Genotypic Frequency

X Xia, University of Ottawa, Ottawa, ON, Canada
CR Primmer, University of Turku, Turku, Finland

This article is a revision of the previous edition article by A Clark, volume 2, p 873, © 2001, Elsevier Inc.

Glossary

Cladogenic event An event in which a gene or genomic lineage diverges into two or more lineages.
Parameter estimation Statistical protocol for obtaining the point and interval estimate of the parameter.
Population parameter A quantity of interest in a population to be estimated from a sample taken from the population.

The genotypic frequency is the frequency of a particular genotype in a population. The conceptual definition of a genotype is the genetic makeup of a cell, an organism, or an individual. As individuals are almost always unique except those from clonal reproduction, the conceptual definition would imply that all genotypic frequencies would be $1/N$, where N is the population size. In population genetics, a genotype is often defined for one or a few loci in diploid species, so that there are three genotypes (AA, Aa, and aa) for a single locus with two different alleles (A and a). With two loci each with two alleles, we would have nine different genotypes, AABB, AaBB, aaBB, …, aabb.

Genotypic frequencies are most frequently used to compute allele frequencies and the inbreeding coefficient which is the basis for other F statistics in population genetics. For a single locus with two alleles (A and a), and the counts of the three genotypes (N_{AA}, N_{Aa}, and N_{aa}) being 30, 50, and 20, respectively, we have genotypic frequencies $P_{AA} = 0.3$, $P_{Aa} = 0.5$, and $P_{aa} = 0.2$. The allele frequencies p_A and p_a are then

$$p_A = \frac{2N_{AA} + N_{Aa}}{2(N_{AA} + N_{Aa} + N_{aa})} = 0.55;\ p_a = 1 - p_A = 0.45 \quad [1]$$

The conversion from genotypic frequencies to allele frequencies leads to significant loss of information which is often measured by Shannon entropy (H). The information contained in the genotypic frequencies and allele frequencies, designated by $H_{genotype}$ and H_{allele}, respectively, are

$$\begin{aligned} H_{genotype} &= -(P_{AA}\log_2(P_{AA}) + P_{Aa}\log_2(P_{Aa}) + P_{aa}\log_2(P_{aa})) \\ &= 1.4855\ (\text{bits}) \\ H_{allele} &= -(p_A\log_2(p_A) + p_a\log_2(p_a)) = 0.9928\ (\text{bits}) \end{aligned} \quad [2]$$

If the allele frequencies are equal, then the information in the allele frequencies will be exactly 1 bit and the information in the expected genotypic frequencies assuming the Hardy–Weinberg equilibrium (i.e., $P_{AA} = 0.25$, $P_{Aa} = 0.5$, and $P_{aa} = 0.25$) will be exactly 1.5 bits. The loss of information during the conversion from genotypic frequencies to allele frequencies implies that genotypic frequencies cannot be recovered from allele frequencies. This has implications for us to choose parameter estimators. When a population parameter (e.g., genetic distance between two populations) can be estimated by either genotypic frequencies or allele frequencies, the estimation based on the genotypic frequencies is always more preferable.

When two loci each with two alleles are completely linked, they are equivalent to a single locus with two alleles, with only three genotypes. If the two loci are not linked, then there would be nine genotypes and the information in the genotypic frequencies will be the summation of information in the genotypic frequencies for each locus. The information is equivalent to genetic variance, of which the additive component is directly related to directional selection. This is why increased recombination, by increasing information, can lead to more efficient selection.

Genotypic frequencies are also frequently used for estimating the inbreeding coefficient and the associated F statistics. The inbreeding coefficient (F), in the case of one locus with two alleles, is defined by

$$\begin{aligned} P_{AA} &= p_A^2(1-F) + p_A F \\ P_{Aa} &= 2p_A p_a(1-F) \\ P_{aa} &= p_a^2(1-F) + p_a F \end{aligned} \quad [3]$$

The log-likelihood function for estimating p_A and F is therefore

$$\ln L = \ln\left(P_{AA}^{N_{AA}} P_{Aa}^{N_{Aa}} P_{aa}^{N_{aa}}\right) \quad [4]$$

Taking the partial derivatives of $\ln L$ with respect to p_A and F, setting the partial derivatives to zero and solving the two resulting simultaneous equations, we obtain p_A and F as functions of genotypic frequencies:

$$\begin{aligned} p_A &= \frac{2N_{AA} + N_{Aa}}{2(N_{AA} + N_{Aa} + N_{aa})} \\ F &= \frac{4N_{AA}N_{aa} - N_{Aa}^2}{2N_{AA}N_{Aa} + 4N_{AA}N_{aa} + N_{Aa}^2 + 2N_{Aa}N_{aa}} \end{aligned} \quad [5]$$

The maximum likelihood approach is particularly useful when computing an estimate of F from multiple loci because the likelihood function for multiple loci is just the product of the likelihood functions for individual loci. Suppose we have locus 1 with $N_{AA} = 1469$, $N_{Aa} = 138$, $N_{aa} = 5$, and locus 2 with $N_{BB} = 100$, $N_{Bb} = 60$, $N_{bb} = 5$. The F value, when estimated separately for each locus, would be 0.0227 for locus 1 and −0.0879 for locus 2. The likelihood estimate of F from the two loci is 0.0108, which is closer to the estimate from locus 1 because locus 1 has more data and therefore contributes more to the final estimate than locus 2.

One difficulty with genotype characterization is the presence of hidden alleles. Suppose a locus with three alleles A, B, and C, with alleles A and B producing distinct protein bands (electromorphs) in electrophoresis and allele C producing no band. This will lead to scoring AC and BC heterozygotes as AA and BB homozygotes. When the frequency of allele C is low, the sample may contain no CC individual. Even if a few CC individuals have been sampled, the lack of bands for these CC individuals may be attributed to experimental errors instead of alerting us to the presence of hidden alleles.

Theory of statistical estimation can help us detect the presence of hidden alleles and estimate their frequencies. Here, we illustrate a simple case involving the three-allele system mentioned above by assuming the Hardy–Weinberg equilibrium. We have a three-allele hypothesis (H_3) and a two-allele hypothesis (H_2) and need to decide which one describes the data better. Suppose we obtain 30 individuals with a single A band, 30 with a single B band, and 40 with both A and B bands ($N_{A?} = 30$, $N_{B?} = 30$, $N_{AB} = 40$). With H_2, which assumes $N_{A?} = N_{AA}$ and $N_{B?} = N_{BB}$, the proportion of AA, AB, and BB genotypes is p_A^2, $2p_Ap_B$, and p_B^2, respectively. So the log-likelihood function for estimating p_A is

$$\ln L_{H_2} = N_{AA} \ln (p_A^2) + N_{AB} \ln (2p_A p_B) + N_{BB} \ln (p_B^2) \quad [6]$$

which will lead to $p_A = 0.5$ and $\ln L_{H_2} = -110.903\,55$. We do not need to estimate p_a because $p_a = 1 - p_A$.

With H_3, we need to estimate p_A and p_B. The log-likelihood function is

$$\ln L_{H_3} = N_{A?} \ln (p_A^2 + 2p_A p_C) + N_{AB} \ln (2p_A p_B) + N_{B?} \ln (p_B^2 + 2p_B p_C) \quad [7]$$

which will lead to $p_A = p_B = 0.4667$ and $\ln L_{H_3} = -109.623\,26$.

We can use the likelihood ratio test to evaluate which of the two hypotheses are significantly better than the other. With the large sample approximation, $\chi^2 = 2(\ln L_{H_3} - \ln L_{H_2})$ follows the chi-square distribution with Δp degrees of freedom (DF, where Δp is the difference in the number of parameters between the two hypotheses, and is 1 in our case as H_3 has one more allele frequency to estimate than H_2). With $\chi^2 = 2.5606$ and DF = 1, $p = 0.1096$, and we cannot reject H_2 in favor of H_3 at the 0.05 level.

An alternative to significance test in model selection is to use the information theoretic indices such as Akaike information criterion (AIC) defined as

$$\mathrm{AIC} = -2 \ln L + 2N_p \quad [8]$$

where N_p is the number of parameters in the model, being 1 in H_2 and 2 in H_3. The smaller the AIC value, the better the model. With AIC equals 223.8071 for H_2 and 223.2465 for H_3, the criterion slightly favors H_3.

One may also use the least-squares method for parameter estimation and model selection. The residual sum of squares (RSS) for H_2 and H_3 are, respectively,

$$\begin{aligned} \mathrm{RSS}_{H_2} &= (N_{AA} - p_A^2 N)^2 + (N_{AB} - 2p_A p_B N)^2 + (N_{BB} - p_B^2 N)^2 \\ \mathrm{RSS}_{H_2} &= [N_{A?} - (p_A^2 + 2p_A p_C)N]^2 + (N_{AB} - 2p_A p_B N)^2 \\ &\quad + [N_{B?} - (p_B^2 + 2p_B p_C)N]^2 \\ \text{where } N &= N_{A?} + N_{AB} + N_{B?} \end{aligned} \quad [9]$$

To obtain p_A with H_2, we take the derivative of RSS_{H_2} with respect to p_A, set the derivative to zero, and solve for p_A. This results in $p_A = 0.5$ and $\mathrm{RSS}_{H_2} = 150$. To obtain p_A and p_B with H_3, we take partial derivatives of RSS_{H_3} with respect to p_A and p_B, set the partial derivatives to zero, and solve for p_A and p_B from the two resulting simultaneous equations. This leads to $p_A = p_B = 0.4454$ and $\mathrm{RSS}_{H_3} = 0.4821$. Thus, H_3 fits the date much better than H_2. The information theoretic indices can also be used with RSS by a relationship between RSS and likelihood, with the result favoring H_3.

Differences in genotypic frequencies among different populations can be used to study genetic divergence, calculate genetic distances, build phylogenetic trees, and date cladogenic events using the genetic distances.

See also: Allele Frequency; Alleles; Genotype; Hardy–Weinberg Law; Least Squares; Locus; Maximum Likelihood.

Further Reading

Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723.

Burnham KP and Anderson DR (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer.

Xia X (2009) Information-theoretic indices and an approximate significance test for testing the molecular clock hypothesis with genetic distances. *Molecular Phylogenetics & Evolution* 52: 665–676.

Xia X and Yang Q (2011) A distance-based Least-square method for dating speciation events. *Molecular Phylogenetics & Evolution* 28: 1827–1834.