sample mean",main="Norm pop, n=10", xlim=c(6,13), ylim = c(0, 300),
col = "blue")
> # draw a vertical line at the true value
> abline(v=pop.norm.mu,lwd=2)
> qqnorm(samp.norm10$mean,col = "blue", ylim =c(6, 14), main = "Norm pop, n=10")
> qqline(samp.norm10$mean)
> #-------------------------------------------------------------------------------------------------
----
> ### Now for samples of size 30
> hist(samp.norm30$mean,xlab="Sample mean",main="Norm pop, n=30", xlim=c(6,13), ylim = c(0, 300),
col = "green")
> # draw a vertical line at the true value
> abline(v=pop.norm.mu,lwd=2)
> qqnorm(samp.norm30$mean,col = "green",ylim =c(6, 14), main = "Norm pop, n=30")
> qqline(samp.norm30$mean)
```



```
> # Now a boxplot to compare all three distributions side by side
> par(mfcol=c(1,1),mar=c(3.5,3.5,1.5,0.5),mgp=c(3.5,0.5,0))
> boxplot(samp.norm5$mean, samp.norm10$mean,samp.norm30$mean, ylab="Sample mean", xlab="Sample si
ze",col = c("red","blue", "green"),
+         names=c("Norm pop, n=5","Norm pop, n=10","Norm pop, n=30"), main="Normal population",yl
im=c(6,14))
```



C5 (3) Describe how the sampling distribution of the sample mean changes with increasing sample size, including the shape, variability, and center of the distribution.

Histograms

When n (sample size) = 5, small, the histogram is wide. Showing greater variability for the mean values. When n (sample size) = 30, large, the histogram is narrow, showing that most data are within the same bins. Hence narrow x-range from 9-12 for n = 30, vs 7 -13, for n = 5. All histograms are unimodal and bell-shaped, resembling a normal distribution, as they are all drawn from a normal population. Their median correspond to the entire population's median

Qqplots

The sample with size 30, has the y-intercept, mean, higher than the rest, at 8.7, and closer to the entire population's mean of 10. The n=30 scatterplot is somewhat left skewed, with both ends of the plots falling below that of the sample population mean qqline.

At size n = 10, intercept is 7.7, which varies slightly further from the entire population's mean of 10. This scatterplot shows a distribution that is slightly right-skewed, with both ends of the plot being above the sample population mean qqline.

At n = 5, intercept is about 7, which varies greatest from the entire population's mean of 10 compared to the other two distributions. The corresponding scatter plot seems to be leptokurtic, with more data below normal at the offset, and more data somewhat above normal at the right side of the plot—though it is hard to confirm this, since some of the data falls on the qqline at the right end of the plot.

Boxplots

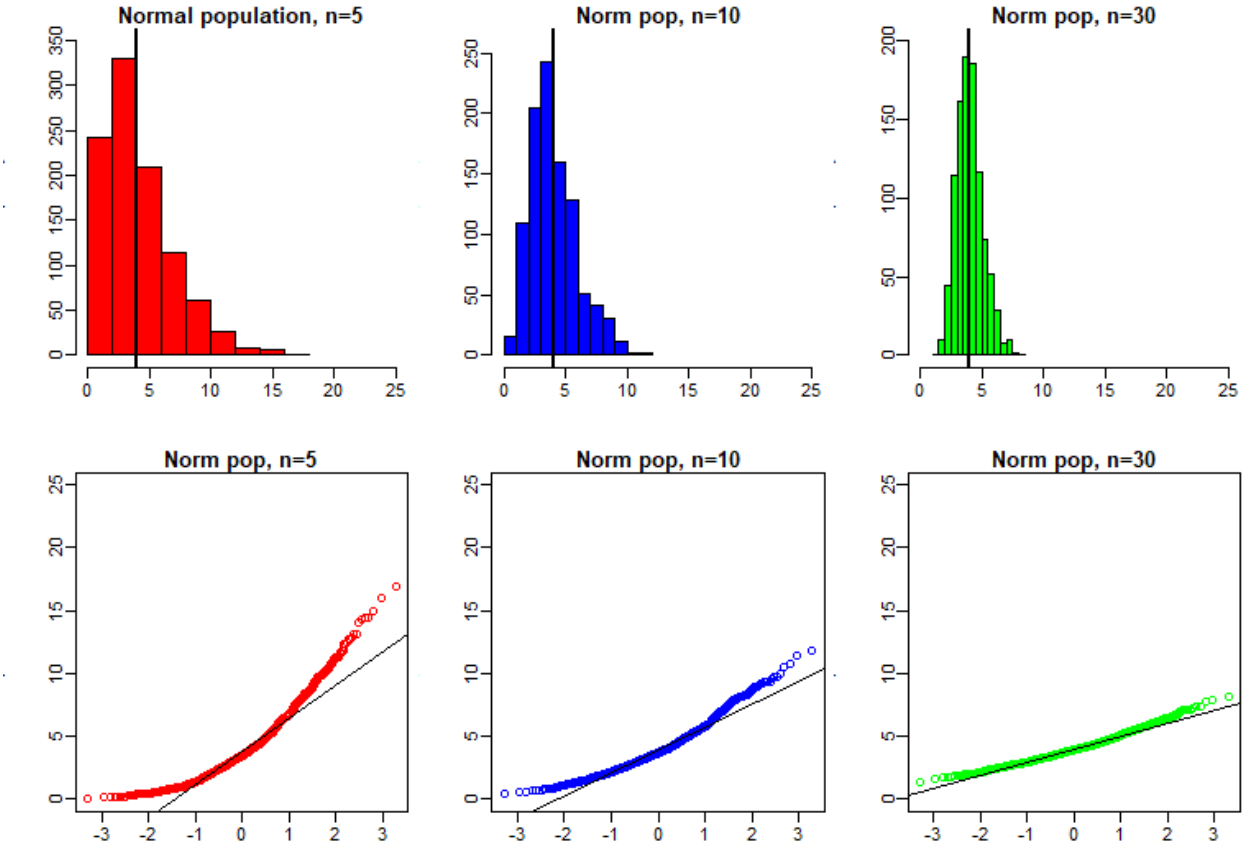
At n = 5, the boxplot is comparatively taller to n=10 and n=30. This confirms the dispersion of the data as shown by the corresponding histogram. With greater spread or variability at size n =5, than at n=30. Also echoed by the outliers, with outliers at n = 30, being much more closer to the entire population's median value, 10, than those of sample pop n = 10 and sample pop n=30.

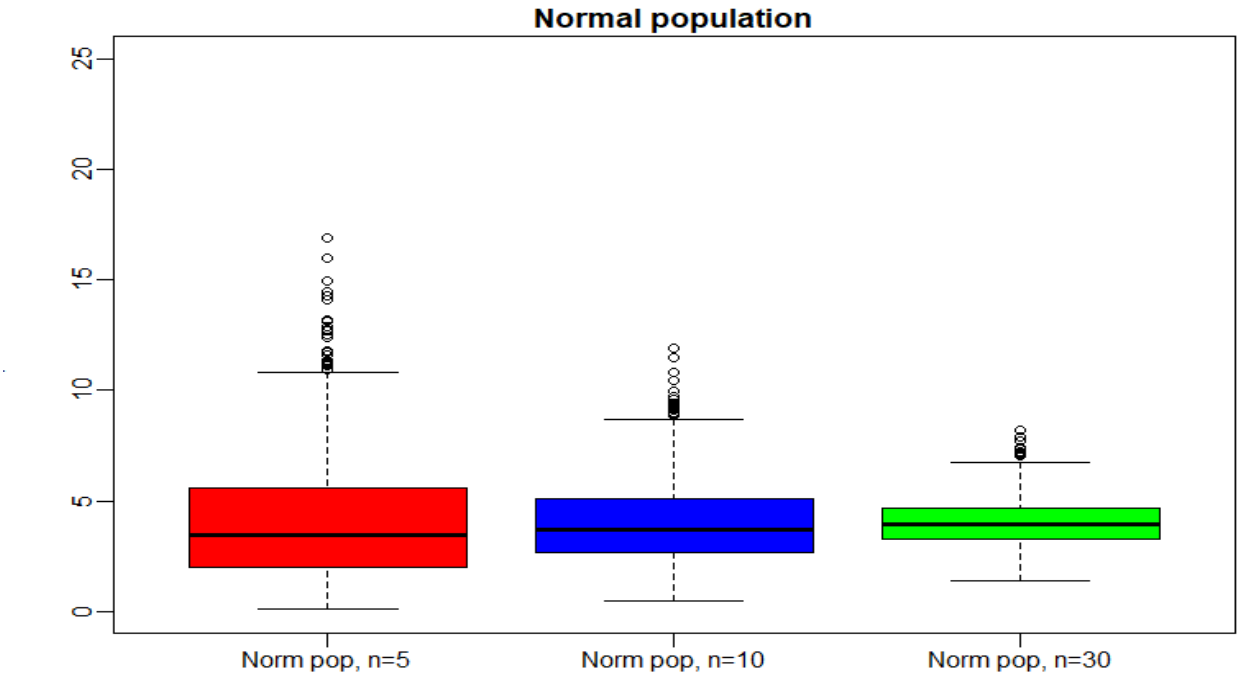
All three plots show that a sample size of n = 30, the mean closely resembles the corresponding population's mean than at n<30.

Now modify the graph code above to visualize the distribution of the sample variance when sampling from a normal distribution. Note. For the sample variance change the axis limits to be between 0,25, and make sure instead of pop.norm.mu you're looking at the value of pop.norm.var. We will use these plots to compare variability of the sampling distribution, which requires all are created on the same scale.

```
> ### Size 5
> par(mfcol=c(2,3),mar=c(3.5,3.5,1.5,0.5),mgp=c(3.5,0.5,0))
> hist(samp.norm5$var,xlab="Sample Variance",main="Normal population, n=5", xlim=c(0,25),ylim = c
(0, 350), col = "red")
> # draw a vertical line at the true value
> abline(v=pop.norm.var,lwd=2)
> qqnorm(samp.norm5$var,col = "red",ylim =c(0, 25), main="Norm pop, n=5")
> qqline(samp.norm5$var)
> # Entire Population Variance
> pop.norm.var
[1] 3.997102
> # Sample Population Variance n = 5
> var5 = var(samp.norm5$var)
> var5
[1] 7.892729
> #-------------------------------------------------------------------------------------------
> ### Now for samples of size 10
> hist(samp.norm10$var,xlab="Sample Variance",main="Norm pop, n=10", xlim=c(0,25), ylim = c(0, 26
0), col = "blue")
> # draw a vertical line at the true value
> abline(v=pop.norm.var,lwd=2)
> qqnorm(samp.norm10$var,col = "blue", ylim =c(0, 25), main = "Norm pop, n=10")
> qqline(samp.norm10$var)
> # Entire Population Variance
> pop.norm.var
[1] 3.997102
> # Sample Population Variance n = 10
> var10 = var(samp.norm10$var)
> var10
[1] 3.667556
> #-------------------------------------------------------------------------------------------
> ### Now for samples of size 30
> hist(samp.norm30$var,xlab="Sample Variance",main="Norm pop, n=30", xlim=c(0,25), ylim = c(0, 20
0), col = "green")
> # draw a vertical line at the true value
> abline(v=pop.norm.var,lwd=2)
> qqnorm(samp.norm30$var,col = "green",ylim =c(0, 25), main = "Norm pop, n=30")
> qqline(samp.norm30$var)
> # Entire Population Variance
> pop.norm.var
[1] 3.997102
> # Sample Population Variance n = 10
> var30 = var(samp.norm30$var)
> var30
[1] 1.150167
```

C6 (3) Provide the graph generated by the code above for the distribution of the sample sample variance





C7 (3) Describe how the sampling distribution of the sample sample variance changes with increasing sample size, including the shape, variability, and center of the distribution.

Histogram

All 3 histograms are right skewed and unimodal. With greater variability seen in the spread of the sample population n=5. Sample population with n=30, is the one that seems to have a histogram that closely resembles that of the underlying normal population.

Qqplot

All Qq plots confirm the right-skewedness shown by the three histograms. The qqplots at their start and end points are above the qqline of the sample variance. Also the y-intercept of the n=30 qqplot, which is a little over 0, is much closer to the underlying populations variance of 3.9971 than the other two plots' y-axis intercept, which appear to occur at a value below 0.

Boxplot

The variance of the sample variances from the normal population also echo the same pattern seen with the means. The boxplots show that there is greater variability with the variance from the sample n =5, than with at n = 30. Entire population variance is about 3.9971. Median of the boxplots seem to be close to this value, with n=30 closest of all three boxplots.

Also the boxplots show how the sample variance distributions are all right skewed, with the outliers for all three lying on the upper end of the boxplots.

The variance and mean distributions confirm that at n=30, the underlying population is closely resembled than at n<30.