BINOMIAL DISTRIBUTION

The binomial distribution is a particular type of discrete pmf.

It describes random variables which satisfy the following conditions:

1. You perform n identical experiments (called *trials*), where you know n before starting.
2. The random variable counts the number of times a particular event is seen in all of the trials.
3. Each trial is independent of every other.
4. The probability of seeing the event you're interested in (p) is the same for every trial.

Example

- I flip a coin 50 times; let X be the number of times I see heads.
- I inspect 10 products from an assembly line; let Y be the number of products which fail.
- A family has 4 children; let Z be the number of boys.

Counterexamples

The following are **not** binomial random variables:

1. I flip a coin until I see heads; let X be the number of flips. (n is not fixed.)
2. I deal 7 cards; let Y be the number of spades in my hand. (The trials are not independent.)
3. I play in a 5-round tournament; let Z be the number of matches I win. (The probability of winning is different for each opponent.)

Regardless of the details of the specific case, in terms of probability **all binomial random variables behave in the same way**, determined by two parameters: n and p.

Without loss of generality, we say each experiment has only two possibilities: **success** (you see the outcome you're interested in) or **failure** (you don't.)

This is the first **distribution family** we see. There are many more, which describe different kinds of random variables. For each distribution family, we will derive formulas for the pmf, expected value, and variance. Whenever you recognize that a random variable belongs to one of these families, you can use these formulas directly.

Let's say I know n and p. What is the pmf of X?

Step one: what are the possible values X can take? $X \in \{0, 1, \ldots, n\}$

Step two: what is the probability that X takes any of these values?

Example

I inspect 10 products from an assembly line; let Y be the number of products which fail. The probability any product passes is 0.9. What is the probability that $Y = 2$?

The probability of the first two products failing and the last eight products passing is $(0.1)^2(0.9)^8$.

However, I don't care about the order in which the products pass and fail. There are $\binom{10}{2}$ ways in which I can see two products pass and eight fail.

Each of these orderings has the same probability. Therefore
$P(Y = 2) = \binom{10}{2}(0.1)^2(0.9)^8 = 0.19$

For any binomial random variable X we have

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

for any $k \in \{0, 1, \ldots, n\}$.

More Examples

- I flip a coin 50 times; the probability of seeing 20 heads is $\binom{50}{20}(1/2)^{20}(1/2)^{30} = 0.042$
- The probability of winning a casino game is 0.4. If I play 10 times, the probability I win just once is $\binom{10}{1}(0.4)^1(0.6)^9 = 0.040$
- A family has 4 children; the probability of having 2 boys is $\binom{4}{2}(1/2)^2(1/2)^2 = 0.375$

The **Bernoulli** distribution is the special case of the binomial distribution when $n = 1$.

HYPERGEOMETRIC DISTRIBUTION

Remember the four properties of the binomial distribution:

1. You perform n identical experiments (called *trials*), where you know n before starting.
2. The random variable counts the number of times a particular event is seen in all of the trials.
3. Each trial is independent of every other.
4. The probability of seeing the event you're interested in (p) is the same for every trial.

The hypergeometric distribution changes properties 3 and 4.

Rather than every trial being completely independent, imagine that we're sampling from a population **without replacement**.

One of our binomial counterexamples was "how many spades will I see in a seven-card hand?" Each card you are dealt changes the probability of seeing spades for the rest of your hand.

The properties of the hypergeometric distribution can be stated as follows:

1. I sample k objects from a finite population which has b successes and r failures. (k plays the role of n in binomial)
2. The random variable counts the number of times a particular event is seen in all of the trials.
3. Each sample of k objects is equally likely to be chosen.

This is approximately a binomial distribution when b, r, and k are large.

What is the probability of being dealt 3 aces in a hand of 5 cards?

In terms of the hypergeometric distribution, we have $b = 4$ (successes in population), $r = 48$ (failures in population), and $k = 5$ (sample size).

Since each hand of 5 cards is equally likely, we can use a counting argument like that used to derive properties of the binomial distribution.

$$P(A) = \frac{N(A)}{N}$$

The number of 5-card hands is $\binom{52}{5} = 2598960$.

Each hand we are interested in has 3 aces, and 2 non-aces.

The number of ways to choose the 3 aces is $\binom{4}{3} = 4$. (Why?)

The number of ways to choose the 2 non-aces is $\binom{48}{2} = 1128$.

By the fundamental principle of counting, $N(A) = 4 \times 1128 = 4512$.

Therefore, the probability of seeing three aces is $4512/2598960 = 0.00174$, or roughly 1 in 576.

We can generalize this procedure to find the probability that $X = x$ when X is a hypergeometric random variable.

The number of different samples we could have seen is $\binom{b+r}{k}$.

Each hand we are interested in has x successes, and $k - x$ failures.

The number of ways to choose the x successes is $\binom{b}{x}$.

The number of ways to choose the $k - x$ failures is $\binom{r}{k-x}$.

By the fundamental principle of counting, the number of samples with x successes is $\binom{b}{x}\binom{r}{k-x}$.

$$P(X = x) = \frac{\binom{b}{x}\binom{r}{k-x}}{\binom{b+r}{k}}$$

Example

Out of the 46 students in class, 10 will come to my office hours this week. Assuming that there are 20 students in architectural engineering, and that the students who come to my office hours are a random sample, what is the probability that four of them are from architectural?

Define a student from architectural engineering as a "success." Then $b = 20$, $r = 26$, $k = 10$, and $x = 4$.

$$P(X = 4) = \frac{\binom{20}{4}\binom{26}{6}}{\binom{46}{10}} = 0.27$$

NEGATIVE BINOMIAL

Remember the four properties of the binomial distribution:

1. You perform n identical experiments (called *trials*), where you know n before starting.
2. The random variable counts the number of times a particular event is seen in all of the trials.
3. Each trial is independent of every other.
4. The probability of seeing the event you're interested in (p) is the same for every trial.

The negative binomial distribution changes properties 1 and 2.

Rather than counting the number of successes in a predetermined number of trials, we want to know how many trials it will take before we see a predetermined number of successes.

Specifically, a negative binomial random variable counts **the number of trials until the m-th success** (including the last one), where m is given.

I flip a coin until I see 3 heads. How many times do I have to flip the coin?

You are interviewing students to hire 3 interns. Assume that the probability each interviewee is an acceptable hire is 30%. What is the probability I will have to interview 8 students in total?

In terms of the negative binomial distribution, we have $m = 3$ (the number of successes before I stop) and $p = 0.3$ (probability that each interviewee is an acceptable hire).

I know that the last student I interview was a "success." Therefore, of the preceding students, there must be 5 who were rejected, and 2 who were hired, for a total of 8 interviews.

The probability that the 8th interview was a success is 0.3.

The probability of having 2 hires from the first 7 is binomial, with $n = 7$, $p = 0.3$, and $x = 2$.

Therefore, the probability is $\left(\binom{7}{2}(0.3)^2(0.7)^5\right)(0.3)$.

For any negative binomial random variable, we have

$$P(X = k) = \binom{k-1}{m-1} p^m (1-p)^{k-m}$$

for $k \in \{m, m+1, \ldots\}$ and 0 otherwise.

The negative binomial distribution is also known as the **Pascal** distribution.

The **geometric** distribution is the special case when $m = 1$.

Some books define the negative binomial distribution slightly differently, to count the number of *failures* before the m-th success rather than the number of *trials*. The concepts are the same, but the formulas will be slightly different (the random variables will differ by m).

POISSON DISTRIBUTION

One common scenario involves *random occurrences over time*. Examples:

- How many customers arrive at a checkout line in an hour?
- How many vehicles drive past a traffic signal during 15 minutes?
- How many goals will be scored in a soccer match?

None of the distributions we've seen so far represent this situation.

Enter the Poisson distribution

The Poisson distribution is used to count the number of times an event occurs, assuming

1. Occurrences of the event are *independent* of each other.
2. More than one of these events cannot occur simultaneously.

These two assumptions hold true for the scenarios we saw earlier.

- How many transactions will a cashier handle in an hour?
- How many vehicles drive past a traffic signal during 15 minutes?
- How many goals will be scored in a soccer match?

Here are a few counterexamples which do not satisfy the Poisson assumptions:

- How many buses will pass the bus stop in an hour?
- How many people arrive at a checkout line in an hour?

Why?

To derive the Poisson pmf, assume we know the average number of occurrences λ over the time period, and divide the time period into n equal intervals.

Then $P(k)$ can be found using a binomial distribution. The Poisson pmf is found by taking the limit as $n \to \infty$.

If there are n equal intervals, and the average rate is λ, this means $p = \lambda/n$.

For the binomial distribution $P(k) = \binom{n}{k} p^k (1-p)^{n-k}$.

$$\lim_{n\to\infty} P(k) = \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n(n)\cdots(n)} \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{n}$$

$$= \frac{e^{-\lambda}\lambda^k}{k!}$$

Assume that the number of students who stop by my office hours is a Poisson random variable with $\lambda = 2$. What is the probability that no students stop by my office hours? The probability that between 3 and 5 students stop by?

$$P(X = 0) = \frac{e^{-2}2^0}{0!} = 0.135$$

$$P(X = 3) + P(X = 4) + P(X = 5) = \frac{e^{-2}2^3}{3!} + \frac{e^{-2}2^4}{4!} + \frac{e^{-2}2^5}{5!} = 0.307$$

The number of callers to a technical support line is a Poisson random variable. On average, there are 2 calls per hour. What is the probability that there will be at least 2 calls during an eight-hour shift?

For the eight-hour shift, the average number of occurrences is 16, so $\lambda = 16$.

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - \frac{e^{-16} 16^0}{0!} + \frac{e^{-16} 16^1}{1!} = 0.999998$$

For the Poisson PMF to be valid, $P(X = k) \geq 0$ for all k, and $\sum_k P(X = k) = 1$. Are these true?