Multinomial Distribution

For a SNP with alleles A, a the three genotypes and their probabilities are

Genotype	Probability
AA	P_{AA}
Aa or aA	P_{Aa}
aa	P_{aa}

The multinomial distribution gives the probability of x of AA, y of Aa and z of aa. The probability of x genotypes AA is $(P_{AA})^x$, etc. The numbers of ways of ordering x, y, z occurrences of the three outcomes is $n!/(x!y!z!)$ where $n = x + y + z$.

The multinomial probability is:

$$\Pr(x, y, z) \quad = \quad \frac{n!}{x!y!z!}(P_{AA})^x(P_{Aa})^y(P_{aa})^z$$

Multinomial Variances and Covariances

If $\{P_i\}$ are the probabilities for a series of categories, the sample proportions $\tilde{P}_i$ from a sample of n observations have these properties:

$$
\begin{aligned}
\mathcal{E}(\tilde{P}_i) &= P_i \\
\text{Var}(\tilde{P}_i) &= \frac{1}{n}P_i(1-P_i) \\
\text{Cov}(\tilde{P}_i,\tilde{P}_j) &= -\frac{1}{n}P_iP_j,\ \ i\neq j
\end{aligned}
$$

The covariance is defined as $\mathcal{E}[(\tilde{P}_i - P_i)(\tilde{P}_j - P_j)]$.

For the sample counts:

$$
\begin{aligned}
\mathcal{E}(n_i) &= nP_i \\
\text{Var}(n_i) &= nP_i(1-P_i) \\
\text{Cov}(n_i,n_j) &= -nP_iP_j,\ \ i\neq j
\end{aligned}
$$

Allele Frequency Sampling Distribution

If a locus has alleles A and a, in a sample of size n the allele counts are sums of genotype counts:

$$\begin{aligned} n &= n_{AA} + n_{Aa} + n_{aa} \\ n_A &= 2n_{AA} + n_{Aa} \\ n_a &= 2n_{aa} + n_{Aa} \\ 2n &= n_A + n_a \end{aligned}$$

Genotype counts in a random sample are multinomially distributed. What about allele counts? Approach this question by calculating variance of n_A.

Within-population Variance

$$\begin{aligned}
\text{Var}(n_A) &= \text{Var}(2n_{AA} + n_{Aa}) \\
&= \text{Var}(2n_{AA}) + 2\text{Cov}(2n_{AA}, n_{Aa}) + \text{Var}(n_{Aa}) \\
&= 4nP_{AA}(1 - P_{AA}) - 4nP_{AA}P_{Aa} + nP_{Aa}(1 - P_{Aa}) \\
&= 2np_A(1 - p_A) + 2n(P_{AA} - p_A^2)
\end{aligned}$$

This is not the same as the binomial variance $2np_A(1-p_A)$ unless $P_{AA} = p_A^2$. In general, the allele frequency distribution is not binomial.

The variance of the sample allele frequency $\tilde{p}_A = n_A/(2n)$ can be written as

$$\text{Var}(\tilde{p}_A) = \frac{p_A(1 - p_A)}{2n} + \frac{P_{AA} - p_A^2}{2n}$$

Within-population Variance

It is convenient to reparameterize genotype frequencies with the within-population *inbreeding coefficient* f:

$$\begin{aligned} P_{AA} &= p_A^2 + f p_A p_a \\ P_{Aa} &= 2p_A p_a - 2f p_A p_a \\ P_{aa} &= p_a^2 + f p_A p_a \end{aligned}$$

Then the variance can be written as

$$\mathrm{Var}(\tilde{p}_A) = \frac{p_A(1-p_A)(1+f)}{2n}$$

This variance is different from the binomial variance of $p_A(1-p_A)/2n$.

Bounds on f

Since

$$\begin{aligned} p_A \geq P_{AA} &= p_A^2 + f p_A(1 - p_A) \geq 0 \\ p_a \geq P_{aa} &= p_a^2 + f p_a(1 - p_a) \geq 0 \end{aligned}$$

there are bounds on f:

$$\begin{aligned} -p_A/(1 - p_A) \leq & \ f \ \leq 1 \\ -p_a/(1 - p_a) \leq & \ f \ \leq 1 \end{aligned}$$

or

$$\max\left(-\frac{p_A}{p_a}, -\frac{p_a}{p_A}\right) \leq \ f \ \leq 1$$

This range of values is [-1,1] when $p_A = p_a$.

An aside: Indicator Variables

A very convenient way to derive many statistical genetic results is to define an indicator variable x_{ij} for allele j in individual i:

$$x_{ij} = \begin{cases} 1 & \text{if allele is } A \\ 0 & \text{if allele is not } A \end{cases}$$

Then

$$\begin{aligned} \mathcal{E}(x_{ij}) &= p_A \\ \mathcal{E}(x_{ij}^2) &= p_A \\ \mathcal{E}(x_{ij}x_{ij'}) &= P_{AA} \end{aligned}$$

If there is random sampling, individuals are independent, and

$$\mathcal{E}(x_{ij}x_{i'j'}) = \mathcal{E}(x_{ij})\mathcal{E}(x_{i'j'}) = p_A^2$$

These expectations are the averages of values from many samples from the same population.

An aside: Intraclass Correlation

The inbreeding coefficient is the correlation of the indicator variables for the two alleles j, j' at a locus carried by an individual i. This is because:

$$\begin{aligned}\mathsf{Var}(x_{ij}) &= \mathcal{E}(x_{ij}^2) - [\mathcal{E}(x_{ij})]^2 \\ &= p_A(1-p_A) \\ &= \mathsf{Var}(x_{ij'}),\ \ j \neq j'\end{aligned}$$

and

$$\begin{aligned}\mathsf{Cov}(x_{ij}, x_{ij'}) &= \mathcal{E}(x_{ij}x_{ij'}) - [\mathcal{E}(x_{ij})][\mathcal{E}(x_{ij'})],\ \ j \neq j' \\ &= P_{AA} - p_A^2 \\ &= f p_A(1-p_A)\end{aligned}$$

so

$$\mathsf{Corr}(x_{ij}, x_{ij'}) = \frac{\mathsf{Cov}(x_{ij}, x_{ij'})}{\sqrt{\mathsf{Var}(x_{ij})\mathsf{Var}(x_{ij'})}} = f$$

Allele Dosage

The dosage X of allele A for an individual is the number of copies of A (0,1,2) that individual carries (the sum of its two allele indicators).

The probabilities for X are

$$\Pr(X = 0) = P_{aa}, \Pr(X = 1) = P_{Aa}, \Pr(X = 2) = P_{AA}$$

so the expected value of X is $2P_{AA} + P_{Aa} = 2p_A$.

The expected value of X^2 is $4P_{AA} + P_{Aa} = 2(p_A + P_{AA})$ and this leads to a variance the dosage for an individual of

$$\text{Var}(X) \;=\; 2P_{AA} + 2p_a - 4p_A^2 = 2p_A(1 - p_A)(1 + f)$$

We will come back to this result, but note here that the f term is usually not included in genetic data analysis packages.