Spring 2015 STAT 350 Project 1 (270 points)

Due Thursday, Feb. 26, 2015

Objectives: Understanding the Central Limit Theory

Instructions

- Groups of 3 – 4 students are required.
- Only PDF files are accepted.
- NO late work is accepted.
- Names and sections of all students in the group are on top of each page (all students must have the same professor, though it is acceptable to be in different sections)
- Statement of contribution for each student (submitted separately).
- Put all code in appendix; nothing is required in the main body. Code is required.
- Your report should be in the same order as the questions posed. Clearly label each part.
- All discussion should be in complete English sentences.

Only one report should be submitted per group with each person submitting their own statement of contribution. Everything should be submitted in Blackboard. For the person who is submitting the report, you can add a separate attachment for the statement of contribution. The statement of contribution should consist of what each student did in the project and if there were any problems with the group as a whole. This statement should not be shared with your group mates. Please include all of your group mates (and sections) in the header of the statement.

If you have any question about the project, please post your comment on piazza (preferred), ask TAs in office hours, or discuss it with your instructor.

It is acceptable for different parts of the project to use different software packages. Please read the tutorials for details in how to use the various software packages

1) (30 points) standard normal, sample sizes n = 1, 2, 6, 10
2) (30 points) uniform (0,1), sample sizes n = 1, 2, 9, 16, μ = ½, σ = $\sqrt{\frac{1}{12}}$
3) (50 points) gamma α = 5.4, β = 1, μ = 5.4, σ = $\sqrt{5.4}$, sample sizes n = 1, 5, 10, 20, 30, 40, … until the shape becomes normal
4) (60 points) exponential λ = 2, μ = 0.5 = σ, sample sizes n = 1, 5, 10, 20, 30, 40, 50, … until the shape becomes normal.
5) (30 points) Poisson, λ= 1, sample size n = 1, 2, 5, 10, 20, … until the shape becomes normal.
6) (40 points) Rolling a 2 on a fair six-sided die n times, n = 1, 2, 5, 10, 15,… until the sample becomes normal.

Grading information: (Sample point distributions)

For each of the distributions, we will be grading on the following information:

- (5 pts.) Procedure or code on obtaining the distribution (more than one software package may be used for each part). Only one code needs to be provided per part unless more than one software package is used for each part. Then both codes need to be provided. If too much information is provided in the code or the code does not match the output, take off 1 pt per error to maximum of 4 pts. Take off 5 pts. if the code is not provided at all. If the code is not in an appendix, see below for how many points should be taken off.

Spring 2015 STAT 350 Project 1 (270 points)

Due Thursday, Feb. 26, 2015

- (5 pts.) The Summary table + concluding sentence; 4 pts. for the table, 1 pt for the concluding sentence. Work needs to be required at least once for how the theoretical standard deviation is calculated. Take off 2 points ONCE if it is not provided. They also need to provide the output at least once in the project for the experimental data. Take off 2 points ONCE if it is not provided. For parts 1 and 2, take off 1 pt for each row that is NOT included. For parts 3 – 6 take off 1 pt if at least one row is not included up to 3 pts if only the first and/or last rows are included. The concluding sentence should be the equivalent of 'theory and experiment match.
- (5 pts.) Concluding sentence (for the plots). This should consist of at what number of averages did the distribution become normal and if there are any 'strange occurrences'. An example of a 'strange occurrence' is that at n_1 the shape is normal and at $n_2 > n_1$ the shape is not normal. There should also be a statement stating that the experimental and theoretical means and standard deviations match. Take off 1 point if this is not in understandable English or is not in complete sentences.
- (the rest, approximately 2 - 3 points for each pair) The plots (histograms and normal quantile plots) and whether they are normal or not. All of the appropriate pairs need to be included. The number of points taken off if a pair of graphs is not included depends on the number of pairs that are supposed to be there. If they do not state whether a pair of graphs is normal or not, take off 1 pt. Only the word 'normal' needs to be included; not a complete sentence.

In addition to the points mentioned below, you will be graded on organization and style for an additional 30 points. These points will consist of whether the organization of the report is easy to read and the items are in the correct order, whether the student names and sections are at the beginning of the report and if we receive the statement of contribution from each of the students.

Organization: 30 points:

Order of paper: 5 pts. This would include not putting the code in the appendix, the parts are not in numerical order, etc.

Names on top with sections (if appropriate): 5 pts. Only Dr. Sellke's and Dr. Findsen's sections need the appropriate section. This is all or nothing.

Statement of contribution: 20 pts. Please read these. If this statement is not provided, then that student ONLY loses the 20 points from the project. Please skim these to be sure that all group members contribute. If not all of the group members contribution according to these statements, please refer this to your instructor for how many points to provide on this part for non-performing students ONLY.

Note: If a student does not have a group, then take off 27 points. This includes the 25 points above for the names on the top and the statement of contribution. However, still grade 5 points for the order of the paper.

In the answer key, I am providing a sample table, the code for both SAS and R (for average = 1 only) at the end, some of the graphs and approximate values of when the distributions become normal for each of the parts. If you have any questions, please contact me.

Remember:

1. Work is required for the theoretical standard deviation.

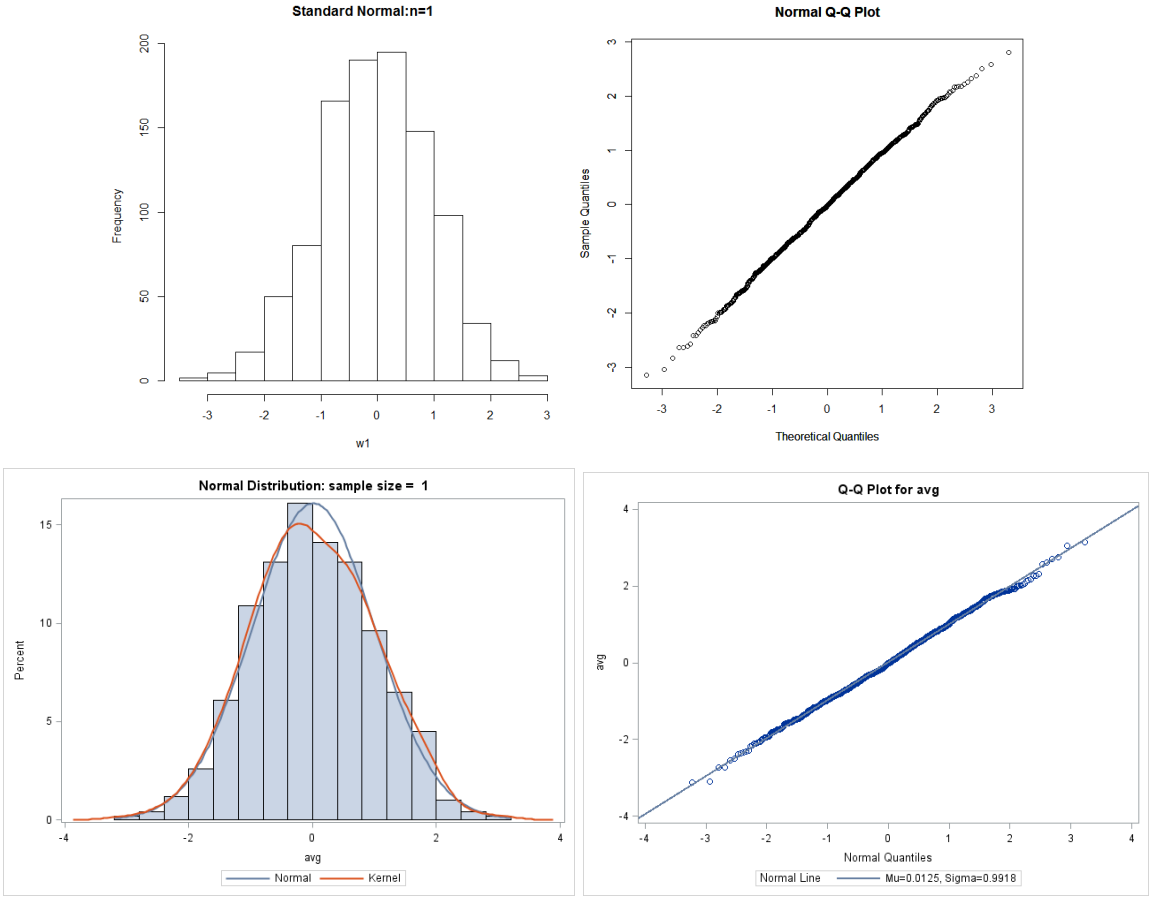
Spring 2015 STAT 350 Project 1 (270 points)

Due Thursday, Feb. 26, 2015

2. Descriptions are required for the table.
3. They need to state whether each set of graphs is normal or not.

1) (30 points) standard normal, sample sizes n = 1, 2, 6, 10
Sample output:

average=1



Based on the histogram and QQPlot, this dataset looks reasonable normal. Of course, the dataset was generated from the normal distribution

Equation for theoretical standard deviation:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

Sample Table:

n	Mean of 1000 sample means	Theoretical mean	Std. dev of 1000 sample means	Theoretical standard deviation
1	0.01249	0	0.9918	1
2	0.0761	0	0.7008	.707
6	0.0027	0	0.4054	.408
10	0.0115	0	0.3208	.316

Yes, the experimental values are close to the theoretical ones.

Sample Code:

SAS

```
%Let repeats = 1000;
%Let norm = rand ('Normal',0,1);

%LET n=1;
data normaldist;
do j=1 to &repeats by 1;
   avg=0;
   do i=1 to &n by 1;
      avg=avg+&norm;
   end;
   avg=avg/&n;
   output; drop i; drop j;
end;
title1 'Normal Distribution: sample size = ' &n;
proc univariate data = normaldist;
  qqplot avg/normal (mu=est sigma=est);
  run;
proc sgplot data = normaldist;
  histogram avg;
  density avg;
  density avg / type=kernel;
run;
```

R

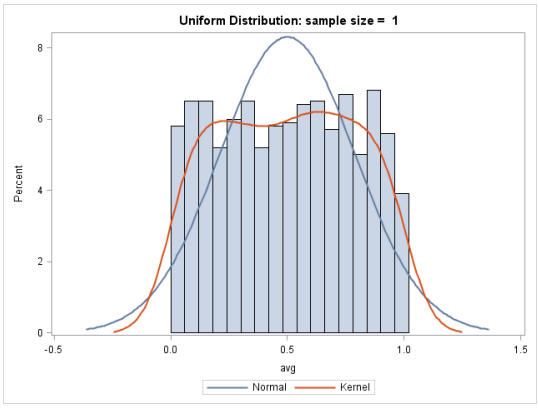
```
n <- 1000 #the number of repeats (not to be changed)

average <- 1
data.vec <- rnorm(n*average,mean=0,sd=1) #creates the random data
data.mat <- matrix(data.vec, ncol = average) #separates the data into columns
#apply(matrix, c(1, 2) == c("row", "column"), function)
avg <- apply(data.mat, 1, mean) #performs the averaging
meana = mean(avg)
meana
stda = sd(avg)
stda
hist(avg, main="Standard Normal:n=1",freq=F)
curve(dnorm(x,mean=meana,sd=stda),col="blue",lwd=2,add=T)
lines(density(avg),col="red",lwd=2)
dev.new()
qqnorm(avg,main= "Standard Normal:n=1 ")
qqline(avg)
```

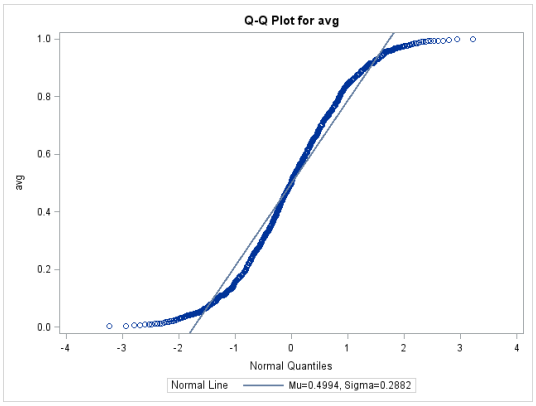
Spring 2015 STAT 350 Project 1 (270 points)

Due Thursday, Feb. 26, 2015

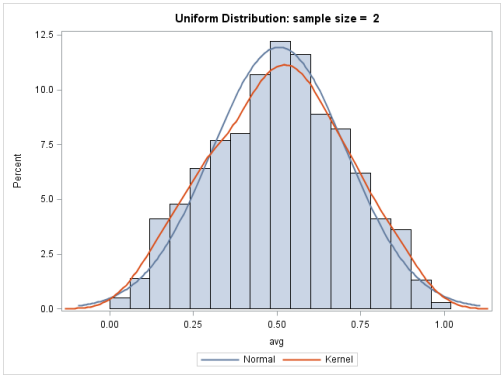
2) (30 points) uniform (0,1), sample sizes n = 1, 2, 9, 16, μ = ½, σ = $\sqrt{\frac{1}{12}}$

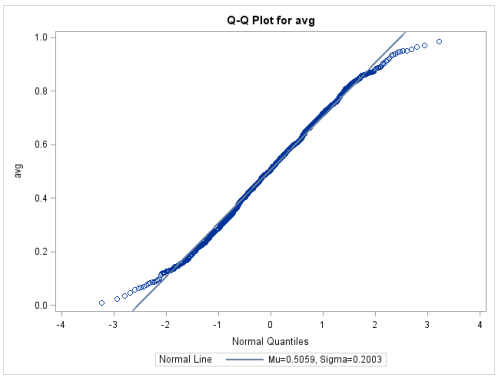
Sample output:



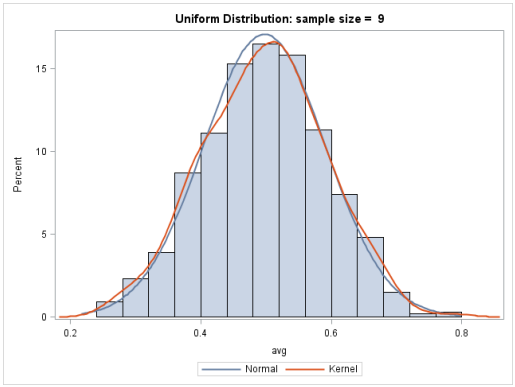


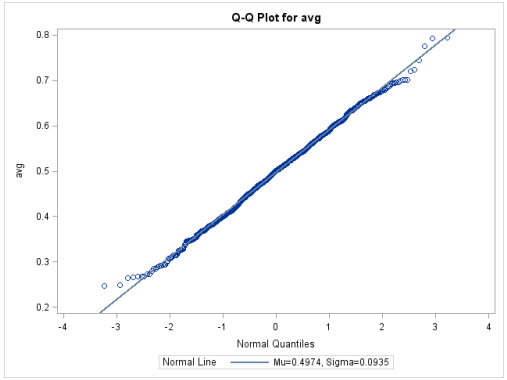
The histogram's lack of tails easily invalidates the normality assumption.





Since there appears to be concavity of the QQ Plot, the sample has increased its normality, but a larger value of n should be attempted.





Normality appears to be reasonable at this point.

N = 12 is definitely normal.

Spring 2015 STAT 350 Project 1 (270 points)

Due Thursday, Feb. 26, 2015

Sample Table:

n	Mean of 1000 sample means	Theoretical mean	Std. dev of 1000 sample means	Theoretical standard deviation
1	0.4994	.5	0.2881	.2887
2	0.5059	.5	0.2003	.2041
9	0.4974	.5	0.0935	.0962
12	0.5000	.5	0.0792	.0833

Yes, the experimental values are close to the theoretical ones.

Sample Code:

SAS

```
%Let repeats = 1000;
%Let unif = rand ('Uniform');

%LET n=1;
data uniform;
do j=1 to &repeats by 1;
   avg=0;
   do i=1 to &n by 1;
      avg=avg+&unif;
   end;
   avg=avg/&n;
   output; drop i; drop j;
end;
title1 'Uniform Distribution: sample size = ' &n;
proc univariate data = uniform;
  qqplot avg/normal (mu=est sigma=est);
  run;
proc sgplot data = uniform;
  histogram avg;
  density avg;
  density avg / type=kernel;
  run;
```

R

```
n <- 1000 #the number of repeats (not to be changed)

average <- 1
data.vec <- runif(n*average, min=0, max=1) #creates the random data
data.mat <- matrix(data.vec, ncol = average) #separates the data into columns
#apply(matrix, c(1, 2) == c("row", "column"), function)
avg <- apply(data.mat, 1, mean) #performs the averaging
meana = mean(avg)
meana
stda = sd(avg)
stda
hist(avg, main="Uniform:n=1",freq=F)
curve(dnorm(x,mean=meana,sd=stda),col="blue",lwd=2,add=T)
lines(density(avg),col="red",lwd=2)
dev.new()
```

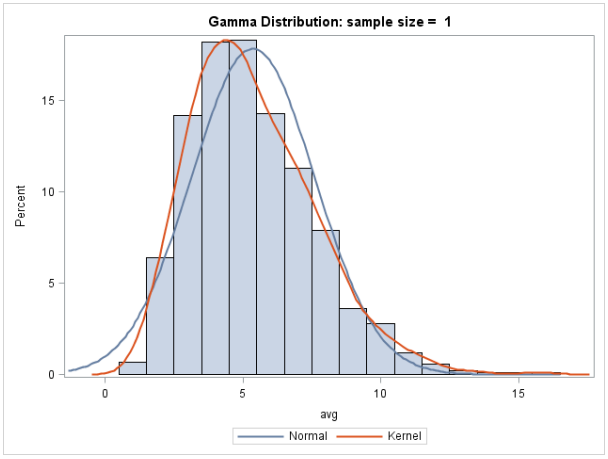
Spring 2015 STAT 350 Project 1 (270 points)

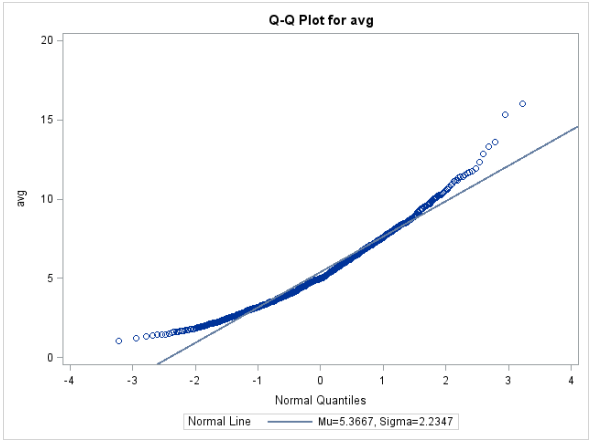
Due Thursday, Feb. 26, 2015

```
qqnorm(avg,main= "Uniform:n=1")
qqline(avg)
```

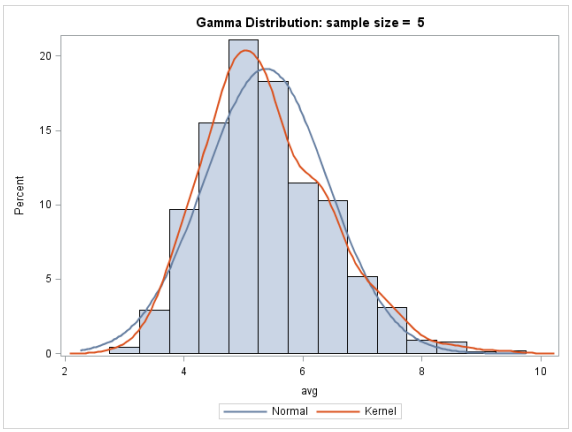
3) (50 points) gamma α = 5.4, β = 1, μ = 5.4, σ = $\sqrt{5.4}$, sample sizes n = 1, 5, 10, 20, 30, 40, … until the shape becomes normal

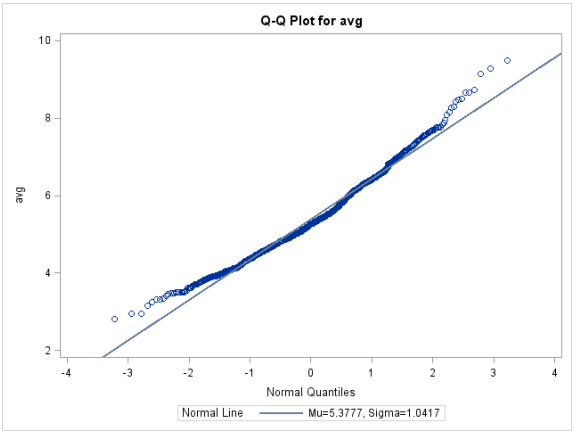
Sample output:





The strong right ward skewness in the histogram and the curvature of the QQ Plot invalidate the normality assumption

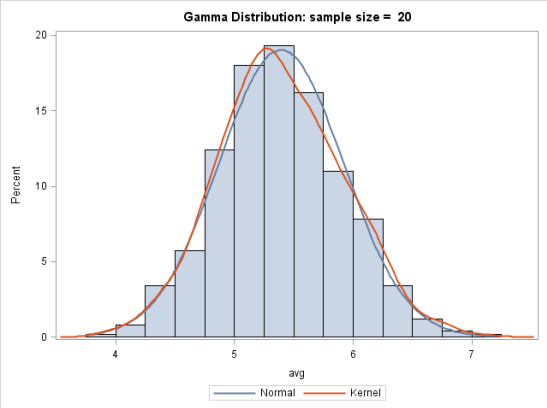


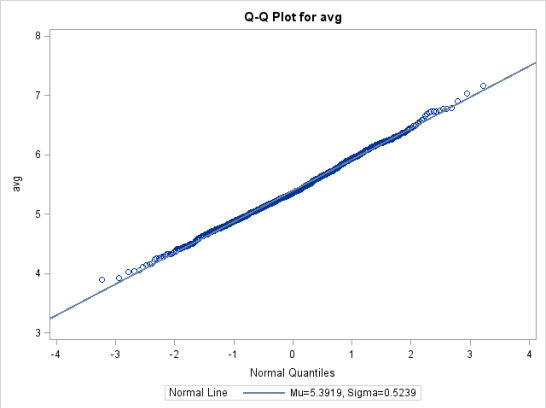


Although this histogram has improved over the previous one, it still maintains slight skewness.

Spring 2015 STAT 350 Project 1 (270 points)

Due Thursday, Feb. 26, 2015





Normality is a reasonable assumption

Sample Table:

n	Mean of 1000 sample means	Theoretical mean	Std. dev of 1000 sample means	Theoretical standard deviation
1	5.3667	5.4	2.2347	2.324
5	5.3777	5.4	1.0417	1.039
10	5.4195	5.4	0.7395	.7348
20	5.3919	5.4	0.5239	.5196
30	5.4419	5.4	0.4242	.4243
40	5.4155	5.4	0.3756	.3674
50	5.4091	5.4	0.3252	.3286

Yes, the experimental values are close to the theoretical ones.

Sample Code:

SAS

```
%Let repeats = 1000;
%Let gam = rand ('Gamma',5.4);
*Note: rand ('Gamma',5.4,1) also will work with SAS 9.4;

%LET n=1;
data gamma;
do j=1 to &repeats by 1;
   avg=0;
   do i=1 to &n by 1;
      avg=avg+&gam;
   end;
   avg=avg/&n;
   output; drop i; drop j;
end;
title1 'Gamma Distribution: sample size = ' &n;
proc univariate data = gamma;
  qqplot avg/normal (mu=est sigma=est);
  run;
proc sgplot data = gamma;
  histogram avg;
  density avg;
```

Spring 2015 STAT 350 Project 1 (270 points)

Due Thursday, Feb. 26, 2015

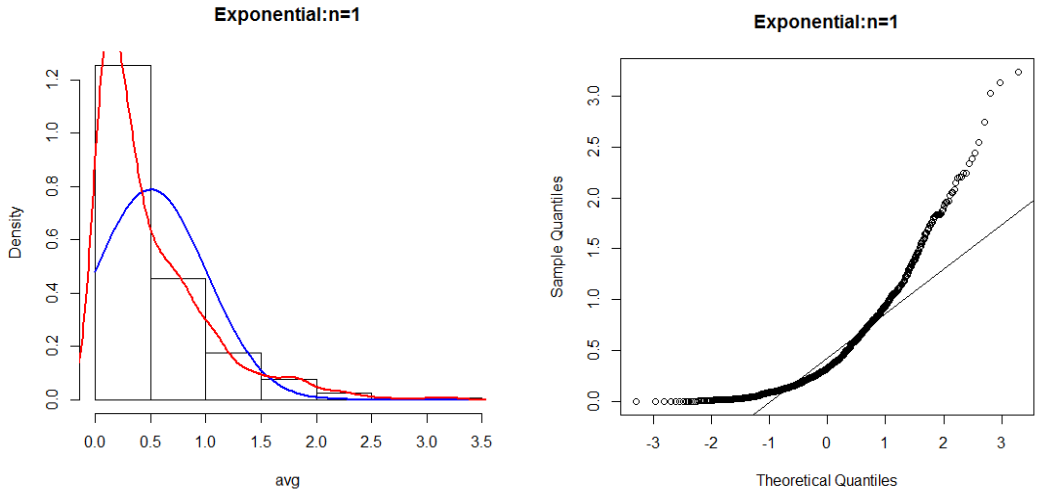
```
  density avg / type=kernel;
  run;
```

R

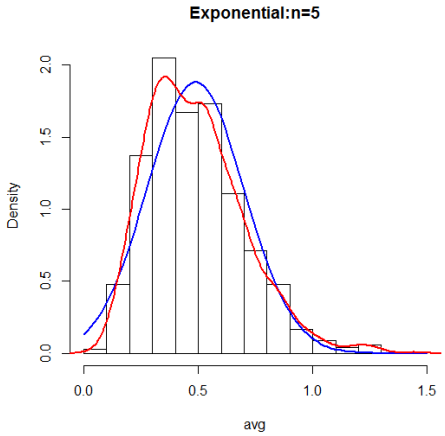
```
n <- 1000 #the number of repeats (not to be changed)

average <- 1
data.vec <- rgamma(n*average,5.4,rate=1) #creates the random data
data.mat <- matrix(data.vec, ncol = average) #separates the data into columns
#apply(matrix, c(1, 2) == c("row", "column"), function)
avg <- apply(data.mat, 1, mean) #performs the averaging
meana = mean(avg)
meana
stda = sd(avg)
stda
hist(avg, main="Gamma:n=1",freq=F)
curve(dnorm(x,mean=meana,sd=stda),col="blue",lwd=2,add=T)
lines(density(avg),col="red",lwd=2)
dev.new()
qqnorm(avg,main= "Gamma:n=1 ")
qqline(avg)
```

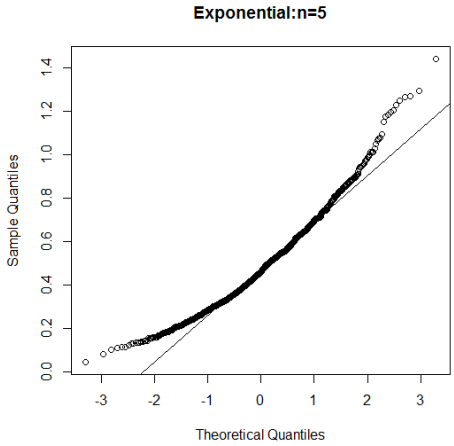
4) (60 points) exponential λ = 2, μ = 0.5 = σ, sample sizes n = 1, 5, 10, 20, 30, 40, 50, … until the shape becomes normal.

Sample output:

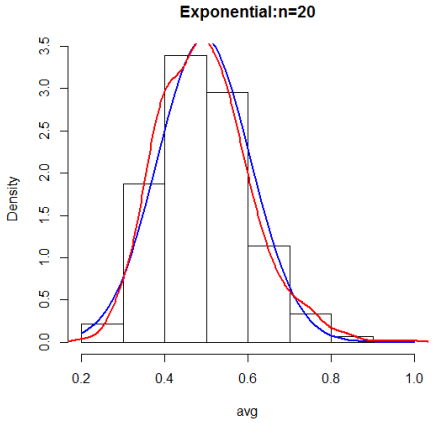


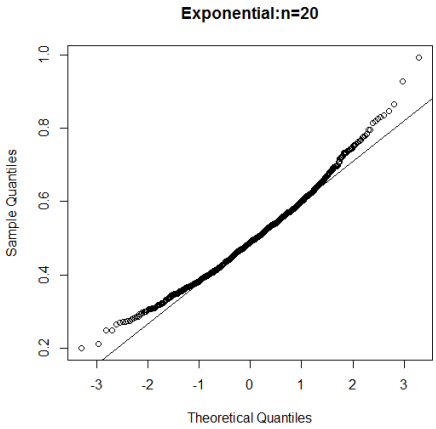
This is right skewed; clearly not normal.



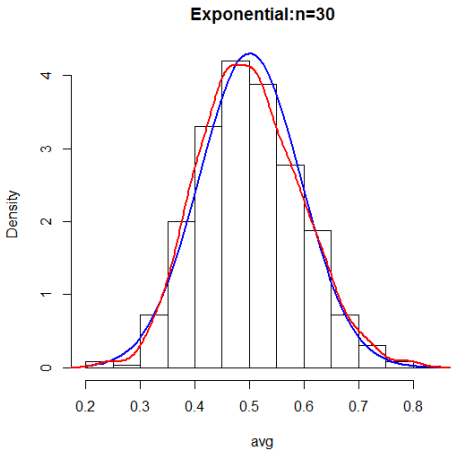


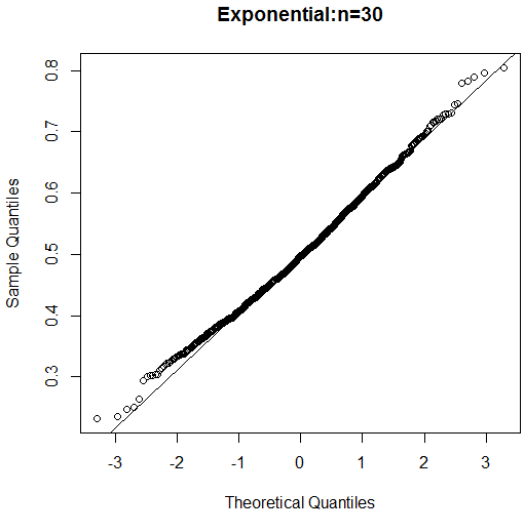
This is still right skewed and not normal.





This is still right skewed, but less so. It is still not normal.

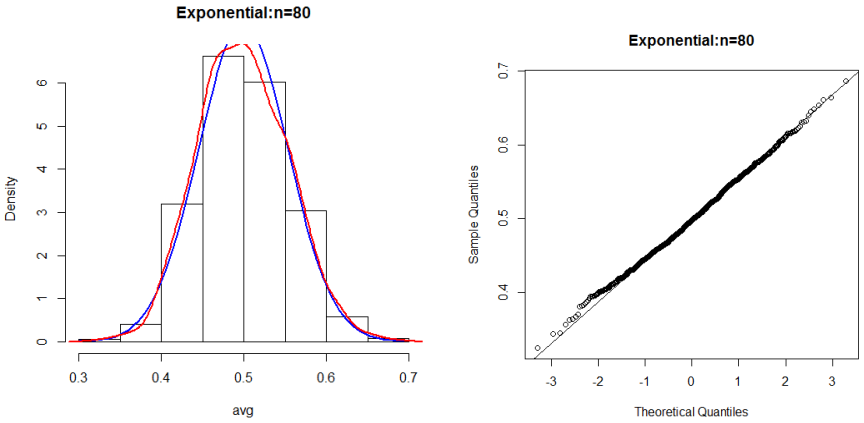




Normality is very reasonable for this pair of graphs. However, usually the histogram looks close but the QQplot still shows right skewedness.

Spring 2015 STAT 350 Project 1 (270 points)

Due Thursday, Feb. 26, 2015



Normality is very reasonable in these pair of graphs. It is possible that the students did go higher than 80 to get this though..

Sample Table:

n	Mean of 1000 sample means	Theoretical mean	Std. dev of 1000 sample means	Theoretical standard deviation
1	0.5023	.5	0.5053	.5
5	0.4876	.5	0.2122	.2236
10	0.5003	.5	0.1561	.1581
20	0.4945	.5	0.1111	.1118
30	0.5005	.5	0.0927	.0913
80	0.4994	.5	0.0543	.0559

Yes, the experimental values are close to the theoretical ones.

Sample Code:

SAS

```
%Let repeats = 1000;
%Let exp = rand ('Exponential',0.5);
%LET n=1;
data expo;
do j=1 to &repeats by 1;
   avg=0;
   do i=1 to &n by 1;
      avg=avg+&exp;
   end;
   avg=avg/&n;
   output; drop i; drop j;
end;
title1 'Exponential Distribution: sample size = ' &n;
proc univariate data = expo;
  qqplot avg/normal (mu=est sigma=est);
  run;
proc sgplot data = expo;
  histogram avg;
  density avg;
```

Spring 2015 STAT 350 Project 1 (270 points)

Due Thursday, Feb. 26, 2015

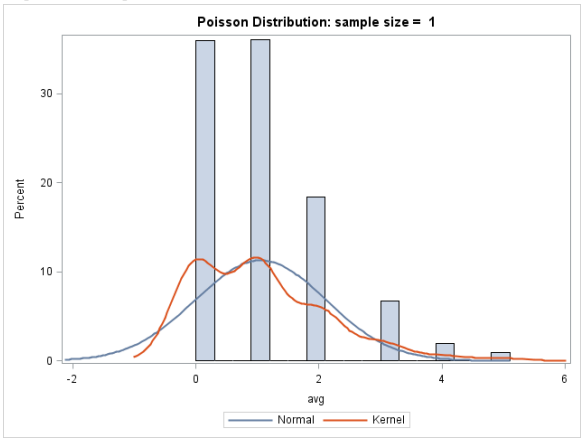
```
  density avg / type=kernel;
  run;
```

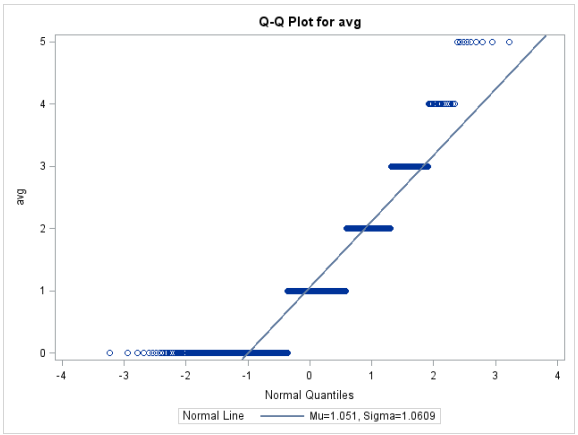
R

```
n <- 1000 #the number of repeats (not to be changed)

average <- 1
data.vec <- rexp(n*average,rate = 2) #creates the random data
data.mat <- matrix(data.vec, ncol = average) #separates the data into columns
#apply(matrix, c(1, 2) == c("row", "column"), function)
avg <- apply(data.mat, 1, mean) #performs the averaging
meana = mean(avg)
meana
stda = sd(avg)
stda
hist(avg, main="Exponential:n=1",freq=F)
curve(dnorm(x,mean=meana,sd=stda),col="blue",lwd=2,add=T)
lines(density(avg),col="red",lwd=2)
dev.new()
qqnorm(avg,main= "Exponential:n=1 ")
qqline(avg)
```

5) (30 points) Poisson, λ= 1, sample size n = 1, 2, 5, 10, 20, … until the shape becomes normal.

Sample output:

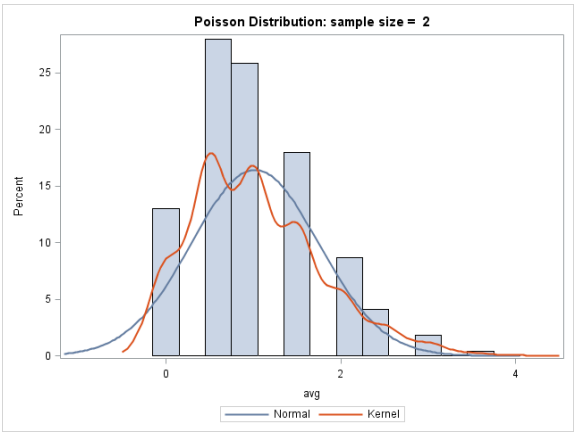


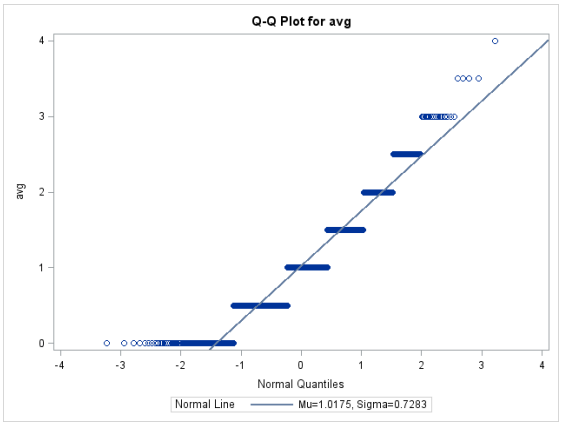


This histogram is not even remotely normal and it looks discrete.

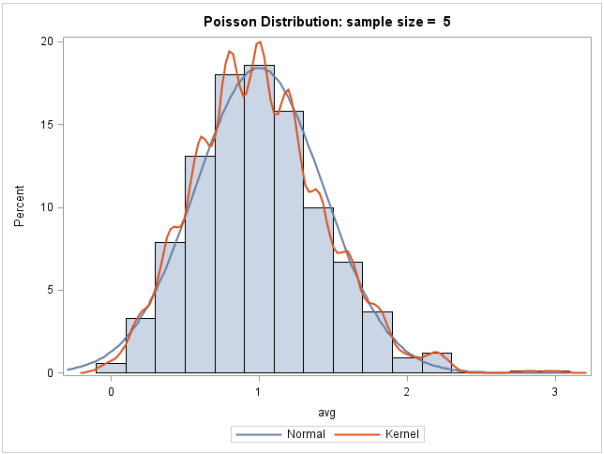
Spring 2015 STAT 350 Project 1 (270 points)

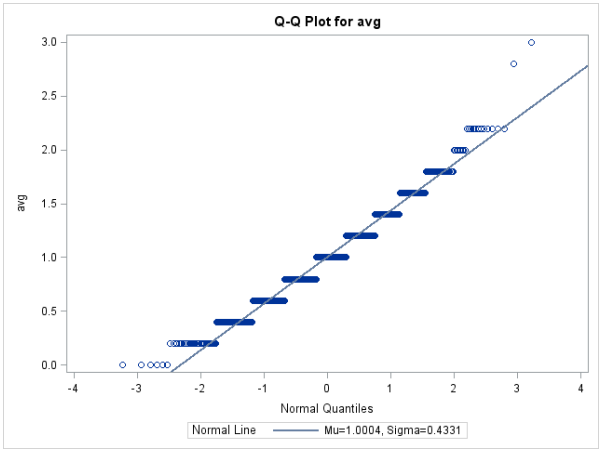
Due Thursday, Feb. 26, 2015



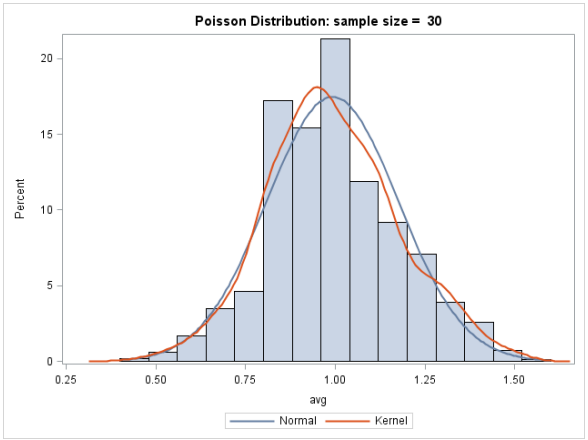


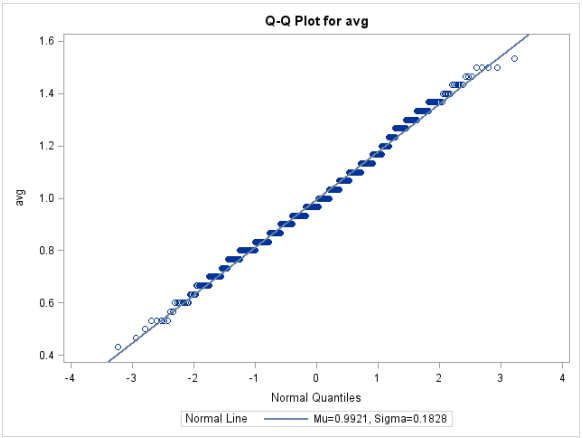
This is still right skewed and still looks discrete.





There is still moderate right skewness and only the qqplot looks discrete..

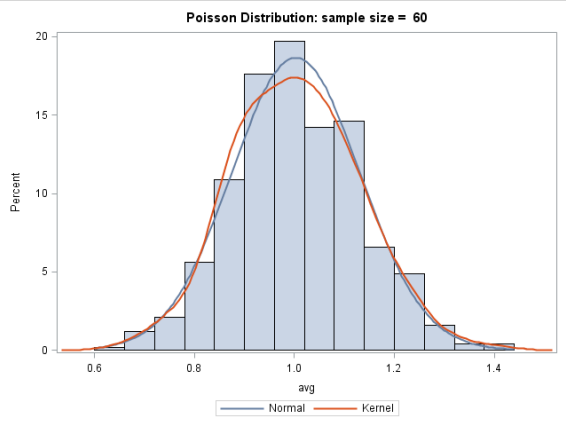


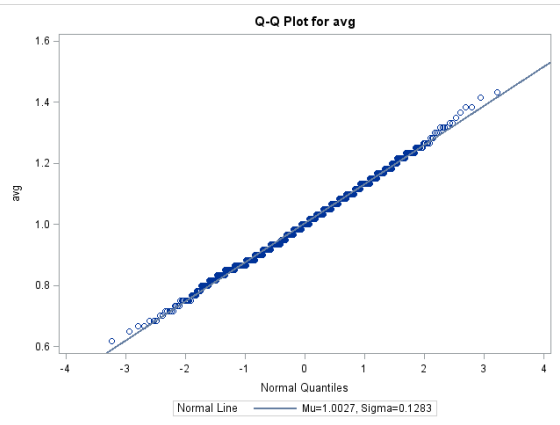


Normality appears to be reasonable at this point and the qqplot looks more continuous.

Spring 2015 STAT 350 Project 1 (270 points)

Due Thursday, Feb. 26, 2015





Normality appears to be reasonable at this point and the distribution looks more continuous also.

Sample Table:

n	Mean of 1000 sample means	Theoretical mean	Std. dev of 1000 sample means	Theoretical standard deviation
1	1.051	1	1.0609	1
2	1.0175	1	0.7283	0.7071
5	1.0004	1	0.4331	0.4472
10	1.0076	1	0.3239	0.3162
20	1.01825	1	0.2288	0.2236
30	0.9921	1	0.1828	0.1826
40	0.9961	1	0.1576	0.1581
60	1.0027	1	0.1283	0.1291

Yes, the experimental values are close to the theoretical ones.

Sample Code:

SAS

```
%Let repeats = 1000;
%Let pois = rand ('Poisson',1);

%LET n=1;
data poisson;
do j=1 to &repeats by 1;
   avg=0;
   do i=1 to &n by 1;
      avg=avg+&pois;
   end;
   avg=avg/&n;
   output; drop i; drop j;
end;
title1 'Poisson Distribution: sample size = ' &n;
proc univariate data = poisson;
  qqplot avg/normal (mu=est sigma=est);
  run;
```

Spring 2015 STAT 350 Project 1 (270 points)

Due Thursday, Feb. 26, 2015

```
proc sgplot data = poisson;
  histogram avg;
  density avg;
  density avg / type=kernel;
  run;
```

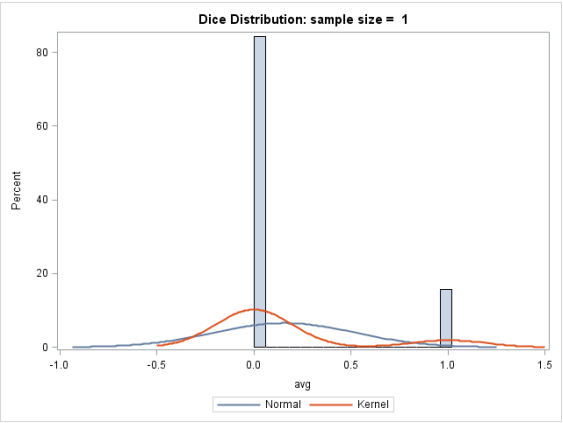
R

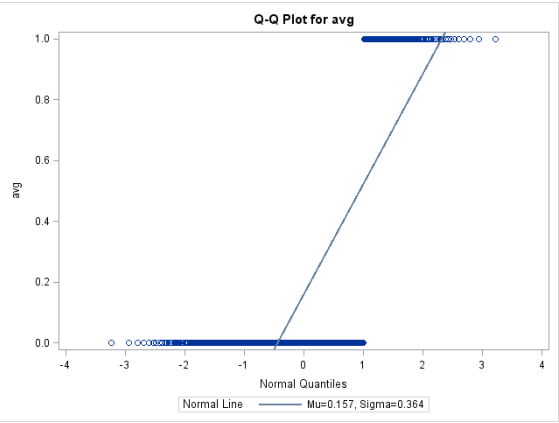
```
n <- 1000 #the number of repeats (not to be changed)

average <- 1
data.vec <- rpois(n*average,1) #creates the random data
data.mat <- matrix(data.vec, ncol = average) #separates the data into columns
#apply(matrix, c(1, 2) == c("row", "column"), function)
avg <- apply(data.mat, 1, mean) #performs the averaging
meana = mean(avg)
meana
stda = sd(avg)
stda
hist(avg, main="Poisson:n=1",freq=F)
curve(dnorm(x,mean=meana,sd=stda),col="blue",lwd=2,add=T)
lines(density(avg),col="red",lwd=2)
dev.new()
qqnorm(avg,main= "Poisson:n=1 ")
qqline(avg)
```

6) (40 points) Rolling a 2 on a fair six-sided die n times, n = 1, 2, 5, 10, 15,… until the sample becomes normal.

Note: When running the code multiple times for the same 'n', I get greatly different values. Therefore, the final value of n where the distribution becomes normal will vary greatly.

Sample output:

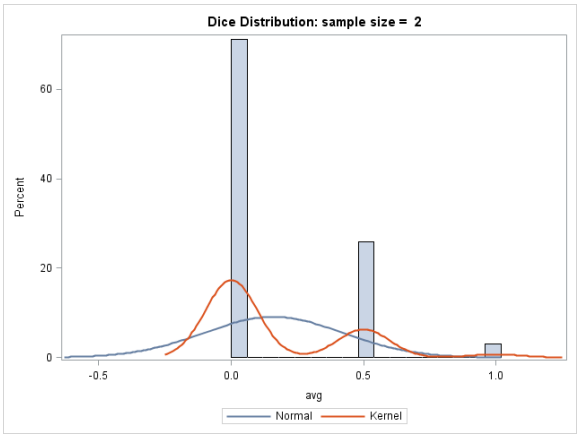


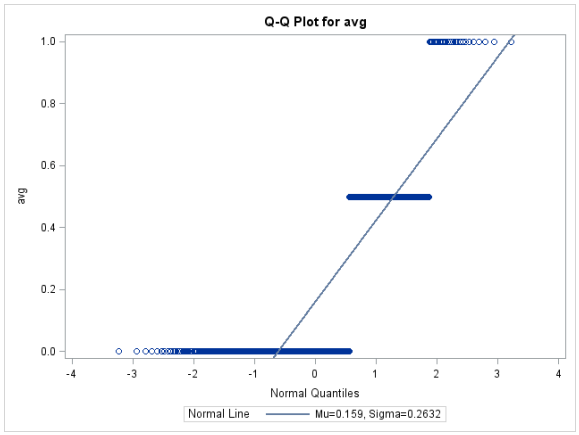


This plot is clearly not normal for every possible reason.

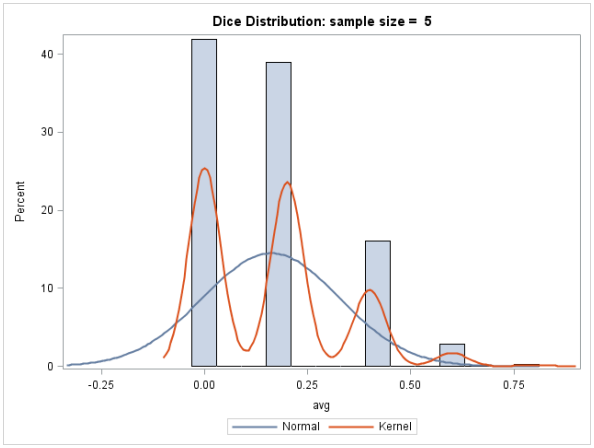
Spring 2015 STAT 350 Project 1 (270 points)

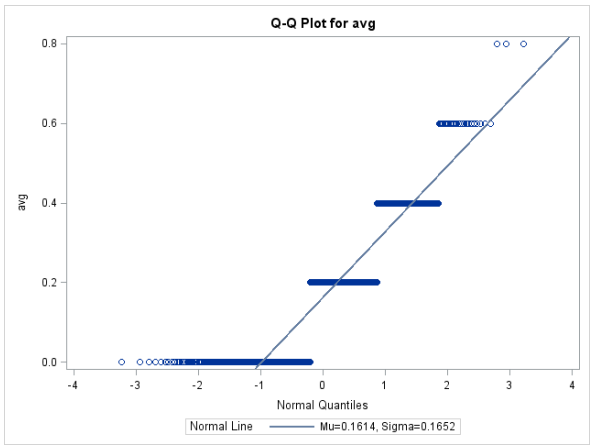
Due Thursday, Feb. 26, 2015



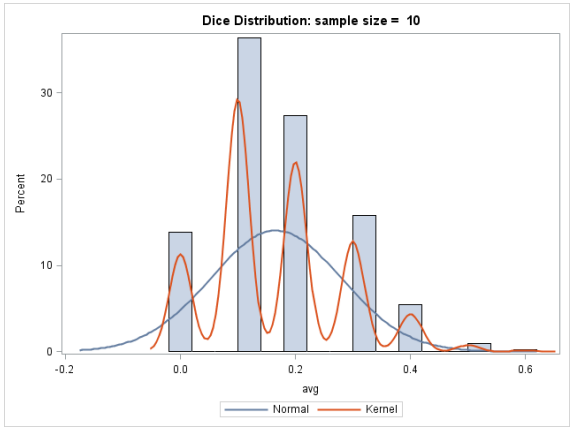


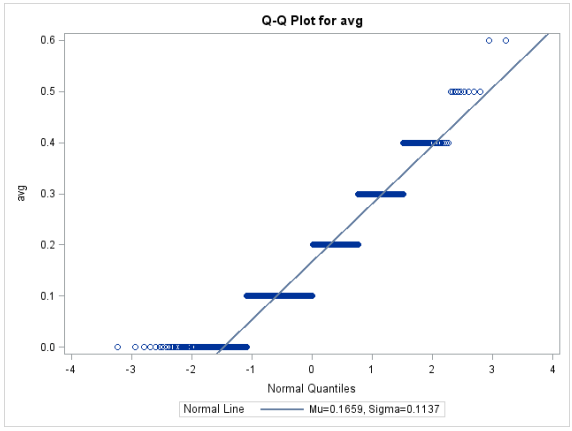
This plot is still clearly not normal for every possible reason.





This plot is still clearly not normal for every possible reason.

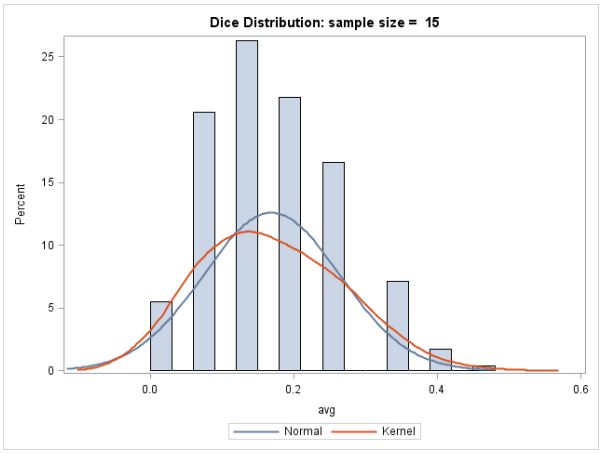


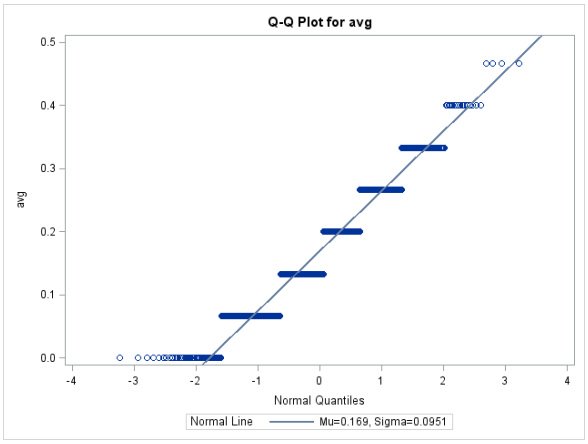


The distribution is still discrete but only slightly right skewed.

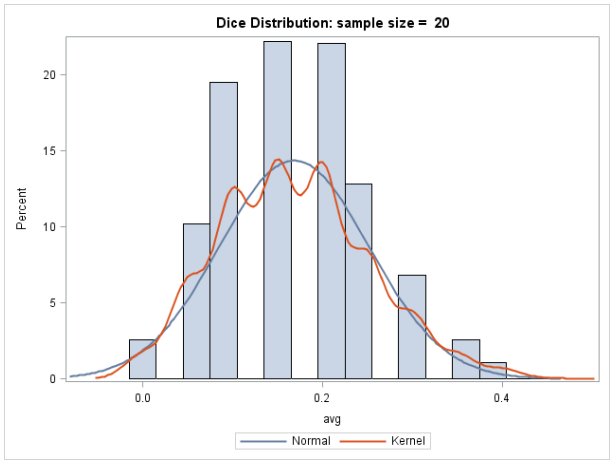
Spring 2015 STAT 350 Project 1 (270 points)

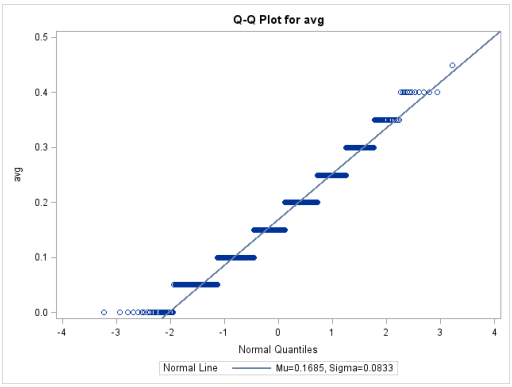
Due Thursday, Feb. 26, 2015



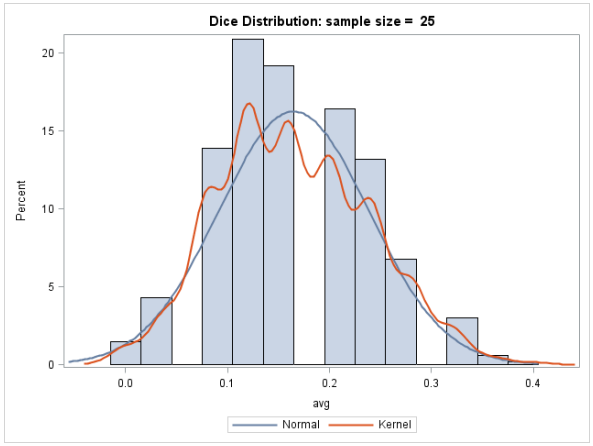


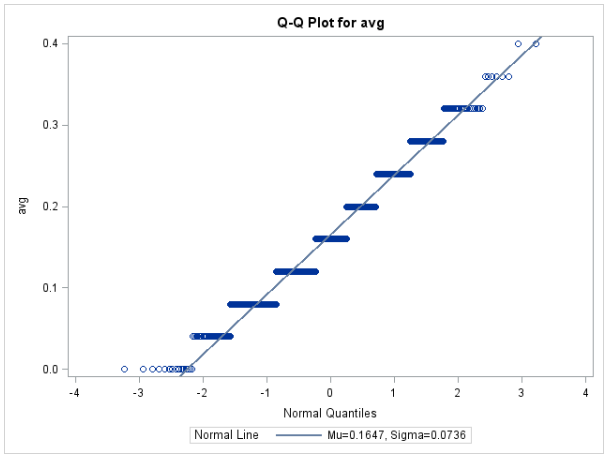
This distribution is almost normal, but still looks discrete.





This distribution is almost normal, but still looks discrete.

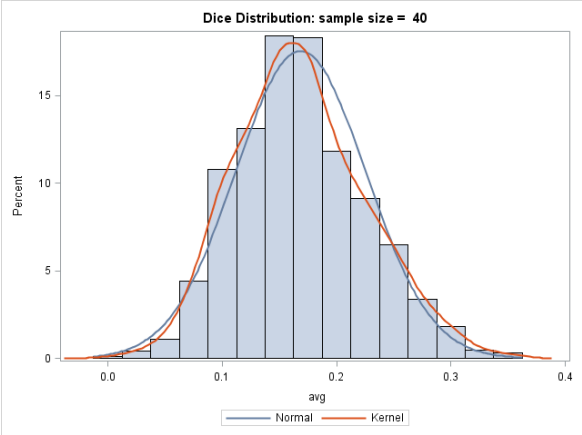


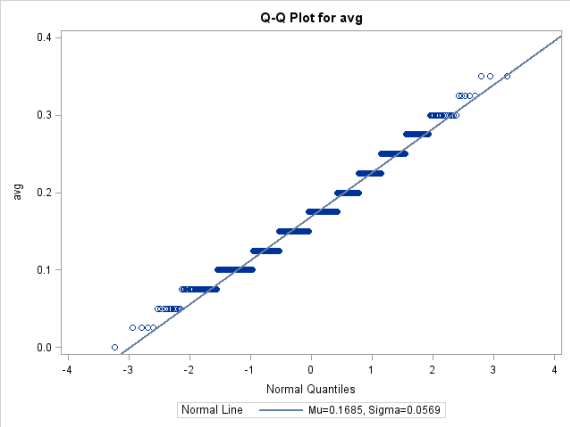


This distribution is still slightly right skewed (but very close to normal). I would like to do one more because of the holes in the histogram.

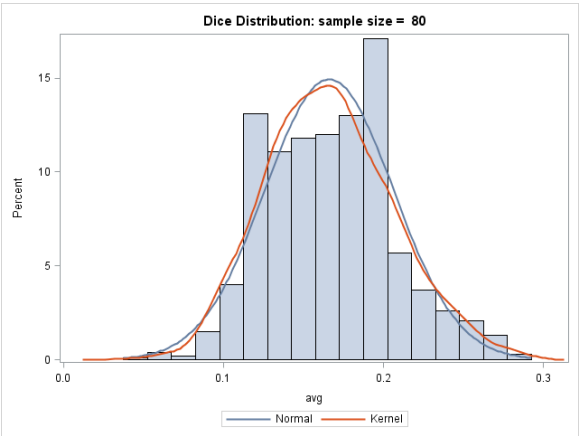
Spring 2015 STAT 350 Project 1 (270 points)

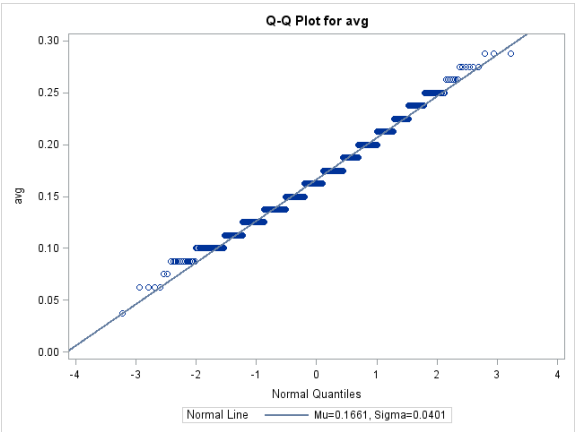
Due Thursday, Feb. 26, 2015





This distribution is very slightly right skewed from the QQ plot, but looks symmetrical from the histogram so I would state that this one is normal even though the QQ plot looks discrete.





Even though the shape of the histogram does not look normal, both the curves do match and the QQplot is fairly straight. The fact that the QQplot still looks discrete is not relevant.

Sample Table:

n	Mean of 1000 sample means	Theoretical mean	Std. dev of 1000 sample means	Theoretical standard deviation
1	0.157	.1667	0.3640	.3727
2	0.159	.1667	0.2632	.2636
5	0.1614	.1667	0.1652	.1667
10	0.1659	.1667	0.1137	.1179
15	0.169	.1667	0.0951	.0962
20	0.16845	.1667	0.0833	.0833
25	0.1647	.1667	0.0736	.0745
40	0.1685	.1667	0.0569	.0589
80	0.1661	.1667	0.0401	.0417

Yes, the experimental values are close to the theoretical ones. This is a binomial distribution with n = 1, p = $\frac{1}{6}$

Spring 2015 STAT 350 Project 1 (270 points)

Due Thursday, Feb. 26, 2015

Sample Code:

SAS

```
%Let repeats = 1000;
%Let dice = int(6*rand('Uniform')) + 1;

%LET n=1;
data dice;
do j=1 to &repeats by 1;
   avg=0;
   do i=1 to &n by 1;
      answer = 0;
       x = &dice;
       if x = 2 then answer = 1;
      avg=avg+answer;
   end;
   avg=avg/&n;
   output; drop i; drop j;
end;
title1 'Dice Distribution: sample size = ' &n;
proc univariate data = dice;
  qqplot avg/normal (mu=est sigma=est);
  var avg;
  run;
proc sgplot data = dice;
  histogram avg;
  density avg;
  density avg / type=kernel;
  run;
```

R

```
n <- 1000 #the number of repeats (not to be changed)

average <- 1
dice <- sample(1:6,n*average,replace=T) #simulates the die role
data.vec <- rep(0,n*average) #initializes the vector to be 0
for (i in 1:(n*average)) {
   if (dice[i]==2) data.vec[i] <- 1
data.mat <- matrix(data.vec, ncol = average) #separates the data into columns
#apply(matrix, c(1, 2) == c("row", "column"), function)
avg <- apply(data.mat, 1, mean) #performs the averaging
meana = mean(avg)
meana
stda = sd(avg)
stda
hist(avg, main="Dice Roll:n=1",freq=F)
curve(dnorm(x,mean=meana,sd=stda),col="blue",lwd=2,add=T)
lines(density(avg),col="red",lwd=2)
dev.new()
qqnorm(avg,main= "Dice Roll:n=1 ")
qqline(avg)
```