Please note that the solution key is presented based on the assignment version where Part F is a bonus problem. If your group have five members, please see the assignment document to see how the points are allocated.

Lab 4 (100 points + 20 points BONUS) – Central Limit Theorem
Objectives: A better understanding of the Central Limit Theorem

This is a group lab so only one report should be submitted per group. There should be 3 – 4 people in each group. It is acceptable that each person does one or two distributions and then discuss the results with the rest of their group to write a combined summary statement. Different software packages may be used in this lab.

The following is a review from Chapter 7. To help you understand the Central Limit Theorem, you are going to be simulating the distribution of the sample mean ($\bar{X}$) for four different distributions: normal, uniform, gamma, and Poisson. The distribution of $\bar{X}$ is called a sampling distribution. For each distribution, the sampling mean and standard deviation are:

$$\mu_{\bar{X}} = \mu_X, \ \ \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} \qquad \text{Equations 1}$$

where $\mu_{\bar{X}}$ is the mean of the sampling distribution, μ_X (or μ) is the mean of the population, $\sigma_{\bar{X}}$ is the standard deviation of the sampling distribution, σ_X (or σ) is the standard deviation of the population, and n is the number of data points averaged. When n is large, the distribution of $\bar{X}$is approximately normal, that is

$$\bar{X} \dot\sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{Equation 2}$$

Here is how you will visualize the sampling distribution of the mean:

For each of the distributions, begin by creating 1000 random samples, each of size $n.$ Then, for each of the 1000 samples, you will calculate the sample average, $\bar{X}$. After calculating 1000 different $\bar{X}$'s, you will be able to make a *histogram* and *normal probability plot* of the $\bar{X}$ values and thus visualize the distribution of $\bar{X}$. The goal is to see what value of n is large enough for the distribution of $\bar{X}$ to become approximately normal. Notice that this value of n depends on the population distribution. To determine the value of n required, your simulations will start from a small n and progress to larger n's. You will assess the normality based on the plots for each n and continue until either you have finished the values of n listed or increased the values until observing sufficient normality in the plots. The tutorial explains how to do this for each given n.

For each of the distributions below, you will complete the following:

1. (5 points) Code
   You only need to provide one code listing for each distribution (i.e. you don't need to repeat the code for each choice of n).

2. (10 points) Histogram/normal probability plots
   For each of the values of n, submit a histogram (with the two colored curves) and a normal probability plot. For each of the graph pairs, indicate whether they appear sufficiently normal

or not. No explanation is required. Make sure you increase n until the distribution of $\bar{X}$ appears sufficiently normal.

3. (5 points) Summary table
This table contains the experimental mean and standard deviation calculated from the data (output is required for each value of n) and the theoretical mean and standard deviation calculated from Equations 1 (with work for one of the values for each distribution where $n \neq 1$). The format for this table for Part B is below. Make sure you increase n until the distribution of $\bar{X}$ appears sufficiently normal.

For standard normal Part B:

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				
6				
10				

The distributions and the values of n that you are required to use (the number of samples to average) are below: I have included the population mean and standard deviation for the distribution that we have not covered in class.

A. (10 points) Online Prelab

B. (20 points) Standard Normal Distribution. $n = 1$, 2, 6 and 10.

C. (20 points) Uniform distribution over the interval (0, 5). $n = 1$, 2, 10 and 15.

D. (20 points) Gamma distribution with parameters $\alpha = 3$ and $\beta = 2$. $n = 1$, 5, 10, 20, 40, and continue in intervals of 20 if needed until the shape becomes normal. This distribution has population mean and standard deviation of $\mu = \frac{3}{2}$, $\sigma = \frac{\sqrt{3}}{2}$.

E. (20 points) Poisson distribution with parameter $\lambda = 3$. $n = 1$, 5, 10, 20, 40, and continue in intervals of 20 if needed until the shape becomes normal.

F. (BONUS: 20 points) Exponential with parameter $\lambda = 3$ $n = 1$, 5, 10, 20, 40, and continue in intervals of 20 if needed until the shape becomes normal.

G. (10 points) Concluding remarks.

Please write a conclusion summarizing the information in Parts B, C, D, and E (and F if performed). Please include comments on each of the following (at least one sentence or table for each):

- Whether Equations 1 are valid for all values of n for all of the distributions
- What happens as n increases for each of the distributions?
- **Include a table** of what value of n is considered 'large' for each distribution.
- Finally write a concluding sentence that will provide a 'rule of thumb' for an estimate of what value of n is 'large enough' – in the sense that $\bar{X}$ becomes approximately normal – given the shape of a specific parent distribution. This way, if you know the shape of the population distribution, you will have a feel of how large n should be for $\bar{X}$ to be approximately normal and hence justify the appropriateness of many statistical tests.
-

Solution

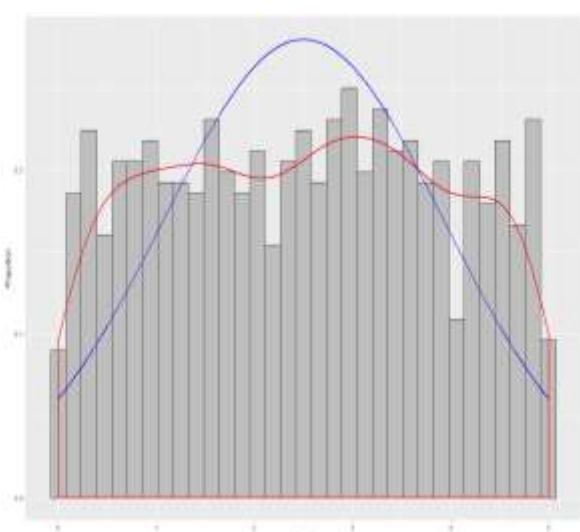
The theoretical values in the summary table should always be close to the experimental values. This fact is not dependent on whether the shape of the sampling distribution is normal or not. The fact that the distribution is discrete also does not change the result.

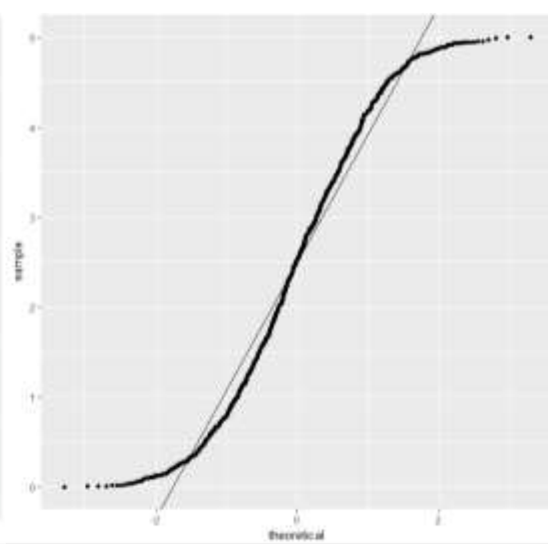
All of the distributions will eventually become normal with the only thing that changes is the value of n when this occurs. The table of possible n's is included below:

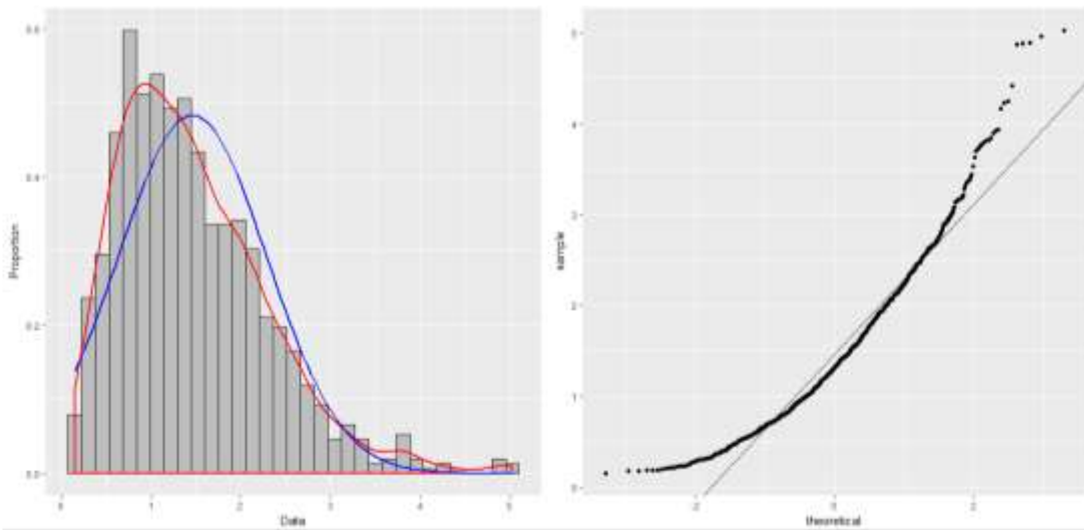
Distribution	Normal	Uniform	Gamma $\alpha = 3$, $\beta = 2$	Poisson $\lambda = 3$	Exponential $\lambda = 3$
range of n	always	~10	20 - 40	10 - 40	60 - 80

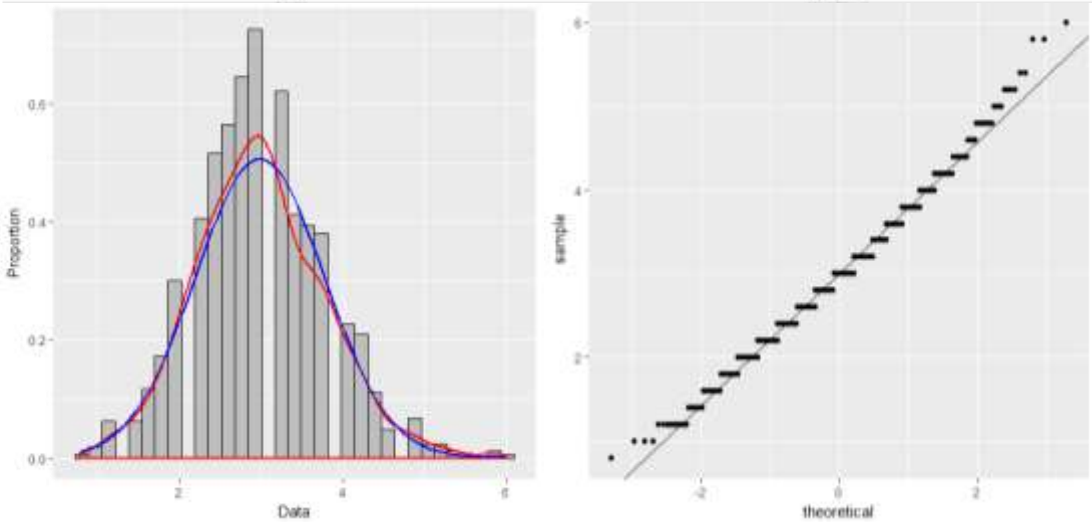
The rule of thumb depends on your observations, but in general 30 – 40 (like what is stated in the textbook) is good enough except if the distribution is highly skewed like the exponential.

<u>Not Normal</u>

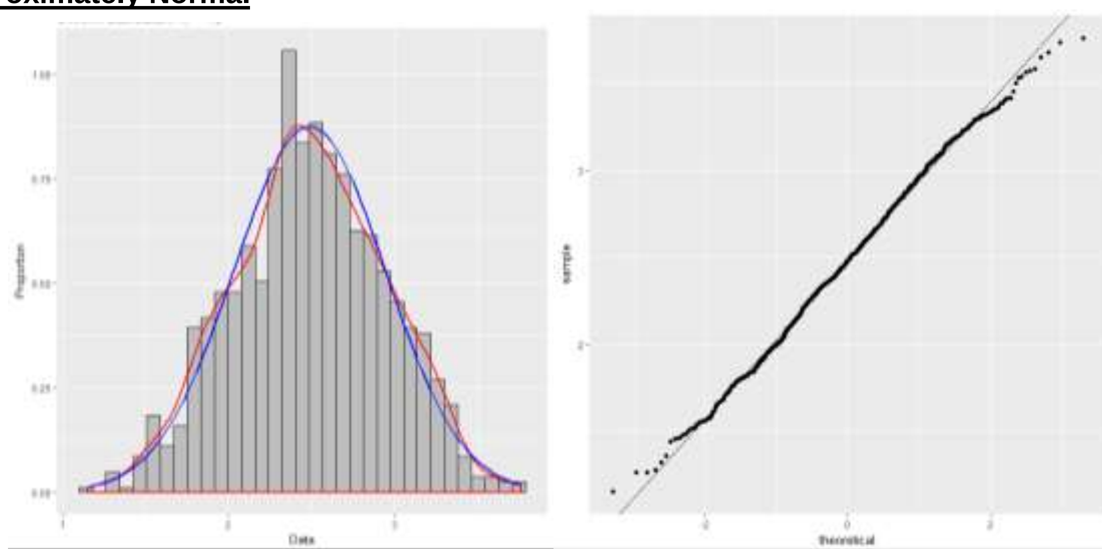


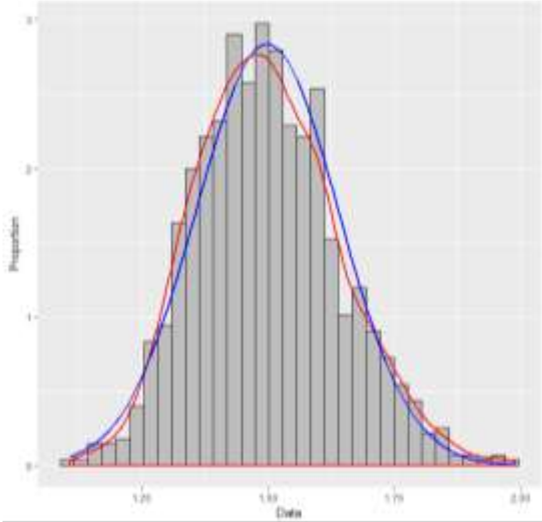


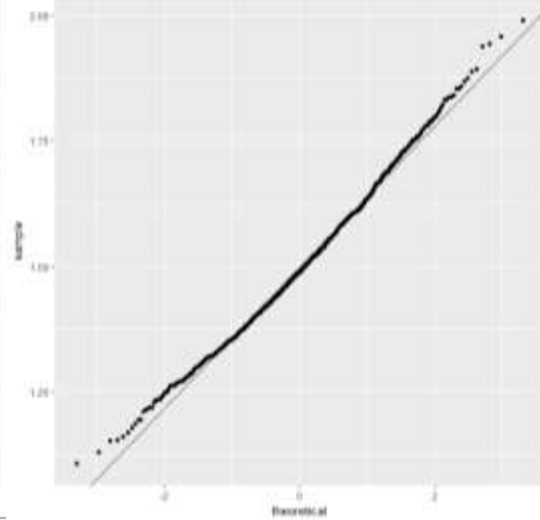


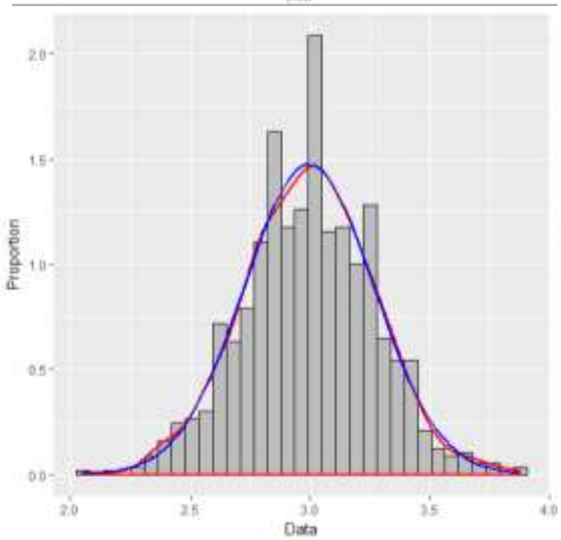


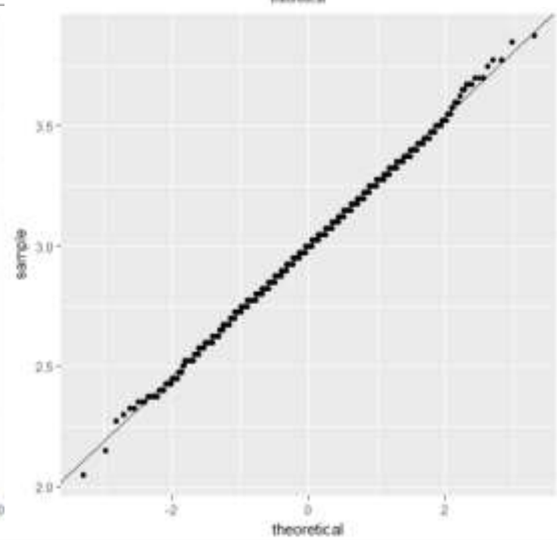
Approximately Normal

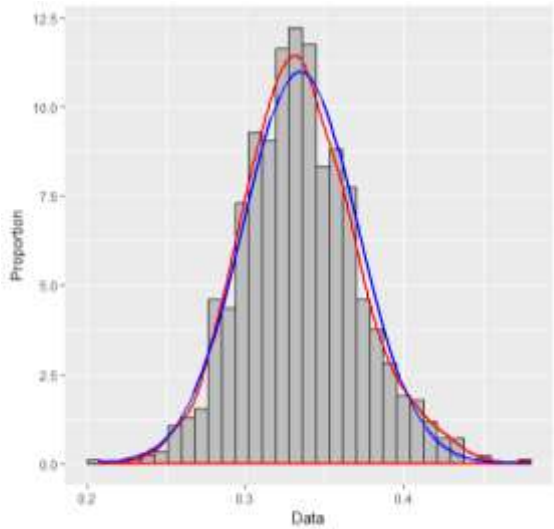


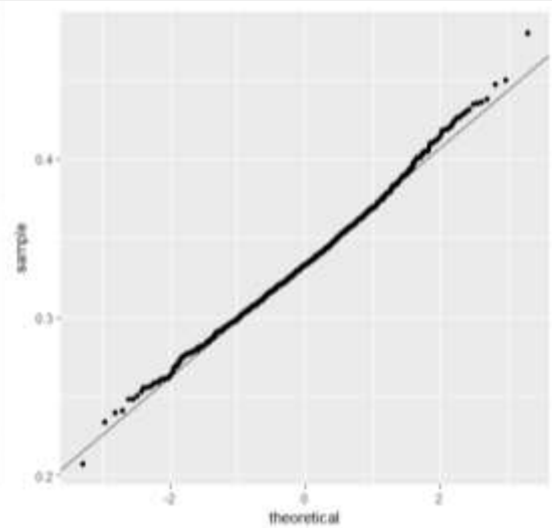

Proportion
Data


sample
theoretical


Proportion
0.0
0.5
1.0
1.5
2.0
2.0
2.5
3.0
3.5
4.0
Data


sample
2.0
2.5
3.0
3.5
-2
0
2
theoretical


Proportion
0.0
2.5
5.0
7.5
10.0
12.5
0.2
0.3
0.4
Data


sample
0.2
0.3
0.4
-2
0
2
theoretical