Table 1. Observed counts of ABO blood type counts in a British population

Phenotype (i)	Genotype	Probability	Count (n_i)	Frequency (f_i)
A	AA, AO	p^2+2pr	$n_A = 44$	0.26994
B	BB, BO	q^2+2qr	$n_B = 27$	0.16564
AB	AB	$2pq$	$n_{AB} = 4$	0.02454
O	OO	r^2	$n_O = 88$	0.53988
Sum		1	$n = 163$	1

Note.— [Perhaps use instead the following data from Morton (1964) and Yasuda and Kimura (1968), with $n = 2128$ for a Brazilian population: $n_A = 725$, $n_B = 258$, $n_{AB} = 72$ and $n_O = 1073$, with the MLEs $\hat{p} = 0.209131$, $\hat{q} = 0.080801$, and $\hat{r} = 0.710068$. Perhaps include that dataset as an exercise.]

Applications used in the book

Example 1. Estimation of allele frequencies in the ABO blood group

The ABO blood type is determined by the presence or absence of the A and B antigens on erythrocytes. It is controlled by a single gene (the ABO gene) with three alleles: I^A, I^B, and i. Here I stands for isoagglutinogen or antigen, while i means absence of either antigen. For convenience we write the three alleles as A, B, and O. As both A and B alleles are dominant over O, genotypes AA or AO both have the same phenotype (type A), and individuals with BB or BO have type B. At Hardy-Weinberg equilibrium, the genotype and phenotype frequencies are given as functions of the frequencies of the three alleles, p, q, and $r = 1-p-q$. The data, $X = (n_A, n_B, n_{AB}, n_O)$, are counts of the four blood types. The probability of observing the data is given by the multinomial distribution with four categories

$$p(X|p,q) = (p^2+2pr)^{n_A} \cdot (q^2+2qr)^{n_B} \cdot (2pq)^{n_{AB}} \cdot (r^2)^{n_O} \tag{1.7}$$

(see table 1). Our objective is to estimate allele frequencies p and q using the observed blood-type counts in table 1 for a British population (Cavalli-Sforza and Bodmer, 1971).

Heuristics. There are three free observed proportions (f_i) and two unknowns, so one does not expect a perfect match between the expected and observed proportions. As a heuristic estimator, we can let $\hat{r} = \sqrt{f_O} = 0.734765$, then calculate $\hat{p}$ and $\hat{q}$ from f_A and f_B, ignoring f_{AB}. We can check how far away $2\hat{p}\hat{q}$ is from f_{AB}. One can massage the estimates to be more consistent with the observed proportions.

Maximum likelihood. The ML method estimates the parameters by maximizing the likelihood function $L(p,q) = p(X|p,q)$, or its logarithm, i.e., the log-likelihood function

$$\ell(p,q) = n_A \log(p^2+2pr) + n_B \log(q^2+2qr) + n_{AB} \log(2pq) + n_O \log(r^2). \tag{1.8}$$

In theory one can set the derivatives of the log likelihood with respect to p and q to 0 and solve the resulting equations but those are not tractable analytically. Numerical optimization algorithms could be used solve the 2-D optimization problem. Here we use the *gene-counting algorithm*

Table 2. The gene-counting algorithm for the ABO allele frequencies

Round (k)	$h_A^{(k)}$	$h_B^{(k)}$	$p^{(k)}$	$q^{(k)}$	$r^{(k)}$	ℓ
0	0.5	0.5	0.21472	0.13650	0.64877	−181.022
1	0.14199	0.09519	0.16640	0.10298	0.73062	−175.505
2	0.10224	0.06583	0.16104	0.10054	0.73842	−175.449
3	0.09832	0.06374	0.16051	0.10037	0.73912	−175.448
4	0.09795	0.06358	0.16046	0.10036	0.73918	−175.448
5	0.09791	0.06357	0.16045	0.10036	0.73919	−175.448
6	0.09791	0.06357	0.16045	0.10036	0.73919	−175.448

developed by CAB Smith of University College London (Ceppelini *et al.*, 1955, Smith, 1957; see also Yasuda and Kimura, 1968). This is a version of the Expectation-Maximization (EM) algorithm, a broadly applicable algorithm for computing maximum likelihood estimates from incomplete data. EM has been discovered many times in different disciplines, and is particularly widely used in finite-mixture problems (Dempster *et al.*, 1977), but the version developed in genetics appears to be the earliest. Here we describe the algorithm. The algorithm is known to be *non-decreasing*; in other words, the likelihood will only increases but not decrease during the iteration and will converge to the MLE. However here we will not attempt a proof; see Dempster *et al.* (1977).

The idea is that if we knew the genotypes of blood types A and B, we could calculate the allele frequencies by simple counting. Let h_A be the proportion of homozygote AA among individuals having blood type A, and h_B be the proportion of BB among type B. Then

$$\begin{aligned} p &= \frac{1}{2n}\Big[n_{AB} + n_A(1+h_A)\Big], \\ q &= \frac{1}{2n}\Big[n_{AB} + n_B(1+h_B)\Big]. \end{aligned} \tag{1.9}$$

Furthermore, given the allele frequencies $(p,q,r = 1-p-q)$, h_A and h_B can be calculated from their definitions as

$$\begin{aligned} h_A &= \frac{p^2}{p^2+2pr} = \frac{p}{p+2r}, \\ h_B &= \frac{q^2}{q^2+2qr} = \frac{q}{q+2r}. \end{aligned} \tag{1.10}$$

Thus we have an iterative algorithm going through eqs. 1.9 and 1.10. We can start with initial values for (h_A,h_B), calculate (p,q) using eq. (1.9), and then update (h_A,h_B) using eq. (1.10). Repeat until the parameters stabilize. Table 1 shows the progress of the algorithm, with the starting values $(h_A,h_B) = (0.5,0.5)$. We print out the log likelihood as well although the algorithm does not require its explicit calculation. The MLEs are $\hat{p} = 0.16045$ and $\hat{q} = 0.10036$ (with $\hat{r} = 0.73919$), at which the maximized log likelihood (eq. (1.8)) is $\ell(\hat{p},\hat{q}) = -175.44834$. One can also lower the log likelihood by $\frac{1}{2}\chi^2_{2,5\%} = \frac{1}{2}\times 5.99 = 3.00$ to construct a 95% confidence region for p and q.

Bayesian solution. In a Bayesian analysis, we assign the uniform Dirichlet prior

$$f(p,q) = 2, \quad p>0, q>0, p+q<1. \tag{1.11}$$

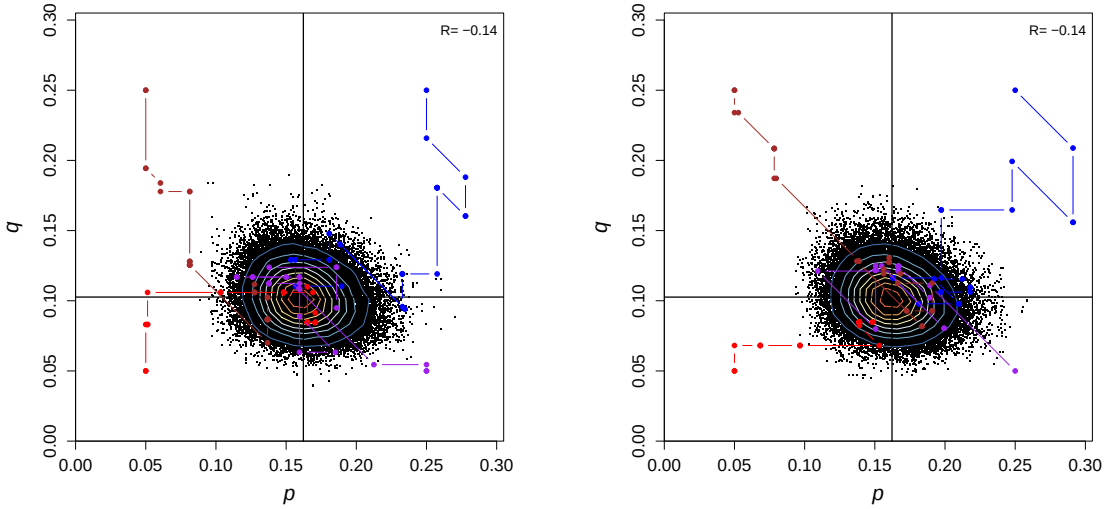

Figure 1. Joint posterior of p and q for alleles A and B in the ABO blood-group example, using (**a**) the random-scan and (**b**) systematic-scan algorithms. The first 30 iterations of four runs with different starting values are show. The density contours are constructed using 2-D kernel-density smoothing using a long run with 10^6 samples.

We consider an MCMC algorithm that consists of three simple steps:

- Step 1 (diagonal move): change p and q with their sum fixed.
- Step 2 (vertical move): change q and r with their sum fixed.
- Step 3 (horizontal move): change r and p with their sum fixed.

The three steps make diagonal, vertical, and horizontal moves, respectively, in the p–q plane. We use the same window size $w = 0.125$ for the three moves. The posterior means and 95% equal-tail intervals are in table 3. Posterior means are 0.1623 0.1026 0.7351 for p, q, and r, respectively.

The acceptance rates are 0.3576, 0.4178, 0.4957 for the three steps. As different parameters have slightly different precisions, use of the same sliding window w in all three steps is not optimal. Acceptance should be the lowest for the first move (since p and q have narrow CIs), and highest for the third move (since r and p have large CIs). In other words, our step is too large for the first move and too small for the third move.

We use either random scan or systematic scan, with the initial state $(p,q) = (0.25, 0.25)$. With the random scan, each MCMC iteration consists of one of the three steps sampled at random. Run time on a PC for 10^6 iterations using the R code was 6.5s. With systematic scan, each MCMC iteration consists of the three steps in a fixed or random-permuted order. Acceptance rates are the same as before. Run time was 15.8s, about 3 times as long as for random scan. Thus systematic scan takes about three times as much computation, and is about 3 times as efficient as random scan.

C program: Running time for 10^8 iterations is 12s for random scan, 27s for systematic scan, and 33s for systematic scan with permutation of the three steps. The C program is about 50 times faster than the R code.

Table 3. Summary of MCMC algorithms applied to the ABO blood-group data

		CI		Efficency		
Parameter	Posterior mean (CI)	width	SD	random	systematic1	systematic2
p	0.1622 (0.123, 0.206)	0.083	0.0212	0.102	0.116	0.114
q	0.1026 (0.071, 0.138)	0.067	0.0172	0.127	0.143	0.141
r	0.7351 (0.682, 0.784)	0.102	0.0254	0.086	0.100	0.099
Run time ($N = 10^6$), R				6.5s	15.8s/3	
Run time ($N = 10^8$), C				12s	27s/3	33s/3

Note.— Systematic2 shuffles the three steps. Efficiency for systematic-scan is calculated by sampling after each step. Thus each iteration of the systematic scan with 3 steps is as efficient as 3 iterations of the random scan. Random permutation of the three steps in the systematic scan had no impact on mixing efficiency. [From this table we should be able to say something about the relative efficiency of $1 \times p$ D move vs. $p \times 1$D moves in a model with p parameters?]

Example 2. Change-point process model for coal mine fatalities data

From Green (1995). A data set that has been frequently used in illustrating new methods for change-point analysis is the point process of dates of serious coal-mining disasters between 1851 and 1962, given by Raftery and Akman (1986). In contrast to some other previous analyses of these data, we will work in continuous time, with the points recorded in days rather than years. Figure 1 displays the dates of the 192 disasters in these 112 years = 40 907 days as a jittered dot plot, together with the cumulative counting process, shown as a dotted line. For data points $i = 1, 2, n$ from a Poisson process with rate given by the function $X(t)$, the log-likelihood is

Example 3. Change point process model for archaeological data

From Brooks (1998). A simple change-point model. To explain, in greater detail, exactly how to formalize statistical problems for use with MCMC algorithms, we follow Buck *et al.* (1993) who used a non-linear regression model with a single change-point, to describe the shape of prehistoric corbelled tombs. Data were collected concerning the shapes of corbelled tombs at sites throughout the Mediterranean. The data consist of a series of measurements of depth from the apex of these tombs, d, and the corresponding radius r; see Buck *et al.* (1993) and Fig. 2. Having obtained data from tombs at different sites, we try to model the data from each tomb separately, to ascertain whether or not different civilizations used similar technologies to construct the tombs, resulting in tombs of similar shapes and therefore models with similar fitted parameters.

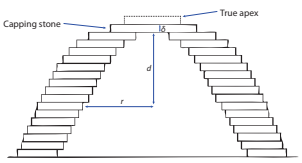

Figure 2. Shape of the inside of an ideal corbelled tomb with a capping stone.