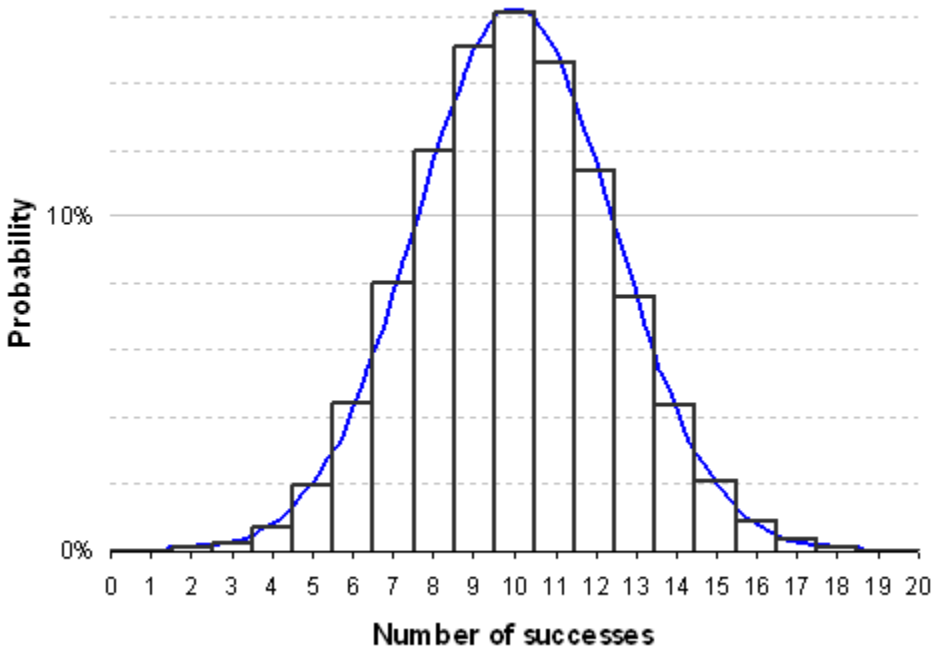
A Justification for the Sample Size Required for Proportion Work

Whenever we deal with samples, we're trying to estimate their related parameter values. In order to do this well, we need a good sample. But what does "good" mean? Well, random's a start; a random selection from a population ensures proper representation. However, we also need ***enough*** random data points. In class, we will see why the magic number 30 is enough when you're dealing with population means. This little paper will show you why you need to ensure that ***np*** and ***nq*** must both exceed 5 (or, better, 10) when you're dealing with population proportions.

Remember, we use the normal distribution to approximate the binomial in these proportional studies[1]. As such, we need to make sure that we get enough data to justify this replacement. In class, we looked at histograms to see why larger samples did the trick, and it seemed that, as soon as ***np*** and ***nq*** got large enough, the binomial curves (regardless of how skewed from 0.5 the values of ***p*** and ***q*** were) began to look normal. But why?

Well, we have to come to grips with something here. A normal (bell – shaped) curve extends infinitely in both directions from - ∞ to ∞; however, a binomial distribution does not. It starts at 0, and goes to ***n***, whatever that ***n*** may be. You might remember this from class; the "bell curves" that we looked at...stopped. They didn't extend infinitely left and right like usual bell curves:



Yes, we can draw a bell curve on top of the binomial histogram, but the tails that extend left of 0 and right of ***n*** must be "removed". That's OK; there isn't a whole lot of "tail" there, anyway. Remember? More than 2 standard deviations from the mean, there's only about 5% of the area, and outside 3 standard deviations, there's only about 0.5%.

[1] Eventually, this whole "replacement" might go away, as computing technology gets better, cheaper, and more ubiquitous.

Now, ***your*** book requires that $np \geq 5$. This is based on the idea that any data outside of 2 standard deviations is "unusual". Basically, it means that, if the data are symmetric, the average of the data should be centered, and that center should be roughly equidistant from 0 and n (your sample size), when viewed graphically. In the graph above, for example, the average is 10, equidistant from 0 and 20 (however, when sampling data, we never hit the center perfectly).

The standard deviation (not to be confused with the collective term "standard deviations") is a distance. All we have to do is ensure that the average is far enough to the right of 0. Your book uses 2 of these distances as a marker, so we just need to make sure that the average is larger than 2 standard deviations[2], or

$$\mu > 2\sigma$$

Now, let's play a little, algebraically:

$np > 2\sqrt{npq}$ (substitute the values for the mean and standard deviation of the binomial)

$n^2 p^2 > 4npq$ (square both sides)

$np > 4q$ (divide out the common factors)

Now, since q is a fraction less than 1, you should believe that $4q < 4$. So, if we require that $np > 4$, we have most definitely satisfied that $np > 4q$. So, your book just rounds the 4 up to a 5. Why? I don't know...but it makes the argument even safer.

To get $nq > 5$ (that is, the right tail), we just need to see that the average needs to be two standard deviations <u>below</u> n...that is:

$$\mu + 2\sigma < n$$

Let's play some more:

$np + 2\sqrt{npq} < n$ (substitute the values for the mean and standard deviation of the binomial)

$n(1-p) > 2\sqrt{npq}$ (move some things around and factor)

$nq > 2\sqrt{npq}$ (remember that $p + q = 1$)

$n^2 q^2 > 4npq$ (square both sides)

$nq > 4p$ (divide common factors)

From this point, the argument follows from above.

The same argument can be used if, instead of 2 σ's, you go 3. Then, the algebra is the same, but you end up with 9's (instead of 4's) in the proof. Then, texts usually round to 10[3]. So there you have it!

---

[2] This is for the left tail...we'll do the right one in a minute.

[3] Maybe it's because the numbers 5 and 10 are "nicer" than 4 and 9 (I'm surmising here; not being trained as a statistician, I try to find the mathematical truth. Sometimes I realize that there's a little hand waving going on).