

Basics of Probability Theory

A) Counting

In a population of size N , how many ways are there to choose x individuals (i) with replacement; or (ii) without replacement?

With replacement, there are N possible choices for the first individual, N possible choices for the second individual, N possible choices for the third individual, etc. Thus, the number of possible ways to choose x individuals is N^x .

Without replacement, there are N possible choices for the first individual, $(N - 1)$ possible choices for the second individual, $(N - 2)$ possible choices for the third individual, etc. Thus, the number of possible ways to choose x individuals is

$$N*(N - 1)*(N - 2)*\dots (N - x + 1) = N!/(N - x)!$$

Note that in the results above we assume that each different order in choosing x individuals counts as a different way (e.g., $\{A,B,C,D\}$ counts as a different outcome than $\{B,A,C,D\}$). What if we don't care about order (i.e. if $\{A,B,C,D\}$ counts as the same type of outcome as $\{B,A,C,D\}$ or any other order of the these four)?

We will only concern ourselves with the case of sampling *without replacement* (i.e., no duplicates). Given that you have a set of x unique individuals how many different ways could you order them? Well, there are x choices for the first spot, $(x - 1)$ for the second spot, etc. Thus there are

$$x*(x - 1)*\dots 1 = x!$$

ways to rearrange each a sequence of x different individuals. Let y be the number of different *unordered* ways to sample x individuals. Then

$$y*x! = N!/(N - x)!$$

$$y = \frac{N!}{x!(N - x)!} \equiv \binom{N}{x}$$

The last term is read as “ N choose x ” and is extremely useful.

Q1.1) You want to make a haploid model with loci A, B, and C, having 2, 3, and 4 alleles, respectively. How many haplotypes do you have to deal with?

Q1.2) You want to make a diploid model with just 1 locus but having 3 alleles. Assuming no parent-of-origin effects (A_1/A_2 and A_2/A_1 are equivalent) how many genotypes must you track? More generally, what if there are i alleles? (Hint: count the number of homozygotes and heterozygotes separately.)

Q1.3) You want to make a diploid model with k di-allelic loci. Assuming no cis-trans effects or parent-of-origin effects (e.g., genotypes AB/ab , ab/AB , Ab/aB , aB/Ab are all equivalent), of how many genotypes must you keep track?

Q1.4) Epistasis is the term used to describe interactions that occur between loci. Imagine that some trait is controlled by 50 genes. How many different two-way epistatic interactions are needed to fully describe this model? How many 10-way epistatic interactions are needed? How many 40-way epistatic interactions? How many 48-way epistatic interactions?

b) Basic rules of probability theory

The probability of an event A, must lie in the interval $0 \leq P(A) \leq 1$.

The sum of all possible mutually exclusive events is 1. For a discrete probability distribution with N different possible outcomes,

$$\sum_{i=1}^N P(x_i) = 1$$

For a continuous probability distribution (e.g., normal distribution) with the *probability density function* $p(x)$

$$\int_{-\infty}^{+\infty} p(x) dx = 1$$

If event A occurs with probability $P(A)$, then its **complement** (i.e., A does NOT occur) has probability $1 - P(A)$.

The **conditional probability** that A occurs given that B occurs is

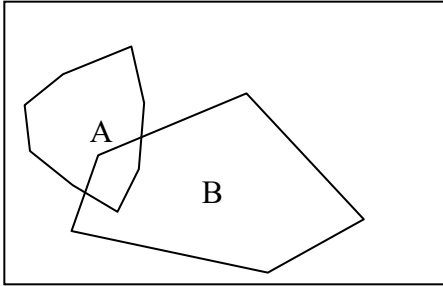
$$P(A|B) = P(AB)/P(B).$$

Events A and B are **independent** if the occurrence of B provides no information about whether A will occur (and vice versa). That means that $P(A|B)$ should simply be $P(A)$, so we can write

$$P(A|B) = P(AB)/P(B) = P(A) \text{ which can be rearranged to give}$$

$$P(AB) = P(A) P(B)$$

That is, if A and B are independent, then the probability that both A and B occur is simply the product of each event. This result is extremely important.



Regardless of the relationship between A and B, it is always true that the **marginal** probability of A is

$$P(A) = P(AB) + P(A \text{ \& not B})$$

The probability of event A *or* B is (consider the diagram)

$$P(A \text{ or } B) = P(A \text{ \& not B}) + P(B \text{ \& not A}) + P(AB)$$

$$\begin{aligned} P(A \text{ or } B) &= (P(A) - P(AB)) + (P(B) - P(AB)) + P(AB) \\ &= P(A) + P(B) - P(AB). \end{aligned}$$

By definition, $P(AB) = 0$ for **mutually exclusive** events. Thus, if A and B are mutually exclusive, the probability of A or B is simply the sum of their marginal probabilities:

$$P(A \text{ or } B) = P(A) + P(B)$$

Q1.5) If 5% of men in a population are both taller than 6' and have red hair, and 50% of men are taller than 6'. What is the probability that a man has red hair given that he is over 6'? What can be said about the total frequency of redheaded men in this population?

Q1.6) Are mutually exclusive events independent? Prove it. (To prove something is true you need to be able to show that it holds in general. To prove something is false, you need only find a single example that shows the statement is false.)

A **random variable** is something whose value depends on the outcome of some stochastic event. We assign a number to different outcomes. For example, in a coin flip we might say $x = 1$ for a heads and $x = 0$ for a tails (i.e., x is a random variable, the value of which depends on the outcome of the coin flip). Alternatively, if we are rolling a 6-sided die, x might simply be the number that shows on the die (i.e., x can be 1, 2, 3, 4, 5, or 6).

A **probability distribution** is a function that gives the probability of a certain event occurring (i.e., that the random variable will take a particular value). For example, on a

regular six-sided die, the probability that the number obtained on a single role will be X is given by the function

$$P(X = x) = \begin{cases} 1/6 & \text{for } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

c) Means, (co)variances, and central moments

For a discrete probability distribution with outcomes x_i the expected value (e.g., the mean or average) of x is

$$E(x) \equiv \sum_i P(x_i)x_i \equiv \bar{x}$$

or for a continuous distribution

$$E(x) \equiv \int (p(x)x)dx \equiv \bar{x}$$

Helpful rules when working with expectations:

As we work through the derivations below you will see that the following rules are true. Learning these rules will help you solve problems more quickly.

- (1) The expectation of a sum is equal to the sum of the expectations of each part.

$$\text{For example, } E[ax + b^3x^2 + y - d] = E[ax] + E[b^3x^2] + E[y] - E[d]$$

- (2) A constant can always be factored out of an expectation.

$$\text{For example, if } b \text{ is a constant and } x \text{ is random variable } E[b^3x^2] = b^3E[x^2]$$

- (3) The expectation of a constant is equal to a constant.

$$\text{For example, if } d \text{ is a constant } E[d] = d.$$

$$(\text{We could use rule \#2 to get } E[d] = dE[1] = d)$$

Let's consider what happens if we want to take the expected value of a linear function of x such as $y = ax + b$ where a and b are constants (NOT random variables).

$$\begin{aligned} E(y) &= \sum_i P(x_i)y_i = \sum_i P(x_i)(ax_i + b) = \sum_i P(x_i)(ax_i) + \sum_i P(x_i)(b) \\ &= a \sum_i P(x_i)x_i + b \sum_i P(x_i) \\ &= a\bar{x} + b \end{aligned}$$

The expected values of higher powers of x can also be useful.

For example,

$$E(x^2) \equiv \sum_i P(x_i)x_i^2$$

or more generally,

$$E(x^k) \equiv \sum_i P(x_i) x_i^k$$

This is the k^{th} moment of the distribution

It is often more useful to study the deviations from the mean value:

$$\tilde{x} \equiv x - \bar{x}$$

The moments of these deviations are called **central moments**.

$$\begin{aligned} E(\tilde{x}) &= \sum_i P(x_i) \tilde{x}_i = \sum_i P(x_i) (x_i - \bar{x}) = \sum_i P(x_i) x_i - \sum_i P(x_i) \bar{x} \\ &= \bar{x} - \bar{x} = 0 \end{aligned}$$

The second central moment is also known as the **variance**

$$\begin{aligned} V(x) &\equiv E[\tilde{x}^2] \\ &= E[(x - \bar{x})^2] = E[x^2 - 2x\bar{x} + \bar{x}^2] \\ &= E[x^2] - 2\bar{x}E[x] + \bar{x}^2 \\ &= E[x^2] - \bar{x}^2 \end{aligned}$$

We can read this as: “The variance is equal to expected value of the square minus the square of the expected value.”

For a multivariate distribution (e.g., the distribution of arm lengths, x , and leg lengths, y), there is another central moment, the **covariance**

$$\begin{aligned} C(x, y) &\equiv E[\tilde{x}\tilde{y}] \\ &= E[(x - \bar{x})(y - \bar{y})] = E[xy - x\bar{y} - y\bar{x} + \bar{x}\bar{y}] \\ &= E[xy] - \bar{x}E[y] - \bar{y}E[x] + \bar{x}\bar{y} \\ &= E[xy] - \bar{x}\bar{y} \end{aligned}$$

Note that the covariance of x with itself is the variance, i.e., $C(x, x) = V(x)$.

Often we are interested in the properties of functions of random variables. For example, the expected value of the product $z = x * y$

Higher central moments are calculated analogously.

Q1.7) If $y = ax + b$, where a and b are constants, and the variance of x is $V(x)$, prove that the variance of y is $a^2V(x)$.

Q1.8) Let $z = x + y$ where x and y are random variables. What is the variance of z ? It is often very helpful to redefine random variables as deviations from their mean, e.g., use the relationship $x = \tilde{x} + \bar{x}$

Q1.9) Assuming b is a constant, show that $C(x, b) = 0$.

d) Bernoulli, binomial, and multinomial distributions

A Bernoulli variable x is 1 (success) with probability p and is 0 (failure) with probability $q = 1 - p$.

A Bernoulli variable could describe the outcome of a coin toss (heads or tails) or choosing a random allele (A or a).

Q1.10) Prove that $E(x) = p$ and that $V(x) = pq$.

A binomial variable is the sum of n independent Bernoulli events. For example, the number of heads in 10 coin tosses or the number of A alleles in a random sample of 30 gametes. What is the probability of getting y successes in n trials?

$$P(Y = y) = \binom{n}{y} p^y q^{n-y}$$

Q1.11) Explain why the equation above for $P(Y = y)$ makes sense. Hint: Think of it as two parts: the binomial coefficient as one part and the $p^y q^{n-y}$ as the other part.

Q1.12) Show that $E(y) = np$ and that $V(y) = npq$. Hint: Recall that the binomial is the sum of n Bernoulli variables.

We have been assuming that each event can be described as being either a success (1) or a failure (0). The multinomial is similar to the binomial except that each underlying event can have one of several types of outcomes. For example, imagine that there are k different types of alleles A_1, A_2, \dots, A_k that occur with frequencies p_1, p_2, \dots, p_k , respectively. In a sample of n random gametes, the probability of sampling these alleles in abundances y_1, y_2, \dots, y_k is given by

$$P(y_1, y_2, \dots, y_k) = \frac{n!}{\prod_i y_i!} \left(\prod_i p_i^{y_i} \right) \text{ where } \prod_i x_i = x_1 * x_2 * \dots * x_k$$

This equation describes the multinomial distribution.

For the multinomial distribution, $E(y_i) = np_i$, $V(y_i) = np_i q_i$, and $C(y_i, y_j) = -np_i p_j$ where $q_i = 1 - p_i$.