The Binomial Distribution

XDASI Fall 2021

Contents

References

- **Whitlock & Schluter, Chapter 7: Analyzing Proportions**
- **Optional**: **Chapter 3.1-3.2** *Foundational and Applied Statistics for Biologists with R* (Ken Aho)

Overview

What is a binomial distribution? "Binomial" means "**two names**", and it describes the probability distribution for a series of ***two mutually exclusive events***.

Simple examples include the number (or proportion) of:

- heads in a bunch of coin tosses
- times you get 6 (vs. some other number) in 10 rolls of a fair die
- green M&Ms (vs. other colors) in a bag of mixed colors
- times a professor emails you back when you try to get hold of them repeatedly
- etc.

Each of these scenarios has a certain ***fixed probability of success***. For a coin toss, $p(heads) = 0.5$; the probability of any particular face value of a die is $p(6) = 1/6$; for green M&Ms it is $1/[number\ of\ colors]$ (assuming an even distribution of each); the probability of raising a response from your professor might be $\ldots\ p(response) = ?$.

⇒ ***Question: What are some examples of biological questions that follow a binomial distribution?***

Answer

- The number of lines (animals, cells, etc.) with successful CRISPR edits out of X attempts.
- The number of PCR reactions showing an insert out of X cloning attempts.
- The probability that progeny of a cross will show a particular phenotype, given two different alleles inherited from the parental (or grand-parental) generation.
  - *For example, the probability that the F_2 generation of Mendel's cross between true-breeding round and wrinkled peas would be round or wrinkled (depending on which one is dominant).*
- *Other ideas?*

Success or failure? In statistics lingo, these binary alternative outcomes are called **"successes"** vs. **"failures"** in a series of ***random trials*** (where the total probability of all possible events, of course, is 1).

The binomial is one of the most common discrete probability distributions and allows us to answer questions of the form,

What is the total probability of getting X (or $>, \geq, <, \leq$ X) successes out of out of Y trials?

In order to understand how the binomial distribution works, we need to consider the joint probabilities of multiple independent events, as well as all the ways it is possible to get different outcomes. Let's review some basic mathematics underlying combinatorial probabilities, and then apply these concepts to the binomial distribution.

Permutations

A ***permutation*** is a particular ***ordering*** of objects, like a DNA or protein sequence.

Permutations of n objects

The number of possible ***orderings*** of n objects is "n factorial", which is written as "***n!***". For example:

- There are 6 ways to arrange 3 nucleotides to make a codon: $3! = 6$.
    - One possible set is $S = A, G, U$, and the codons you can make from these are: AGU (Ser), AUG (Met), GAU (Asp), GUA (Val), UAG (Stop), UGA (Stop).
- 5 amino acids can be arranged in 120 different ways: $5! = 120$.
    - This will apply even if two or more a.a.'s are the same, as long as we consider each one as an individual, e.g. RE_1E_2DY.

Sampling without replacement

If you randomly pick a ***subset*** of k objects from a ***finite set*** of n objects, the then ***the number of possible sequences is much larger***.

For example, if you you select 5 amino acids at random from a peptide that is 22 residues long, there will be 22 choices for the first a.a. That will leave 21 choices left for the next one, and so on. Because the number of residues to choose from decreases by one each time, the number of possible sequences of any 5 a.a. picked at random will be $22 * 21 * 20 * 19 * 18$.

We can use a very simple mathematical trick to find a convenient way to express this. We just multiply $22 * 21 * 20 * 19 * 18$ by 1, but we write it as $1 = \frac{17!}{17!}$. So, we can now write out the number of permutations, $nPerm$, as:

$$\begin{aligned} nPerm &= 22 * 21 * 20 * 19 * 18 * \frac{17!}{17!} \\ &= 22 * 21 * 20 * 19 * 18 * \left(\frac{17 * 14 * ... * 2 * 1}{17 * 14 * ... * 2 * 1} \right) \\ &= \frac{22!}{17!} = \frac{n!}{(n-k)!} \end{aligned}$$

where k is the number of objects picked (a specific ***subset*** of individual objects), and n is the total number of objects to pick from (the ***universal set***).

Probabilities of permutations

The ***probability*** of picking **one item** from a set of n items is $1/n$, i.e. the inverse of the number of items to choose from. If we extend this to a set of k out of n items, the probability of a particular ***ordering*** of the k items is the **product** of the individual probabilities of each independent random event.

However, because n is finite, the probability for each successive event is not constant! For the peptide example, the probability of a picking the first a.a. is $\frac{1}{22}$; for the second it is $\frac{1}{21}$, and so on.

We can use the same trick we used above to express this in general terms:

$$\begin{aligned} \frac{1}{nPerm} =& \frac{1}{22} * \frac{1}{21} * \frac{1}{20} * \frac{1}{19} * \frac{1}{18} \\ =& \frac{17!}{22!} = \frac{(n-k)!}{n!} \end{aligned}$$

So we see that the ***probability*** of any permutation of of independent random variables drawn from a finite set is just the ***inverse*** of the number of possible permutations. In such a case we are ***sampling without replacement***: the number of possible choices decreases by one with each successive event, and the corresponding probabilities change accordingly.

Combinations

Now let's say we don't care about the ***ordering*** of those 5 amino acids at all; we just want to know how probable it is that we pick **any particular set of 5 a.a.**. Instead of 5! = 120 possible orderings of 5 a.a., we now have just **one set** of 5 a.a. in no particular order: $S = \{x_1, x_2, x_3, x_4, x_5\}$. So, we just divide our permutations by 5! to get the number of combinations.

More generally, for n items, this reduces the number of possibilities by $k!$, since we collapse all of the independent possible orderings (***permutations***) into just one set (***combination***).

$$nComb = \frac{nPerm}{5!} = \frac{22!}{5! * 17!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

This is the **binomial coefficient**, which we call "***n*** choose ***k***".

Binomial probability

When we perform more than one random trial, we are interested in the probability of obtaining one particular outcome relative to all possible outcomes.

For example, if you want to make a strain with a GFP-tagged protein by CRISPR, and your probability of getting a transformant is 20%, then if you test 10 independent lines, the probability that the first 3 lines you pick were transformed and the next 7 were not is $Pr[T]^3 * Pr[!T]^7$.

This is one ***permutation*** of possible outcomes. But this represents only one of the ways you could get 3 transformants! It could be that the last 3 you picked were transformed, or the 1st, 4th, and 7th, etc. To calculate the probability of getting **any 3** transformants out of 10, we need to know two things:

1. The **number of ways** we can get 3 successes and 7 failures out of 10 trials.
2. The **probability** of getting exactly 3 successes and 7 failures out of 10 trials.

How many ways it is possible to get this number of transformants? This is where the binomial coefficient comes in! We could use **decision trees** to visualize how many possibilities exist for the different permutations of outcomes, but this approach becomes increasingly unwieldy as the number of independent trials grows. Instead, we can just use the formula we learned for ***combinations***:

$$nComb = \frac{nPerm}{3!} = \frac{10!}{3!7!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

The **binomial coefficient** provides a general solution to the question, *"How many ways are there to get k successes and $n - k$ failures out of n trials?"*

We can also use a simple formula to generalize the probability of a particular outcome. Since each random trial is independent, we multiply the probability of success for each trial by the number of successes, and the probability of failure by the number of failures. For our example, this is:

$$\begin{aligned} Pr[transformed]^3 * Pr[NOT \;\; transformed]^7 &= Pr[success]^3 * Pr[failure]^7 \\ &= Pr[success]^3 * Pr[failure]^{10-3} \end{aligned}$$

since the number of failures is just $10 - number\ of\ successes$. This can be expressed more generally as follows:

$$\begin{aligned} Pr[(k\ successes)\ \cap\ (n-k\ failures)] &= Pr[success]^k * Pr[failure]^{n-k} \\ &= \pi^k(1-\pi)^{n-k} \end{aligned}$$

where π is shorthand for $Pr[success]$, $(1-\pi)$ is $Pr[failure]$ (since these are mutually exclusive and add to 1), k is the number of successes, and n is the total number of trials, so that $n-k$ is the number of failures.

Now we can put this all together, and say that ***the total probability of getting any 3 successes out of 10 trials*** is:

$$Pr[3\ \ successes\ \ out\ \ of\ \ 10\ \ trials] = \binom{10}{3} * (0.2)^3 * (0.8)^7$$

Or more generally,

$$Pr[k\ \ successes\ \ out\ \ of\ \ n\ \ trials] = \binom{n}{k}\pi^k(1-\pi)^{n-k}$$

Binomial proportions

If you want to find out the total probability that you will get **3 or less** transformants, then you need to add up all the possible ways you could get one, two, or three transformants. This will be just the sum of outcomes for $k=1$, $k=2$, and $k=3$:

$$\begin{aligned} Pr[\leq 3\ \ successes\ \ out\ \ of\ \ 10\ \ trials] =& \binom{10}{0} * 0.2^0 * (0.8)^{10} + \\ & \binom{10}{1} * 0.2^1 * (0.8)^9 + \\ & \binom{10}{2} * 0.2^2 * (0.8)^8 + \\ & \binom{10}{3} * 0.2^3 * (0.8)^7 \\ =& \sum_{k=0}^{3} \binom{10}{k} * 0.2^k * 0.8^{1-k} \end{aligned}$$

More generally, we can write:

$$Pr[\leq k\ \ successes\ \ out\ \ of\ \ n\ \ trials] = \sum_{0}^{k} \binom{n}{k}\pi^k(1-\pi)^{n-k}$$

The Bernoulli distribution

The Bernoulli distribution is a **discrete distribution** that describes the ***probability of success for a single trial of a binary random variable***. If we encode the outcomes as either 1 (“success”) or 0 (“failure”), then we can write a formula for this as:

$$f(x) = P(X=x) = \pi^x(1-\pi)^{1-x}, \quad x \in \{0,1\}$$

where π represents the probability of “success”, and ranges from zero to one: $0 \leq \pi \leq 1$.

The formal way of writing this function seems kind of complicated, but it’s really pretty simple. Since x can only take on values of 0 or 1, and there is only one trial, $f(x)$ can take on only one of two values:

$$
\begin{aligned}
f(1) &= P(X = 1) = \pi \\
f(0) &= P(X = 0) = 1 - \pi
\end{aligned}
$$

The CDF is:

$$F(X) = \begin{cases} 1 - \pi & x = 0 \\ 1 & x = 1 \end{cases}$$

What does this mean in practice? For a fair coin toss, the chance of getting heads or tails is the same: $P(X = 1) = P(X = 0) = 0.5$.

Illustration

Let's say you are trying to fuse a GFP tag onto the end of a CDS in *C. elegans* using CRISPR, and the efficiency is around 20% (this may be unrealistically high, depending on how well CRISPR is currently working in your lab, but let's go with it anyway). This means that if you pick only one worm, the chance of recovering a line WITH the GFP tag is 20%, and the chance of that worm NOT having the GFP tag is 80%. So,

$$
\begin{aligned}
P(X = 1) &= 0.2 \\
P(X = 0) &= 0.8
\end{aligned}
$$

The Binomial distribution

What if you are not just interested in a single Bernoulli trial, but you want to know how many worms you will have to pick to get at least three independent GFP lines, or to have an 80% chance of getting a GFP transformant? Being able to figure out the answer to this kind of question can help you plan your experiments better.

This is what the binomial distribution is for! The binomial describes the ***sum of many Bernoulli trials***, and it is one of the most fundamental distributions in probability theory. More precisely, it gives the probability of a particular number of "successes" x, given n i.i.d. ("independent and identically distributed") Bernoulli trials.

The binomial distribution is a function of two parameters: a random variable, n, and the probability of success for a single trial, π. We denote it as $X \sim BIN(n, \pi)$, where the tilde means that the random variable X "follows" the binomial distribution.

Probability mass function (PMF)

The binomial probability mass function (PMF) is:

$$f(x) = P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad x \in \{0, 1, ..., n\}$$

The above equation simply says that, in order to find the probability of a particular outcome x, we need to multiply three things:

- the probability of x "successes", (π^x),
- the probability of $n - x$ "failures", $(1 - \pi)^{n-x}$, and

- the total number of ways this number of "successes" can happen, $\binom{n}{x}$.

This is the same as the binomial probability we discussed above! Breaking this equation down into its component parts, we see that:

1. The term $\binom{n}{x}$ is the **binomial coefficient**, "n choose x". It is the ***number of possible combinations*** of x successes (and $n - x$ failures) out of n Bernoulli trials.
2. The rest of the equation is just the **probability** of ***one of these combinations***: $\pi^x(1-\pi)^{n-x}$. Each trial is independent, so we follow the **Product Rule** to find the probability that any number of trials had a particular outcome:
   - Since each trial has π probability of success, the probability of x successes is π^x.
   - Similarly, for x successes there are $n - x$ failures, each with probability $1 - \pi$, so the probability of $n - x$ failures is $(1 - \pi)^{n-x}$.

The mean and variance of a Binomial distribution are: $\mu = E(X) = np$ and $\sigma^2 = V(X) = np(1 - p)$.

Cumulative Distribution Function (CDF)

The full CDF for the binomial distribution gives the total probability and its equation is:

$$F(X) = \sum_{x=0}^{n} \binom{n}{x} \pi^x (1 - \pi)^{n-x} = 1$$

If we are interested to find the total probability that $X \leq x_i$, this would be:

$$P(X \leq x) = \sum_{x_i=0}^{x} \binom{n}{x_i} \pi^{x_i} (1 - \pi)^{n-x_i}$$

Fortunately, R has the built-in function `binom()` family of functions that let us compute these probabilities automatically, so that we don't have to do all of this by hand!

Illustration

Let's continue the CRISPR example above. You probably need to pick more than one worm to find your GFP strain! Let's say you pick 3 worms. What's the probability that two out of the three will be transformants?

First, let's ask how likely it is that the first two worms you pick will be transformants, and the third will not? Well, that works out to $(0.2) * (0.2) * (0.8) = 0.032$.

But, you're not done yet! You need to consider ***how many ways*** there are to get 2 out of 3 transformants. Let's work this out using Set Theory:

- There is only one way to get zero transformants in three tries: $S = \{000\}$.
- There are three ways to get one GFP strain and two non-GFP strains: $S = \{100, 010, 001\}$.
- Similarly, there are 3 ways to get 2 GFP strains and 1 non-GFP strain: $S = \{110, 101, 011\}$.
- Finally, there is only one way to get 3 transformants in three tries: $S = \{111\}$.

This is what the $\binom{n}{k}$ part of the equation is for! Now we are ready to solve the problem:

$$P(X = 2) = \binom{3}{2}(0.2)^2(0.8) = 3 * 0.032 = 0.096$$

If you pick only three worms, you'll have about a 1 in 10 chance of finding exactly two transformants.

However you're probably more interested in the chance of finding ***at least*** 2 transformants (it's always good to have more than one independent CRISPR line!). To do this, we will need to use the CDF. Specifically, we are interested in the ***upper-tail*** probability that we will find *more than one* transformant. To do this, we add up the *lower-tail probabilities* for zero or one transformants, and subtract the sum from 1:

$$\begin{aligned} P(k \leq 1) &= \binom{3}{0}(0.8)^3 + \\ &= \binom{3}{1}(0.2)(0.8)^2 \end{aligned}$$

So, $P(X \leq 1) = 0.512 + 0.384 = 0.896$ and $P(X > 1) = 1 - (0.512 + 0.384) = 0.104$. This is slightly better, but not much!

You can compute this using the R function for the PDF. Fortunately, it gives the same result!

```
# cumulative upper-tail probability of getting MORE THAN one transformant: P(X > 1)
1-pbinom(1,3,0.2)                # 1 minus the lower-tail probability
## [1] 0.104
pbinom(1,3,0.2,lower.tail=FALSE) # this is equivalent
## [1] 0.104
```

How many worms would you have to pick to guarantee an 80% chance of getting at least two transformants? It would be pretty tedious to calculate this out by hand, especially as the number of trials increases.

```
# probability of getting at least 2 transformants for different numbers of worms picked
pbinom(1,3,0.2,lower.tail=FALSE)
## [1] 0.104
pbinom(1,5,0.2,lower.tail=FALSE)
## [1] 0.26272
pbinom(1,8,0.2,lower.tail=FALSE)
## [1] 0.4966835
pbinom(1,11,0.2,lower.tail=FALSE)
## [1] 0.6778775
pbinom(1,14,0.2,lower.tail=FALSE) # you should check at least 14 if you want 2 or more!
## [1] 0.8020879

# you can find the number of transformants corresponding to the 80% quantile given some sample size
qbinom(0.8, 13, 0.2, lower.tail=FALSE)
## [1] 1
qbinom(0.8, 14, 0.2, lower.tail=FALSE)
## [1] 2
```

The Binomial Theorem

One interesting feature of the Binomial distribution is that it is **symmetric**. This means that the number of ways you can get exactly 2 out of 3 "successes" is the same as the number of ways you can get exactly 1 out of 3 successes.

In the example above, if $Pr[success]$ were 0.5 instead of 0.2, then you'd have the same chance of finding exactly one or exactly two transformants, since the chance of success or failure would be the same.

Pascal's Triangle

As we learned above, the term "n choose k" has a special name, the **Binomial coefficient**. For any n, it is possible to work out the number of possible ways of achieving any outcome using Pascal's Triangle:

$$\begin{array}{ccc} \binom{0}{0} & = & 1 \\ \binom{1}{0}\binom{1}{1} & = & 1 \quad 1 \\ \binom{2}{0}\binom{2}{1}\binom{2}{2} & = & 1 \quad 2 \quad 1 \\ \binom{3}{0}\binom{3}{1}\binom{3}{2}\binom{3}{3} & = & 1 \quad 3 \quad 3 \quad 1 \end{array}$$

The top line represents the possible combinations for $n = 0$, the second line for $n = 1$, etc. Notice that the sum of two components at a higher level of Pascal's triangle equals the component of the next lower level that is situated directly beneath and between them.

It is easy to see that the Bernoulli distribution is a special case of the Binomial distribution with $n = 1$. There's only one way to get one success or one failure in one trial!

Binomial Expansion

If we consider two Bernoulli trials (where the probability of a success for each trial is π), we can use the product rule to work out the probability of two successes (call these A), two failures (call these B), or one success and one failure:

$$\begin{aligned} P(A \cap A) &= P(A) * P(A) && = \pi^2 \\ P(A \cap B) &= 2 * P(A) * P(B) && = 2\pi(1-\pi) \\ P(B \cap B) &= P(B) * P(B) && = (1-\pi)^2 \end{aligned}$$

There is only one way to get two successes or two failures ($S = \{00\}$ or $S = \{11\}$), and two ways to get one of each ($S = \{10, 01\}$). This pattern should look familiar to you; it's the same as for the binomial expansion:

$$(x+y)^2 = x^2 + 2xy + y^2$$

The expansion above is a special case of the **Binomial Theorem** for n independent trials, $(x+y)^n$, where $n = 2$. What would this look like for 3 trials?

$$(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

Do you start to see a pattern here? The binomial theorem can be extended to any arbitrary number of trials using the product rule for independent events. The binomial coefficients can be found using the "n choose k" shortcut, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Instead of writing out Pascal's Triangle to get the number of possible combinations for each outcome, we just use the general formula for arbitrary n and k:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

The second expression is equivalent to the first because the Binomial distribution is symmetric.

More generally, we can say that $(x+y)^n$ can be expressed as a sum of terms of the form $ax^b y^c$, where $b+c=n$ and each a is a ***binomial coefficient*** that is a specific positive integer that depends on n and b (or, equivalently, n and c).

In the limit, a binomial distribution looks very much like a normal, or Gaussian distribution. We will convince ourselves of this in a future class exercise.

Postscript

Maximum likelihood

Sometimes we have some data, but we don't know the probability p, so we want a way to estimate this population parameter given the data. We can use **maximum likelihood estimation (MLE)** to find the most likely value of p, given the data. We will come back to this topic toward the end of the course.

Multinomial distribution

If there are multiple mutually exclusive alternatives, the ***multinomial*** distribution is used to describe probabilities for more than two mutually exclusive events. The binomial distribution is actually a special case of the multinomial distribution where the number of alternatives is 2.

Examples include:

- The probabilities of different phenotypic outcomes for more than one independently segregating trait (e.g. yellow/green and smooth/wrinkly peas).
- The joint probability that a group of randomly selected people belong to different ABO blood groups: A, B, AB, O. Here, we have four categories, $k=4$, and the probability of each blood group, p_i, is different.

The algebraic expansion for n trials would then be: $p_1 + p_2 + p_3 + p_4)^n$.

The joint probability function for a multinomial distribution with $i \in \{1, 2, ..., k\}$ categories, each with frequency x_i and probability p_i, is:

$$P(x_1, x_2, ..., x_k | p_1, p_2, ..., p_k) = \binom{n}{x_1, x_2, ..., x_k} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$$

where $\sum_{i=1}^{k} x_i = n$ and $\sum_{i=1}^{k} p_i = 1$.

The multinomial is often used as an approximation in genetics since the sampling proportion is small relative to the entire population (so we can treat the data as if we are sampling without replacement).

We will not cover multinomial distributions in this course, but it's good to know they exist!