According to Theorem 4 in Schweinsberg (2003), if $\mu > 1$ and

$$P(X_1 > k) \sim Ck^{-a},\ k \to \infty \tag{7.24}$$

for some finite constants $C > 0$ and $a > 0$, then the weak convergence (7.21) holds with the coalescent limit R_t, depending on the parameter value a:

- If $a \geq 2$, the limit R_t is the Kingman coalescent.
- If $1 \leq a < 2$, the limit is the Beta$(2-a, a)$-coalescent and, in particular, if $a = 1$, it is the Bolthausen–Sznitman coalescent already mentioned. If $1 < a < 2$, the timescale N/σ_N^2 is proportional to N^{a-1}, and $N/\sigma_N^2 \sim \ln N$ in the case $a = 1$ (see Box 7.1).
- If $0 < a < 1$, the limit process belongs to a certain one-parameter class of coalescent processes with simultaneous multiple mergers.

7.2 Ancestral Inference in Branching Processes

S. Tavaré

7.2.1 Introduction

The topic of inference for branching processes is classic and many articles and books have been devoted to it. Common themes include estimation of the offspring mean, the offspring distribution, and the age of the process (cf. Stigler 1970; Guttorp 1991, 1995). In this subsection we illustrate some computational approaches to ancestral inference for branching processes when the effects of mutations among individuals in the population are taken into account. Our examples are from population genetics (in which the timescale is of the order of thousands of years) and from cancer biology (in which the timescale is of the order of years). The techniques illustrated here are but the tip of the inferential iceberg, but they serve to illustrate the crucial interplay between the simulation of a stochastic model and any inference about its parameters.

7.2.2 Inference in the coalescent

Coalescent trees. In Section 7.1 the *coalescent* was introduced as a model for ancestral relationships among a set of chromosomal segments sampled from an evolving population. In the case of a population that has a constant but large number N of chromosomal segments, we showed that when time is measured in units of N generations, the coalescent tree of a sample of n segments can be described as follows. We begin with n tips and wait for an amount of time T_n that has an exponential distribution with mean $2/n(n-1)$ time units before choosing at random two of the tips to coalesce. The coalescent tree now has $n-1$ nodes (which corresponds to $n-1$ ancestors of the sample), and we then wait a further time T_{n-1} that has an exponential distribution with mean $2/(n-1)(n-2)$ time units until, once again, choosing at random two of the nodes to coalesce. We can continue this description using mutually independent exponential random variables, the waiting time while there are j ancestors of the sample having a mean of $2/j(j-1)$ time units. Eventually, the segments in the sample can be traced back to a common

Tavaré S (2005). Ancestral Inference in Branching Processes. In: Branching Processes: Variation, Growth, and Extinction of Populations, Haccou P, Jagers P & Vatutin VA, pp. 208–217. Cambridge University Press.
© International Institute for Applied Systems Analysis

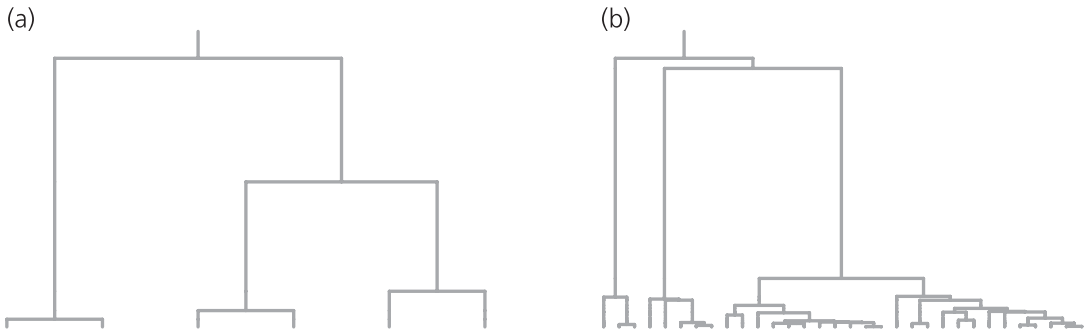

Figure 7.2 Coalescent trees for samples of size (a) 6 and (b) 32 from a population of constant size.

ancestor. Figure 7.2 shows two simulated coalescent trees for samples of size $n = 6$ and $n = 32$.

The height of the coalescent tree, which is the time to the most recent common ancestor (T_{MRCA}) of the sampled segments, is

$$W_n = T_n + T_{n-1} + \cdots + T_2 \,, \tag{7.25}$$

and the length of the tree is

$$L_n = nT_n + (n-1)T_{n-1} + \cdots + 2T_2 \,, \tag{7.26}$$

the sum of the lengths of all the branches in the tree. The means of W_n and L_n are given by

$$\mathbb{E}[W_n] = 2\left(1 - \frac{1}{n}\right), \quad \mathbb{E}[L_n] = 2\sum_{j=1}^{n-1} \frac{1}{j} \,, \tag{7.27}$$

these being multiplied by N to convert coalescent time into generations.

Mutation in the coalescent. The variation observed in the chromosomal segments in the sample is a consequence of mutation in the ancestry of the sample. There are many models for the effects of such mutations, depending on the type of data under consideration. In this subsection we use the so-called *infinitely-many-sites model*, the simplest description of variation in a set of DNA sequences. We suppose that mutations occur only at locations in the DNA segment at which mutations have not occurred before. The sequences in the sample then exhibit a number of *segregating sites*, positions in the DNA at which the members of the sample are not identical. In modern parlance, such locations are called single nucleotide polymorphisms (SNPs). A consequence of this description is that each mutation which occurs in the ancestry of the sample results in a SNP.

The rate at which mutations occur in the region is determined by the compound parameter θ, defined by

$$\theta = 2Nu \,, \tag{7.28}$$

where u is the mutation probability in the region per segment per generation. Mutations are superimposed on the coalescent tree of the sample according to Poisson

processes of rate $\theta/2$, independently in each branch. It follows that the number S_n of SNPs in the sample has a distribution determined by the length L_n of the coalescent tree; given $L_n = l$, S_n has a Poisson distribution with mean

$$\mathbb{E}[S_n \mid L_n = l] = \theta l/2 \, . \tag{7.29}$$

Inference about θ and W_n. In this subsection we illustrate a computational technique to simulate observations from the posterior distribution of (θ, W_n), given that $S_n = k$. To do this, set $T = (T_n, T_{n-1}, \ldots, T_2)$ and note that

$$\begin{aligned} f(\theta, T \mid S_n = k) &\propto \mathbb{P}(S_n = k \mid \theta, T)\pi(\theta, T) \\ &= \mathrm{Po}(\theta L_n/2)\{k\}\, \pi(\theta, T) \, , \end{aligned} \tag{7.30}$$

where we define

$$\mathrm{Po}(\lambda)\{k\} = e^{-\lambda}\frac{\lambda^k}{k!} \tag{7.31}$$

with $\mathrm{Po}(0)\{0\} = 1$. In Equation (7.30), $\pi(\theta, T)$ denotes the prior distribution of (θ, T), which is typically the product of the prior π for θ and the "prior" for T, determined by the coalescent model. The prior for θ can be used to incorporate known information about θ. For example, in many problems the size of the mutation rate u in Equation (7.28) is known, at least approximately, as is that of N. This information can be used to design the prior π. A common alternative is to use an uninformative prior for θ, in the form of a density uniform over an interval.

In practice, the density implicit in Equation (7.30) is hard to evaluate in a useful form and it is much simpler to simulate observations from the distribution instead. This is achieved readily by the rejection algorithm:

- A1. Simulate θ from $\pi(\theta)$ and $t = (t_n, \ldots, t_2)$ from the coalescent model, and calculate $l = nt_n + \cdots + 2t_2$;
- A2. Accept (θ, t) with probability

  $$h = \mathrm{Po}(\theta l/2)\{k\} \, , \tag{7.32}$$

  and return to A1.

Accepted observations clearly have the required density, as can be seen by simple calculation. We make three observations about this approach.

First, it is more efficient to replace h in A2 by h/c where

$$c = \max_{\theta, l} \mathrm{Po}(\theta l/2)\{k\} = \mathrm{Po}(k)\{k\} \, , \tag{7.33}$$

which can result in considerable gains of speed.

Second, it is not necessary to compute the probability h in A2. Instead, the number of mutations k' on the tree of length l can be simulated, and A2 replaced with

- A2′. Accept (θ, t) if $k' = k$.

In this example, h can be computed easily so this simulation-based approach is not necessary. However, the alternative approach is far more general than the first

Table 7.1 Inference about θ and W for Yakima data based on 5000 simulated values.

	$T_{MRCA}W$	Mutation rate θ
First quartile	1.05	0.019
Mean	1.68	0.024
Median	1.46	0.023
Third quartile	2.07	0.029

because the likelihood h does not need to be known (in theory or computationally) to use the method. Note, though, that the gains in speed mentioned in the first observation do not seem to be available in this approach.

Third, there is no need to restrict the algorithm to a coalescent with constant size. All that is required to handle the case of deterministic fluctuations in population size is to simulate from the appropriate coalescent distribution for T. In a similar way, we can simulate observations from the posterior distribution of the coalescent topology (and not just the branch lengths); all that is required is to simulate a coalescent tree and proceed as before. Many other applications of these and related algorithms can be found in Tavaré *et al.* (1997).

Example 7.2 To illustrate these ideas, we use some molecular data obtained as part of a larger study on mitochondrial variation observed in Amerindian populations in the USA (Ward *et al.* 1991; Shields *et al.* 1993). Among the aims of this study was the development of methods to infer population history from DNA sequence variation, and in particular to gain an understanding of the way in which the Americas were settled. A convenient place to read more about this field of research is the 2 March 2001 issue of *Science*.

The particular data we use for illustration here comprise a set of $n = 42$ Yakima mitochondrial DNA sequences, each of length 360 base pairs, given in Shields *et al.* (1993; see also Markovtsova *et al.* 2000a, p. 404). The observed base frequencies in the sequences are

$$(\pi_A, \pi_G, \pi_C, \pi_T) = (0.328, 0.113, 0.342, 0.217)\ , \tag{7.34}$$

and there are 20 distinct sequences and 31 SNPs in the sample.

In the absence of other information, we chose a wide uniform prior for θ, and used a constant population size coalescent to model T. The results of 5000 accepted runs of the algorithm are given in Table 7.1 and Figure 7.3. The posterior distribution of W does not differ enormously from its prior determined by the coalescent model. The parameter N is approximately 600, so if we assume a generation time of 20 years, the mean height of the coalescent tree is about 20 000 years.

◇ ◇ ◇

7.2.3 Approximate Bayesian computation

The Yakima data used in Example 7.2 have been discussed in a coalescent framework by Markovtsova *et al.* (2000a, 2000b), where the posterior distributions of θ and W_n were found by Markov chain Monte Carlo methods using the full sequence data rather than the summary statistic $S_n = 31$. One reason for basing our inference on statistics such as S_n, rather than the full data, is a practical one: we

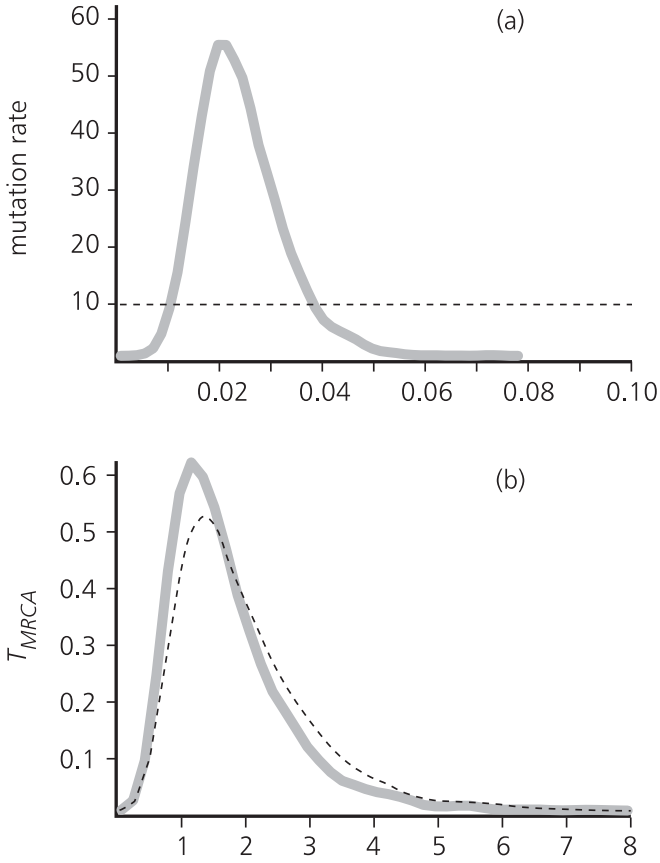

Figure 7.3 Posterior density of θ (a) and W_n (b). Dotted lines show prior density.

hope to generate observations much more quickly than when using other stochastic computation methods. In exchange, we are left with observations from the density $f(\theta, W \mid S_n = k)$ as opposed to the full density $f(\theta, W_n \mid \mathcal{D})$, where $\mathcal{D}$ denotes the complete sequence data. Approaches that use summary statistics for inference are called approximate Bayesian computation (ABC). The consequences of such reductions can be complicated and unexpected; see Beaumont *et al.* (2002) for a number of related examples and other approaches, as well as historical references on ABC.

7.2.4 Inference for tumor histories

In the next example we adapt the same type of approach to a discrete setting that involves inferences about the history of a tumor.

The data and the problem. It is difficult to infer tumor histories by using direct observation of a patient. Adenomas, thought to be precursors of cancer, are removed if they are detected, and the amount of time required to observe the entire progression of a cancer may be many decades. To overcome the limitations of direct observation it is possible to exploit the pattern of mutations observed in an adenoma or a cancer (Tsao *et al.* 2000). These mutations can be used to estimate the age of the adenoma or cancer, in much the same way as we used variation in mitochondrial sequences to infer aspects of the history of the Yakima. The timescale

of the cancer example is of the order of years, in contrast to the Yakima example, which is of the order of tens of thousands of years.

In this example, we study a class of colon cancers known as mutator phenotype cancers. These colorectal cancers have lost DNA mismatch repair (MMR), so they are less able to repair errors during DNA replication. These cancers also have greatly elevated mutation rates. The consequences can be observed most easily in microsatellite (MS) loci. These loci, which may be thought of as runs of a short motif such as CA, show dramatic expansions and contractions in size over small numbers of cell divisions.

It is these mutations that we use to track the history of a cancer. We are able to measure the length variation in a series of such MS loci sampled from cells in a tumor (Tsao *et al.* 2000). The problem is to estimate the time since MMR was knocked out; that is, to estimate the age of the tumor.

Once MMR is lost in a parent cell, the descendant cells derived from it by mitotic division eventually form a final clonal expansion that originates from a single cell and results in a detectable tumor (which we assume has an average size of about 1 cm^3, or about 10^9 cells). Using the MS variation, we estimate the number of divisions Y_0 between loss of MMR and the initiation of the final clonal expansion, and the number of generations Y_1 from that event until the tumor is observed at biopsy.

Once more, this is an "ancestral inference" problem, in which the desired posterior is $f(Y_0, Y_1 \mid \mathcal{D})$. The data $\mathcal{D}$ come from L MS loci, the first of which is measured in n_1 tumor chromosomes, the second from n_2 tumor chromosomes, ..., and the Lth from n_L chromosomes. The total number of chromosomes used is then

$$n = n_1 + \cdots + n_L \, . \tag{7.35}$$

In the studies reported below, we sampled X chromosome MS loci from male patients. As males have a single copy of their X chromosome, we can identify each sampled X chromosome with a single cell. This simplifies the required analysis.

We know the somatic size of each locus (i.e., the number of repeats at each locus prior to loss of MMR). All repeat lengths are measured relative to this baseline size. For each MS locus, we are able to estimate the mean MS lengths, $m_1, m_2, \ldots, m_L$, and the variances of these lengths, $s_1^2, s_2^2, \ldots, s_L^2$. These data are, in turn, summarized by the two statistics

$$\begin{aligned} S^2_{\text{alleles}} &= \text{ average of } s_1^2, \ldots, s_L^2 \, ; \\ S^2_{\text{loci}} &= \text{ variance of } m_1, \ldots, m_L \, . \end{aligned} \tag{7.36}$$

A model for tumor evolution. One question that has to be addressed is the model used to describe the evolution of the tumor from loss of MMR until detection. The relative sizes of S^2_{alleles} and S^2_{loci} give a hint.

In Figure 7.4, taken from Figure 1 in Tsao *et al.* (2000), the results of 1000 simulations of $L = 20$ MS loci measured in $n_i = 25$ chromosomes are summarized (the method used to perform the simulations is given in Section 7.2.5). The

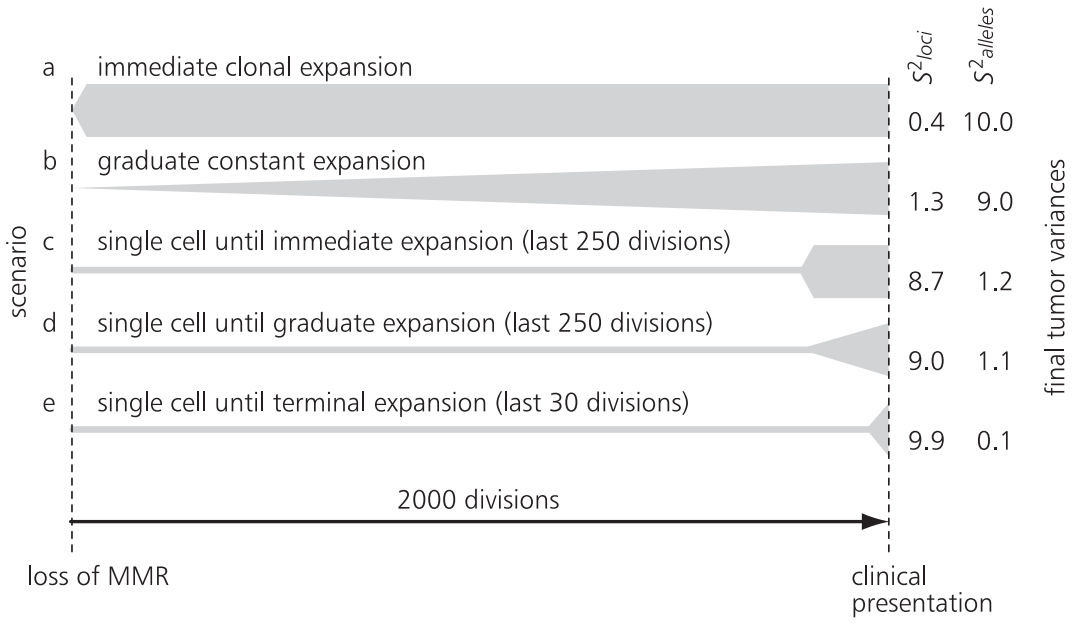

Figure 7.4 Simulations of MS mutation. Different patterns of MS mutations are summarized by the values of S^2_{alleles} and S^2_{loci}. All simulations use 2000 divisions, but different tumor histories. Further details are given in the text. Modified from Tsao *et al.* (2000).

simulations assume a symmetric stepwise mutation model with the chance of the addition and of the loss of one repeat being 0.0025, to give a total mutation rate of 0.005 per division. In each scenario a total of 2000 divisions is assumed, and the final tumor size is, on average, one billion cells.

The results show that it is possible to infer, in broad terms, the form of the tumor history by measuring its MS alleles and estimating S^2_{alleles} and S^2_{loci}. In the analysis that follows, we use the model that corresponds to scenarios (d) and (e) in Figure 7.4: a single progenitor cell lineage that lasts for Y_0 divisions, and a terminal expansion described by symmetric binary splitting for Y_1 generations with parameters chosen to make the average size of the tumor a billion cells. Some experimental justification for this model is given in Figure 2 of Tsao *et al.* (2000).

The genealogy of a sample from a branching process. The data we collect come from a few hundred cells sampled from a tumor that contains about a billion cells. To simulate observations on the MS loci observed in the tumors we could simulate the entire tumor history and then subsample these cells, or we could generate the history of the sample only. The latter approach is the one we used to describe the coalescent: we generate the genealogical history of the cells in the sample, then simulate the effects of mutations at the MS loci in this shared ancestry. This results in a sample of MS loci in the cells in the sample.

Methods to generate the genealogical history of a sample from a branching process are described in Weiss and von Haeseler (1997) in the context of the

polymerase chain reaction (PCR), and in Tsao *et al.* (2000) in the present context. Suppose we want to simulate the history of a sample taken after g generations, going back to time 0. The basic idea is to use three passes to generate the MS sample: in the first phase, the numbers of cells that have 0, 1, and 2 descendant cells in generation 1, generation 2, ... , generation g are simulated. This results in a collection of family-size statistics (M_{j0}, M_{j1}, M_{j2}), $j = 0, 1, \ldots, g-1$, where M_{jl} is the number of families of size l born to cells in generation j. In a given generation, $j+1$ say, there are $M_{j+1} = M_{j1} + 2M_{j2}$ cells. To generate M_{jl} requires knowledge of the offspring distribution in each generation (which may differ across generations). The branching property means that if the total number of cells in generation j is M_j, the number of cells that have 0, 1, and 2 descendants in the next generation is multinomially distributed with parameters M_j and p_0, p_1, p_2, these being the probabilities of 0, 1, or 2 descendants, respectively, from a given cell in generation j.

The second stage reconstructs the genealogy of the sample taken from generation g using the family sizes (M_{j0}, M_{j1}, M_{j2}) in the order $j = g-1, g-2, \ldots, 0$. If the sample has n cells at time g, we assign the n cells at random to ancestors in generation $g-1$, in accordance with the numbers $M_{g-1,1}$ and $M_{g-1,2}$. Using a "balls in urns" analogy, this is equivalent to choosing without replacement n balls from $M_{g-1,1} + M_{g-1,2}$ urns, $M_{g-1,1}$ of which contain one ball, and $M_{g-1,2}$ of which contain two balls. This done, we count the number n_{g-1} of distinct ancestors (i.e., the number of different urns sampled) in generation $g-1$, and repeat the assignment of these cells to their parental cells using the counts $M_{g-2,1}$ and $M_{g-2,2}$. Continuing back in this way to time 1 produces a genealogical tree of the sample.

The third stage starts from the top of this genealogical tree by assigning MS lengths to each of the ancestral cells at time 0, and then runs the mutation process down the branches of the tree until arriving at the n cells in the sample at time g. The mutation mechanism we use here is the simplest of a large number of models that have been used in the literature: a MS locus inherits the same length as its parent, plus the addition or deletion of a single motif caused by errors in MMR.

Before exploiting this approach to infer the age of a tumor, we note that the algorithm used to generate the history of a sample of cells can be adapted to arbitrary branching processes. The branching property is the only key assumption: given the history of the process up to time j, the individuals in generation j produce offspring independently and with identical distributions (which may depend on the history up to time j). In particular, the resultant branching process need not even be Markovian. This provides plenty of flexibility to analyze samples from extraordinarily complicated processes about which theoretical results are few and far between. The approach can also be modified to generate genealogical histories of samples from multi-type branching processes.

7.2.5 Example

This example comes from data at 23 loci measured in an adenoma. The sample sizes at each locus varied between 10 and 33, and the observed summary statistics were $S^2_{\text{alleles, obs}} = 0.828$, $S^2_{\text{loci, obs}} = 6.229$.

As described above, we assumed a simple symmetric step-wise mutation model for each MS, with an overall mutation rate of 0.005 per replication. We used uniform priors for Y_0 and Y_1, with ranges (100, 2100) and (25, 400), respectively.

For the ABC approach we simulated observations from the priors for Y_0 and Y_1, and then simulated the history of the n cells that were sampled. Given this genealogy we simulated L MS loci using the given mutation model, and calculated the simulated values $S^2_{\text{alleles, sim}}$ and $S^2_{\text{loci, sim}}$ of the statistics in Equation (7.36). The values of Y_0 and Y_1 were accepted if

$$\left| \frac{S^2_{\text{loci, sim}}}{S^2_{\text{loci, obs}}} - 1 \right| + \left| \frac{S^2_{\text{alleles, sim}}}{S^2_{\text{alleles, obs}}} - 1 \right| < \epsilon \ , \tag{7.37}$$

where ϵ is a tuning parameter. Large values of ϵ accept most values and so reconstruct the prior, whereas as $\epsilon \to 0$ only those values of Y_0 and Y_1 that reproduce the data S^2_{alleles} and S^2_{loci} exactly are accepted. The trade-off is in picking values of ϵ that lead to a reasonable number of accepted values in a given time, as well as a reasonable approximation to the required posterior. In the example below ϵ is set to 0.1, which corresponds to an acceptance rate of about 0.6%.

In Table 7.2 summary statistics for the posterior distributions of Y_0, Y_1, and the age $Y = Y_0 + Y_1$ of the tumor are given. The corresponding posterior densities, based on 1000 simulated observations, are given in Figure 7.5. The posterior density of Y is shown in Figure 7.6. A 95% credible interval for Y is (895, 2197) divisions. Assuming one division per day, this translates into an interval of (2.5, 6.0) years, and a mean posterior age of 4.2 years.

In Tsao *et al.* (2000) a different statistical approach was used to assess variability in the estimate of Y. Using data combined from two regions of the adenoma, they found an estimated age of 1300 divisions (3.6 years), with an estimated 95% confidence interval of (1.3, 5.2) years. Despite the different statistical approaches these results are consistent with each other.

We remark that the approach outlined here can also be used to investigate the robustness of the modeling assumptions. All that needs to be changed are the details of the branching process being used and the mutation model; some of this is described in Tsao *et al.* (2000). Different statistical approaches can also be explored in this way, such as by using different metrics in Equation (7.37) and different summaries of the data.

Table 7.2 Inference about Y_0 and Y_1 for adenoma data based on 1000 accepted values. $Y = Y_0 + Y_1$ is the age of the tumor.

	Y_0	Y_1	Y
First quartile	1077.0	170.0	1255.0
Mean	1343.9	186.0	1529.8
Median	1325.0	184.0	1514.0
Third quartile	1614.3	200.0	1790.0

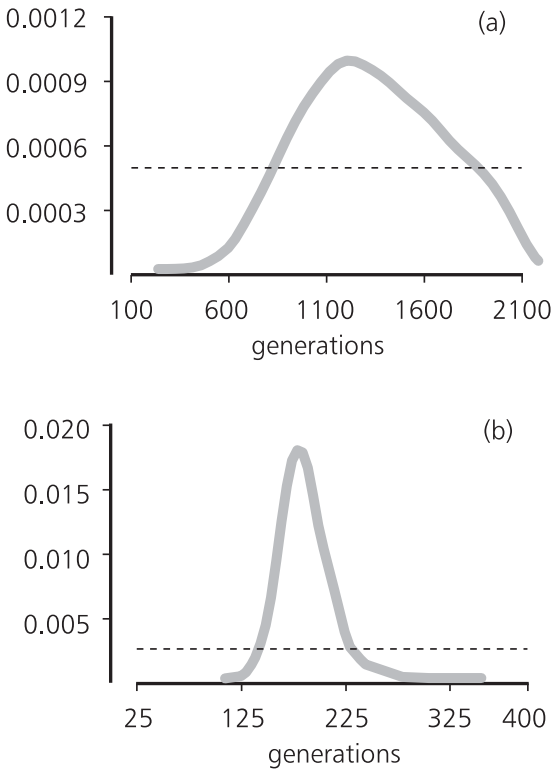

Figure 7.5 Posterior density of Y_0 (a) and Y_1 (b). Dotted lines show prior density.

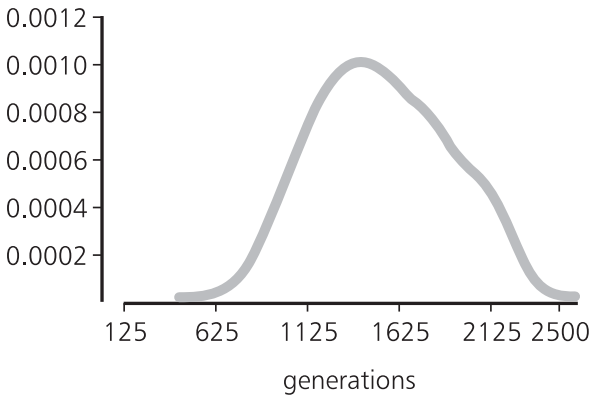

Figure 7.6 Posterior density at age Y of the tumor.

References

References in the book in which this section is published are integrated in a single list, which appears on pp. 295–305. For the purpose of this reprint, references cited in the section have been assembled below.

Beaumont MA, Zhang W & Balding DJ (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**:2025–2035

Guttorp P (1991). *Statistical Inference for Branching Processes*. New York, NY, USA: John Wiley & Sons

Guttorp P (1995). *Stochastic Modeling of Scientific Data*. London, UK: Chapman & Hall

Markovtsova L, Marjoram P & Tavaré S (2000a). The age of a unique event polymorphism. *Genetics* **156**:401–409

Markovtsova L, Marjoram P & Tavaré S (2000b). The effects of rate variation on ancestral inference in the coalescent. *Genetics* **156**:1427–1436

Shields GF, Schmeichen AM, Frazier BL, Redd A, Vovoeda MI, Reed JK & Ward RH (1993). mtDNA sequences suggest a recent evolutionary divergence for Beringian and Northern North American populations. *American Journal of Human Genetics* **53**:549–562

Stigler SM (1970). Estimating the age of a Galton–Watson branching process. *Biometrika* **57**:505–512

Tavaré S, Balding DJ, Griffiths RC & Donnelly P (1997). Inferring coalescence times for molecular sequence data. *Genetics* **145**:505–518

Tsao J, Yatabe Y, Salovaara R, Järvinen HJ, Mecklin J, Altonen LA, Tavaré S & Shibata D (2000). Genetic reconstruction of individual colorectal tumor histories. *Proceedings of the National Academy of Sciences of the USA* **97**:1236–1241

Ward RH, Frazier BL, Dew K & Pääbo S (1991). Extensive mitochondrial diversity within a single Amerindian tribe. *Proceedings of the National Academy of Sciences of the USA* **88**:8720–8724

Weiss G & von Haeseler A (1997). A coalescent approach to the polymerase chain reaction. *Nucleic Acids Research* **25**:3082–3087