Review: Probability and the Binomial Distribution

Kris Gunsalus

9/24/2020

Contents

Random variables

When we perform experiments, we are doing "**random trials**" with two or more possible "**outcomes**" that cannot be known with certainty in advance of measurement or observation.

Different possible outcomes are called "**events**". This might be something like "Probability that an individual carries allele X", or "Brood size is between 20-50% of WT", etc.

When we want to evaluate our results, for example in hypothesis testing, we want to know the **frequency** with which different outcomes are obtained. This is a **probability.**

Set theory and notation

Sets

In order to compute probabilities we need to be able to enumerate **all possible outcomes**.

The set of possible outcomes, or the **sample space**, is the universal set S. For **discrete** data, we can enumerate the elements of the set in curly brackets:

$$S = \{x_1, x_2, ...x_n\}$$

Where n is the total number of possible distinct outcomes. Each x_i is an **element** of S, and we write this as:

$$x_i \in S$$

An **event** is any plausible subset of possible outcomes. For example, $A = \{x_1\}$, $A = \{x_1, x_2\}$, and $A = \{x_1, x_2, x_3\}$ are different **events**, and all of these are **subsets** of S.

For example, say you have made a genetic cross to get a 3xFLAG-tagged gene into a mutant background. You test 8 progeny by PCR to see which ones have he 3xFLAG tag in them. The set of PCR reactions is

$S = \{P_1, P_2, P_3, ...P_8\}$. One possible outcome is that that the first and third reactions have the 3xFLAG in them, and the rest don't. This subset would be $S = \{P_1, P_3\}$.

Probability of an event

The **probability** is the **proportion** of random trials with a particular outcome. The **probability of event** A is its frequency of occurrence, relative to all other outcomes:

$$Pr[A] = \frac{N[A]}{N}$$

Where $N[A]$ is the number of times A was observed, and N is the total number of observations.

The probability of any event ranges from 0 to 1, and the probability of the full sample space $Pr[S] = 1$.

So, the probability that some event does **NOT** occur is $1 - Pr[A]$.

If your 3xFLAG strain was a heterozygote, then the chance of any single PCR reaction being positive is $Pr[3xFLAG] = 0.5$, and $Pr[NOT \;\; 3xFLAG] = 1 - 0.5 = 0.5$.

Venn diagrams: Addition and multiplication rules

Mutually exclusive vs. independent

Two events are **mutually exclusive** if they cannot occur at the same time (for example, a particular allele cannot be both A1 and A2):

$$Pr[A \;\; AND \;\; B] = Pr[A \cap B] = 0$$

If two events are unrelated, and knowing something about one gives no information about the other, then they are **independent**:

$$Pr[A \;\; AND \;\; B] = Pr[A \cap B] = Pr[A] * Pr[B]$$

Many combinations of genetic traits are **independent** (e.g. Mendel's green and wrinkly peas).

Sometimes, however, two genes will show a ***genetic interaction***, in which case the independence rule is violated. When this happens, a functional association between these genes will show up as a deviation from independence.

Addition rule: A or B (or both)

If two things are **mutually exclusive**, then the probability that ***either one*** occurs is simply their sum:

$$Pr[A \;\; OR \;\; B] = Pr[A \cup B] = Pr[A] + Pr[B]$$

If two things are **NOT** mutually exclusive, then to find their **union** (the chance that either or both has occurred) we need to subtract the probability that they are co-occurring from the total probability that each one occurs (or else we would be counting the overlap twice).

$$Pr[A \;\; OR \;\; B] = Pr[A \cup B] = Pr[A] + Pr[B] - Pr[A \;\; AND \;\; B] = Pr[A] + Pr[B] - Pr[A \cap B]$$

We apply this **general addition rule** whenever two events are **NOT** mutually exclusive.

Permutations

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

The number of possible orderings of n objects is "n factorial", which is written as $n!$. For example, if you have 5 playing cards, then you can arrange them in 120 different ways: $5! = 120$.

If you randomly pick a subset of k items from a larger set of items n, then the number of possible sequences is much larger. For example, if you pick 5 cards from a deck of 52, then you will have 52 choices on the first pick, 51 choices, on the second pick, and so on. The number of ways to order 5 cards picked at random will be $52*51*50*49*48$.

We can use a very simple mathematical trick to find a convenient way to express this. We just multiply by 1, but we express it as

$$\frac{47!}{47!} = \frac{47 * 46 * ... * 2 * 1}{47 * 46 * ... * 2 * 1}$$

So, we can now write out the number of choices as:

$$nPerm = 52 * 51 * 50 * 49 * 48 = 52 * 51 * 50 * 49 * 48 * \frac{47 * 46 * ... * 2 * 1}{47 * 46 * ... * 2 * 1} = \frac{52!}{47!} = \frac{n!}{n - k!}$$

The **probabiity** of picking a particular set of 5 cards in some order is:

$$\frac{1}{52} * \frac{1}{51} * \frac{1}{50} * \frac{1}{49} * \frac{1}{48} = \frac{1}{nPerm} = \frac{47!}{52!} = \frac{n - k!}{n!}$$

Combinations

Now, say we don't care about the ordering of those 5 cards at all; we just want to know how probable it is that we pick any 5 cards. Instead of $5! = 120$ possible orderings, we now have just one set of 5 cards: $S = \{1, 2, 3, 4, 5\}$ in no particular order.

This reduces the number of possibilities by 5. So, we just divide our permutations by 5! to get the number of combinations:

$$nComb = \frac{nPerm}{5!} = \frac{52!}{5! * 47!} = \frac{n!}{k!(n - k)!} = \binom{n}{k}$$

This is the **binomial coefficient**, "$\boldsymbol{n}$ choose $\boldsymbol{k}$".

Binomial coefficient

When we perform more than one random trial, we are interested in the total probability that something occurs our of all the 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 different lines, the probability that the first 3 lines you pick were transformed and the next 7 were not is:

$$Pr[T] * Pr[T] * Pr[T] * Pr[!T] * Pr[!T] * Pr[!T] * Pr[!T] * Pr[!T] * Pr[!T] * Pr[!T] = Pr[T]^3 * Pr[!T]^7$$

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.

In terms of computing probabilities, we are interested in the **number of successes** for a series of **random trials**. In this example, that is 3 transformants out of 10.

So we need to know **how many ways** it is possible to get this number of transformants. This is where the **binomial binomial coefficient** comes in.

We can use **decision trees** to visualize how many possibilities exist for the different permutations of outcomes, and thus the number of possible combinations. However, this approach becomes increasingly unwieldy as the number of independent trials grows.

Instead, we can just use the formula we learned for combinations:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Binomial probability

To calculate the total probability of getting 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. To get this number, we just use the **binomial coefficient**, which provides a general solution to the question, *"How many ways are there to get k successes and $n-k$ failures out of n trials?"*
2) The **probability** of getting exactly 3 transformants in a single experiment. In the above example, we already found that this works out to $Pr[T]^3 * Pr[!T]^7$.

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 times the number of successes, and the probability of failure times the number of failures. For our example, this is:

$$Pr[transformed]^3 * Pr[NOT \;\; transformed]^7 = Pr[success]^3 * Pr[failure]^7$$

Since $Pr[failure]$ just equals $1 - Pr[success]$, this can be expressed more generally as follows:

$$\pi^k \pi^{n-k}$$

Where we use π as shorthand for $Pr[success]$, k for the number of successes, and n for the total number of trials, so that $n-k$ equals $Pr[failure]$.

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 how probable it is 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 $k = 3$:

$$Pr[\leq 3 \ successes \ out \ of \ 10 \ trials] = \binom{10}{0} * 0.2^0 * (0.8)^{10} + \binom{10}{1} * 0.2^1 * (0.8)^9 + \binom{10}{2} * 0.2^2 * (0.8)^8 + \binom{10}{3} * 0.2^3 * (0.8)^7 = \sum_0^3 \binom{}{}$$

More generally, we can write:

$$Pr[\leq k \ \ successes \ \ out \ \ of \ \ n \ \ trials] = \sum_0^k \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

Fortunately, R has built-in functions that let us compute these proportions, so that we don't have to do all of this by hand.