Stats Lab 2

Ruolin Wang, Ruth Risberg, Iman Ebrahimi

October 2022

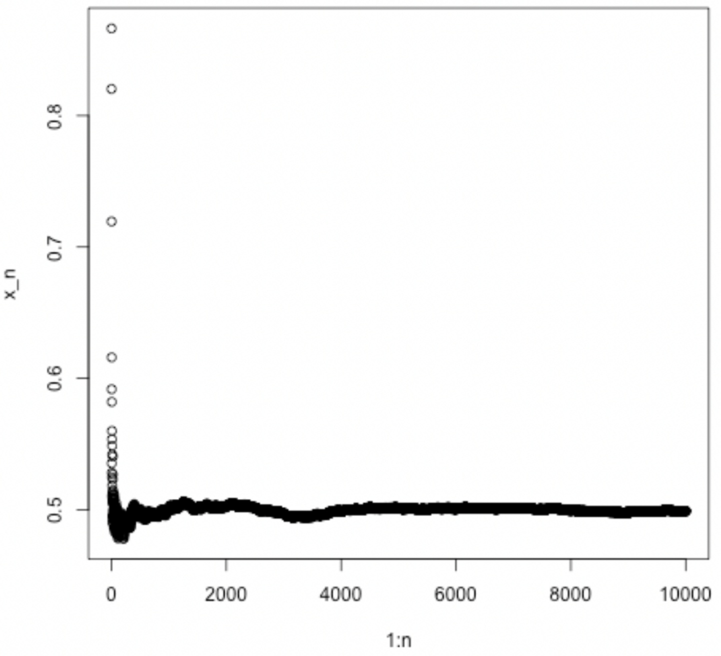
1 The law of large numbers

1. Generate 10000U$(0, 1)$ distributed r.v.'s and plot the sequence of averages $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ against n.

Q1 What does the graph seem to converge to? Calculate (analytically) $\mathbb{E}\left(X_1\right)$.

The graph seem to converge to $\frac{1}{2}$.
Since X_i is $Un(0, 1)$distributed, therefore

$$\mathbb{E}\left(X_1\right) = \int_0^1 x\frac{1}{1-0} = \frac{1}{2}$$



Q2 How does this relate to the statement of the LLN?

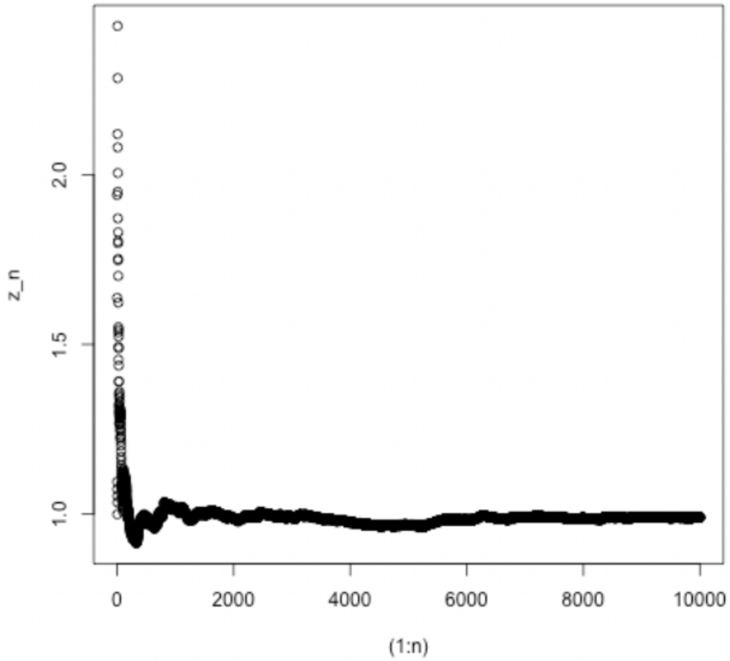
The average value of the sum of identical independent uniform distributions converges to the expected value of the single distribution in probability when n is large.

2. Generate $10000 N(0,1)$-distributed r.v.'s and form $Z_n = n^{-1}\sum_{i=1}^{n} X_i^2$, and plot Z_n versus n.

Q1 What does the graph seem to converge to? Calculate $\mathbb{E}\left(X_1^2\right)$.

The graph seems to converge to 1.

$$\mathbb{E}\left(X_1^2\right) = \int_{-\infty}^{+\infty} x^2 \frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-0}{1}\right)^2} = \frac{1}{2} - (-\frac{1}{2}) = 1$$



Q2 How does this relate to the statement of the LLN.

The average value of the sum of identical independent normal distributions squared seems to equal $\mathbb{E}\left(X_1^2\right)$. So Z_n converge to $\mathbb{E}\left(X_1^2\right)$ in L^2 when n is large.

Q3 Explain how this can be used to give approximations for $\mathbb{E}(g(X))$ for functions g
(see also part 4 below.) For which functions g can this be done? Give a restriction on g that ensures that this can be done. Analytically when we apply some function $g(x)$ to the numbers from the interval $[-l, l]$, tt has to be finite continuous function defined on $[-l, l]$. And continuity asserts that g(x) could be calculated and $g(x)$ has to be a finite number. In our case it means that functions like $\frac{1}{x}$ or $ln(x)$ are prohibited.
We can use the Monte-Carlo approximation like question 4
3. Simulate 10000 coin tosses, with

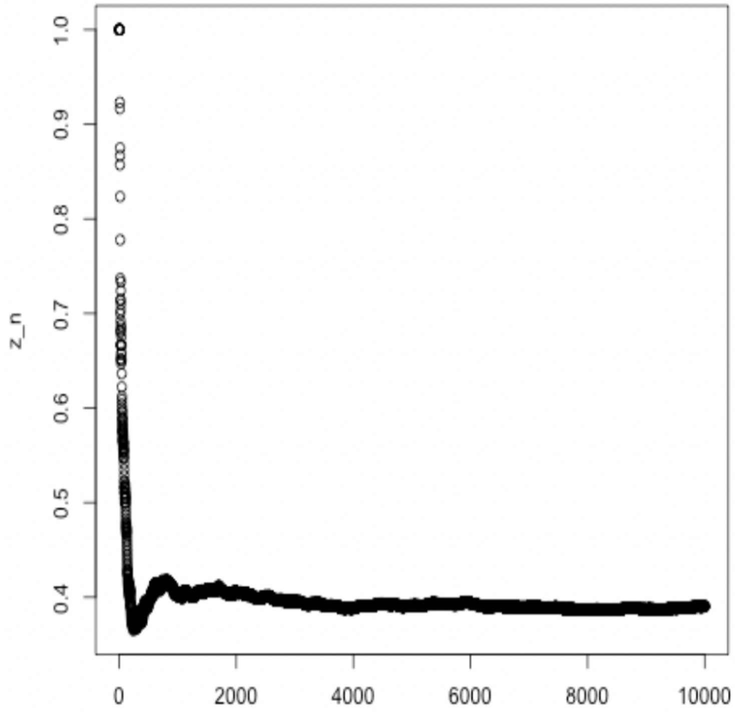
$$X_i = \begin{cases} 0 & \text{if tail up in toss number } \quad i \\ 1 & \text{if head up in toss number } \quad i \end{cases}$$

for $i = 1, \ldots, 10000$. The model is that $P($ head $) = 0.39$ och $P($ tail $) = 0.61$. Form $X_n = n^{-1} \sum_{i=1}^{n} X_i$ and plot versus n.

Q1 What does the graph converge to? Calculate (analytically) $\mathbb{E}(X_1)$.

The graph seems to converge to 0.4(0.39)very closely.

$$\mathbb{E}(X_1) = 0 \times 0.61 + 1 \times 0.39 = 0.39$$



Q2 How does the result relate to what the LLN says?

The average value of sum of identical independent Bernoulli distributions with parameter 0.39 converges to expected value of a single Bernoulli distribution when n is large

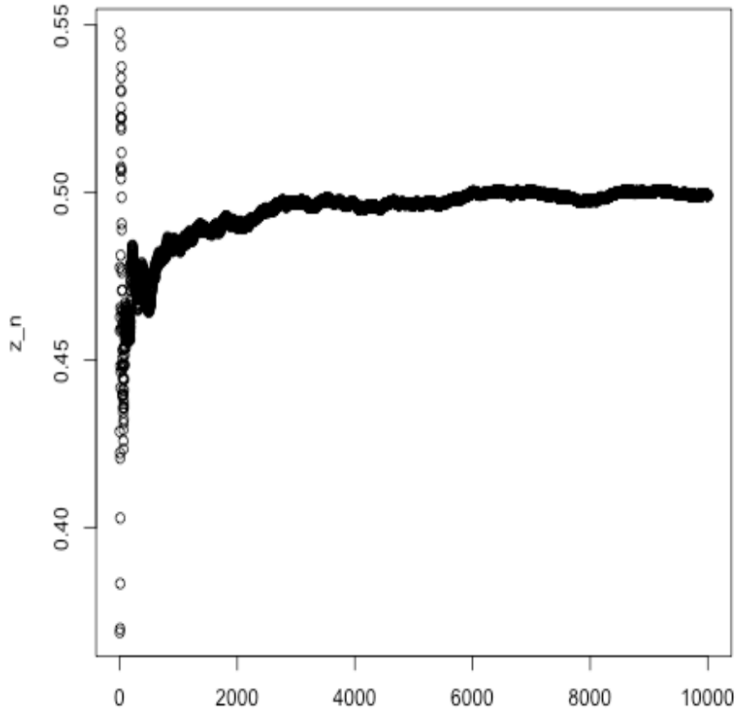
Q3 Explain how this can be used to get approximations to $P(A)$ for arbitrary events A.

To get approximations to $P(A)$ for arbitrary events A, we can sum a large number of identical independent distributions on that event and then take an average value of them. Then the graph will converge to $P(A)$ which gives us an approximation for unknown $P(A)$.

4. Let us study the integral $I = \int_0^1 e^{-x^2} dx$. There is no closed form expression for I, since we lack an elementary expression for the anti-derivative of e^{-x^2}. We will do a so called Monte-Carlo approximation of I. We can write $I = \mathbb{E}(g(X))$, with $X \in \mathrm{U}(0,1)$ and $g(x) = e^{-x^2}$. Run the following commands.

Q1 What numerical value does the graph seem to converge to?

The value seems to converge to 0.5 base on the code.



Q2 Use the LLN to justify the approximation to I.

$E[g(x)] = \int g(x) f_x(x) dx$, since x is uniformly distributed.

$$\begin{aligned} \int f_x(x) = \frac{1}{b-a}, \int_a^b g(x) dx &= (b-a) \int g(x) \frac{1}{b-a} dx \\ &= (b-a) \cdot E[g(x)] \\ &\approx (b-a) \cdot \frac{1}{n} \sum_{i=1}^{n} g(x_i) \end{aligned}$$

So as we can see the integral can be estimated using law of large number.

2 The central limit theorem (CLT)

Q1 Derive an analytic expression for $\mathbb{E}(F_n(x))$.

$$\begin{aligned}\mathbb{E}(F_n(x)) &= \mathbb{E}\left(\frac{1}{n}\sum_{n}^{i=1} 1_{\{X_i \leq x\}}\right) \\ &= \mathbb{P}(X_i \leq x)\end{aligned}$$

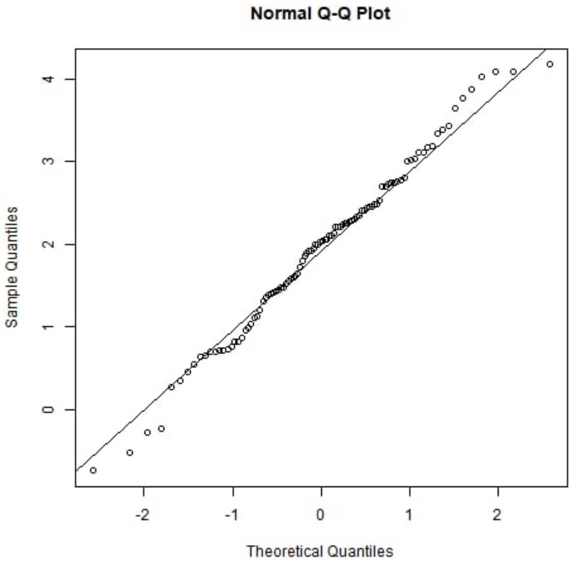
Q2 Use the LLN to explain why $F_n(x)$ can be used as approximation of F.

The LLN says that for a sequence of random variables like our sequence of X_i, the value of the function F_n will be between $F - h$ and $F + h$ where h is some fixed arbitrarily small number, with a high probability given that n is large enough. This means that we can give F_n a high chance of being arbitrarily close to F given a high enough value of n.

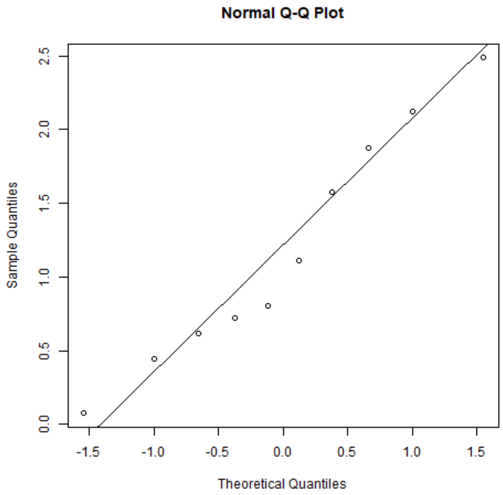
Q3 For a fixed x_0: In what sense does $F_n(x_0)$ converge to $F(x_0)$ when $n \to \infty$?

The probability that $F_n(x_0)$ will be in any fixed interval around $F(x_0)$ goes to infinity.

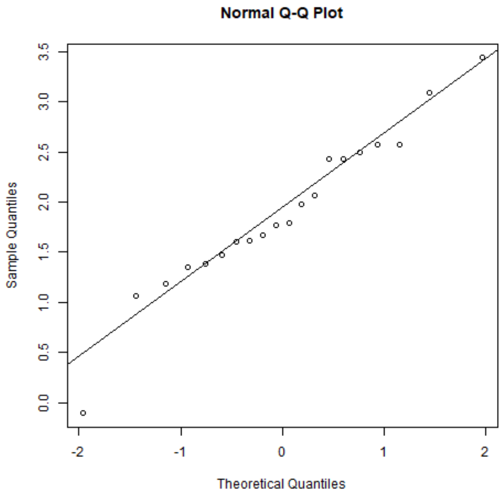
1. Generate $n = 100$ normal distributed r.v.'s with expectation $\mu = 2$ and variance $\sigma^2 = 1$. Do a normal probability plot of these.



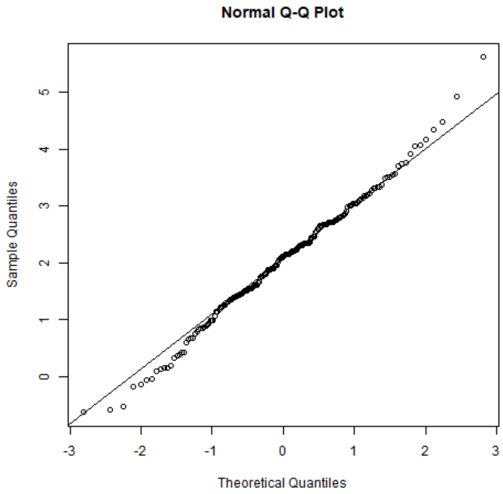
Q1. Do the above with values $n = 10, 20, 200, 1000$. What is the outcome?



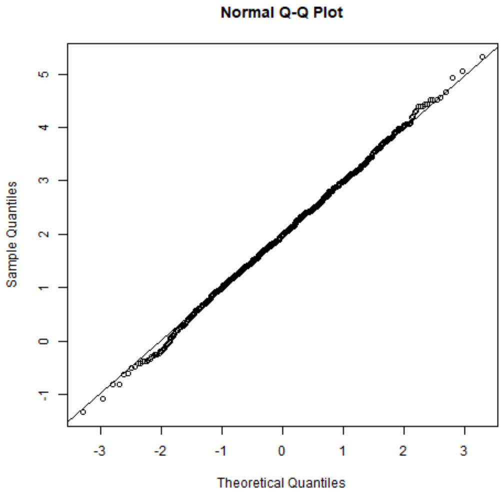
(a) n = 10



(b) n = 20



(c) n = 200

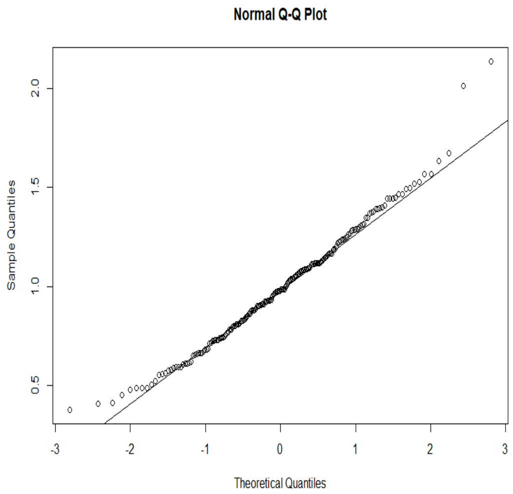


(d) n = 1000

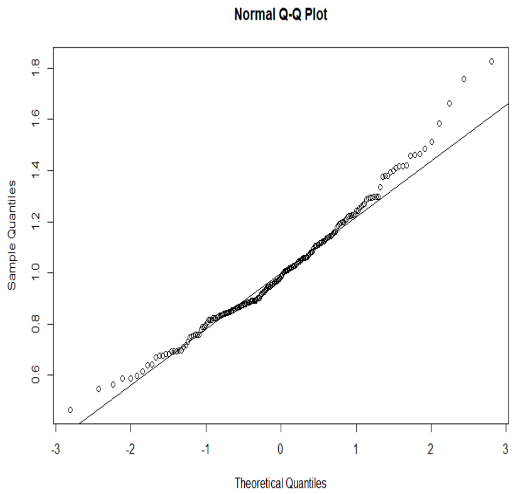
We observe that as the sample size (n) gets larger, the distribution of the sample means tends towards a normal distribution.

2. Generate $n = 100$ exponentially distributed r.v.'s with expectation 1 and form their average $\bar{X}_n$. Do this $m = 200$ times so that you get 200 averages $\bar{X}_n^1$, ..., $\bar{X}_n^{200}$

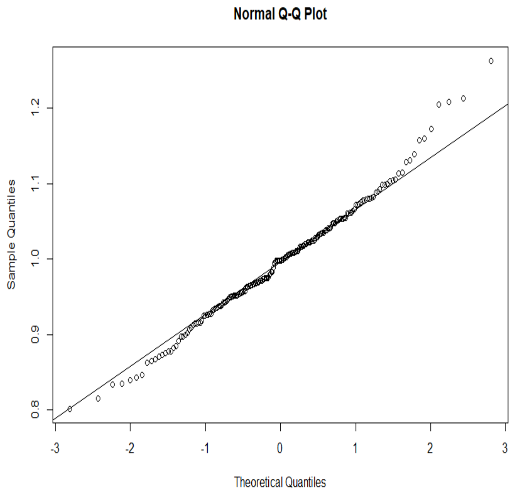
Q1. Do the above with values $n = 10, 20, 200, 1000$. What is the outcome?



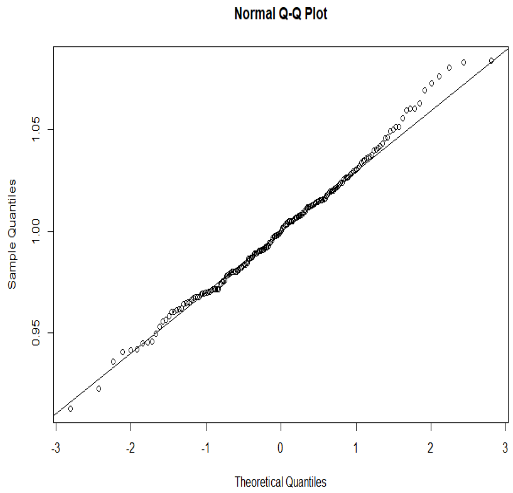
(a) n = 10



(b) n = 20



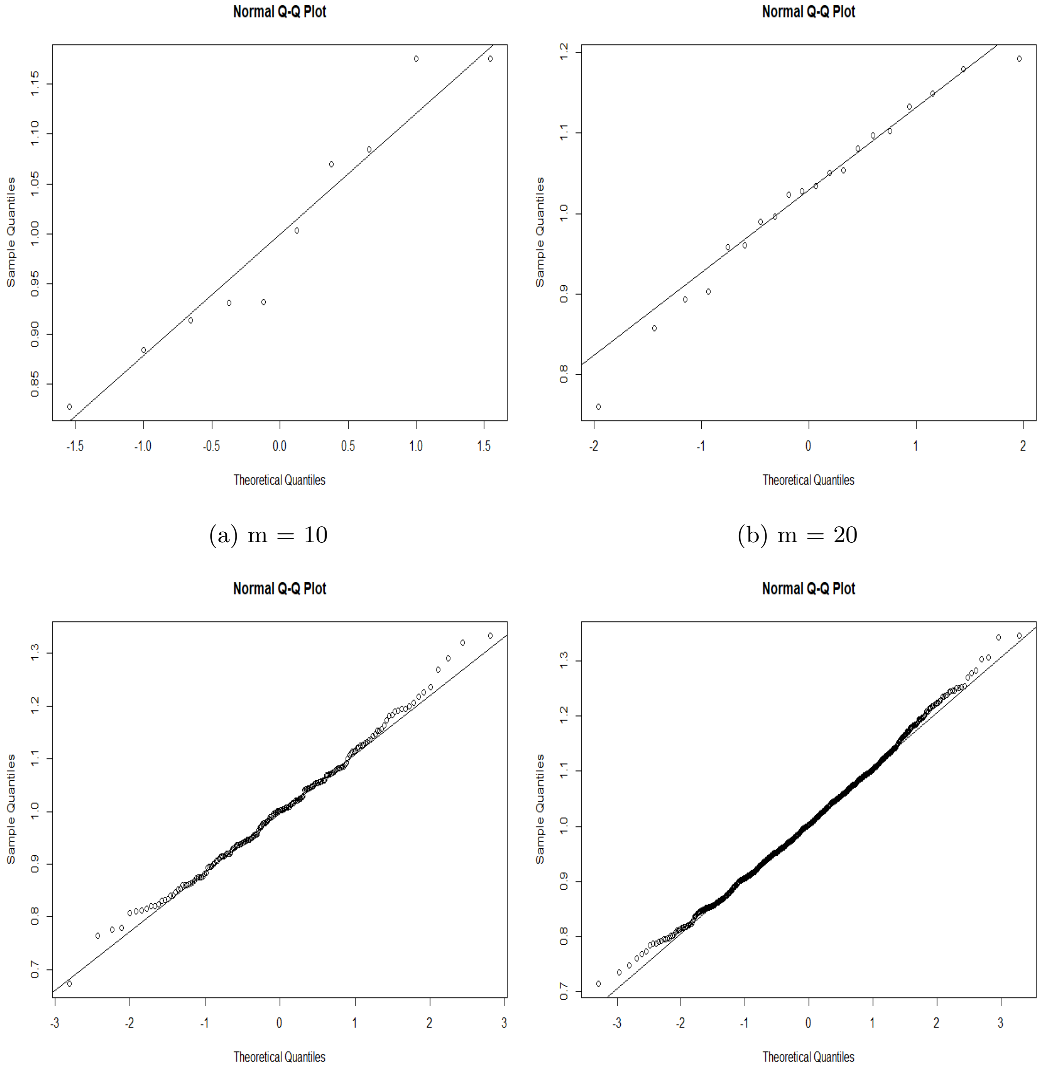
(c) n = 200



(d) n = 1000

The data points become less spread out for higher values of n. This can be seen by looking at the scale of the vertical axes.

Q2. Do the above with varying m. What is the outcome?



(a) m = 10

(b) m = 20

(c) m = 200

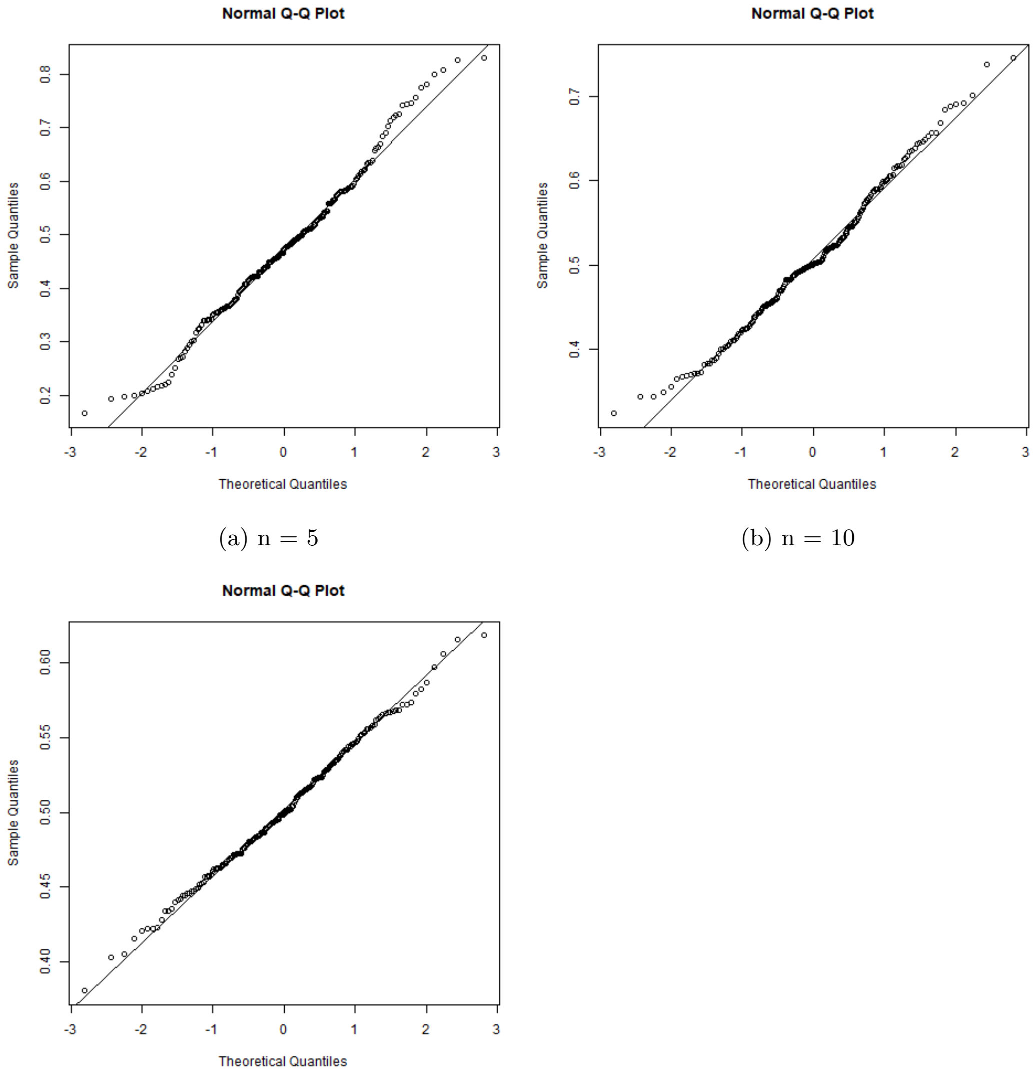
(d) m = 1000

m is the number of data points, so a higher value of m means more points in the plot and it becomes easier to see potential trends.

Q3. Explain how your results relate to the CLT. Explain what happens qualitatively as $n \to \infty$. and as $m \to \infty$

Since the n random variables whose average makes up each of the m points in the plots fulfill the requirements of the Central Limit Theorem, testing if these points follow a normal distribution is a good test to illustrate the CLT. As n goes to infinity, each point becomes the average of more and more separate random variables, and thus by the LLN it becomes closer and closer to its expected value. This means that when n goes to infinity, the value of each point will be exactly its expected value. As m goes to infinity, the number of points will be infinite so we would need to plot the relative density of points instead of a scatter plot. Using this new plot we would be able to see exactly how big the probability is for a point to be in any given area.

3. Do the above with $U(0,1)$-distributed r.v.'s and varying $n = 5, 10, 50$. Answer Q1, Q2 and Q3 as above for this case.



(a) n = 5

(b) n = 10

(c) n = 50

Q1. By observing the scale of the vertical axes, it becomes clear that as n increases, the points tend more towards the mean and are less spread out.

Q2. We observe that the greater our number of data points (m) becomes, the more points are there to be seen. Hence, the potential trends are easier to observe as m increases.

Q3. As n goes to infinity, it appears that every point tends to being the average of X_n random variables. According to the Law of Large Numbers, X_n tends to the expected value $E(x)$. As m goes to infinity, the number of points increases. Hence by use of a plot to show the relative density of points, it becomes more clear how great the probability for a point could be in any given area.

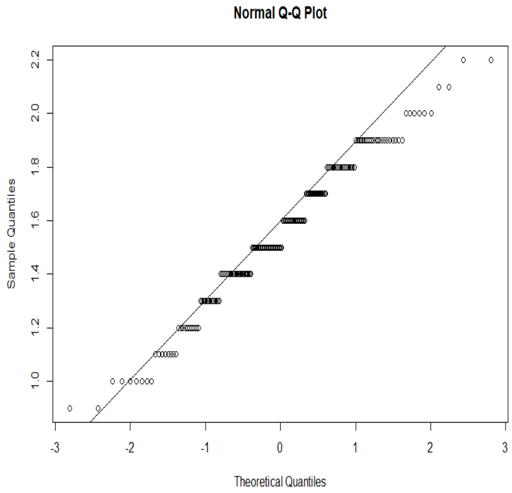
Q4. What, if any, is the difference with the previous exercise?

Exercise number 2 studies exponentially distributed random variables, with the conclusion that "although the Random Variables were exponentially distributed, their means are normally distributed."
In this exercise we had a uniform distribution. Despite that all the $X_n's$ are uniformly distributed (i.e there is an equal probability of selecting values between 0 and 1), The means are again normally distributed (Which means the CLT does not care how the r.v.s are distributed, the means' distributions are always Gaussian).

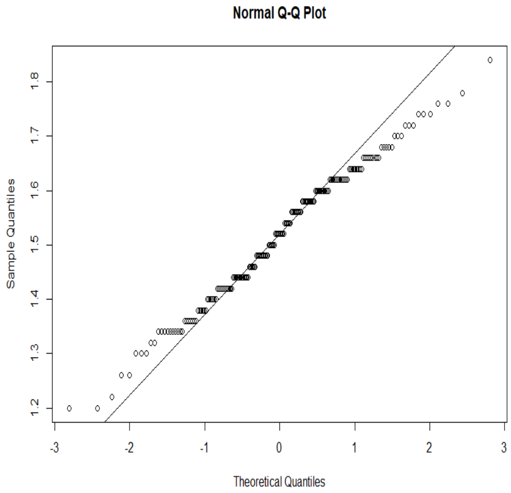
We can conclude that the CLT applies in the same way even though the r.v.s that we take the average of have different distributions.

4. Generate $n = 100 Bin(nn, pp)$-distributed r.v.'s with $nn = 3$, and $pp = 0.5$ and their average $\bar{X}_n$. Do this $m = 200$ times. Do normal probability on these.

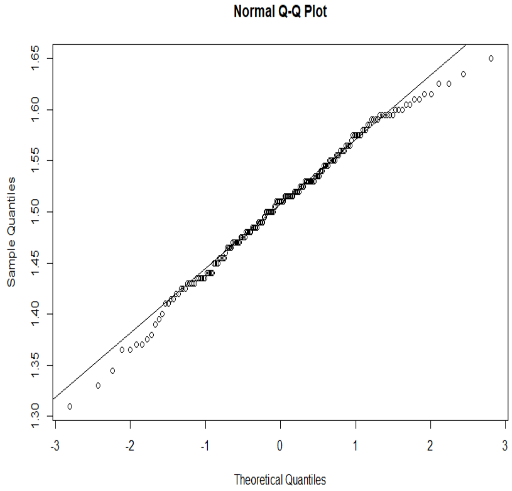
Q1. Vary $n = 10, 50, 200$. What is the result?



(a) n = 10



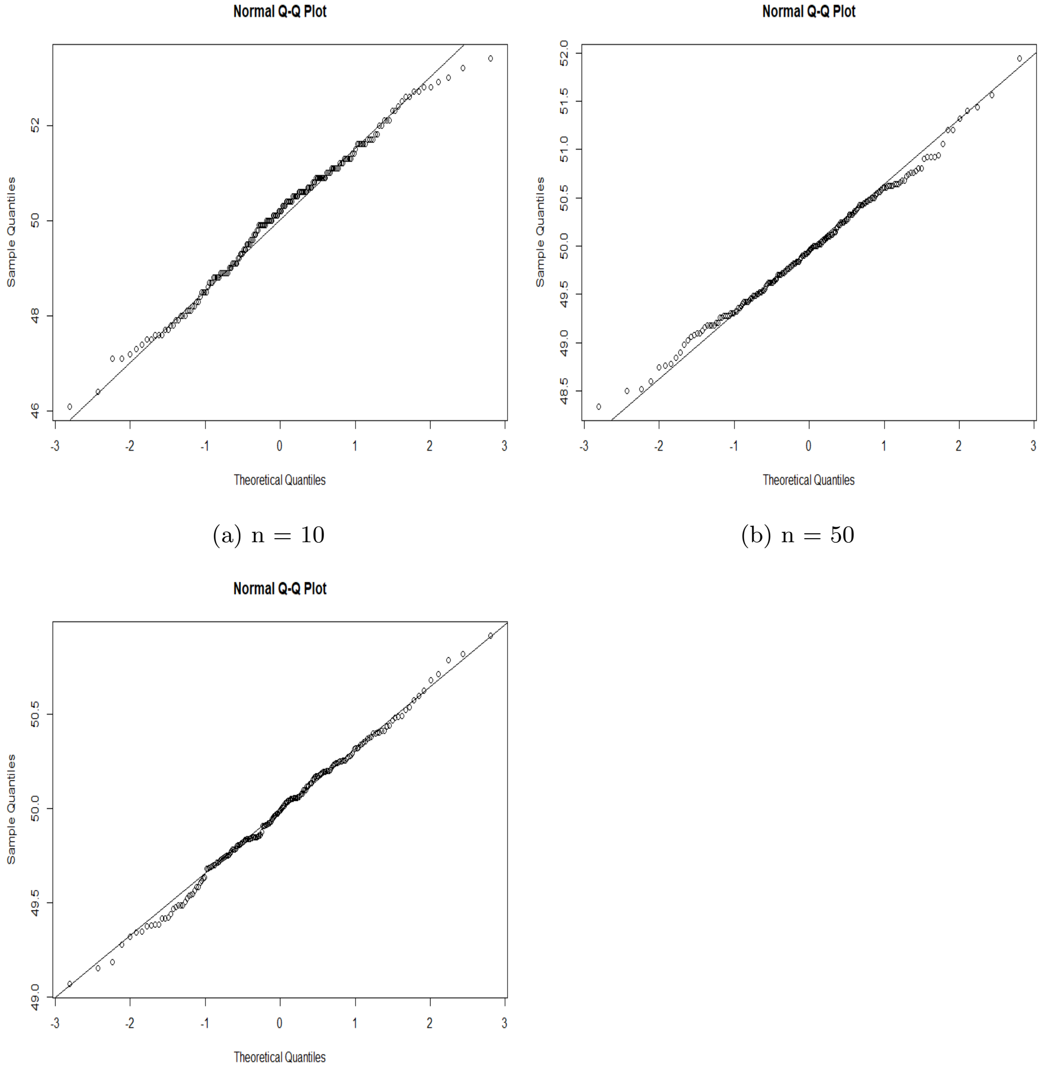
(b) n = 50



(c) n = 200

For greater values of n the graphs looks a lot smoother since it includes data of averages of a lot more values.

Q2. What happens if you change to $nn = 100$, and vary $n = 10, 50, 200$?



(a) n = 10

(b) n = 50

(c) n = 200

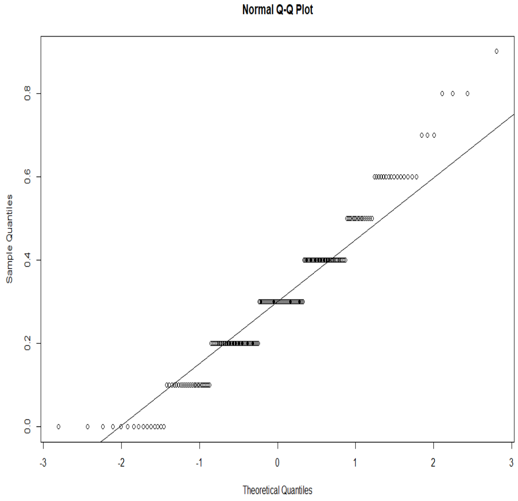
A real life example of this binomial distribution would be a series of nn fair coins tossed n times. Using this analogy we can easily see why a greater value of nn makes the random variables take greater values and a greater number of different possible values.

Q3. Can you use the CLT to explain the qualitative difference between the distributions $Bin(3, 0.5)$ and $Bin(100, 0.5)$?

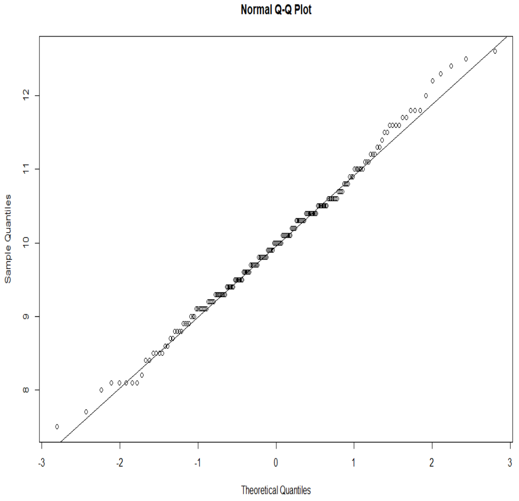
The difference between n and nn is simply that the random variables are multiplied by nn but not by n. So an increase of nn essentially means an increase of the amount of random variables we take the average of before applying the CLT. Therefore, an increased value of nn should mean a better approximation of a normal distribution.

5. Try to imagine what can happen if you change the binomial distribution in the previous exercise to $Bin(100, 0.1)$. Think first, and then do a run and answer Q1 and Q2 in the previous exercise.

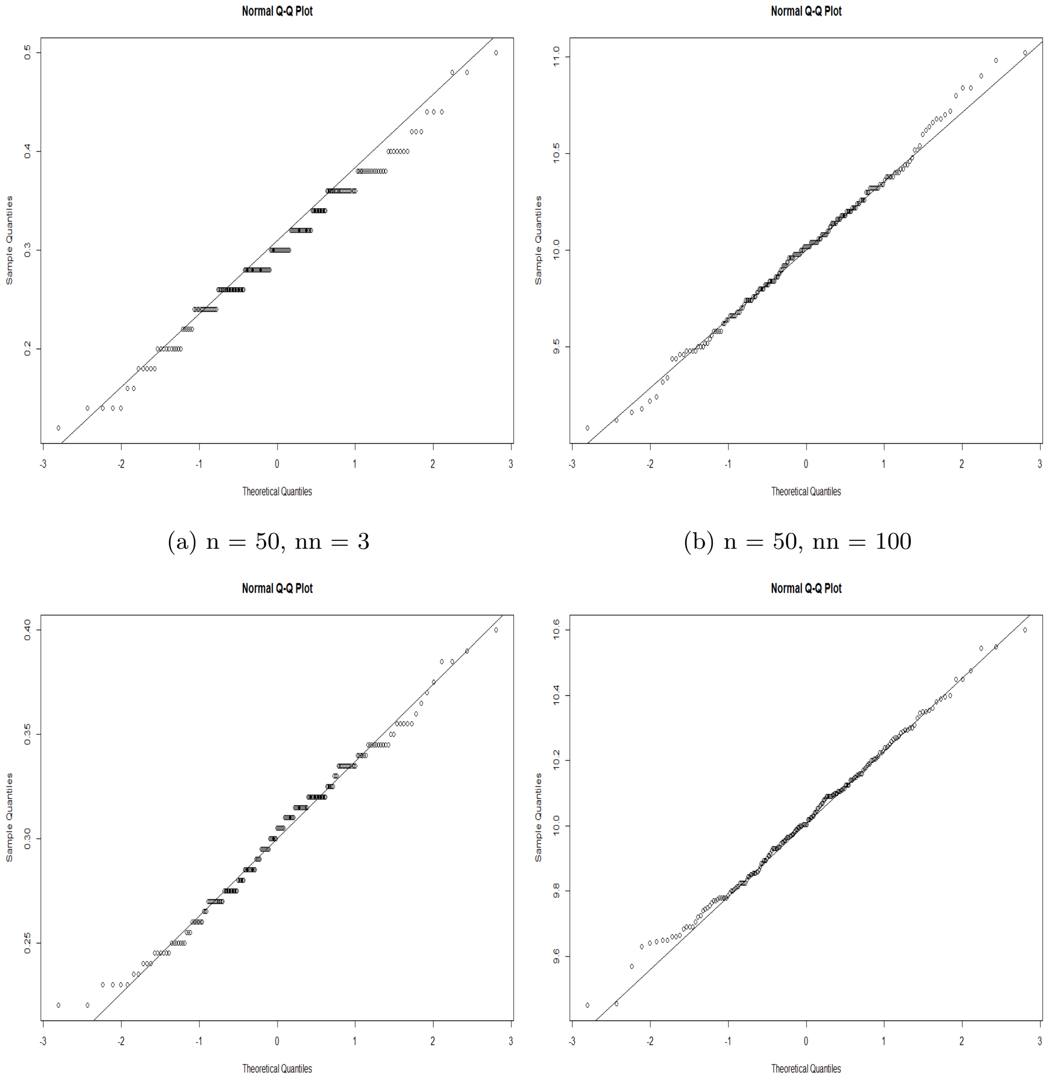
Q1 and Q2
As also observed in the previous exercise, the increase of n would result in a much smoother graph, the distance between every "step" of the graph becomes smaller as it consists data of more average points. The difference from the last exercise is that the random variables take smaller values.(See more graphs on next page)



(a) n = 10, nn = 3



(b) n = 10, nn = 100



(a) n = 50, nn = 3

(b) n = 50, nn = 100

(c) n = 200, nn = 3

(d) n = 200, nn = 100