

Principles of population genetic inference

Gil McVean

Department of Statistics, University of Oxford

January 29, 2002

Contents

1	Statistical inference	2
1.1	Principles of inference	2
1.2	The unusual nature of population genetic data	4
1.3	The null-model principle	5
2	Genes in populations	5
2.1	The Wright-Fisher model	5
2.2	The coalescent	7
2.3	Effective population size	9
2.4	Variation in coalescence times	10
3	Properties of the n-coalescent	11
3.1	How much variability?	12
3.2	How much variability in levels of variability?	12
3.3	Other mutation models	13
4	Summarizing data	14
4.1	Summary statistics	14
4.2	Graphical representation of data	15
5	Stochastic simulation	17
6	Estimating parameters	18
6.1	Moment estimators	19
6.2	Likelihood	21
6.3	Bayesian inference	23
6.4	Strengths and weaknesses of coalescent theory	24

1 Statistical inference

1.1 Principles of inference

Inference is the process of extracting information from a sample about the forces that generated the data. There is, of course, no universally accepted way of doing inference; in fact there are several different philosophies. But underlying all approaches there are certain key ideas. The elements of inference, and how they are influenced by the peculiar nature of population genetic data, will be the focus of this lecture.

The process of statistical inference can be summed up by the flow chart on slide 2, along with the various issues that need to be dealt with at each stage. The process of inference (parametric, or model-based, inference to be more precise) starts by building a model of the world within which to work. A model has three elements; rules, parameters and quantities. The rules determine how the parameters of the model influence the way in which the quantities behave. In population genetics, the parameters are things like mutation rates, migration rates, and selection coefficients; the rules govern how organisms interact, reproduce, move and die; and the quantities are the genes (or chromosomes, alleles, etc.).

Once we have a model of the world from which to work, we then need to explore the properties of the model. By this I mean we need to explore how variation in the parameters leads to variation in the quantities of interest. How best to summarize data so that it is informative about the underlying parameters. How to represent data graphically, so that important patterns can be identified. An important part of this process is simulation; the generation of artificial data sets from the model. Simulation allows us to explore the properties of the model when analytical expressions cannot be derived, as is usually the case in population genetics. The phase of model exploration also allows us to refine the basic model if we feel it is inadequate in some fashion. Usually a model will have been formulated with some prior knowledge, or intuitive understanding, of the process of interest. If the model

produces some very strange and unlikely results, then it is probably time to go back to the drawing board.

If we have decided that our model is a reasonable approximation to reality, the next step is to start looking at real data. The central idea here is that we wish to estimate the parameters of the model from the sample by finding the set of parameter values that, within the rules of the model, are most compatible with the data we have observed. How to gauge “compatibility” is a problem that has generated much argument over the years. Three ideas that we will explore in regards to population genetics are: moment methods, in which observed quantities are equated to expectations; likelihood, in which the estimated parameters are those most likely to have generated exactly the data you observed; and Bayesian methods, in which the sample data is used to update prior information on the parameters of interest. The strengths and weaknesses of each method will become apparent as we progress.

Almost as important as finding the most compatible parameter values is the need to place some indication of how much faith to put in the estimates. Usually, this takes the form of identifying a range of possible parameter values that are reasonable, given the data, with the hope that the true value lies somewhere within this interval. Again, how this degree of uncertainty is defined varies with the method of parameter estimation.

Although it may seem that we have achieved what we set out to do, the process of statistical inference is far from complete. Having used an explicit model of the population genetic process as the basis for estimating parameters, those estimates are only as good as the model is an accurate representation of biological reality. In practice we can never completely model the evolutionary process, all we can hope for is that we have captured the important features. It is therefore of critical importance that we ask how well the data is explained by the model. In particular, we may wish to look for unusual parts of the data (outliers) and test for greater variation in aspects of the data than we should expect (heterogeneity). Another model-testing procedure that has proved extremely successful in population genetics is to com-

pare two estimators of the same parameter (which use different aspects of the data) and to ask whether the difference between the estimates is greater than we should expect. Detecting deviations from the assumed model is an extremely important part of inference, as it can lead to the discovery of interesting phenomena.

In all probability we will need to refine our model on the basis of what was learned from the goodness of fit tests. This probably means introducing new parameters into the model, which in turn will lead to another period of model exploration, data collection, parameter estimation and more model testing. The process could go on for ever.

Finally, it is worth sounding a note of caution about how much analysis to do on a single data set. First, the goodness-of-fit tests you are going to carry out should be decided on before the collection of data. If you decide on analyses after looking at the data (i.e. on a *post hoc* basis), you run the risk of finding what you want to find, rather than a meaningful deviation from the assumed model. This is also true of model refinement. Data should be used to identify new hypotheses and suggest model improvements that are then tested on newly collected data.

1.2 The unusual nature of population genetic data

Before describing the models used in population genetics it is interesting to consider what there is about population genetic data that makes the process of statistical inference unusual.

In many experiments, such as looking for the effective of a fertilizer on crop growth, or assessing the impact of a drug on disease treatment, the data consists of information about one or a few measured variables (height, weight, time of disease progression, etc.) collected from a large number of independent replicates, and the models used to describe the data are the standard tools of statistics; Normal distributions, logistic regression, etc. Consequently, the dimensions of the sample space (essentially the range of possible outcomes) is fairly small, and it is possible to use all possible information for inference (within the rules of the model).

In population genetics, you have a very different situation. The data collected often consists of sequences (or alleles) from a single locus, which represents the outcome of a single evolutionary history (in effect we have a single data point). The range of possible outcomes for any given set of evolutionary parameters is enormous (i.e. the sample space has many dimensions), and it is usually very difficult to devise methods of analysis that use all possible information in the sample. In many ways, the analysis of population genetic data bears greater resemblance to particle physics than biology or medicine.

1.3 The null-model principle

A complete model of the evolutionary process would include descriptions of the rates of mutation, migration, recombination and natural selection at every nucleotide position in every genome of every organism that has ever lived. In other words, it is impossible to achieve. However, what we can hope to achieve is a representation of biological complexity that includes the key processes and the key levels of heterogeneity in the parameters that govern the processes. Where then should we start?

Given that our aim is to reduce the complexity of evolution down to its key features, the obvious place to start is to devise the simplest possible model. In other words, we shall start with the premise that *nothing interesting ever happens in evolution*, then by asking how reality differs from this null model, we can hopefully identify the key processes that make evolution interesting.

2 Genes in populations

2.1 The Wright-Fisher model

The null model in population genetics is called the Wright-Fisher (WF) population model, after the two great theoreticians who introduced and explored the theoretical properties of the model. The model has many assumptions, almost none of which

are true for any single biological population

- Constant diploid population of size $N(2N$ alleles)
- Non-overlapping generations
- Random mating with respect to both geographic location and genotype
- Sexual reproduction with the possibility of selfing
- No migration to or from other populations
- Mutations are neutral and occur at a constant rate μ per generation

Reproduction works by randomly selecting a gamete from one individual, and then randomly selecting another gamete from a second individual (which may be the same as the first) to make one offspring. This is repeated N times, until a daughter population of N individuals has been generated. It is easy to see why this type of model is called a bean-bag model, because its properties are exactly reproduced by having a bag with $2N$ beans (some of which may be a different colour if there are several alleles in the population), and picking $2N$ beans with replacement in order to make a daughter population.

Let us ignore mutations for the time being. We can follow the fate of genes over several generations by iterating the WF model. Over time, some lineages leave no offspring and die out, while others are successful and leave many descendants. If we look at all genes in the population at the present day, then we can trace their genealogy back through that of previous generations. Looking back in time, the number of distinct ancestors to the current population is reduced as genes in the current population find their common ancestors. Eventually, we find a single gene in an ancestor from which all genes in the current population are found. This individual is called the most recent common ancestor (often written as the MRCA).

However, when we look at population genetic data, we do not sample every chromosome in the population, we take a sample. So we can think of the history, or

genealogy, of our sample as being embedded in the broader genealogy of the entire population. Indeed for the purposes of interpreting patterns of genetic variability in our sample, the only mutations that will influence our sample are those that occurred in the ancestral chromosomes. Mutations that occurred in non-ancestral chromosomes will leave no trace in the sample (assuming mutations are neutral).

The idea that we need only consider events in the genes that are ancestral to those in the sample is formulated in coalescent theory (Kingman, 1982; Hudson, 1990). The central idea behind coalescent theory is that the vast majority of events that have occurred in the history of the population have left no impression on the genes in our sample, so we might as well not bother with trying to model them. Rather, we can just trace the lines of descent back through ancestral populations, until they find their common ancestors - or coalesce.

2.2 The coalescent

It is worth being more formal about the foundations of coalescent theory, mainly so that we are aware of when its assumptions might break down. The key ideas behind the theory are:

- Mutations are neutral, therefore have no influence on the probability that an individual reproduces
- If the allelic state of a population has no influence on the reproductive success of chromosomes that are ancestral to our sample, we need only consider the genealogy of genes that are ancestral to our sample
- We can model the genealogical process in a probabilistic manner by considering the times to common ancestor events in WF populations.

To illustrate these ideas, we will start by thinking about the coalescent process for a sample of two genes. In one generation, in a WF population, the probability that two chromosomes, picked at random from the population, came from the same

ancestral chromosome, is the probability that they came from the same parent (remember that we are allowing for self-fertilization) times the probability that they can from the same chromosome in that individual. These probabilities are $1/N$ and $1/2$ respectively (in the WF model) so the probability of a coalescent event is just $1/2N$ and the probability of coming from different chromosomes in the parental population is one minus this quantity. What about the next generation back in time? If a coalescent event occurred, we need not go further back in time, but if the two genes did not coalesce, we are left with exactly the same situation as we had at the beginning, with exactly the same probabilities. So the probability of coalescing t generations ago is the probability of not coalescing for the first $t - 1$ generations, times the probability that a coalescent event occurred in the t th generation.

$$P(\text{coalesce at } t) = \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N} \quad (1)$$

The distribution of coalescence times is geometric; the average is $E[t] = 2N$, but the distribution has a long tail of large coalescent times, such that 63% of all coalescence times are less than the average.

The coalescent creates a genealogical tree (albeit a rather simple one in this case). We can now consider how this tree relates to the genetic differences we expect to see between a pair of sequences sampled at random, by iterating our favourite mutation model along the tree, adding mutations by some stochastic process to generate a possible sample from the population. The conventional mutation model associated with the coalescent process is the infinite-sites model. The probability of a mutation occurring in a single generation is μ (this is the mutation rate for the entire gene), so the number of mutations occurring during t generations is Poisson distributed with mean μt (the Poisson nature results directly from the independence of mutations). The expected number of differences between two sequences (often called diversity) is therefore the product of the expected time to coalescence and the mutation rate

$$E[\pi] = 4N\mu \quad (2)$$

In population genetics, the mutation rate almost only ever enters equations in its product with the population size. For this reason, and also because it is the single most important quantity in population genetics, the product is usually written as a single parameter $\theta = 4N\mu$.

2.3 Effective population size

Because there is a simple, linear relationship between the population size and the number of differences you expect to see between two chromosomes sampled at random from a WF population, you might hope to estimate the mutation rate by comparing levels of diversity in species of different population sizes (assuming a constant mutation rate). However, empirical studies of variability in a wide range of species show levels of diversity vary considerably less than census population sizes. For example, diversity in humans is $1/10^{th}$ that of *Drosophila melanogaster*, but for every human there must be at least 1000 *Drosophila*.

Clearly, the assumptions of the WF model are incorrect. Some ugly piece of biology is getting in the way. However, we have no desire to reject the model outright, because it leaves us with no alternative framework for interpreting patterns of genetic variability. Can we find some way of keeping the simplicity of the model structure, but introducing important biological complications?

The solution is so subtle that it is almost semantic. It turns out that the effect of many biological complications on genetic variability, such as variation in reproductive success, fluctuations in population size over time, and differences in the number of breeding males and females, is to make the population behave as if it were still our naive WF model, but one in which the population size is smaller than the true census population size. This transformed population size is called the effective population size and is written as N_e . Consequently, the amount of diversity we expect to see in populations is $E[\pi] = 4N_e\mu$ and θ is really defined as $\theta = 4N_e\mu$.

You may be feeling a little queasy by this piece of algebraic backhand. No one

really knows why diversity varies so much less than census population size; some think the cause is recurrent adaptive evolution (Gillespie, 2000), which would have a large impact on coalescent theory if true. Our stance for the moment will be that using N_e rather than N seems to solve one nasty problem, so we shall use it, but we should also be on the lookout for ways of testing the idea.

2.4 Variation in coalescence times

With this slightly revised view of the coalescent, we can finish off coalescent theory for two sequences by considering the variation one might expect to see in pairwise differences. There are two sources of stochasticity; variation in coalescence time and variation in the mutation process.

To derive the full distribution of pairwise differences, we first need to introduce a trick that is a feature of coalescent theory. If the effective population size is very large, then the rate of coalescence is slow and, if we blur our eyes a bit, we can pretend that time is continuous rather than discrete (formally, we rescale time in units of $\tau = t/2N_e$, and let $N_e \rightarrow \infty$). In this limit the geometric distribution of coalescence times becomes exponential; $\phi(\tau) = e^{-\tau}$. The probability of observing d differences between two sequences is given by the integral over coalescence times of the probability of d mutations given τ from the Poisson mutation process

$$\begin{aligned}
 P(d \text{ mutations}) &= \int P(d|\tau)\phi(\tau)d\tau \\
 &= \int \frac{e^{-\theta\tau}(\theta\tau)^d}{d!}e^{-\tau}d\tau \\
 &= \left(\frac{1}{1+\theta}\right)\left(\frac{\theta}{1+\theta}\right)^d
 \end{aligned} \tag{3}$$

This is also a geometric distribution, and consequently has considerable variance. Examples of the distribution with $\theta = 0.1, 1$ and 5 are shown. For $\theta = 5$ the distribution is extremely broad, and there is a considerable probability of observing either 0 or 10 or more differences in a pairwise comparison.

3 Properties of the n -coalescent

We now wish to consider the coalescent for more than two sequences - a process known as the n -coalescent. The mechanics of the process are very similar, but it is important to be aware of the assumptions that underlie the theory. The critical assumptions are

- Lineages coalesce independently
- No more than one coalescent event can occur in a single generation
- The time-scale is so large that it can be represented as continuous

None of these assumptions are particularly restrictive, and are equivalent to saying that the sample is small selection from a very large population. In other words, coalescent theory is applicable to almost all situations, except when the sample size represents a considerable fraction of the entire population (e.g. for captive populations or very rare species). Under these circumstances, it is likely that more than one coalescent event occurs in recent generations.

If these assumptions are not violated, it becomes a simple matter to work out both the distribution of times to coalescence, and properties of the number of mutations in a sample. In any generation, the probability of any given pair of lineages coalescing is $1/2N_e$ and there are $n(n-1)/2$ such pairs. So the probability of any one of the pairs coalescing is

$$P(\text{Coalescence}) = \frac{n(n-1)}{4N_e}$$

and the expected time to coalescence is this reciprocal of this; $E[T_n] = 4N_e/n(n-1)$. During the time while there are n lineages, the expected number of mutations that occur is $n\mu E[T_n]$, so the expected number of mutations contributed by the time when there are n lineages is $\theta/(n-1)$.

What happens after the first coalescence? A key property of the coalescent process is that the only factor influencing the probabilities of mutation and coalescence is the current state of the sample. So the coalescent process for a sample

of size n after the first coalescent event is identical to the coalescent process for a sample of size $n - 1$. This lack of memory is called Markovian, and is critically important, because it means in order to follow the fate of lineages back in time, all we need know is the number of lineages currently active in the population.

3.1 How much variability?

The Markovian property of the coalescent means that the times between successive coalescent events are independent of one another, and the expected number of mutation events in the entire history of the sample is the sum of the expected number of mutations for each step in the coalescent history. Writing S for the number of segregating sites in the sample, under the infinite-sites model, the expectation is

$$E[S] = \theta \sum_{i=1}^{n-1} 1/i \quad (4)$$

This result was actually first derived by Watterson (1975) without using coalescent theory. The expected time (in units of $2N_e$ generations) until all lineages have coalesced into a single ancestor (the time of the MRCA) is

$$E[T_{MRCA}] = 2 \left(1 - \frac{1}{n} \right) \quad (5)$$

The chart shows how the expected number of segregating sites and the expected time until the MRCA increase with sample size. Two things are of note. First, as the sample size increases, the expected number of segregating sites increases, but at a decreasing rate. As a consequence, doubling the sample size does not double the number of segregating sites you expect to find. Second, the time until the MRCA very rapidly reaches a nearly constant level, approaching a maximal value of $4N_e$. This is because for anything but very small sample sizes, there is a high probability that the MRCA of the sample is the same as the MRCA of the entire population.

3.2 How much variability in levels of variability?

As with the pairwise differences, the stochastic nature of the coalescent and mutational processes generates considerable variance in the number of segregating

sites that may be observed for any given sample size and value of θ . We can use the properties of compound distributions to obtain the variance in the number of segregating sites (all times are rescaled in units of $2N_e$ generations):

$$\sigma_S^2 = \frac{\theta}{2} E[T_{tree}] + \frac{\theta^2}{4} \sigma_{T_{tree}}^2$$

where T_{tree} is the total time in the tree (see any probability text book for how this result is derived using generating functions). Because the coalescence times are independent, the variance in the total time is the sum over coalescent events when there are i lineages

$$\sigma_{T_{tree}}^2 = \sum_{i=2}^n i^2 \sigma_{\tau(i)}^2$$

Where $\sigma_{\tau(i)}^2$ is the variance in the coalescent time when there are i lineages, which is $1/i^2(i-1)^2$. Putting the elements together gives

$$\sigma_S^2 = \theta \sum_{i=1}^{n-1} 1/i + \theta^2 \sum_{i=1}^{n-1} 1/i^2 \quad (6)$$

Again, Watterson (1975) first derived this result without using coalescent theory. In fact, it is possible to derive the entire probability distribution for the number of segregating sites in a sample without recombination using a recursive method first developed by Tavaré (1984). As we might expect from the case of $n = 2$, there is considerable variation in the number of segregating sites one might expect for a given set of parameter values.

3.3 Other mutation models

The results so far presented have all concerned the infinite-sites model of sequence evolution (Kimura, 1969). However, there are other important models of the way in which sequences change that we may also wish to study.

Much of the early work in population genetic inference used the infinite-alleles model (IAM), which can be used to model protein polymorphisms. In this model all new mutations change the state of the allele to one not currently present in the sample (for example, an electrophoretic variant with a different mobility). Under

this model, the probability that two alleles picked at random from the population are different is $\theta/(1 + \theta)$. The most famous result concerning the IAM model is due to Ewens (1972) who showed that the expected number of different alleles in a sample is

$$E[K] = 1 + \frac{\theta}{1 + \theta} + \frac{\theta}{2 + \theta} + \dots + \frac{\theta}{n - 1 + \theta} \quad (7)$$

Although allozyme studies now play a minor role in population genetics, a novel application of the result is that it also predicts the number of unique haplotypes in a sample (only if there is no recombination).

Another widely used mutation model is the step-wise mutation model (Kimura and Ohta, 1978), for microsatellite mutation. In the model, changes of only 1 repeat unit can occur, at a rate of μ per microsatellite per generation. The natural way of summarizing information on genetic variability is to use the variance in repeat number across samples. Moran (1975) showed that the variance in repeat length is

$$E[\sigma_L^2] = \theta/4$$

without the use of coalescence theory. A coalescent derivation was first produced by Slatkin (1995).

4 Summarizing data

Thinking back to our flow-chart of statistical inference, we are still very much in the phase of examining the properties of our model. What we need to do next is to find ways of summarizing information on the observed patterns of variability.

4.1 Summary statistics

A statistic is broadly defined as any function of the data (for example, the letter of the first base sequenced is a statistic!). However, what we are interested in is capturing the information in a sample, and summarizing it in an efficient and interpretable manner.

A summary statistic is a way of collapsing information from a high-dimension sample space into a single number that is informative about the parameters of the model. If all possible information about a parameter is contained within the summary statistic, it is called a *sufficient statistic*. For example, if data are generated by a Normal distribution, the average of a set of observations contains all the information about the mean of the distribution.

In population genetics, it is very rare to find sufficient statistics. Rather, we have to look for a set of statistics that summarize different aspects of the data, and that bear a intuitive relation to the data. Ideally, we wish to use as few statistics as possible, to maximize the amount of information extracted from the sample, with the minimum of overlap between statistics. Statistics whose expectations can be calculated analytical are particularly useful, as these can be used for parameter estimation. For example, we have already shown that the expectations of the differences between pairs of sequences, the total number of segregating sites, and the number of haplotypes bear a simple relationship to the value of θ . They are therefore a good choice for summary statistics, although when the mutation rate is low, the number of haplotypes will be limited by the number of segregating sites (i.e. there will be much shared information).

4.2 Graphical representation of data

It is very often a great aid to data interpretation to have some graphical means to display information. Because we have been thinking about samples in a genealogical manner, a natural way of summarizing the data is in the form of a tree. Note this only applies when there is no recombination.

However, it is rare that we can draw a single resolved tree from population genetic data. We may have multiple copies of identical genes, or situations where the internal branching order cannot be resolved. One way of representing this uncertainty is to draw trees in which unresolved branches are drawn as multifurcations, and the nodes are defined by shared mutations. The lengths of branches, and the

ages of mutations, can be estimated using coalescent inference (see below). Such representations are known as gene trees (Griffiths, 2001). The idea is very similar to using parsimony, but unlike parsimony, branch lengths, and ages of mutations, are inferred with an explicit model of the genealogical process.

It is worth noting some of the differences between phylogenetic tree reconstruction and drawing gene trees. In most phylogenetic methods, no explicit model of species (or gene) birth and death processes is used; in effect our prior is that all phylogenies are equally likely. In gene trees, we have prior information about the genealogical process that we wish to incorporate into our analysis (an idea that is essentially a Bayesian notion - see later). As an extreme example of when you might get different results from phylogenetic and population genetic methods, consider the case where there are no polymorphic sites in the sample; as occurred in one of the earliest studies on human Y chromosome variation (Dorit et al., 1995). A phylogenetic method would say that the most likely time of the MRCA was yesterday. In contrast, if we incorporate a population genetic model, we get a much more sensible answer (Donnelly et al., 1995).

In cases where there is recombination, or where there is recurrent mutation, there is often no unique tree that is compatible with the data. In such circumstances, it may be possible to represent some of the uncertainty about multiple trees using networks. In particular, split-tree decomposition can generate reticulate structures in the presence of conflicting phylogenetic signals, but will generate well-behaved trees if the data can be accommodated by a single tree. Networks can be a good way of depicting rare recombination events, and homoplasies. However, many different permutations of data can give the same network, and there is no clear relationship between features of the networks and parameters of evolutionary models. For cases where there is considerable recombination, networks are unlikely to improve one's understanding.

A third possible representation to use haplotype plots to represent sequence variation. The idea here is to represent segregating sites as blocks of different

colour. One benefit of such plots is that strong haplotype structure (linkage disequilibrium stretching over multiple markers) becomes obvious by visual inspection. Furthermore, such plots can highlight regions of genes with unusual patterns of variability. For example, in the *LPL* data, haplotype structure appears to change abruptly in the middle of the sequence.

5 Stochastic simulation

So far we have tried to explore properties of the WF model analytically; by deriving expressions for quantities that we can measure. However, often it is not possible to be so elegant, and simple algebraic formulations cannot be found. In such circumstances (usually the case when features such as geographic subdivision and natural selection are included), stochastic simulation of the data is the only possible way of exploring the model (and also estimating parameters).

The ease of simulation is one of the greatest strengths of coalescent theory (Hudson, 1993). Because you only need keep track of chromosomes that are ancestral to the sample, simulations can be run backwards in time, starting from the present and choosing ancestors and times of coalescence. Even if there are a few hundred chromosomes, simulations take a matter of milliseconds on today's computers. By contrast, if you were to try to simulate entire WF populations, the computational burden makes simulating more than a few thousand chromosomes unfeasible. Unfortunately, for some types of problems with natural selection, this is the only possibility.

One way of simulating data that is highly efficient is to jointly simulate the coalescent history and the mutational process (Griffiths and Tavaré, 1994). The simulation process can be summarised as follows

1. If there are n chromosomes, the rate rate of coalescence is $n(n - 1)/2$ and the rate of mutation is $n\theta/2$.
2. Choose a time for the next event by picking an exponentially distributed

random number with rate equal to the sum of the rates of coalescence and mutation.

3. Decide whether the event was a coalescence or a mutation; the probability that the event was a coalescence is $(n - 1)/(n - 1 + \theta)$ and the probability that the event was a mutation is $\theta/(n - 1 + \theta)$.
4. If the event was a coalescence, choose two lineages to coalesce at random, reduce the number of lineages by one. If the event was a mutation, choose a lineage at random to have experienced the mutation and mutate the sequences in the sample to which the lineage is ancestral. Return to step 1.
5. Repeat until there is only one lineage left (the MRCA).

Coalescent simulation can easily be adapted to incorporate important deviations from the WF model, such as geographic subdivision, population growth and bottlenecks, recombination and some types of natural selection. In many ways, the ease of simulation is the single most important aspect of the coalescent. This is particularly important when it comes to estimating parameters.

6 Estimating parameters

By now we have a fairly good understanding of the WF model, and how coalescent theory can be used to predict the behaviour of samples from the population. We can now turn our attention to estimating the parameters of the model; in the simplest formulation (no selection, migration, recombination, etc.) there is only one parameter to the model, θ . In the last section of this lecture we will explore different approaches that one can take to the estimation of this parameter. Very similar considerations apply to the estimation of multiple parameters in more complex models. To illustrate the different ideas we will consider a toy data set of 10 chromosomes simulated under the coalescent model with $\theta = 7$.

6.1 Moment estimators

The simplest manner to estimate a parameter is to equate the expectation of a statistic with its observed value, and to invert the relationship. For the case of no recombination, we have found three different statistics whose expectation is a simple function of θ . We will consider two different ways of estimating the parameter

$$\begin{aligned}\hat{\theta}_\pi &= \frac{2}{n(n-1)} \sum_{ij} \pi_{ij} \\ \hat{\theta}_W &= S/a_1\end{aligned}$$

Where $a_1 = \sum_{i=1}^{n-1} 1/i$ and π_{ij} is the number of differences between sequences i and j . We could also use the number of distinct haplotypes as an estimator of θ , however, as the relationship we derived is only valid for the case of no recombination, it is not a generally applicable estimator. If we consider the toy example, the two different estimators give $\hat{\theta}_\pi = 5.9$ and $\hat{\theta}_W = 5.0$.

What should we make of the two estimates? They are fairly similar (which is good), but neither is that close to the true value. Clearly, we need to be able to provide an expression of how much faith we have in the estimate, and, if possible, suggest a range of values within which the true value probably lies, given our estimate.

In the world of moment estimators, an expression of faith about the accuracy of the estimate comes down to estimating the variance of the estimator. The variance of an (unbiased) estimator is given by

$$\sigma_{\hat{\theta}}^2 = E[(\theta - \hat{\theta})^2]$$

In words, the variance of an estimator is the expected squared difference between the true parameter and the estimated parameter. The variances of the two estimators we are considering were derived by Tajima (1983) and Watterson (1975)

respectively

$$\begin{aligned}\sigma_{\hat{\theta}_\pi}^2 &= \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2 \\ \sigma_{\hat{\theta}_W}^2 &= \frac{1}{a_1}\theta + \frac{a_2}{a_1^2}\theta^2\end{aligned}$$

Where $a_2 = \sum_{i=1}^{n-1} 1/i^2$. For a given value of θ , the variance in the estimator decreases as the sample size increases for both estimators. Clearly, this is a good property (the more data we collect, the better an estimate we get). However, whereas the variance of $\hat{\theta}_W$ tends to zero as our sample size gets infinitely large, the variance of $\hat{\theta}_\pi$ levels off. This means that even if we collected an infinitely large sample of sequences, if we estimated θ using average pairwise differences, we would still have some uncertainty about the true value. In contrast, if we estimated θ from the number of segregating sites, our estimate would converge on the true value with certainty. The property of zero variance for an infinitely large sample size is called *consistency*, and it is a highly desirable property for estimators to have. For this reason, it is better to estimate θ from the number of segregating sites.

In the toy data set, the standard deviations of the two estimators are $\sigma_{\hat{\theta}_\pi} = 3.5$ and $\sigma_{\hat{\theta}_W} = 2.5$. If we wish to convert this variance into a confidence interval (CI), the best we can do is to assume that the estimator is Normally distributed, in which case the 95% CI is $\hat{\theta} \pm 2\sigma$. Unfortunately, for small values of θ , the Normal distribution is a poor approximation, so the estimated 95% CI will not have the correct coverage properties. A better estimate of the 95% CI can be obtained by simulation.

Moment estimators are intuitive and easy to calculate. However, unless the estimators are Normally distributed, it is difficult to translate their variances into sensible confidence intervals. Moment estimation can also become difficult when there are multiple parameters to estimate, because it is often not possible to find a set of parameters which give expectations equal to the values of multiple observed statistics. In a world of Normal distributions, this problem is circumvented by using

least squares estimation. However, in population genetics, statistics are unlikely to follow Normal distributions, and multiple parameter estimation using moment methods is problematic.

6.2 Likelihood

Another way of estimating parameters is to use the concept of likelihood, to which Fisher made many important contributions; see Edwards (1992) for a good introduction to the theory of likelihood. The key idea behind likelihood is that we can represent the compatibility of different parameter values with the data by the probability that exactly the set of data we observed was generated. Our best estimate of parameters is those that give the greatest probability of observing the data; the maximum likelihood estimate.

Actually, rather than deal in exact probabilities, likelihood typically works with relative probabilities, in which case the probability need only be known up to a constant of proportionality. Furthermore, for good statistical reasons, the log of the relative likelihoods is typically used, often called the *score*. Although obvious, it is also worth stating that the use of likelihood requires an explicit probabilistic model of the process of interest.

How does likelihood work in the case of population genetics? Coalescent theory provides a probabilistic description of samples from populations, so we can calculate the probability of observing various aspects of the data. For example, we discussed earlier a recursion derived by Tavaré (1984) which generates the full probability distribution for the number of segregating sites in a sample. We can use this recursion to calculate the likelihood surface for different values of θ . For the toy data set, the maximum likelihood (ML) estimate of θ is 5.2. As before, we also wish to express how much faith to put in our estimate. In likelihood, this is easy. The likelihood curve has an explicit meaning, such that a point whose score is 6.91 units less than that of the ML value is 100 times less likely to have generated the data. Often people will present ML estimates ± 2 -unit intervals. The reason for

doing this comes from the Normal behaviour of the likelihood surface under the asymptotic limit of infinite data. In population genetics, you are never near this limit (as I said before, you really only have a single data point), so application of asymptotic theory should be treated with caution.

We have obtained a ML estimate of θ from the number of segregating sites in the sample. However, by ignoring how these mutations are distributed between chromosomes, we are clearly throwing away information. Given that the coalescent fully describes the genealogical and mutational processes, how can we go about using all the information in the data?

The answer comes back to stochastic simulation. The set of genealogies that can generate a given set of data is enormous, in fact it is infinitely large. So it is theoretically impossible to enumerate all possible sets of histories (genealogies plus mutations) that are compatible with the data, adding up their likelihoods. However, for any given data set, there is a much smaller set of histories that will contribute the majority of the likelihood. If we can bias our search to such histories, we might be able to approximate the true likelihood by something that is very close.

Without going into the details, the idea that we can calculate full likelihoods by simulating histories that are likely to have generated the data is one of the most important areas of research in modern population genetics. The first attempt to implement Monte Carlo estimation of likelihoods was that of Griffiths and Tavaré (1994). Although there is some fairly complex maths involved, the central idea is very simple; starting with the data, you go back in time proposing events (coalescent events and mutations) that are compatible with the data. For each run, you calculate the likelihood and weight this by the probability of choosing that history given your proposal algorithm. Do this many thousands or millions of times for a given set of parameter values, and you have an approximation to the likelihood.

The result of applying this approach to the full-likelihood estimation (i.e. using all the information in the sample) of θ in the toy data set is shown on the same chart as that from using the number of segregating sites. The ML estimate is very similar

($\hat{\theta} = 4.7$), and the 2-unit support interval is slightly smaller (1.9 – 12.2). Two points are worth noting; first we have not gained much more information about θ from using all the data, as opposed to the number of segregating sites. Second, even when we have used all possible information, there is still a large degree of uncertainty about the true value of θ .

What is the point of trying to use full likelihood methods if simpler approaches give us almost the same answers? The answer is that coalescent likelihood methods can be used to analyze many aspects of the data, not just estimate θ . For instance, in many situations, the age of specific mutations is of considerable interest (mutations that define haplotypes specific to a geographical area, or the mutations thought to contribute to disease susceptibility, for example). Given an estimate for θ , the age of mutations can be estimated by looking at the distribution of times at which the mutation of interest occurred in the simulations that were used to estimate the likelihood (Griffiths, 2001). Another aspect of the data we may also be interested in is the time to the MRCA of the sample.

6.3 Bayesian inference

In likelihood analysis we can calculate the relative probabilities that a certain set of parameter values gives a particular sample. However, in many cases, we may have some prior knowledge about parameter values, and it makes sense to try to incorporate this information into the estimation procedure.

Bayesian inference is the branch of statistical theory aimed at incorporating prior information into the estimation procedure. The theory is named after the Revd Thomas Bayes (1702-1761). If we define D as the data and θ as the set of parameter values we wish to estimate, Bayes theorem states

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} \quad (8)$$

In words, the equation states that the probability of the parameters given the data, $P(\theta|D)$, is equal to the prior probability of the parameters (obtained from previous

experience), $P(\theta)$, times the likelihood, $P(D|\theta)$, divided by the total probability of the data, $P(D)$, which is obtained by summing the product of the prior and the likelihood over all values of θ .

The critical difference between likelihood and Bayesian inference is that in a Bayesian world, parameters are themselves drawn from a higher level probability distribution, and the method of inference forces us to decide how much we know about parameters *before* the analysis, which in turn tells us how much knowledge we have gained from looking at the data. In likelihood, parameters have fixed, unique values (which we assume we know nothing about), and the goal is to represent belief in the different possible values given the data. There are many deep and philosophical arguments between adherents of the two schools of thought. In this course, we will take a pragmatic approach, and use Bayesian inference when it is reasonable to think we have prior knowledge of some parameters of interest.

By way of example, suppose we actually know something about the effective population size and mutation rates in humans, and wish to use this information to inform our estimation of θ in the toy example. We will say that N_e in humans is somewhere between 5,000 and 50,000, but that within that interval all values are equally likely. For the mutation rate, which can be estimated from interspecific divergence, we will say that our prior knowledge can be represented by a log-normal distribution with mean 1.5×10^{-9} per site per generation and standard deviation of 1 log units. The combined information gives a prior for θ which has a maximum at 10. When combined with the likelihood calculation from *GENETREE*, the posterior distribution of θ has maximum at 7.9 and the middle 95% weight of the distribution lies between 5.0 and 14.4. Clearly, the incorporation of prior information can have a large impact on the estimation of parameters.

6.4 Strengths and weaknesses of coalescent theory

As you will probably have gathered, coalescent theory is an immensely powerful way of looking at population genetic data. In that it considers only the history of

chromosomes ancestral to the sample, it is very efficient. It is fast and easy to simulate from, so that we can build up an understanding of the properties of the WF model, even when analytical results cannot be obtained. As you will learn about in the next few lectures, it is very flexible, and can readily incorporate biologically important complications of the basic model such as recombination, population growth, bottlenecks, geographical structure and some forms of natural selection (selective sweeps and balancing selection). Furthermore, full likelihood analysis using coalescent theory uses all possible information in the data, and can be used to estimate the ages of mutations and the times to common ancestry for samples, or subsamples.

Are there any weaknesses of the theory? The only important weakness is that when the fate of lineages depend on their allelic state, i.e. there is natural selection, and the strength of selection is fairly weak, such that genetic drift is also important, coalescent theory cannot be used, but see (Neuhauser and Krone, 1997). In addition, full likelihood analysis can be very computationally intensive (although this is an inherent feature of the population genetic process), and devising efficient algorithms is a challenging task (Stephens and Donnelly, 2000; Fearnhead and Donnelly, 2001).

References

- Donnelly P, Tavaré S, Blading DJ, Griffiths RC (1995). Estimating the age of the common ancestor of men from the ZFY intron. *Science* 272:1357–1359.
- Dorit RL, Akashi H, Gilbert W (1995). Absence of polymorphism at the ZFY locus on the human Y chromosome. *Science* 268:1183–1185.
- Edwards AWF (1992). *Likelihood* (2nd ed.). Baltimore: John Hopkins University Press.

- Ewens WJ (1972). The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* 3:87–112.
- Fearnhead P, Donnelly PJ (2001). Estimating recombination rates from population genetic data. *Genetics* 159:1299–1318.
- Gillespie JH (2000). Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* 155:909–919.
- Griffiths RC (2001). Ancestral inference from gene trees. In P. Donnelly (Ed.), *Genes, Fossils and Behaviour. An Integrated Approach to Human Evolution*, Volume 310 of *NATO Science Series: Life sciences*, Japan. Ohmsha.
- Griffiths RC, Tavaré S (1994). Simulating probability distributions in the coalescent. *Stochastic Process. Appl.* 13:235–248.
- Hudson RR (1990). Gene genealogies and the coalescent process. In D. Futuyama and J. Antonovics (Eds.), *Oxford Surveys in Evolutionary Biology*, Volume 7, pp. 1–44. Oxford University Press.
- Hudson RR (1993). The how and why of generating gene genealogies. In A. G. Clark and N. Takahata (Eds.), *Mechanisms of Molecular Evolution*, pp. 23–36. Tokyo: Japanese Scientific Societies Press.
- Kimura M (1969). The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations. *Genetics* 61:893–903.
- Kimura M, Ohta T (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci. U.S.A.* 75:2868–2872.
- Kingman JFC (1982). The coalescent. *Stoch. Proc. Appl.* 13:235–248.
- Moran PAP (1975). Wandering distributions and the electrophoretic profile. *Theor. Pop. Biol.* 8:318–330.

- Neuhauser C, Krone S (1997). The genealogy of samples in models with selection. *Genetics* 154:519–534.
- Slatkin M (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139:457–462.
- Stephens M, Donnelly P (2000). Inference in molecular population genetics. *J. R. Statist. Soc. B* 62:1–30.
- Tajima F (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.
- Tavaré S (1984). Line-of-descent and genealogical processes, and their applications in population genetics processes. *Theor. Pop. Biol.* 26:119–164.
- Watterson GA (1975). On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* 7:256–276.