FUN FACTS ABOUT DISCRETE RANDOM VARIABLES AND LOGS

RANDOM VARIABLES. A random variable is a variable whose value is determined by random chance (often denoted by a capital letter). A discrete random variable can assume one of only a discrete set of values (e.g., \( X = \) "the outcome of the roll of a die" or \( Y = \) "the number of \( A_1 \) alleles found at locus A in a randomly chosen individual"). In contrast, a continuous random variable can assume any one of a continuous set of values (e.g., \( Z = \) "body weight of a randomly chosen individual"). An event is an outcome of a random variable at a trial (e.g., the event "\( X = 6 \)" in the die example).

For now, we'll consider only discrete random variables:

PROBABILITY, FREQUENCY, and PROBABILITY DISTRIBUTION. The probability of an event is simply the fraction of times it is expected to occur among a set of trials. The frequency of an event is the fraction of times the event is observed to occur among a set of trials. These two terms become interchangeable when considering an extremely large number of trials. The probability of event \( A \) is often denoted \( \Pr(A) \). A probability distribution is the set of probabilities associated with the outcomes of a random variable. For example, \{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\} is the probability distribution associated with the random variable \( X \) above.

INDEPENDENCE. Two events \( A \) and \( B \) are defined as being independent if the probability that both occur is equal to the product of the probabilities that each occurs separately. That is,

\[
\Pr(A \text{ and } B) = \Pr(A) \cdot \Pr(B).
\]

EXPECTATION or MEAN. The expectation (or, roughly speaking, the mean) of a random variable \( X \) is given by the equation

\[
E(X) = \overline{X} = \sum_{i=1}^{\infty} f_i X_i,
\]

where \( X_1, X_2, \ldots \) are the different values that \( X \) can take, and \( f_i \) is the probability that \( X \) takes value \( X_i \). Important properties of expectations include the following. If \( a \) is a constant, then

\[
E(aX) = aE(X).
\]

If \( Y \) is another random variable, then

\[
E(X + Y) = E(X) + E(Y).
\]

VARIANCE. The variance of a random variable is defined as

\[
\text{Var}(X) = E[(X - \overline{X})^2] = \sum_{i=1}^{\infty} f_i (X_i - \overline{X})^2.
\]

Another, computationally convenient way to write this is:

\[
\text{Var}(X) = E(X^2) - [E(X)]^2 = \overline{X^2} - (\overline{X})^2 = \sum_{i=1}^{\infty} f_i X_i^2 - \left[ \sum_{i=1}^{\infty} f_i X_i \right]^2.
\]
Variances are never negative; they range from 0 to $+\infty$. The concept of variance is fundamental to evolution in general and population genetics in particular. (In fact, the mathematical concept of variance first appeared in the first modern paper on population genetics, published by R. A. Fisher in 1918!) Sometimes we'll write $\sigma^2_x$ for $\text{Var}(X)$.

**COVARIANCE, INDEPENDENCE, and CORRELATION:** The covariance of two random variables is

$$\text{Cov}(X, Y) = E[(X - \overline{X})(Y - \overline{Y})] = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} h_{ij}(X_i - \overline{X})(Y_j - \overline{Y}),$$

where $h_{ij} = \Pr(X = X_i$ and $Y = Y_j)$. An equivalent formula for the covariance is

$$\text{Cov}(X, Y) = E[XY] - \overline{X} \cdot \overline{Y} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} h_{ij}X_i Y_j - \left(\sum_{i=1}^{\infty} f_i X_i\right) \left(\sum_{j=1}^{\infty} g_j Y_j\right) = \overline{X} \cdot \overline{Y} - \overline{X} \cdot \overline{Y}.$$

where $f_i = \sum_{j=1}^{\infty} h_{ij} = \Pr(X = X_i)$ and $\Pr(Y = Y_j) = g_j = \sum_{i=1}^{\infty} h_{ij}$. Covariances, unlike variances, can be positive or negative and range from $-\infty$ to $+\infty$. Sometimes we'll write $\sigma_{xy}$ for $\text{Cov}(X, Y)$. Notice that the variance of a random variable is the same thing as its covariance with itself.

An important fact is that the sum of two random variables $X$ and $Y$ has the variance

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y).$$

If two random variables $X$ and $Y$ are independent, then $\text{Cov}(X, Y) = 0$ and so $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. (Caution: a zero covariance between two random variables does not necessarily imply the converse, that they are independent.)

The covariance of $X$ and $Y$ can be rescaled relative to their variances. This produces a measure called the correlation that ranges from $-1$ to $+1$. The correlation between $X$ and $Y$ is

$$r_{xy} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Notice that the sign of the correlation is always the same as the sign of the covariance. Also notice that the correlation does not depend on the units of $X$ and $Y$ (i.e., it is "dimensionless").

**LOGARITHMS.** We'll run into some logarithms. These will always be "natural" logs, i.e. logarithms with base $e (= 2.7)$. When the number $e$ is raised to the $x$ power, it can be written $e^x$ or $\exp[x]$; these mean the same thing. Some of the basic laws of logarithms are:

\[
\begin{align*}
\ln(e^a) &= a \\
\ln(ab) &= \ln(a) + \ln(b) \\
\ln(k^a) &= a \ln(k)
\end{align*}
\]

$\ln(e) = 1$ and $\ln(1) = 0$

[Don't take my word for it, prove these for yourself as a "warm-up" exercise! Hint: write $a = e^x$ and $b = e^y$ so that $x = \ln(a)$ and $y = \ln(b).$]

**WARNING!!!:** $\ln(a + b) \neq \ln(a) + \ln(b)$