Vol.9.Issue.2.2021 (April-June)

http://www.bomsr.com
Email:editorbomsr@gmail.com

RESEARCH ARTICLE

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

THE NORMAL APPROXIMATION TO THE BINOMIAL DISTRI BUTION – AN EXPLORATORY STUDY

RAMNATH TAKIAR
Scientist G – (Retired)
National Centre for Disease Informatics and Research (NCDIR),
Indian Council of Medical Research (1978-2013)
Bangalore – 562110, Karnataka, India.
Email: ramnathtakiar@gmail.com, ramnath_takiar@yahoo.co.in
DOI:10.33329/bomsr.9.2.47

ABSTRACT

The present paper explore empirically the relationship between the Binomial and the Normal distribution. The paper also attempts to find out the values of "n" and "p" above which the Binomial distribution can safely be approximated to the Normal distribution? For different combinations of n (20 and 50), p (0.1, 0.2, 0.5) and the number of successes (X = 0,1,2,3…n), the probabilities are calculated using both the Binomial and the Normal distribution approach and comparisons are made. For all the combinations of n (20, 50) and P(0.1, 0.2, 0.5), the expected frequency distributions of the Binomial and the Normal are found to be comparable. The Chi-square test revealed no significant differences between the expected frequencies of the Binomial and the Normal distribution. According to the literature available, for a good approximation of Binomial distributrion by the Normal distribution, n should be atleast 50. However, the results of the present study suggest that for the Normal approximation of the Binomial distribution n can be as low as 20 and P ≥ 0.1, thus giving rise to a new condition that is np ≥ 2. There is no need to put any condition on nq. For P=0.5, the Normal and the Binomial distributions are found be comparable even when n=7.

INTRODUCTION

One of the popular discrete distribution is the Binomial distribution. It often deals with the occurring or non-occuring of an event popularly called as success or a failure of an event. The Binomial distribution can be typically characterized by the two parameters namely "n" and "p" where "n" represents the number of trials and "p" represents the fixed probability of occurring of the desired event. Using the Binomial distribution approach, one can easily determine the probability of occurring a specified number of successes in a specified number of trials. For example, if a coin is tossed 8 times, we can calculate the probability of getting the desired number of successes say 3 or 4 heads. The formula to calculate the number of successes can be given as follows:

$$P(X = x) = {}^{n}C_{x}\, p^{x}\, q^{(1-x)} \quad \text{for } x = 1,2,3\ldots\ldots n.$$

Where "x" denotes the desired number of successes, "p" is the fixed probability of getting a success and "n" is the number of trials made. As long as we are looking for the exact number of successes, the calculation is relatively simple, however when we wish to go for a group of successes, say more than 4 or less than 6 heads in a trial of 8, the number of probability calculations required, accordingly, would be 4 or 6. Imagine, if we are looking for 100 trials and wish to estimate the probability of getting atleast 40 heads, then, it will require the calculations of atleast 40 probabilities, using the above given formula. Therefore, for the large n and for the large number of outcomes as success, the very calculation of probability of success would become very difficult and challenging. In such a situation, the normal approximation to binomial distribution is advocated.

The normal distribution is the most important, widely used and a very popular continuous distribution in statistics. It is characterized by two parameters, the mean (μ) and the standard deviation (σ). Further, it is a symmetric distribution with mean, median and mode being equal. In medical and social fields, many variables are known to follow the normal distribution like height, birth weight, blood pressure, IQ score, job satisfaction and SAT score.

The general rule of thumb to use normal approximation to binomial distribution is that $np \geq 5$ and $nq \geq 5$ (Chaudhary 2021, Statistic How to 2021, VrcAcademy 2020, Helm 2008). In order that above two conditions are met, n should be atleast 50 (STAT 414 2021). In literature, we can also find the conditions that $np \geq 10$ and $nq \geq 10$ for the Normal approximation to the Binomial distribution implying that n should be greater than 100 (Taylor 2018).

In view of the different coditions being stated by the different authors for the Normal approximation to the Binomial distribution, it is thought worth while to explore the empirical relationship between the Normal and the Binomial probabilities for selected combinations of n and p. The paper aims to find out:

- How does the Binomial and Normal Probabilities compare with the varying n and p values?
- What is the extent of linear relationship between both the probabilities with the varying n and p values ?

METHODOLOGY

Let X be a Bernoulli variable. The probability of X = x successes can be obtained by the probability distribution of Binomial distribution, given by

$$P (X = x) = {}^{n}C_{x}\, p^{x}\, q^{(1-x)} \text{ for } x = 1,2,3\ldots\ldots n.$$

where "p" is the fixed probability of getting the success and "n" is the number of trials made.

To use the normal approximation for the binomial distribution, we have to first, find Z score corresponding to given "x" using the formula: $Z = \frac{(x - np)}{\sqrt{npq}}$

Then, find the probability for calculated Z, using the table of Normal probabilities or using the function in Excel.

Continuity Correction for Normal Approximation

Binomial distribution is a discrete distribution whereas the normal distribution is a continuous distribution. Hence, while calculating various Normal probabilities as the normal approximation to the Binomial distribution, there is a need to apply a correction known as the continuity correction. The continuity corrections as applicable are shown below (Chaudhary 2021, VrcAcademy 2020)

- $P\,(X = A) = P\,(A - 0.5 < X < A + 0.5)$
- $P\,(X < A) = P\,(X < A - 0.5)$
- $P\,(X \leq A) = P\,(X \leq A + 0.5)$
- $P\,(A < X \leq B) = P\,(A + 0.5 < X < B + 0.5)$
- $P\,A \leq X < B) = P\,(A - 0.5 < X < B - 0.5)$
- $P\,(A \leq X \leq B) = P\,(A - 0.5 < X < B + 0.5)$

For the study purposes, first the "P" value was selected such as P = 0.1, then for n = 20 and 50 and for varying values of x = 0,1,2, 3,….n, the Binomial and the Normal probabilities are calculated until one of the probabilities becomes 0. The probabilities so calculated are tabulated and shown in a table. Similar exercise was again repeated for P = 0.2 and 0.5.

In calculation of the Binomial and the Normal probabilities, the excel function keys are used as shown below (Microsoft Corporation 2019):

BINOM.DIST(number_s, trials, probability_s, cumulative)

Example: When n = 20 and P = 0.1 and X = 0, the following inputs can be used in the above formula, as shown below:

BINOM.DIST(0, 20, 0.1, FALSE) will yield 0.122 as the required probability.

NORM.DIST(x, mean, standard_dev, cumulative)

Example: When n = 20 and P = 0.1 and X = 0, the following inputs can be used in the above formula, as shown below:

On applying the continuity correction, we get X = - 0.5 and X = 0.5

NORM.DIST(0.5, 2, 1.342, FALSE) will yield 0.132 as the required probability.

NORM.DIST(- 0.5, 2, 1.342, FALSE) will yield 0.031 as the required probability.

So, P (X = 0) = P (X = 0.5) – P (X = - 0.5) = 0.132 – 0.031 = 0.101

All the probabilities obtained for the Binomial and the Normal distribution, for varying x (1,2,3….n) and P (0.1, 0.2 and 0.5) are pooled for each n, separately and the Correlation (r) and the Slope (b) are obtained, and the Regression equation is formulated. The Scatter graph along with the trend line is also plotted for n = 20 and n=50, to show the linear relationship between the Binomial

and the Normal probabilities. The r^2 calculated shows the percentage variation explained in Y by X where X is the Binomial Probability and Y is the Normal probability.

From the Binomial and Normal probabilities calculated for given "n" and "P", for various "x" values, are multiplied by "n" to get the expected frequencies. As a Golden rule, if the expected frequency distributions of the Binomial and the Normal are found to be comparable then it is assumed that the Normal approximation for the Binomial distribution is holding good for the given "n" and "P". For example, when n = 20 and P = 0.1, the Binomial probability for X = 0 is 0.122 and the Normal probability is 0.101. So, when you multiply both the probabilities with 20, we get the expected frequencies as 2.44 and 2.02, respectively which on rounding become equal to 2. Proceeding in a similar way, for different value of X (0,12,3…n), we get the expected frequencies for the Binomial and the Normal distributions.

The expected frequency distribution obtained for both the Binomial and the Normal distribution is converted to feasible Chi-square contingency table and the significance of the calculated Chi-square was attempted. If the calculated Chi-square value for the given table is found to be non-significant then it is assumed that those two distributions are comparable and the Normal approximation to Binomial distribution can be carried out safely and with the confidence, Otherwise, the Normal approximation is considered as not good enough for the Binomial distribution.

For the sample size of 20, a probability of 0.05 can be considered as the least significant probability as on multiplying 20*0.05, we get 1. The least significant probability can be assumed to be that probability which on multiplying with the selected sample size yields the difference of 1 in the expected frequencies of both the distributions. So, unless both the probabilities are differing by more than 0.05, they do not show different rounded off frequencies. It is to be noted that by above definition, the least significant probability for the sample size of 50 would be 0.02 and for the sample size of 100, it would be 0.01. While looking for visual differences between the Binomial and Normal probabilities, the least significant probability should be kept in mind. Ultimately, a non-significant difference in the expected frequency distributions would indicate whether the probabilities under consideration for the Normal and the Binomial distribution are actually comparable or not?

RESULTS

The comparison of the Binomial and the Normal Probabilities when n = 20 and P = 0.1 is shown in Table 1.

It is clear from the Table 1, that the Binomial and the Normal Probabilities are quite close to each other as further evident by the expected frequencies. From the above frequency distributions, the 3 X 2 Contingency table is made and then from the calculated x^2 value, its significance is established (Table 2). The x^2 test suggested no significant differences between the expected frequencies.

The comparison of the Binomial and the Normal Probabilities when n = 20 and P = 0.2 is shown in Table 3. By the x^2 test, both the expected frequency distributions are found to be comparable.

The comparison of the Binomial and the Normal Probabilities when n=20 and P=0.5 is shown in Table 4. By the x^2 test, both the expected frequency distributions are found to be comparable.

The comparison of the Binomial and the Normal Probabilities when n = 50 and P = 0.1 is shown in Table 5. By the x^2 test, both the expected frequency distributions are found to be comparable.

Table 1: The Comparison of the Binomial and the Normal Probabilities (P = 0.1 and n = 20)

	Probability		Rounded off Frequency	
	Binomial	Normal	Binomial	Normal
0	0.122	0.101	2	2
1	0.27	0.223	5	5
2	0.285	0.291	6	6
3	0.19	0.223	4	4
4	0.09	0.101	2	2
5	0.032	0.027	1	1
≥ 6	0.011	0.004	0	0

Table 2: The Contingency table for the Expected Frequencies
(n = 20 and P = 0.1)

X value	Binomial	Normal	Total
0 - 1	7	7	14
2	6	6	12
≥ 3	7	7	14
Total	20	20	40

$x^2(2) = 0$; *P value = 1.0*

Table 3: The Comparison of the Binomial and the Normal Probabilities (P = 0.2 and n = 20)

X	Probability		Rounded off Frequency	
	Binomial	Normal	Binomial	Normal
0 - 1	0.07	0.075	1	2
2	0.137	0.12	3	2
3	0.205	0.189	4	4
4	0.218	0.22	4	4
5	0.175	0.189	4	4
6	0.109	0.12	2	2
7	0.055	0.056	1	1
≥ 8	0.031	0.031	1	1

$x^2(1) = 0$; *P value =1.0*

Table 4: The Comparison of the Binomial and the Normal Probabilities (P = *0.5* and n = 20)

X	Probability		Rounded off Frequency	
	Binomial	Normal	Binomial	Normal
≤ 6	0.058	0.058	1	1
7	0.074	0.073	2	2
8	0.12	0.119	2	2
9	0.16	0.16	3	3
10	0.176	0.177	4	4
11	0.16	0.16	3	3
12	0.12	0.119	2	2
13	0.074	0.073	2	2
≥ 14	0.058	0.061	1	1

$x^2(2) = 0$; *P value = 1.0*

Table 5: The Comparison of the Binomial and the Normal Probabilities (P = 0.1 and n = 50)

X	Probability		Rounded off Frequency	
	Binomial	Normal	Binomial	Normal
≤ 2	0.112	0.115	6	6
3	0.139	0.12	7	6
4	0.181	0.167	9	8
5	0.185	0.186	9	9
6	0.154	0.167	8	8
7	0.108	0.12	5	6
8	0.064	0.07	3	4
9	0.033	0.033	2	2
≥ 10	0.024	0.017	1	1

$x^2(6) = 0.304$; *P value = 0.999*

The comparison of Binomial and Normal Probabilities when n=50 and *P = 0.2* is shown in Table 6. By the x^2 test, both the expected frequency distributions are found to be comparable.

The comparison of Binomial and Normal Probabilities when n = 50 and P = 0.5 is shown in Table 7. By the x^2 test, both the expected frequency distributions are found to be comparable.

The linear relationship between the Binomial and the Normal probabilities, pooled for n=20 and P is 0.1, 0.2 and 0.5, is shown in Fig. 1. The r^2 in this case was observed to 0.962 which is quite high by any standards. The slope is 0.969 which is close to 1. In Fig. 2, for n = 50 and P = 0.1, 0.2, 05 (pooled), the r^2 is observed to be 0.977. The slope of 0.971 suggests that when the Binomial probability changes by 0.1-unit, the Normal probability changes by 0.097 unit.

Table 6: The Comparison of the Binomial and the Normal Probabilities (P = 0.2 and n = 50)

X	Probability		Rounded off Frequency	
	Binomial	Normal	Binomial	Normal
≤ 4	0.018	0.026	1	1
5	0.03	0.03	1	1
6	0.055	0.052	3	3
7	0.087	0.08	4	4
8	0.117	0.11	6	6
9	0.136	0.132	7	7
10	0.14	0.14	7	7
11	0.127	0.132	6	7
12	0.103	0.11	5	5
13	0.075	0.08	4	4
14	0.05	0.052	3	2
15	0.03	0.03	2	2
≥ 16	0.028	0.025	1	1

$x^2(6) = 0.13$; *P value = 0.999*

Table 7: The Comparison of the Binomial and the Normal Probabilities (P = 0.5 and n = 50)

X	Probability		Rounded off Frequency	
	Binomial	Normal	Binomial	Normal
≤ 18	0.032	0.058	2	2
19	0.027	0.027	1	1
20	0.042	0.042	2	2
21	0.06	0.06	3	3
22	0.079	0.079	4	4
23	0.096	0.096	5	5
24	0.108	0.108	5	5
25	0.112	0.112	6	6
26	0.108	0.108	5	5
27	0.096	0.096	5	5
28	0.079	0.079	4	4
29	0.06	0.06	3	3
30	0.042	0.042	2	2
31	0.027	0.027	1	1
32	0.016	0.016	1	1
≥ 33	0.016	0.016	1	1

$x^2(8) = 0$; *P value = 1.0*

Fig. 1: The Linear Relationship between the Binomial and the Normal Probabilities for varying P and n=20

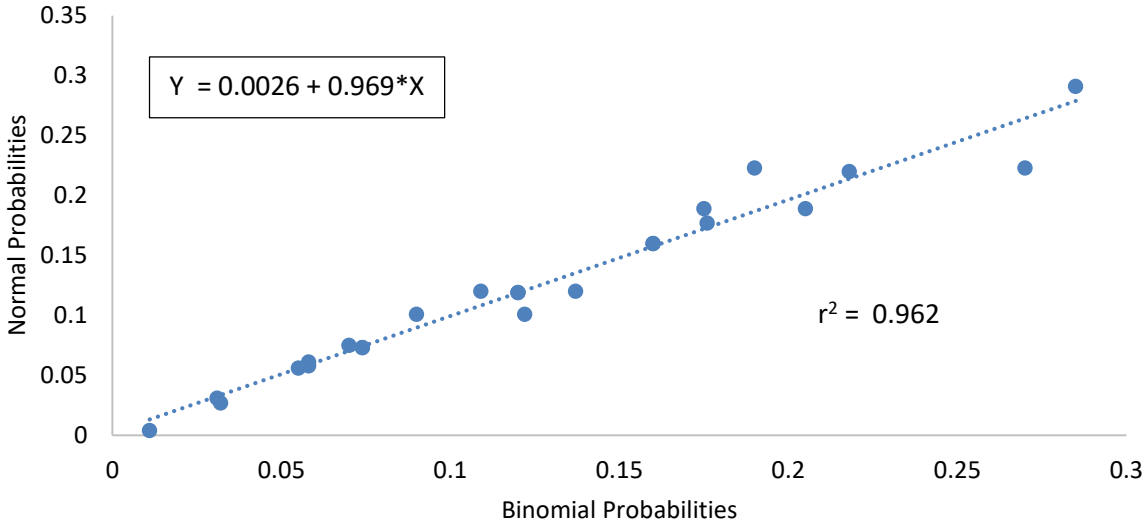

Fig. 2: The Linear Relationship between the Binomial and the Normal Probabilities for varying P and n=50

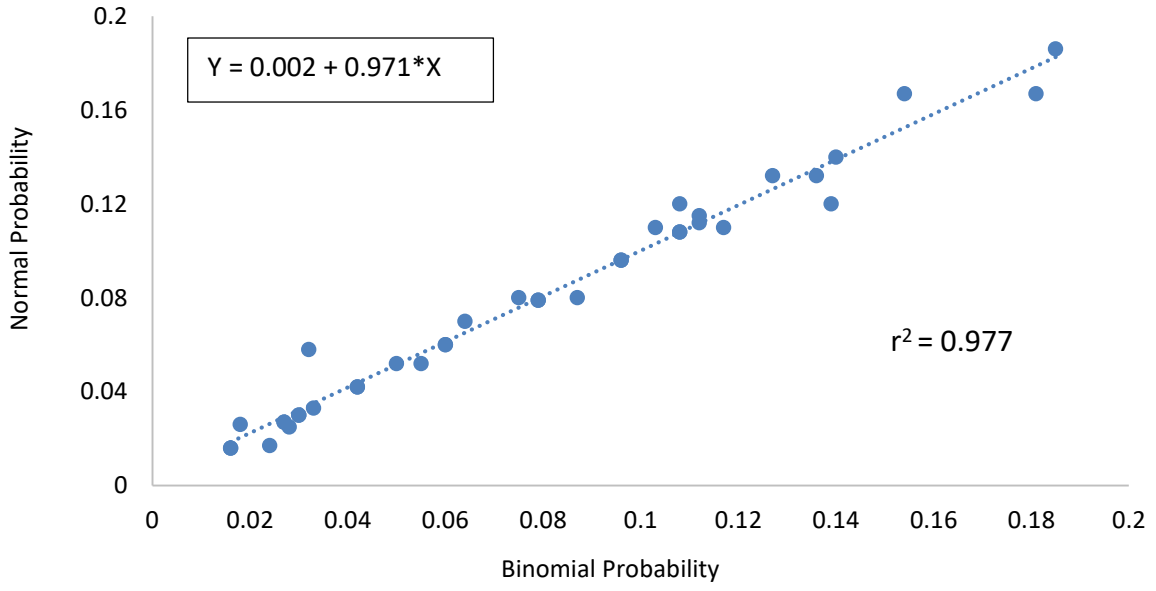

Table 8: The Comparison of the Binomial and the Normal Probabilities when n = 7 and P = 0.5 for Varying X values

Distribution	≤ 1	2	3	4	5	≥ 6
Binomial	0.063	0.164	0.273	0.273	0.164	0.063
Normal	0.064	0.16	0.275	0.275	0.16	0.064

DISCUSSION

The present study has clearly demonstrated that even when n = 20 and P = 0.1, the Normal approximation to the Binomial distribution can be carried out. It was observed that the probabilities obtained for the Binomial and Normal distribution may show a little variation but when they are converted to the expected frequencies, they show a remarkable resemblance to each other. Further by the Chi-square test, they confirm to be comparable and non-significant.

As mentioned in the beginning, for the Normal approximation to the Binomial distribution, different authors stated different conditions. The present study indicates that for n ≥ 20 and for P ≥ 0.1, the normal approximation can safely be carried out. Especially when P = 0.5, for n as low as 7, both the probabilities can be demonstrated to be quite similar, as shown in the Table 8.

The Correlation graph between the Binomial and the Normal probabilities suggests a very strong correlation between the two probabilities and a slope close to 0.97 indicating that there could be on an average, at most 3% error if we go for the Normal approximation to the Binomial probabilities.

Conclusion

The results of the present study clearly suggest that the Normal approximation to the Binomial distribution can safely be carried out when n ≥ 20 and for any P ≥ 0.1. When *P* = 0.5, the Normal approximation to Binomial distribution can hold good even when n is as low as 7.

REFERENCES

[1]. Chaudhary R (2021): Normal Approximation To Binomial with Examples https://www.vrcbuzz.com/normal-approximation-binomial-calculator-with-examples/ Accessed on 23rd May 2021.

[2]. Helm (2008): The Normal Approximation to the Binomial Distribution, Section 39.2

https://learn.lboro.ac.uk/archive/olmp/olmp_resources/pages/workbooks_1_50_jan2008/Workbook39/39_2_norm_apprx_bnml_dist.pdf, Accessed on 23rd May 2021.

[3]. Microsoft Corporation, 2019. Microsoft Excel, Available at: https://office.microsoft.com/excel.

[4]. STAT 414 – PennState (2021): Normal Approximation to Binomial, Eberly College of Science.

https://online.stat.psu.edu/stat414/lesson/28/28.1, Accessed on 23rd May 2021.

[5]. Statistics How to (2021):Normal approximation Binomial

https://www.statisticshowto.com/probability-and-statistics/binomial-theorem/normal-approximation-to-the-binomial/, Accessed on 23rd May 2021.

[6]. Taylor C (2018): The Normal Approximation to the Binomial Distribution, ThoughtCo, Aug. 27, 2020, thoughtco.com/normal-approximation-to-the-binomial-distribution-3126589.

Accessed on 23rd May 2021.

[7]. VrcAcademy (2021), Normal Approximation to Binomial distribution calculator with Examples

https://vrcacademy.com/calculator/normal-approximation-binomial-distribution-calculator/

Accessed on 23rd May 2021.