PART II

Validating Central Limit theorem by Simulation of samples from Binomial Distribution and approximating by Poisson and Normal curve

Central Limit Theorem

Let $\{X_1, X_2, \dots, X_n\}$ be a random sample of size n, i.e. a sequence of independent and identically distributed random variables drawn from a population distribution with expected values given by μ and finite variance given by σ^2.

Define the sample average by, $S_n := \frac{X_1+X_2+\cdots+X_n}{n}$. By the law of large numbers, it shows that if $n \to \infty$, its expected value turns out to be μ. More precisely, it states that as n gets larger, the distribution of difference between the sample average S_n and μ, multiplied by $\sqrt{n}$ (i.e. $\sqrt{n}(S_n - \mu)$), approximates normal distribution with mean 0 and variance σ^2. This is same as stating that, S_n approximately follows normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

Approximating Binomial by Normal

In this section, I will try to analyse the effects of central limit theorem using Binomial distribution. Basically, the population is a set of Bernoulli trials and we know that sum of independent Bernoulli trials follows Binomial distribution. Hence, as the central limit theorem suggests, we will try to fit Normal distribution by generating random samples from binomial distribution.

For generating random samples from binomial distribution, I have considered seed as 13 and generated 1000 sample observations from each class.

Symmetric Binomial Distribution (p = ½):

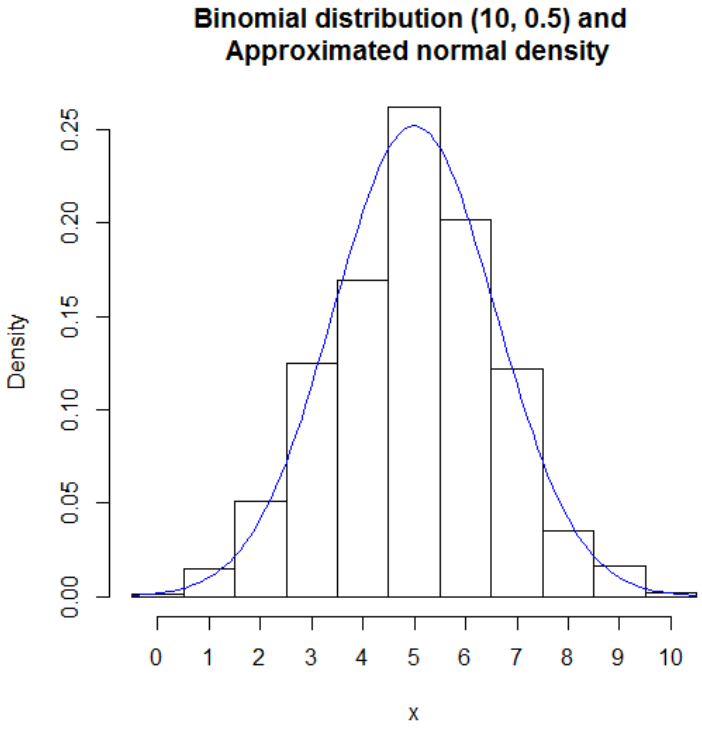
We will first consider symmetric case of binomial distribution and then will try to approximate that by normal curve.

Binomial (n = 10, p = 1/2)

We have expected mean as (n*p) = 5 and variance as (n*p*q) = 2.5. We have 11 different classes for 0 to 10, and the following is the table containing expected frequencies from normal (mean = 5, variance = 2.5) distribution and observed frequencies:

Value	Observed Frequency	Expected Frequency
0	1	2
1	15	12
2	51	43
3	125	114
4	169	205
5	262	248
6	202	205
7	122	114
8	35	43
9	16	12
10	2	2
Total	1000	1000

For calculating χ^2 test statistic in this case, one has to coalesce first two and last two classes as they have expected frequencies less than 5. Finally, the goodness of fit test statistic comes out to be $\chi^2_{obs} = 13.18$ on 8 degrees of freedom. The p-value of χ^2 is 0.1058046, quite small. We conclude that the fit is moderate in that case. The figure for this case is given below:

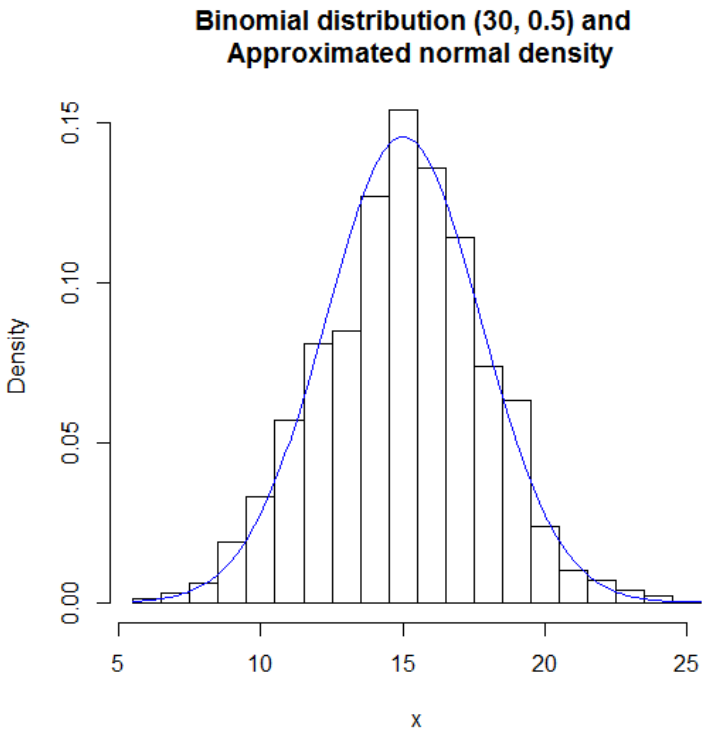


Binomial (n = 30, p = 1/2)

We have expected mean as (n*p) = 15 and expected variance as (n*p*q) = 7.5. We have 16 different classes from 8 to 23. The frequency distribution is as follows (The expected frequencies are from normal distribution with the aforementioned mean and variance):

Value	Observed Frequency	Expected Frequency	Value	Observed Frequency	Expected Frequency
≤8	10	9	16	136	136
9	19	13	17	114	111
10	33	28	18	74	80
11	57	50	19	63	50
12	81	80	20	24	28
13	85	111	21	10	13
14	127	136	≥22	13	9
15	154	146	**Total**	1000	1000

The goodness of fit chi-square test statistic comes out to be 18.08 on 14 degrees of freedom. The p-value associated with chi-square is 0.2031615. This is good enough to conclude that normal approximation does its job in fitting binomial distribution. However, as the central limit theorem suggests, the increment of n as binomial parameter does make the fitting better.



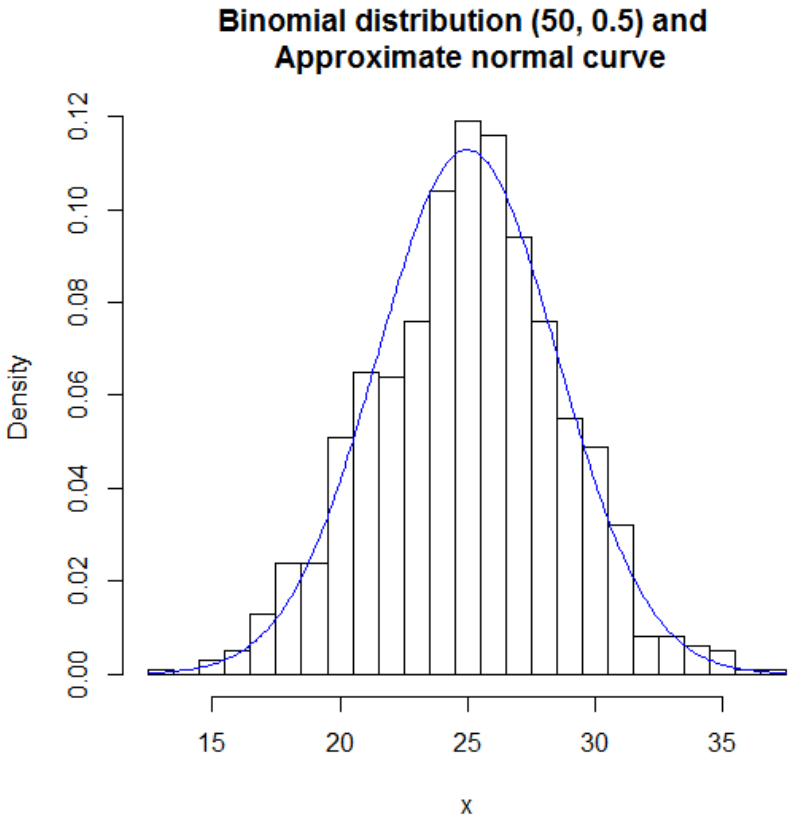
Binomial (n = 50, p = 1/2):

When the parameters of binomial distribution is 50 and probability of success is 1/2 , I have got values ranges from 13 to 36. So, I have considered two values grouped together to keep the number of classes moderate. The frequency distribution is as follows;

Value	Observed Frequency	Expected Frequency
12.5- 14.5	2	1
14.5- 16.5	8	7
16.5- 18.5	37	25
18.5- 20.5	75	69
20.5- 22.5	129	138
22.5- 24.5	180	204
24.5- 26.5	235	221
26.5- 28.5	170	175
28.5- 30.5	104	101
30.5- 32.5	40	43
32.5- 34.5	14	13
34.5- 36.5	6	3
Total	1000	1000

Based on the above table, one has to coalesce first two and last two classes to get expected frequencies more than 5. After that, we get observed value of chi-square test statistic is 11.52 on 9 degrees of freedom. The p-value turns out to be, 0.241779 which is clearly higher than the previous one indicating we have a better fit of binomial distribution by normal approximation.

The figure is given below:



Binomial (n = 100, p = 1/2):

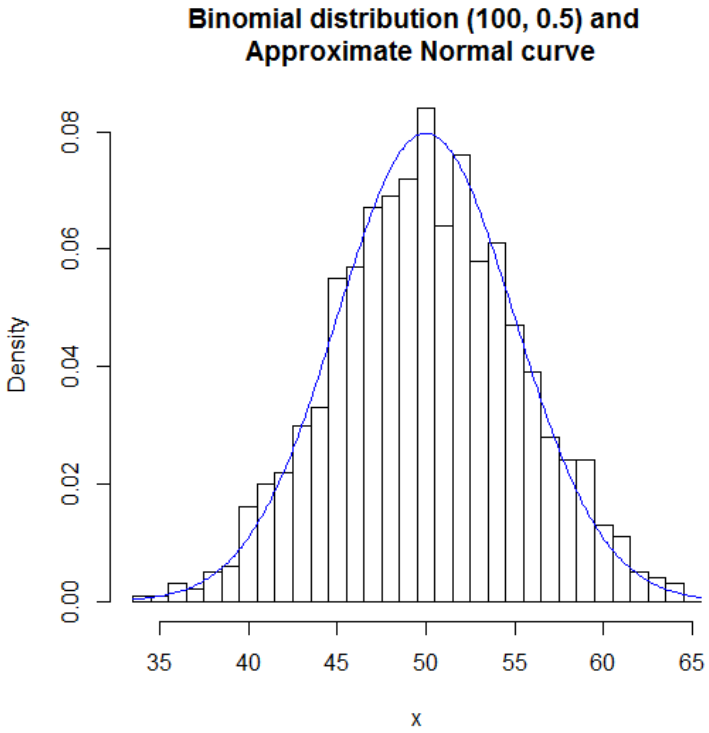
For this case, the range of the observed values are from 35 to 64. Hence, we have taken two values together as a group to make the number of classes reasonable. The expected mean is 50 and expected variance is 25.

The frequency distribution table is as follows:

Value	Observed Frequency	Expected Frequency	Value	Observed Frequency	Expected Frequency
34.5- 36.5	5	4	50.5- 52.5	140	152
36.5- 38.5	7	7	52.5- 54.5	119	124
38.5- 40.5	22	18	54.5- 56.5	86	87
40.5- 42.5	42	38	56.5- 58.5	52	52
42.5- 44.5	63	69	58.5- 60.5	37	27
44.5- 46.5	112	106	60.5- 62.5	16	12
46.5- 48.5	136	140	62.5- 64.5	7	6
48.5- 50.5	156	158	**Total**	1000	1000

The chi-square goodness of fit test statistic comes out to be 8.925083 on degrees of freedom 14, extremely good fitting of the binomial samples by normal approximation. As the observed chi-square lies on the left hand side of the expected one, so we calculate the left hand tail probability i.e. the p-value which is 0.8358196.

The figure for this case is as follows:



Binomial (n = 1000, p = 1/2):

In this case, the range of the sample taken is between 450 and 550. Here, I approximate it by using a normal distribution with mean 500 and variance 250. We get the following frequency table:

Value	Observed Frequency	Expected Frequency
450 - 460	6	6
460 - 470	39	23
470 – 480	86	74
480 – 490	164	161
490 – 500	245	236
500 – 510	222	236
510 – 520	141	161
520 – 530	76	74
530 – 540	18	23
540 - 550	3	6
Total	1000	1000

The chi-square test statistic turns out to be surprisingly 19.43149 on 9 degrees of freedom. Clearly, as it seems not a very good fit, the p-value of the test statistic assures it as by being a very low value 0.021764. It shows that, the fit is not so good contrary to the theory.

The problem with this is that classification of values in low number of groups is averaging out the effects of its neighbourhood values. So, the distribution jumps between different points of the approximation distribution.

This is curtained by the following two figures below:

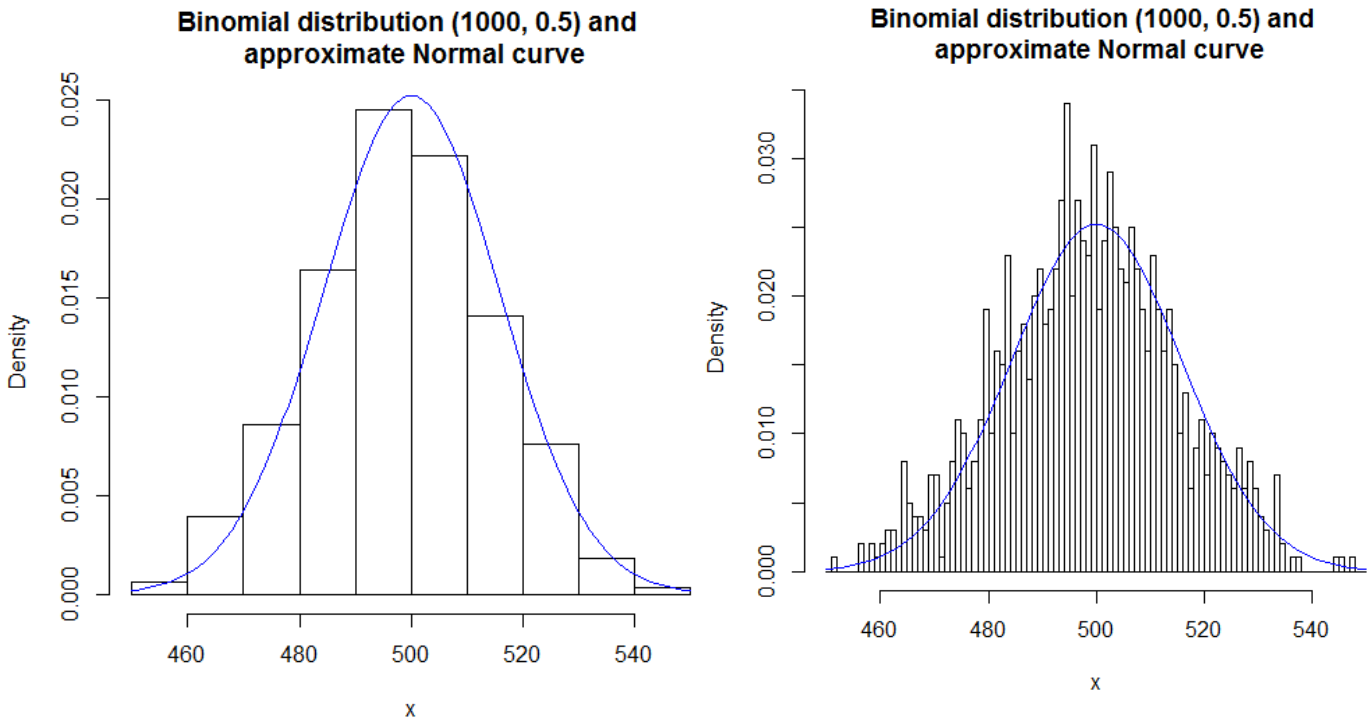


Figure: The effect of classification into groups in fitting approximated Normal distribution to Binomial distribution

Asymmetric Binomial distribution with p = 0.1

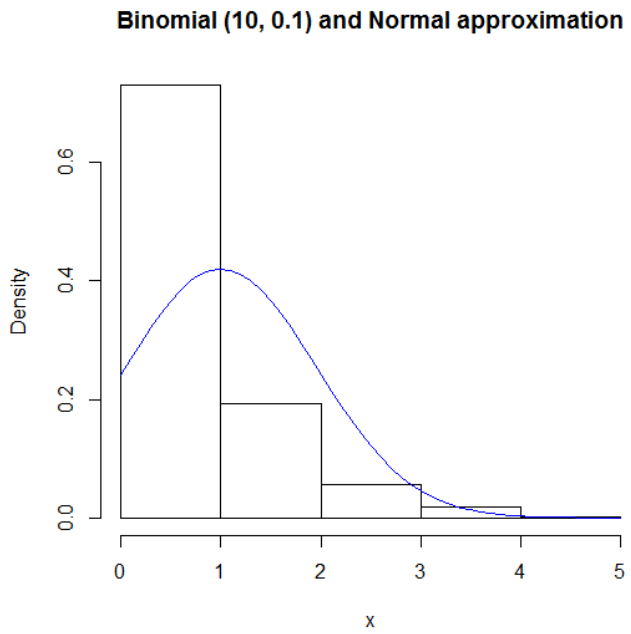
Binomial (n = 10, p = 0.1)

For the asymmetric case with probability of success being 0.1, and size of sample space i.e. the parameter n = 10, the theory suggests that normal approximation should not be good with such small n. Let us try fitting a normal distribution with mean (n*p) = 1, and variance (n*p*q) = 0.9. We get the following table:

Value	Observed Frequency	Expected Frequency
0	342	299
1	389	401
2	193	242
3	56	53
4	18	5
5	2	0
Total	1000	1000

The chi-square test statistic comes out to be 61.63435 on 4 degrees of freedom. It is certainly a bad fit. The p-value turns out to be 1.315059×10^{-12}, which is surely very negligible probability. Therefore, the situation goes well with the theory.

The figure for this case is given below:



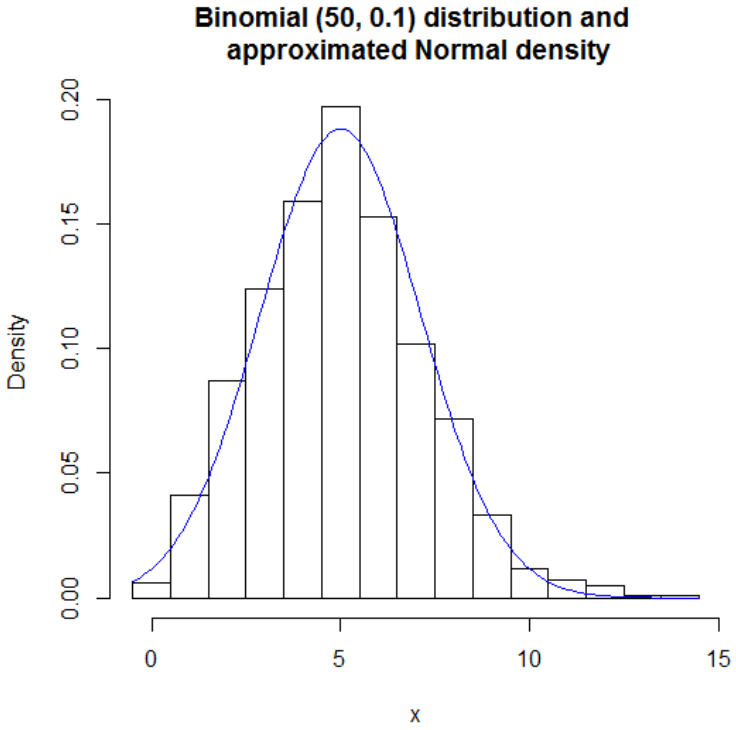
Binomial (n = 50, p = 0.1):

In this case, the expected mean is 5 and the expected variance is 4.5. In this case, the sample values ranges from 0 to 14, altogether making only 15 classes. The frequency distribution is as follows:

Value	Observed Frequency	Expected Frequency
0	6	17
1	41	33
2	87	70
3	124	120
4	159	167
5	197	186
6	153	167
7	102	120
8	72	70
9	33	33
10	12	12
11	7	4
12	5	1
13	1	0
14	1	0
Total	1000	1000

As we see the last four classes has expected frequency less than 5, so one need to coalesce them together to find out goodness of fit chi-square test statistic. For this case, it is 34.48351 on 11 degrees of freedom. Clearly, the normal approximation is still a bad fit for this case. However, we have p-value as 3.0137×10^{-4}, still very negligible probability to assert that the fit is no good at all.

The approximated curve along with the histogram is given below:



This figure explains how increasing sample size can make the histogram closer to normal density. We expect if we increase n more, it will give better result.

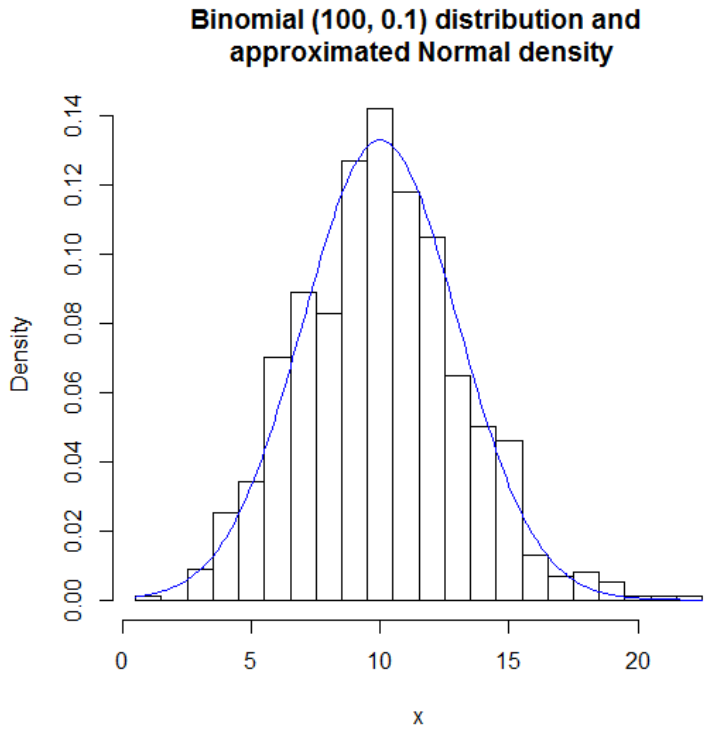
Binomial (n = 100, p = 0.1):

In this case, the expect mean is 10 and the expected variance is 9. We get to see that the sample generated has range over 2 to 22, containing only 21 classes. We will not group them as to see the actual effect. (Grouping may lead us to some problem as mentioned in the symmetric case.) The frequency distribution is given below:

Value	Observed Frequency	Expected Frequency	Value	Observed Frequency	Expected Frequency
2	1	7	13	65	81
3	9	9	14	50	55
4	25	18	15	46	33
5	34	33	16	13	18
6	70	55	17	12	9
7	89	81	18	5	4
8	83	106	19	2	2
9	127	125	20	1	1
10	142	132	21	1	0
11	118	125	22	1	0
12	105	106	**Total**	1000	1000

The chi-square test statistic comes out to be in this case about 31.3688 on 17 degrees of freedom. The associated p-value is 1.800453×10^{-2}, asserting that still the normal approximation is no good. It really shows p-value increases only a little and also the rate of convergence to normal density is slower in this asymmetric case.

The following figure contains histogram and the approximated normal density curve.



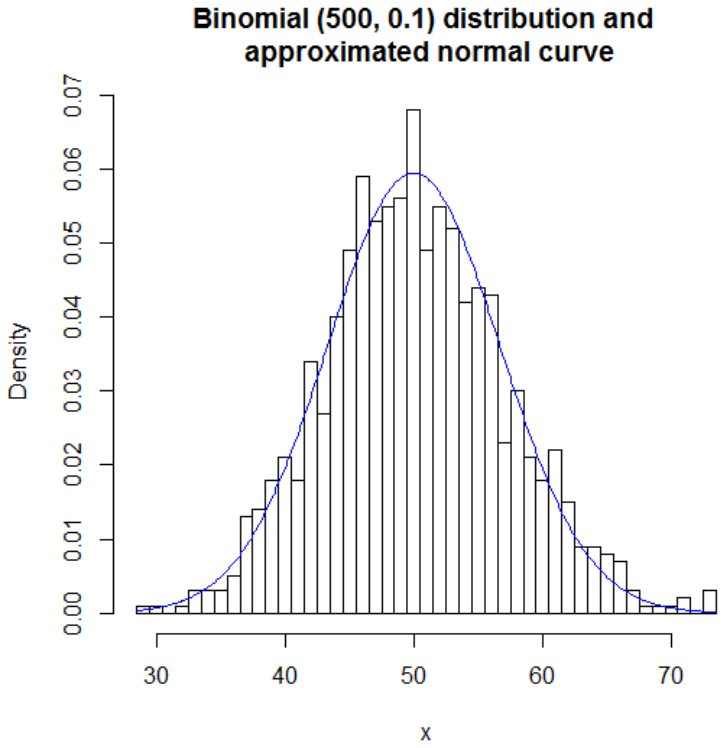
Binomial (n =500, p = 0.1):

In this case, the expected mean is 50 and the expected variance is 45. We have values ranges over 29 to 72, so we will try to consider them into groups so that we have less number of classes. We have considered class width to be 2 and the frequency distribution is given below:

Class	Observed Frequency	Expected Frequency	Class	Observed Frequency	Expected Frequency
28.5 – 30.5	2	1	50.5 – 52.5	104	116
30.5 – 32.5	3	3	52.5 – 54.5	94	104
32.5 – 34.5	6	6	54.5 – 56.5	87	85
34.5 – 36.5	8	12	56.5 – 58.5	53	64
36.5 – 38.5	27	21	58.5 – 60.5	39	44
38.5 – 40.5	39	35	60.5 – 62.5	37	28
40.5 – 42.5	52	53	62.5 – 64.5	18	16
42.5 – 44.5	67	74	64.5 – 66.5	15	8
44.5 – 46.5	108	95	66.5 – 68.5	4	4
46.5 – 48.5	109	111	68.5 – 70.5	1	2
48.5 – 50.5	124	118	70.5 – 72.5	3	0

The chi-square test statistic in this case becomes, 21.09421 on 17 degrees of freedom. This shows that the fit is considerably better than the previous ones. We see that, the p-value of observed chi-square turns out to be 0.224096, large enough to accept that the fit is good. This shows essentially that the convergence of binomial for asymmetric case is slower than symmetric ones, it starts showing significant sign of convergence to normal approximation when n is about 500, while it was only 30 for the symmetric case.

The figure for histogram along with approximated curve is given below:



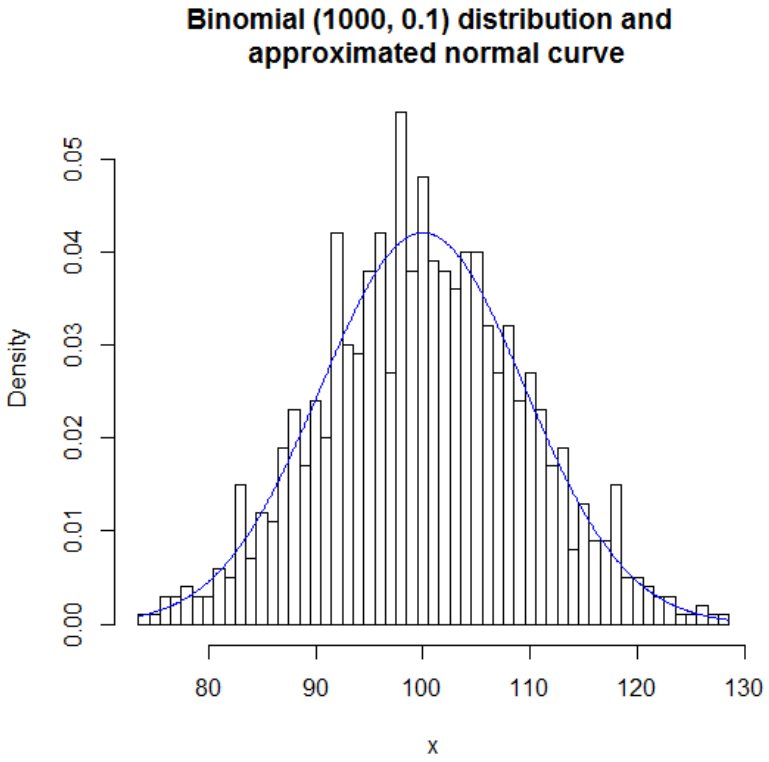
Binomial (n = 1000, p = 0.1):

When the parameters of the binomial distribution is 1000 and probability of success is 0.1, the sample found ranges over 74 to 127. The expected mean is 100 with the expected variance being 90. I have grouped them into classes of width 3 so that the number of classes is reasonable. I have kept the class width small so that the 'grouping effect' discussed for symmetric case does not appear. The distribution is as follows:

Values	Observed Frequency	Expected Frequency	Values	Observed Frequency	Expected Frequency
73.5 – 76.5	5	6	100.5 – 103.5	114	123
76.5 – 79.5	10	9	103.5 – 106.5	112	109
79.5 – 82.5	14	17	106.5 – 109.5	83	88
82.5 – 85.5	34	31	109.5 – 112.5	67	65
85.5 – 88.5	53	50	112.5 – 115.5	40	43
88.5 – 91.5	61	72	115.5 – 118.5	33	26
91.5 – 94.5	101	96	118.5 – 121.5	14	14
94.5 – 97.5	107	115	121.5 – 124.5	7	7
97.5 – 100.5	141	125	124.5 – 127.5	4	4

In this case, the goodness of fit chi-square test statistic surprisingly comes out to be 9.003659 on 16 degrees of freedom. This shows the fitting is extremely good by the normal approximation. The left tail p-value comes out to be 0.9132628, considerably indicating this is a situation of overfitting.

We may take a look at the histogram along with normal density curve in the following figure.



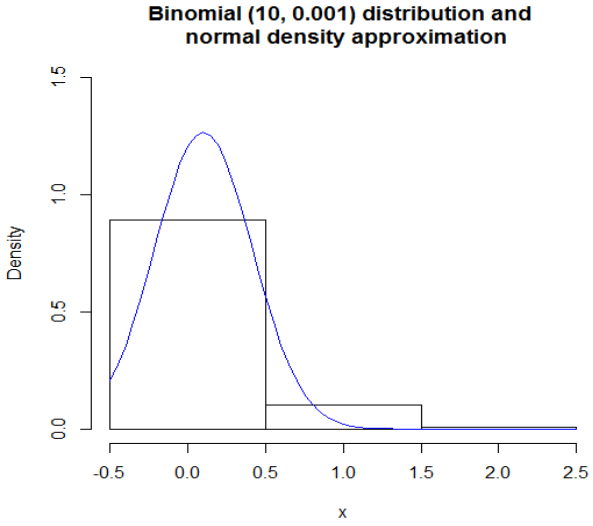
Asymmetric Binomial Distribution (p = 0.01):

Binomial (n = 10, p = 0.01)

Here, the expected mean is 0.1 and expected variance is 0.099, indicating that the binomial distribution should be highly positively skewed and normal approximation should not fit in this case. But, we will see what it gives. The frequency distribution is as follows:

Values	Observed Frequency	Expected Frequency
0	892	898
1	101	102
2	7	0
Total	1000	1000

Contrary to the expectation, the chi-square test statistic turns out to be 0.39303 on 1 degrees of freedom. The p-value turns out to be 0.53071.



The histogram along with the approximated normal curve is given beside. As it seems, the binomial distribution is no way being approximated by normal density.

To explain why this anomaly happened with the chi-square test statistic is that, chi-square is a good measure for goodness of fit when there is a reasonable number of classes. But, here after coalescing one has only 2 classes compared to total frequency 1000, somewhat weakening the effects / significance of goodness of fit measure.

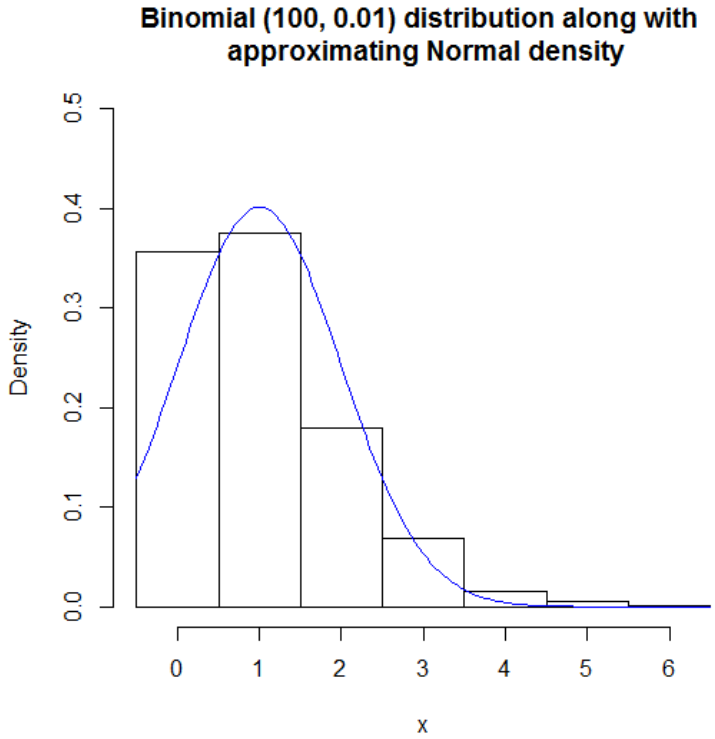
Binomial (n = 100, p = 0.01):

In this case, the expected mean is 1 and expected variance is 0.99. The sample ranges over only the values 0 to 6. We are also expecting a bad fit of normal distribution here, but let's see if the problem with significance with chi-square test statistic remains here also!

Values	Observed Frequency	Expected Frequency
0	356	308
1	375	384
2	179	242
3	68	60
4	15	6
5	6	0
6	1	0
Total	1000	1000

The Pearsonian chi-square test statistic (observed) is 67.829 in this case with degrees of freedom being 4. The corresponding p-value is 6.528111×10^{-14}, being very negligible probability. It implies that the normal approximation is not so good in this case, as suggested by the theory.

The histogram along with approximating curve is given below:



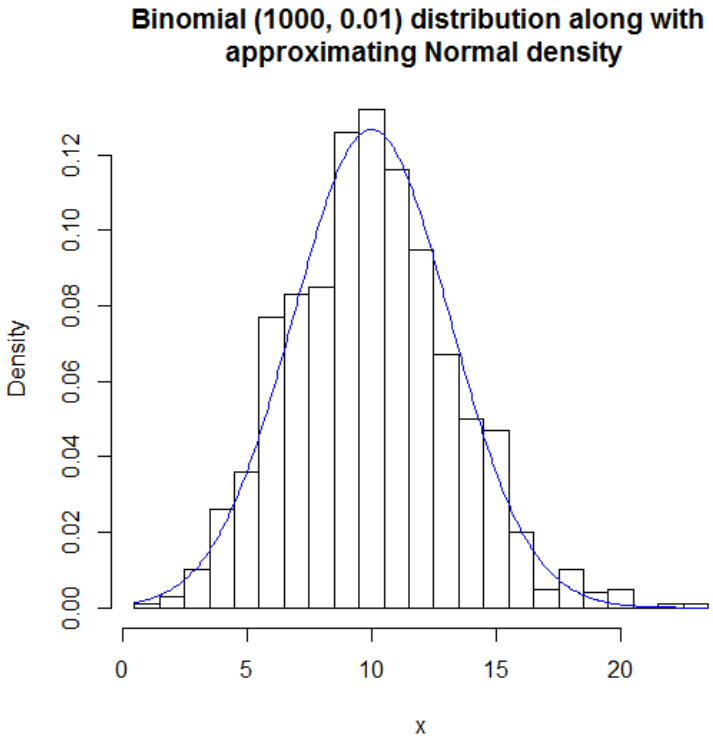
Binomial (n = 1000, p = 0.01)

The expected mean is 10 and expected variance is 9.9 in this situation. The sample here ranges from 1 to 22, and we divide the whole range into classes of width 2.

Class	Observed Frequency	Expected Frequency
0.5 – 2.5	4	5
2.5 – 4.5	36	32
4.5 – 6.5	113	93
6.5 – 8.5	168	184
8.5 – 10.5	258	246
10.5 – 12.5	211	223
12.5 – 14.5	117	137
14.5 – 16.5	67	57
16.5 – 18.5	15	16
18.5 – 20.5	9	7
20.5 – 22.5	2	0
Total	1000	1000

The chi-square test statistic is 14.64759 in this case, on 9 degrees of freedom. The corresponding p-value here is 0.1011364, implying that the fit is good in this case. As seen from the asymmetric case with p = 0.1, that needed 500 sample size as parameters to produce a significant result. We see that, in this case, it needs more than 1000. We may conclude that smaller or very large p needs more and more steps to converge to normal density.

The histogram is as follows:



Binomial (n = 5000, p = 0.01)

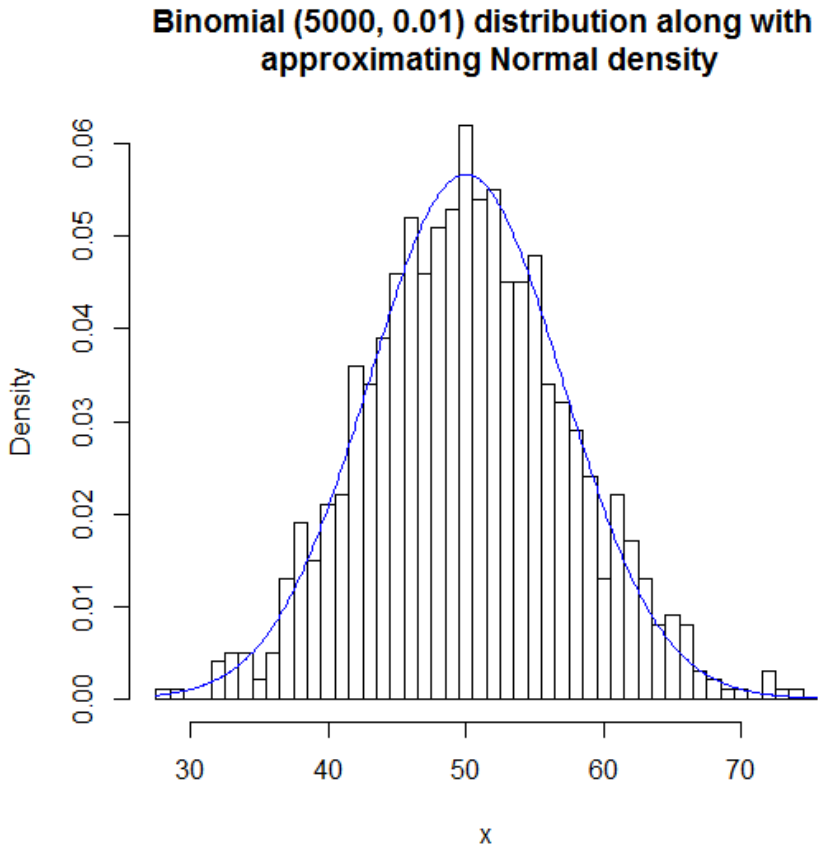
In this case, the expected mean is 50 and expected variance is 49.5. The sample ranges from 28 to 75, hence they are grouped into classes of width 3.

The following is the frequency distribution:

Class	Observed Frequency	Expected Frequency	Class	Observed Frequency	Expected Frequency
27.5 – 30.5	2	4	51.5 – 54.5	145	154
30.5 – 33.5	9	7	54.5 – 57.5	114	118
33.5 – 36.5	12	18	57.5 – 60.5	66	75
36.5 – 39.5	47	40	60.5 – 63.5	52	40
39.5 – 42.5	79	75	63.5 – 66.5	25	18
42.5 – 45.5	119	118	66.5 – 69.5	6	7
45.5 – 48.5	149	154	69.5 – 72.5	4	3
48.5 – 51.5	169	168	72.5 – 75.5	2	1

The goodness of fit chi-square test statistic turns out to be 11.7698 on 12 degrees of freedom. It seems it is a pretty good fit for the binomial distribution with normal approximation. The p-value is 0.46434, significantly implying that the fit is good.

We may assume that, for p = 0.01, the convergence to normal approximation starts from around when n = 5000, larger than the previous convergence criterions.



As suggested by the analysis above, it follows that for binomial distribution, the central limit theorem is valid. (The theory goes consistent with the data). Moreover, there may be some problems due to the groupings into classes and hence it gives rise to some error in the approximation.

We have also seen, as p deviates from the symmetric case to either 0 or to 1, it takes more and more samples of Bernoulli trials (i.e. larger parameter 'n' of binomial distribution) to reach sufficiently close to the approximation.

Approximating Binomial by Poisson

We know that as $n \to \infty$ and $p \to 0$ in binomial distribution, but with $np = \lambda$, it follows Poisson distribution with parameter λ. Here, in this section, we will try to analyse that, with fixed λ and will take 1000 samples from binomial distribution.

Poisson (λ = 5)

Now as we have $\lambda = 5$, we expect that for the poison distribution, first few expected frequencies should be as follows:

Value	Poisson Frequency	Value	Poisson Frequency
0	7	7	104
1	34	8	65
2	84	9	36
3	140	10	18
4	175	11	8
5	175	12	3
6	146	13	1

And the rest of the frequencies are all 0's. Therefore, it is obvious that taking binomial sample size parameter less than 13 would not help a lot. We start with the following;

Binomial (n = 30, p = 1/6)

In this situation, we have the following frequency distribution as follows:

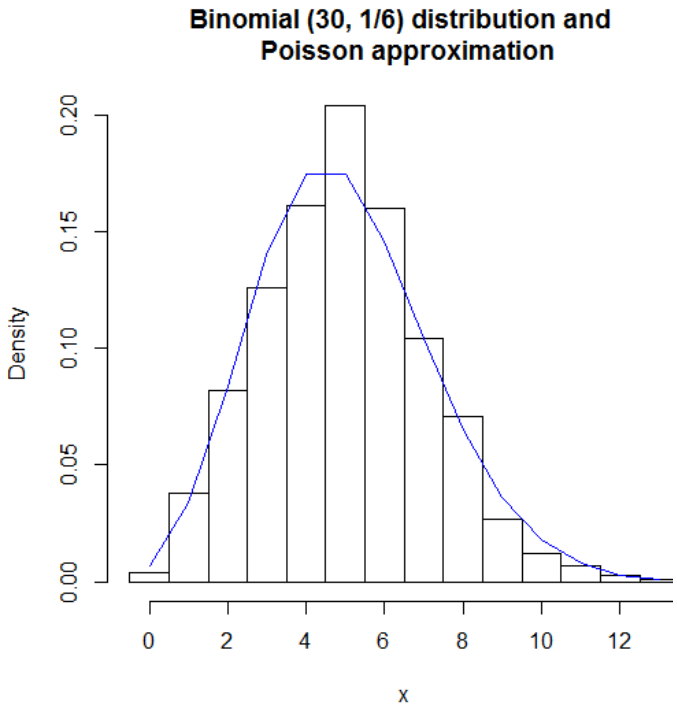
Value	Binomial Frequency	Value	Binomial Frequency
0	4	7	104
1	38	8	71
2	82	9	27
3	126	10	12
4	161	11	7
5	204	12	3
6	160	13	1

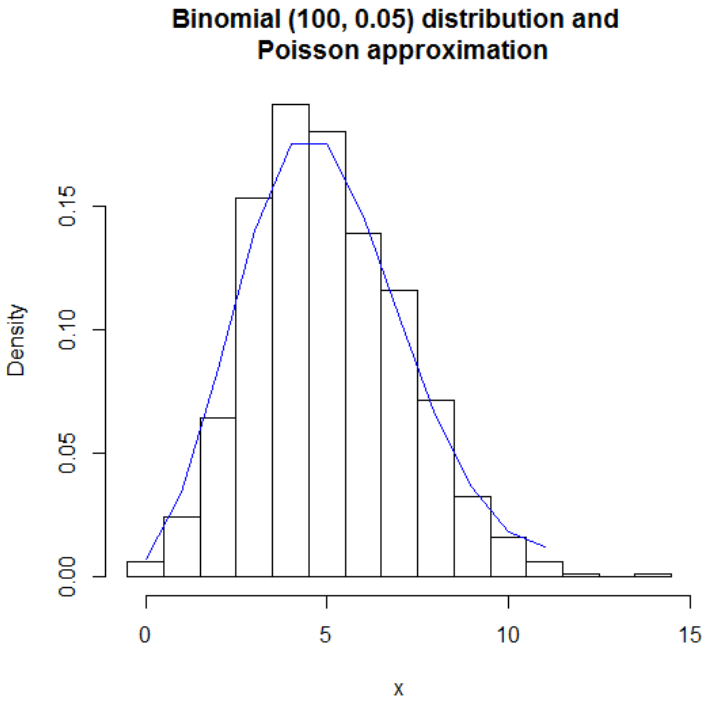
The above table shows, only n = 30 here, gives a pretty good approximation for this Poisson parameter 5. The chi-square test statistic is 15.40095 here, on degrees of freedom 11, with p-value 0.1648671. Clearly, this shows the fit is good enough.

Binomial (n = 100, p =0.05)

The frequency distribution for this case is as follows:

Value	Binomial Frequency	Value	Binomial Frequency
0	6	7	116
1	24	8	71
2	64	9	32
3	153	10	16
4	191	11	6
5	180	12	1
6	139	13	1

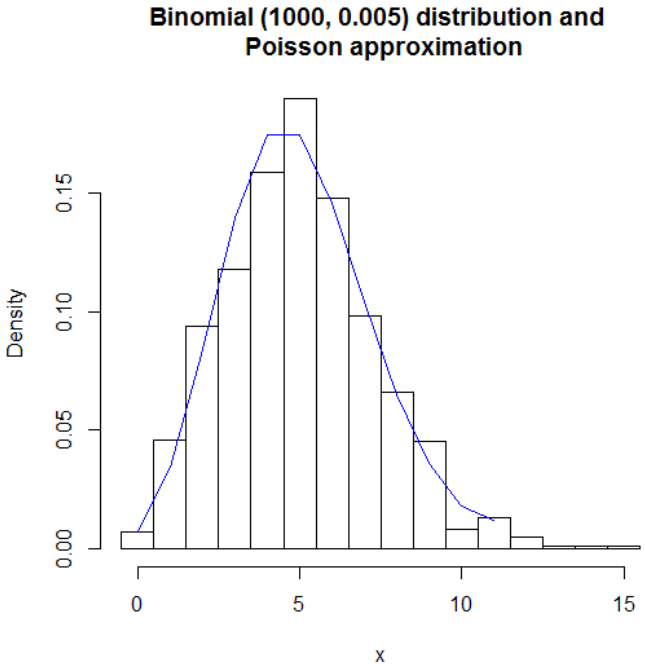




For this case, we have chi-square test statistic as 14.9327 on 11 degrees of freedom, essentially implying we have come only a little way forward to fit the binomial dataset using poison approximation.

Binomial (n = 1000, p = 0.005)

For this case, the chi-square test statistic is coming to be 16.99264 on 11 degrees of freedom again and corresponding p-value is 0.1080935. This shows the Poisson approximation fits the binomial distribution well, but it is not going to be any better if we increase binomial parameter ‘n’. In this case, the histogram along with poisson curve looks like the following:



Conclusion

In the part of approximating binomial distribution by normal density and the validation of central limit theorem is assured through different examples. From these examples, we get that as binomial parameter sample size ‘n’ gets larger and larger, as suggested by the theory, the approximation gets better.

Also, one needs more steps to get closer to the approximation when the underlying binomial distribution is not symmetric. The more it is skewed (negatively or positively), the more time it takes to converge to approximated normal density and hence the lower convergence rate.

We have also, come to know of the fact that very less number of groups compared to the size (or total frequency) gives worse fit to the dataset.

From the part with Poisson distribution approximating binomial distribution, we see that due to the randomness of the samples taken from binomial distribution, increasing binomial parameter sample size ‘n’ does not make any much of a difference to the goodness of fit measure. Maybe for different lambda, we may need different sample size ‘n’ to get better convergence.