Introduction to Bayesian Statistics

Mike Goddard

University of Melbourne and Victorian Institute of Animal Science

Introduction to Bayesian Statistics

The bayesian vs frequentist debate

Heated debates sometimes occur between classical or frequentist statisticians and Bayesian statisticians. These debates are rather philosophical and I will not attempt to resolve them. Instead I will advocate a pragmatic approach which argues that it is sometimes useful to adopt bayesian methods and sometimes frequentist methods.

Bayes theorem

Bayes theorem uses a very simple rule about conditional probabilities:

P(x | y) = P(x and y) / P(y) = P(y | x) P(x) / P(y)

This is best understood with an example. Suppose I have a jar of coins in which 99% are fair coins and 1% are double headed coins. I take a coin at random and toss it 3 times and observe 3 heads. What is the probability that this is a double headed coin?

Let y = the event 3 heads from 3 tosses, x = this is a double headed coin, x' = this is a fair coin. Then P(x) = 0.01, P(y | x) = 1, P(x') = 0.99, P(y | x') = 0.125. All the outcomes of the experiment can be represented in a table:

	P(x or x')	P(y | x or x')	P(y|x)* P(x)
Fair coin	0.99	0.125	0.124
Double headed	0.01	1.0	0.01
Total = P(y)			0.134

Therefore the probability that this is a double headed coin given that I observed 3 heads from 3 tosses is P(x | y) = P(y | x) P(x) / P(y)

= 1.0 * 0.01 / 0.135 = 0.075

That is, despite observing 3 heads, there is still only a small chance that this is a double headed coin because double headed coins are rare.

Bayes theorem is useful because sometimes is easy to calculate P(y | x), but not so easily to calculate P(x | y), as in this case.

Notice that we have calculated the total probability of observing 3 heads, P(y), by adding up the probability of drawing a far coin and then throwing 3 heads plus the probability of drawing a double headed coin and then throwing 3 heads.

We could use Bayes theorem to calculate the probability that the coin is a fair coin
P(x' | y) = P(y | x) P(x') / P(y)
= 0.125* 0.99/ 0.134 = 0.925.
The total probability of observing 3 heads, P(y), is used as the denominator in both calculations. This is a constant in all calculations that we do after the result of the experiment are known, so we can also write Bayes theorem as

P(x | y) is proportional to P(y | x) * P(x).

In other words the odds of the coin being double headed to being a fair coin are 1.0 * 0.01 to 0.125 * 0.99 = 0.01 to 0.124. This is sometimes useful because it is easy to calculate the numerator in Bayes theorem but hard to calculate P(y).

Frequentists agree with and use Bayes theorem. Where they differ from Bayesians is in the situations in which they use it. Bayesians often use the theorem where y is the data observed in an experiment and x is a parameter that they want to estimate. As explained below, frequentists find many such uses of the theorem unacceptable. This is because of the definitions of probability used by the two groups. Frequentists define probability to mean the long term frequency of an event when an experiment is repeated many times. Bayesians allow probability to a subjective statement about how likely you think an event is to occur. Therefore frequentists discriminate sharply between a random variable that can be resampled in every experiment and a parameter that is always the same. They are happy to make probability statements about random variables but not about parameters. Bayesians do not make a sharp distinction between the two.

Estimating a parameter

Frequentist approach

Consider an experiment to estimate the difference in height between men and women in the Australian population. We take a random sample of 10 men and 10 women and measure their heights. We assume a statistical model for this data

$y = u + s + e$

Where

y= height

u= mean height of women

s= the difference in height between men and women in the population

e= an individual's deviation from the average which is assumed to be normally distributed $N(0,\sigma^2)$.

The frequentist estimates s by maximum likelihood. The likelihood of the data is $P(y \mid s)$ considered as a function of s. ML estimation consists of finding the value of s that maximizes $P(y \mid s)$. Call that estimate s-hat. If we do the experiment many times the true value s will always be the same but we will get different values of s-hat. So we can described the distribution of s-hat as $N(s, v^2 = 2\sigma^2/10)$, where v is the standard error of s-hat. Therefore we can make a probability statement about s-hat

$P(s – 2^* v< \text{s-hat} < s+ 2^*v) =0.95$

And we can rewrite this as a confidence interval

$P(\text{s-hat} – 2^*v < s < \text{s-hat} + 2^*v) = 0.95$.

Frequentists are very careful about the meaning of this. To them it is a statement about the probability of s-hat not the probability of s. s is fixed, so it is meaningless to make statements saying that the probability is 95% that s lies between l and h. When giving a confidence interval, what the frequentist means is that if you did the experiment 100 times and calculated a confidence interval each time, then in 95 experiments the confidence interval would include the true value of s.

Bayesian approach

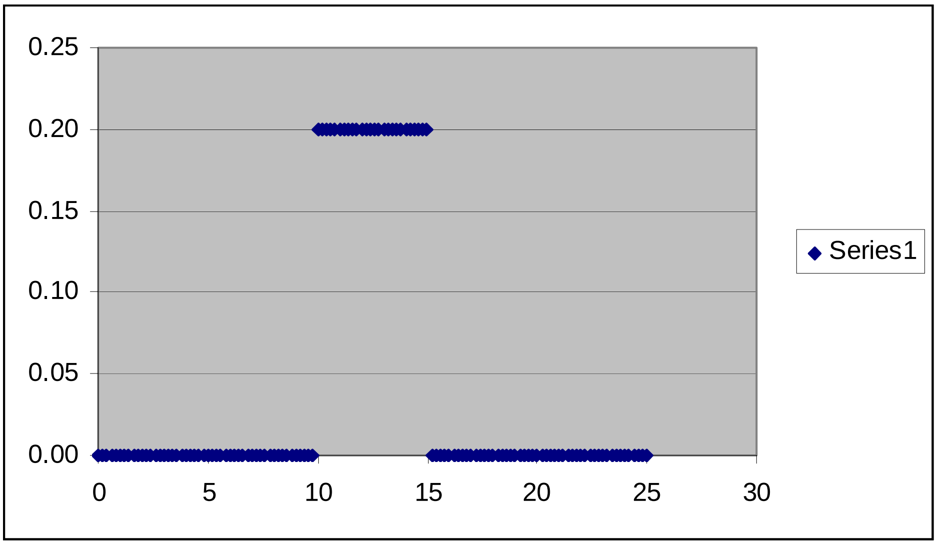
Contrary the frequentist, the bayesian's aim when analysing the experiment is to make a probability statement about the true value of s. She does this using Bayes theorem. That is,

P(s | y) is proportional to P(y | s) * P(s).

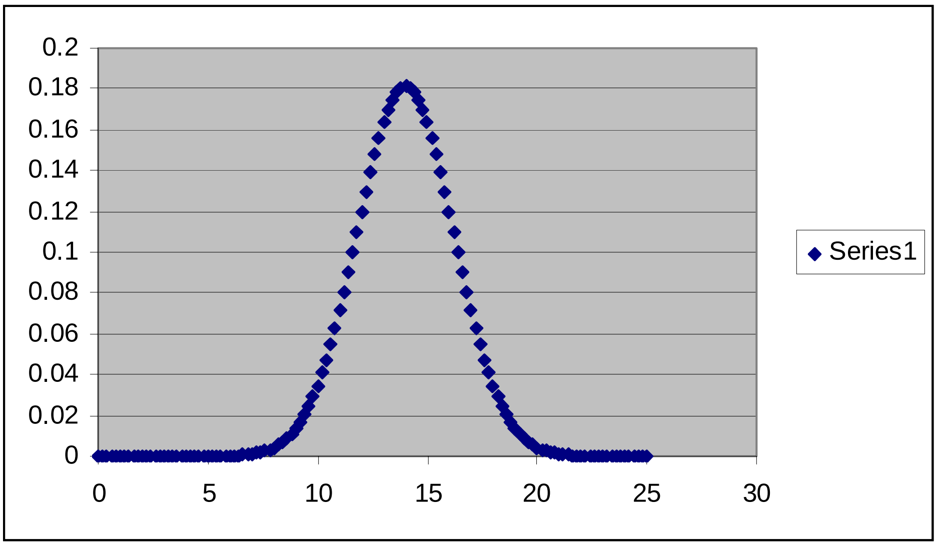
(y and s are continuous variables so their distributions are described by probability density functions and we will interpret the P terms above in that way.)

P(s | y) is called the posterior probability because it is the probability after the experiment has been done. It is calculated from two terms: P(y | s) is the likelihood used by frequentists; P(s) is called the prior probability because it is the probability of s before the experiment was conducted. This gives the bayesian a method to incorporate prior knowledge into the estimate of s.

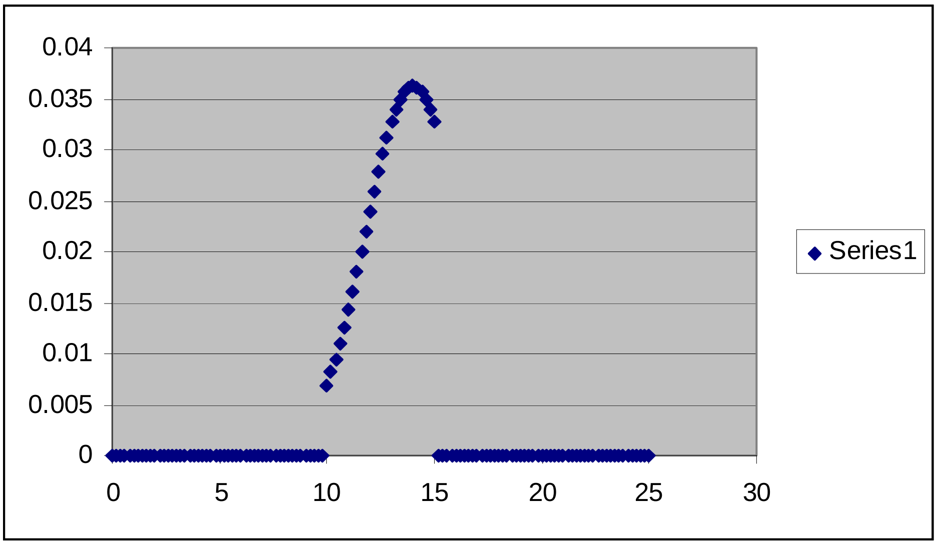
Suppose the bayesian statistician analysing this data thought before the experiment that men were, on average, between 10 and 15 cm taller than women, but within this range she regards all values as equally likely. Her prior P(s) can be drawn



Suppose the mean difference in the sample of 10 men and 10 women is 14 cm with a standard error of 2.2 cm. The likelihood, P(y | s), is proportional to



Multiplying these two terms together we find that the posterior, P(s| y), is proportional to



Except for a constant, necessary to make the area under the curve equal 1.0, this is the posterior distribution of s ie P(s| y). It still says that s must lie between 10 and 15 cm because this was the only range that the prior distribution allowed, but it now says that the true value of s is more likely to be 14 than 10 say. If we want a point estimate of s, we could use the mode of the posterior distribution (14 cm) or the mean (13 cm). In this case, the mode is the same as the frequentist ML estimate. This is not surprising because the likelihood helps to determine the posterior and it is the mode or maximising value of the likelihood that is the ML estimate.

If the bayesian statistician analysing these data had had no idea of the value of s before the experiment, she could have used a prior that had the same value for all values of s. Then the posterior would be simply proportional to the likelihood. It is commonly the case that a bayesian estimate based on an uninformative prior is the same as the ML estimate.

Fixed and random effects

Frequentist approach

Suppose the experiment is to estimate the difference in milk yield between the daughters of bull A and bull B, where bulls A and B are selected at random from the Australian Holstein population. The statistical model for the data might be

$$Y = u + s + e$$

Where

y = milk yield

u = mean of the population

s = effect of the sire on his daughters milk yield

e = error ~ $N(0,\sigma^2)$

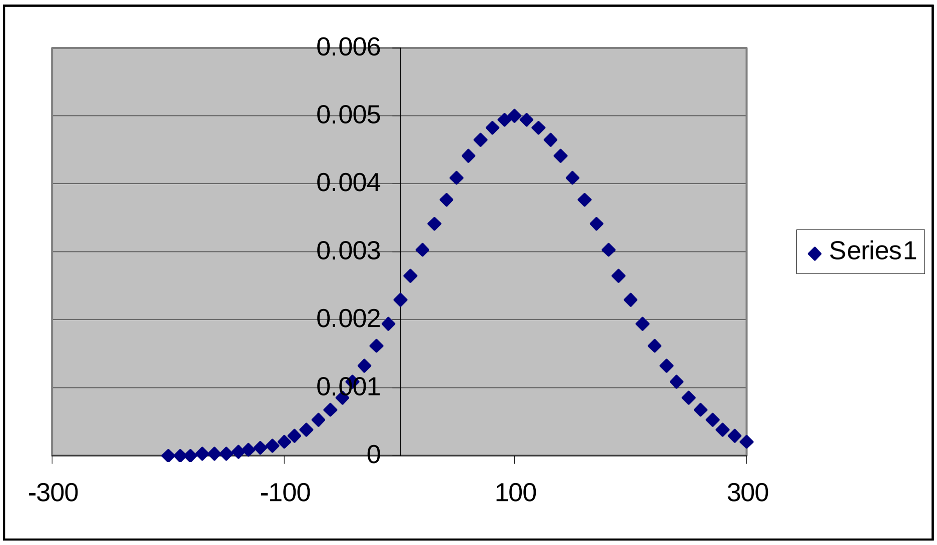
It is normal in this circumstance to treat the sire effects s as a random variable drawn from a distribution that is ~ $N(0,\sigma_s^2)$. Then, if u and the variances σ and σ_s are known, s can be estimated by $E(s \mid y) = [\sum(y-u)] / (n+\lambda)$ ($\lambda= \sigma^2/\sigma_s^2$, n= number of daughters for this sire), which is the usual selection index estimate.

If u, and perhaps other fixed effects, are unknown, then replacing u with uhat, the generalised least squares estimate of u, in the above formula, gives the BLUP solution for s. For instance, if a bull's 100 daughters average 100 L above the population mean and $\lambda = 16$, then s-hat = $[\sum(y-u)] / (n+\lambda) = 100*100/116 = 86$ L.

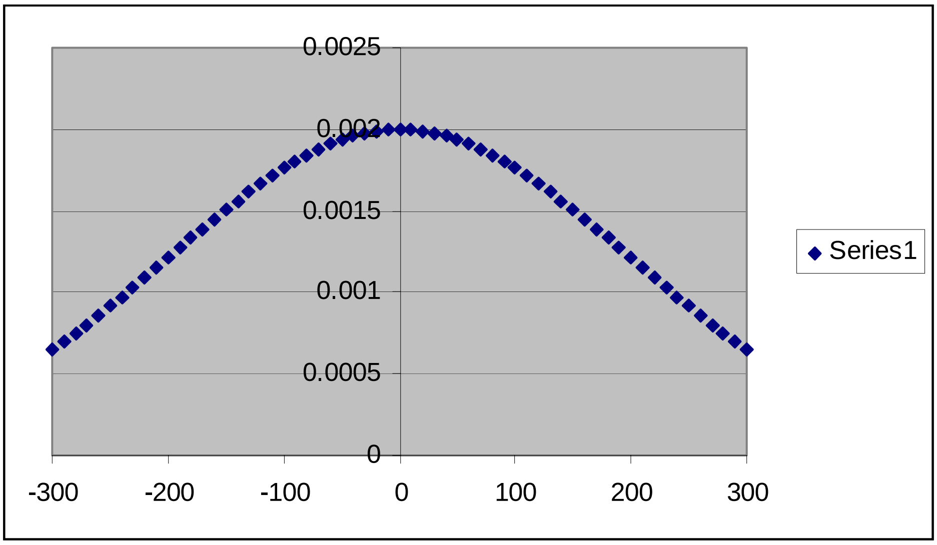
The frequentist uses a different method to estimate s in this example to the method used to estimate s in the previous example, because the effect of a bull is treated as a random effect, whereas the effect of sex is treated as a fixed effect. Because the effect of a sire on his daughter's milk yield is treated as a random variable, it is acceptable to talk about the expectation of this random variable (0) and the expectation conditional on the observed milk yields of 100 daughters (86). In the previous example, the effect of sex on height was considered a fixed effect and therefore it is unacceptable to talk about its distribution and the expectation of that distribution – there is just one true value that we are trying to estimate.

Bayesian approach

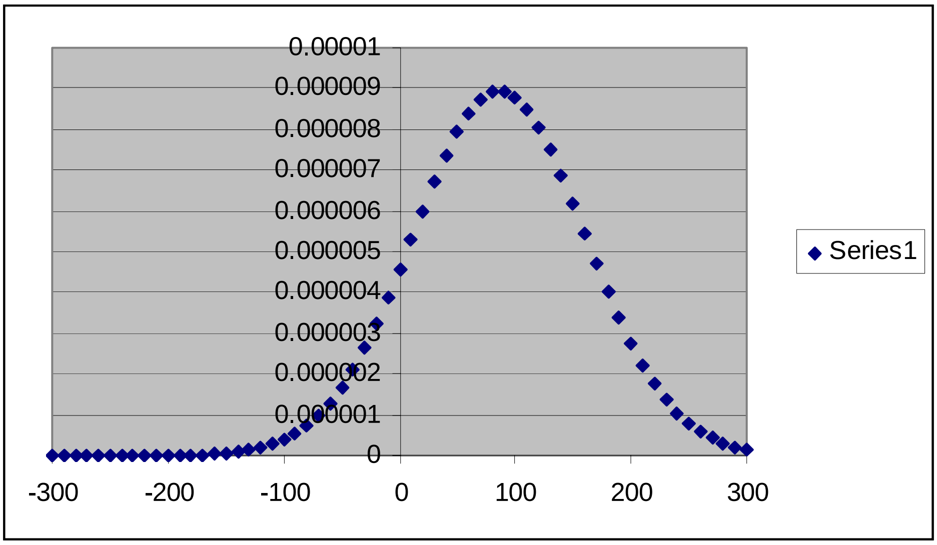
The bayesian approaches this problem in the same way as the previous one. The posterior distribution, P(s | y) is proportional to P(y | s)*P(s). Assuming σ = 800 L and σs = 200 L, the likelihood P(y | s) is



The prior distribution, P(s), is



Multiplying the prior P(s) by the likelihood P(y | s) gives the posterior distribution P(s | y) .



The mean of the posterior is at 86 L, exactly the same as the BLUP estimate of s calculated by our frequentist. In fact, this is a general result – the BLUP estimate and the bayesian estimate are the same. This is because the prior distribution used by the bayesian is also used by the BLUP. Frequentists are happy to use prior distributions for random variables especially when the choice of a prior is well founded. In this case the prior is well founded because we know that a normal distribution with a standard deviation of 200 L is appropriate for sire effects on milk yield. Frequentists are also happy to use posterior distributions for random variables except they would call them conditional distributions (conditional on the observed data).

The posterior distribution contains more information than just its mean. In this case it tells you that the sires effect could be as low as –100 or as high as +300 although these values are unlikely. Because this posterior is a normal distribution, the mean and the variance or prediction error variance (pev), completely describe it. However in general the posterior distribution contains more information than just the mean and pev and , to a bayesian, it contains all the information about the parameter that is known.

Thus one difference between bayesians and frequentists is that bayesians treat all effects as random variables. Rather they argue about whether a particular effect is random or fixed it is more useful to consider the properties on the two types of estimates. If s-hat is a frequentist estimate of a fixed effect (eg a least squares estimate), it has the property

E(s-hat | s) = s.

This means if you do the experiment many times, on average you get the right result. This is the property frequentists call unbiased. If s-hat is a BLUP estimate of the random effect s, it has the property

E(s | s-hat) = s-hat.

This means that if you select the best effect out of many based on s-hat, on average the true effects will be as good as you expected. It is this property that makes BLUP solutions the best criterion on which to select. Least squares estimates do not have this property. Typically the best least squares estimates are over estimated. What is worse, the less reliable a high least squares estimate is, the more it is likely to overestimate the true value. Compared to a least squares estimate, a BLUP estimate is regressed back towards the mean because it incorporates the prior distribution whose mean is zero. The more data accumulates on an effect, the more the likelihood dominates the prior and the less the estimate is regressed back.

Both types of estimates have uses in my opinion. Fixed effect or least squares estimates are a good summary of one experiment whereas bayesian estimates tend to ‘pollute’ the results of this one experiment with past experience as represented in the prior. Bayesian or BLUP solutions are useful if some decision is going to be based on them because they have a smaller prediction error variance than least squares solutions and are more realistic ie the don’t systematically overestimate the benefits from the chosen decision.

These beneficial properties of bayesians estimates are less clear cut if there is no good justification for the prior chosen. However, even a rather vague prior may be a better basis for decisions than no prior at all.

Estimating multiple parameters

Frequentist approach

Often a number of parameters are needed to explain the observed data. In a simple example, we might use a sample from a normal distribution to estimate the mean and the variance of the population. The ML method is to find the estimates of both the mean and the variance that maximize the likelihood. For the variance, this is $\sum(u-\bar{u})^2/n$. This ML estimate is biased. To produce an unbiased estimate it is conventional to divide by (n-1) instead of n. The bias in the ML estimator occurs because it does not allow for the uncertainty in the sample mean ($\bar{u}$). The ML estimator of the variance is the same regardless of whether it uses the true population mean or the sample mean, but the sample mean is closer to the individual sample values than the population mean, and so the estimator of the variance that uses the sample mean underestimates the variance.

Bayesian approach

Using Bayes theorem we obtain the joint posterior distribution of the mean and variance, ie

$$P(\mu,\sigma^2 \mid y) = P(y \mid \mu, \sigma^2) * P(\mu,\sigma^2).$$

If the joint distribution is known, it is possible to calculate the marginal distribution by integrating over one of the variables. For instance, the marginal distribution of σ^2 is

$$P(\sigma^2 \mid y) = \int P(\mu, \sigma^2 \mid y)\, d\mu$$

This marginal distribution can be used to make inferences about σ^2. For instance, the mean of the marginal posterior distribution is the conventional unbiased estimate of

σ^2 that divides by n-1. The marginal distribution is not the same as the distribution conditional on μ = the sample mean, which is centred on the biased ML estimate of σ^2. That is, by integrating out μ, we have taken account of uncertainty in μ when estimating σ^2.

Nuisance parameters

Frequentist approach

Sometimes the model includes parameters that we are not really interested in, but which must be included to give a good fit to the data and hence improve the estimates of other parameters in which we are interested. An example might be the effect of herd-year-seasons on milk yield when we are trying to estimate the sire variance or the breeding values of individual sires. ML deals with this problem in three different ways:

- When estimating breeding values, the nuisance parameters are fitted in the model.
- When estimating the sire variance, the likelihood is partitioned into a part that depends on the nuisance parameters and other fixed effects and a part that does not. The latter likelihood is maximized by the choice of variance estimates which are called Restricted ML or REML estimates.
- When estimating the fixed effects, the individual sire breeding values are regarded as random effects and not included as parameters of the model. The likelihood might include the variance of breeding values but it doesn't include the individual breeding values.

Bayesian approach

The bayesian treats all 3 types of nuisance parameters in the same way – they are included in the model with appropriate priors but integrated out to give the marginal distribution of the parameters that we are interested in.

This points out an important feature of the bayesian approach. Frequentists distinguish sharply between parameters that describe distributions and realisations of random variables from these distributions. For instance, if y ~ N(μ, σ2) then μ and σ are parameters but any particular observed value of y is not. For bayesians it is not necessary to make this distinction – they are all just parameters which are drawn from a prior distribution.

Statistical inference

Frequentist approach

The classical approach is to test a null hypothesis against an alternative hypothesis. The null hypothesis is a simple hypothesis where a parameter that we are interested in has been set to zero or some other fixed value. We test the observed data to see how unlikely it is to have occurred under the null hypothesis. For instance, in the first example, we might test the null hypothesis that there is no difference in height between men and women in Australia. Under the null hypothesis less than 5% of samples would have as big or bigger difference between the sexes in height as we observed. Therefore we reject the null hypothesis and accept the alternative hypothesis that there is a difference between the sexes in height.

Bayesian approach

For a bayesian, all our knowledge about a parameter is represented in its posterior distribution. For instance, the posterior for the difference in height between sexes has zero density at zero, because the prior had zero density at zero. That is, the statistician believed before the experiment that there was no chance that there was no sex effect on height. The posterior shows that a sex difference of 14 cm is six times as likely as a difference of 10 cm. This is not the same as performing a significance test and, in general, bayesians are disinclined to perform significance tests. However it is often possible to define a range of values that includes 95% of the posterior distribution and call that a 95% confidence interval. A null hypothesis that lies outside that range would be rejected.

My simplistic conclusion is that the bayesian approach is well suited to estimation problems, especially where the estimate is to be used to make a practical decision. The frequentist approach is well suited to testing hypotheses about the nature of the world.

Computational problems

Calculating the posterior distribution and integrating out some parameters may be difficult to do. Often it is impossible to find a formula that gives the solution. Bayesians have invented various methods to overcome this difficulty:

- Chose priors that make the algebra easy. So-called conjugate prior distributions have the property that, when combined with a particular distribution for the data, they yield a recognised distribution for the posterior. For instance, if the data are normally distributed and a normal prior is used for a parameter such as the sex effect on height, then the posterior distribution of that parameter is also normal.
- Numerical integration. If you can calculate the height of the posterior distribution at every point, you can integrate it over nuisance parameters using numerical integration such as Simpson's rule.
- Simulation. If you can draw random samples from the posterior distribution you can use the sample to approximate the distribution. For instance, the mean of many samples is a good approximation to the mean of the distribution. This is what Markov Chain Monte Carlo (MCMC) methods do.

Gibbs Sampling

MCMC

Gibbs sampling is one of several Markov Chain Monte Carlo (MCMC) methods. The aim of these methods is to take samples from the posterior distribution. They are called Monte Carlo because they involve drawing random numbers from specified distributions and Markov chain because each sample depends on the previous sample. MCMC methods can be used by frequentists but, because they involve sampling from the posterior distribution, they fit more conveniently into the statistical tool box of bayesians.

The value of MCMC methods is that they can be used in analyses that are very difficult to perform analytically. They work by breaking a very complex problem down into a series of simple steps. Their popularity has grown in recent years because they often require a lot of computer power to generate many samples and this is only practical with fast computers. I will describe Gibbs sampling because it is the most popular MCMC method in genetic analysis but other methods such as the Metropilis-Hastings algorithm are also used.

Gibbs sampling

In Gibbs sampling we draw a random sample of one parameter at a time assuming that the current values of all the other parameters are correct. Then we go on to the next parameter and cycle through all the parameters many times. Although it is not obvious, this procedure generates samples from the posterior distribution of all the parameters, provided some conditions discussed later are met.

A very simple example of Gibbs sampling

I will illustrate Gibbs sampling with a trivially simple example so that you can see how it works. Assume we have observed that an offspring of sire A mated to dam B carries a lethal recessive gene. The frequency of this recessive in the population is known to be 0.1.What is the probability that this lethal gene is carried by the sire?

We define two variables:

- The genotype of the dam (Gd) which is either normal (++) or carrier (+m)
- The genotype of the sire (Gs) which is either ++ or +m

The Gibbs sampler consists of sampling these from their conditional distributions where we condition on the observed data (Go= genotype of offspring = +m) and the genotype of the mate. That is, when we sample the genotype of the sire it is from P(Gs | offspring, Gd). When we sample the genotype of the dam we sample from P(Gd | offspring, Gs). Therefore before we implement the sampling we need to work out these probabilities. It is convenient to use Bayes theorem to do this. That is

P(Gs | Go, Gd) = P(Go | Gs, Gd) *P(Gs) / P(Go | Gd)

Where the priors are P(Gs =++) = 0.9*0.9 =0.81, P(Gs = +m) = 2*0.9*0.1 =0.18 and the same for Gd. The likelihoods are

P(Go = +m | Gs = ++, Gd = ++) =0
P(Go = +m | Gs = ++, Gd = +m) = 0.5
P(Go = +m | Gs = +m, Gd = ++) = 0.5
P(Go = +m | Gs = +m, Gd = +m) = 0.5

Using these probabilities, the conditional probabilities for Gs are calculated below:

Gd = ++

Gs	P(Gs)	P(Go=+m | Gs, Gd)	P(Gs)*P(Go=+m | Gs,Gd)	P(Gs | Go, Gd)
++	0.81	0	0	0
+m	0.18	0.5	0.09	1
total = P(Go = +m | Gd = ++)			0.09	

Gd = +m

++	0.81	0.5	0.405	0.82
+m	0.18	0.5	0.09	0.18
total = P(Go = +m | Gd = +m)			0.495	

In summary, if the dam is ++ the sire must be +m; if the dam is +m the probability that the sire is +m is 18%. Due to symmetry, these conditional probabilities are the same with Gs and Gd reversed.

Now we can implement the sampling scheme. However, if we start by sampling Gs, we need a starting value for Gd. Let us arbitarily start with Gd = ++. This means we sample Gs from the conditional distribution P(Gs | Go = +m, Gd = ++). From the table above we sample Gs = ++ with probability 0 and +m with probability 1. Therefore we sample Gs = +m. Now sample Gd from P(Gd | Go = +m, Gs = ++m) which has probability Gd = ++ of 0.82 and Gd = +m of 0.18. Suppose we sample Gd = ++. Now continue sampling Gs and then Gd. Table 1 gives an example of the results from one run of 20 cycles of this Gibbs chain.

Table 1 samples from the Gibbs sampler

Cycle	Gd	Gs
1	++	+m
2	++	+m
3	++	+m
4	++	+m
5	++	+m
6	+m	++
7	+m	++
8	+m	+m
9	+m	++
10	+m	+m
11	++	+m
12	++	+m
13	++	+m
14	+m	++
15	+m	++
16	+m	++
17	+m	++
18	+m	+m
19	++	+m
20	++	+m

In these 20 cycles, we sampled Gs = ++ 7times and Gs=+m 13 times. Therefore, using these samples we would estimate that the P(Gs = +m | Go=+m) is 13/20 = 0.65. However two things are wrong with this implementation of the Gibbs sampler.

Burn-in

Firstly, at the beginning we arbitarily set Gd=++. In the long run this starting value wont affect the distribution that the chain reaches because eventually the chain converges to the true distribution, but for the first few cycles the starting value does affect the results. This can be clearly seen in this example because, by starting with Gd=++, we forced Gs=+m. The normal method to avoid the final result being influenced by the starting values in to discard the first few samples which are called 'burn-in'.

Auto-correlation between cycles and reducibility

The second problem is that the Gibbs chain does not move freely between possible values of Gs. It tends to get stuck in runs where Gs = ++, Gd = +m or runs where Gs=+m and Gd=++. This is a common problem in Gibbs chains. It is described by saying there is an auto-correlation between one cycle and the next. This is not surprising since only one parameter is changed at a time. The simplest solution is to do a lot of cycles, so that the runs of one kind average out with the runs of the other kind. This solution works provided all possible solutions (Gs, Gd pairs in this case) can be reached from any starting position. If this is not the case, some possible solutions never get tested and so can't appear in the final sample. Such a Gibbs chain is called reducible and can't be used to estimate the parameters. Our Gibbs chain is irreducible and does reach all possible solutions, so valid estimates of Gs and Gd would be obtained if we used more cycles and discarded the first few cycles as burn-in.

Joint sampling of more than one parameter

Some Gibbs chains are formally irreducible, but the auto-correlation is so high that it would take too many cycles to explore the whole parameter space. In this case the gibbs chain must be redesigned to reduce the auto-correlation between cycles. One way to do this is to sample more than one parameter at a time. In our simple example we could sample Gs and Gd together using the following table of conditional probabilities:

Gs	Gd	P(Gs,Gd)	P(Go=+m | Gs,Gd)	P(Gs,Gd | Go= +m)	
++	++	0.81*0.81	0	0	/0.162 = 0
++	+m	0.81*0.18	0.5	.073	/0.162 = 0.45
+m	++	0.18*0.81	0.5	.073	/0.162 = 0.45
+m	+m	0.18*0.18	0.5	.016	/0.162 = 0.10
total				.162	

From these conditional probabilities we can sample Gs and Gd simultaneously. Since these are the only two parameters in the model, there is no dependency of one cycle on the next and so no auto-correlation. In this case we can see that the Gibbs chain would sample (Gs, Gd) = (++,+m) 45% of the time, (+m, ++) 45% and (+m, +m) 10%. Therefore the probability that Gs = +m is 0.45 +0.10 = 0.55. In this case we have computed the conditional probability without needing the Gibbs chain, but where there are many parameters, we usually can’t sample them all simultaneously, but sampling them more than one at a time often reduces the auto-correlation between cycles and means that less cycles are needed.

Number of Gibbs cycles needed

The more cycles are used the more accurate will be a mean based on these cycles. If there was no auto-correlation between cycles, it would be easy to calculate the standard error of the mean from the variance across cycles and the number of cycles using the usual formula. However typically there is an auto-correlation. One practical strategy is to run more than one chain, starting from different positions, and compare the answers from different chains. If these do not agree well enough, then longer chains or more chains are needed.

The length of burn-in needed also depends on the auto-correlation, so it is reasonable to discard the first 10% of samples once a chain of sufficient length has been performed.

An example with a continuous variable

Suppose we wish to estimate the mean and variance of a population from a sample of 10 observations. We assume the observations (y) are normally distributed $y \sim N(\mu, \sigma^2)$ and independent of each other. The gibbs sampler will sample μ conditional on the current value of σ and then σ conditional on the current value of μ, leading to a chain of values μ1, σ1, μ2, σ2 … which, after a burn-in, will be a sample from the posterior distribution. As usual we derive the conditional posterior distributions by applying Bayes theorem to the prior distribution for each parameter and the likelihood of the data given the parameters. Since they are normally distributed, the likelihood of the 10 observations

$P(y \mid \mu, \sigma)$ is proportional to $\sigma^{-n} \exp\{ - \Sigma(y-\mu)^2 / (2\sigma^2) \}$ where n=10.

We will assume that we have little knowledge of the prior distributions of μ and σ and therefore use flat, uninformative priors P(μ) = constant and $P(\sigma^2)$ = constant. Consequently the posterior distributions $P(\mu \mid y, \sigma)$ and $P(\sigma^2 \mid y, \mu)$ are both proportional to the likelihood.

When the likelihood formula above is viewed as a distribution of μ it shows that the posterior $P(\mu \mid y, \sigma)$ is a normal distribution with mean equal to the sample mean (Σy/n) and variance = σ^2/n. Thus at each cycle of the Gibbs chain we sample μ from $N(\Sigma y/n, \sigma^2/n)$ where σ is the last value of σ sampled.

When the likelihood formula above is viewed as a distribution in σ^2 it shows that the posterior $P(\sigma^2 \mid y, \mu)$ is an inverse chi-square distribution with n-2 degrees of freedom scaled by the sample sum of squares, $\Sigma(y-\mu)^2$. Therefore at each cycle of the Gibbs chain we sample a chi-squared variate, invert it and multiply it by $\Sigma(y-\mu)^2$, where μ is the last value of μ sampled. This gives us a sample from the posterior distribution of μ and σ that we can summarize by a mean and a standard error for both μ and σ^2 if we wish.

Gibbs sampling in a linear model

The Gibbs chain for a sample from a single population can be extended to a more complex experiment with normally distributed errors:

$Y = Xb + e$

Where

y = a vector of observations

b = a vector of parameters

X = a design matrix

e = a vector of independent errors $\sim N(0,\sigma^2)$

The likelihood conditional on b,

$P(y \mid b)$ is proportional to $\sigma^{-n} \exp\{-(y-Xb)'V^{-1} (y-Xb)/2\}$ where $V=I\sigma^2$.

The prior P(b) will be assumed to be $b \sim N(\beta, W)$. Some elements of b can have uninformative priors, in which case the corresponding diagonal element of W is infinite. Other elements of b can have informative priors. For instance, elements that are breeding values can have $b \sim N(0,A\sigma_a^2)$. The variances σ^2 and σ_a^2 can have flat priors or scaled inverted chi-square priors to reflect prior knowledge about them.

If the variances in W were known, we could solve a system of mixed model equations:

$(X'V^{-1}X + W^{-1})b = X'V^{-1} y$

or $\quad C\, b = z$

Then the posterior distribution of b is $N(C^{-1} z, C^{-1})$.

Because the variances are not known, we use a Gibbs chain. With the current values of the variances, sample each element of b in turn from $b \sim N(b_i, c_{ii}^{-1})$ where b_i is the solution of the equation

$c_i'\, b = z_i$ with all other elements of b equal to their current value in the chain and c_{ii}^{-1} is the inverse of the diagonal element i. [This is the distribution of b_i if all other parameters were known and hence the data could be corrected for them.]

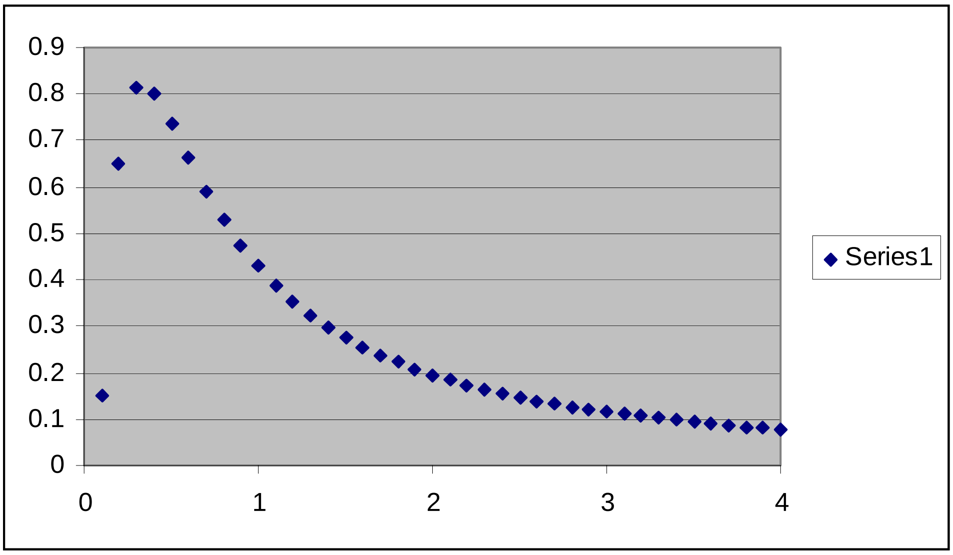
Now sample the variances. This is relatively easy because we have already sampled the true values of variables such as breeding values, and so we can calculate their variance just as we did above when dealing with a single sample. If flat priors were used for the variances, they are sampled from inverted chi-squared distributions multiplied by the appropriate sample sum of squares. For instance,

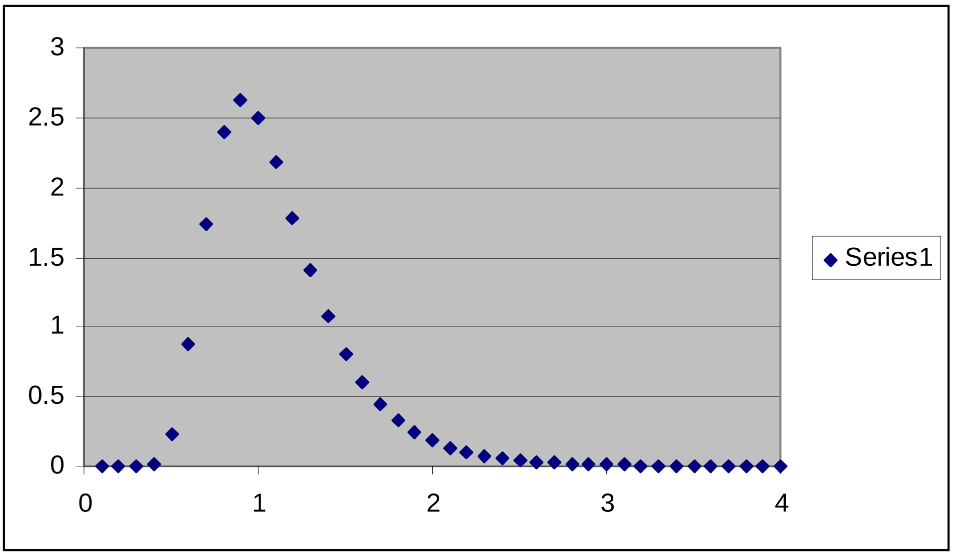
σ^2 is sampled from an inverted chi-square with n-2 degrees of freedom multiplied by e'e where n is the order of the vector e = y-Xb;

σ_a^2 is sampled from an inverted chi-square with n_a-2 degrees of freedom multiplied by $b'A^{-1}b$ where b= the vector of breeding values of size n_a.

If informative priors are used for the variances, it is convenient to use a scaled inverted chi-squared distribution as a prior. That is, $\sigma_i^2 \sim X_{vi}^{-2} * S_i$. This describes a distribution with a mean of $S_i/(v_i-2)$ and a variance that depends on $1/(v_i-4)$. Therefore distributions with v<2 have an infinite mean and v<4 have an infinite variance, but they are still proper distributions and can be used as priors. Below are graphed the inverted chi-squared distributions with 1 and 20 degrees of freedom each scaled by their own degrees of freedom. From these you can see that the distribution with 20 degrees of freedom is narrower than that with 1 degree of freedom. Therefore the larger the degrees of freedom used, the more informative the prior, implying that we know the variance quite well before the experiment. Thus the prior describes a belief that the expected variance is about S/(v-2) and our confidence in this belief is signified by v.

Figure 1 Inverted chi-squared distributions with 1 df and with 20 df and scaled by 20





The advantage of using an inverted chi-square distribution as a prior for variances is that, for normally distributed data, the posterior is also an inverted chi-square. If the prior for the error variance has scaling factor S and the degrees of freedom v, then the posterior for σ^2, P(σ^2 | y, all other parameters) is an inverted chi-square scaled by (e'e + S) and with (n +v) degrees of freedom. Similarly if the prior for genetic variance is an inverted chi-square with degrees of freedom v_a and scaling factor S_a, then the posterior of σ_a^2 is an inverted chi-square scaled by ($a'A^{-1}a + S_a$) and with ($n_a + v_a$) degrees of freedom. From these formulae it is clear that the scaling factor can be viewed as a total sum of squares from the prior and the data and similarly the degrees of freedom is a total of those in the prior and the data. Consequently the scaling factor divided by the degrees of freedom is a weighted mean of the variance implied by the prior and by the data, and is the expected value when sampling from the scaled chi-squared distribution.

This gibbs chain demonstrates a common feature of gibbs sampling. All variables are treated alike; there is no distinction between fixed and random effects or parameters and random variables. By sampling some variables, such as breeding values, we make it much easier to sample other variables such as the variance of breeding values. That is, the conditional distributions (eg genetic variance conditional on a sample of breeding values) are easier to calculate than the marginal distributions (eg genetic variance conditional only on the data).This is analogous to the use of 'missing data' algorithms such as the EM algorithm. Usually by sampling well chosen missing variables we make sampling other variables simpler but we increase the auto-correlation between cycles. Therefore we may be forced to choose between a simple sampling scheme with a long computing time and a more difficult sampling scheme with reduced computing time.

Exercises for Gibbs sampling course

1. Four animals form the pedigree

A

C B
(1,1)

D
(1,1)

Animals C and D have been genotyped for a bi-allelic marker at which the allele frequencies are p= p(1) = 0.1, 1-p = p(2)= 0.9. What are the genotype probabilities for animals A and B that have not been genotyped?

With this small pedigree, the genotype probabilities could easily be calculated using segregation analysis and a peeling algorithm, but in large pedigrees with loops it may be better to use gibbs sampling. We will use gibbs sampling on this small pedigree.

It is easy to calculate the genotype probabilities of an animal if its parents, mates and progeny have known genotypes, because, if these are known, the genotypes of other animals have no effect. For instance, if the genotype of animal B was known, we would not need to consider the genotype of D when calculating the genotype probabilities for animal A. Similarly, if the genotype of A was known, we would not have to consider the genotype of C when calculating the genotype probabilities for animal B.

Therefore we will use a gibbs chain that samples the genotype of animals A and B in each cycle. Although inefficient, this can be done in excel using the IF function to account for the two possible genotypes of A when calculating genotype probabilities of B and the two possible genotypes of B when calculating genotype probabilities of A.

The steps in each cycle are:

1. Calculate the probability of the possible genotypes for animal A assuming the most recently sampled genotype for B.
2. Sample one of the possible genotypes for A according to the probabilities just calculated.
3. Repeat steps 1 and 2 for B assuming the genotype just sampled for A.

Notice from the genotypes of C and D, that A and B can only have genotypes 11 or 12, so these are the only ones we need to consider when sampling. [In the following I will use the symbol A to mean the genotype of A , etc for B, C, D].

The genotype probabilities are calculated as follows:

P(A | C, B) is proportional to P(A, B , C) = P(A) * P(B | A) * P(C | A)

Where

$P(A=11) = p^2$, $P(B = 11 \mid A =11) = p$, $P(C = 11 \mid A=11) = p$,
$P(B = 12 \mid A =11) = (1-p)$,

$P(A=12) =2p(1-p)$, $P(B=11 \mid A=12) = p/2$, $P(C=11 \mid A=12) = p/2$,
$P(B=12 \mid A=12) = 0.5$,

Calculate P(A, B, C) for the current genotype of B and for both possible genotypes of A. Then

$$P(A=11 \mid B, C) = \frac{P(A=11, B, C)}{P(A=11, B, C) + P(A=12, B, C)}$$

$$P(A=12 \mid B, C) = \frac{P(A=12, B, C)}{P(A=11, B, C) + P(A=12, B, C)}$$

To sample A given these two probabilities:

1. sample a random number $u \sim U(0,1)$ ie distributed evenly between 0 and 1. [This can be done in excel using rand()].
2. If $P(A=11 \mid B, C) > u$, sample A =11; otherwise sample A=12.

P(B | A, D) is proportional to $P(B, A, D) = P(D \mid A, B) * P(A, B)$
$= P(D \mid B) * P(B \mid A) * P(A)$
Which because P(A) is constant during these calculations
is proportional to $P(D \mid B) * P(B \mid A)$

where

$P(D=11 \mid B=11) = p$, $P(B=11 \mid A=11) = p$
$P(B=11 \mid A=12) = p/2$,

$P(D =11 \mid B=12) = p/2$, $P(B=12 \mid A=11) = 1-p$
$P(B=12 \mid A =12) =0.5$,

Calculate P(A, B, D) for the current genotype of A and for both possible genotypes of B. Then

$$P(B=11 \mid A, D) = \frac{P(B=11, A , D)}{P(B=11, A, D) + P(B=12, A, D)}$$

$$P(B=12 \mid A, D) = \frac{P(B=12, A, D)}{P(B=11, A, D) + P(B=12, A, D)}$$

Sample B with these 2 probabilities.

Repeat the sampling of A and B many times and calculate the posterior distribution of A and B.

2. Two inbred lines of mice are crosses to produce F1's and they are mated to produce F2's. The parent lines have genotypes aabb and AABB. In the F2's the number of progeny of each genotype out of 100 born are:

	AA	Aa	aa
BB	16 (0)	8 (1)	1 (2)
Bb	8 (1)	34 (0 and 2)	8 (1)
bb	1 (2)	8 (1)	16 (0)

We can deduce the recombinations that produce all the F2 genotypes except the AaBb genotype and this number is given in brackets below the number of progeny. When you consider linkage phase, the genotype AaBb is actually a mixture of two genotypes
A B and A b
a b   a B
In the first no recombinations have occurred while in the second 2 recombinations have occurred. If we could distinguish these two genotypes it would be easy to estimate the recombination rate by simply counting the number of recombinations in the 200 gametes that produced the 100 progeny. However we cant distinguish these 2 genotypes. Therefore we will use a gibbs chain, that samples two variables – the recombination rate and the number of each of the 2 genotypes among the AaBb mice.

Let
r = recombination rate
N0 = number of A B progeny that have resulted from gametes with 0 recombinations,
a b
N2 = number of A b progeny (that have resulted from gametes with 2 recombinations)
a B

The probability of A B progeny is 0.5 $(1\text{-}r)^2$, the probability of A b progeny is 0.5 r^2,
a b   a B
So, conditional on the fact that one of these two outcomes has occurred whenever we observe a AaBb mouse,
N0 | r is distributed binomially with proportion p = $(1\text{-}r)^2 / ((1\text{-}r)^2 + r^2)$ and number of trials =34 ie N0 | r ~ B (34, p). Thus we sample N0 from a binomial distribution and then N2= 34 – N0.

Once N2 has been sampled, we know we have observed a total of
Nr =2*2 + 32 + N2*2 recombinations out of 200 gametes. Using Bayes theorem, P(r | Nr) is proportional to P(Nr | r) P(r). If we assume a flat prior for r ie P(r) = constant, then P(r| Nr) is proportional to P(Nr | r). Nr | r is distributed as a binomial, ie P(r | Nr) = ${}^{200}C_{Nr}\, r^{Nr} (1\text{-}r)^{200\text{-}Nr}$. When viewed as a function of r, this is a beta distribution with parameters Nr and 200-Nr. So we sample r from this beta distribution.

One method to sample from a known distribution with cumulative distribution function $F(x) = P(X < x)$ is as follows:

3. sample a random number $u \sim U(0,1)$ [There are common computer functions to do this].
4. Find the value of the variable of x such that $F(x) = u$. This requires using an inverse cumulative distribution function $I(u) = x$. [Computer functions exist for many common distributions].

Normally a gibbs sampler would be implemented using a computer programming language such as fortran or C++. However it is possible to use Excel which we will do. Set up a spread sheet that repeatedly samples N2 | r and then r | N2 and observe the posterior distribution of r. Excel has an inverse cumulative beta distribution function called betainv, an inverse bimonial called critbinom and a random number generator called rand().

3. A sample of size n is taken from a population and measuring for a variable y that is normally distributed ie $y \sim N(\mu, \sigma^2)$. Estimate the mean and variance and confidence intervals for these two parameters.

This is easy to do without gibbs sampling but we will use gibbs sampling to illustrate the process. We will sample μ assuming σ is known and σ^2 assuming μ is known.

Sample μ
If the true variance σ^2 is known, the sample mean y-hat ~ $N(\mu, \sigma^2/n)$. That is,
$P(\text{y-hat} = x \mid \mu, \sigma) = f(x; \mu, \sigma^2/n)$ where $f(x; \mu, \sigma^2)$ is the normal probability density function at point x if the mean= μ and the variance = σ^2.

Therefore, $P(\mu \mid \text{y-hat}, \sigma)$ is proportional to $P(\text{y-hat} \mid \mu, \sigma) * P(\mu)$. We will assume a flat prior $P(\mu)$ = constant, so
$P(\mu \mid \text{y-hat}, \sigma)$ is proportional to $P(\text{y-hat} \mid \mu, \sigma) = f(\text{y-hat}; \mu, \sigma^2)= f(\mu; \text{y-hat}, \sigma^2)$. Therefore we sample μ by sampling from a normal distribution with mean y-hat and variance σ^2. This can be done by the usual means of sampling $u \sim U(0,1)$ and then finding μ such that $F(\mu; \text{y-hat}, \sigma^2)=u$ using a cumulative inverse normal distribution function where $F(\mu; \text{y-hat}, \sigma2)= P(X < \mu \mid X \sim N(\text{y-hat}, \sigma^2))$ is the cumulative normal distribution function.[See the function norminv in excel].

<u>Sample σ2</u>
If the true mean μ is known, the sample variance $s^2 = \Sigma(y-\mu)^2/n \sim X^2 \sigma^2 /n$ where X^2 is a chi-squared with n degrees of freedom. Assuming a flat prior for σ^2, the posterior for σ^2 is $X^{-2} ns^2$ where X-2 is an inverted chi-square with n-2 df. Therefore to sample σ^2, sample a chi-square with n-2 df and calculate $\sigma^2 = ns^2 / X^2$.[Excel has chiinv function to help sample chi-squared].