Contents

Part II

Frequency Estimation and HWE

1 Characterizing a Population

1.1 Fundamentals

Fundamental rules of genetics

Modeling the genetics of populations starts with the fundamental and actually quite simple and solid rules of genetic transmission. We will pose these laws as probability rules.

1. Law of Segregation A diploid parent is equally likely to pass along either of its two alleles to its offspring.

$$P(\text{pass copy } 1) = P(\text{pass copy } 2) = \frac{1}{2}$$

2. Law of Random Union Gametes unite randomly. So, for example, allele A_1 is no more likely to unite with allele A_1 than A_2.

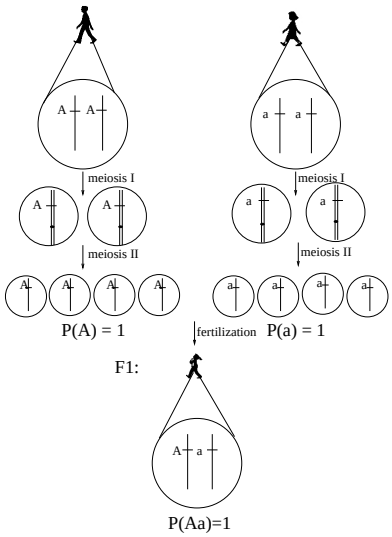
$$\begin{aligned} P(\text{offspring is } A_1A_1) &= P(\text{father passes } A_1) \times P(\text{mother passes } A_1) \\ P(\text{offspring is } A_1A_2) &= P(\text{father passes } A_1) \times P(\text{mother passes } A_2) \\ &\quad + P(\text{mother passes } A_1) \times P(\text{father passes } A_2) \end{aligned}$$

(3.) Law of Independent Assortment Mendel's second law was partially disproven, and we will discuss it in more detail later when we consider multiple loci.

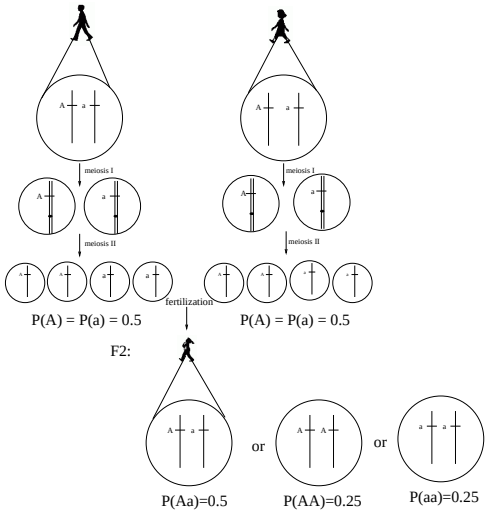
We can visualize these laws by observing what happens during particular kinds of controlled *crosses*. Crosses of *inbred* lines to make *hybrid* individuals are common strategy in plant breed. This kind of cross takes

homozygotes and makes heterozygotes, as shown below as we follow a cross of inbreds through two generations, F1 and F2. We will see later in this class why *inbreeding* leads to homozygous individuals and why the resulting heterozygotes tend to have desirable traits.

Segregation & Random Union (F1)



Segregation & Random Union (F2)



Genetic Composition of a Population

Starting from these basic rules, we are ready to begin the study of population genetics. I have said repeatedly that population genetics is the study of genetic diversity in populations, so clearly we must begin the process of characterizing the diversity that exists in population. How do we summarize genetic diversity in a population?

- Number of alleles at a locus.
- The frequency of alleles at a locus.

- The frequency of genotypes at a locus.

The frequency of genotypes provides additional information about the population. To illustrate this point, consider the following two populations that have the same allele frequencies, but *very* different genotype frequencies.

	A_1A_1	A_1A_2	A_2A_2
Population 1	50	0	50
Population 2	25	50	25

Notation for Population Summaries

Diallelic locus: Imagine a locus A with two possible alleles: A_1 and A_2

Muliallelic locus: A locus B with alleles B_k for $i = 1, 2, \ldots, K$.

- **Parameters**: properties of the population that can never actually be observed.
  - Population size: N
  - Population frequency of genotype at a locus:
    - $P_{A_1A_1}, P_{A_1A_2}, P_{B_1B_5}$, etc.
    - Or P_{11}, P_{12}, etc. when the locus is assumed.
  - Population frequency of allele at a locus:
    - $p_{A_1}, p_{A_2}, p_{B_k}$, etc.
    - Or p_1, p_2, p_k when the locus is assumed.
- Note the relationship between genotype and allele frequencies:

$$p_u = P_{uu} + \sum_{u<v} \frac{1}{2} P_{uv}$$

I will use this equation to illustrate the use of the basic probability rules. Consider the *probabilistic experiment* of drawing a random allele from the poulation. The probability you draw A_u is p_u. Drawing a random allele from the population can be reduced to a two part experiment. First you draw a random diploid individual. Then you draw one of their two alleles at random. Consider the list of all possible genotypes, e.g. $\{A_uA_u, A_uA_v, \ldots, A_vA_v, A_vA_w, \ldots\}$: they are mutually exclusive and exhaustive outcomes for the first part of the experiment, drawing a random diploid individual. Conditional on the genotype, the probability of drawing allele u is 1, if the genotype is A_uA_u, $\frac{1}{2}$ if the genotype is A_uA_v for $v \neq u$, and 0 otherwise (we have just defined a conditional pmf). Therefore, the allele frequency follows from the LTP version using conditional probabilities:

$$p_u = P(\text{draw } u \mid A_uA_u)P(A_uA_u) + \sum_{u<v} P(\text{draw } u \mid A_uA_v)P(A_uA_v)$$

and the formula follows. If this is not clear as a bell, you need more practice in applying the basic rules of probability.

1.2 Hardy Weinberg Equilibrium

Hardy Weinberg Equilibrium HWE - History

- G. H. Hardy, a mathematician, who wanted to counter the suggestion that any dominant trait should rise to a proportion of 75%.

Under what type of cross would you expect 75% dominant? The F2 generation has $75\% = 50\% + 25\%$ with dominant trait.

- W. Weinberg, an obstetrician, wanted to know if bearing twins was a Mendelian trait.
- Note, **evolution** is the change in allele frequencies in a population over time. Under what conditions does evolution *not* occur? What is the *zero force* law for population genetics? Hardy and Weinberg derived this law, and it is probably the first major result every population geneticist learns.

Haploid Population

We will begin our quest for the zero force law by looking for the conditions that lead to constant allele frequencies in an asexual haploid population. To reproduce, the individuals in this population simply "clone" themselves.

- Suppose a population consists of two types of individuals (e.g. green, yellow).
- Suppose all individuals in the population reproduce simultaneously.
- Let $N_1(t)$ and $N_2(t)$ be the counts of each type of individual at generation t.
- Then, $p_1(t) = \frac{N_1(t)}{N_1(t)+N_2(t)}$ is the population allele frequency of of allele 1 at generation t.
- Assume each individual in generation t has exactly W_t offspring. (Note: even with environmental fluctuation, if the number of offspring per individual are iid random variables, the law of large numbers implies that an average W_t offspring will be produced per individual, per generation and the result is the same.)
- What does the population look like in generation $t + 1$?

Change in Allele Frequency in One Generation

A linear recurrence equation for counts across generations:

$$\begin{aligned} N_1(t+1) &= W_t N_1(t) \\ N_2(t+1) &= W_t N_2(t) \end{aligned}$$

To see if the allele frequency is changing (evolution), consider the allele frequency in the next generation

$$p_1(t+1) = \frac{N_1(t+1)}{N_1(t+1)+N_2(t+1)} = \frac{W_t N_1(t)}{W_t N_1(t)+W_t N_2(t)} = \frac{N_1(t)}{N_1(t)+N_2(t)} = p_1(t)$$

The results can be generalized to populations consisting of k different types of individuals. The fundamental assumptions have been:

- All individual types produce the same W_t offspring at the tth generation or the population is large enough so that environmental fluctuations average out.
- There is no mutation during offspring production.
- The population is closed (no imigration or emigration from other populations).

Linear recurrence relation [footnote]

- A *linear recurrence relation* on a sequence of numbers $N(1), N(2), \ldots, N(t), \ldots$ expresses $N(t)$ as a first-degree polynomial of $N(k)$ with $k < t$.

$$N(t) = AN(t-1) + BN(t-2) + CN(t-3) + \cdots$$

- A *first-order* linear recurrent relation involves only the preceding number in the sequence:

$$N(t) = AN(t-1) + B$$

- Given an *initial condition* $N(0) = N_0$ and $A \neq 1$, there is a unique solution to the first-order linear recurrence relation.

$$N(t) = \left(N_0 + \frac{B}{A-1}\right) A^t - \frac{B}{A-1}$$

Proof of linear recurrence relation solution [footnote]

By induction, show that it is true for $t = 1$:

$$\begin{aligned} N(1) &= \left(N_0 + \frac{B}{A-1}\right) A^1 - \frac{B}{A-1} \\ &= N_0 A + \frac{B}{A-1} A - 1 \\ &= N_0 A + B. \end{aligned}$$

Then suppose the solution is

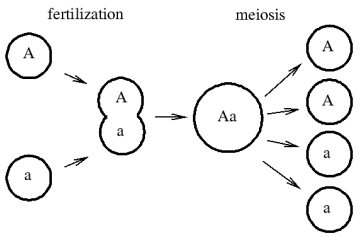
$$N(t) = \left(N_0 + \frac{B}{A-1}\right) A^t - \frac{B}{A-1},$$

and show that $N(t+1)$ satisfies the desired equation.

$$\begin{aligned} N(t+1) &= AN(t) + B \\ &= A\left[\left(N_0 + \frac{B}{A-1}\right) A^t - \frac{B}{A-1}\right] + B \\ &= \left(N_0 + \frac{B}{A-1}\right) A^{t+1} - \frac{AB}{A-1} + \frac{B(A-1)}{A-1} \\ &= \left(N_0 + \frac{B}{A-1}\right) A^{t+1} - \frac{B}{A-1}. \end{aligned}$$

Sexually Reproducing Diploid Population

Now we consider the extra complications of diploid populations. We will start with the allele frequencies (we can think of this as the allele frequencies in the pool of all gametes produced by the adults) in one generation and follow these alleles through fertilization, maturation to adulthood, and until gametes are produced again, through meisosis, in the next generation.



- Suppose the population consists of two genotypes A_1 and A_2.
- Let $p_1(t)$ be the proportion of A_1 gametes produced by this tth generation. We'll also let the genotype frequencies at generation t be $P_{11}(t)$, $P_{12}(t)$, and $P_{22}(t)$.
- **Goal:** Prove that under certain conditions the allele and genotype frequencies are unchanging generation after generation.

Following population through one generation

Assuming...

- identical allele frequencies in both sexes
- random mating
- infinite population size
- no viability differences

Generation t Allele Frequencies				**Generation $t+1$ Genotype Frequencies**	
Maternal		**Paternal**			
Gamete	**Prob.**	**Gamete**	**Prob.**	**Offspring**	**Probability**
A_1	$p_1(t)$	A_1	$p_1(t)$	A_1A_1	$P_{11}(t+1) = p_1(t)p_1(t)$
A_2	$1-p_1(t)$	A_2	$1-p_1(t)$	A_2A_2	$P_{22}(t+1) = [1-p_1(t)][1-p_1(t)]$
A_1	$p_1(t)$	A_2	$1-p_1(t)$	A_1A_2	$P_{12}(t+1) = p_1(t)[1-p_1(t)]$
A_2	$1-p_1(t)$	A_1	$p_1(t)$	A_2A_1	$P_{21}(t+1) = [1-p_1(t)]p_1(t)$

In the left side of the table, we consider all possible alleles that mother and father could contribute to the offspring and the probability of each type of allele. On the right, we consider the genotypes possible in the offspring and the resulting genotype probabilities. But, we don't distinguish the two heterozygotes, so

Generation $t+1$ Genotype Probabilities	
A_1A_1	$P_{11}(t+1) = p_1^2(t)$
A_1A_2	$P_{12}(t+1) = 2p_1(t)[1-p_1(t)]$
A_2A_2	$P_{22}(t+1) = [1-p_1(t)]^2$

Further assuming...

- equal fertility
- no mutation
- infinite populations,

each offspring produces gametes according to the following conditional probabilities:

		Conditional Gamete Probs.	
Genotype	**Probability (from above)**	$P(A \mid$ **genotype**$)$	$P(a \mid$ **genotype**$)$
A_1A_1	$P_{11}(t+1) = p_1^2(t)$	1	0
A_2A_2	$P_{22}(t+1) = [1-p_1(t)]^2$	0	1
A_1A_2	$P_{12}(t+1) = 2p_1(t)[1-p_1(t)]$	0.5	0.5

Applying the law of total probability with conditioning an all the possible genotypes producing alleles through meiosis, allele frequencies of generation $t+1$ are:

Genotype	Probability
A_1	$p_1(t+1) = 1 \times p_1^2(t) + 0.5 \times 2p_1(t)[1-p_1(t)] = p_1(t)$
A_2	$[1-p_1(t+1)] = 1 \times [1-p_1(t)]^2 + 0.5 \times 2p_1(t)[1-p_1(t)] = 1-p_1(t)$

We have proven that allele frequencies in the gametes are constant across generations. Furthermore, since genotype frequencies are a fixed function of allele frequencies from the preceding generation, they also cannot change across generations. The preceding constitutes a proof for the following theorem.

Hardy Weinberg Theorem

The **Hardy-Weinberg (HW) assumptions** are:

- No difference in genotype proportions between the sexes.
- Synchronous reproduction at discrete points in time (discrete generations).
- Infinite population size (so that small variabilities are erased in the average).
- No mutation.
- No migration (precisely no immigration and balanced emigration).
- No selection (precisely no differences in fertility and viability).
- Random mating.

Theorem (1908): Given all the assumptions above, then the allele and genotype frequencies are at Hardy-Weinberg equilibrium (HWE) (unchanging from generation to generation). If the frequencies are perturbed, they will return to equilibrium (not necessarily the same equilibrium) in a single generation.

Proof: We have just proven $p_1(t+1) = p_1(t)$, i.e. that allele frequencies do not change from generation to generation. Furthermore, $P_{11}(t+1) = p_1^2(t)$, $P_{22}(t+1) = [1-p_1(t)]^2$, and $P_{12}(t+1) = 2p_1(t)[1-p_1(t)]$.

One can also achieve the proof by starting from genotype frequencies in one generation and showing they are equivalent to the genotype frequencies in the following generation. This proof requires considering all the mating types and their probabilities, e.g. $A_1A_2 \times A_1A_2$ has probability $P_{12} \times P_{12}$ while $A_1A_1 \times A_1A_2$ has probability $2P_{11}P_{12}$.

Mating Table

A *mating table* is a useful construct for tracking how genotypes in one generation produce genotypes in the next generation. Since this approach of tracking the population from one generation to the next is used repeatedly in deriving population genetics results, I reproduce a sample mating table below.

The mating table shows two probability mass functions. The first is the probability of all possible mating types in the preceding generation. Below, those mating probabilities, shown in column 2, are derived assuming random mating. Second, the conditional probability mass function of the offspring genotypes given the mating type. Here, this conditional probability is derived from the Law of Segregation and Law of Random Union. Consideration of mutation during meiosis, unequal viability, and other effects would lead to a different conditional distribution.

		Offspring Genotype		
M = Mating Type	Probability	$P(A_1A_1 \mid M)$	$P(A_1A_2 \mid M)$	$P(A_2A_2 \mid M)$
$A_1A_1 \times A_1A_1$	P_{11}^2	1	0	0
$\times A_1A_2$	$2P_{11}P_{12}$	$\frac{1}{2}$	$\frac{1}{2}$	0
$\times A_2A_2$	$2P_{11}P_{22}$	0	1	0
$A_1A_2 \times A_1A_2$	P_{12}^2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$\times A_2A_2$	$2P_{12}P_{22}$	0	$\frac{1}{2}$	$\frac{1}{2}$
$A_2A_2 \times A_2A_2$	P_{22}^2	0	0	1

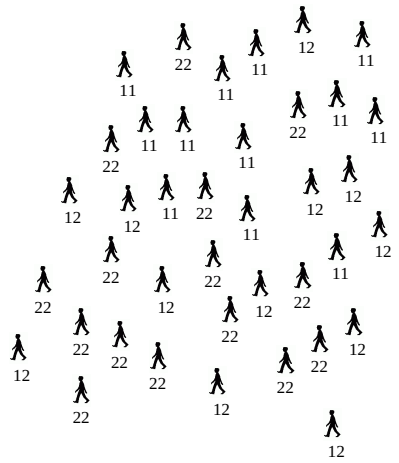
With the mating table in hand, recurrence equations for the genotype probabilities of the next generation given the genotype probabilities of the preceding generation can be derived using the law of total probability with conditioning on the mating types. For example,

$$P_{11}(t+1) = 1 \times P_{11}^2(t) + \frac{1}{2} \times 2P_{11}(t)P_{12}(t) + \frac{1}{4} \times P_{12}^2(t)$$

summing down the third column. Since allele probabilities are just functions of the genotype probabilities, recurrence relations for the allele probabilities are easily available.

Consider This "Population"

While this population is clearly not infinite, as required for HWE, we will use this population to demonstrate how genotype frequencies that are not at HWE achieve HWE in one generation when the randomness of segregation, union, and survival are removed. Infinite populations effectively remove this randomness by averaging random effects over countless individuals.



Population Genotype Frequencies

A count of genotypes leads to the population counts:

$$\begin{aligned} N_{11} &= 12 \\ N_{12} &= 12 \\ N_{22} &= 15 \\ N &= 39 \end{aligned}$$

implying the population genotype and allele frequencies:

$$\begin{aligned} P_{11} &= 0.31 \qquad & p_1 &= \frac{2 \times 12 + 12}{2 \times 39} = \frac{36}{78} \approx 0.46 \\ P_{12} &= 0.31 & & \\ P_{22} &= 0.38 \qquad & p_2 &= \frac{2 \times 15 + 12}{2 \times 39} = \frac{42}{78} \approx 0.54. \end{aligned}$$

Note, the HWE is not satisfied, for example $P_{11} \neq p_1^2$.

Next Generation

To remove randomness from this very finite population, let's suppose that this population produces an infinite pool of gametes in exactly the allele proportions expected from the calculations above. In a truly finite population, random events will allow some individuals to contribute more to the gamete pool and contributions will vary somewhat from the proportions above. In the next generation, when these alleles unite randomly, the genotype frequencies (before the population is whittled back down to finite size) will be:

$$\begin{aligned} P_{11}(1) &= p_1^2 = 0.46^2 = 0.21 \\ P_{12}(1) &= 2p_1p_2 = 2 \times 0.46 \times 0.54 = 0.50 \\ P_{22}(1) &= p_2^2 = 0.54^2 = 0.29 \\ \text{total} &= 1 \end{aligned}$$

The Hardy-Weinberg equations are satisfied and these adults will produce gametes with unchanging proportions p_1 and p_2 for the next generation.

Generalization to Multiple Alleles

Suppose there are $k > 2$ different alleles $A_1, A_2, \ldots, A_k$ with population frequencies $p_1, p_2, \ldots, p_k$. Then, upon HWE, the diploid genotype frequencies are:

$$\begin{aligned} P_{ii} &= p_i^2 \qquad \text{for } i = 1, 2, \ldots, k \\ P_{ij} &= 2p_ip_j \quad \text{for } i \neq j \in \{1, 2, \ldots, k\}. \end{aligned}$$

A very compact version of the proof distinguishes $P_{ij} = P_{ji}$ and recognizes that if the previous generation was a product of random mating, then $P_{ij} = p_ip_j$, so the allele frequency in the next generation is

$$\begin{aligned} p_i(t+1) &= P_{ii} + \frac{1}{2}\sum_{i \neq j} P_{ij} \\ &= \frac{1}{2}\sum_{j=1}^{n} 2p_ip_j \\ &= p_i \sum_{j=1}^{n} p_j = p_i \qquad \square \end{aligned}$$

Implications of HWE

- Under the appropriate conditions, genotype frequencies can be predicted from allele frequencies.
- Therefore, we need only track the allele frequencies when analyzing populations satisfying the assumptions.
- Mendelian reproduction does not favor one allele over another, hence there will be no loss of genetic variability from generation to generation.
- The dominant phenotype will not always make up 75% of the population. Indeed, only when $p_{A_1} = 0.5$.
- Evidence of HWE in a population does *not* imply the HWE assumptions are true.
- However, lack of HWE in a population implies that at least *one* of the HWE assumptions has been violated. This observation allows us to *detect* forces acting on populations and *estimate* their magnitude by checking the extent to which the equilibrium is disrupted.

Synchronous Reproduction

We will be spending a lot of time in this course relaxing the assumptions of HWE one at a time, sometimes two at a time, but we will not consider the assumption of synchronous reproduction. I will try to argue below that it doesn't really matter what we assume about reproduction timing, as long as the population has been reproducing under all the other assumptions for a long enough time.

- We have made the assumption of synchronous reproduction. What happens when this assumption is violated?
- If you assume individuals live an exponentially distributed lifetime and then reproduce, then the HWE will be achieved when the last individual from the founding population dies. It could take a very long ime for this goal to be achieved.
- Exponentially distributed lifetimes are not usually applicable to biological populations. More complex models are difficult mathematically.

1.3 Estimating frequencies

1.3.1 Preliminaries

Application: Sampling and Frequency Estimation

Here is a reading that demonstrates how estimating allele frequencies in a population can be used to at least pose, if not answer, thought provoking questions about human evolution.

Walter E. Nance and Michael J. Kearsey (2004) Relevance of Connexin Deafness (DFNB1) to Human Evolution. *Am. J. Hum. Genet.* **74**:1081-1087.

- Mutations at over 100 loci (*plural* of locus) can cause deafness. (Not Mendelian!)
- Hypothesize that decreasing adverse effects of deafness (less severe selection) and *assortative mating* (we'll discuss, but you can read as *nonrandom mating*) on deafness can increase the incidence of the most common deafness allele in the population.
- In particular, they hypothesize that the incidence of deafness has increased since the introduction of sign language for these reasons.
- They use simulation to quantify the effect of these forces.

Estimation

In general, if we want to use evidence against HWE to estimate forces acting on populations, then we first need methods to estimate allele and genotype frequencies.

Recall that *statistics* are functions defined on data, i.e. *variables* measured on individuals of a *sample* collected from a *population* in some kind of *probabilitistic experiment*. *Estimators* are statistics meant to *estimate* a population parameter. An *estimate* is the value returned by an *estimator* for a particular data set. An *estimator*, and *statistics* in general, are random variables with *sampling distributions*, and therefore have *expectations* and *variances*, among other properties. We use these properties to assess the behavior of estimators.

I remind you of the various criteria we have for comparing among alternative estimators and choosing best estimators.

- **consistent**: estimator is consistent if it is is more and more accurate as n increases, specifically $\lim_{n\to\infty} P\left(|\hat{\theta}_n - \theta| < \epsilon\right) = 1$, where $\hat{\theta}_n$ is the estimate computed on a sample of size n.

- **unbiased**: $E(\hat{p}) = p$.
- **estimator variance**: $E\left[(\hat{p} - E(\hat{p}))^2\right]$.
- **efficient**: an estimator whose variance achieves the minimum possible variance.
- **sufficient**: a statistic is sufficient for a parameter if it contains all the information in a sample about that parameter.
- Result: There is an efficient estimator only if there is a sufficient statistic.

The Added Randomness of Population Genetics

Clearly, our estimates of population genetics parameters will vary from sample-to-sample, but there is another source of variation that sometimes needs to be considered, leading to two sources of variation in population genetics:

- **Statistical:** Traditional variation and uncertainty caused by taking a sample of size n from a population of size $N >> n$ and summarized in the *sampling distribution*.
- **Genetic**: life is stochastic
  - Reproduction and genetic transmission are random processes following precise, but nevertheless stochastic probability rules.
  - The population we study arose as a realization of this random process.
  - The variation resulting from this genetic sampling is important when:
    - predicting the genetic future of the population, and
    - studying the processes (including estimation of population genetics parameters) that gave rise to this population, and others like it.

Genetic randomness is only rarely considered in population genetics. The main reason is that, just as for statistical randomness, we cannot estimate it unless we have replication or solid theory (e.g. CLT concluding sample mean $\bar{X}$ has an approximate normal sampling distribution) for what variation would occur if we had replicates. In most cases, there are no replicates of the population under study. For example, there are no replicates of the U.S population. And while there is some simple theory about how populations arise, usually it is woefully invalid. Thus, in most cases we estimate population parameters only accounting for variation caused by sampling.

Summary Statistics for Populations

We will use the following notation to indicate common statistics computed from a sample.

- Sample size: n (number of diploids sampled)
- Sample counts of genotypes ($n_{A_1A_1}$, $n_{A_1A_2}$, $n_{A_2A_2}$ or n_{11}, n_{12}, n_{22}).
- Sample counts of alleles (n_{A_1}, n_{A_2} or n_1, n_2). Notice, allele counts are obtained from genotype counts:

$$\begin{aligned} n_1 &= n_{12} + 2n_{11} \\ n_2 &= n_{12} + 2n_{22} \end{aligned}$$

- Don't forget, there are $2n$ total alleles when n diploids are sampled, making the sample allele proportion $\tilde{p}_u = \frac{n_u}{2n}$.
- Sample frequencies (denoted by tilde)

$$\begin{aligned} \tilde{p}_{A_1} &= \frac{n_{A_1}}{2n} \quad \text{or } \tilde{p}_1 \\ \tilde{P}_{A_1A_2} &= \frac{n_{A_1A_2}}{n} \quad \text{or } \tilde{P}_{12} \end{aligned}$$

A Statistical Model of Genotype Sampling

When we sample individuals from a population, we do not traditionally sample with replacement, i.e. we generally eliminate the chance of sampling one individual more than once. Sampling without replacement from a finite population changes the genotype frequencies after each individual is sampled. However, if the population is large enough the change is negligible.

Given population frequencies P_{11}, P_{12}, P_{22} we could model the statistical sampling process as $Mult(n, P_{11}, P_{12}, P_{22})$ if the population size is large enough. More generally, if there are k classes of individuals, and the ith class exists in proportion Q_i in the population, the counts of individuals in each class are distributed as Multinomial.

- Multinomial distribution: $Mult(n, Q_1, Q_2, \ldots, Q_k)$

$$L(n_1, n_2, \ldots, n_k; Q_1, \ldots, Q_k) = P(n_1, n_2, \ldots, n_k) = \frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} Q_i^{n_i}$$

- Binomial distribution: $Bin(n, Q)$ applies when there are two categories

$$P(n_1, n - n_1) = \frac{n!}{n_1!(n - n_1)!} Q^{n_1} (1 - Q)^{n - n_1}$$

Reminder: Some Important Properties of Expectations and Variances

Some basic facts about expectations and variance are repeated as particularly relevant for this section.

- Expectation is a *linear* function on random variables, meaning for two random variables X and Y and constants a and b,

$$E[aX + bY] = aE[X] + bE[Y]$$

- Another formula for variance.

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

- Another formula for covariance.

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

- Variance is not a linear function on random variables, but there is a rule for random variables X and Y and constants a and b:

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

1.3.2 Sample Proportions

Estimating Multinomial Probabilities

We now propose the *sample proportion* as an estimator of population frequencies with multinomial data. Below, we examine the general properties of this estimator before proceeding to genetics applications.

- **Mean counts**: one can use the definition of expectation and Multinomial pmf to show:

$$E(n_i) \quad = \quad nQ_i$$

- Therefore, we can show the sample proportion is an *unbiased* estimate of population frequency.

$$E\left(\tilde{Q}_i\right) = E\left(\frac{n_i}{n}\right) = \frac{1}{n} E\left(n_i\right) = Q_i$$

- **Variance in counts**:

$$\text{Var}(n_i) = nQ_i(1 - Q_i)$$

- Therefore, the variance of our estimator is:

$$\text{Var}\left(\tilde{Q}_i\right) = \text{Var}\left(\frac{n_i}{n}\right) = \frac{1}{n^2}\text{Var}(n_i) = \frac{1}{n}Q_i(1 - Q_i) \tag{1}$$

Estimating Covariances and Correlations

Continuing, we can compute the covariance and correlation of pairs of sample statistics.

$$\begin{aligned}
E(n_i n_j) &= \sum_{r=0}^{n}\sum_{s=0}^{n-r} rsP(n_i = r, n_j = s) \\
&= n(n-1)Q_iQ_j \\
E\left(\tilde{Q}_i\tilde{Q}_j\right) &= \frac{n-1}{n}Q_iQ_j \\
\text{Cov}(n_i, n_j) &= -nQ_iQ_j \\
\text{Cov}(\tilde{Q}_i, \tilde{Q}_j) &= -\frac{1}{n}Q_iQ_j \\
\text{Corr}\,(n_i, n_j) &= \frac{\text{Cov}(n_i, n_j)}{\sqrt{\text{Var}(n_i)\text{Var}(n_j)}} \\
&= \text{Corr}(\tilde{Q}_i, \tilde{Q}_j)
\end{aligned}$$

[Suggestion: work through the mathematical details.]

Genotype Frequency Estimation

If genotypes are observed directly (e.g. all alleles are codominant), then genotype counts n_{11}, n_{12}, *etc*. follow a multinomial sampling distribution and genotype frequencies can be estimated by the unbiased sample proportion, e.g.

$$\tilde{P}_{12} = \frac{n_{12}}{n}$$

and the sampling variance is, directly from above,

$$\text{Var}\left(\tilde{P}_{12}\right) = \frac{1}{n}P_{12}\left(1 - P_{12}\right)$$

Allele Counts and Allele Frequency Estimation

Remembering, allele counts are obtained from genotype counts:

$$n_u = 2n_{uu} + \sum_{v<u} n_{uv}$$

we'll also show that the sample proportion for alleles is a reasonable estimate of the population allele frequency.

- Expected allele counts:

$$\begin{aligned}
E\left(n_u\right) &= E\left(2n_{uu} + \sum_{v<u} n_{uv}\right) \\
&= 2E\left(n_{uu}\right) + \sum_{v<u} E\left(n_{uv}\right) \\
&= 2nP_{uu} + \sum_{v<u} nP_{uv} = 2np_u
\end{aligned}$$

- Therefore, sample allele frequencies are unbiased estimates of population allele frequencies.

$$E\left(\tilde{p}_u\right) = E\left(\frac{n_u}{2n}\right) = \frac{1}{2n} \times 2np_u = p_u$$

Variance of Allele Frequency Estimator

$$\begin{aligned}
\text{Var}(n_u) &= \text{Var}\left(2n_{uu} + \sum_{v<u} n_{uv}\right) \\
&\quad \text{Apply formula for the variance of sums of random variables} \\
&\quad \text{and use the expectation/variance results for multinomial counts.} \\
&= 2n\left(p_u + P_{uu} - 2p_u^2\right) \\
\text{Var}\left(\tilde{p}_u\right) &= \text{Var}\left(\frac{n_u}{2n}\right) = \frac{1}{2n}\left(p_u + P_{uu} - 2p_u^2\right) \qquad (2)
\end{aligned}$$

[Suggestion: work through the mathematical details.]

Variance Estimation Now that we have our first estimator, we take a moment to discuss how to compute the sampling variance of estimators, covering a variety of methods, including several advanced methods. None of these advanced methods are needed for the sample proportion because direct analytic formulae are available, but the methods will become useful for more complicated estimators to come.

Importance of Estimating Sampling Variance

Computing the variance of an estimator tells us how estimates and inferences based on estimates will differ among samples. In the case of allele frequencies, if we estimate $\hat{p}_u = 0.3$, we are foolhardy to assume $p_u = 0.3$ without further consideration because uncertainty may be so large to make $p_u = 0.9$ nearly as likely.

- To actually use the variance (covariance, etc) formulas already derived requires knowledge of the population parameters, which of course, we don't have.
- Substitute sample proportions $\tilde{p}_u$ and $\tilde{P}_{uu}$ into the variance/covariance formulas to obtain an estimates

$$\widehat{\text{Var}}\left(\tilde{p}_u\right) \text{ and } \widehat{\text{Var}}\left(\tilde{P}_{uu}\right)$$

- If the sampling distribution is approximately normal (see if CLT applies to your estimator), then confidence intervals for the estimates can be obtained: The population parameter ϕ has approximately $(1-\alpha)\%$ chance of falling in the interval

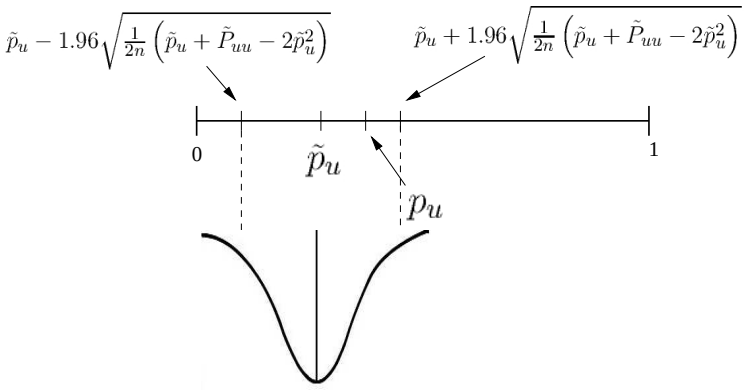
$$\hat{\phi} \pm z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}\left(\hat{\phi}\right)}.$$

  where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution, obtain in **R** as

```
R> qnorm(1-alpha/2)
```

Confidence Interval

If random variable $X \sim f(x)$, then one way to obtain a confidence interval at confidence level $1-\alpha$ (say $1-\alpha = 0.95$) is to find the critical values of $f(x)$ such that an area the size of $\frac{\alpha}{2}$ is clipped from the two tails of the distribution. If the random variable X is estimator of population parameter θ, then we interpret the confidence interval as defining a region that is sure to contain the true population paremeter θ in $(1-\alpha)\%$ of all possible samples. For more information, the Wikipedia entry (http://en.wikipedia.org/wiki/Confidence_interval) covers all the important points.



General Strategies for Computing Variance Estimates

As we discuss the following methods for estimating estimator variance, we'll use examples involving sample proportions for genetics to illustrate them, even when variances are available through simpler methods.

- Analytically compute the variance, then substitute estimates for population parameters. We used this approach to get Var($\tilde{p}_u$).
- Use indicator variables. Another analytic solution sometimes helpful for dealing with count data. We'll demonstrate.
- Approximate delta method (and for multinomial counts, Fisher's approximation).
- Approximate computational methods.
  - Jacknife
  - Bootstrap

Indicator Variables - Estimating Covariance of Allele Proportions

Let X_{ij} be a random indicator variable that is 1 if the jth allele in the ith individual is A_1 and 0 otherwise. Let Y_{ij} be a random indicator variable that is 1 if the jth allele in the ith individual is A_2 and 0 otherwise. Given these definitions

$$\begin{aligned}\tilde{p}_1 &= \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{2} X_{ij} \\ \tilde{p}_2 &= \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{2} Y_{ij}\end{aligned}$$

so we can compute

$$E\left(\tilde{p}_1\tilde{p}_2\right) = \frac{1}{4n^2}E\left[\left(\sum_i\sum_j X_{ij}\right)\left(\sum_i\sum_j Y_{ij}\right)\right]$$

Taking expectations of indicator variables is very easy:

$$\begin{aligned}E\left(X_{ij}\right) &= 1 \times P\left(X_{ij}=1\right) + 0 \times P\left(X_{ij}=0\right) \\ &= P\left(X_{ij}=1\right) = p_1\end{aligned}$$

We conclude (after algebra) that

$$E\left(\tilde{p}_1\tilde{p}_2\right) = p_1p_2 + \frac{1}{4n}\left(P_{12} - 4p_1p_2\right)$$

The covariance is then

$$\begin{aligned}\text{Cov}\left(\tilde{p}_1, \tilde{p}_2\right) &= E\left(\tilde{p}_1\tilde{p}_2\right) - p_1p_2 \\ &= \frac{1}{4n}\left(P_{12} - 4p_1p_2\right)\end{aligned}$$

[Suggestion: work through the mathematical details.]

Delta Method

- Let T be a function of the data, specifically the counts n_i: $T(n_1, n_2, \ldots)$.
- By Taylor's series:

$$\text{Var}(T) \approx \sum_i \left(\frac{\partial T}{\partial n_i}\right)^2 \text{Var}\left(n_i\right) + \sum_i \sum_{j \neq i} \frac{\partial T}{\partial n_i}\frac{\partial T}{\partial n_j}\text{Cov}\left(n_i, n_j\right)$$

  and replace n_i in the derivatives with $E(n_i) = nQ_i$ for multinomial counts.

- In addition, equations for variances and covariances of multinomial counts

$$\begin{aligned}\text{Var}\left(n_i\right) &= nQ_i\left(1 - Q_i\right) \\ \text{Cov}\left(n_i, n_j\right) &= -nQ_iQ_j\end{aligned}$$

  we have

$$\text{Var}(T) \approx n\sum_i \left(\frac{\partial T}{\partial n_i}\right)^2 Q_i - n\left(\sum_i \frac{\partial T}{\partial n_i}Q_i\right)^2$$

Fisher's Approximate Variance Formula

When $T(n_1, n_2, \ldots)$ is homogeneous of degree zero (i.e. $T(\beta n_1, \beta n_2, \ldots,) = T(n_1, n_2, \ldots)$), then.

$$\text{Var}(T) \approx n\sum_i \left(\frac{\partial T}{\partial n_i}\right)^2 Q_i - n\left(\frac{\partial T}{\partial n}\right)^2$$

where the second term is needed only when T explicitly involves the sample size n. In addition, terms with higher power of n in the denominator (e.g. $\frac{1}{n^2}$) are ignored in the derivative functions.

The above approximation works when

- T is a ratio of functions of the same order in the counts n_i, or
- counts n_i in T only appear divided by the total sample size n.

Application: Fisher's Approximation

Example:

$$
\begin{aligned}
\tilde{P}_{12} &= \frac{n_{12}}{n} \\
\frac{\partial T}{\partial n_{12}} &= \frac{1}{n} \\
\frac{\partial T}{\partial n} &= \frac{-n_{12}}{n^2} = \frac{-P_{12}}{n} \\
\text{Var}\left(\tilde{P}_{12}\right) &\approx n \times \left(\frac{1}{n}\right)^2 P_{12} - n \times \left(\frac{P_{12}}{n}\right)^2 \\
&= \frac{1}{n} P_{12}\left(1 - P_{12}\right)
\end{aligned}
$$

which agrees with eq. (1)

Other Methods for Confidence Intervals

What can one do when the sample size is small ($n < 30$) or when no formula for the variance can be derived?

Jackknife

- You begin with a sample of observations $X_1, X_2, \ldots, X_n$ of size n.
- You use these data to calculate an estimate $\hat{\phi}$.
- Compute n new estimates $\hat{\phi}_{(i)}$ where the ith estimate is calculated using all the data minus the ith data point, e.g. $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$.
- Compute their average
$$\hat{\phi}_{(\cdot)} = \frac{1}{n} \sum_i \hat{\phi}_{(i)}$$
- Obtain a less biased estimated:
$$\hat{\phi}_J = n\hat{\phi} - (n-1)\hat{\phi}_{(\cdot)}$$
- Calculate an estimate of the variance of $\hat{\phi}$
$$\text{Var}\left(\hat{\phi}\right)_J = \frac{n-1}{n} \sum_i \left(\hat{\phi}_{(i)} - \hat{\phi}_{(\cdot)}\right)^2$$

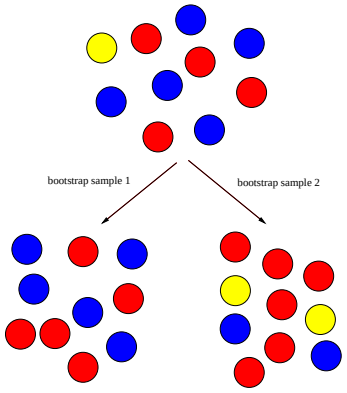
Bootstrap

- Obtain M samples by sampling with replacement from the original data.
- Compute the boostrap estimate $\hat{\phi}_{(i)}$ for each bootstrap dataset.
- Plot histogram of $\hat{\phi}_{(i)}$ for all $i = 1, \ldots, M$ to obtain an approximation to the sampling distribution.
- Estimate bias

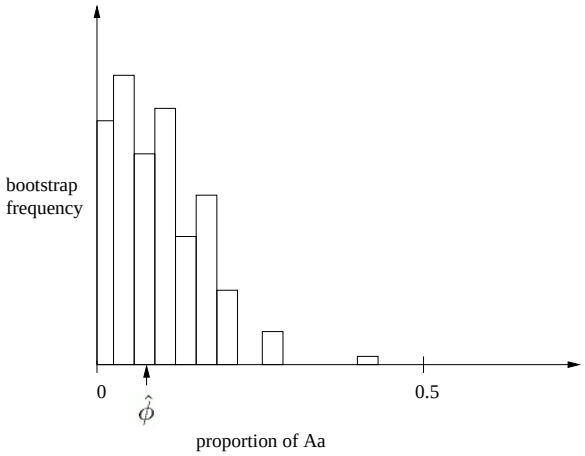
$$\text{bias} \doteq \frac{\sum_{b=1}^{M}(\hat{\phi}_{(b)} - \hat{\phi})}{M}$$

- Estimate variance

$$\text{variance} \doteq \frac{\sum_{b=1}^{M}(\hat{\phi}_{(b)} - \bar{\hat{\phi}}_{(b)})}{M-1}$$



Bootstrap Sampling Distribution



Genetic Sampling Variance

We have only computed among sample within population variances so far. In general, for the cases where between population variance should be considered, we need to do more work with variances. In particular, multiple populations subject to the same evolutionary process will not all share the same exact population allele frequencies. Methods that account for this variation in population allele frequencies will be discussed, but we note that the confidence intervals will be larger when genetic variation is added to sampling variation.

1.3.3 Maximum Likelihood

Maximum Likelihood

We have analyzed the use of sample proportions as estimates of population frequencies. We now consider some alternative estimators.

We already know how to compute maximum likelihood estimates of the multinomial proportions, e.g. genotype frequencies P_{11}, P_{12}, and P_{22}. In particular, the sample proportions *are* the MLEs for multinomial proportions. Here, we will derive MLEs for allele frequencies.

- Suppose the expected proportions Q_i from the multinomial distribution are functions of other population parameters. For example, under HWE

$$\begin{aligned} P_{11} &= p_1^2 \\ P_{12} &= 2p_1p_2 = 2p_1\left(1-p_1\right) \\ P_{22} &= \left(1-p_1\right)^2. \end{aligned}$$

- Suppose we observe counts n_{11}, n_{12}, n_{22}, then the *likelihood* of the data can be written in terms of the allele frequencies:

$$\begin{aligned} L(n_{11}, n_{12}, n_{22}; p_1) &= \frac{n!}{n_{11}!n_{12}!n_{22}!}\left(P_{11}\right)^{n_{11}}\left(P_{12}\right)^{n_{12}}\left(P_{22}\right)^{n_{22}} \\ &= \frac{n!}{n_{11}!n_{12}!n_{22}!}p_1^{2n_{11}}\left[2p_1\left(1-p_1\right)\right]^{n_{12}}\left(1-p_1\right)^{2n_{22}} \end{aligned}$$

[Review] Supports and Scores

- It is usually more convenient to work with $\ln L$, called the *support*.
- The derivatives of the support with respect to the parameters are called the *scores*:

$$S_{p_1} = \frac{\partial \ln L}{\partial p_1}$$

- The *maximum likelihood estimates* are those values of the parameters (e.g. p_1) that maximize the likelihood. They are found by setting the scores equal to 0 and simultaneously solving the resulting system of equations.

Maximum likelihood estimate of p_1

$$\begin{aligned} L(p_1) &= \frac{n!}{n_{11}!n_{12}!n_{22}!}p_1^{2n_{22}}\left[2p_1\left(1-p_1\right)\right]^{n_{12}}\left(1-p_1\right)^{2n_{22}} \\ \ln L(p_1) &= \ln\left(\frac{n!}{n_{11}!n_{12}!n_{22}!}\right) + \left(2n_{11}+n_{12}\right)\ln\left(p_1\right) + \left(n_{12}+2n_{22}\right)\ln\left(1-p_1\right) \\ S_{p_1} &= \frac{2n_{11}+n_{12}}{p_1} - \frac{n_{12}+2n_{22}}{1-p_1} \end{aligned}$$

Solve

$$S_{p_1} = 0$$

to obtain the maximum likelihood estimate

$$\hat{p}_1 = \frac{1}{2n}\left(2n_{11}+n_{12}\right).$$

We know this is the maximum because

$$\begin{aligned} \frac{\partial S_{p_1}}{\partial p_1} &= -\frac{2n_{11}+n_{12}}{p_1^2} - \frac{n_{12}+2n_{22}}{\left(1-p_1\right)^2} \\ &<= 0 \text{ for all } p_1 \end{aligned}$$

[Review] Properties of MLEs

- Do not attempt to estimate parameters that are functions of each other. For example $P_{11} = 1 - P_{12} - P_{22}$ when there are only two alleles and thus three genotypes. (Either write out the functional form, i.e. use $1 - P_{12} - P_{22}$ instead of P_{11} or use Lagrange Multipliers).
- The MLE of a function of parameters is the function of the MLEs. For example,

$$\widehat{p_i^2} = \hat{p}_i^2$$

- The MLE may be biased.
- MLEs are consistent estimators under general conditions, so for very large samples the bias disappears.
- The information of a parameter is the negative second derivative, e.g.

$$I_{p_1} = -\left(\frac{\partial^2 \ln L(p_1)}{\partial p_1^2}\right)$$

- For large samples, the variance of the MLE is the inversed expected information:

$$\text{Var}\left(\hat{p}_1\right) = \frac{1}{E\left[I_{p_1}\right]}$$

- When the likelihood is a function of multiple independent parameters, e.g. p_{11}, p_{12}, the information is a matrix.
- The variance is obtained as the inverse of the expectation of this matrix.
- For large samples, the MLE is approximately normally distributed (and parameter vectors are multivariate normal). For example,

$$\hat{p}_1 \sim N\left(p_1, \left\{E\left[I\left(p_1\right)\right]\right\}^{-1}\right)$$

Example: Variance of $\hat{p}_1$

$$\begin{aligned}
\frac{\partial S_{p_1}}{\partial p_1} &= -\frac{2n_{11} + n_{12}}{p_1^2} - \frac{n_{12} + 2n_{22}}{\left(1 - p_1\right)^2} \\
E\left[-\frac{\partial S_{p_1}}{\partial p_1}\right] &= \frac{2E[n_{11}] + E[n_{12}]}{p_1^2} + \frac{E[n_{12}] + 2E[n_{22}]}{\left(1 - p_1\right)^2} \\
&= \frac{2np_1^2 + 2np_1p_2}{p_1^2} + \frac{2np_1p_2 + 2np_2^2}{\left(1 - p_1\right)^2} \\
&= \frac{2n}{p_1} + \frac{2n}{1 - p_1} \\
&= \frac{2n}{p_1(1 - p_1)} \\
\text{Var}\left(\hat{p}_1\right) &= \frac{p_1(1 - p_1)}{2n}
\end{aligned}$$

Note, that this is the variance for a sample proportion estimator of a multinomial probability [see eq. (1)] as if we had directly sampled alleles (and not individuals) from the population. Under HWE, individuals "sample" their alleles at random from the preceding generation's allele frequencies, and allele frequencies are unchanging. Therefore, under HWE, sampling n individuals is equivalent to sampling $2n$ alleles directly. Note also that this formula is equivalent to eq. (2) if we substitute $P_{11} = p_1^2$ in the former.

Lagrange Multipliers[Footnote]

One method for dealing with relationships among the parameters utilizes the Lagrange multiplier λ. For example, for the general multinomial likelihood, modify the log likelihood to obtain

$$\ln L = C + \sum_i n_i \ln(Q_i) + \lambda \left(1 - \sum_i Q_i\right).$$

Then the scores

$$S_{Q_i} = \frac{n_i}{Q_i} - \lambda$$

yield the algebraic system

$$n_i - \lambda Q_i = 0.$$

Find λ by summing all scores

$$\sum_i n_i - \lambda \sum_i Q_i = n - \lambda = 0$$

and

$$\hat{Q}_i = \frac{n_i}{n}.$$

Bailey's Method for MLEs

When the number of independent parameters is equal to the number of independent pieces of information (degrees of freedom) in the data, then Bailey's Method applies.

Bailey's method obtains MLEs by setting observations to their expected values. These equations may be easier to solve than the equations obtained by setting the scores equal to zero.

Example:

Suppose you have derived a model in which there are two unknowns p_1 and f:

$$\begin{aligned} P_{11} &= p_1^2 + p_1(1-p_1)f \\ P_{12} &= 2p_1(1-p_1)(1-f). \end{aligned}$$

You collect data and produce the following independent genotype counts n_{11}, n_{12}. Bailey's method prescribes:

$$\begin{aligned} E(n_{11}) &= n\left[\hat{p}_1^2 + \hat{p}_1(1-\hat{p}_1)\hat{f}\right] = n_{11} \\ E(n_{12}) &= n\left[2\hat{p}_1(1-\hat{p}_1)(1-\hat{f})\right] = n_{12}. \end{aligned}$$

The MLEs $\hat{p}_1$ and $\hat{f}$ are obtained by solving this system of equations.

- 2 times the first equation plus the second equation eliminates f and yields

$$2\frac{n_{11}}{n} + \frac{n_{12}}{n} = 2\hat{p}_1^2 + 2\hat{p}_1(1-\hat{p}_1) = 2\hat{p}_1$$

- Rearrange yields the usual sample proportions estimator

$$\hat{p}_1 = \tilde{p}_1 = \frac{n_{11}}{n} + \frac{n_{12}}{2n}.$$

- The estimator $\hat{f}$ is obtained by plugging this back into either equation.

$$\hat{f} = \frac{4nn_{11} - (2n_{11} + n_{12})^2}{(2n_{11} + n_{12})(2n - 2n_{11} - n_{12})}.$$

MLEs by Computer: Iterative Formulae

- Sometimes the equations obtained by the preceding methods do not have analytic solutions. What to do?
- Grid methods:
  - Set up a grid of possible values for your parameters (e.g. for p_A a very simply grid would be $0, 0.01, 0.02, \ldots, 0.99, 1$).
  - Evaluate the likelihood at each grid point and plot the results.
- **Newton-Raphson iteration:** Is probably the fastest method, but it is not guaranteed to find the MLE. It can get stuck. To counteract this problem, start from multiple starting conditions.
  - Begin with a guess for your parameter ϕ_0.
  - Use Taylor's theorem to update from your current estimate ϕ_t to your next estimate ϕ_{t+1}.

$$\phi_{t+1} = \phi_t - \frac{S_{\phi_t}}{I(\phi_t)}$$

  - Iterate until $|\phi_{t+1} - \phi_t| < \epsilon$, where ϵ is some tolerance you decide on before iteration starts.
- **EM Algorithm**: we'll discuss next...
- **Rely on someone smarter. R** provides functions like *optim()*, which can do simultaneous optimization over multiple parameters.

Expectation-Maximization (EM) Algorithm

Another iterative algorithm that is useful when the observed data can be considered incomplete, i.e. there is some information you are missing. For count data (from multinomial sampling distribution), missing information means there are categories you cannot observe (two or more categories are subsumed into one).

- Make an initial guess ϕ_0.
- **Expectation step:** Assume that the previous iteration ϕ_t is the true value of the population parameter. Estimate the expected complete data using the conditional probability:

$$P(\text{unobservable category} \mid \text{observable category}, \phi_t)$$

- **Maximization step:** Compute the MLEs using the *complete likelihood* and call them ϕ_{t+1}.

Question

Can you think of examples in genetics where information might be missing?

- Dominant allele - Multiple genotypes appear phenotypically the same. For example, when there are two alleles, genotypes A_1A_1 and A_1A_2 appear the same. Observed data is the dominant phenotype $n_{11,12} = n_{11} + n_{12}$ and the recessive phenotype n_{22}.
- Phase information - $\frac{AB}{ab}$ looks the same as $\frac{Ab}{aB}$.

Estimating Recessive Allele Frequency

Can you think of a way to estimate the recessive allele frequency p_2 given your current knowledge about estimation?

Assume HWE. Then, the two classes we can observe have probabilities

$$\begin{aligned} P_{11} + P_{12} &= p_1^2 + 2p_1(1-p_1) \\ P_{22} &= (1-p_1)^2 \end{aligned}$$

Bailey's method applies because there is one degree of freedom in the data ($n_{11,12} + n_{22} = n$) and there is one parameter p_1 to estimate.

$$n_{22} = n(1-p_1)^2$$

yields

$$\hat{p}_2 = 1 - \hat{p}_1 = \sqrt{\frac{n_{22}}{n}}.$$

Estimating Recessive Frequency Using EM

Despite having just found the mle directly, we will use this example to demonstrate the EM algorithm.

- Again, you must assume HWE. Otherwise, anything is possible.
- Determine the conditional probability P (unobservable category $\mid$ observable category, ϕ):

$P(A_1A_1 \mid A_1A_1 \text{ or } A_1A_2, p_2)$	$P(A_1A_1 \mid A_2A_2, p_2)$
$P(A_1A_2 \mid A_1A_1 \text{ or } A_1A_2, p_2)$	$P(A_1A_2 \mid A_2A_2, p_2)$

$\frac{P(A_1A_1 \mid p_2)}{P(A_1A_1 \text{ or } A_1A_2 \mid p_2)}$	0
$\frac{P(A_1A_2 \mid p_2)}{P(A_1A_1 \text{ or } A_1A_2 \mid p_2)}$	0

Facts used:

$$\begin{aligned} P(A \mid B, C) &= \frac{P(A \text{ and } B \mid C)}{P(B \mid C)} \\ \{A_1A_1 \cap \{A_1A_1 \cup A_1A_2\}\} &= \{A_1A_1\} \end{aligned}$$

$\frac{[1-p_2]^2}{1-p_2^2} = \frac{1-p_2}{1+p_2}$	0
$\frac{2p_2[1-p_2]}{1-p_2^2} = \frac{2p_2}{1+p_2}$	0

Facts used:

$$\begin{aligned} P(A_1A_1) &= [1-p_2]^2 \\ P(A_1A_1 \cup A_1A_2) &= [1-p_2]^2 + 2p_2[1-p_2] \end{aligned}$$

Expectation Step

Let $p_2(t)$ be the value of our recessive allele estimate in the tth step of the iteration.

$$\begin{aligned} n_{11}^* = E\left(n_{11} \mid n_{11,12}, n_{22}, p_2(t)\right) &= \sum_{i=0}^{n_{11,12}} iP\left(i \mid n_{11,12}, p_2(t)\right) \\ &= \frac{n_{11,12}\left[1 - p_2(t)\right]}{1 + p_2(t)} \\ n_{12}^* = E\left(n_{12} \mid n_{11,12}, n_{22}, p_2(t)\right) &= \frac{2n_{11,12}p_2(t)}{1 + p_2(t)} \end{aligned}$$

Maximization Step

Pretend you observed the data $n_{11}^*, n_{12}^*, n_{22}$, where $n_{11,12} = n_{11}^* + n_{12}^*$. Find the maximum likelihood estimate of the nth iteration $\hat{p}_2$.

$$\begin{aligned} \hat{p}_2 &= \tilde{p}_2(t) = \frac{n_{12}^* + 2n_{22}}{2n} \\ &= \frac{1}{2n}\left[\frac{2n_{11,12}p_2(t)}{1 + p_2(t)} + 2n_{22}\right] \end{aligned} \tag{3}$$

This MLE becomes the new value of our recessive allele frequency:

$$p_2(t+1) = \hat{p}_2$$

Summary: EM Iteration

- Guess a value for $p_2(0)$, say 0.5.
- In the tth iteration, compute the expected data n_{11}^*, n_{12}^*, given $p_2(t)$.
- Given the complete set of expected plus real data $n_{11}^*, n_{12}^*, n_{22}$, compute the MLE $\hat{p}_2$.
- Set the next iterate estimate $p_2(t+1)$ to the MLE $\hat{p}_2$.
- Iterate until $|p_2(t+1) - p_2(t)| < \epsilon$, where ϵ is some small tolerance, like 0.0001.

Variance for the EM Estimates

We now have an MLE estimate of p_2. To complete our estimation procedure, we need a variance for $\hat{p}_2$. Generally, the problem of obtaining variances for EM-obtained estimates is difficult. Your options are:

- Substitute EM estimates into the information matrix and invert.
- Use the resampling techniques. For each bootstrap or jackknife sample, you need to run EM to compute another MLE.

Analysis of Iterative Schemes

- Cautions about iterative schemes
  - Iteration scheme may not converge.

- Iteration scheme may converge but not to the maximum.
- Use multiple starting values or consider solving to find the fixed point.

- When might maximum likelihood fail
  - No sampling distribution is known so the likelihood is not available.
  - The sample is small and the MLE is biased.

Variance for $\hat{p}_2$

To complete our estimation of p_2, we now derive derive and study its variance. Please note that the mle we computed before using Bailey's method

$$\hat{p}_2 = \sqrt{\frac{n_{22}}{n}}$$

also satisfies the EM recurrence relation (eq. 3) when the algorithm has converged, i.e. when $p_2(t+1) = p_2(t)$. In this special case, when the MLE can be obtained analytically, we also see that the EM equations can be analytically solved, and there is no need for the algorithm! Anyway, now we need a variance...

Can we use Fisher's approximation?

Yes.

$$\begin{aligned}
\text{Var}(T) &\approx n\sum_i \left(\frac{\partial T}{\partial n_i}\right)^2 Q_i - n\left(\frac{\partial T}{\partial n}\right)^2 \\
T &= \sqrt{\frac{n_{22}}{n}} \\
\text{Var}\left(\hat{p}_2\right) &\approx n\left(\frac{\partial T}{\partial n_{22}}\right)^2 P_{22} - n\left(\frac{\partial T}{\partial n}\right)^2
\end{aligned}$$

$$\text{Var}\left(\hat{p}_2\right) \approx n\left(\frac{\partial T}{\partial n_{22}}\right)^2 P_{22} - n\left(\frac{\partial T}{\partial n}\right)^2$$

$$\begin{aligned}
\frac{\partial T}{\partial n_{22}} &= \frac{1}{2n}\sqrt{\frac{n}{n_{22}}} \\
\frac{\partial T}{\partial n} &= -\frac{1}{2}\sqrt{\frac{n}{n_{22}}}\frac{n_{22}}{n^2}
\end{aligned}
\qquad\Bigg|\qquad
\begin{aligned}
\left(\frac{\partial T}{\partial n_{22}}\right)^2 &= \frac{1}{4n_{22}n} \\
\left(\frac{\partial T}{\partial n}\right)^2 &= \frac{n_{22}}{4n^3}
\end{aligned}$$

$$\begin{aligned}
\left(\frac{\partial T}{\partial n_{22}}\right)^2\Bigg|_{E[n_{22}]} &= \frac{1}{4n^2P_{22}} \\
\left(\frac{\partial T}{\partial n}\right)^2\Bigg|_{E[n_{22}]} &= \frac{P_{22}}{4n^2}
\end{aligned}$$

What do we do now?

Substitute in an estimate for P_{22}:

$$\text{Var}\left(\hat{p}_2\right) \quad \hat{=} \quad \frac{1}{4n} - \frac{n_{22}}{4n^2}$$

Also note we can write the variance estimator in terms of p_2 using HWE:

$$\text{Var}\left(\hat{p}_2\right) \quad \approx \quad \frac{1}{4n}\left(1 - p_2^2\right)$$

Comparison of Variances

- The variance of $\hat{p}_2$ for codominant alleles is smaller than the variance for dominant alleles:

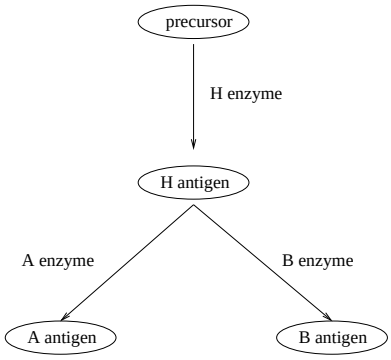
$$\begin{aligned} \frac{\text{Var}_{\text{dominant}}\left(\hat{p}_2\right)}{\text{Var}_{\text{codominant}}\left(\hat{p}_2\right)} &\approx \frac{\frac{1}{4n}\left(1 - p_2^2\right)}{\frac{1}{2n}p_2\left(1 - p_2\right)} \\ &= \frac{\left(1 + p_2\right)}{2p_2} \\ &= \frac{1}{2p_2} + \frac{1}{2} \\ &> \frac{1}{2} + \frac{1}{2} = 1 \end{aligned}$$

  for all $p_2 \in [0, 1]$.

ABO locus biology

The following locus is classically used as a didactic example to illustrate the EM algorithm in genetics.

- ABO locus is on chromosome 9. It encodes for an enzyme that modifies H antigen.
- There are three alleles possible at this locus. A modifies antigen H to produce antigen A. B modifies antigen H to produce antigen B. O does not encode a working enzyme, leaving H antigen untouched.
- An antigen is a substance that can stimulate the production of antibodies.
- A, B, and H antigens are also produced by plants. Our bodies create antibodies to all except the ones we recognize as "self."
- Our immune system's antibodies will attack blood containing antigens that are foreign to us.



Counting Genotypes

- If there are 3 alleles at a locus, how many genotypes are possible?

$$6 : AB, AO, BO, AA, BB, OO$$

- In general, if there are n alleles at a locus, how many genotypes are possible?

$$\binom{n}{2} + n = \binom{n+1}{2} = \frac{(n+1)!}{2!(n-1)!} = \frac{(n+1)n}{2}$$

- How many of these are distinguishable?

  There are four groups: $\{AA, AO\}, \{BB, BO\}, \{AB\}, \{OO\}$.

Review Slide

- If HWE assumptions are met, then $P_{ij} = 2p_i p_j, i \neq j$ and genotype and allele frequencies are unchanging in time.
- We sought estimates of genotype frequencies P_{ij} and allele frequencies p_i.
- Estimators
  - maximum likelihood
    - Set score equal to zero.
    - Bailey's method
    - Iterative: Newton-Raphson, EM, others (e.g. `optim()`)
  - method of moments
  - Bayesian
- Obtaining variance of estimators
  - direct calculation
  - direct calculation with indicator variables
  - approximate analytic: Delta method with Fisher's approximation
  - approximate computational: bootstrap and jacknife
  - via information matrix for mles

1.3.4 Method of Moments

Method of Moments

Definition: *moment*

The nth moment of random variable X is $E[X^n]$. The nth *central* moment is $E[(X-\mu)^n]$, where $\mu = E[X]$ is the first moment.

Definition: *sample moment*

The kth sample moment of a sample $X_1, \ldots, X_n$ of size n is

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

It is easy to show that the sample moment is *unbiased* for the population moment.

Definition: *method of moments*

The *method of moments* produces estimates of population parameters by setting sample moments equal to population moments.

Please note that Bailey's method is a special case of MOM-estimation and shows that MOM estimators can be MLEs, in some cases.

MOM: Advantages and Disadvantages

- **Less accurate/efficient.** MLEs have higher probability of being close to the quantities to be estimated.
- **Easier.** In some cases, likelihood equations may be intractable without computers, but MOMs can be quickly and easily calculated by hand.
- **Initialization.** MOMs can be used to initialize MLE iterative solutions.
- **Invalid.** Sometimes, particularly for n small, the MOMs can fall outside the parameter space; not true for MLEs.
- **No MLEs.** When estimating other structural parameters (e.g., parameters of a utility function, instead of parameters of a known probability distribution), appropriate probability distributions may not be known, and moment-based estimates may be preferred to MLE.

Example: MOM Estimate of f

Examine the expectations of moments of sample proportions and play around until you can solve these equations for an unknown parameter.

As an example, recall our model (presented so far without motivation):

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

Generally, for any allele u, $\text{Var}(\tilde{p}_u) = E[\tilde{p}_u^2] - p_u^2$, so

$$\begin{aligned} E\left(\tilde{p}_u^2\right) &= p_u^2 + \frac{1}{2n}\left(p_u + P_{uu} - 2p_u^2\right) \\ &= p_u^2 + \frac{1}{2n}\left(p_u + p_u^2 + p_u(1-p_u)f - 2p_u^2\right) \\ &= p_u^2 + \frac{1}{2n}p_u\left(1-p_u\right)\left(1+f\right). \end{aligned}$$

Also,

$$E\left(\tilde{P}_{uu}\right) = p_u^2 + p_u\left(1-p_u\right)f.$$

Cast around for something to do with these expression to isolate f. Hmmmm....let's sum over alleles...

$$E\left(\sum_u \tilde{p}_u^2\right) = F \quad = \quad \sum_u p_u^2 + \frac{1}{2n}\sum_u p_u\left(1-p_u\right)\left(1+f\right).$$

And...let's sum over all homozygotes:

$$\begin{aligned} E\left(\sum_u \tilde{P}_{uu}\right) = G &= \sum_u p_u^2 + \sum_u p_u\left(1-p_u\right)f \\ &= \sum_u p_u^2 + f\left(1 - \sum_u p_u^2\right). \end{aligned}$$

Now, if we want an estimator for f, we need to work these equations until all occurrences of p_u are eliminated.

$$\frac{G - F + \frac{1}{2n} - \frac{1}{2n}G}{1 - F - \frac{1}{2n} + \frac{1}{2n}G} \quad = \quad f$$

1.3.5 Bayesian Method

Bayesian Approach

- The *frequentist* approach we have been using focuses on the quantity $L(X, \theta) = P(X; \theta)$, the likelihood.
- Using Bayesian notation, this likelihood can be rewritten as

  $$L(X, \theta) = P(X \mid \theta)$$

  because it is the probability of the data *conditioning* on the model and values for its parameters.
- Some people consider it to be much more natural to think about the following conditional probability

  $$P(\theta \mid X),$$

  the probability that the population parameter takes on a particular value given that we have collected some data X.
- **Bayes Theorem** relates these two conditional probabilities

  $$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)}$$

- **Prior.** Before we perform an experiment or collect data we have some *prior* knowledge or beliefs about the parameter values. For example, at the very least you know $P_{AA} \in [0, 1]$. One summarizes this information in a *prior distribution* $P(\theta)$.
- **Posterior.** Data is collected and our *prior* beliefs are updated to our *posterior* beliefs $P(\theta \mid X)$ using Bayes' theorem. The *posterior distribution* provides us with the state of knowledge about the population parameters after we have examined the data.
- Bayes theorem follows much the same logic we humans use to gain knowledge in the world. We begin with certain beliefs and understandings $P(\theta)$. We collect more information X. We update our beliefs to $P(\theta \mid X)$.
- The denominator in Bayes theorem can be problematic. When possible, it is obtained as:

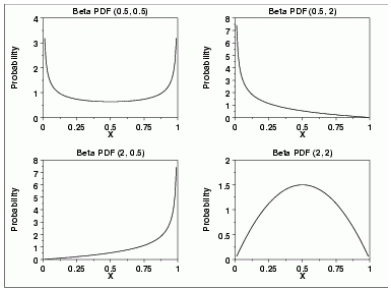
  $$P(X) = \int_{\theta} P(X \mid \theta)P(\theta)d\theta.$$

Bayesian Procedure: Overview

- Specify a reasonable prior distribution.
- Determine your sampling distribution and hence your likelihood.
- Compute the posterior distribution. Analytic solution may not be possible, but in the last couple of decades Markov Chain Monte Carlo, a numerical approximation, has allowed for the flouring of the Bayesian technique.
- Compute Bayesian estimators (e.g. mean, mode, median, etc) from the posterior distribution.

Bayesian Method: Allele Estimation

Let's use the Bayesian approach to estimate the allele frequency p_1 at a diallelic locus.



Select a beta prior distribution (mathematical convenience):

$$P(p_1) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_1^{\alpha-1}(1-p_1)^{\beta-1} = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} p_1^{\alpha-1}(1-p_1)^{\beta-1}$$

Compute the posterior distribution:

$$\begin{aligned} P(p_1 \mid n_1, n_2) &\propto \frac{2n!}{n_1!(2n-n_1)!} p_1^{n_1}(1-p_1)^{2n-n_1} \times \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} p_1^{\alpha-1}(1-p_1)^{\beta-1} \\ &= \frac{\Gamma(\alpha+\beta+2n)}{\Gamma(\alpha+n_1)\Gamma(\beta+2n-n_1)} p_1^{\alpha+n_1-1}(1-p_1)^{\beta+2n-n_1-1} \end{aligned}$$

Compute Bayesian estimators: For example, the posterior mean

$$\begin{aligned} E(p_1 \mid n_1, n_2) &= \frac{\alpha+n_1}{\alpha+\beta+2n} \\ &= \frac{\alpha}{\alpha+\beta+2n} + \frac{2n}{\alpha+\beta+2n}\hat{p}_1 \end{aligned}$$

Bayesian Method: Advantages/Disadvantages

- How does one invent a prior distribution? How might incompetent selection of prior distributions affect the estimation? The results are not objective, since another prior can give you different results.
- On the other hand, often prior information is available. It's logical to use it then. It can enhance the power of the data, for example when you data agrees with the bulk of prior information.
- The $(1-\alpha)\%$ Bayesian confidence intervals are interpreted as "the parameter falls within these limits with $(1-\alpha)\%$ probability." This cannot be said of frequentist confidence intervals where $(1-\alpha)\%$ of similarly constructed confidence intervals are expected to contain the true parameter.