Jordan Mayer (Lab 071-1:30pm, LEC 076-1:30pm)
Harley Jo Rowland (Lab 071-1:30pm, LEC 076-1:30pm)
Qihang Xu(Lab Section 071-1:30pm, LEC 076-1:30pm)
Ernest Lee (Lab Section 051-3.30pm, LEC 050-3.30pm)

STAT 350 Lab 4

B. (20 points) Standard Normal Distribution.

1.

```
> SRS <- 1000
> n <- 1
> normal <- paste("Normal distribution: averaged over ", n)
> data.vec <- rnorm(SRS*n, mean = 0, sd = 1)
> data.mat <- matrix(data.vec, nrow = SRS)
> avg <- apply(data.mat, 1, mean)
```

```
> mean(avg)
[1] 0.0119638
> sd(avg)
[1] 1.011365
```

```
> library(ggplot2)
Warning message:
package 'ggplot2' was built under R version 3.4.3
> windows()
> xbar <- mean(avg)
> s <- sd(avg)
> ggplot(data.frame(avg=avg), aes(x=avg)) +
+     geom_histogram(aes(y=..density..),bins = sqrt(length(avg))+2,
+                    fill = "grey",col = "black") +
+     geom_density(col = "red", lwd = 1) +
+     stat_function(fun=dnorm, args=list(mean=xbar, sd=s), col="blue",
+                   lwd = 1) +
+     ggtitle(normal) +
+     xlab("Data") +
+     ylab("Proportion")
> ggplot(data.frame(avg=avg), aes(sample=avg)) +
+     stat_qq() +
+     geom_abline(slope = s, intercept = xbar) +
+     ggtitle(normal)
```

Mean and SD for n = 2,6,10:

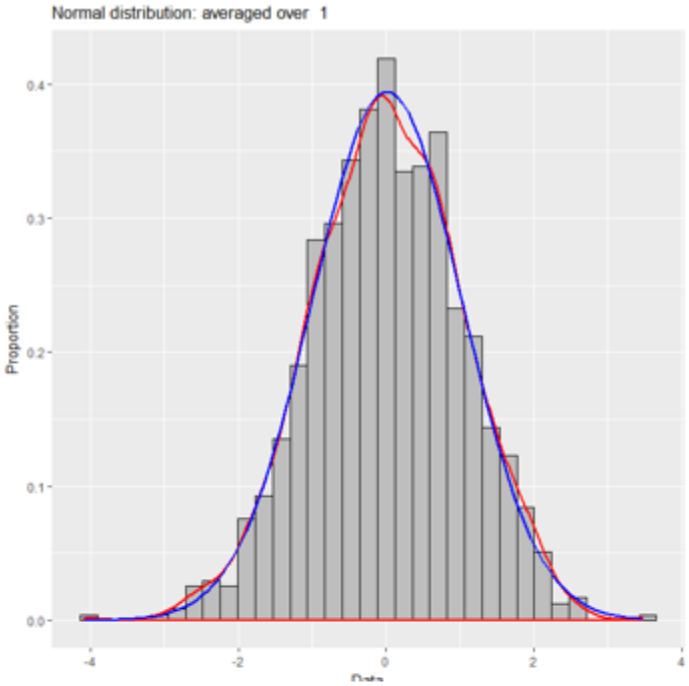
```
> mean(avg)
[1] 0.007291169
> sd(avg)
[1] 0.6995682
```

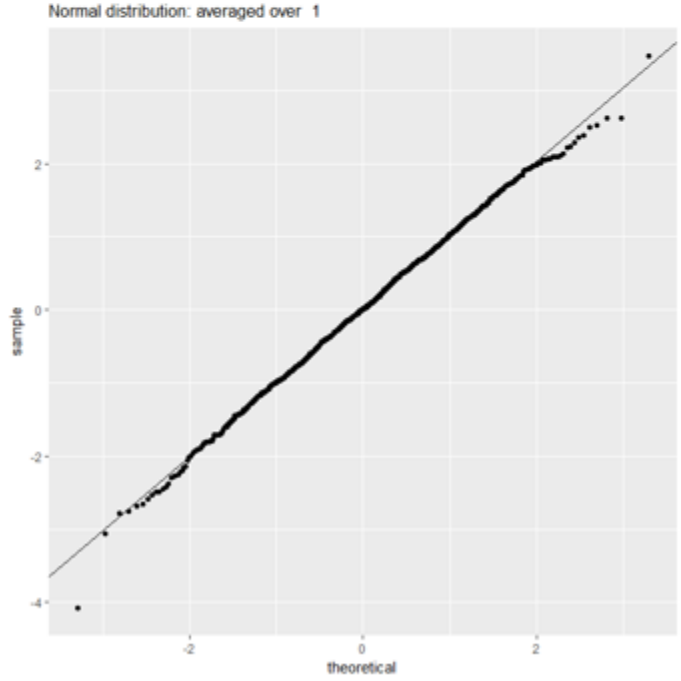
```
> mean(avg)
[1] -0.001516932
> sd(avg)
[1] 0.4117159
```

```
> mean(avg)
[1] -0.003300118
> sd(avg)
[1] 0.3249846
```

2.

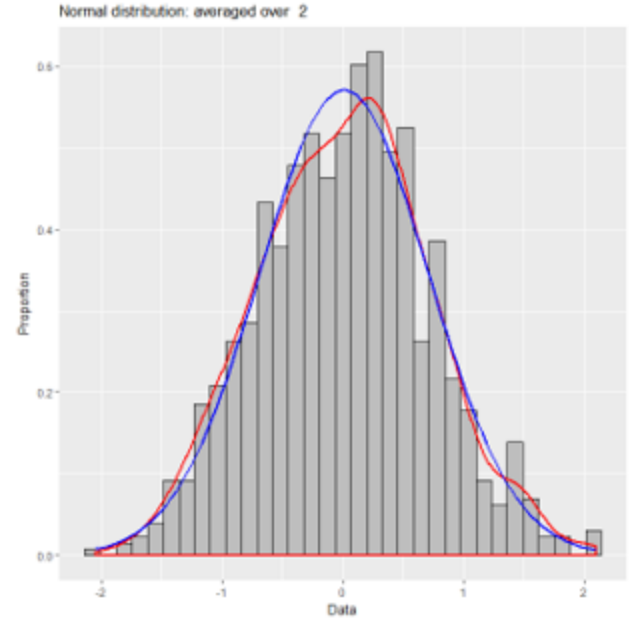
n=1

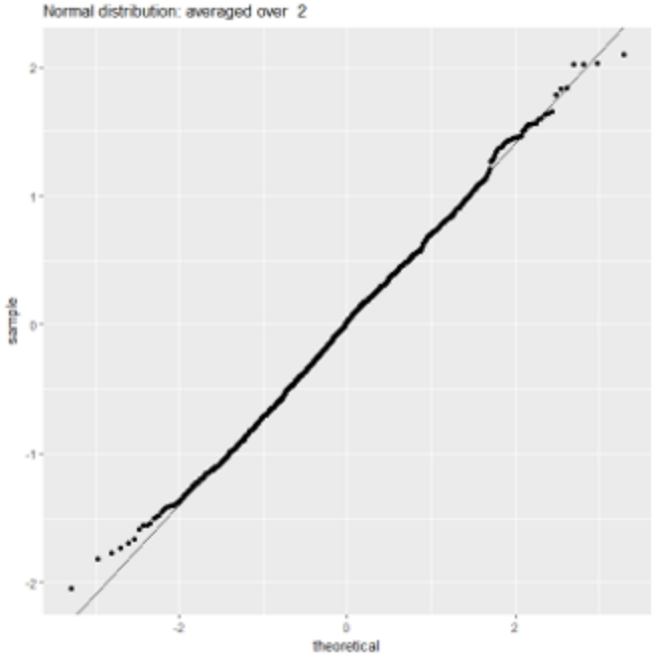




Both the graphs appear sufficiently normal.

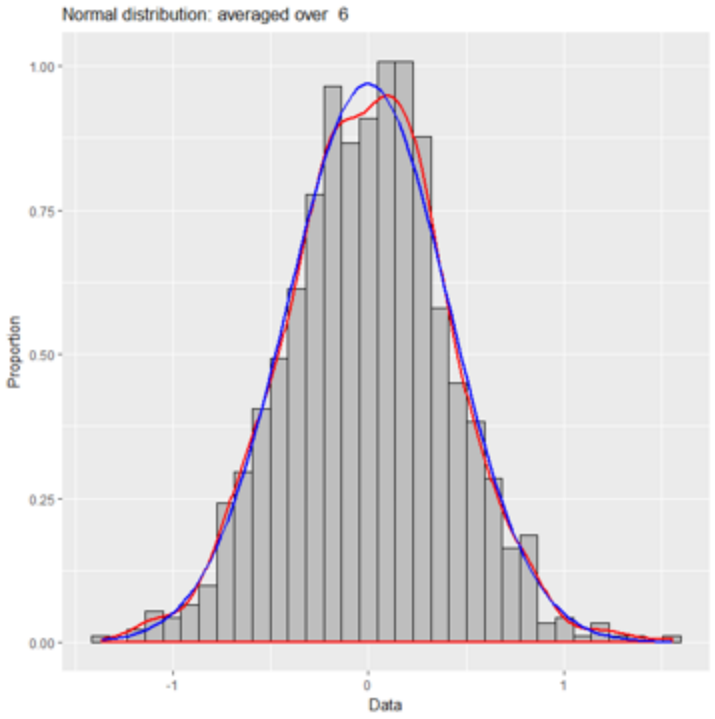
n=2

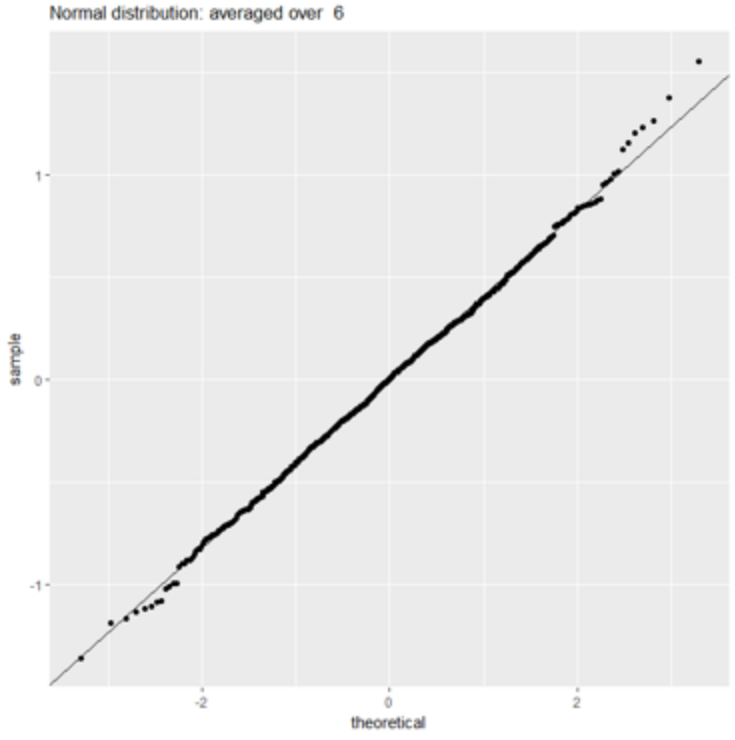




Both the graphs appear sufficiently normal.

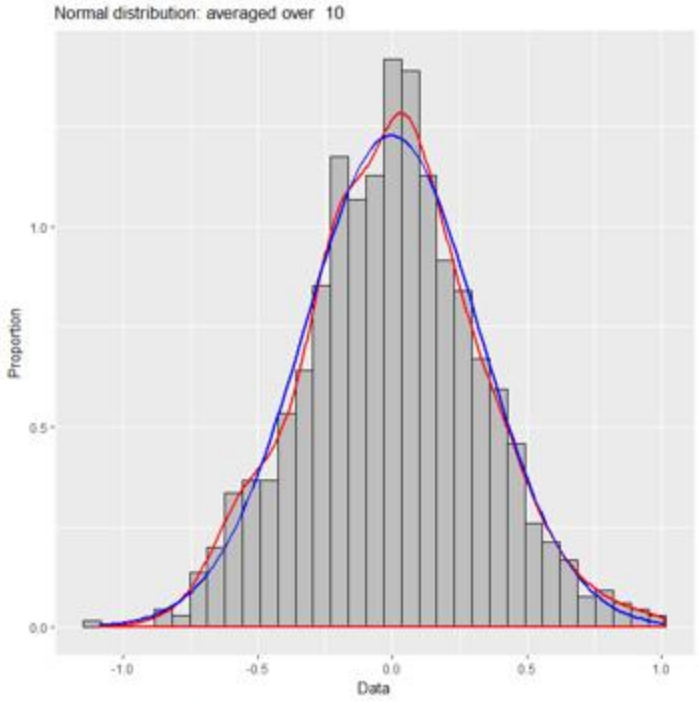
n=6

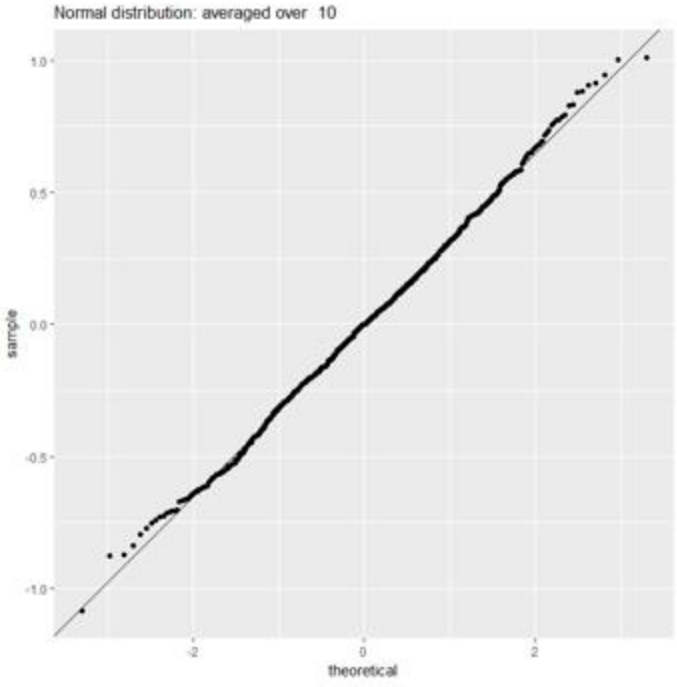




Both the graphs appear sufficiently normal.

n=10





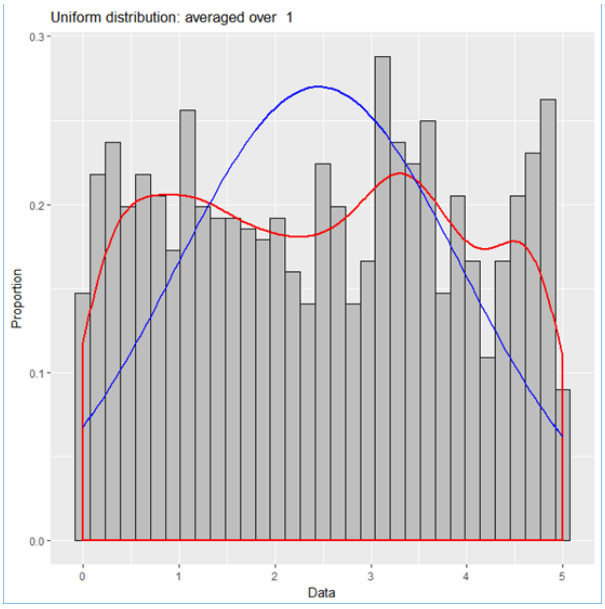
Both the graphs appear sufficiently normal.

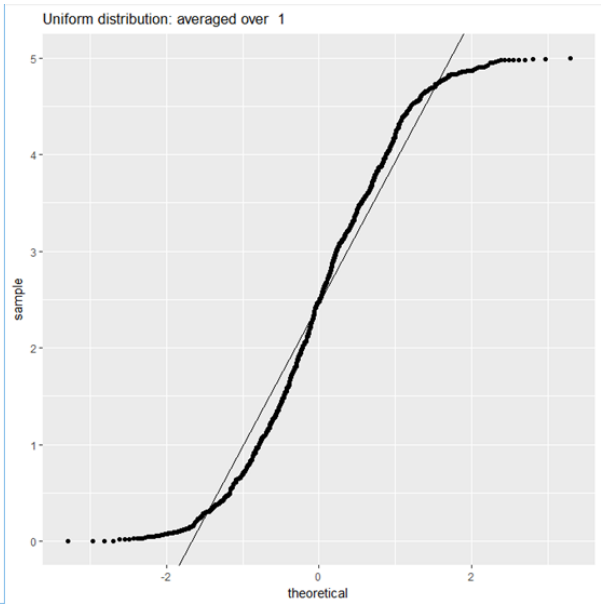
3.

n	Experimental mean of your 1000 $\bar{x}$ (from output)	Theoretical mean (Equations 1)	Experimental standard deviation of your 1000 $\bar{x}$ (from output)	Theoretical standard deviation (Equations 1)
1	0.0119638	0	1.011365	1
2	0.0072912	0	0.6995682	1/sqrt(2) = 0.7071068
6	-0.0015169	0	0.4117159	1/sqrt(6) = 0.4082483
10	-0.0033001	0	0.3249846	1/sqrt(10) = 0.3162278

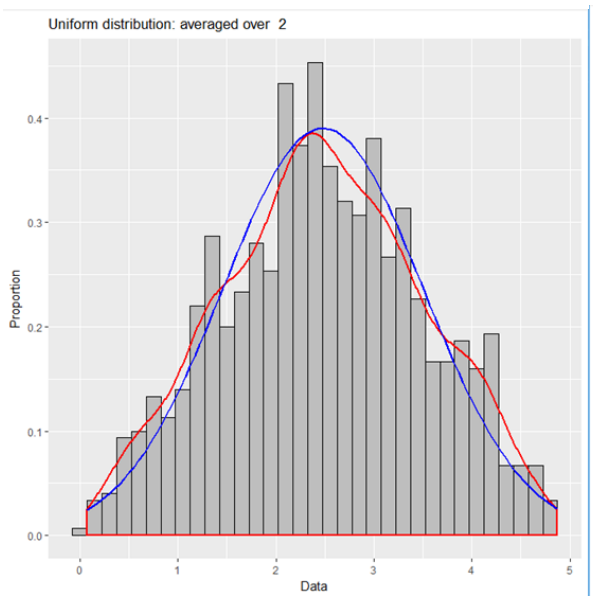
C. (20 points) Uniform distribution

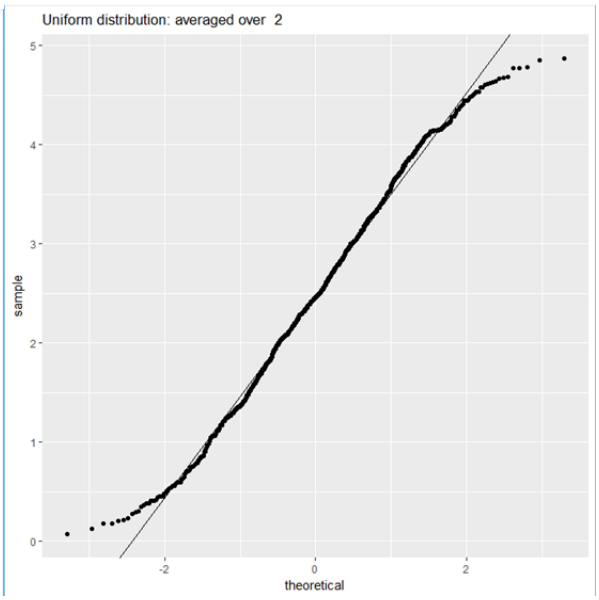
```
>  SRS <- 1000
> n <- 1
> title <- paste("Uniform distribution: averaged over ", n)
> data.vec <-runif(SRS*n,min=0, max = 5)
> data.mat <-matrix(data.vec, nrow = SRS)
> avg <- apply(data.mat,1,mean)
> mean(avg)
[1] 2.458981
> sd(avg)
[1] 1.477483
> xbar = mean(avg)
> s = sd(avg)
> library(ggplot2)
> windows()
> ggplot(data.frame(avg=avg), aes(x=avg)) +
+     geom_histogram(aes(y=..density..),bins = sqrt(length(avg))+2,
+                    fill = "grey",col = "black") +
+     geom_density(col = "red", lwd = 1) +
+     stat_function(fun=dnorm, args=list(mean=xbar, sd=s), col="blue",
+                   lwd = 1) +
+     ggtitle(title) +
+     xlab("Data") +
+     ylab("Proportion")
> ggplot(data.frame(avg=avg), aes(sample=avg)) +
+     stat_qq() +
+     geom_abline(slope = s, intercept = xbar) +
+     ggtitle(title)
```



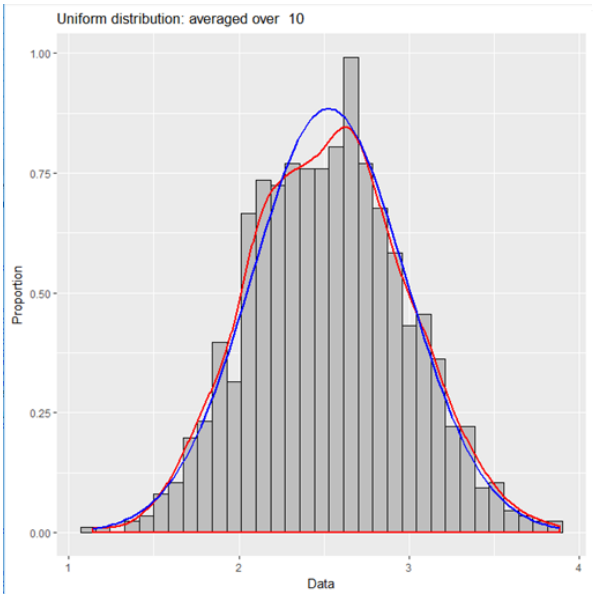


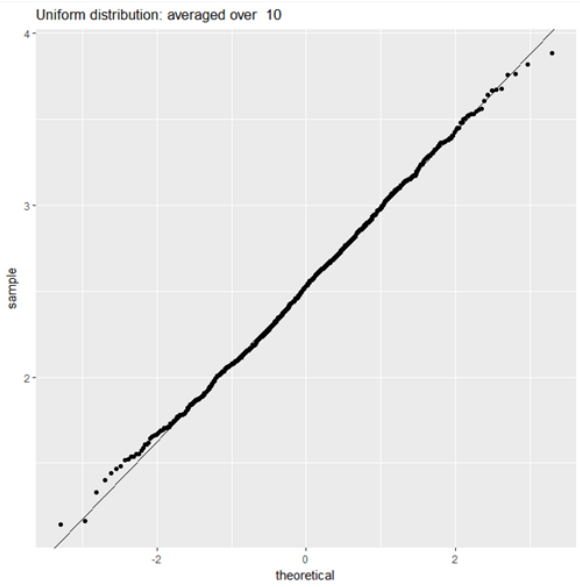
Both still look like uniform distribution.



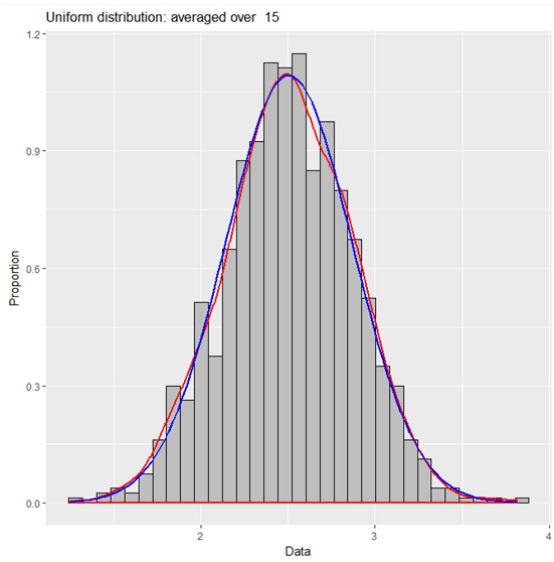


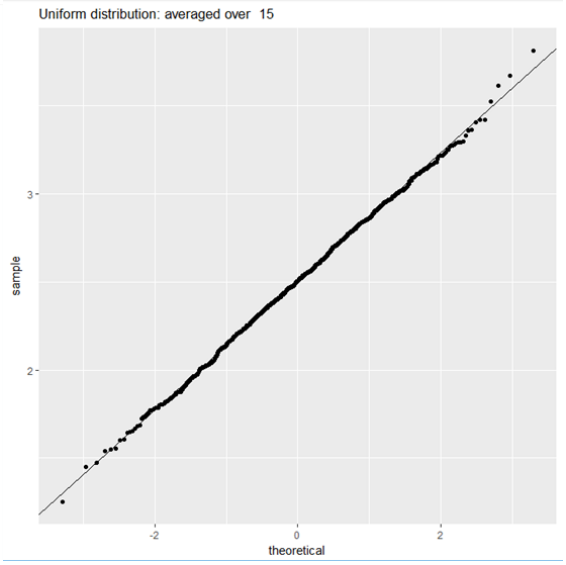
Both of the graphs are somewhere in between uniform and normal.





Graphs are close to uniform distribution.





Graphs are reasonably normal.

n	Experimental mean of your 1000 $\bar{x}$ (from output)	Theoretical mean (Equations 1)	Experimental standard deviation of your 1000 $\bar{x}$ (from output)	Theoretical standard deviation (Equations 1)
1	2.458981	(0+5)/2 =2. 5	1.477483	Sqrt((5-0)^2/12)/1=1.443375673
2	2.480496	2.5	1.022201	1.443375673/sqrt(2) = 1.020620726
10	2.526079	2.5	0.4509768	1.443375673/sqrt(10)=0.4564354646
15	2.504869	2.5	0.3652246	1.443375673/sqrt(15)=0.3726779962

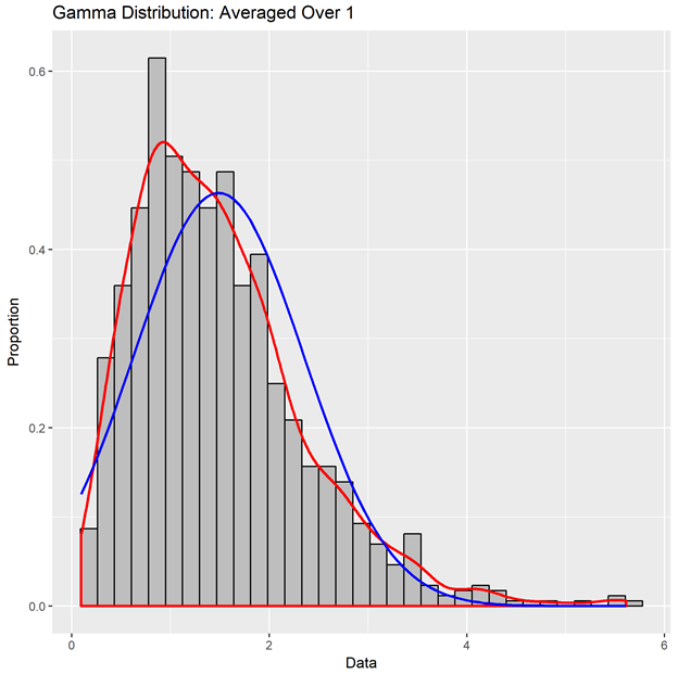
D. (20 points) Gamma distribution

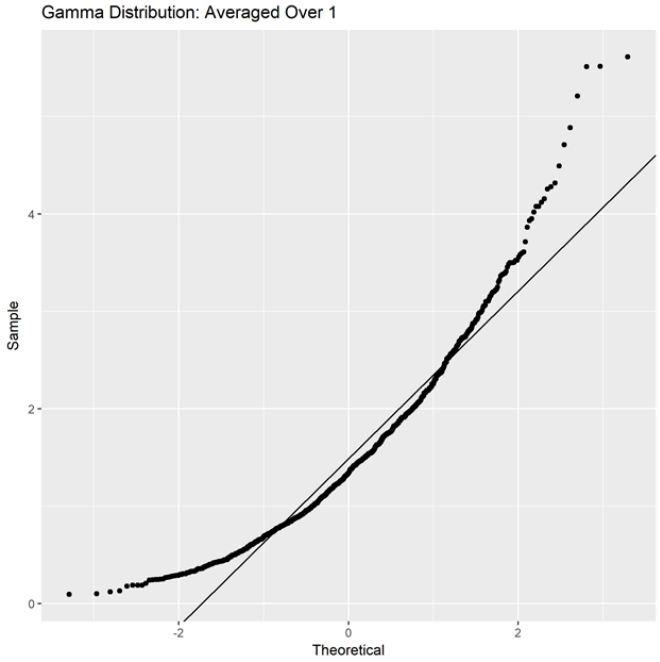
1. Code

```
# set working directory to Lab 04
setwd("C:/Users/jordan/Google Drive/Courses Spring 2018/STAT 350/STAT 350
Labs/Lab 04")
# set up ggplot2 for plotting
library(ggplot2)

###
# PART D: Gamma distribution with parameters alpha=3, beta=2
#
# For a gamma distribution with alpha = 3 and beta = 2,
# generate 1000 random samples of size n = 1, 5, 10, 20, 40, 60, 80, ...
#
# Then for each sample:
# find sample average and
# create histogram (with two colored curves) and
# normal probability plot of sample mean
# for each graph pair, indicate whether they appear sufficiently normal.
# NOTE: do not need to explain judgment of normality.
#
# Do this until n is large enough that distribution of sample mean
# appears normal.
###

samples <- 1000   # number of samples to collect
sizes <- c(1, 5, 10, 20, 40, 60, 80, 100)  # sample sizes

# create graph pair for each sample size
for (n in sizes) {
  title <- paste("Gamma Distribution: Averaged Over", n)  # title for graphs

  # generate data and calculate means
  gamma.vec <- rgamma(samples*n, 3, rate=2)  # random gamma data
  gamma.mat <- matrix(gamma.vec, nrow=samples)  # separate data into rows
  gamma.means <- apply(gamma.mat, 1, mean)  # calculate means

  # create histogram
  hist <- ggplot(data.frame(gamma.means=gamma.means),aes(gamma.means))+
    geom_histogram(aes(y=..density..),bins=sqrt(samples)+2,
                   fill="grey",col="black")+
    geom_density(col="red",lwd=1)+
    stat_function(fun=dnorm,args=list(mean=mean(gamma.means),
                                      sd=sd(gamma.means)),
```

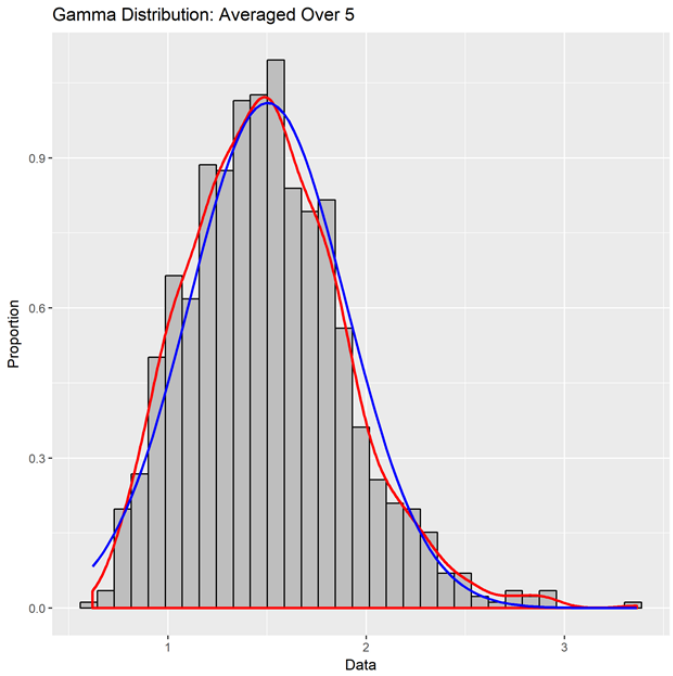
```
                  col="blue",lwd=1)+
    ggtitle(title)+
    xlab("Data")+
    ylab("Proportion")
  ggsave(hist, filename=paste("gammaHist",n,".png",sep=""))

  # create normal probability plot
  qq <- ggplot(data.frame(gamma.means=gamma.means),aes(sample=gamma.means))+
      stat_qq()+
    geom_abline(slope=sd(gamma.means),intercept=mean(gamma.means))+
    ggtitle(title)+
    xlab("Theoretical")+
    ylab("Sample")
  ggsave(qq, filename=paste("gammaQQ",n,".png",sep=""))
  print(paste("n = ", n))
  print(paste("mean = ", mean(gamma.means)))
  print(paste("sd = ", sd(gamma.means)))
}
```

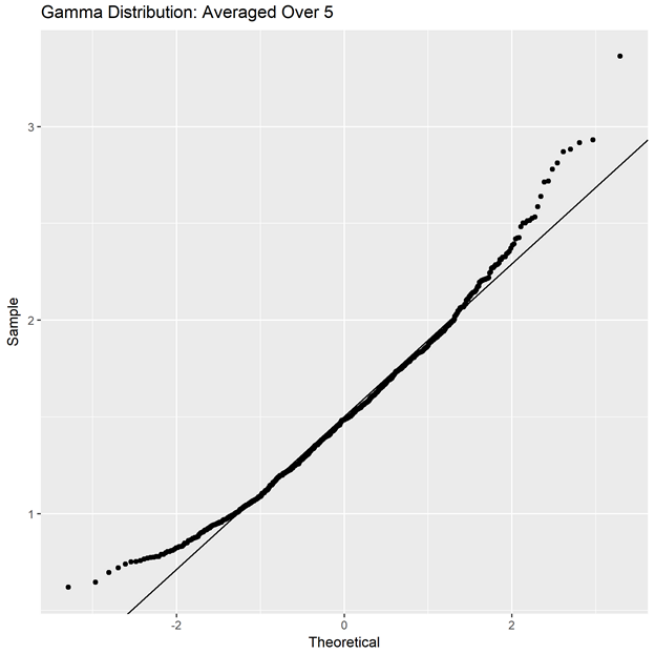
2. Histogram/normal probability plots



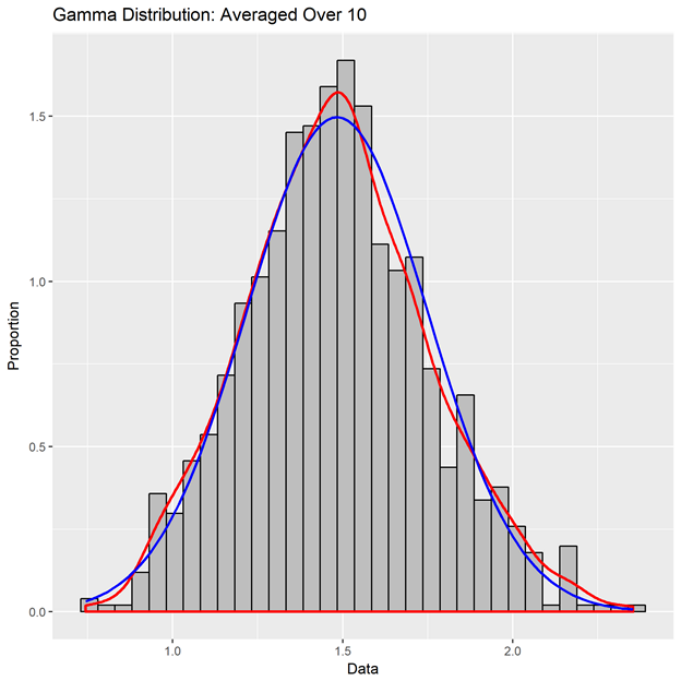


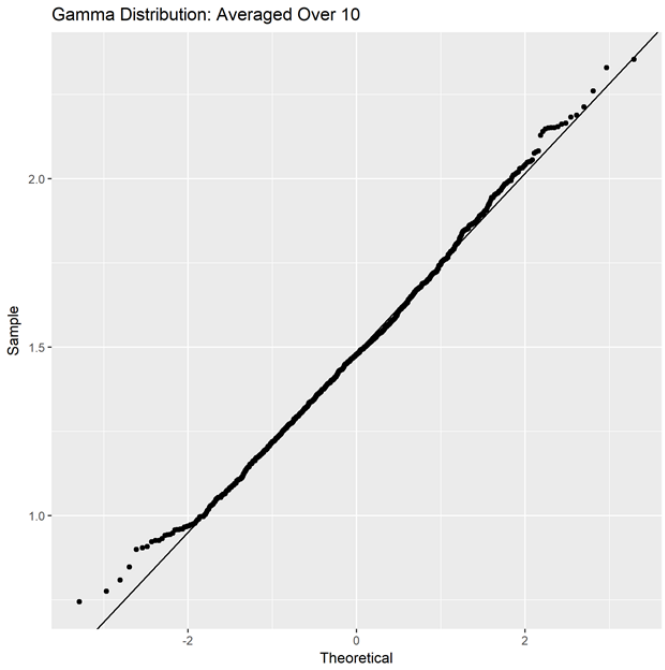
Not sufficiently normal



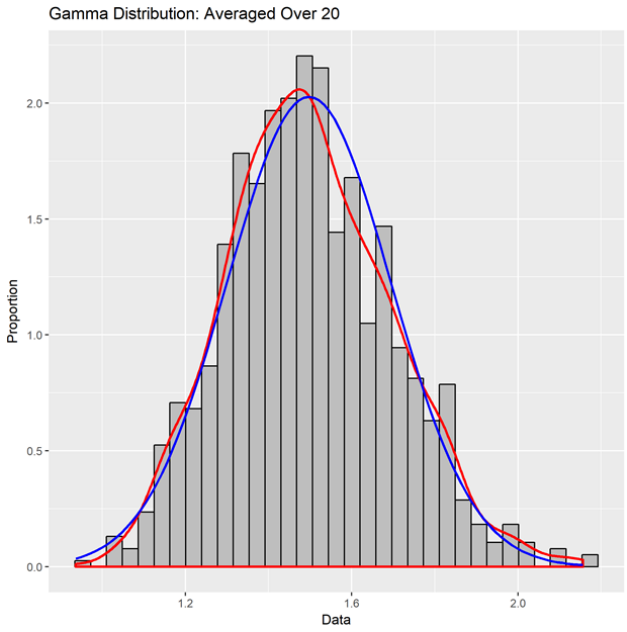


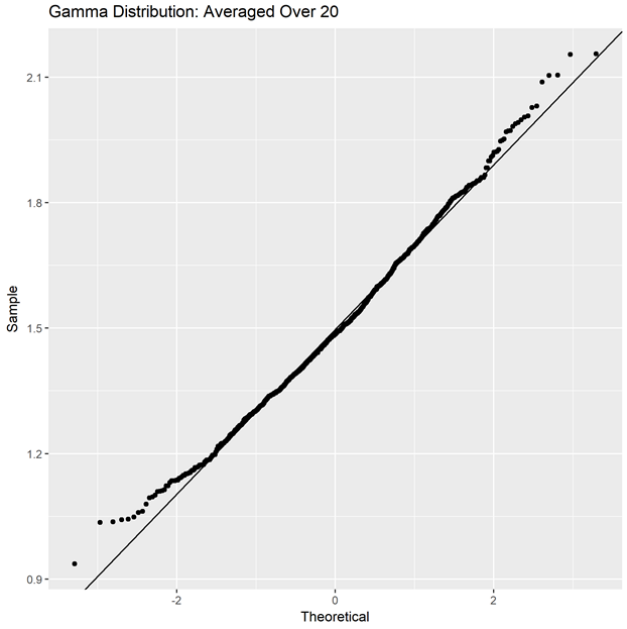
Not sufficiently normal



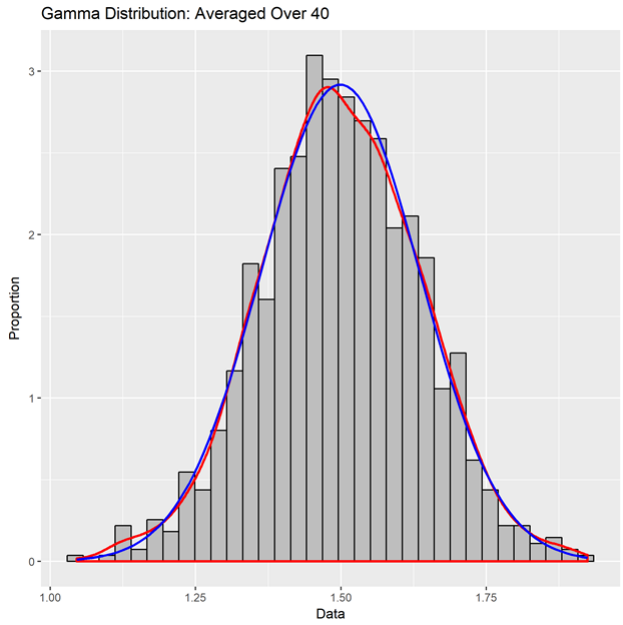


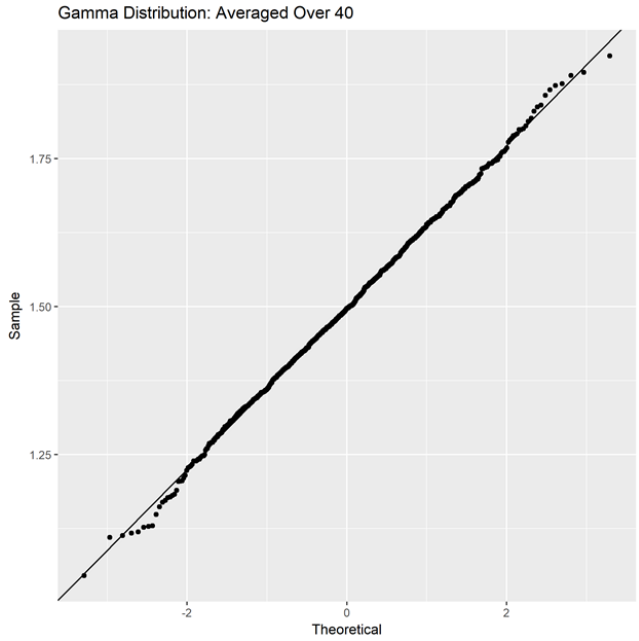
Not sufficiently normal



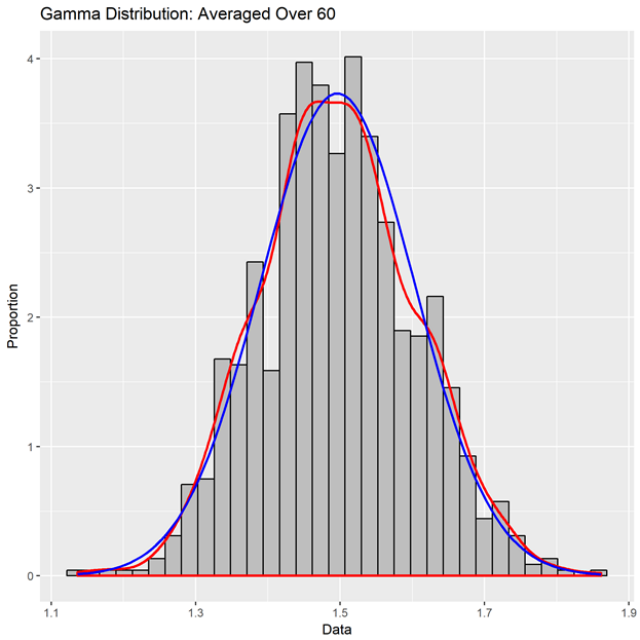


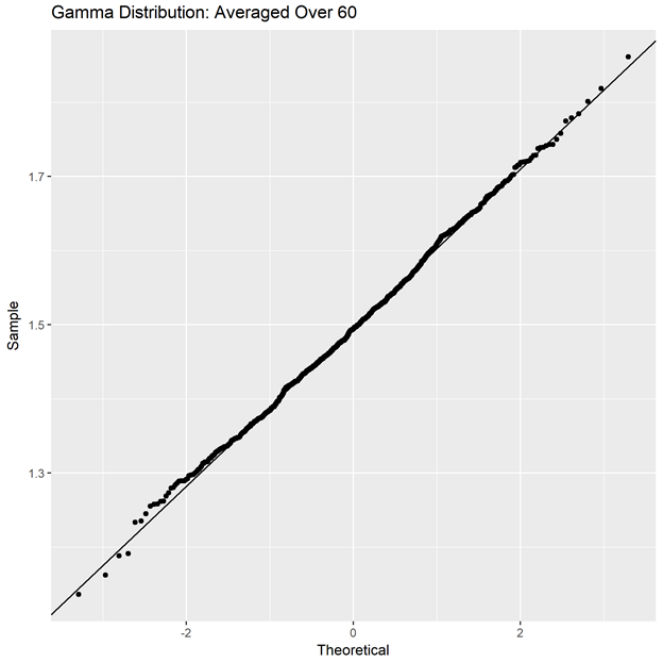
Not sufficiently normal





Not sufficiently normal, but pretty close





Approximately normal

3. Summary table

n	**experimental mean of 1000 (from output)**	**theoretical mean (given)**	**experimental standard deviation of 1000 (from output)**	**theoretical standard deviation (given and equation)**
1	1.4934	1.5	0.863	sqrt(3)/2 = 0.8660
5	1.4946	1.5	0.383	0.8660/sqrt(5) = 0.3873
10	1.5076	1.5	0.2721	0.8660/sqrt(10) = 0.2739
20	1.4951	1.5	0.1899	0.8660/sqrt(20) = 0.1936
40	1.5027	1.5	0.1354	0.8660/sqrt(40) = 0.1369
60	1.5018	1.5	0.1169	0.8660/sqrt(60) = 0.1118

E. (20 points) Poisson distribution

1.

```
#Set WD & Initialie ggplot
setwd("~/Desktop/School Work/Purdue/Spring 2018/Stat 350/Labs/Lab 4")
library(ggplot2) #Run for each session

#E-Poissons Dist
SRS <- 1000 # number of samples

n<-1 #1, 5, 10, 20, 40, and continue in intervals of 20 if needed until the shape becomes normal
title_E <- paste("Poisson Distribution: Averaged over", n)

#Calculates average data sample
data.vec <- rpois(SRS*n, 3) #creates random data for Poisson
data.mat <- matrix(data.vec, nrow = SRS) #separates the data into rows
Avg_E <- apply(data.mat, 1, mean)

#Q2(Histogram)
xbar_E <- mean(Avg_E)
s_E <- sd(Avg_E)

quartz()
ggplot(data.frame(Avg_E=Avg_E), aes(x=Avg_E)) +
  geom_histogram(aes(y=..density..),bins = sqrt(length(Avg_E))+2,
                 fill = "grey",col = "black") +
  geom_density(col = "red", lwd = 1) +    #Density is blue
  stat_function(fun=dnorm, args=list(mean=xbar_E, sd=s_E), col="blue",   #Normal function is blue
                lwd = 1) +
  ggtitle(title_E) +
  xlab("Data") +
  ylab("Proportion")
#Q2B(Normal Probability Plot)
quartz()
ggplot(data.frame(Avg_E=Avg_E), aes(sample=Avg_E)) +
  stat_qq() +
  geom_abline(slope = s_E, intercept = xbar_E) +
  ggtitle(title_E)

n
xbar_E
s_E
```

```
> n
[1] 1
> xbar_E
[1] 3.093
> s_E
[1] 1.783124
```

```
> n
[1] 5
> xbar_E
[1] 2.9954
> s_E
[1] 0.7649078
```

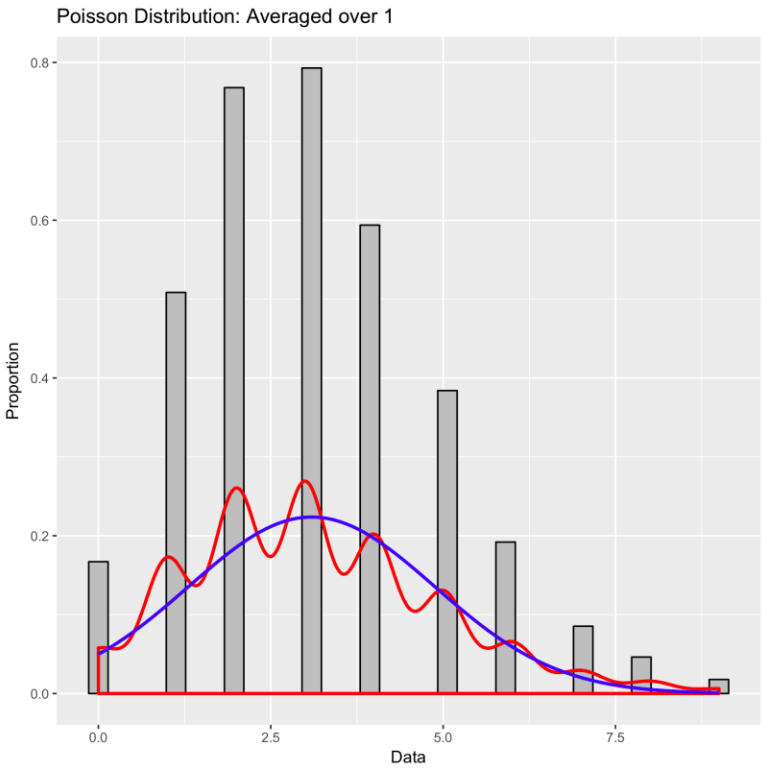
```
> n
[1] 10
> xbar_E
[1] 3.0166
> s_E
[1] 0.5572028
```

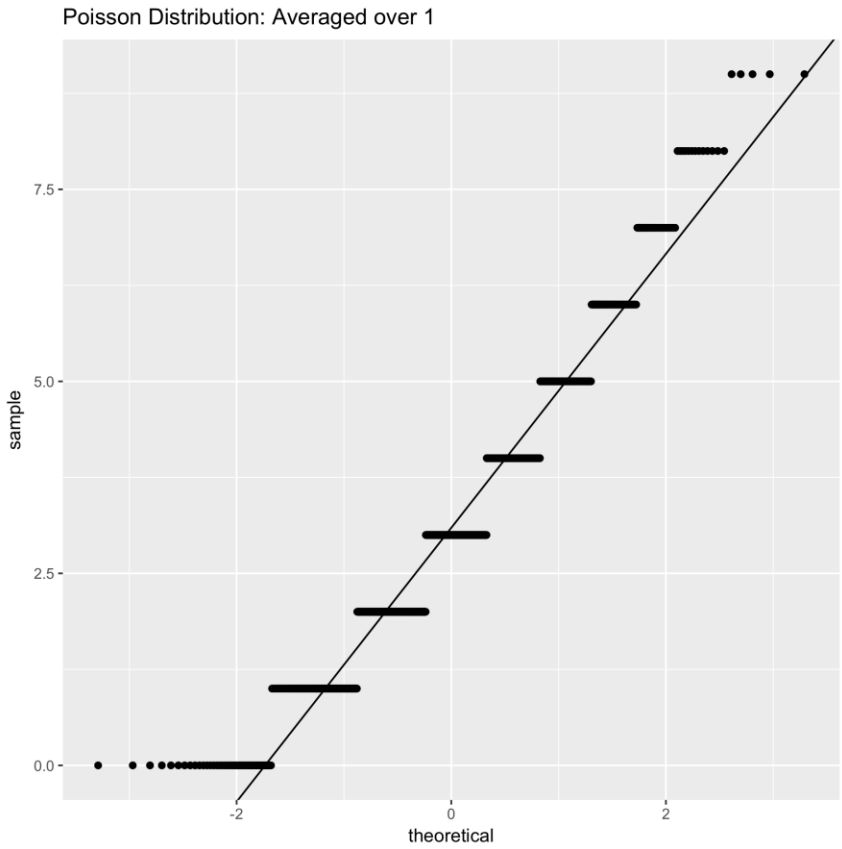
```
> n
[1] 20
> xbar_E
[1] 3.0236
> s_E
[1] 0.3916266
```

```
> n
[1] 40
> xbar_E
[1] 3.0053
> s_E
[1] 0.2748954
```

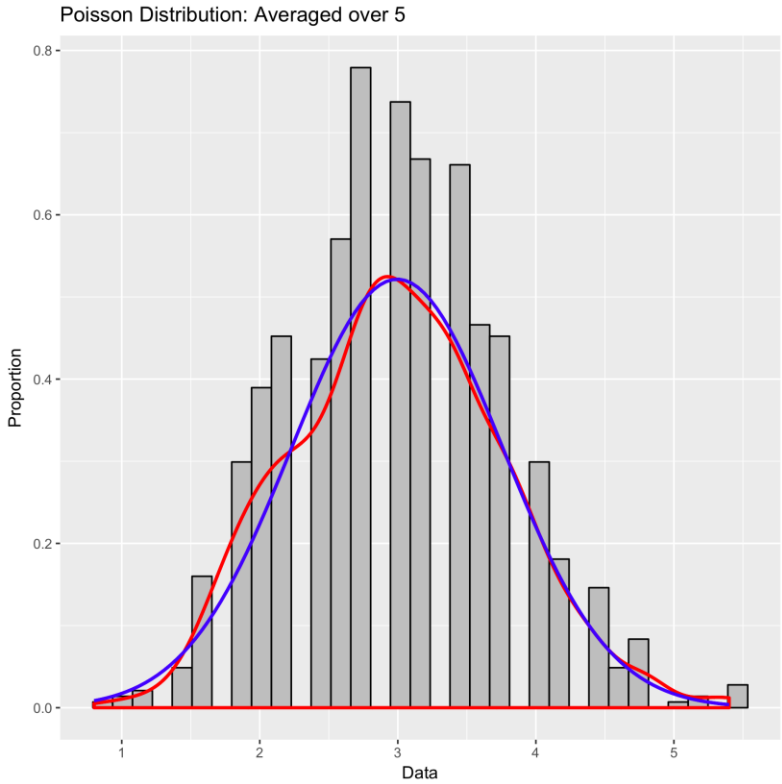
```
> n
[1] 60
> xbar_E
[1] 2.9877
> s_E
[1] 0.2201791
```

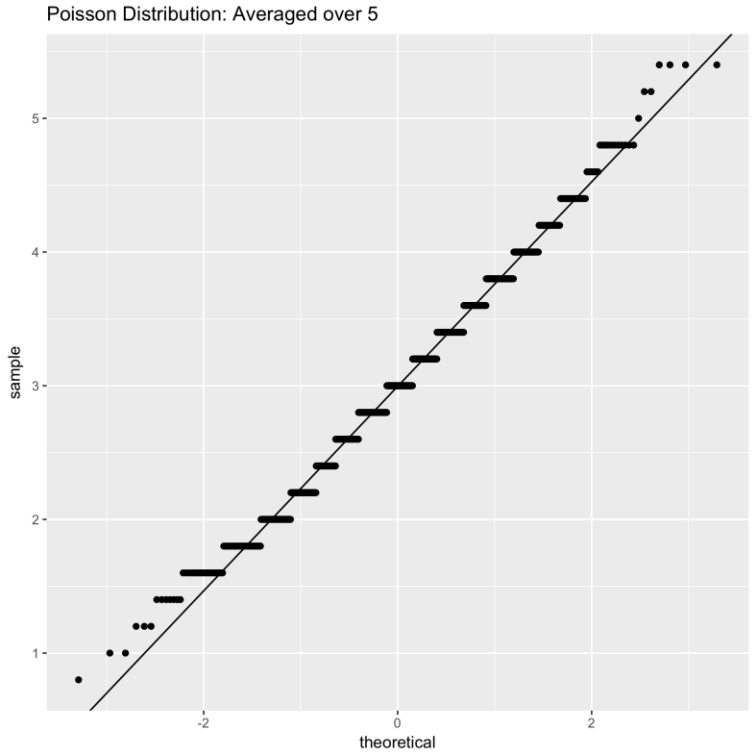
n=1





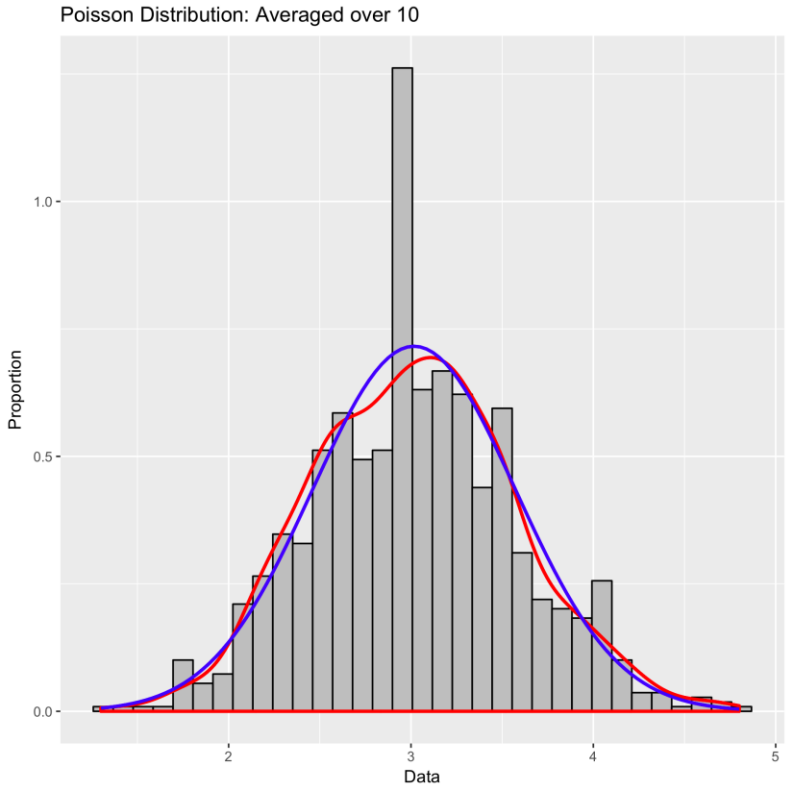
Not sufficiently normal, the QQ plot is very abnormal
n=5

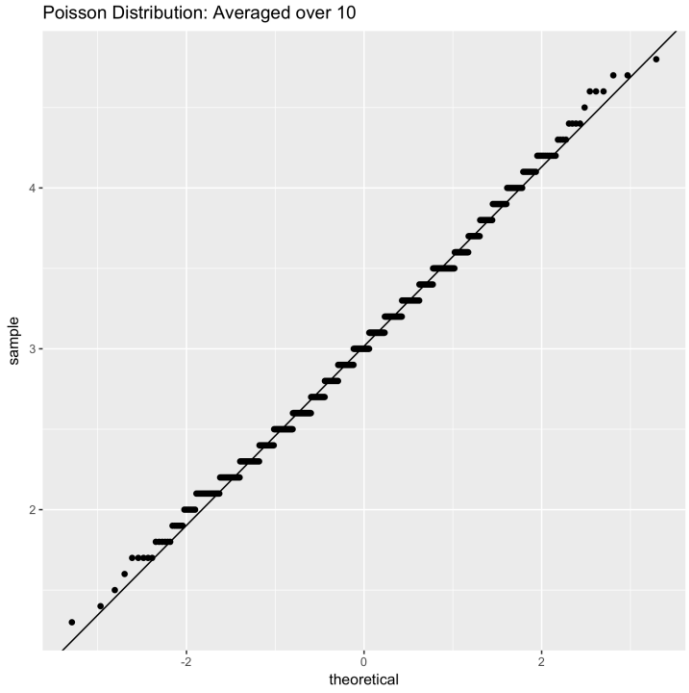




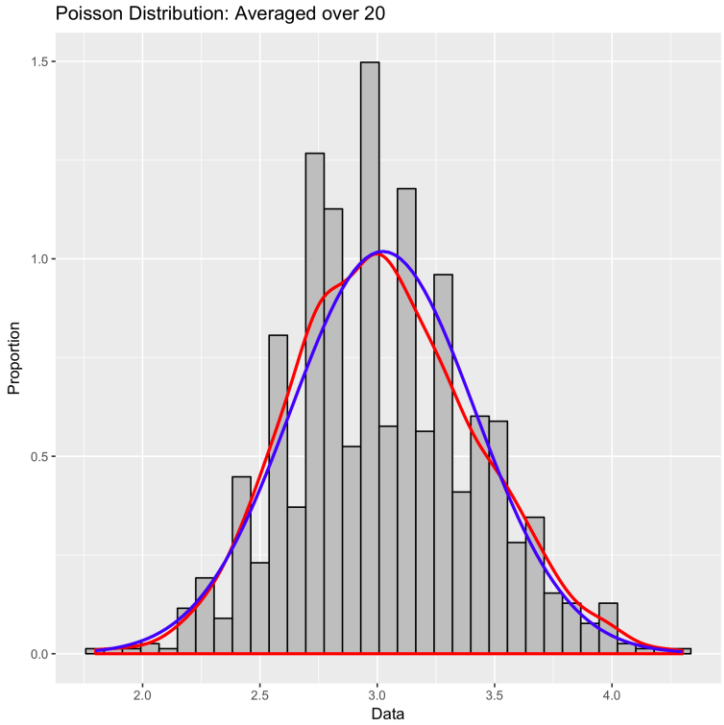
Not sufficiently normal, the red and blue lines are close but the histogram has spaces, the QQ plot is abnormal it has many horizontal steps.

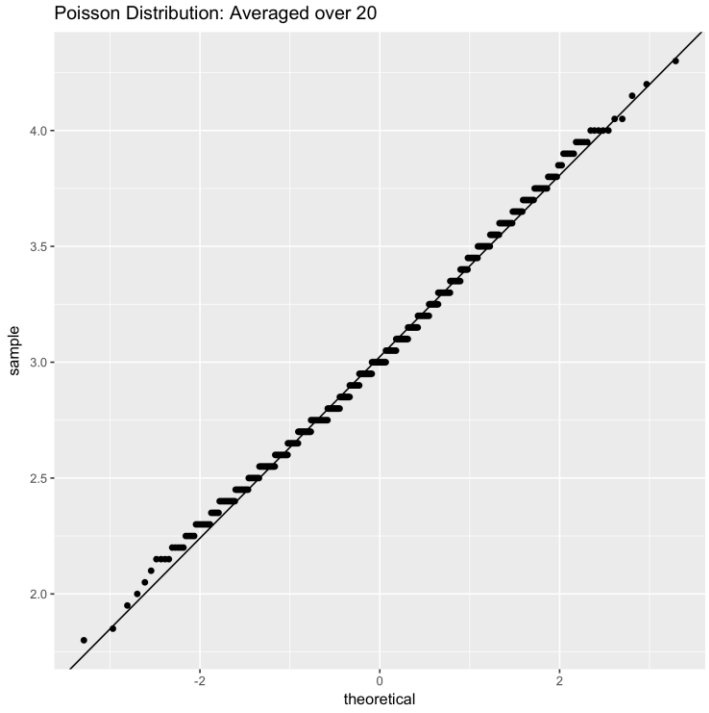
n=10



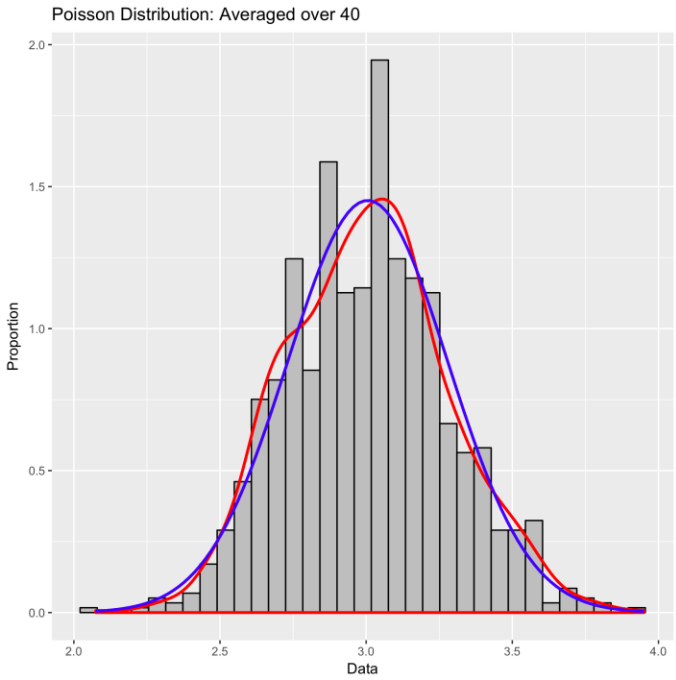


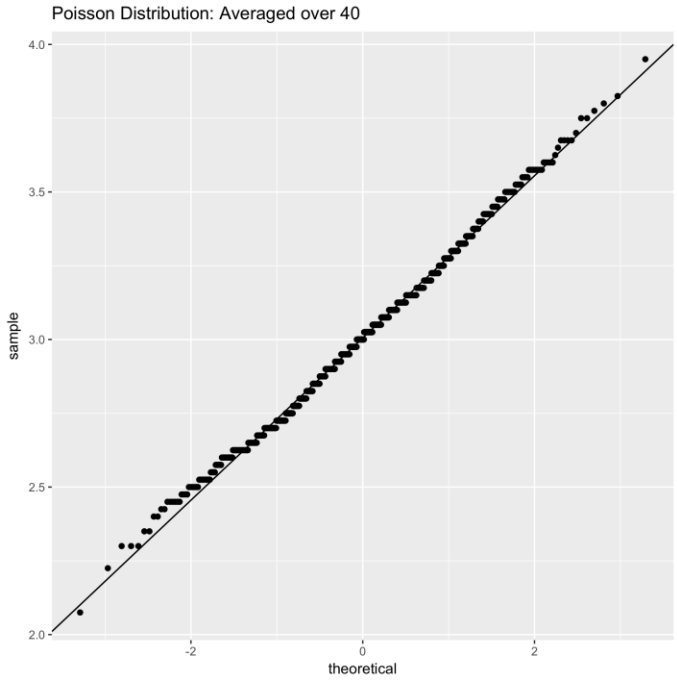
Not sufficiently normal, the red and blue lines are close but the histogram has one abnormally large bar in the middle, the QQ plot is very abnormal it has many horizontal steps which are shorter than before.
n=20





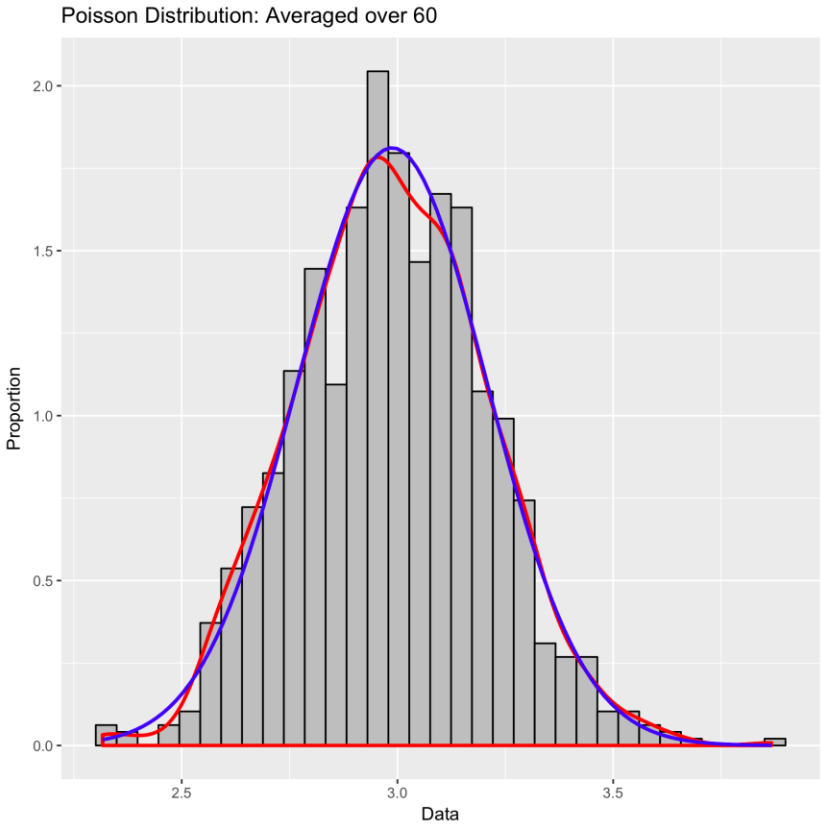
Not sufficiently normal, the red and blue lines are close but the histogram has alternating high and low bars, the QQ plot is very abnormal it has many horizontal steps with the data points more separated on either side of the line.
n=40

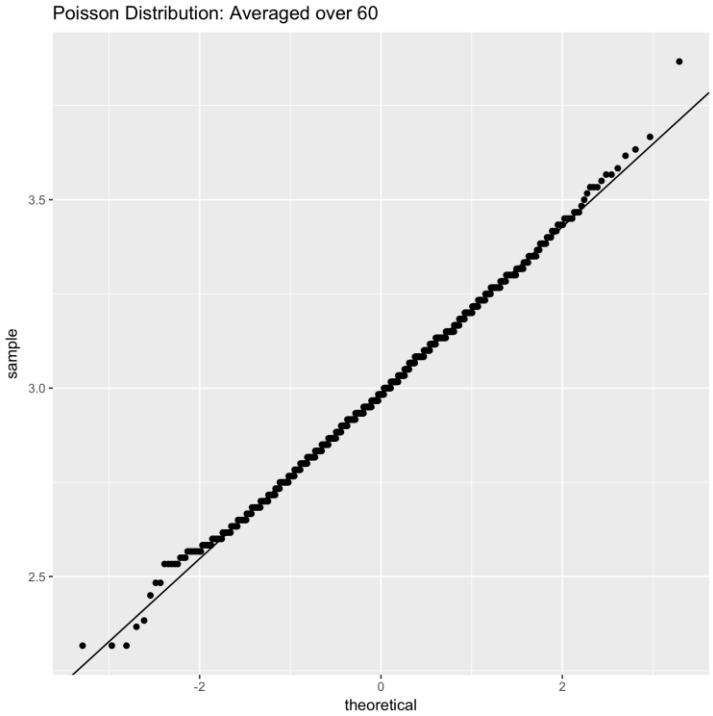




Both the QQ plot and histogram look normal, with a slight variation from normal in the histogram and some spaces between horizontal steps still in the QQ plot.

n=60





Both the QQ plot and histogram look normal

n	experimental mean of your 1000 (from output)	theoretical mean (Equations 1)	experimental standard deviation of your 1000 (from output)	theoretical standard deviation (Equations 1)
1	3.093	3	1.783124	1.732
5	2.9954	$=\lambda=3$	0.7649078	$=\frac{\sqrt{3}}{\sqrt{5}}=0.7746$
10	3.0166	3	0.5572028	0.5477
20	3.0236	3	0.3916266	0.3873
40	3.0053	3	0.2748954	0.2739
60	2.9877	3	0.2201791	0.2236

For Poisson Distribution:

$\mu_{\square} = \square = \square_{\underline{\square}}$ $\sigma_{\square} = \sqrt{\square}, \square\square\ \square_{\underline{\square}} = \frac{\sqrt{\lambda}}{\sqrt{\square}}$

F. (BONUS: 20 points) Exponential distribution

1.

```
> SRS <- 1000
> n <- 1
> exp <- paste("Exponential distribution: averaged over ", n)
> data.vec  <-rexp(SRS*n, 3)
> data.mat <- matrix(data.vec, nrow = SRS)
> avg <- apply(data.mat, 1, mean)
```

```
> mean(avg)
[1] 0.3441099
> sd(avg)
[1] 0.330395
```

```
> library(ggplot2)
Warning message:
package 'ggplot2' was built under R version 3.4.3
> windows()
> xbar <- mean(avg)
> s <- sd(avg)
> ggplot(data.frame(avg=avg), aes(x=avg)) +
+     geom_histogram(aes(y=..density..),bins = sqrt(length(avg))+2,
+                    fill = "grey",col = "black") +
+     geom_density(col = "red", lwd = 1) +
+     stat_function(fun=dnorm, args=list(mean=xbar, sd=s), col="blue",
+                   lwd = 1) +
+     ggtitle(exp) +
+     xlab("Data") +
+     ylab("Proportion")
> ggplot(data.frame(avg=avg), aes(sample=avg)) +
+     stat_qq() +
+     geom_abline(slope = s, intercept = xbar) +
+     ggtitle(exp)
```

Mean and SD for n = 5,10,20,40,60,80 :

n=5

```
> mean(avg)
[1] 0.3317099
> sd(avg)
[1] 0.1491399
```

n=20

```
> mean(avg)
[1] 0.3398453
> sd(avg)
[1] 0.07585229
```

n=60

```
> mean(avg)
[1] 0.3323412
> sd(avg)
[1] 0.04330211
```

n=10

```
> mean(avg)
[1] 0.333619
> sd(avg)
[1] 0.1033288
```

n=40

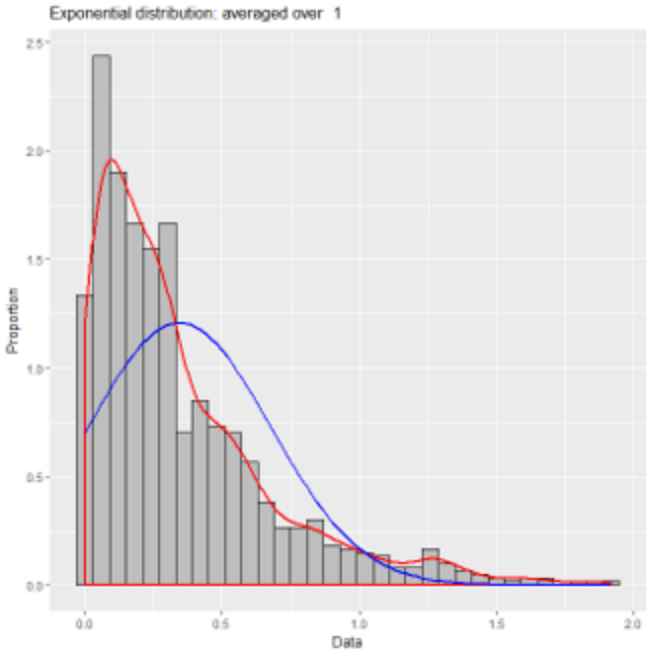
```
> mean(avg)
[1] 0.3328251
> sd(avg)
[1] 0.0535324
```

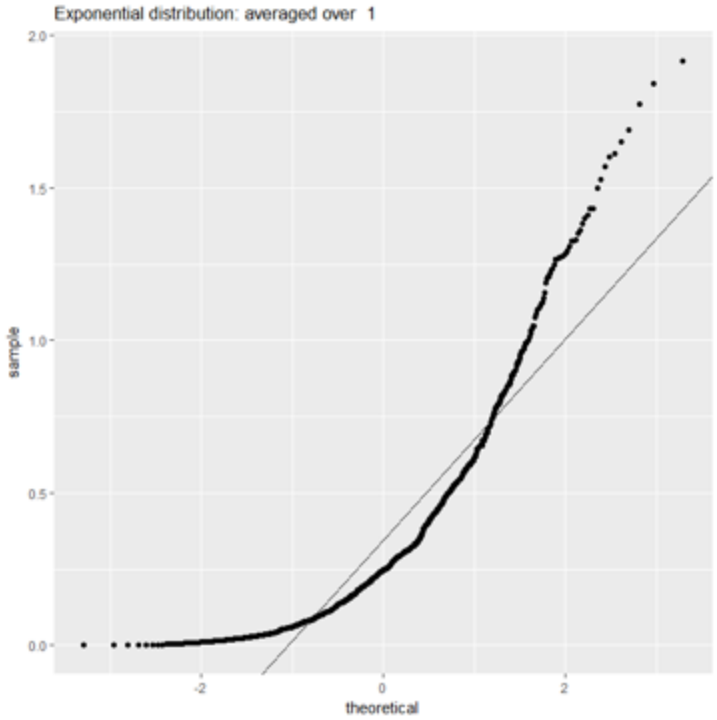
n=80

```
> mean(avg)
[1] 0.3333074
> sd(avg)
[1] 0.03673066
```

2.

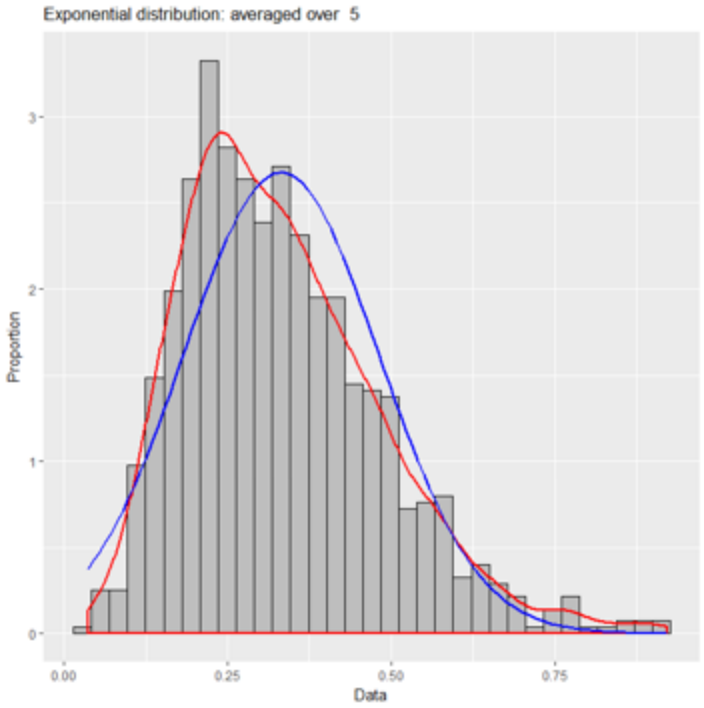
n=1

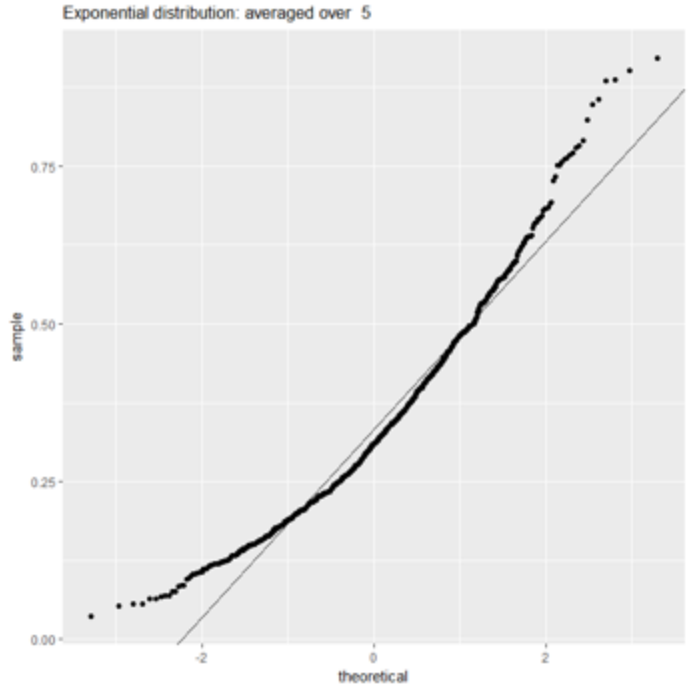




The graphs do not appear sufficiently normal.

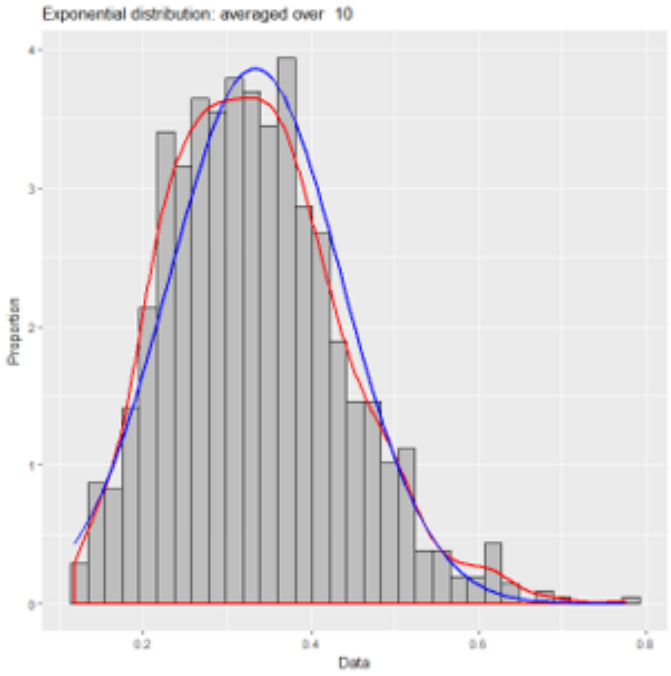
n=5

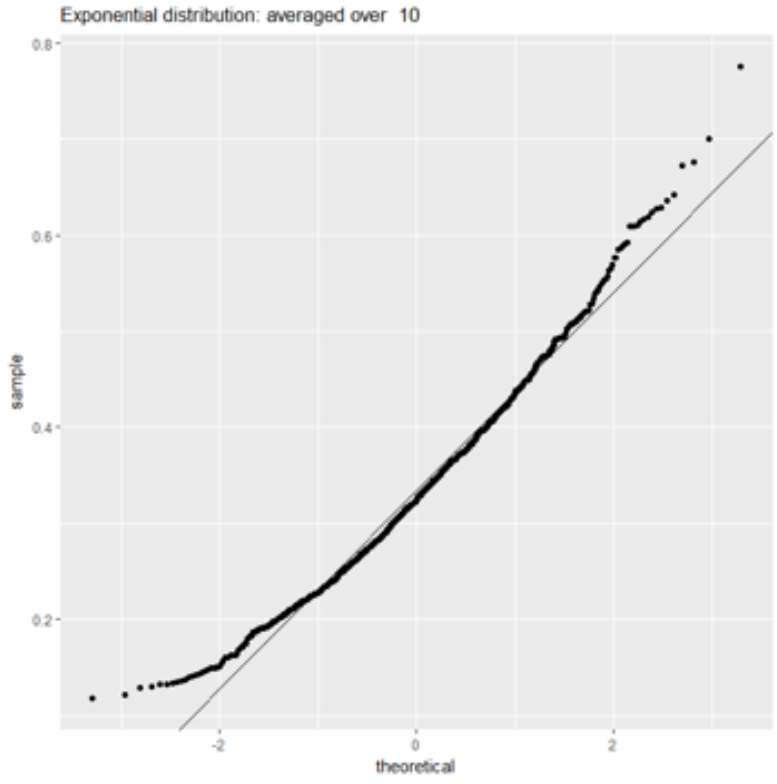




The graphs do not appear sufficiently normal.

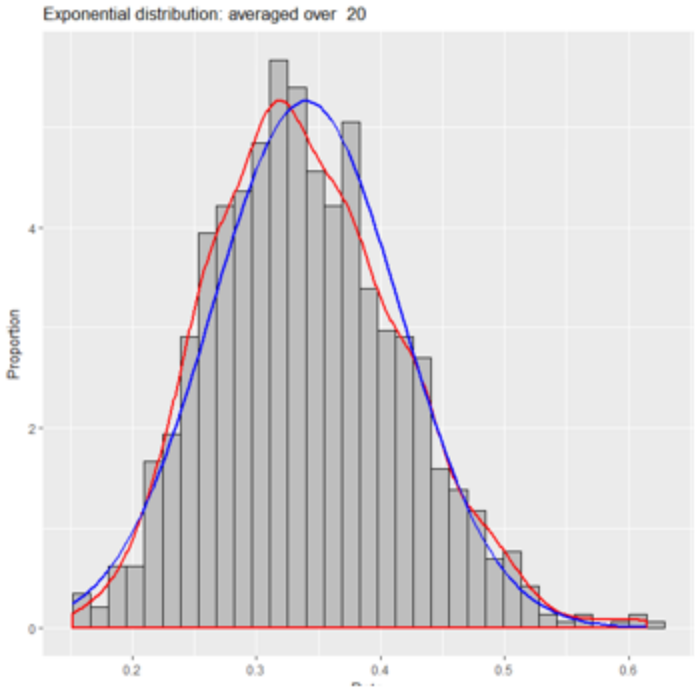
n=10

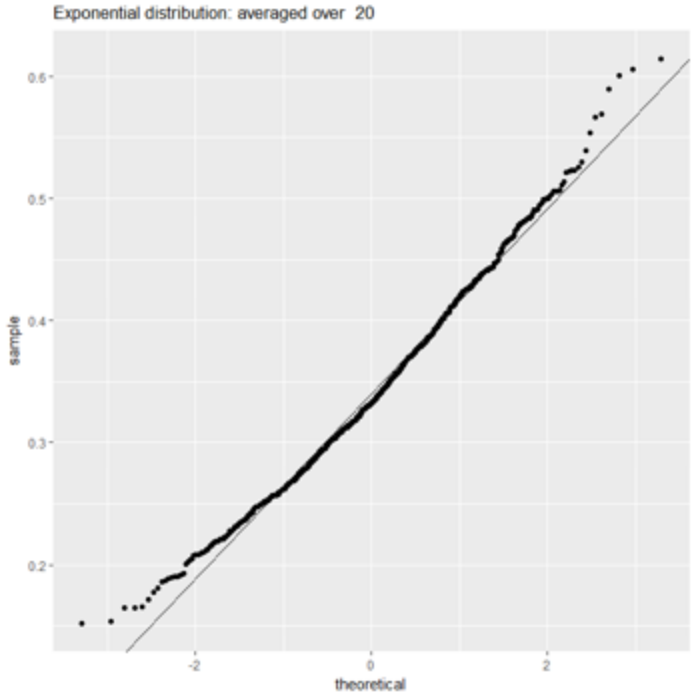




The graphs do not appear sufficiently normal.

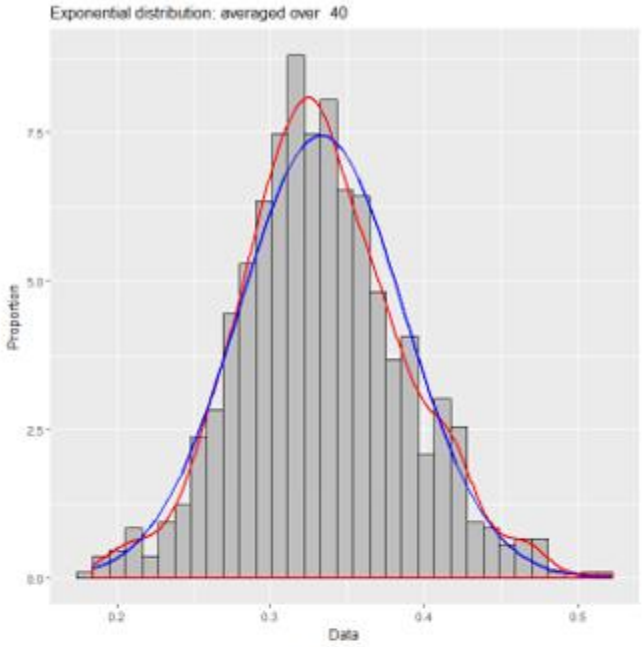
n=20

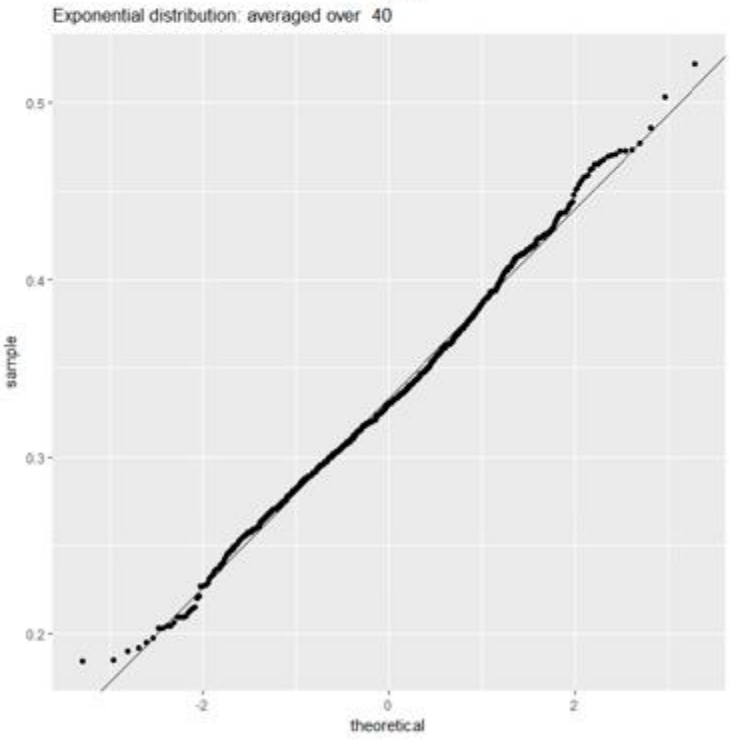




The graphs do not appear sufficiently normal.

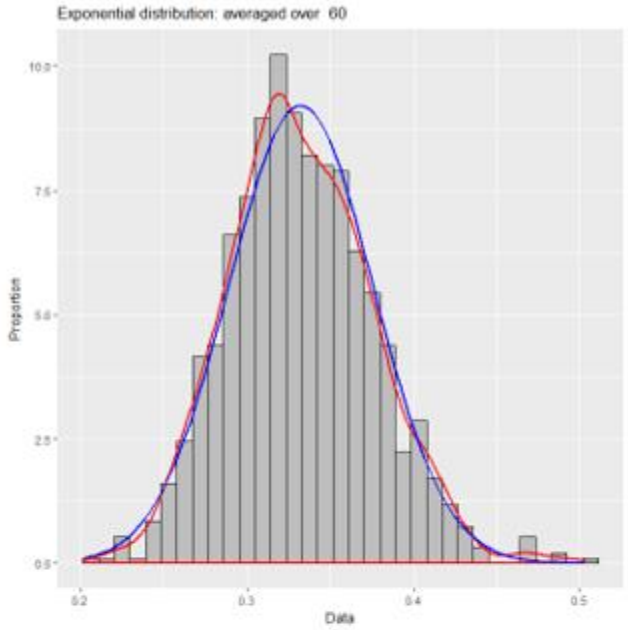
n=40

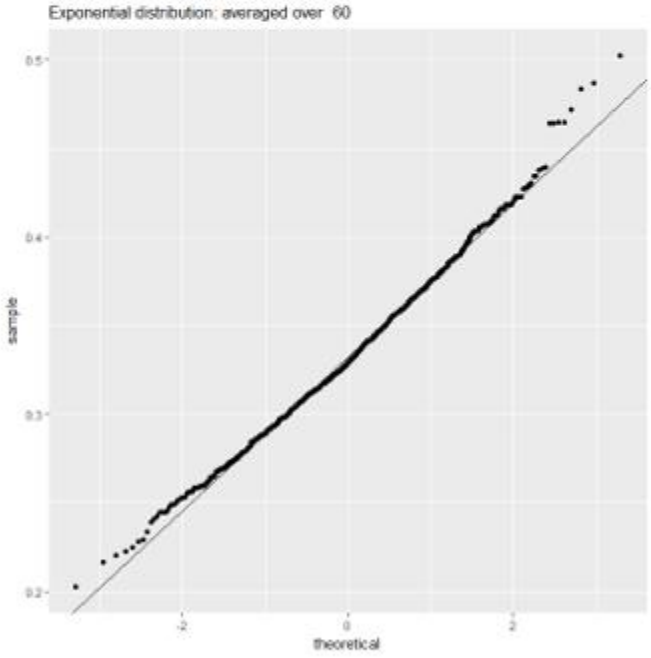




The graph appears slightly normal.

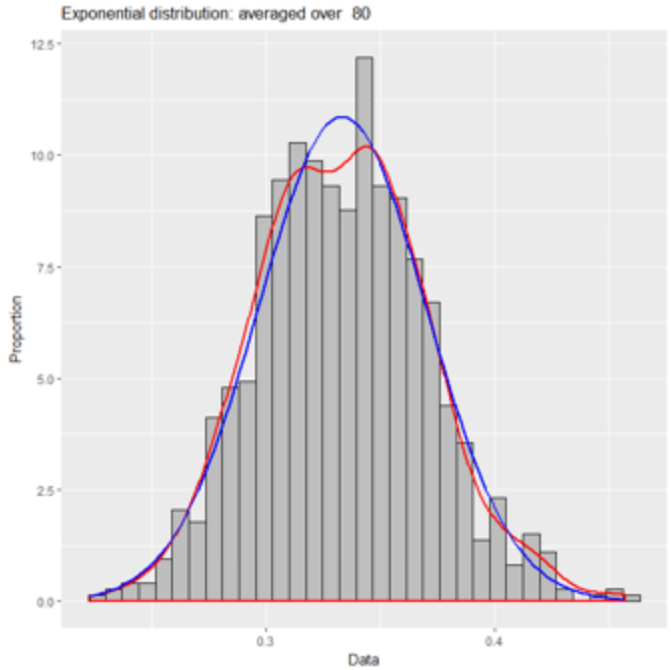
n=60

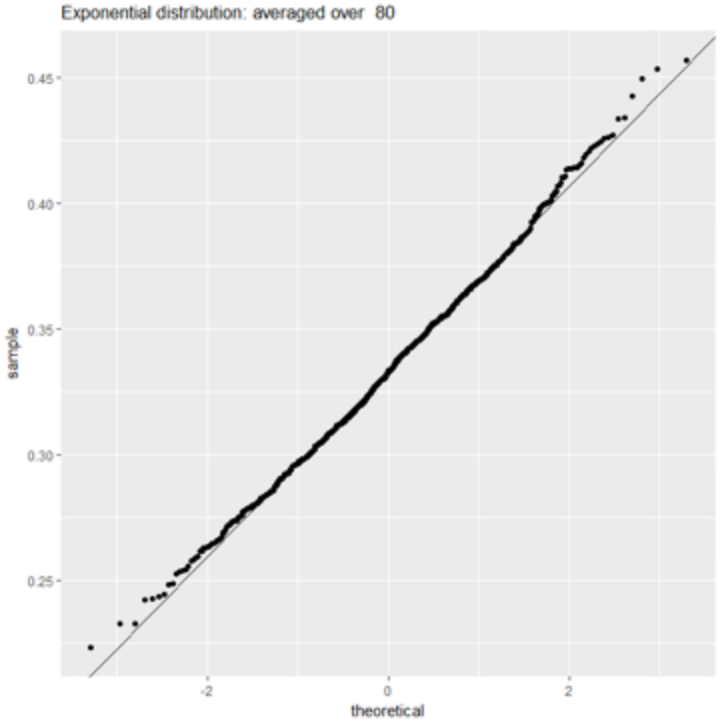




The graphs appears somewhat sufficiently normal.

n=80





The graphs appear sufficiently normal.

3.

n	Experimental mean of your 1000 $\bar{x}$ (from output)	Theoretical mean (Equations 1)	Experimental standard deviation of your 1000 $\bar{x}$ (from output)	Theoretical standard deviations (Equations 1)
1	0.3441099	1/3 = 0.3333333	0.3303950	1/3 = 0.3333333
5	0.3317099	1/3 = 0.3333333	0.1491399	(1/3)/sqrt(5) = 0.1490712
10	0.3336190	1/3 = 0.3333333	0.1033288	(1/3)/sqrt(10) = 0.1054093
20	0.3398453	1/3 = 0.3333333	0.0758523	(1/3)/sqrt(20) = 0.0745356
40	0.3328251	1/3 = 0.3333333	0.0535324	(1/3)/sqrt(40) = 0.0527046
60	0.3323412	1/3 = 0.3333333	0.0433021	(1/3)/sqrt(60) = 0.0430331
80	0.3333074	1/3=0.3333333	0.0367307	(1/3)/sqrt(80) = 0.0372678

G. (10 points) Concluding remarks

Part G (Normal Distribution)

For a normal distribution, Equations 1 is valid for all values of n.

For a normal distribution, as n increases, the shape of the histogram and probability plot remains the same shape. Their mean remains the same, but the standard deviation decreases as n increases. This is due to Equations 1.

n	"size"
1	small
2	small
6	medium
10	large

Rule of Thumb: For a normal distribution, it does not rely on how large n is as its distribution is already of a normal variety. Therefore, even if n = 1, its population distribution and probability plot is already normal.

Uniform distribution:

1.Equations are valid.
2. As n increases, the histogram and probability plot gain more normality. The mean remains the same, standard deviation decreases as n increases.
3.N = 15 is considered large.

n	"size"
1	Small
2	Small
10	Medium
15	large

Rule of thumb: An n value of 15 or above should be considered "large enough" for X bar to become approximately normal.

G (Gamma distribution)

1. Equations 1 appear to be valid for all values of *n*, with experimental and theoretical values being approximately equal.
2. As *n* increases, the distribution of the sample mean becomes closer to normal. Numerically, the standard deviation decreases, while the mean remains relatively unchanged.
3. N = 60 can be considered large for gamma distribution.

n	"Size"
1	small
5	small
10	medium
20	medium
40	almost large
60	large

4. As a rule of thumb, N should be larger than approximately 60 before we can say the sample mean will have an approximately normal distribution for a population with a gamma distribution.

E.Poisson Distribution:

1. Equation 1 is valid for all values of n for poisson distribution.
2. For Poisson's distribution as the value of n increases, the histogram and probability plot become more normal. The mean remains around a value of 3 for all values of n, while the standard deviation decreases from 1.78 to 0.22 as the value of n increases.
3. n=60 is considered large, the QQ plots are abnormal for all values of n below this.

n	"Size"
1	small
5	medium
10	medium
20	medium
40	almost large

	enough
60	large

4. Rule of Thumb: An n value of 60 is large enough for the distribution to becomes approximately normal producing a histogram and probability plot resembling those for the normal distribution.

Part G (Exponential Distribution)

For an exponential distribution, Equations 1 is valid for all values of n.

As n increases, the histogram changes from a right-skewed graph to a normal distribution graph. Its probability plot changes from an exponential shape (probability plot for a right-skewed distribution) to a normal probability plot. The distribution's mean remains the same, whereas its standard deviation decreases due to Equations 1.

n	"size"
1	small
5	small
10	small
20	medium
40	medium
60	large
80	large

Rule of Thumb: For an exponential distribution, a n value of above 60 (>60) should be large enough to produce a histogram that resembles a normal distribution shape and a normal probability plot.

Overall

Distribution	**Value of *n* for approximately normal sample mean distribution**
Standard normal	1
Uniform	15
Gamma	60
Poisson	60
Exponential	60

As a general rule of thumb: the closer a population distribution is to a normal distribution, the higher *n* should be in order to ensure that the sample mean distribution becomes approximately normal. As *n* increases, every population distribution will have a sample mean distribution approaching normality; some population distributions will simply achieve approximately normal sample mean distributions faster than others.