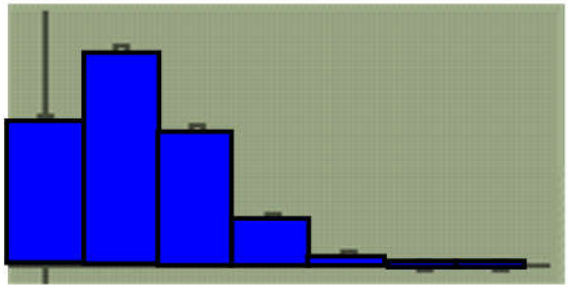
Derivation of the Success/Fail criteria for Normal approximation to a Binomial distribution

If we have a Binomial setting with a small number of trials and a probability of success not near 0.5 (for example, n=6, p=0.2) we could use a calculator binompdf function to determine the probability of each possible outcome. If you use a Ti-83/84 and load values 0-6 into L1, and specify L2=binompdf(6, 0.2, L1) to calculate corresponding binomial model probability values, you can then display a scatterplot of (L1,L2) and see the probability distribution (histogram bars added here to the scatterplot):

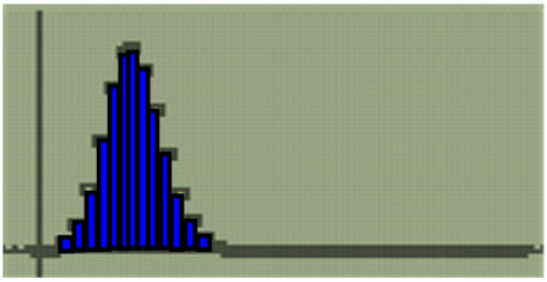
X	0	1	2	3	4	5	6
P	*binompdf* (6, 0.2, 0)	*binompdf* (6, 0.2, 1)	*binompdf* (6, 0.2, 2)	*binompdf* (6, 0.2, 3)	*binompdf* (6, 0.2, 4)	*binompdf* (6, 0.2, 5)	*binompdf* (6, 0.2, 6)



This distribution is highly skewed right because with low probability of success, it is more likely to have 0 or 1 success out of 6 than 5 or 6 out of 6 successes.

But the situation changes if we increase the number of trials. Keeping p=0.2, but increasing n to 40 gives the following Binomial probability distribution:

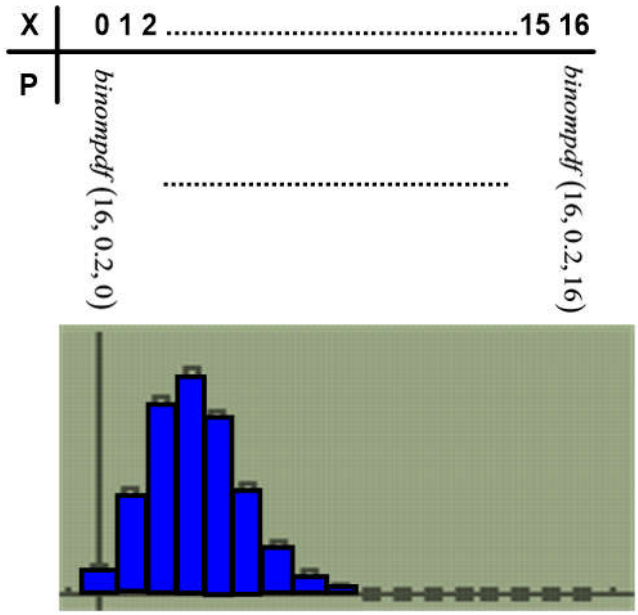
X	0 1 2	.....	.....39 40
P	*binompdf* (40, 0.2, 0)	.....	*binompdf* (40, 0.2, 40)



Even though the only thing that changes is the number of trials, the distribution is very symmetrical (even with p=0.2), looks much like a Normal distribution, and we could use a Normal distribution to approximate this Binomial distribution.

At what number of trials would we have a Binomial distribution that is close enough to Normal that we could approximate it with a Normal distribution? The value is somewhere between 6 and 40 and it may vary depending upon the p value.

Let's explore this further by looking at an intermediate case: n=16, p=0.2:

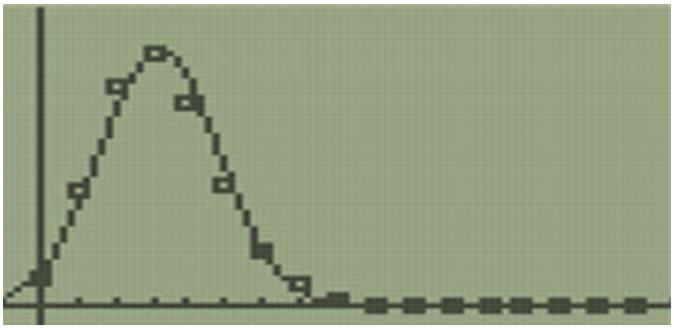


This distribution is still skewed, but only slightly - the values are 'bunched up' toward the low end of the distribution (because p=0.2, a low value).

We have equations for the mean and standard deviation of Binomial distributions:

$$\mu = np = 16(.2) = 3.2$$
$$\sigma = \sqrt{npq} = \sqrt{16(.2)(.8)} = 1.6$$

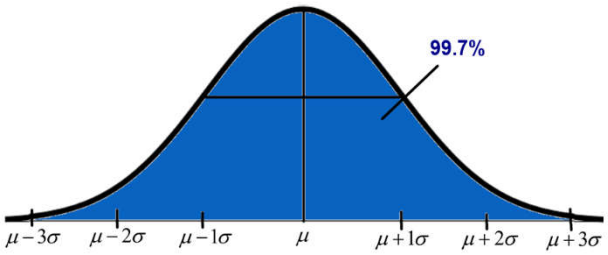
Let's add a Normal distribution curve with this mean and standard deviation to the Binomial distribution scatterplot (I'll omit the histogram bars so we can see things better):



The Normal curve aligns fairly well, but one issue is that the Binomial outcomes are from 0 to 16, while the Normal curve is a model with a domain from $-\infty$ *to* ∞.

That means that there is a mismatch at the low end: the portion of the Normal distribution extending below 0 is not correctly modeling the Binomial distribution. A good criteria we could use to make this a better fit would be to require that little or no Normal distribution extends past the outcome boundaries of the Binomial distribution.

We know that for all Normal distributions, 99.7% of the population is within 3 standard deviations of the mean:



So what if we required that the lower boundary of the Binomial outcomes be at least 3 standard deviations below the mean? Then practically none of the Normal distribution would be 'sticking out' past the Binomial distribution lower end.

The lower end of a Binomial distribution is 0, so we can just require that the point on the Normal distribution 3 standard deviations below the mean be at least zero (so the Binomial zero is past 3 standard deviations from the Normal mean):

$$\mu - 3\sigma > 0$$

$$\mu > 3\sigma$$

Substituting expressions for the mean and standard deviation:

$$np > 3\sqrt{npq}$$

$$(np)^2 > \left(3\sqrt{npq}\right)^2$$

$$n^2 p^2 > 9npq$$

$$np > 9q$$

q will always be ≤ 1, *so this is equivalent to*:

$$np > 9$$

Then, to make things easier to remember, we typically just use $np \geq 10$. The number of successes must be at least 10 in order to approximate a Binomial distribution with a Normal distribution.

And because we could have had a high p-value, a left-skewed Binomial distribution, and be bumping up against the upper outcome limit for the Binomial distribution, using similar reasoning, we need to require the number of failures to also be at least 10: $nq \geq 10$.