ALLELE FREQUENCIES

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

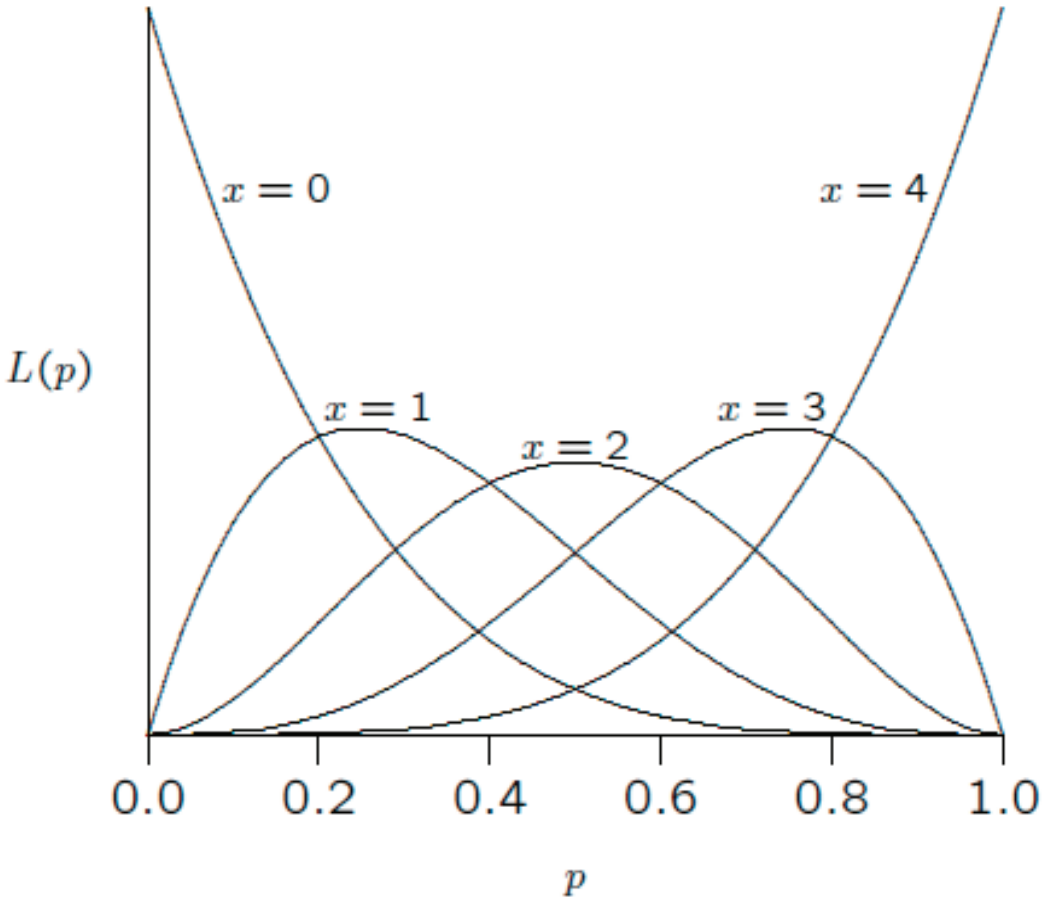
The quantity $\Pr(x|p)$ is the *probability of the data*, x successes in n trials, when each trial has probability p of success.

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 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% 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 $x = 0$.

Exact Confidence Intervals: Two-sided

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 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.