TMATH 390 R Lab 6

Using simulated data to explore sampling distributions

C1 (4). Submit the R script you used to complete this computer lab assignment.

In this lab we will use R to simulate samples of different sizes from a normal distribution. We will then explore the distribution of the sample mean and the sample variance.

Sampling distributions

We will set things up for the current environment. We will be sampling from a normal distribution. Let's create "populations" of each distribution from which we will sample.

```
# first let's get a clean slate on the
# environment
rm(list=ls())
# Each population will have 100000 individuals (N)
# available to be sampled from
N = 100000
# now let's create a normally distributed population
# with a mean of 10 and a standard deviation of 2
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)
```

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

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.

Here is how you set the x-axis limits:

```
par(mar=c(3.5,3.5,1.5,0.5),mgp=c(2.5,0.5,0),mfrow=c(2,2))
# xlim=c(lower,upper) is the flag used to customize the
# x-axis limits. Substitute ylim=c(lower,upper) for the y-axis
hist(pop.norm,xlim=c(0,20),main="Normally distributed population")
```

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

For each population we will take 1000 independent samples, of different sizes (5,10,30). To sample from each population, we will use the sample function. For example, here is a single sample of size 5 from the normal population, from which we calculate the summary statistics.

```
# the sample function takes as arguments
# the object from which we're sampling
sample1=sample(pop.norm,size=5)
# here is the sample
sample1
```

```
## [1]  6.121971 11.334937  7.235422  5.883025 11.722741
```

```
mean(sample1)
```

```
## [1] 8.459619
```

```
median(sample1)
```

```
## [1] 7.235422
```

```
var(sample1)
```

```
## [1] 8.129377
```

```
sd(sample1)
```

```
## [1] 2.851206
```

We want to repeat this process 1000 times, storing the value for the sample mean ($\bar{x}$), the sample median ($\tilde{x}$), the sample variance (s^2) and the sample standard deviation (s), for each of the sample sizes. Here is an example for the normal distribution with sample size 5:

```
# first we create an empty data frame to store all
# of our values
samp.norm5=data.frame(mean=rep(NA,1000),
median=rep(NA,1000),var=rep(NA,1000),sd=rep(NA,1000))
# now we have a for loop that we use to repeat
# the sampling 1000 times. We index our loop with "i"
for(i in 1:1000) #index i starting at 1, until we hit 1000
{
  #draw the random sample
  tmp.samp=sample(pop.norm,size=5)
  # and enter the values for the summary statistics
  # in the data frame
  samp.norm5$mean[i]=mean(tmp.samp)
  samp.norm5$median[i]=median(tmp.samp)
  samp.norm5$var[i]=var(tmp.samp)
  samp.norm5$sd[i]=sd(tmp.samp)
}
```

Execute this code, then modify it to produce samples of size 10, and then samples of size 30. This requires changing the names of all of your objects so they have unique names (e.g., samp.norm5 changes to samp.norm10, or samp.norm30). And change the size argument in the sample function to the sample size of interest (10 or 30).

Now we want to look at the sampling distribution of our estimators. We can use histograms, boxplots, and normal QQ plots to look at the distribution of the sample mean and the distribution of the sample standard deviation with increasing sample size.

Here is an example to visualize the distribution of the estimators, as simulated above. Note we have set the x-axis limits to the same as for the population graph above. We can use the boxplot to directly compare the distributions. Use this code to create your graph, and make sure that the graphing window is large enough so all of the graphs are readable!

```
# note: mfcol tells R to fill in the panels column
# by column rather than by row
par(mfcol=c(2,4),mar=c(3.5,3.5,1.5,0.5),mgp=c(2.5,0.5,0))
hist(samp.norm5$mean,xlab="Sample mean",main="Normal population, n=5",
     xlim=c(0,20))
# draw a vertical line at the true value
abline(v=pop.norm.mu,lwd=2)
qqnorm(samp.norm5$mean,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(0,20))
# draw a vertical line at the true value
abline(v=pop.norm.mu,lwd=2)
qqnorm(samp.norm10$mean,main="Norm pop, n=10")
qqline(samp.norm10$mean)
### Now for samples of size 30
hist(samp.norm30$mean,xlab="Sample mean",main="Norm pop, n=30",
     xlim=c(0,20))
# draw a vertical line at the true value
abline(v=pop.norm.mu,lwd=2)
qqnorm(samp.norm30$mean,main="Norm pop, n=30")
qqline(samp.norm30$mean)

# Now a boxplot to compare all three distributions side by side
boxplot(samp.norm5$mean,samp.norm10$mean,samp.norm30$mean,
        ylab="Sample mean",xlab="Sample size",names=c("Pop","5","10","30"),
        main="Normal population",ylim=c(0,20))
```

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

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.

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.

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.