The Binomial Setting and Distribution

Reading

- Section 3.4.1

Practice Problems

3.6.4 (Page 163) 3.25, 3.26, 3.31, 3.32, 3.33, 3.34

Notes

The Binomial Setting

The *Binomial Setting* is a probability model that describes a situation that occurs frequently, and abstractly it is meant to "emulate" the process of "flipping a coin a fixed number of times".

Take a moment to think of what is important from a mathematical point of view, when we consider flipping a coin a fixed number of times.

> **Binomial Setting**
>
> 1. We have a fixed number of **trials** (the coin flips). We denote it by n.
> 2. Each trial has exactly two possible outcomes. We call one **success** and the other **failure**. These are often called *Bernoulli trials*.
> 3. The probability of success is the same on each trial. We denote it by p.
> 4. The trials are *independent* of each other: Success on the first trial does not alter the chances of success on the second trial.
>
> We could therefore say that the binomial setting is a "fixed number of identical independent Bernoulli trials".
>
> When we are in a binomial setting, there are two primary quantities of interest, and they are both random variables:
>
> - The **number of successes**, usually denoted by X.
> - The **percent of successes**, usually denoted by $\hat{p}$.
>
> We say this X follows a **binomial distribution**, and we write $X \sim B(n, p)$.

When you are given a verbal description of a problem, you have to decide if it "fits" into the binomial setting. To that end, you have to answer each of these questions:

- What would a "trial" be in this problem?
- Do all trials have exactly 2 outcomes?
- What would "success" mean for a trial?
- Do we have a fixed number of trials?
- Do we have the same chance of success for each trial?
- Are the trials independent of each other? What "physical" implications does assuming they are independent have?
- Are we at the end of the day interested in the number of successes, or the percent of successes? If the answer we are after is not related to one of those, then using the binomial setting might not work.

Example:

> We select at random around the U.S. 1000 people and ask them if they support the current president. We want to look at the percent of people who say yes.

Does this fit into a binomial setting? Let us see:

1. We can consider each person we pick and ask as one trial.
2. They have exactly two options, to say yes or to say no.
3. For our purposes we will associate "success" with a yes answer, because we are interested at the end of the day in the percent of people who said yes. So a person answering yes is success for that trial.
4. There is a fixed number of trials, $n = 1000$, since we are asking exactly 1000 people.
5. Each person is selected in exactly the same way, at random among all people, so the chance of success is the same for each trial, and is equal to the percent of people among the entire population who would answer yes to this question. Whatever that percent is, that is our p.
6. Are the trials independent of each other? Does who we have selected for the first say 5 trials have an effect on what might happen on the 6th trial? To an extent it does: We can't pick those 5 people again, so the pool has changed slightly. Given how big the population of the United States is, this change is hardly noticeable. So we can assume the trials are independent.

Let us elaborate on this last point:

When we sample from a population, we change the population size for any subsequent samples. So after we select 3 people from a group, this changes what options we have for the 4th person. How much of an effect this has on the calculations depends on how many people we have altogether.

Suppose for example that we have overall 20 people, and we are considering their gender. Say that to begin with 10 are men and 10 are women. What happens after we have selected three people? What are the chances that the fourth one we pick is a man?

There are two extreme cases: One is if the first three people we picke were all men. Then we have only 7 men left, and 17 people to choose from total, so the chance that the fourth person is a man is $7/17 = 41.18\%$. At the other extreme we have the case where the first three people were all women, and then the chance that the fourth person is a man is $10/17 = 58.82\%$.

So this is a case where we clearly don't have independence of the trials: The chances that the fourth selection is a man depend a whole lot on what happened on the first three trials.

If on the other hand we had 10000 people to begin with, so 50000 of them are men and 50000 are women, then what happens after 3 are picked? On the one extreme, if we pick 3 men, then we have 49997 left, out of 99997, so the chance that the fourth is a man is $49997/99997 = 49.9985\%$. On the other extreme, if we pick 3 women, then we still have 50000 men left and the chance that the fourth is a man is $50000/99997 = 50.0015\%$.

So in this second case the probability changes so little that we might as well pretend that it does not change at all. So for all practical purposes, we can assume the trials are independent.

> If the population size is at least around 20 times larger than the sample size, then we can assume that when sampling from that population each person sampled is independent of the previously selected persons.
>
> The change in population size during sampling is always one source of possible lack of independence, and this is the rough rule that tells us when this is not a problem. You still have to worry about other sources of lack of independence!

Here are some of the problems that might land us in a setting that does not fit the binomial:

1. We might be making a variable number of trials (say if we flip a coin until it comes out heads).
2. There might be more than 2 possible outcomes for each trial (e.g. in voting we might have to consider people who would upstain, or who would vote for an independent candidate).

3. The probability of success might not be the same across trials, for instance if we are considering students' gender, and we pick one student at random from the CC, then the next student at random from a specific fraternity house, then the third student at random from a specific sorority house etc. The chances of getting a male student are different in those 3 cases.

4. There might be reasons to doubt the assumption of independence of trials. For instance if the sample size is too large relative to the population size. But there are other sources of dependence. For example if we pick at random a group of students as they come out of a building, and we ask 2 of them for their gender, then those two selections are probably not independent of each other, as students tend to some extent to spend more time with other students of the same gender.

The Binomial Distribution

When we are in a binomial setting, with parameters n (number of trials) and p (probability of success on a given trial), we are typically interested in the variable X that denotes the number of successes that occur. This is similar to rolling a die say 10 times and counting how many times we get heads.

This is a random variable, as the outcome can vary each time we try it out, but we will describe its distribution precisely.

> We say X follows a **binomial distribution**, and we write $X \sim B(n, p)$.
>
> The main formula for computing using the binomial distribution is:
>
> $$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

So we have a nice compact formula for the chances that X would equal exactly k, or in other words that we would get *exactly* k successes.

Recall the definition of $\binom{n}{k}$. It is a special symbol that indicates "the number of ways of choosing k out of n objects". It is formally computed as:

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)}{k \cdot (k-1) \cdot (k-2) \cdots \qquad 1}$$

Simply put there are exactly k terms top and bottom; the bottom terms go from k and down by 1 till we hit 1, the top terms start at n and go down by 1 matching the bottom terms.

Two extreme cases are when $k = 0$, in which case the answer is automatically 1, and when $k = n$, in which case again the answer is 1.

The explanation of the formula is simple: In order to get exactly k successes we must get k successes, hence the p^k factor, and we must get $n-k$ failures, hence the $(1-p)^{n-k}$ factor. We then have to count this product once for each way we could arrange for the k successes and $n-k$ failures to occur, which is what $\binom{n}{k}$ measures.

Let us do a simple example, in a binomial with $n = 6$ and $p = 0.3$. We have:

$$P(X = 0) = 1 \cdot 0.3^0 \cdot 0.7^6 = 0.1176$$
$$P(X = 1) = 6 \cdot 0.3^1 \cdot 0.7^5 = 0.3025$$
$$P(X = 2) = 15 \cdot 0.3^2 \cdot 0.7^4 = 0.3241$$

and so on. Fill in the rest of the table! Check that the answers add up to 1.