ALLELE FREQUENCIES

Bruce Weir

April 9, 2024

Lincoln (Plant and Food research)

Probability

Probability provides the language of data analysis.

Equiprobable outcomes definition:
Probability of event E is number of outcomes favorable to E divided by the total number of outcomes. e.g. Probability of a head $= 1/2$.

Long-run frequency definition:
If event E occurs n times in N identical experiments, the probability of E is the limit of n/N as N goes to infinity.

Subjective probability:
Probability is a measure of belief.

First Law of Probability

Law says that probability can take values only in the range zero to one and that an event which is certain has probability one.

$$\begin{cases} 0 \leq \Pr(E) \leq 1 \\ \\ \Pr(E|E) = 1 \text{ for any } E \end{cases}$$

i.e. If event E is true, then it has a probability of 1. For example:

$$\Pr(\text{Seed is Round}|\text{Seed is Round}) \;=\; 1$$

Second Law of Probability

If G and H are mutually exclusive events, then:

$$\Pr(G \text{ or } H) \;=\; \Pr(G) + \Pr(H)$$

For example,

$$\Pr(\text{Seed is Round or Wrinkled}) \;=\; \Pr(\text{Round}) + \Pr(\text{Wrinkled})$$

More generally, if $E_i, i = 1, \ldots r$, are mutually exclusive then

$$\begin{aligned} \Pr(E_1 \text{ or } \ldots \text{ or } E_r) &= \Pr(E_1) + \ldots + \Pr(E_r) \\ &= \sum_i \Pr(E_i) \end{aligned}$$

Complementary Probability

If $\Pr(E)$ is the probability that E is true then $\Pr(\bar{E})$ denotes the probability that E is false. Because these two events are mutually exclusive

$$\Pr(E \text{ or } \bar{E}) = \Pr(E) + \Pr(\bar{E})$$

and they are also exhaustive in that between them they cover all possibilities – one or other of them must be true. So,

$$\begin{aligned} \Pr(E) + \Pr(\bar{E}) &= 1 \\ \Pr(\bar{E}) &= 1 - \Pr(E) \end{aligned}$$

The probability that E is false is one minus the probability it is true.

Third Law of Probability

For any two events, G and H, the third law can be written:

$$\Pr(G \text{ and } H) \;=\; \Pr(G)\Pr(H|G)$$

There is no reason why G should precede H and the law can also be written:

$$\Pr(G \text{ and } H) \;=\; \Pr(H)\Pr(G|H)$$

For example

$$\begin{aligned} & \Pr(\text{Seed is round \& is type AA}) \\ = \;& \Pr(\text{Seed is round}|\text{Seed is type AA}) \times \Pr(\text{Seed is type AA}) \\ = \;& 1 \times p_A^2 \end{aligned}$$

Independent Events

If the information that H is true does nothing to change uncertainty about G, then

$$\Pr(G|H) \;=\; \Pr(G)$$

and

$$\Pr(H \text{ and } G) \;=\; \Pr(H)\Pr(G)$$

Events G, H are independent: "Product Rule".

Law of Total Probability

If $G, \bar{G}$ are two mutually exclusive and exhaustive events ($\bar{G}$ = not G), then for any other event E, the law of total probability states that

$$\Pr(E) = \Pr(E|G)\Pr(G) + \Pr(E|\bar{G})\Pr(\bar{G})$$

This generalizes to any set of mutually exclusive and exhaustive events $\{S_i\}$:

$$\Pr(E) = \sum_i \Pr(E|S_i)\Pr(S_i)$$

For example

$$\begin{aligned}
\Pr(\text{Seed is round}) &= \Pr(\text{Round}|\text{Type AA})\Pr(\text{Type AA}) \\
&\quad + \Pr(\text{Round}|\text{Type Aa})\Pr(\text{Type Aa}) \\
&\quad + \Pr(\text{Round}|\text{Type aa})\Pr(\text{Type aa}) \\
&= 1 \times p_A^2 + 1 \times 2p_A p_a + 0 \times p_a^2 \\
&= p_A(2 - p_A)
\end{aligned}$$

Bayes' Theorem

Bayes' theorem relates $\Pr(G|H)$ to $\Pr(H|G)$:

$$\begin{aligned}\Pr(G|H) &= \frac{\Pr(GH)}{\Pr(H)}, \text{ from third law} \\ &= \frac{\Pr(H|G)\Pr(G)}{\Pr(H)}, \text{ from third law}\end{aligned}$$

If $\{G_i\}$ are exhaustive and mutually exclusive, Bayes' theorem can be written as

$$\Pr(G_i|H) = \frac{\Pr(H|G_i)\Pr(G_i)}{\sum_i \Pr(H|G_i)\Pr(G_i)}$$

Bayes' Theorem Example

Suppose G is event that a man has genotype A_1A_2 and H is the event that he transmits allele A_1 to his child. Then $\Pr(H|G) = 0.5$.

Now what is the probability that a man has genotype A_1A_2 given that he transmits allele A_1 to his child?

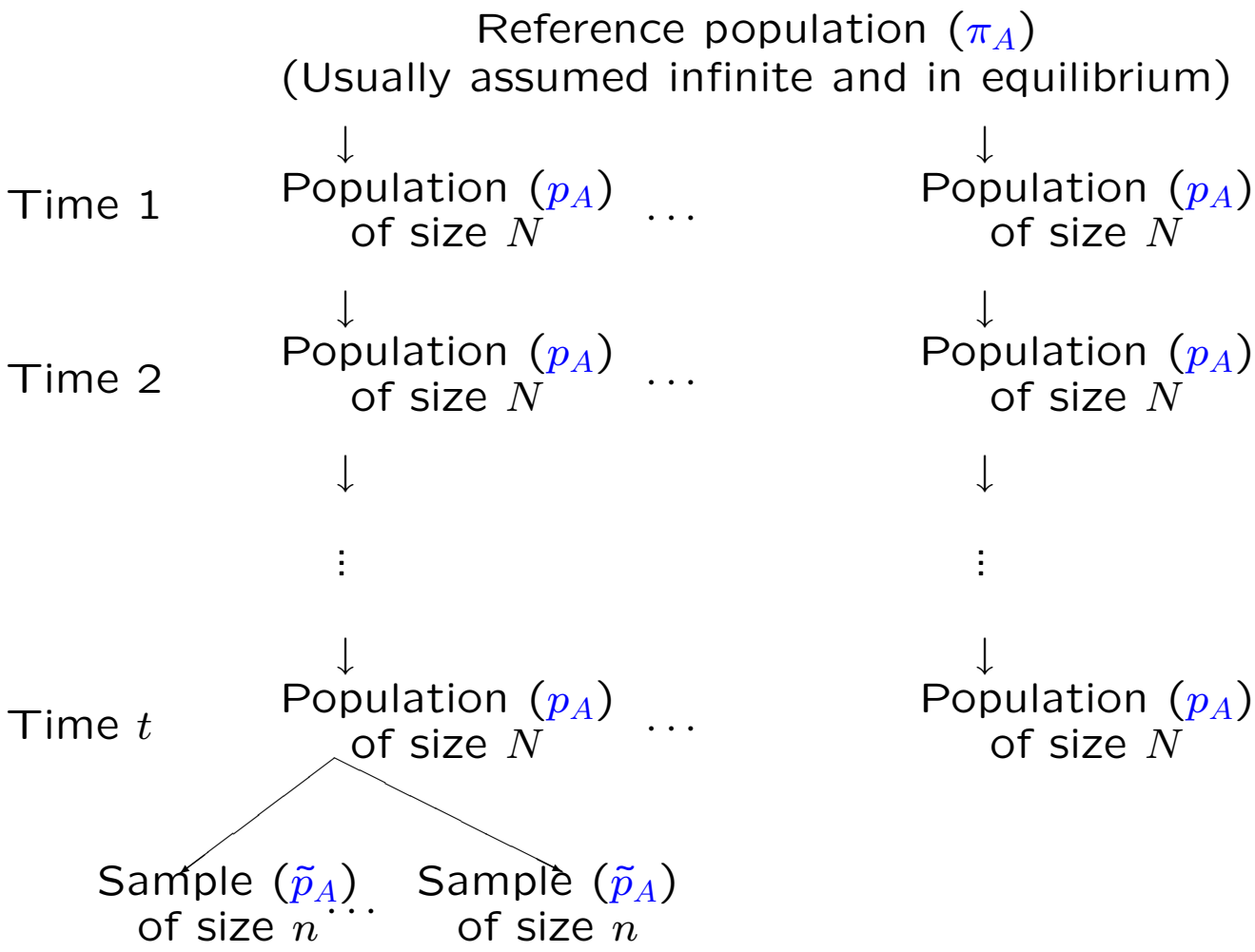
$$\begin{aligned} \Pr(G|H) &= \frac{\Pr(H|G)\Pr(G)}{\Pr(H)} \\ &= \frac{0.5 \times 2p_1p_2}{p_1} \\ &= p_2 \end{aligned}$$

Sampling

Statistical sampling: The variation among repeated samples from the same population ("fixed" sampling). Inferences can be made about that particular population.

Genetic sampling: The variation among replicate (conceptual) populations ("random" sampling). Inferences are made to all populations with the same pedigree.

Classical Model



Probability π_A, population proportion p_A, sample frequency $\tilde{p}_A$ for allele A.

Aside: Coalescent Theory

An alternative framework works with genealogical history of a sample of alleles. There is a tree linking all alleles in a current sample to the "most recent common ancestral allele." Allelic variation is due to mutations since that ancestral allele.

The coalescent approach requires mutation and may be more appropriate for long-term evolution and analyses involving more than one species. The classical approach allows mutation but does not require it: within one species variation among populations may be due primarily to drift.

Binomial Distribution

Properties of Estimators

Consistency	Increasing accuracy as sample size increases
Unbiasedness	Expected value is the parameter
Efficiency	Smallest variance
Sufficiency	Contains all the information in the data about parameter

Binomial Distribution

Most population genetic data consists of numbers of observations in some categories. The values and frequencies of these counts form a *distribution*.

Toss a coin n times, and note the number of heads. There are $(n+1)$ outcomes, and the number of times each outcome is observed in many sets of n tosses gives the sampling distribution. Or: sample n alleles from a population and observe x copies of type A.

Binomial distribution

If every toss has the same chance p of giving a head:

Probability of x heads in a row of *independent* tosses is

$$p \times p \times \ldots \times p \;=\; p^x$$

Probability of $n - x$ tails in a row of *independent* tosses is

$$(1-p) \times (1-p) \times \ldots \times (1-p) \;=\; (1-p)^{n-x}$$

The number of ways of ordering x heads and $n - x$ tails among n outcomes is $n!/[x!(n-x)!]$.

The binomial probability of x successes in n trials is

$$\Pr(x|p) \;=\; \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Binomial Likelihood

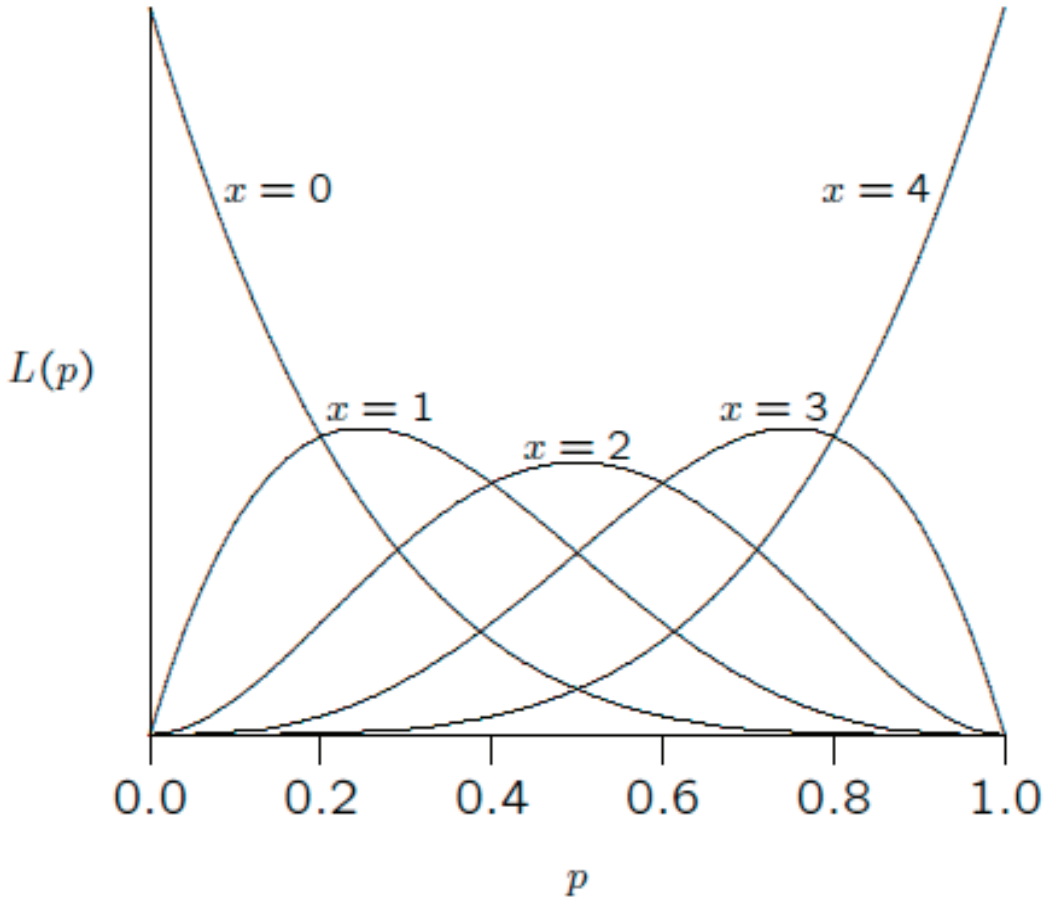
The quantity $\Pr(x|p)$ is the *probability of the data*, x successes in n trials, when each trial has probability p of success.

The same quantity, written as $L(p|x)$, is the *likelihood of the parameter*, p, when the value x has been observed. The terms that do not involve p are not needed, so

$$L(p|x) \quad \propto \quad p^x(1-p)^{(n-x)}$$

Each value of x gives a different likelihood curve, and each curve points to a p value with maximum likelihood. This leads to *maximum likelihood estimation*.

Likelihood $L(p|x, n = 4)$



Binomial Mean

If there are n trials, each of which has probability p of giving a success, the *mean* or the *expected number* of successes is np.

The *sample proportion* of successes is

$$\tilde{p} \;=\; \frac{x}{n}$$

(This is also the maximum likelihood estimate of p.)

The expected, or *mean*, value of $\tilde{p}$ is p.

$$\mathcal{E}(\tilde{p}) \;=\; p$$

Binomial Variance

The expected value of the squared difference between the number of successes and its mean, $(x - np)^2$, is $np(1-p)$. This is the *variance* of the number of successes in n trials, and indicates the spread of the distribution.

The statistical sampling variance of the sample proportion $\tilde{p}$ is

$$\text{Var}(\tilde{p}) = \frac{p(1-p)}{n}$$

Normal Approximation

Provided np is not too small (e.g. not less than 5), the binomial distribution can be approximated by the normal distribution with the same mean and variance. In particular:

$$\tilde{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

To use the normal distribution in practice, change to the *standard normal* variable z with a mean of 0, and a variance of 1:

$$z = \frac{\tilde{p} - p}{\sqrt{p(1-p)/n}}$$

For a standard normal, 95% of the values lie between ± 1.96. The normal approximation to the binomial therefore implies that 95% of the values of $\tilde{p}$ lie in the range

$$p \pm 1.96\sqrt{p(1-p)/n}$$

Confidence Intervals

A 95% confidence interval is a variable quantity. It has end-points which vary with the sample. It is expected that 95% of samples will lead to an interval that includes the unknown true value p.

The standard normal variable z has 95% of its values between -1.96 and $+1.96$. This suggests that a 95% Wald confidence interval for the binomial parameter p is

$$\tilde{p} \pm 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$$

Confidence Intervals

For samples of size 10, the 11 possible confidence intervals are:

$\tilde{p}$	Confidence Interval
0.0	$0.0 \pm 2\sqrt{0.000} = (0.00, 0.00)$
0.1	$0.1 \pm 2\sqrt{0.009} = (0.00, 0.29)$
0.2	$0.2 \pm 2\sqrt{0.016} = (0.00, 0.45)$
0.3	$0.3 \pm 2\sqrt{0.021} = (0.02, 0.58)$
0.4	$0.4 \pm 2\sqrt{0.024} = (0.10, 0.70)$
0.5	$0.5 \pm 2\sqrt{0.025} = (0.19, 0.81)$
0.6	$0.6 \pm 2\sqrt{0.024} = (0.30, 0.90)$
0.7	$0.7 \pm 2\sqrt{0.021} = (0.42, 0.98)$
0.8	$0.8 \pm 2\sqrt{0.016} = (0.55, 1.00)$
0.9	$0.9 \pm 2\sqrt{0.009} = (0.71, 1.00)$
1.0	$1.0 \pm 2\sqrt{0.000} = (1.00, 1.00)$

Can modify interval a little by extending it by the "continuity correction" $\pm 1/2n$ in each direction.

Confidence Intervals

To be 95% sure that the estimate is no more than 0.01 from the true value, $1.96\sqrt{p(1-p)/n}$ should be less than 0.01. The widest confidence interval is when $p = 0.5$, and then the sample size should satisfy

$$0.01 \geq 1.96\sqrt{0.5 \times 0.5/n}$$

which means that $n \geq 10,000$. For a width of 0.03 instead of 0.01, $n \approx 1,000$ as is common in public opinion surveys.

If the true value of p was about 0.05, however,

$$\begin{aligned} 0.01 &\geq 2\sqrt{0.05 \times 0.95/n} \\ n &\geq 1,900 \approx 2,000 \end{aligned}$$

Exact Confidence Intervals: One-sided

The normal-based confidence intervals are constructed to be symmetric about the sample value, unless the interval goes outside the interval from 0 to 1. They are therefore less satisfactory the closer the true value is to 0 or 1.

More accurate confidence limits follow from the binomial distribution exactly. For events with low probabilities p, how large could p be for there to be at least a 5% chance of seeing no more than x (i.e. $0, 1, 2, \ldots x$) occurrences of that event among n events. If this upper bound is p_U,

$$\sum_{k=0}^{x} \Pr(k) \geq 0.05$$

$$\sum_{k=0}^{x} \binom{n}{k} p_U^k (1-p_U)^{n-k} \geq 0.05$$

If $x = 0$, then $(1-p_U)^n \geq 0.05$ if $p_U \leq 1 - 0.05^{1/n}$ and this is 0.0295 when $n = 100$. More generally, $p_U \approx 3/n$ when $x = 0$.

Aside: Two-sided Exact Confidence Intervals

A two-sided interval is bounded above by p_U for which there is at least a 2.5% chance of seeing no more than x (i.e. $0, 1, 2 \ldots x$) occurrences, and is bounded below by p_L for which there is at least a 2.5% chance of seeing at least x (i.e. $x, x+1, x+2, \ldots n$) occurrences:

$$\sum_{k=0}^{x} \binom{n}{k} p_U^k (1-p_U)^{n-k} \geq 0.025$$

$$\sum_{k=x}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} \geq 0.025$$

If $x = 0$, then $(1-p_U) \geq 0.025^{1/n}$ and this gives $p_U \leq 0.036$ when $n = 100$.

If $x = n$, then $p_L \geq 0.975^{1/n}$ and this gives $p_L \geq 0.964$ when $n = 100$.

Exact CIs for $n = 10$

One-sided			Two-sided			
x	$\tilde{p}$	p_U	x	p_L	$\tilde{p}$	p_U
0	0.00	0.26	0	0.00	0.00	0.31
1	0.10	0.39	1	0.00	0.10	0.45
2	0.20	0.51	2	0.03	0.20	0.56
3	0.30	0.61	3	0.07	0.30	0.65
4	0.40	0.70	4	0.12	0.40	0.74
5	0.50	0.78	5	0.19	0.50	0.81
6	0.60	0.85	6	0.26	0.60	0.88
7	0.70	0.91	7	0.35	0.70	0.93
8	0.80	0.96	8	0.44	0.80	0.97
9	0.90	0.99	9	0.55	0.90	1.00
10	1.00	1.00	10	0.69	1.00	1.00

The two-sided exact CI is not symmetrical around $\tilde{p}$.

Bootstrapping

An alternative method for constructing confidence intervals uses *numerical resampling*. A set of samples is drawn, with replacement, from the original sample to mimic the variation among samples from the original population. Each new sample is the same size as the original sample, and is called a *bootstrap sample*.

The middle 95% of the sample values $\tilde{p}$ from a large number of bootstrap samples provides a 95% confidence interval.

Allele Frequency Sampling

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}

For a sample of n *independently* sampled individuals, 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:

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

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:

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

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

For the sample counts:

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

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 statistical sampling 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_{jk} for allele k in individual j:

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

Then

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

If there is random sampling, individuals are independent, and

$$\mathcal{E}(x_{jk}x_{j'k'}) = \mathcal{E}(x_{jk})\mathcal{E}(x_{j'k'}) = p_A^2$$

These expectations are the averages of values from many samples from the same population. (Later work on Genetic Relatedness Matrices does not assume independent individuals.)

An aside: Intraclass Correlation

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

$$\begin{aligned} \mathrm{Var}(x_{jk}) &= \mathcal{E}(x_{jk}^2) - [\mathcal{E}(x_{jk})]^2 \\ &= p_A(1-p_A) \\ &= \mathrm{Var}(x_{jk'}), \;\; k \neq k' \end{aligned}$$

and

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

so

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

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 of 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.

Maximum Likelihood Estimation: Allele Data

For a sample of n *independent* alleles, the likelihood of p_A when there are n_A alleles of type A is

$$L(p_A|n_A) \;=\; C(p_A)^{n_A}(1-p_A)^{n-n_A}$$

and this is maximized when

$$\frac{\partial L(p_A|n_A)}{\partial p_A} = 0 \quad \text{or when} \quad \frac{\partial \ln L(p_A|n_A)}{\partial p_A} = 0$$

Now

$$\ln L(p_A|n_A) \;=\; \ln C + n_A \ln(p_A) + (n-n_A)\ln(1-p_A)$$

so

$$\frac{\partial \ln L(p_A|n_A)}{\partial p_A} \;=\; \frac{n_A}{p_A} - \frac{n-n_A}{1-p_A}$$

and this is zero when $p_A = n_A/n$. The MLE of p_A is its sample value: $\widehat{p}_A = \tilde{p}_A$.

Maximum Likelihood Estimation: Genotype Data

If $\{n_i\}$ are multinomial with parameters n and $\{P_i\}$, then the MLE's of P_i are n_i/n. This will always hold for genotype proportions, but not always for allele proportions.

For two alleles, the MLE's for genotype proportions are:

$$\begin{aligned} \hat{P}_{AA} &= n_{AA}/n \\ \hat{P}_{Aa} &= n_{Aa}/n \\ \hat{P}_{aa} &= n_{aa}/n \end{aligned}$$

Does this lead to estimates of allele proportions and the within-population inbreeding coefficient?

Maximum Likelihood Estimation: f

Because

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

The likelihood function for p_A, f is

$$\begin{aligned} L(p_A, f) &= \frac{n!}{n_{AA}! n_{Aa}! n_{aa}!}[p_A^2 + p_A(1-p_A)f]^{n_{AA}} \\ &\quad \times [2p_A(1-p_A)f]^{n_{Aa}}[(1-p_A)^2 + p_A(1-p_A)f]^{n_{aa}} \end{aligned}$$

and it is difficult to find, algebraically, the values of p_A and f that maximize this function or its logarithm.

There is an alternative way of finding maximum likelihood estimates in this case: equating the observed and expected values of the genotype frequencies.

Bailey's Method

Because the number of parameters (2) equals the number of degrees of freedom in this case, we can just equate observed and expected genotype proportions based on the estimates of p_A and f:

$$\begin{aligned} n_{AA}/n &= \hat{p}_A^2 + \hat{f}\hat{p}_A(1-\hat{p}_A) \\ n_{Aa}/n &= 2\hat{p}_A(1-\hat{p}_A) - 2\hat{f}\hat{p}_A(1-\hat{p}_A) \\ n_{aa}/n &= (1-\hat{p}_A)^2 + \hat{f}\hat{p}_A(1-\hat{p}_A) \end{aligned}$$

Solving these equations (e.g. by adding the first equation to half the second equation to give solution for $\hat{p}_A$ and then substituting that into one equation):

$$\hat{p}_A = \frac{2n_{AA} + n_{Aa}}{2n} = \tilde{p}_A$$

$$\hat{f} = 1 - \frac{n_{Aa}}{2n\tilde{p}_A(1-\tilde{p}_A)} = 1 - \frac{\tilde{P}_{Aa}}{2\tilde{p}_A\tilde{p}_a}$$

Method of Moments

An alternative to maximum likelihood estimation is the method of moments (MoM) where observed values of statistics are set equal to their expected values regardless of degrees of freedom. In general, this does not lead to unique estimates or to estimates with variances as small as those for maximum likelihood.

(Bailey's method is for the special case where the MLEs are also MoM estimates.)

Aside: Method of Moments

For the inbreeding coefficient at loci with m alleles A_u, two possible MoM estimates are (for large sample sizes)

$$\hat{f}_{\text{LH1}} = \frac{\sum_{u=1}^{m}(\tilde{P}_{uu} - \tilde{p}_u^2)}{\sum_{u=1}^{m}\tilde{p}_u(1-\tilde{p}_u)}$$

$$\hat{f}_{\text{LH5}} = \frac{1}{m-1}\sum_{u=1}^{m}\left(\frac{\tilde{P}_{uu} - \tilde{p}_u^2}{\tilde{p}_u}\right)$$

These both have low bias. Their variances depend on the value of f.

For loci with two alleles, $m = 2$, the two moment estimates are equal to each other and to the maximum likelihood estimate:

$$\hat{f}_{LH1} = \hat{f}_{LH5} = 1 - \frac{\tilde{P}_{Aa}}{2\tilde{p}_A\tilde{p}_a}$$

Li CC, Horvitz DG. 1953. Am J Human Genetics 5:107-16. Equations 1 and 5.

Aside: MLE for Recessive Alleles

Suppose allele a is recessive to allele A, and a sample of n individuals has n_{aa} recessive homozygotes. The genotypes of the other $(n-n_{Aa})$ individuals can be AA or Aa. If there is Hardy-Weinberg equilibrium, the likelihood for the two phenotypes is

$$\begin{aligned} L(p_a) &= (p_a^2)^{n_{aa}}(1-p_a^2)^{n-n_{aa}} \\ \ln[L(p_a)] &= 2n_{aa}\ln(p_a) + (n-n_{aa})\ln(1-p_a^2) \end{aligned}$$

Differentiating wrt p_a:

$$\frac{\partial \ln L(p_a)}{\partial p_a} = \frac{2n_{aa}}{p_a} - \frac{2p_a(n-n_{aa})}{1-p_a^2}$$

Setting this to zero leads to an equation that can be solved explicitly: $p_a = \sqrt{n_{aa}/n}$.

Aside: EM Algorithm for Recessive Alleles

An alternative way of finding maximum likelihood estimates when there are "missing data" involves *Estimation* of the missing data and then *Maximization* of the likelihood. For a locus with allele A dominant to a the missing information is the counts of the AA and Aa genotypes. Only the joint count $(n - n_{aa})$ of $AA + Aa$ is observed.

Estimate the missing genotype counts (assuming independence of alleles) as proportions of the total count of dominant phenotypes:

$$\begin{aligned} n_{AA} &= \frac{(1-p_a)^2}{1-p_a^2}(n-n_{aa}) = \frac{(1-p_a)(n-n_{aa})}{(1+p_a)} \\ n_{Aa} &= \frac{2p_a(1-p_a)}{1-p_a^2}(n-n_{aa}) = \frac{2p_a(n-n_{aa})}{(1+p_a)} \end{aligned}$$

Aside: EM Algorithm for Recessive Alleles

Maximize the likelihood (using Bailey's method):

$$\begin{aligned}\widehat{p}_a &= \frac{n_{Aa} + 2n_{aa}}{2n} \\ &= \frac{1}{2n}\left(\frac{2p_a(n - n_{aa})}{(1 + p_a)} + 2n_{aa}\right) \\ &= \frac{2(np_a + n_{aa})}{2n(1 + p_a)}\end{aligned}$$

An initial estimate p_a is put into the right hand side to give an updated estimated $\widehat{p}_a$ on the left hand side. This is then put back into the right hand side to give an iterative equation for p_a.

This procedure also has explicit solution $\widehat{p}_a = \sqrt{n_{aa}/n}$.

EM Algorithm for Two Loci

A more interesting application of the EM algorithm is the estimation of two-locus gamete frequencies from unphased genotype data. For locus **A** with alleles A, a and locus **B** with alleles B, b, the ten two-locus frequencies are:

Genotype	Actual	Expected	Genotype	Actual	Expected
AB/AB	P_{AB}^{AB}	p_{AB}^2	AB/Ab	P_{Ab}^{AB}	$2p_{AB}p_{Ab}$
AB/aB	P_{aB}^{AB}	$2p_{AB}p_{aB}$	AB/ab	P_{ab}^{AB}	$2p_{AB}p_{ab}$
Ab/Ab	P_{Ab}^{Ab}	p_{Ab}^2	Ab/aB	P_{aB}^{Ab}	$2p_{Ab}p_{aB}$
Ab/ab	P_{ab}^{Ab}	$2p_{Ab}p_{ab}$	aB/aB	P_{aB}^{aB}	p_{aB}^2
aB/ab	P_{ab}^{aB}	$2p_{aB}p_{ab}$	ab/ab	P_{ab}^{ab}	p_{ab}^2

EM Algorithm for Two Loci

Gamete frequencies are marginal sums:

$$\begin{aligned} p_{AB} &= P_{AB}^{AB} + \frac{1}{2}(P_{Ab}^{AB} + P_{aB}^{AB} + P_{ab}^{AB}) \\ p_{Ab} &= P_{Ab}^{Ab} + \frac{1}{2}(P_{AB}^{Ab} + P_{ab}^{Ab} + P_{aB}^{Ab}) \\ p_{aB} &= P_{aB}^{aB} + \frac{1}{2}(P_{AB}^{aB} + P_{ab}^{aB} + P_{Ab}^{aB}) \\ p_{ab} &= P_{ab}^{ab} + \frac{1}{2}(P_{Ab}^{ab} + P_{aB}^{ab} + P_{AB}^{ab}) \end{aligned}$$

Arrange the gamete frequencies as a two-way table to show that only one of them is unknown when the allele frequencies are known:

p_{AB}	p_{Ab}	p_A
p_{aB}	p_{ab}	p_a
p_B	p_b	1

EM Algorithm for Two Loci

The two double heterozygote counts n_{ab}^{AB}, n_{aB}^{Ab} are "missing data."

Assume initial value of p_{AB} and Estimate the missing counts as proportions of the total count n_{AaBb} of double heterozygotes:

$$\begin{aligned} n_{ab}^{AB} &= \frac{2p_{AB}p_{ab}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} n_{AaBb} \\ n_{aB}^{Ab} &= \frac{2p_{Ab}p_{aB}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} n_{AaBb} \end{aligned}$$

and then Maximize the likelihood by setting

$$\begin{aligned} p_{AB} &= \frac{1}{2n}\left(2n_{AB}^{AB} + n_{Ab}^{AB} + n_{aB}^{AB} + n_{ab}^{AB}\right) \\ \text{or} & \\ n_{AB} &= 2n_{AB}^{AB} + n_{Ab}^{AB} + n_{aB}^{AB} + n_{ab}^{AB} \end{aligned}$$

Example

As an example, consider these data:

	BB	Bb	bb	Total
AA	$n_{AABB} = 0$	$n_{AABb} = 0$	$n_{AAbb} = 2$	$n_{AA} = 2$
Aa	$n_{AaBB} = 1$	$n_{AaBb} = 3$	$n_{Aabb} = 4$	$n_{Aa} = 8$
aa	$n_{aaBB} = 0$	$n_{aaBb} = 1$	$n_{aabb} = 4$	$n_{aa} = 5$
Total	$n_{BB} = 1$	$n_{Bb} = 4$	$n_{bb} = 10$	$n = 15$

There is one unknown gamete count: e.g. $x = n_{AB}$ for AB:

	B	b	Total
A	$n_{AB} = x$	$n_{Ab} = 12 - x$	$n_A = 12$
a	$n_{aB} = 6 - x$	$n_{ab} = x + 12$	$n_a = 18$
Total	$n_B = 6$	$n_b = 24$	$2n = 30$

$$0 \leq x \leq 6$$

Example

EM iterative equation:

$$
\begin{aligned}
x' &= 2n_{AABB} + n_{AABb} + n_{AaBB} + n_{AB/ab} \\
&= 2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{2p_{AB}p_{ab}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} n_{AaBb} \\
&= 0 + 0 + 1 + 3 \times \frac{2x(x+12)}{2x(x+12) + 2(12-x)(6-x)} \\
&= 1 + \frac{3x(x+12)}{x(x+12) + (12-x)(6-x)}
\end{aligned}
$$

Example

A good starting value would assume independence of A and B alleles: $x = 2n * p_A * p_B = (30 \times 12/30 \times 6/30) = 2.4$. Successive iterates are:

Iterate	x	$x/2n$
1	2.4000	0.0800
2	2.5000	0.0833
3	2.5647	0.0855
4	2.6063	0.0869
5	2.6327	0.0878
6	2.6494	0.0883
7	2.6600	0.0887
8	2.6667	0.0889
9	2.6709	0.0890
10	2.6736	0.0891
11	2.6752	0.0892
12	2.6763	0.0892
13	2.6769	0.0892
14	2.6773	0.0892
15	2.6776	0.0893
16	2.6778	0.0893
...	...	...