SAS Project (26.5 pts + 1 Bonus)
Due Oct 12

1.2 DATALINES:

Problem 1 (1 pts.)
Modify the datalines code (a1.sas) to create a new dataset in SAS using the data from Exercise 2.2.4 (p. 39) in section 2.2:
Name the dataset and variables appropriately for the dataset (see page 31). Points will be taken off if you name the dataset a1 and the variable x.
Your submission should consist of your code and the print out of your dataset.

SAS Code:

```
data Gpig;
input nerve @@;
datalines;
23 30 54 28 31 29 34 35 30
27 21 43 51 35 51 49 35 24
26 29 21 29 37 27 28 33 33
23 37 27 40 48 41 20 30 57
;
run;

proc print data=Gpig; run;

quit;
```

SAS Output:

Obs	nerve
1	23
2	30
3	54
4	28
5	31
6	29
7	34
8	35
9	30
10	27
11	21
12	43

13	51
14	35
15	51
16	49
17	35
18	24
19	26
20	29
21	21
22	29
23	37
24	27
25	28
26	33
27	33
28	23
29	37
30	27
31	40
32	48
33	41
34	20
35	30
36	57

Problem 2 (1 pts.)

Modify the datalines code (a1.sas) to create a new dataset in SAS using the data from Exercise 2.4.3 (p. 51) where the X values are listed in the problem and all of the Y values are '1':

Name the dataset and variable appropriately for the dataset (see page 30). Points will be taken off if you name the dataset a1 and the variables x and y.

Your submission should consist of your code and the print out of your dataset.

SAS Code:

```
data sheep;
input milk group @@;
datalines;
56.5 1 89.8 1 110.1 1 65.6 1  63.7 1 82.6 1
75.1 1 91.5 1 102.9 1 44.4 1 108.1 1
;
run;

proc print data=sheep; run;

quit;
```

SAS Output:

Obs	milk	group
1	56.5	1
2	89.8	1
3	110.1	1
4	65.6	1
5	63.7	1
6	82.6	1
7	75.1	1
8	91.5	1
9	102.9	1
10	44.4	1
11	108.1	1

1.3. INFILING:

Problem 3 (1 pts.)
Consider the number of dendrite branches from 36 nerve cells from brains of newborn guinea pigs presented in Exercise 2.2.4 (p. 39). This data file is called E2.2.4.dat. Modify the SAS infiling code (a2.sas) to infile this dataset into SAS.
Your submission should consist of your code and the log file. Note: your log file should only be for this problem and contain no errors in it.

SAS Code:

```
data Gpig;
infile 'H:\My Documents\Stat 503\SAS\E.2.2.4.dat';
input nerve;
run;

proc print data=Gpig;
run;

quit;
```

SAS Log file

```
52   data Gpig;
53   infile 'H:\My Documents\Stat 503\SAS\E.2.2.4.dat';
54   input nerve;
55   run;

NOTE: The infile 'H:\My Documents\Stat 503\SAS\E.2.2.4.dat' is:
      Filename=H:\My Documents\Stat 503\SAS\E.2.2.4.dat,
      RECFM=V,LRECL=256,File Size (bytes)=144,
      Last Modified=11Oct2012:13:50:31,
      Create Time=11Oct2012:13:50:31

NOTE: 36 records were read from the infile 'H:\My Documents\Stat 503\SAS\E.2.2.4.dat'.
      The minimum record length was 2.
      The maximum record length was 2.
NOTE: The data set WORK.GPIG has 36 observations and 1 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.00 seconds

56
57   proc print data=Gpig;
58   run;

NOTE: There were 36 observations read from the data set WORK.GPIG.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds

59
60   quit;
```

Note: the numbers in front of each line will vary depending on what you have done in SAS before this question.

2.1. HISTOGRAMS:

Problem 4 (1 pts.)

Produce a histogram that matches the histogram that you created in the homework for Problem 2.2.4. I have removed the SAS code that I previously posted in the key; however, if the code is identical, that is, all of the variable, data set, and titles are the same, you will receive half credit for this problem.

Your submission should consist of your code and histogram.

SAS Code:

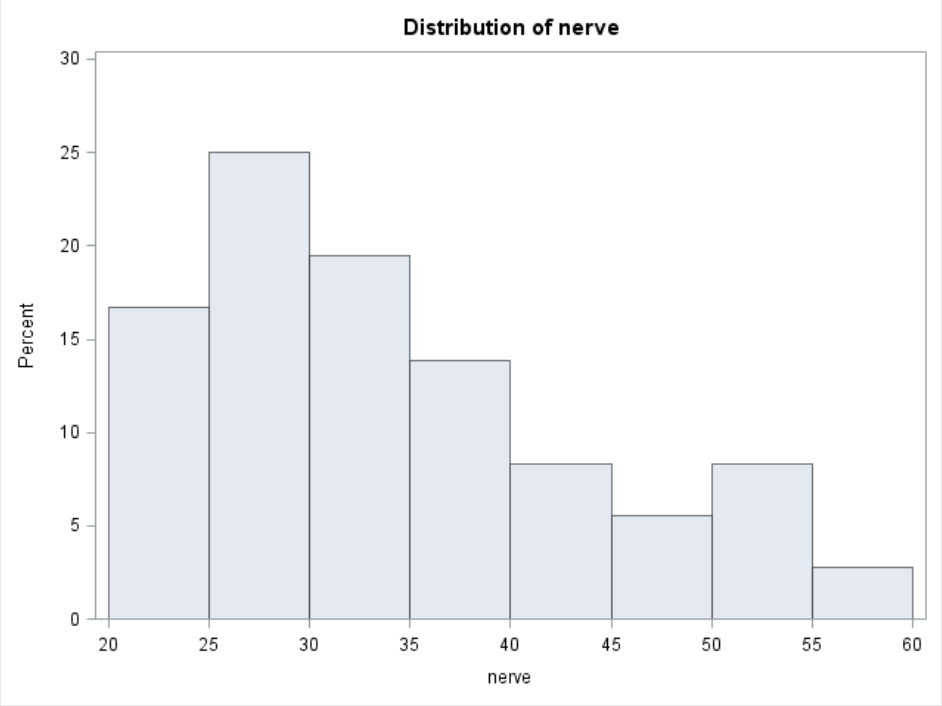
```
data Gpig;
infile 'H:\My Documents\Stat 503\SAS\E.2.2.4.dat';
input nerve;
```

```
run;
proc print data=Gpig;
run;

title1 'Relative percentage of Branches from each nerve';
proc univariate data=Gpig noprint;
histogram nerve / endpoints = 20 to 60 by 5;
run;

quit;
```

SAS Histogram

Relative percentage of Branches from each nerve

The UNIVARIATE Procedure



2.2. BOXPLOTS:

Problem 5 (1 pts.)

Consider the study of milk production in sheep presented in Exercise 2.4.3 (p. 51) of our text. Modify the SAS code (a4.sas) to reproduce modified boxplots that you created in your homework. Again, please change the title of the plot. I have removed the SAS code that I previously posted in the key; however, if the code is identical, that is, all of the variable, data set, and titles are the same, you will receive half credit for this problem.

Your submission should consist of your code and the one figure of the boxplot.

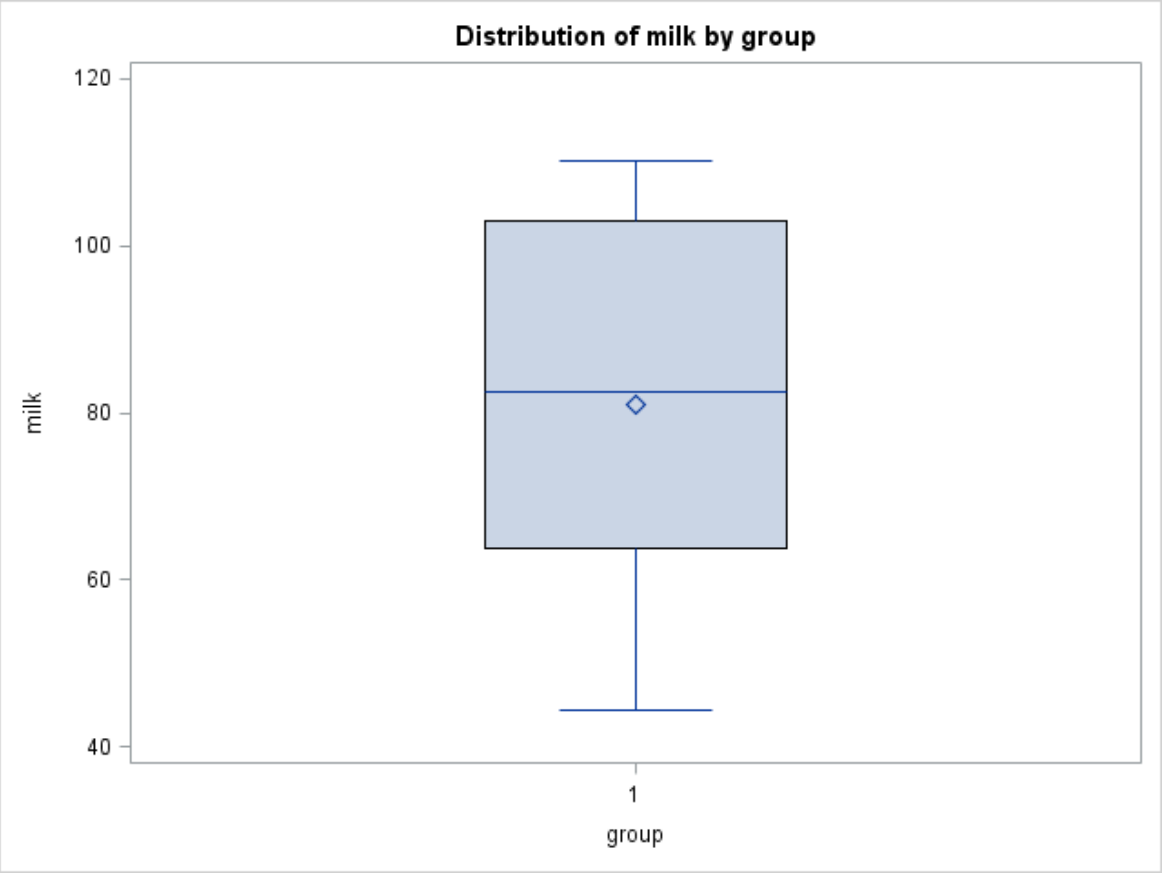
SAS Code:

```
data sheep;
infile 'H:\My Documents\Stat 503\SAS\E.2.4.3.dat';
input milk group;
```

```
run;

title1 'Boxplot of milk production';
proc boxplot data=sheep ;
plot milk*group/boxstyle=schematic idsymbol=circle;
run;

quit;
```

SAS Boxplot:

Boxplot of milk production



Bonus B1 (1 pts.) Consider the Radish Growth using three different light treatments presented in Example 2.5.3 (p. 55) of our text. (dataset Ex2.5.3.dat). Modify the SAS infiling code (a4.sas) to produce side-by-side modified boxplots as in Figure 2.5.3. In the data set for this problem, the second variable is text so be sure to include the $ after it. Again, please change the title of the plot.

Your submission should consist of your code and the one figure of the side-by-side boxplots.

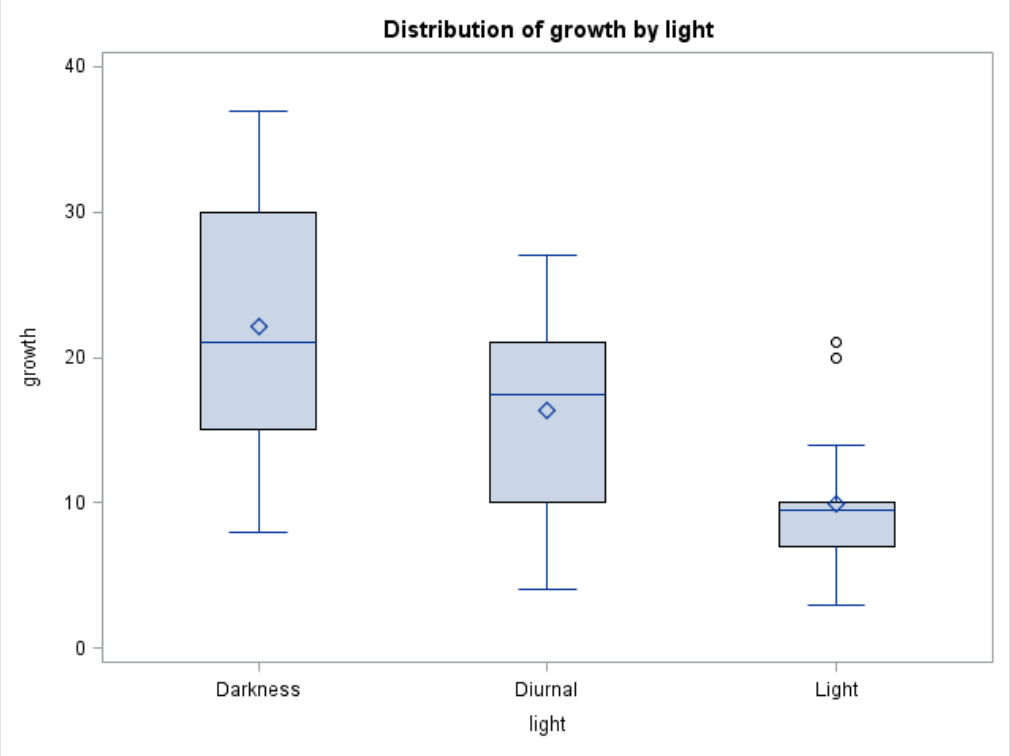
SAS code:

```
data radish;
infile 'H:\My Documents\Stat 503\SAS\Ex2.5.3.dat';
input growth light$;
```

```
run;

title1 'Radish Growth';
proc boxplot data=radish;
plot growth*light/boxstyle=schematic idsymbol=circle;
run;

quit;
```

SAS Boxplot:

Radish Growth

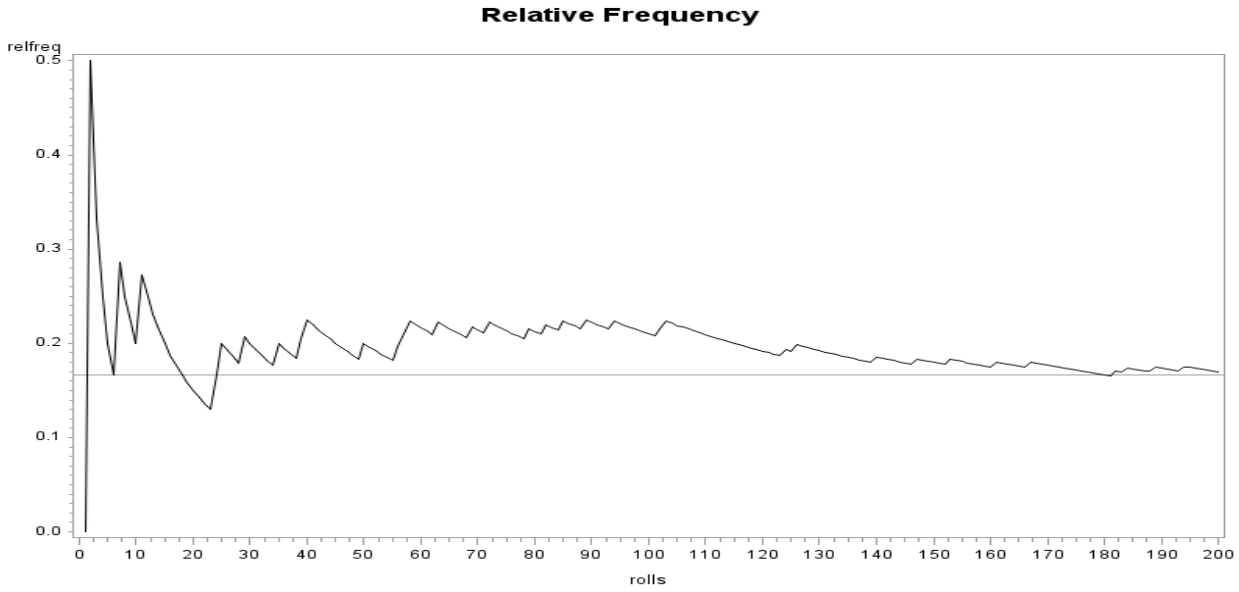


3.1 Long-Run Convergence:

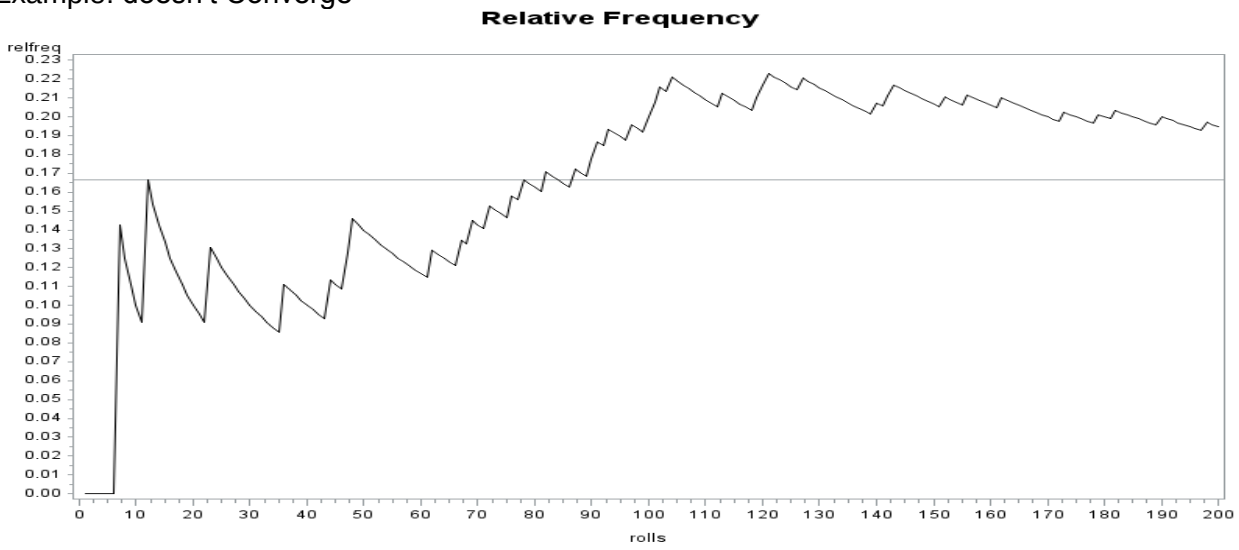
Problem 6 (1.5 pts.)
Run the learning code 10 times. How many of the plots stabilize to the correct value? Your submission should consist of at least one of the plots that converges and at least one of the plots that doesn't converge in addition to the number of plots that stabilize to the correct value. The input code is not required because you are using the learning code with no modifications.

Number of plots that stabilize: any number between 1 and 9 is acceptable.

Example: Converge:



Example: doesn't Converge



4.1. Generate random samples using specific distributions:

Problem 7 (3 pts.)

a) Run b4.sas three times. For each of the times, compare the calculated mean and standard deviation with the true values by stating the difference between each of the values and the true value.

b) Run b5.sas three times. For each of the times, compare the calculated mean and standard deviation with the true values by stating the difference between each of the values and the true value.

Your submission should consist of the output from the means for each of the 6 runs (no printout of the generated data set) in addition to the information requested above which can be hand written or typed next to the corresponding output. The input code is not required because you are using the learning code with no modifications.

a) theoretical values:

mean = np = (40)(0.5) = 20

$standard\ deviation = \sqrt{np(1-p)} = \sqrt{(40)(0.5)(0.5)} = \sqrt{10} = 3.1627766$

Analysis Variable : x1

N	Mean	Difference from theoretical	Std Dev	Difference from theoretical	Minimum	Maximum
20	19.5000000	19.5 – 20 = -0.5	3.9669689	3.9670-3.1628 =0.8042	13.0000000	26.0000000

Analysis Variable : x1

N	Mean	Difference from theoretical	Std Dev	Difference from theoretical	Minimum	Maximum
20	20.3500000	20.35–20 =0.35	3.5433407	3.5433-3.1628 =0.3805	15.0000000	25.0000000

Analysis Variable : x1

N	Mean	Difference from theoretical	Std Dev	Difference from theoretical	Minimum	Maximum
20	20.3500000	20.35–20 =0.35	3.2650461	3.2650-3.1628 =0.1022	14.0000000	27.0000000

Your values will be different from mine.

b) theoretical values: mean = 15.6, standard deviation = 0.4

Analysis Variable : x1

N	Mean	Difference from theoretical	Std Dev	Difference from theoretical	Minimum	Maximum
20	15.6496507	15.6497–15.6 =0.0497	0.3003365	0.3003-0.4 =-0.0997	15.2262461	16.3156827

Analysis Variable : x1

N	Mean	Difference from theoretical	Std Dev	Difference from theoretical	Minimum	Maximum
20	15.4254546	15.4255–15.6 =-0.1745	0.3334666	0.3335-0.4 =-0.0665	14.8299346	15.9751559

Analysis Variable : x1

N	Mean	Difference from theoretical	Std Dev	Difference from theoretical	Minimum	Maximum
20	15.5587709	15.5588–15.6 =-0.0412	0.4395944	0.4396-0.4 =0.0936	14.8114249	16.5079992

5.1 QQplots: Normal

Problem 8 (2.5 pts.)

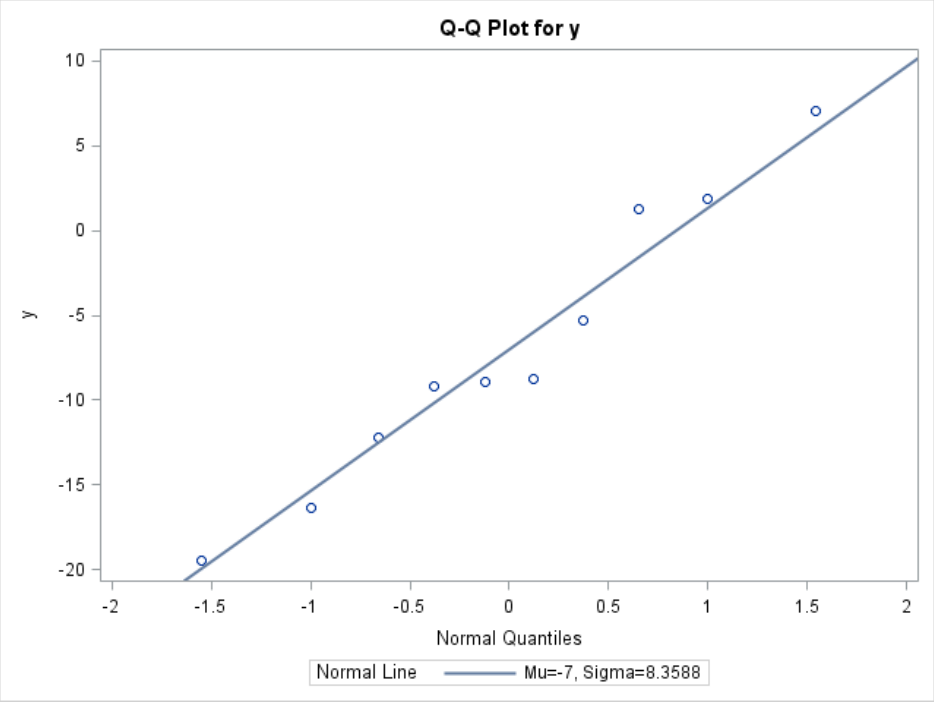
Run a8.sas 3 times with each of 10, 100 and 1000 data points for a total of 9 different times. For each of the runs, save the QQplot. All of these QQplots are from a normal distribution, therefore, all of them form a 'good' line. Which number of data points has the 'best' line? Hint: Besides looking at the line itself, look at the estimated value of μ and σ. Which is the 'worst'? If you would have just looked at the 'worst' line, would you think it came from a normal distribution if you didn't know that it came from a normal distribution? Why or why not?

Your submission should consist the 'best' line from each of the 3 different number of data points from the QQPlot generated from SAS and the answer to the questions above.

Note: I am just providing sample plots:

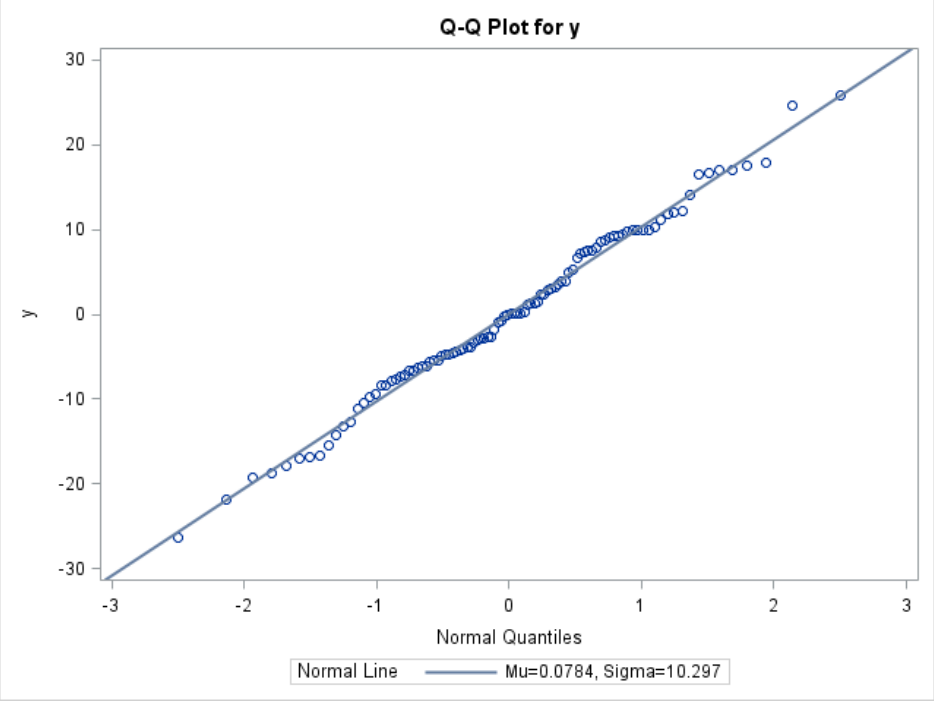
Best for n = 10

QQplot Normal Distribution, n = 10



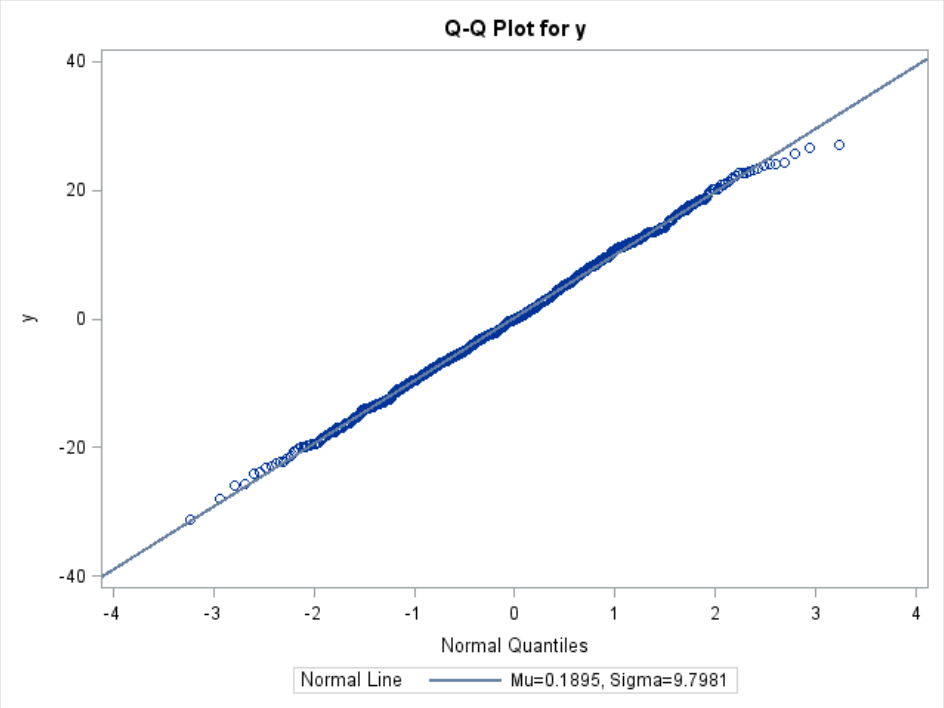
Best for n = 100

QQplot Normal Distribution, n = 100



Best for n = 1000

QQplot Normal Distribution, n = 1000

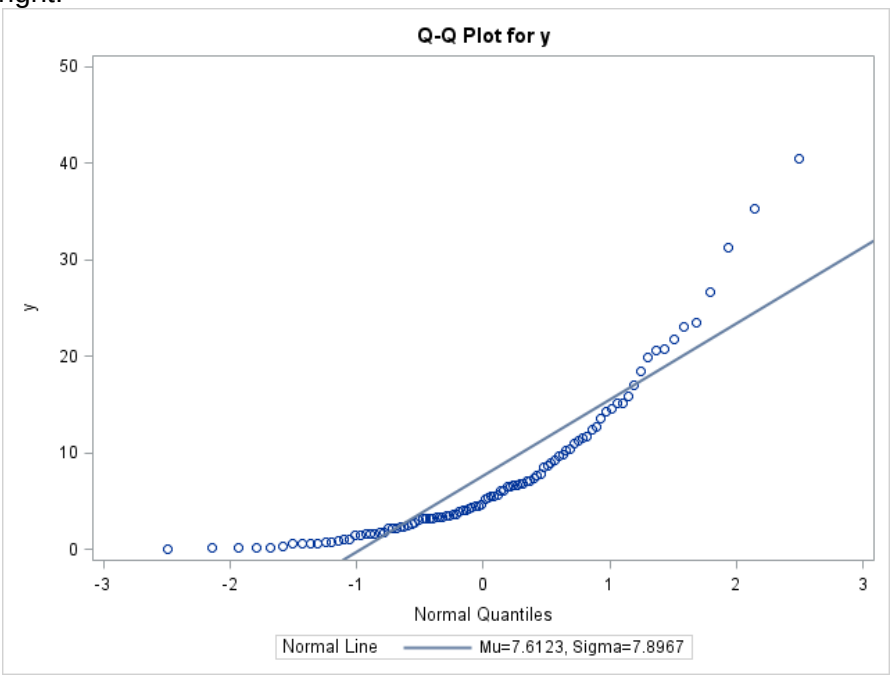


The number of data points that has the best line is: 1000 points.
The number of data points that has the worst line: 10 points.
Does this line come from a normal distribution? The QQplot from the n = 10 set looks right skewed even though I know that it came from a normal distribution.

5.2 QQplots: Not Normal

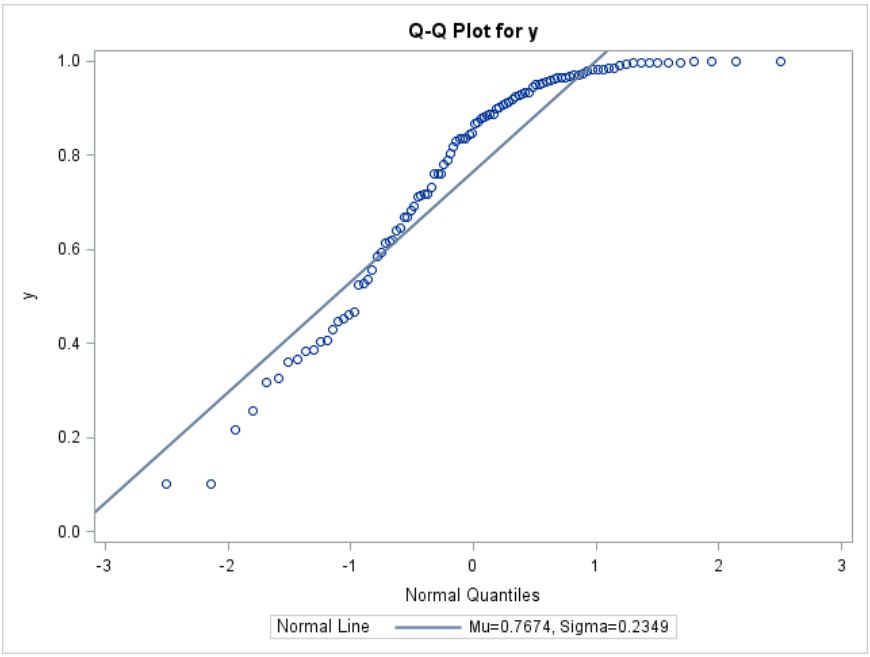
Problem 9 (4.5 pts.)
Run a9.sas 3 times with each of the different not normal distributions. Describe the shape of the QQplot for each of the types of distributions. Go back to the previous problem; do any of the 9 QQplots display a not normal distribution.
Your submission should consist of the most represented line for each of the distributions with a description of the line. In addition, if any of the 'Normal' QQplots display a non-normal distribution, present that with the distribution that it represents.

right:



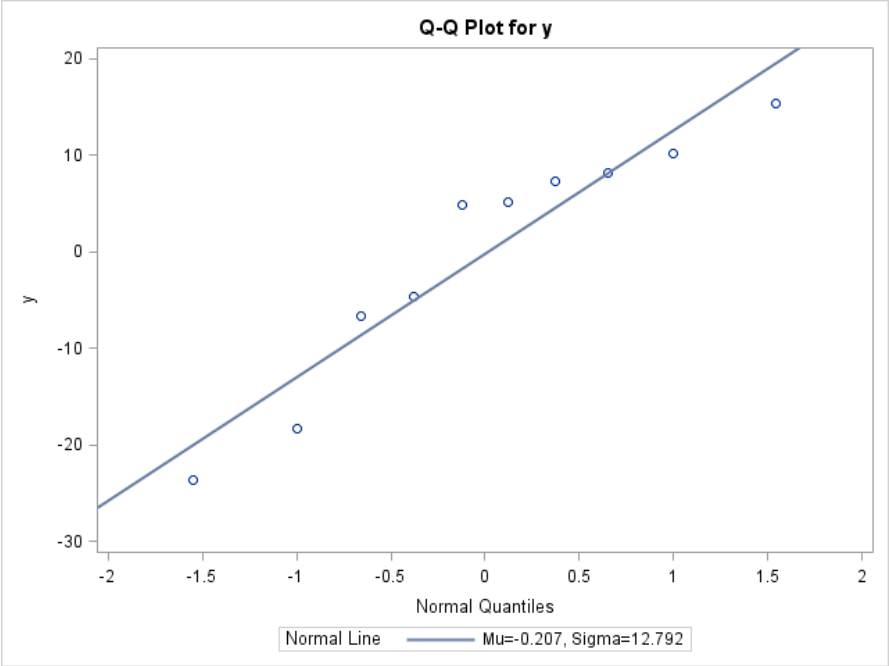
This is curved upward.

left:

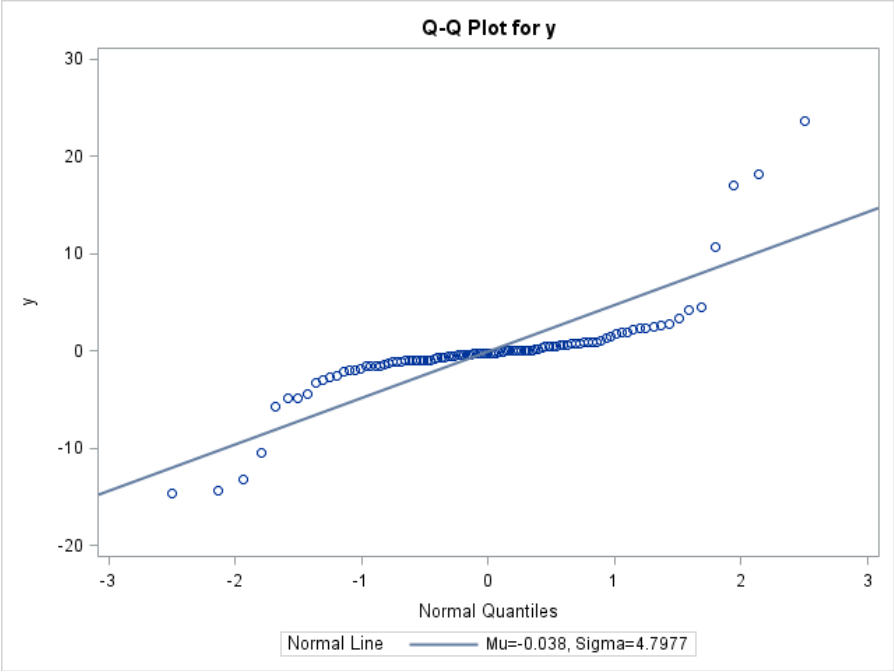


curved downward

This is the normal plot which has a similar shape

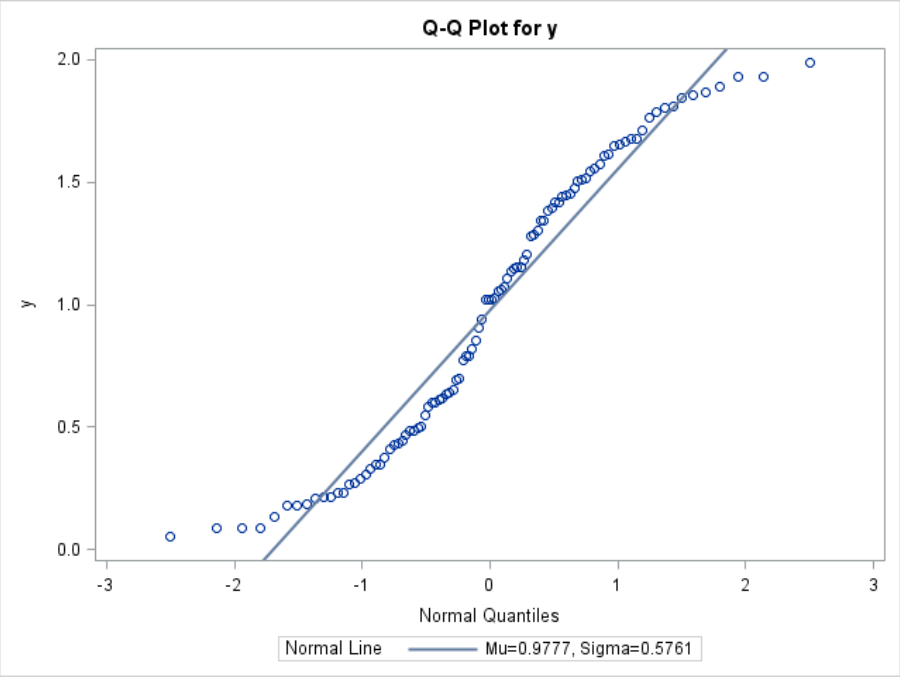


long:



S shaped with pointing upward at high values and downward at low values.

short:



S shaped with pointing to the right at high values and to the left at low values.

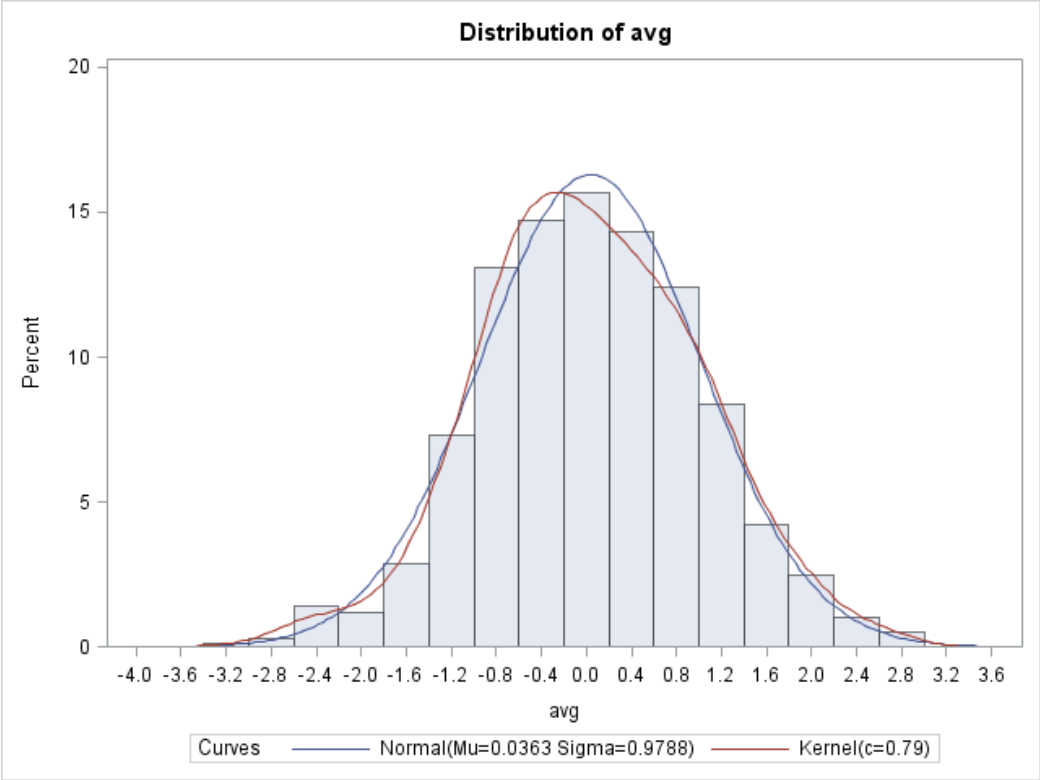
6.Sampling Distributions:

Problem 10 (6 pts.)

Run the preceding learning code for n = 1, n = 2 and n = 6 for both normal distribution and the uniform distribution

Your submission should consist of the six generated histograms, the six outputs from proc means, two tables consisting of the three means from the output and the theoretical value and the three standard deviations from the output and the theoretical value for both the normal distribution and the uniform distribution. The theoretical value for the standard deviation for each run is the standard error given the appropriate value of n. Does the calculated values match the theoretical values? In addition, please indicate which of the graphs look like they are a normal distribution (this can be written directly on the output). No code is required because you are making only minor modifications to the learning code.

Normal:
n = 1



looks normal

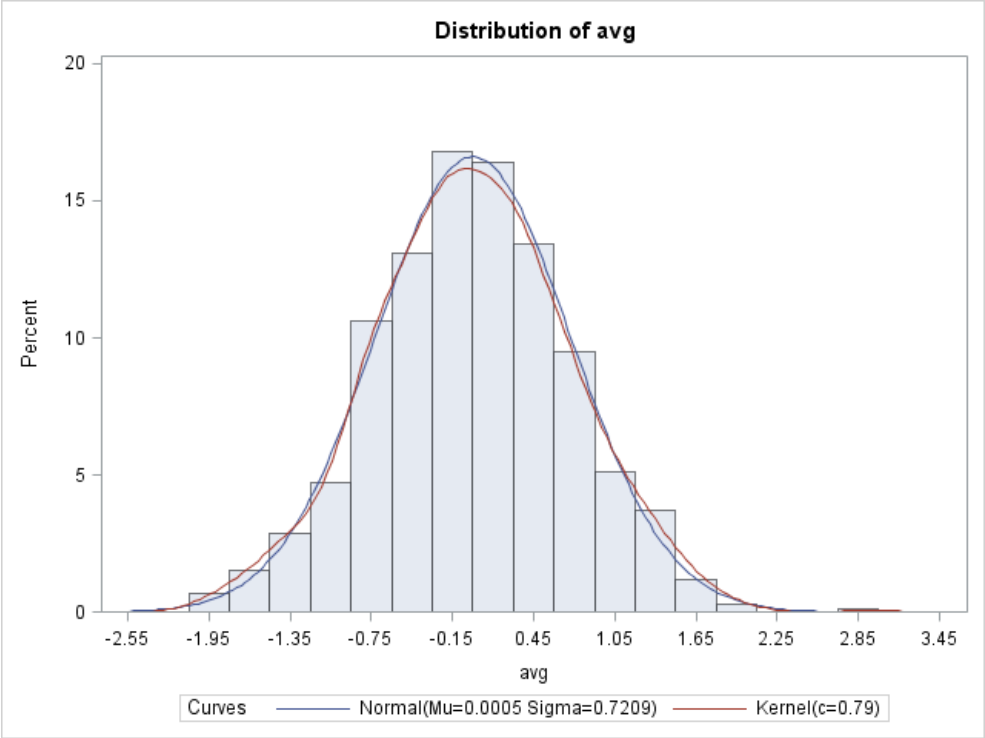
Normal Distribution repeats = 1

The MEANS Procedure

Analysis Variable : avg

N	Mean	Std Dev	Minimum	Maximum
1000	-0.0307342	0.9818657	-3.1328369	3.7313060

n = 2



looks normal

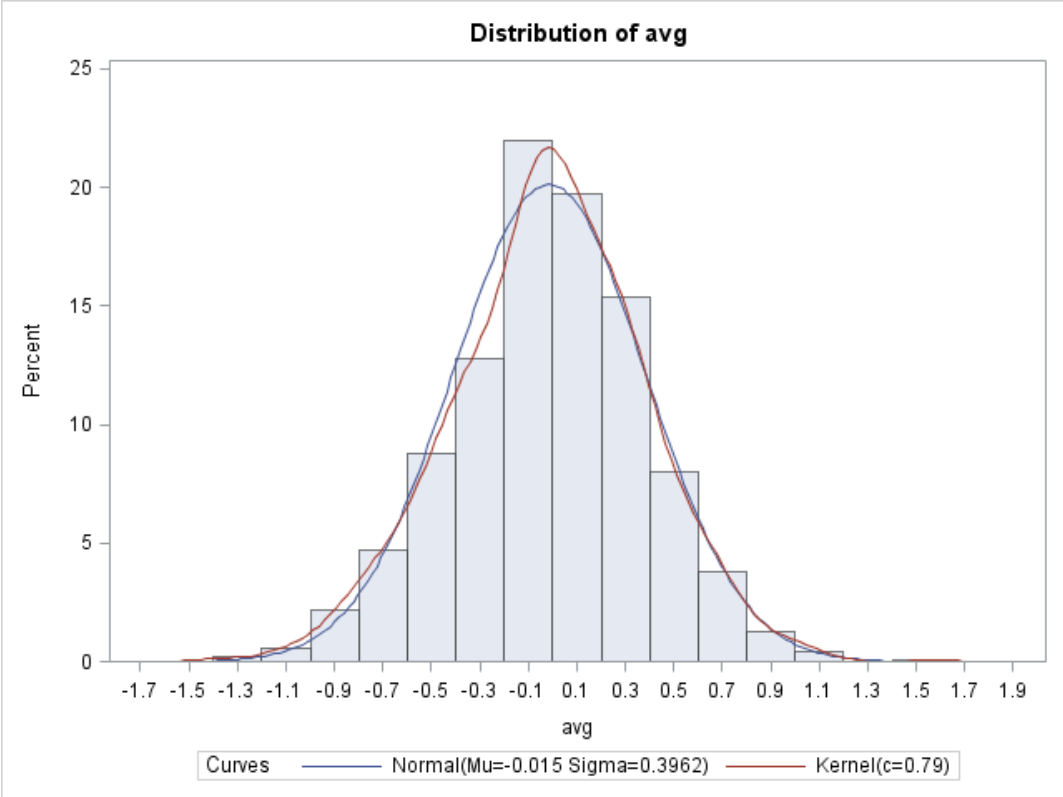
Normal Distribution repeats = 2

The MEANS Procedure

Analysis Variable : avg

N	Mean	Std Dev	Minimum	Maximum
1000	0.000544750	0.7209258	-2.0583407	2.9479325

n = 6



looks normal

Normal Distribution repeats = 6

The MEANS Procedure

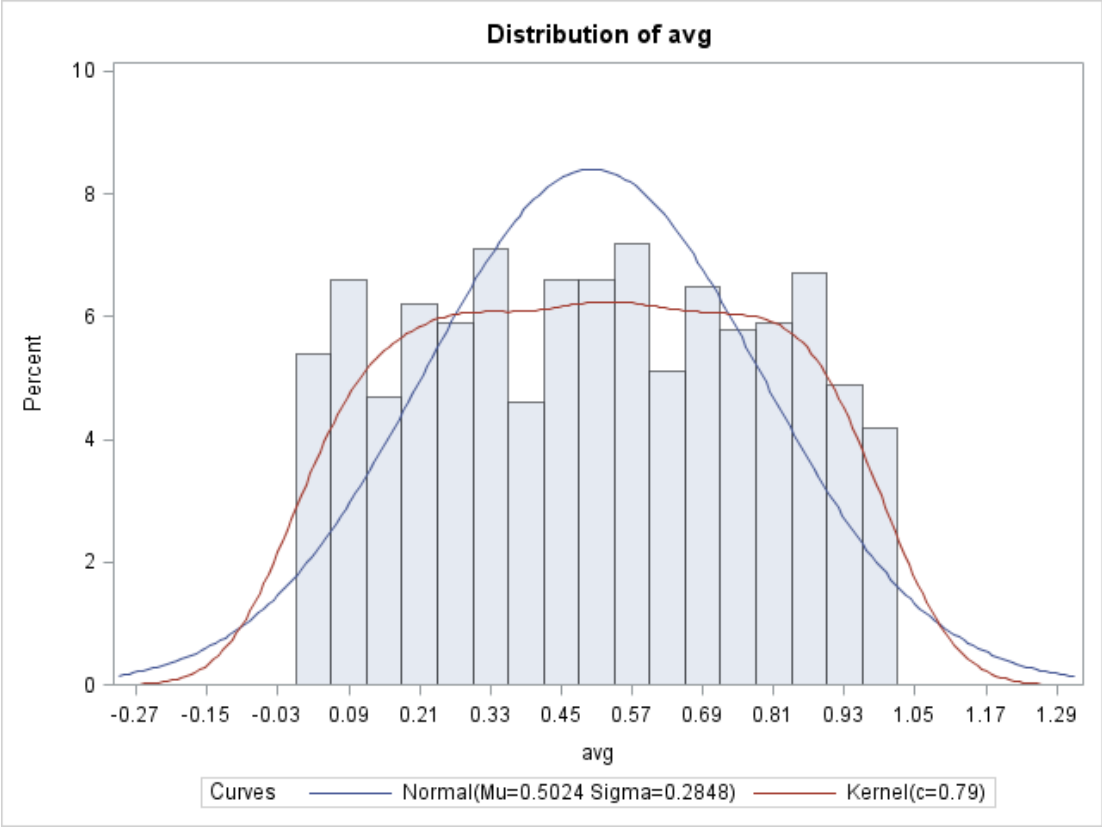
Analysis Variable : avg

N	Mean	Std Dev	Minimum	Maximum
1000	-0.0152099	0.3961601	-1.3799908	1.5750348

n	mean	theoretical mean	SD	theoretical SD
1	-0.0307342	0	0.9818657	$\frac{1}{\sqrt{1}} = 1$
2	0.000544750	0	0.7209258	$\frac{1}{\sqrt{2}} = 0.707107$
6	-0.152099	0	0.3961601	$\frac{1}{\sqrt{6}} = 0.40825$

It looks like the calculated values match the theoretical values.

Uniform:
n = 1



does not look normal

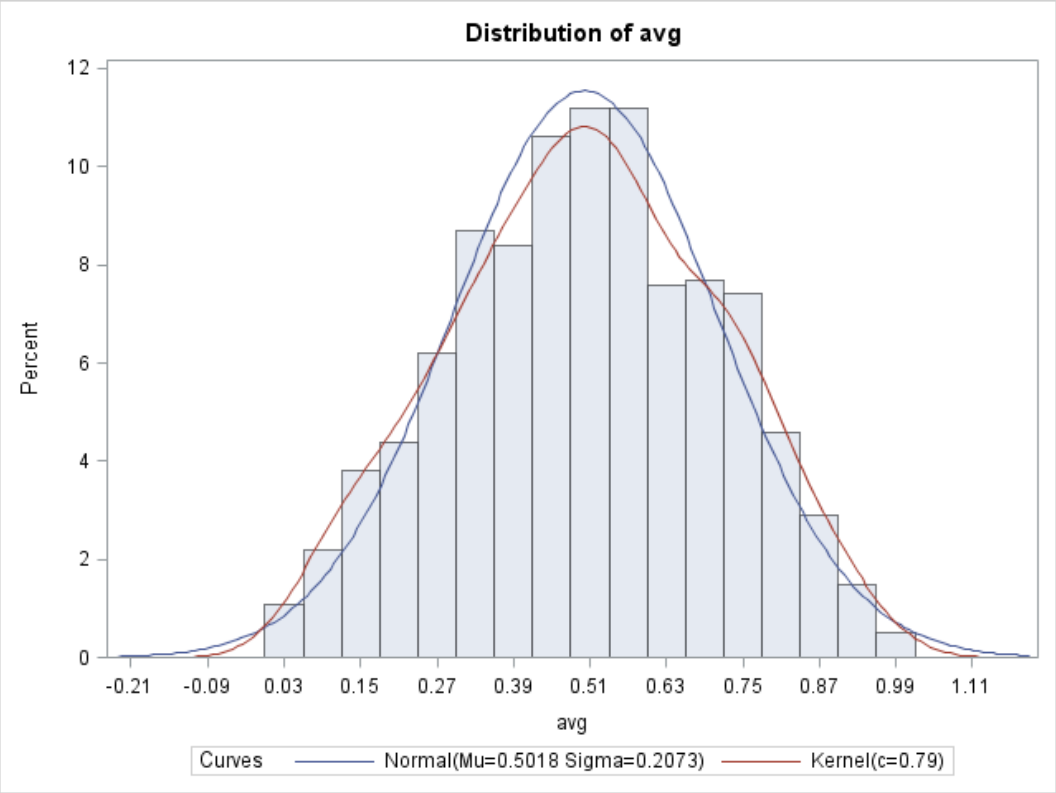
Uniform Distribution repeats = 1

The MEANS Procedure

Analysis Variable : avg

N	Mean	Std Dev	Minimum	Maximum
1000	0.5024465	0.2847927	0.0013425	0.9994848

n = 2



does not look normal but close

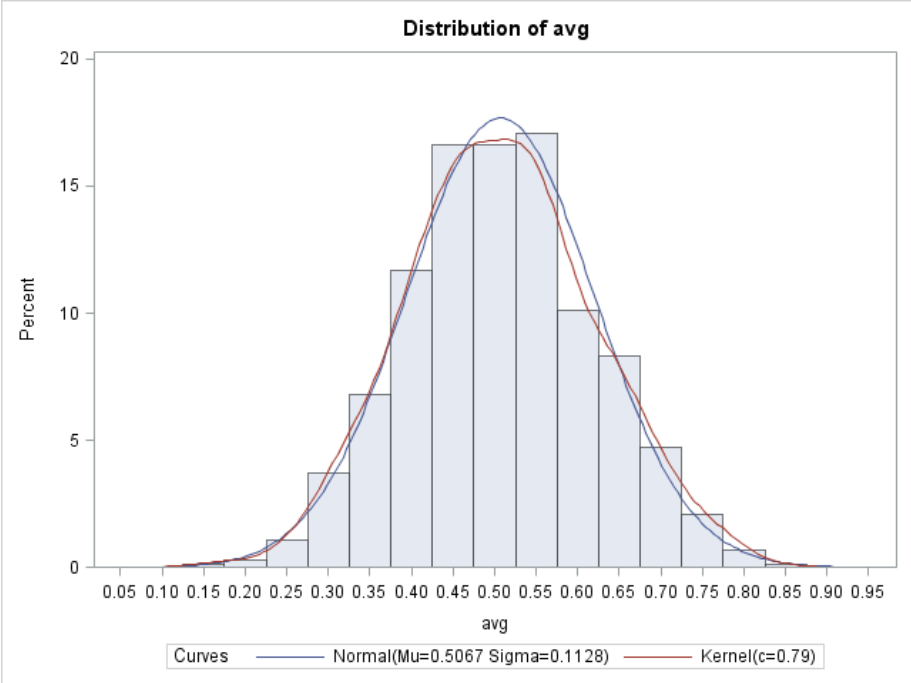
Uniform Distribution repeats = 2

The MEANS Procedure

Analysis Variable : avg

N	Mean	Std Dev	Minimum	Maximum
1000	0.5018143	0.2073308	0.0033105	0.9906815

n = 6



looks normal

Uniform Distribution repeats = 6

The MEANS Procedure

Analysis Variable : avg

N	Mean	Std Dev	Minimum	Maximum
1000	0.5066988	0.1128144	0.1453892	0.8510052

n	mean	theoretical mean	SD	theoretical SD
1	0.5024465	0.5	0.2847927	$\sqrt{\frac{1}{12}} = 0.288975$
2	0.5018143	0.5	0.2073308	$\frac{\sqrt{\frac{1}{12}}}{\sqrt{2}} = 0.204124$
6	0.5066988	0.5	0.1128144	$\frac{\sqrt{\frac{1}{12}}}{\sqrt{6}} = 0.11785$

It looks like the calculated values match the theoretical values.