Estimation

Estimation

- All probability models depend on parameters.
  E.g.,
  Binomial depends on probability of success π.
  Normal depends on mean μ, standard deviation σ.
- Parameters are properties of the "population" and are typically unknown.
- The process of taking a sample of data to make inferences about these parameters is referred to as "estimation".
- There are a number of different estimation methods … we will study two estimation methods:

  Maximum likelihood (ML)
  Bayes

Maximum Likelihood

Fisher (1922) invented this general method.

Problem: Unknown model parameters, θ.

Set-up: Write the probability of the data, $\boldsymbol{Y}$, in terms of the model parameter and the data, $P(Y,\theta)$.

Solution: Choose as your estimate the value of the unknown parameter that makes your data look as likely as possible. Pick $\hat{\theta}$ that maximizes the probability of the observed data.

The estimator $\hat{\theta}$ is called the maximum likelihood estimator (MLE).

Maximum Likelihood - Example

Data: Y_i = 0/1 for i = 1, 2,….n (independent)

Model: $Z = \sum_i Y_i$ ~ Binomial(n,π)

Probability: Let's fix the number in the sample at n = 20. The resulting model for Z is Binomial with size 20 and success probability π.

The ***probability distribution function*** is:

$$P(Z;\pi) = \binom{20}{Z} \pi^Z (1-\pi)^{(20-Z)}$$

where Z is the variable and **π is fixed**.

The ***likelihood function*** is the same function:

$$L(\pi;Z) = \binom{20}{Z} \pi^Z (1-\pi)^{(20-Z)}$$

except now π is the variable and **Z is fixed**.

Maximum Likelihood - Example

Two ways to look at this:

- Fix π and look at the probability of different values of Z:

$$\pi = 0.1$$

Z	$P(Z, \pi)$
0	0.122
1	0.270
2	0.285
3	0.190
4	0.090
5	0.032

- Fix Z and look at the probability under different values of π (this is called the likelihood function):

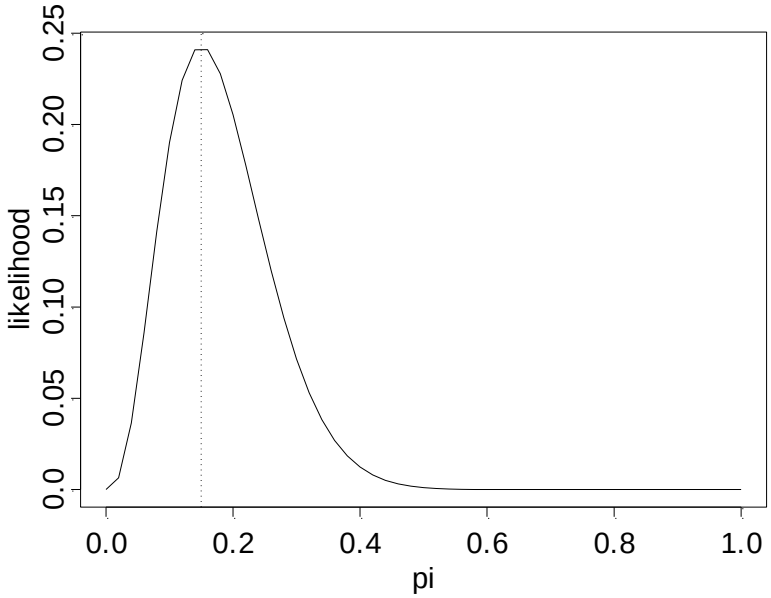
$$Z = 3$$

π	$P(Z, \pi)$
0.01	0.001
0.05	0.060
0.10	0.190
0.20	0.205
0.30	0.072
0.40	0.012

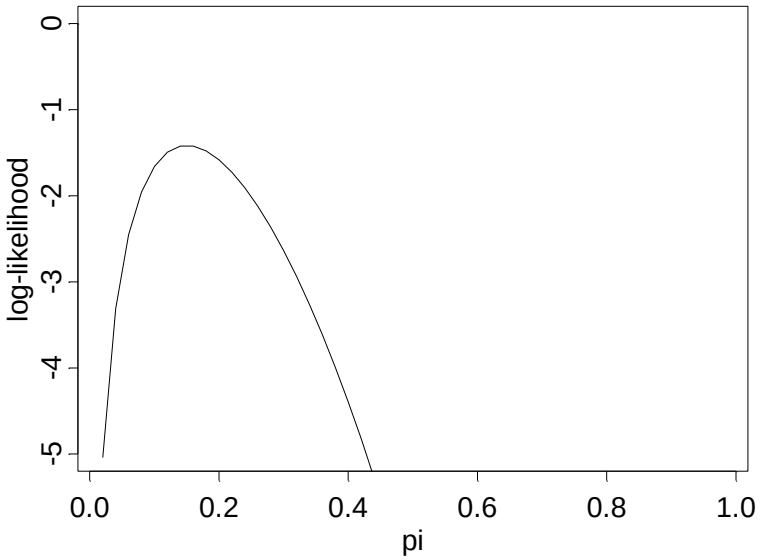
Maximum Likelihood - Example

If you observe the data $Z = 3$ then the likelihood function is shown in the plots below:

P(Z=3) as function of π



log P(Z=3) as function of π



Maximum Likelihood - Example

- We can use elementary calculus (an oxymoron?) to find the maximum of the (log) likelihood function:

$$\frac{d \log L}{d\pi} = 0$$

$$\frac{d}{d\pi} Z \log\pi + (20-Z)\log(1-\pi) = 0$$

$$\frac{Z}{\pi} - \frac{(20-Z)}{1-\pi} = 0$$

$$\hat{\pi} = \frac{Z}{20}$$

- Not surprisingly, the likelihood in this example is maximized at the observed proportion, 3/20.
- Sometimes (e.g. this example) the MLE has a simple closed form. In more complex problems, numerical optimization is used.
- Computers can find these maximum values!

Maximum Likelihood - Notation

$L(\theta)$ = Likelihood as a function of the unknown parameter, θ.

$l(\theta) = \log(L(\theta))$, the log-likelihood.

Usually more convenient to work with analytically and numerically.

$S(\theta) = dl(\theta)/d\theta$ = the "score".

Set $dl(\theta)/d\theta = 0$ and solve for θ to find the MLE.

$I(\theta) = -d^2l(\theta)/d\theta^2$ = the "information".

If evaluated at the MLE, then $-d^2l(\theta)/d\theta^2$ is referred to as the observed information;
$E(-d^2l(\theta)/d\theta^2)$ is referred to as the expected or Fisher information.

$Var(\theta) = I^{-1}(\theta)$ (in most cases)

Maximum Likelihood - Example

$$L(\pi) = \binom{20}{Z} \pi^{Z} (1-\pi)^{(20-Z)}$$

$$\mathrm{l}(\pi) = Z \log(\pi) + (20-Z)\log(1-\pi)$$

$$S(\pi) = \frac{Z}{\pi} - \frac{(20-Z)}{1-\pi} \quad \Rightarrow \quad \pi = \frac{Z}{20}$$

$$I(\pi) = \frac{Z}{\pi^2} + \frac{(20-Z)}{(1-\pi)^2}$$

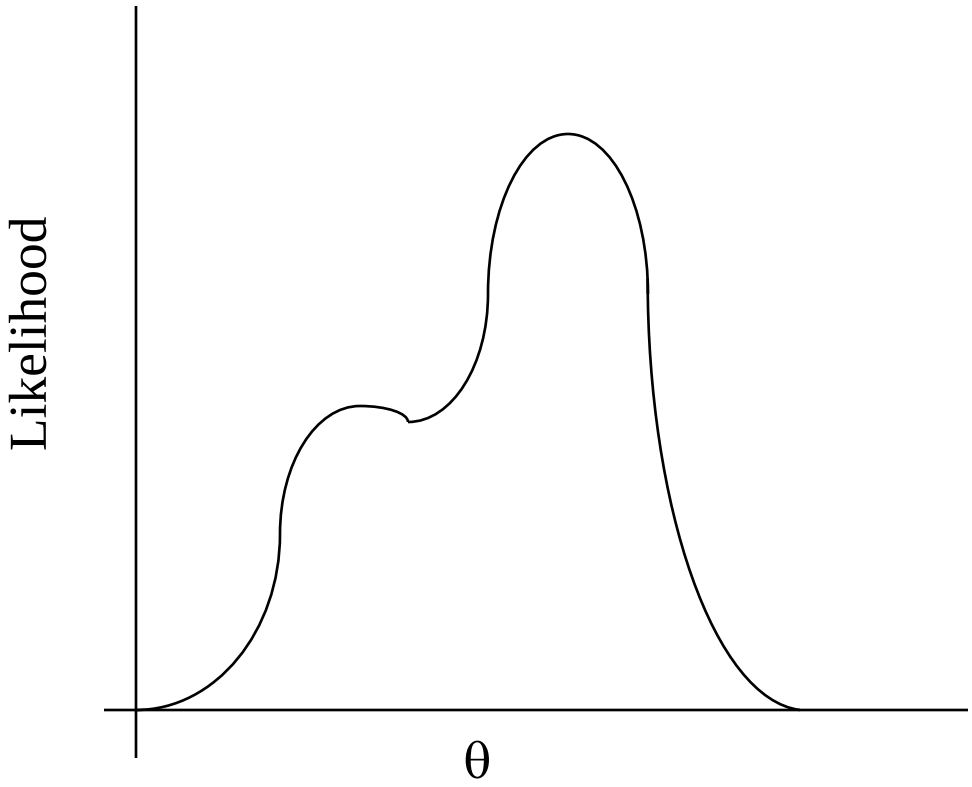
$$E(I(\pi)) = \frac{20\pi}{\pi^2} + \frac{(20-20\pi)}{(1-\pi)^2}$$

$$= \frac{20}{\pi(1-\pi)}$$

(note: constant dropped from $\ell(\pi)$)

Numerical Optimization

- In complex problems it may not be possible to find the MLE analytically; in that case we use numerical optimization to search for the value of θ that maximizes the likelihood
- A common problem with maximum likelihood estimation is accidentally finding a local maximum instead of a global one; solution is to try multiple starting values



Comments:

- Maximum likelihood estimates (MLEs) are always based on a probability model for the data.
- Maximum likelihood is the "best" method of estimation for any situation that you are willing to write down a probability model (so generally does not apply to nonparametric problems).
- Maximum likelihood can be used even when there are multiple unknown parameters, in which case θ has several components

$$(\text{ie. } \theta_0, \theta_1, \text{K}, \theta_p).$$

- The MLE is a "point estimate" (i.e. gives the single most likely value of θ). In lecture 5 we will learn about interval estimates, which describe a range of values which are likely to include the true value of θ. We combine the MLE and Var(θ) to generate these intervals.
- The likelihood function lets us compare different models (next).

Model Comparisons

Q: Suppose we have two alternative models for the data; in each case we use maximum likelihood to estimate the parameters. How do we decide which model fits the data "better"?

A: First thought - compare the likelihoods.

- Larger likelihood is better, but …
- the tradeoff is larger likelihood $\Leftrightarrow$ more complex model.
- How to choose?

A common approach is to "penalize" the likelihood for more complex models (i.e. more parameters).

The AIC and BIC are two examples of penalized likelihood measures.

The LOD ("log odds") score can be thought of as a special case (1 parameter) of a penalized likelihood.

Example – LOD scores

Suppose we have a sample of size N gametes in which the number of recombinants (R) and nonrecombinants (N-R) for two loci can be counted. Let θ be the recombination fraction between the two loci. Then the probability of the data can be modeled using the binomial distribution:

$$P(R) = \binom{N}{R} \theta^R (1-\theta)^{N-R}$$

The situation of no linkage corresponds to θ = 0.5, so we can express the models as

Model 1: θ = 0.5

Model 2: θ anywhere between 0 and 0.5

Example – LOD scores

Model 1: The situation of no linkage corresponds to θ = 0.5. If we substitute this into the likelihood equation, we get

$$\log_{10} L_1 = R\log_{10} 0.5 + (N-R)\log_{10} 0.5$$
$$= N\log_{10} 0.5$$

This model has 0 (free) parameters.

Model 2: The log-likelihood when θ is unrestricted is

$$\log_{10} L_2 = R\log_{10} \theta + (N-R)\log_{10}(1-\theta)$$

This model has 1 parameter.

Taking the derivative and solving for θ gives

$$\hat{\theta} = \frac{R}{N}$$

If we substitute this back into the log-likelihood, we get …

$$\log_{10} L_2 = R\log_{10} {}^{R}\!/_{N} + (N-R)\log_{10}(1-{}^{R}\!/_{N})$$

Example – LOD scores

The LOD score is

$$\text{LOD} = (\log_{10} L_2 - \log_{10} L_1)$$

$$= R \log_{10}\left(\frac{R}{N-R}\right) + N \log_{10}\left(\frac{N-R}{0.5N}\right)$$

Large values of the LOD score (> 3) are considered evidence of linkage (i.e. the penalty is 3).

(As we will see, this is a pretty big hurdle to overcome.)

Example – LOD scores

E.g. N = 50 and R = 18

$\hat{\theta}$ = 18/50 = 36%

$\log_{10}L_1 = -15.0$

$\log_{10}L_2 = -14.2$

LOD = -14.2 – (-15.0) = 0.8

$\Rightarrow$ No evidence of linkage; conclude $\theta = .5$

Model Comparisons – AIC, BIC

AIC – Akaike's Information Criterion

BIC – Bayes Information Criterion

$$AIC = 2\,\lambda(\theta) - 2k$$

$$BIC = 2\,\lambda(\theta) - k\log(n)$$

k = # parameters

(natural logs now)

- Use to compare a series of models. Pick the model with the largest AIC or BIC
- Larger model $\Rightarrow$ larger likelihood (typically)
- Therefore, "penalize" the likelihood for each added parameter
- AIC tries to find the model that would have the minimum prediction error on a new set of data.
- BIC tries to find the model with the highest "posterior probability" given the data
- Typically, BIC is more conservative (picks smaller models)

Model Comparisons – AIC, BIC

Example – Recombinants (N=50, R = 18)

log(L1)= -34.66

(natural logs now)

log(L2) = -32.67

	$\theta = .5$	θ arb	
AIC	-2*34.66 = **-69.32**	-2*32.67 - 2	= **-67.34**
BIC	-2*34.66 = **-69.32**	-2*32.67 - log(50)	= **-69.25**

AIC $\Rightarrow$ pick $\theta = .36$

BIC $\Rightarrow$ pick $\theta = .36$ (but almost tied)

Bayes Estimation

Recall Bayes theorem (written in terms of data X and parameter θ):

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int_{\theta} P(X|\theta)P(\theta)}$$

Notice the change in perspective - θ is now treated as a random variable instead of a fixed number.

P(X|θ) is the likelihood function, as before.

P(θ) is called the *prior distribution* of θ.

P(θ | X) is called the *posterior distribution* of θ.

Based on P(θ | X) we can define a number of possible estimators of θ. A commonly used estimate is the maximum a posteriori (MAP) estimate:

$$\hat{\theta}_{MAP} = \max_{\theta} P(\theta|X)$$

We can also use P(θ | X) to define "credible" intervals for θ.

Bayes Estimation

Comments:

- The MAP estimator is a very simple Bayes estimator. More generally, Bayes estimators minimize a "loss function" – a penalty based on how far $\hat{\theta}$ is from θ (e.g. Loss =$(\widehat{\theta} - \theta)^2$).
- The Bayesian procedure provides a convenient way of combining external information or previous data (through the prior distribution) with the current data (through the likelihood) to create a new estimate.
- As N increases, the data (through the likelihood) overwhelms the prior and Bayes estimator typically converges to the MLE
- Controversy arises when P(θ) is used to incorporate subjective beliefs or opinions.
- If the prior distribution P(θ) is simply that θ is uniformly distributed over all possible values, this is called an "uninformative" prior, and the MAP is the same as the MLE.

Bayes Estimation

Example

Suppose a man is known to have transmitted allele A1 to his child at a locus that has only two alleles: A1 and A2. What is his most likely genotype?

Soln. Let X represent the paternal allele in the child and let θ represent the man's genotype:

$$X = A1$$

$$\theta = \{A1A1,\ A1A2,\ A2A2\}$$

We can write the likelihood function as:

$$P(X \mid \theta = A1A1) = 1$$

$$P(X \mid \theta = A1A2) = .5$$

$$P(X \mid \theta = A2A2) = 0$$

Therefore, the MLE is $\theta = A1A1$.

Bayes Estimation

Suppose, however, that we know that the frequency of the A1 allele in the general population is only 1%. Assuming HW equilibrium we have

$$P(\theta = A1A1) = .0001$$

$$P(\theta = A1A2) = .0198$$

$$P(\theta = A2A2) = .9801$$

This leads to the posterior distribution

$$\begin{aligned} &P(\theta = A1A1 \mid X) \\ &\quad = P(X \mid \theta = A1A1)\, P(\theta = A1A1) / P(X) \\ &\quad = 1 * .0001 / .01 = .01 \end{aligned}$$

$$\begin{aligned} &P(\theta = A1A2 \mid X) \\ &\quad = P(X \mid \theta = A1A2)\, P(\theta = A1A2) / P(X) \\ &\quad = .5 * .0198 / .01 = .99 \end{aligned}$$

$$P(\theta = A2A2 \mid X) = 0$$

So the Bayesian MAP estimator is $\theta = A1A2$.

Exercise: redo assuming the man has 2 children who both have the A1 paternal allele.

Summary

- Maximum likelihood is a method of estimating parameters from data
- ML requires you to write a probability model for the data
- MLE's may be found analytically or numerically
- (Inverse of the negative of the) second derivative of the log-likelihood gives variance of estimates
- Comparison of log-likelihoods allows us to choose between alternative models
- Bayesian procedures allow us to incorporate additional information about the parameters in the form of prior data, external information or personal beliefs.

Problem 1

Suppose we are interested in estimating the recombination fraction, θ, from the following experiment. We do a series of crosses: AB/ab x AB/ab and measure the frequency of the various phases in the gametes (assume we can do this). If the recombination fraction is θ then we expect the following probabilities (sorry, I can't explain these…):

phase	probability (*4)
AB	$3 - 2\theta + \theta^2$
Ab	$2\theta - \theta^2$
aB	$2\theta - \theta^2$
ab	$1 - 2\theta + \theta^2$

Suppose we observe (AB,Ab,aB,aa) = (125,18,20,34). Use maximum likelihood to estimate θ.

Solution to problem 1

$\Pr(\text{data} \mid \theta) \propto (3-2\theta+\theta^2)^{AB}\,(2\theta-\theta^2)^{Ab}\,(2\theta-\theta^2)^{aB}\,(1-2\theta+\theta^2)^{ab}$

$l(\theta) = AB\log(3-2\theta+\theta^2) + (Ab+aB)\log(2\theta-\theta^2) + ab\log(1-2\theta+\theta^2)$

$$\frac{dl(\theta)}{d\theta} = \frac{2AB(\theta-1)}{3-2\theta+\theta^2} + \frac{2(Ab+aB)(1-\theta)}{2\theta-\theta^2} + \frac{2ab(\theta-1)}{1-2\theta+\theta^2} = 0$$

Numerical solution gives θ = .21

$$\frac{d^2l(\theta)}{d\theta^2} = \frac{AB(1+2\theta-\theta^2)}{[3-2\theta+\theta^2]^2} - \frac{(Ab+aB)}{\theta^2} - \frac{ab}{(1-\theta)^2}$$

$$I = E\left(-\frac{d^2\ell(\theta)}{d\theta^2}\right) = -\mathrm{N}*\left(\frac{1+2\theta-\theta^2}{3-2\theta+\theta^2} + \frac{4(1-\theta)}{\theta} + 1\right)$$

= N*16.6

Var(θ) = 1/213.6 = .00468

Problem 2

Every human being can be classified into one of four blood groups: O, A, B, AB. Inheritance of these blood groups is controlled by 1 gene with 3 alleles: O, A and B where O is recessive to A and B. Suppose the frequency of these alleles is r, p, and q, respectively (p+q+r=1). If we observe (O,A,B,AB) = (176,182,60,17) use maximum likelihood to estimate r, p and q.

Solution to problem 2

First, we use basic genetics to find the probability of the observed phenotypes in terms of the unknown parameters. Assuming random mating, we have:

Genotype	prob.	Phenotype	prob.
OO	r^2	O	r^2
AA	p^2		
AO	2pr	A	$p^2 + 2pr$
BB	q^2		
BO	2qr	B	$q^2 + 2qr$
AB	2pq	AB	2pq

$\Pr(\text{data} \mid \theta) \propto (r^2)^O (p^2+2pr)^A (q^2+2qr)^B (2pq)^{AB}$

$l(p,q,r) = 2O\log(r) + A\log(p^2+2pr) + B\log(q^2+2qr) + AB\log(p) + AB\log(q)$

To estimate p, q and r, we need to maximize $l(p,q,r)$ subject to the constraint p+q+r=1. This constraint makes the problem a bit harder …. one approach is to just put r = 1-p-q in the likelihood so we have just 2 parameters … p and q. Then

$$\frac{dl}{dp} = -\frac{2O}{r} + \frac{2Ar}{p(2r+p)} - \frac{2Bq}{q(2r+q)} + \frac{AB}{p} = 0$$

$$\frac{dl}{dq} = -\frac{2O}{r} - \frac{2Ap}{p(2r+p)} + \frac{2Br}{q(2r+q)} + \frac{AB}{q} = 0$$

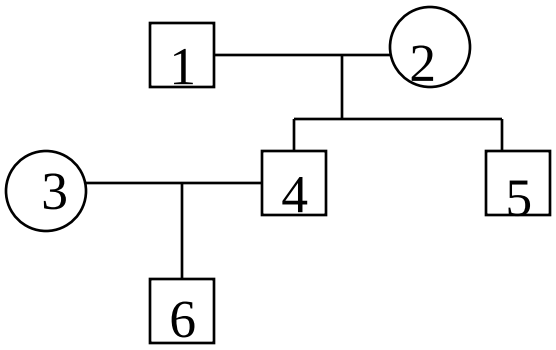
For (O,A,B,AB) = (176,182,60,17), this gives

p = .264 q = .093 r = .642

Further analysis would take 2nd derivatives to find the information and, therefore, the variances of the estimates.

Problem 3

Suppose we have the following simple pedigree.



Define the phenotype of person i as H_i and the genotype as G_{iH} How can we use maximum likelihood to estimate parameters of the *penetrance function*, $\Pr(H \mid G; \theta)$?

Solution to problem 3

- If we knew all the genotypes the problem would be "easy". We would simply write down the log-likelihood and maximize it numerically or analytically:

$$l(\theta) = \sum_i \log \Pr(H_i \mid G_i)$$

- If we don't know the genotypes (only data are the phenotypes), then we must maximize

$$l(\theta) = \log \Pr(H)$$

where H represents the collection of all 6 phenotypes. The general idea is to use the total probability rule to write

$$\begin{aligned}\Pr(H) &= \sum_G \Pr(H \mid G) \Pr(G) \\ &= \sum_{G_1,G_2,G_3,G_4,G_5,G_6} \left\{ \prod_i \Pr(H_i \mid G_i) \right\} \Pr(G_1, G_2, G_3, G_4, G_5, G_6)\end{aligned}$$

Further simplification is achieved by writing

$$\begin{aligned}\Pr(G_1,G_2,G_3,G_4,G_5,G_6) = \Pr(G_6 \mid G_1,G_2,G_3,G_4,G_5) \Pr(G_5 \mid G_1,G_2,G_3,G_4) \Pr(G_4 \mid G_1,G_2,G_3,G_4) \times \\ \Pr(G_3 \mid G_1,G_2) \Pr(G_2 \mid G_1) \Pr(G_1)\end{aligned}$$

Since the genotype of each individual is determined only by his/her parents

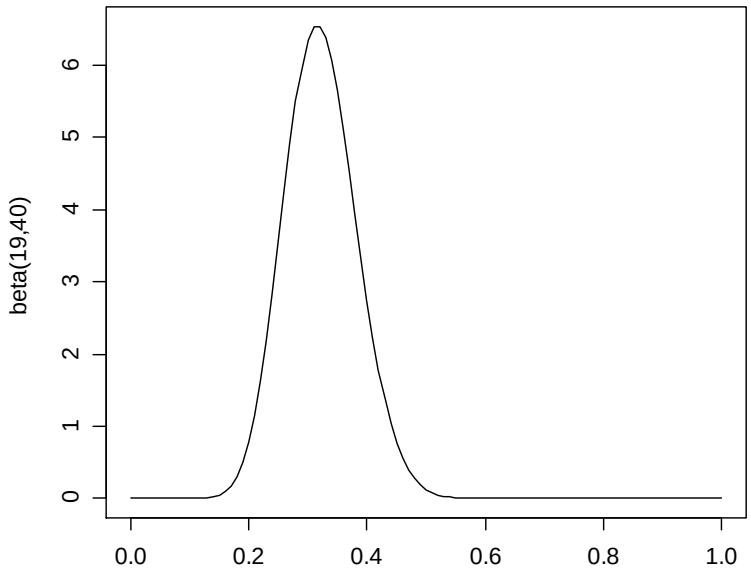
$$\Pr(G_1,G_2,G_3,G_4,G_5,G_6) = \Pr(G_6 \mid G_3,G_4) \Pr(G_5 \mid G_1,G_2) \Pr(G_4 \mid G_1,G_2) \Pr(G_3) \Pr(G_2) \Pr(G_1)$$

Given the inheritance probabilities ($\Pr(G_i \mid G_j, G_k)$) and population frequencies of the genotypes ($\Pr(G_i)$), we have a fully specified model and can maximize the likelihood using a computer.

Problem 4

Suppose we wish to estimate the recombination fraction for a particular locus. We observe N = 50 and R = 18. Several previously published studies of the recombination fraction in nearby loci (that we believe should have similar recombination fractions) have shown recombination fractions between .22 and .44. We decide to model this prior information as a beta distribution (see

http://en.wikipedia.org/wiki/Beta_distribution) with parameters a = 19 and b = 40:



Find the MLE and Bayesian MAP estimators of the recombination fraction. Also find a 95% confidence interval (for the MLE) and a 95% credible interval (for the MAP)

Solution to problem 4

The data follow a binomial distribution with N = 50, R = 18 and the prior information is captured by a beta distribution with parameters a = 19, b = 40:

$$P(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

$$P(X \mid \theta) = \frac{N!}{R!(N-R)!}\theta^{R}(1-\theta)^{N-R}$$

Working through Bayes theorem, we find …

$$P(\theta \mid X) = \frac{\Gamma(N+a+b)}{\Gamma(a+R)\Gamma(N-R+b)}\theta^{a+R-1}(1-\theta)^{N-R+b-1}$$

which is another beta distribution with parameters (a+R) and (N-R+b). The mode of the beta distribution with parameters α and β is $(\alpha-1)/(\alpha+\beta-2)$ so

$$\hat{\theta}_{MAP} = \frac{a+R-1}{N+a+b-2} = \frac{36}{107} = .336$$

Also, we can find the 2.5th and 97.5th percentiles of the posterior distribution (95% credible interval): [.23 - .40]

For comparison the MLE is 18/50 = 0.36 with a 95% confidence interval of [.23 - .49]