Picking the Correct Distribution
Binomial, Negative Binomial, Geometric, or Hypergeometric?

Students often become confused when trying to decide whether a random variable in a word problem fits a binomial distribution, negative binomial, geometric or hypergeometric. This paper will explain the similarities and differences between these four related distributions.

First, a binomial random variable must have n independent trials, they must be Bernoulli trials (i.e., two choices only...1 or 0, heads or tails, yes or no, etc.), and the probability of a "success" must be the same on each trial. We call the probability of success p. In order for the probability of success p to be constant, this means that we are sampling from a very large population, so that picking one sample does not materially affect the probability of success on the next sample. An example of a very large population would be the population of all fish in Lake Superior. The population is large enough that taking a few fish out of the lake, even without replacing them, does not materially affect the probabilities on the next pick.

A binomial random variable can tell us the probability of obtaining k successes out of n trials. For example, if we pick 20 people out of a large population, and we know that there is a probability of 40% that any given member of the population smokes, we can define a random variable X as the number of people from our sample of 20 people that smoke. We can calculate $P(X = 0)$ (no one in the sample of 20 people smokes), $P(X = 3)$ (exactly 3 people out of the sample of 20 smoke), or $P(X > 10)$ (the probability that more than 10 people in the sample of 20 people smoke). We can calculate any of these probabilities, and any similar probability as well. In other words, we can calculate the probability of the "number of successes" in n trials. The number of trials is fixed, and the number of successes is random. In fact, our random variable X is defined as the number of successes out of the n trials.

On the other hand, suppose we want to know how many people we need to pick out of population in order to find 2 people who smoke. This is sort of the "inverse" of the problem defined by the binomial distribution. Here, we select people until we get exactly 2 who smoke, and then we stop selecting. We are interested in how many trials we will need on the average in order to obtain k successes. Say the first person we pick smokes, the second doesn't, the third doesn't, and then the fourth does. We now have our two "successes," and so we would stop right there. We would conclude that 4 selections were required in order to obtain the two successes. We can talk about the number of trials required in order to get k successes as a *negative binomial* random variable. Here, the number of successes is fixed, and the number of trials is random. Our random variable X is defined as the number of *trials* required to obtain k successes.

Now, suppose that we just want to draw from our population until we get the *first* success. This is the same as a negative binomial distribution described above, where the number of successes we want is exactly one. Since this is a simpler problem that occurs

often, a separate distribution has been defined for it. It comes up often in real life, so we've given it its own name. We call this case a *geometric* random variable. Other than the fact that we want exactly one success, this is the same as a negative binomial random variable. X is defined as the number of trials required in order to obtain one success.

Finally, suppose that we have 20 people, and we know that 8 of them are smokers. We pick 5 people at random, and we are interested in how many smokers we are likely to pick. We don't have a *very* large population here, and so this is a case of sampling without replacement. This means that though the problem sounds like it is a binomial random variable in every other way, it is not binomial because we are sampling without replacement from a small (very finite) population. In this case, we say that X (the number of successes we pick out of a population of n trials and S successes) is *hypergeometric*. The main difference here is that the probability of success is not constant as is required for a binomial random variable. It changes every time we pick a sample from the small population. In this case, X is defined as the number of successes in a sample that is drawn from a small population where the total number of successes in the population is explicitly known. Another way to look at X is that it is the number of successes in a sample drawn without replacement, where the probability of a success p is not the same in each trial.

So, all four of these distributions are related to the binomial random variable. The way you know which to use is by the situation. What do you want to know? If you want to know something about the number of successes with a fixed number of trials, then it is binomial or hypergeometric. If you are sampling without replacement from a small population, then it is hypergeometric, while it is binomial if the probability of success is fixed for each trial. This information is usually given in the problem, or can be easily inferred from the information given in the problem.

If you are interested in the number of trials you need to get a fixed number of successes, then you are probably talking about a negative binomial or geometric distribution. In a geometric distribution, you are talking about the number of trials needed to obtain exactly one success. A negative binomial doesn't limit you to one success. It has to do with the number of trials needed to obtain k successes.