6 Things to do with the posterior in the City

Research project is like a big city, full of unseen alleys to be explored. A good prior information may be helpful. For the research, you have probabilistic tools to tackle the uncertainties via updating your probabilities as more data become available when you walk around. So you go on and get posterior probabilities. In what ways can they be summarized or used? What to do with it?

Posterior density is the basic core in every bayesian analysis, but it can serve us in many different ways. The following examples include some common routines: (1) assessing the plausibility of a hypothesis, (2) reporting the 'uncertainty bands' instead of simple point estimates, (3) making a prediction which integrates all involved uncertainties. Comparisons to classical non-bayesian approaches are briefly made.

In more realistic problems, we have a multidimensional unknown parameter, whose posterior distribution needs to be computed, as a starting point for any further assessment. Typically the interest focuses on the marginal distribution of only some parameters:

$$\pi(\theta_1,\ldots,\theta_k \mid \text{data}) = \int\ldots\int \pi(\theta_1,\ldots,\theta_k,\theta_{k+1},\ldots,\theta_n \mid \text{data})\mathbf{d}\theta_{k+1}\ldots\mathbf{d}\theta_n$$

In this way, the rest of the parameters can be seen as only nuisance parameters whose uncertainty just has to be taken into account. For example: with $X_i \sim \mathrm{N}(\mu,\sigma^2)$ we might be interested only in μ, but because σ^2 is unknown, the uncertainty about it needs to be accounted for. Hence, we are interested in the marginal posterior $\pi(\mu \mid X_1,\ldots,X_n) = \int \pi(\mu,\sigma^2 \mid X_1,\ldots,X_n)\mathbf{d}\sigma$. Likewise, with e.g. linear models, we could have a vector or parameters $\beta_1,\ldots,\beta_k$ describing the effect of k explanatory variables which we all want to include, but only some effects might be interesting. The numerical computation of these multidimensional posteriors will be introduced later in the context of Monte Carlo and MCMC simulation.

6.1 Hypotheses

With continuous quantities as r, it is not meaningful to ask e.g. what is the probability $P(r = 0.5)$, because such probability is always zero. The mode of the density shows the value with highest probability density, and thus it provides a 'best guess'. However, a hypothesis about the value of r needs to be constructed as a statement involving intervals. Using posterior density, we can then study what evidence we have to support specific hypotheses. For example, if the hypothesis is that $r < 0.5$, then the prior probability of that hypothesis is

$$P(r < 0.5) = \int_0^{0.5} \pi(r)\mathbf{d}r = 0.5 \qquad \text{(from U(0,1)-prior)}$$

but the posterior probability would be

$$P(r < 0.5 \mid Y, N) = \int_0^{0.5} \mathrm{Beta}(r \mid Y+1, N-Y+1)\mathbf{d}r$$

which is the cumulative probability of the beta-density at $r = 0.5$. The approximate value (0.3125, when $Y = 2, N = 3$) is obtained by typing `pbeta(0.5,Y+1,N-Y+1)`. The posterior probability summarizes the current evidence, but we may also compute posterior odds. The prior odds for the hypothesis were

$$\frac{P(r < 0.5)}{P(r \geq 0.5)} = 1$$

but the posterior odds are only about half of that

$$\frac{P(r < 0.5 \mid Y, N)}{P(r \geq 0.5 \mid Y, N)} = \frac{0.3125}{0.6875} = 0.4545.$$

Hypotheses could also involve comparisons of two quantities. For example, we could study two different bags, each with a different proportion of red balls, r_1 and r_2, and we get some observations from both, (Y_1, N_1) and (Y_2, N_2). The hypothesis could then be e.g. $H_0 : r_1 < r_2$. What is the prior and the posterior probability of the hypothesis? To study this, we can create a new variable: $s = r_1 - r_2$, so that $H_0 : s < 0$. But now the distribution of s is a convolution of two independent distributions and generally it may be difficult to compute analytically, but very easy to simulate.

Hypotheses related to linear models $E(Y_i) = \alpha_0 + \alpha_1 X_i$ could e.g. focus on the slope parameter. $H_0 : \alpha_1 < 0$ and $H_1 : \alpha_1 \geq 0$. Hence, the posterior distribution $P(\alpha_1 < 0 \mid Y, X)$ directly assesses the probability of this hypothesis, $P(H_0 \mid Y, X)$.

Note: frequentist hypothesis testing with p-values gives the probability of more extreme Y than the one observed, *given* null hypothesis: $P(Y$ more extreme than $Y_{\text{obs}} \mid H_0)$. The null hypothesis may thus be rejected (or not), if a more extreme observation than what we had would seem too improbable. The frequentist hypothesis testing does not give probability of the hypothesis. Harold Jeffreys (1939), commented: "an hypothesis that may be true is rejected because it has failed to predict observable results that have not occurred. This seems a remarkable procedure. On the face of it, the evidence might more reasonably be taken as evidence for the hypothesis, not against it. The same applies to all the current significance tests based on P-values".

6.1.1 Example: analysis of birth data

Example from Gelman [?]: the proportion of female births in Germany is 0.485. In a study of a rare condition of pregnancy it was observed that in 980 of such births, 437 were female. That's 0.4459184, which is a little lower than expected. How much evidence this gives for the claim that the proportion of female births in such conditions is lower than in the large population? Assuming uniform prior probability for the female proportion r, the posterior density becomes

$$\pi(r \mid X = 437, N = 980) = \text{Beta}(438, 544).$$

The posterior mean of r is 0.446, and the posterior standard deviation 0.016. The median is 0.446, (`qbeta(0.5,438,544)`). The probability $P(r < 0.485)$ is

$$P(r < 0.485) = \texttt{pbeta(0.485,438,544)} = 0.992826$$

which seems quite high. This result was obtained when the prior was uniform. We can check how much difference does it make if the prior would be more concentrated around population mean 0.485.

$\frac{\alpha}{\alpha+\beta}$	$\alpha+\beta$	posterior median	95%posterior interval
0.5	2	0.446	$[0.415, 0.477]$
0.485	2	0.446	$[0.415, 0.477]$
0.485	5	0.446	$[0.415, 0.477]$
0.485	10	0.446	$[0.415, 0.477]$
0.485	20	0.447	$[0.416, 0.478]$
0.485	100	0.450	$[0.420, 0.479]$
0.485	200	0.453	$[0.424, 0.481]$

The prior mean is outside the 95% interval in all of these. In addition to r, an interesting quantity is the sex ratio $z = (1-r)/r$. Distribution of z could be found using the transformation of variables technique. In practice, it is easier to produce it by simulation techniques.

6.1.2 Winning Monty Hall

Monty Hall problem is a famous game in which you are first offered a choice over 3 boxes, one of which contains a prize and others are empty. Once you have made your initial choice, you are not yet allowed to open your box. Instead, one of the other boxes is shown to be empty by the game master who knows exactly what was placed in each box. You are then asked to make your final choice: do you keep your initially chosen box, or do you change for the other unopened box? The hypothesis under judgement is that A='the prize is in your box already' or B='the prize is in the other box'.

Initially, the probability to make a correct choice is $P(A) = 1/3$, hence $P(B) = 2/3$. We then need to define the conditional probabilities for the data that you'll be shown. Given that the prize is already in your box, the probability that an empty box is revealed to you is surely one: $P(\text{Monty shows empty'} \mid A) = 1$. But since Monty knows exactly the contents of all boxes, there will always be at least one empty box for him that he can reveal. So: $P(\text{'Monty shows empty'} \mid B) = 1$. Now we get $P(B \mid \text{'Monty shows empty'})$

$$= \frac{P(\text{'Monty shows empty'} \mid B)P(B)}{P(\text{'Monty shows empty'} \mid B)P(B) + P(\text{'Monty shows empty'} \mid A)P(A)} = \frac{\frac{2}{3}}{\frac{2}{3}+\frac{1}{3}} = \frac{2}{3}.$$

But let's change the rules! Assume then that Monty is allowed to choose randomly (blindfolded) which one of his boxes he opens. Now we still have $P(\text{Monty shows empty'} \mid A) = 1$, but if the prize is in the other boxes, then $P(\text{'Monty shows empty'} \mid B) = 1/2$. This will change the result:

$$= \frac{P(\text{'Monty shows empty'} \mid B)P(B)}{P(\text{'Monty shows empty'} \mid B)P(B) + P(\text{'Monty shows empty'} \mid A)P(A)} = \frac{\frac{1}{2}\frac{2}{3}}{\frac{1}{2}\frac{2}{3}+\frac{1}{3}} = \frac{1}{2}.$$

We really need to know how the game is played!

6.2 Credible Intervals

We continue with the binomial model of red balls, which led to the posterior of the unknown proportion in the form of a beta-density. Since the expected value of a Beta(α,β)-density is $\alpha/(\alpha+\beta)$, and the mode is $(\alpha-1)/(\alpha+\beta-2)$ the posterior density has mean and mode

$$E(r \mid \alpha, \beta, N, Y) = \frac{Y + \alpha}{N + \alpha + \beta}$$

$$\text{Mod}(r \mid \alpha, \beta, N, Y) = \frac{Y + \alpha - 1}{N + \alpha + \beta - 2}.$$

These are often used as bayesian *point estimates* to summarize the posterior distribution in a single number. The posterior mean can also be written as a weighted average:

$$w\frac{\alpha}{\alpha + \beta} + (1 - w)\frac{Y}{N}, \qquad w = \frac{\alpha + \beta}{\alpha + \beta + N},$$

showing how the prior and the data contribute to the estimate. This is the nice feature of conjugate priors which allows us to explore how much each source of information contributes to the result.

But these are only summaries of the posterior distribution. The full density can always be displayed graphically. Assume again that the first ball drawn is red, and the second ball is also red, but the third turns out white. We can draw the posterior density in each situation by plotting the beta-density. (But we need a software, such as R).

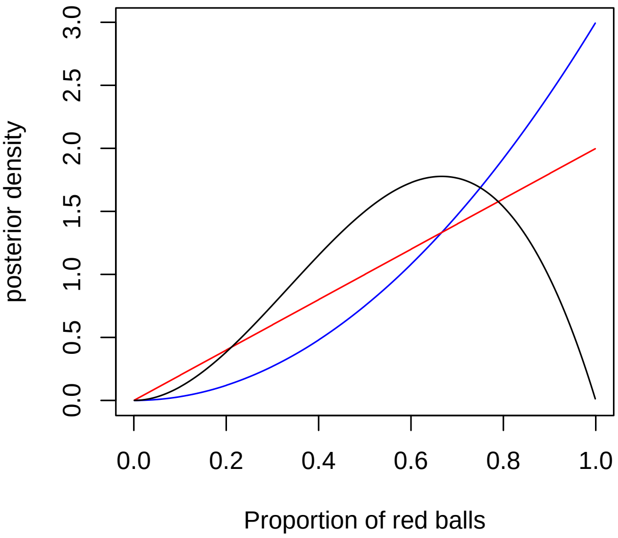


Figure 1: Posterior probability density for the proportion of red balls in an infinitely large bag of infinitely many balls, if one ball is drawn and it is red (red line), and if two balls are drawn and both are red (blue line), and if three balls are drawn and one is white (black line).

A mode shows where the distribution is mostly concentrated, but it does not convey information about how uncertain we are. This is always the problem with point summaries. Hence, variance of a distribution could be reported in addition. However, we are often required to report a region, or interval, to describe the uncertainty. From a posterior distribution we can immediately obtain intervals that contain a specific probability. The interval is usually defined so that the point estimate is somewhere in the middle, but not necessarily exactly in the middle. Any interval $[a, b]$ for which

$$\int_a^b \pi(r \mid \text{data})\mathbf{d}r = Q$$

is said to be a $Q \times 100\%$ *Credible Interval.* This is usually constructed simply by taking $Q/2$ off from both ends of the distribution. But this is not necessarily the shortest possible interval. The shortest Credible Interval is called *Highest Posterior Density Interval* (HPD-interval). The simple Credible Interval is computationally easier to obtain. For standard distributions, it can be calculated by using tabulated (or computerized) quantiles. For example, to compute the 95% CI for the posterior of r with black line in Figure (1) in R-software:

```
> qbeta(c(0.025,0.975),2+1,3-2+1)
[1] 0.1941204 0.9324140
```

And to calculate all 95% Credible Intervals of r for all possible outcomes $x \in [0, N]$:

```
N<-100; y<-0:N
lower<-qbeta(0.025,y+1,N-y+1);
upper<-qbeta(0.975,y+1,N-y+1);
plot(c(y[1],y[1]),c(lower[1],upper[1]),'l',
xlab='Red balls in a sample of N=100',
ylab='Bayesian 95% CI',
xlim=c(0,100),ylim=c(0,1));
for(i in 2:length(y)){
points(c(y[i],y[i]),c(lower[i],upper[i]),'l');
}
```

In comparison, the corresponding HPD interval of r would contain the same probability (e.g. 0.95), but we would need to find such interval that $\pi(r^* \mid X, N) > \pi(r \mid X, N)$ when r^* and r are any values within and outside the interval, respectively.

As a non-bayesian alternative, the exact frequentist 95% Confidence Interval (Clopper-Pearson interval) would be the set

$$\{r : P(Y \leq Y^{obs} \mid N, r) \geq 0.025\} \cap \{r : P(Y \geq Y^{obs} \mid N, r) \geq 0.025\}$$

which could be calculated for every outcome $y \in [0, N]$ as:

```
N<-100; y<-0:N
p<-seq(0,1,by=0.001);
I<-(1-pbinom(y[1]-1,N,p)>0.025)&(pbinom(y[1],N,p)>0.025);
lower<-min(p[I*(1:length(p))]);
upper<-max(p[I*(1:length(p))]);
plot(c(y[1],y[1]),c(lower,upper),'l',
```

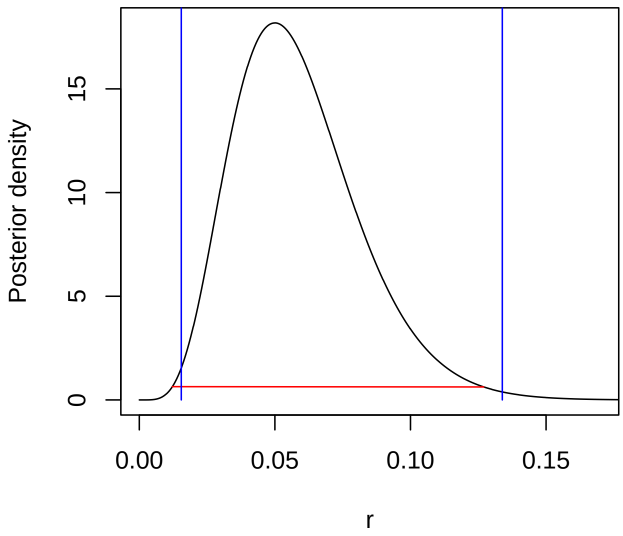


Figure 2: Comparison of HPD credible interval and simple credible interval from Beta(5+1,100-5+1) density. Red line shows 99% HPD interval. The length of 99% HPD CI is 0.1148 compared to 0.1184 of the simple 99% CI.

```
xlab='Red balls in a sample of N=100',
ylab='Freq. 95% CI',xlim=c(0,N),ylim=c(0,1));
for(i in 2:length(y)){
I<-(1-pbinom(y[i]-1,N,p)>0.025)&(pbinom(y[i],N,p)>0.025);
lower<-min(p[I*(1:length(p))]);
upper<-max(p[I*(1:length(p))]);
points(c(y[i],y[i]),c(lower,upper),'l')
}
```

The figure (3) looks very similar in both frequentist and bayesian calculations.

Note the difference of interpretation. In bayesian approach, the unknown proportion r has distribution describing our uncertainty about it. In the frequentist approach, r is fixed unknown constant, and the *interval* is random, and it *would* cover the true unknown value of r in 95% of the situations if the experiment was repeated, but it says nothing about the probability that r belongs to this interval for any given sample Y that actually did occur. (See Bayesian Theory [**?**], page 453).

The bayesian CI was solved by finding the integration limits for the posterior, such that the required probability is achieved between $[a, b]$. In general, the HPD-CI can be a set of distinct intervals if the posterior density happens to be multimodal. Numerical techniques for solving the CI's would require that we can calculate the posterior density function accurately (which was possible above).

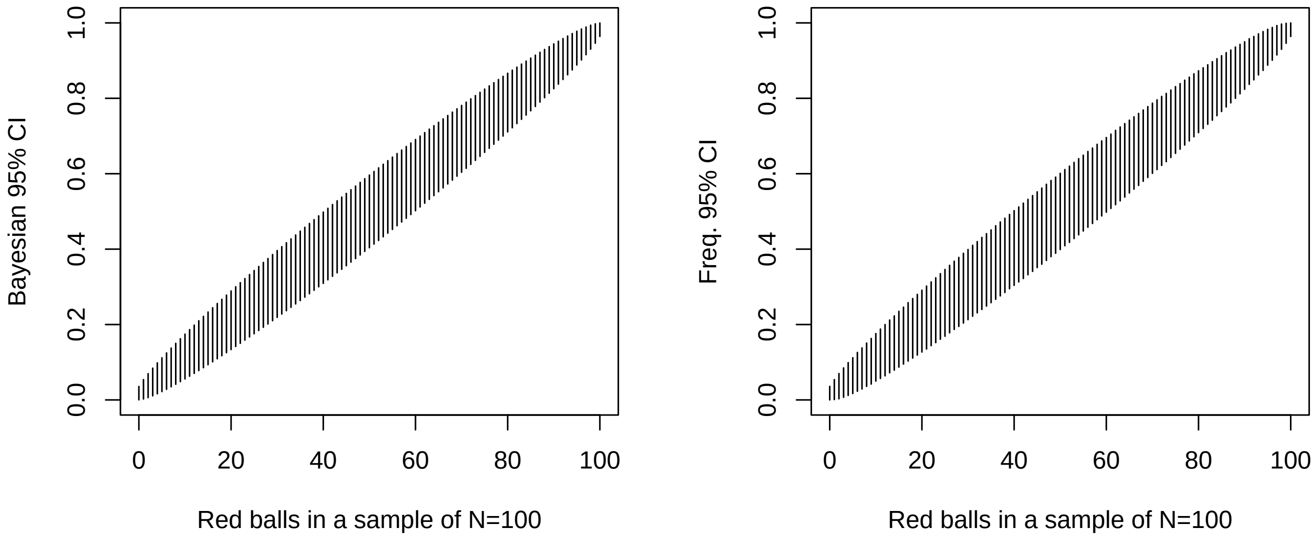


Figure 3: Bayesian Credible Intervals and frequentist Confidence Intervals.

6.2.1 The more data, the narrower CI to expect

Obviously, the resulting width of a CI depends on the amount of information we had. When the amount of data increases, we can expect the posterior to become more peaked, and hence the CI more narrow. On average, this is guaranteed because the prior variance of r can be written as

$$V(r) = E(V(r \mid X)) + V(E(r \mid X))$$

which shows that the posterior variance $V(r \mid X)$ is *expected* to be smaller than the prior variance. We can study the expected width of the CI with different sample sizes N and choose the value of N that gives the required expected width. We postpone this task until the section on Monte Carlo simulation.

6.3 Predictions

While posterior density summarizes our current uncertainty about an unknown quantity, predictions of future experiments and events could sometimes be even more interesting. (Some have even argued that it is the ultimate purpose of modeling). For example, assume that the experiment of drawing balls is to be continued after the first three balls were picked. We should then predict the color of the next ball. Our model tells us that, conditionally on r, the probability of red ball in the next draw is simply r (according to a parametric model and de Finetti). But the true value of r was unknown (and will remain unknown, representing an infinite population). In such parametric model, we could use our current estimate for the parameter, but a fixed point estimate does not account for the fact that we are still uncertain about the parameter. The bayesian approach provides a straightforward probabilistic prediction in which the uncertainty of the parameter is taken into account by integration. In general, assuming Y_i is conditionally independent of any other Y_j, given θ:

$$\pi(Y_{\text{new}} \mid Y_{\text{obs}}) = \int \pi(Y_{\text{new}} \mid \theta) \times \underbrace{\pi(\theta \mid Y_{\text{obs}})}_{\text{uncertainty of } \theta} \mathbf{d}\theta$$

The posterior predictive probability for the next ball to be red is:

$$P(\text{red} \mid Y, N) = \int_0^1 \underbrace{P(\text{red} \mid r)}_{=r} \times \underbrace{\pi(r \mid Y, N)}_{\text{Beta(Y+1,N-Y+1)}} \mathbf{d}r \;\; = E(r \mid Y, N) = \frac{Y+\alpha}{N+\alpha+\beta}$$

which is the same as the posterior mean of parameter r.

Next: consider an experiment where N new balls are to be picked, X of them will be red, so $X \sim \text{Bin}(N, r)$, and our current uncertainty about r is represented by beta-distribution $\text{Beta}(\alpha, \beta)$ (which could be the posterior of r, based on some earlier data). What is the predictive distribution of X in this new experiment?

$$P(X \mid N, \alpha, \beta) = \int_0^1 \underbrace{P(X \mid N, r)}_{\text{Bin}(N,r)} \underbrace{\pi(r \mid \alpha, \beta)}_{\text{Beta}(\alpha,\beta)} \mathbf{d}r$$

$$= \int_0^1 \frac{\Gamma(N+1)}{\Gamma(X+1)\Gamma(N-X+1)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{X+\alpha-1}(1-r)^{N-X+\beta-1} \mathbf{d}r$$

$$= \frac{\Gamma(N+1)\Gamma(\alpha+\beta)}{\Gamma(X+1)\Gamma(N-X+1)\Gamma(\alpha)\Gamma(\beta)} \int_0^1 r^{X+\alpha-1}(1-p)^{N-X+\beta-1}\mathbf{d}r$$

Then, write: $A = X + \alpha$, $B = N - X + \beta$, so that

$$= \frac{\Gamma(N+1)\Gamma(\alpha+\beta)}{\Gamma(X+1)\Gamma(N-X+1)\Gamma(\alpha)\Gamma(\beta)} \underbrace{\int_0^1 \frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)} r^{A-1}(1-r)^{B-1}\mathbf{d}r}_{=1} \frac{\Gamma(A)\Gamma(B)}{\Gamma(A+B)}$$

$$= \frac{\Gamma(N+1)\Gamma(\alpha+\beta)}{\Gamma(X+1)\Gamma(N-X+1)\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(A)\Gamma(B)}{\Gamma(A+B)}$$

$$= \binom{N}{X} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(A)\Gamma(B)}{\Gamma(A+B)}$$

which can also be written using so called *beta-functions:*

$$\binom{N}{X} \frac{\text{beta}(A, B)}{\text{beta}(\alpha, \beta)}$$

This distribution of X is said to be *beta-binomial* distribution. It is sometimes used e.g. in food safety microbial risk assessments to describe e.g. the number of contaminated servings X among N servings, under uncertainty about the true fraction, r, of contaminated servings in a large (infinite) population. In risk assessment literature, the conditional distribution of X (binomial distribution) is often called as the variability distribution of X, and the distribution of r (beta distribution) as the uncertainty distribution of r. Hence, it is often said in RA-literature that 'variability and uncertainty are separated'. In bayesian context, both distributions are expressions of uncertainty (perhaps epistemic uncertainty and aleatoric uncertainty), and the resulting beta-binomial distribution reflects both uncertainties. This can be either prior predictive distribution, or posterior predictive distribution.

As a side step, consider a situation in which we pick N new balls, but assuming that each of the balls is picked from a different population (e.g. different bags) so that for each draw we have Bernoulli-distribution with different parameter r_i. ($\mathrm{Bin}(1, r_i)$). Our uncertainty about all r_i is assumed to be described as some distribution $\pi(r_i)$, (which could be $\mathrm{Beta}(\alpha, \beta)$). What is the distribution of X?

$$P(X \mid N) = \int_0^1 P(X \mid r_1, \dots, r_N) P(r_1, \dots, r_N) \mathbf{d}r_1 \dots \mathbf{d}r_N$$

$$= \int_0^1 \cdots \int_0^1 \binom{N}{X} \prod_{i=1}^{X} r_{k_i} \prod_{i=N-X}^{N} (1 - r_{k_i}) \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r_{k_i}^{\alpha-1} (1 - r_{k_i})^{\beta-1} \mathbf{d}r_{k_1} \dots \mathbf{d}r_{k_N}$$

Here, $k_1, \dots, k_N$ is some permutation of the indices i. After re-arranging the terms in this expression, we get:

$$\binom{N}{X} \int_0^1 \cdots \int_0^1 \prod_{i=1}^{X} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r_{k_i}^{\alpha+1-1} (1 - r_{k_i})^{\beta-1} \prod_{i=N-X}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r_{k_i}^{\alpha-1} (1 - r_{k_i})^{\beta+1-1} \mathbf{d}r_{k_1} \dots \mathbf{d}r_{k_N}$$

and by integrating over each r_i one by one, we get:

$$= \binom{N}{X} E(r_i)^X E(1 - r_i)^{N-X} = \mathrm{Bin}\Big(N, \frac{\alpha}{\alpha+\beta}\Big)$$

This is a distribution that depends on N and the expected value of r_i, so the prior distribution of r_i affects the result via its expected value only. Basically, a version of the same problem is discussed in 'The BUGS project: Evolution, critique and future directions' by Lunn et al, SIM 2009, in the context of Bernoulli model:

```
y[i] ~  dbern(mu[i])
logit(mu[i]) <- dnorm(m.mu,p.mu)
```

which does not make sense because Bernoulli models cannot be over-dispersed because variance is determined by mean. (Actually, the code shown in the paper contains also a syntax error `logit(mu[i]) <- dnorm(m.mu,p.mu)` but that's a different story).

6.4 Approximating posterior density

Posterior density can be approximated by a normal distribution

$$\pi(\theta \mid X) \approx \mathrm{N}(\hat{\theta}, [I(\hat{\theta})]^{-1}),$$

where $\hat{\theta}$ is the posterior mode and $I(\theta)$ is so called *observed information*

$$I(\theta) = -\frac{\mathbf{d}^2}{\mathbf{d}\theta^2} \log \pi(\theta \mid X).$$

The approximation is based on Taylor series expansion of $\log \pi(\theta \mid X)$ centered at the posterior mode, $\hat{\theta}$. For a scalar valued θ this is

$$\log \pi(\theta \mid X) = \log \pi(\hat{\theta} \mid X) + \underbrace{[\frac{\mathbf{d}}{\mathbf{d}\theta} \log \pi(\theta \mid X)]_{\theta=\hat{\theta}}}_{=0} \frac{(\theta - \hat{\theta})}{1!} + [\frac{\mathbf{d}^2}{\mathbf{d}\theta^2} \log \pi(\theta \mid X)]_{\theta=\hat{\theta}} \frac{(\theta - \hat{\theta})^2}{2!} + \cdots,$$

where the first derivative at posterior mode $\hat{\theta}$ is zero. When θ is near the mode, the higher order terms are small compared to the first terms. As a function of θ, the first term in the expression is constant whereas the 2nd order term is proportional to the logarithm of a normal density, which provides the approximation. For a vector valued θ, the Taylor series would be

$$\log \pi(\theta \mid X) = \log \pi(\hat{\theta} \mid X) + \frac{1}{2}(\theta - \hat{\theta})^T \Big[\frac{\mathbf{d}^2}{\mathbf{d}\theta^2} \log \pi(\theta \mid X)\Big]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \cdots.$$

The normal approximation can be a useful benchmark and it gives a quick approximation of the posterior density. For final results, more accurate computations are usually needed. Even so, the first rough estimates can be obtained from the approximation, if only as realistic starting values for more complicated calculations.

6.5 Comment

The above examples required exact or approximate solutions to integrals and posterior densities. The purpose of the examples was to demonstrate how bayesian inference is merely a matter of applying probability calculus to practical problems of quantifying uncertainty. If exact analytical solution becomes difficult to find 'by paper and pencil', the integrals can be calculated using numerical methods available in different softwares. But there are also other ways to approximate. At this point, it is useful to introduce the numerical technique that has unleashed the power of probabilistic modeling: Monte Carlo simulation.

6.6 Exercises

1. Calculate the result of Monty Hall problem assuming that there are N boxes instead of three.

2. What is the *prior predictive* distribution for the number of red balls in N draws, if you assume an 'infinite bag' with proportion of red balls $r \in [0,1]$ and a $\mathrm{U}(0,1)$ prior for r? You can use these results: beta function $\mathrm{Beta}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\mathbf{d}t = \Gamma(x)\Gamma(y)/\Gamma(x+y)$, and $\Gamma(n) = \Gamma(n-1)n$, and $\Gamma(n) = (n-1)!$. (This prior predictive distribution was actually used by T. Bayes to derive the uniform prior, so he was really thinking prior uncertainty in terms of observable predicted outcomes, not directly in terms of the unobservable r).

3. Continue Exercise 3.3. by computing the approximate posterior density using the normal approximation. Plot both densities.
Hint: `plot(la,dgamma(la,a,b),type='l')`, `plot(la,dnorm(la,mu,sig),type='l')`.

4. Write out the *principle* (as a draft of the algorithm) for computing the HPD interval for a unimodal density. (You can implement the algorithm too if you like). Hint: you can imagine the posterior density as a mountain under the ocean, and then the sea level is gradually lowered.

5. Using the R-function for calculating posterior of N, calculate the 95% 'HPD interval' (although the posterior in this case is not a density, and the "interval" is a set of integer values).
Hint: use `A<-sort(p,index.return=TRUE)`, then `attach(A)`, and then check what is `x` and `ix`. Operate with `p` and `ix`. Assume the data: `X<- 1;r<- 0.2;M<- 100`.

6. Disease monitoring. The unknown population prevalence of a disease is p. A random sample of N individuals is drawn from the population. Each individual is tested for the disease (resulting to '+' or '-'), but the test has sensitivity $p_1 = 0.8$, and specificity $p_2 = 0.9$. In other words: $p_1 = P(+ \mid \text{disease})$, $p_2 = P(- \mid \text{no disease})$. The data consist of the number of test positives, $X = 2$, among $N = 100$ tested. The probability of test positive is then $\theta = pp_1 + (1-p)(1-p_2)$. So, the conditional distribution of the data is $\mathrm{Binomial}(N,\theta)$. Previously we solved the posterior of θ (assuming $\mathrm{U}(0,1)$-prior) as $\mathrm{Beta}(X+1, N-X+1)$, but now due to sensitivity and specificity: $0.1 = (1-p_2) \leq \theta \leq p_1 = 0.8$. Hence, we must assume $\mathrm{U}(1-p_2, p_1)$-prior. The posterior of θ is then

$$\pi(\theta \mid X, N, p_1, p_2) = \frac{\mathrm{Beta}(\theta \mid X+1, N-X+1)}{\int_{1-p_2}^{p_1} \mathrm{Beta}(\theta \mid X+1, N-X+1)\mathbf{d}\theta} \qquad ,\theta \in [1-p_2, p_1]$$

What is the posterior density function of the transformed variable p? Calculate the density numerically in R. Hint: use the theorem of variable transform for $p = g(\theta)$. For evaluating the posterior of θ shown above, calculate the normalizing constant in R by `C<-integrate(dbeta,1-p2,p1,X+1,N-X+1)`, then `attach(C)`, and then apply `value` for the constant. Write the posterior of θ as a function in R, and use that for numerical calculation, applying the variable transform theorem.

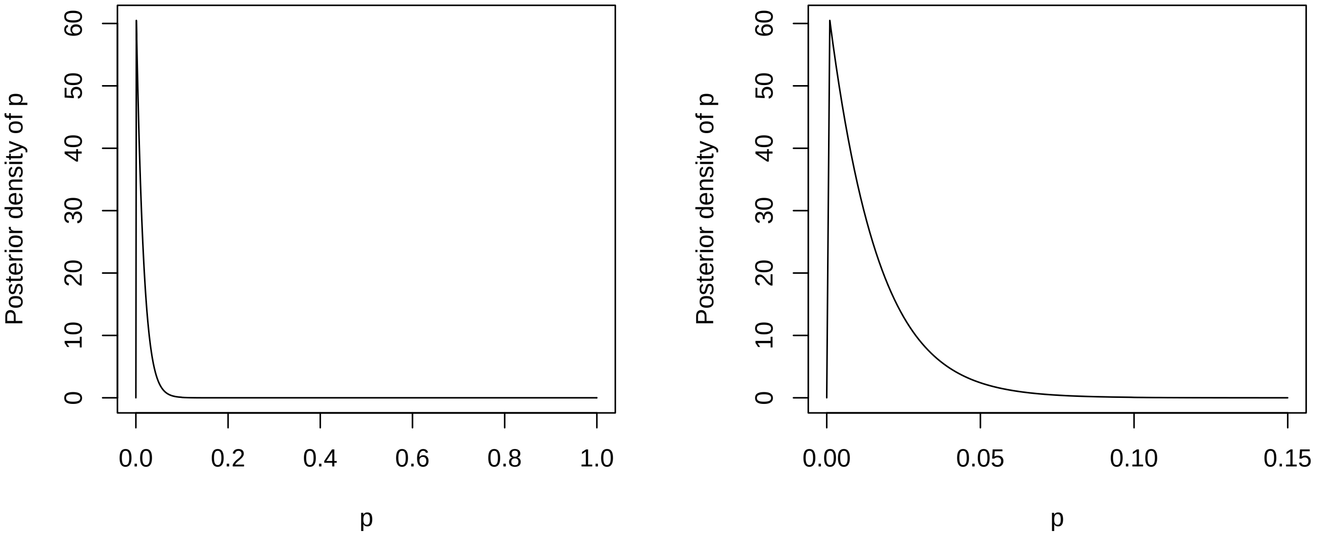


Figure 4: Posterior of the transformed $p = g(\theta)$. Left: the whole density, right: close-up of the peak.

7 Monte Carlo simulation

To begin, the word 'simulation model' is slightly obscure or even misleading. Simulation itself is not a model, but simulation is used as a technique for approximate computations of a particular model. In our case, this model is some probability density, that can be multidimensional. Therefore, the purpose here is *posterior simulation*. Simulation is naturally computer intensive. That is, someone has to program an algorithm for doing it. And it can be more laborious than the specification of the model to be simulated. Perhaps for that reason, attention is (too) easily diverted from the discussion of the *model* to the discussion of the *algorithm* - as the 'simulation model' or 'computer model'. But these are different things! There can be many different algorithms for simulating the same model. Sometimes, it may just be difficult to see from different implementations of the simulation code, that the actual model is essentially or exactly the same. The model is intrinsically related to the scientific question, whereas simulation is a tool to compute something interesting out of the model.

In bayesian inference, simulation is an essential tool because we can then use an infinite variety of models instead of the simple ones that can be solved analytically by using conjugate priors.

7.1 Example: $\pi = 3.14159\ldots$

Any estimation problem could be approached by simulation. For example, the calculation of $\pi = 3.14159\ldots$. Assume a square of size $L \times L$. Assume as big a circle that can fit inside the square. The circle has radius $L/2$. The proportion of the area of the circle to the area of the square is then:

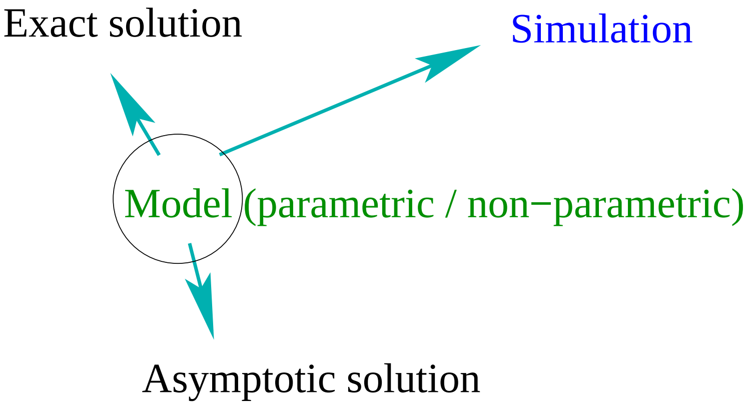


Figure 5: Model and its computation

$$q = \frac{\pi (L/2)^2}{L^2} = \frac{\pi}{4}$$

Imagine then that we can play with darts and our target is the square so that the darts can fall *evenly randomly* all over the square. The probability to hit within the circle is then $q = \pi/4$. After n darts, we count how many times we hit the circle. The percentage of hits, q_n, is an approximation of the exact number $\pi/4$. Therefore,

$$\lim_{n\to\infty} q_n = q,$$

and we can approximate $\pi \approx 4q_n$. This shows the essential principle of Monte Carlo simulation. The more darts we simulate, the more accurate is our approximation. There would be other ways to compute approximations of π, and they can be much more efficient. Monte Carlo simulation is a tool that is usually used when no other tool can help, or when we are lazy to think of alternatives and the computers are conveniently available...

In this example, the model was a 2-dimensional uniform density of variables (X, Y) over a rectangle $[-L/2, L/2] \times [-L/2, L/2]$, and the probability we approximated was

$$P(X^2 + Y^2 < r^2).$$

Simulation in R could be done as:

```
X <- runif(1000000)
Y <- runif(1000000)
q <- sum(X^2+Y^2<1)/1000000
4*q
```

An early example of approximating π by simulation is the Buffon's needle experiment (*Georges Louis Leclerc Comte de Buffon 1707-1788*). In the experiment, we first make parallel lines at equal intervals on a flat surface. Then, a needle is 'randomly' dropped on the surface and we count how often the needle crosses a line.

"Monte Carlo -method" and its systematic use started from A-bomb research, dating back to 1944. The theory was developed by e.g. Fermi, Metropolis & Ulam.

7.2 Simple Monte Carlo simulation of binomial distribution

As stated above, the binomial probability of outcome X is:

$$P(X \mid N, r) = \binom{N}{X} r^X (1-r)^{N-X}$$

There are many readily available tools in different statistical packages that can be used to draw random samples from standard distributions. For example in R:

```
X<-rbinom(1000,N,r)
hist(X,freq=FALSE,min(X):max(X),
xlab='X',ylab='Probability',
main='Empirical distribution')
```

will generate 1000 random values of X from binomial(N, r) and plot the empirical distribution. The more samples, the more accurately the empirical distribution will represent the true distribution.

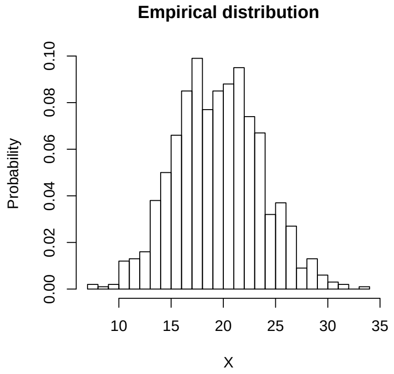


Figure 6: Simulation result of 1000 draws from binomial(100,0.2).

If the binomial distribution is not available, it is sufficient to have the most basic random number generator, the uniform distribution. For example in R:

`for(i in 1:1000){X[i]<-sum(runif(100)<0.2)}`.

This fundamental (uniform) distribution is usually found in most software as the basic random number generator, usually in the form of U(0,1). Some of the generators may not be of good quality, though, depending on the algorithm used. As we know, 'random values' are not truly random in computers.

Instead, they are pseudo-random, which means that they are produced by some deterministic algorithm. Once we have a (good) generator of U$(0,1)$ variables, then, in principle, we can generate all other distributions more or less efficiently and more or less accurately. And once we can simulate some random variable X from its distribution, it is very easy to obtain empirical distributions of any transformations of X. With Monte Carlo simulation, we don't need to solve the analytical form of the probability density of the transformed variable $g(X)$.

7.2.1 Bernoulli probabilities simulated?

Sometimes, programming of 'simulation models' can lead to superfluous models if one is not aware of the mathematical model. For example, the Bernoulli model with 'individual probabilities' explained earlier. Dispersion parameter for such probability is superfluous because Bernoulli variables only depend on the mean, and the variance is a function of the mean.

For a practical verification, you can try making a loop over $i = 1, ..., n$, and generate an 'individual' probability p_i for each trial from a distribution Beta(α, β), and then generate the Bernoulli variable $y_i \sim$ Bern(p_i), and finally compute the sum $x = \sum_1^n y_i$. Repeat this many times and plot the histogram of x. Check that you get the same whenever $\alpha/(\alpha+\beta)$ remains the same. Compare this with the histogram produced by simulating x from Bin$(n, \alpha/(\alpha+\beta)) =$ Bin$(n, E(p))$. They are identical (apart from simulation error, of course). Both specify exactly the same mean for Bernoulli. But a different result is obtained if $q \sim$ Beta(α, β) and $x \sim$ Bin(n, q), which is the beta-binomial model presented earlier.

7.3 Inverse cdf -method

A simple Monte Carlo simulation of a continuous density can be based on solving the inverse of the cumulative density function (cdf). If the cumulative probability function $y = F(x)$ can be inverted analytically, $x = F^{-1}(y)$, then we can draw a value y from uniform density $U(0,1)$ and calculate $x = F^{-1}(y)$. The resulting random values X will have cumulative probability F. Proof: $P(F^{-1}(Y) < x) = P(Y < F(x)) = F(x)$.

7.4 Transformation methods

In principle, it is possible to find some clever transformations of the basic uniform random variable so that the transformed variable has the required distribution. Then, it is only required that we can obtain random variables $X \sim$ U$(0,1)$, and compute the transformation $g(X)$. The problem is how to find what the transformation g should be for a given target distribution. For normal distribution, there are several known transformations, some of them work better than others. For example, if $X_1 \sim$ U$(0,1)$, and $X_2 \sim$ U$(0,1)$ independently, then

$$Y_1 = \sqrt{-2\log(X_1)}\cos(2\pi X_2)$$

$$Y_2 = \sqrt{-2\log(X_1)}\sin(2\pi X_2)$$

can be used as independent random draws from N$(0,1)$ distribution. The idea of transformations is also used in data-analysis when the data distribution appears 'non standard'. After a suitable transformation, it can be made approximately normal, in which case a normal distribution might be chosen as the conditional distribution of data.

7.5 Monte Carlo with standard distributions

Many standard distributions are available in statistical software as R, and we could also use those in WinBUGS/OpenBUGS. These could be used for simulating data from given distributions, or in bayesian inference to simulate from known posterior distribution. This is always possible when using conjugate priors. Just plug in the appropriate parameter values for the standard distribution. With direct Monte Carlo method you can simulate any quantities of interest, and get approximate posterior means, medians, modes, and CIs.

For this, you may need the technical material about analytical solutions of posterior distributions given earlier for binomial, multinomial, poisson, gamma, and normal models.

So, now you need to remember, for e.g. binomial model that:

(1) if prior is $r \sim \text{Beta}(\alpha, \beta)$
(2) if data model is $x \sim \text{Binomial}(n, r)$
(3) then posterior is $r \sim \text{Beta}(x + \alpha, n - x + \beta)$
(4) use any software available to sample from this posterior distribution.
(5) monitor any quantity of interest, $f(r)$, see the Monte Carlo sample histogram.

Price: for each problem, you need to solve the posterior probability density.

7.5.1 Example: virus contamination

Description of problem (adapted from course material from D. Draper 2004): a large vat contains several million litres of milk that is known to be contaminated with some virus, but the level of contamination is unknown. Then, 50 samples, each 1 litre, are taken from the vat and tested. The testing is so reliable that it gives positive result if the number of virus particles is at least 1. It is not possible to distinguish the exact number of virus particles in a positive sample. Problem (1): estimate the concentration of virus particles if there were 7 positive tests.

Assume that one needs to consume at least 8 virus particles in a short time to become infected. Problem (2): if m litres is consumed in a short time, what is the probability of infection? You can try different values for m in the range 4-10.

Estimating concentration in milk

The probability model for the number of virus particles X in a sample volume S litres can be assumed Poisson(λS), where λ is the mean concentration per litre. If the sample size is one litre, then the conditional probability of having no particles is $P(X = 0 \mid S = 1, \lambda) = \exp(-\lambda)$. Therefore, the probability of a positive test result is

$$p = P(X \geq 1 \mid S = 1, \lambda) = 1 - \exp(-\lambda),$$

which also specifies λ as a function of p: $\lambda = -\log(1-p)$. When 50 samples are taken, the probability of 7 positives is Binomial

$$P(Y = 7 \mid N = 50, p) = \text{Binomial}(Y = 7 \mid 50, p).$$

Assuming a uniform prior for p, we get the posterior of p as

$$\pi(p \mid Y = 7, N = 50) = \text{Beta}(7 + 1, 50 - 7 + 1) = \text{Beta}(8, 44).$$

To obtain the posterior distribution of λ by Monte Carlo simulation, just draw many random samples of p from this beta-distribution, and then calculate λ from each sampled value of p, and draw the histogram. (Try using R, WinBUGS, or any suitable software).

Probability of infection

The probability of infection is the probability of having at least 8 particles, in a consumption volume of m.

$$P(\text{infection} \mid \lambda, m) = P(X \geq 8 \mid \lambda, m),$$

where $X \sim \text{Poisson}(\lambda m)$. The probability of infection can be expressed more easily as

$$(\spadesuit)\;\; 1 - P(X \leq 7 \mid \lambda, m) = 1 - \sum_{i=1}^{7} \frac{(m\lambda)^i \exp(-m\lambda)}{i!}.$$

Since λ is uncertain, this expression is also uncertain, and we obtain the final *probability* as

$$P(\text{infection} \mid Y = 7, N = 50, m) = \int_0^\infty P(\text{infection} \mid \lambda, m)\pi(\lambda \mid Y = 7, N = 50)\, \mathbf{d}\lambda.$$

In other words: taking the weighted average of ($\spadesuit$), weighting by the posterior density of λ. Using Monte Carlo, this integral can be calculated approximately by simulating p from the beta density, then calculating λ for each sampled p, and calculating ($\spadesuit$) for each sampled λ, and finally taking the average of the simulated sample of ($\spadesuit$).

Alternatively, if ($\spadesuit$) is interpreted as the proportion of 'servings' (of size m) with infective dose in a large population of all servings produced from this vat, then we might also report the distribution of the *proportion* ($\spadesuit$) with a bayesian CI.

7.6 Rejection sampling

This is a method that produces independent samples from the target distribution, but *indirectly* by generating from an instrumental distribution and discarding some of the results. Therefore, it is not so efficient as direct Monte Carlo method applied to a target density.

The goal is to draw random samples from target density $\pi(x)$. Assume that we have no easy way of sampling directly from it - but we can easily compute the values of that density function for any x. In rejection sampling, we need to find a function $g(x)$ so that g is a probability density over the same support as π, or at least the integral of g is finite so that it can be normalized to one, and

- we can draw samples from $g(x)$ (or a density proportional to g),
- the importance ratio $\pi(x)/g(x)$ must have a known bound. I.e. there must be some known constant M so that $\pi(x)/g(x) \leq M$ for all values x.

The algorithm is then:

1. Sample X from $g(x)$ (or a density proportional to g).
2. With probability $\pi(x)/(Mg(x))$, accept x as a new draw from π, otherwise return to step 1.

This technique requires some preliminary work to select a suitable density g, so we would need to inspect the target density to some extent. At best, if g is closely the same as π, this method can be efficient.

7.7 Markov chain Monte Carlo (MCMC)

This is a version of Monte Carlo sampling in which the consecutive samples are not independent. Hence, if x_i and x_{i+1} are two consecutive samples from the same target distribution, then in Monte Carlo sampling $P(x_i, x_{i+1}) = P(x_i)P(x_{i+1})$, but in MCMC: $P(x_i, x_{i+1}) = P(x_{i+1} \mid x_i)P(x_i)$.

MCMC is based on constructing a Markov chain so that its stationary distribution is the required target distribution.

This means that the ordinary Monte Carlo method is more efficient than MCMC. But in many problems, direct Monte Carlo is not possible. Instead, MCMC is extremely general method and can be widely applied for computing e.g. posterior densities. We only need to be able to compute values of the (unnormalized!) target density at all points x.

But MCMC needs to be applied carefully since there is no guarantee that the results are automatically correct after a *finite* number of iterations. We need to check for possible convergence problems. The general MCMC algorithm is of the form:
1. Set initial value x_1. Set counter $i = 1$.

2. Generate next value, conditionally on the previous: $x_{i+1} \sim f(x \mid x_i)$, set counter $i = i + 1$.

3. Go back to 2, until required sample size is obtained.

There are many different versions of MCMC algorithms, e.g. slice sampling, Gibbs-sampling, Metropolis-algorithm, Metropolis-Hastings algorithm.

MCMC methods always require some form of *convergence diagnostics*. It may happen that the sampler is simply stuck in one part of the parameter space. Then the resulting MCMC sample would not represent the target distribution! Therefore, it is good to investigate using other techniques what the posterior density might look like. Also, it is good to use several different MCMC runs with different starting values. (WinBUGS convergence diagnostics is based on this idea). It is always good to think of simple point estimates that could be calculated from the data. Obviously, sensible point estimates should be roughly in the region where the posterior distribution is. It is not good idea to apply MCMC blindly and take the results automatically, without double checking. However, all convergence diagnostics can only indicate lack of convergence. They can never prove that the target distribution is really correctly and accurately represented by the MCMC sample. The most problematic situations occur if the posterior is multimodal and multidimensional.

7.7.1 Slice sampling

Example: simulating from a truncated normal distribution:

$$\pi(x \mid \mu, \sigma^2)1_{\{x>L\}} \quad \propto e^{-0.5(x-\mu)^2/\sigma^2}1_{\{x>L\}},$$

This could also be done by simulating random draws from an untruncated density, and then accepting only values greater than L. But if L is large, then this algorithm would be running a long time before we have a reasonable sample. Slice sampling is actually one version of Gibbs sampling. The iteration step is as follows (Robert CP, Casella G: Monte Carlo Statistical Methods. Springer 1999):

$$\begin{array}{lll} \text{Step 1:} & x_t \mid z_{t-1} & \sim \mathrm{U}\Big(L, \mu + \sqrt{-2\sigma^2 \log(\sigma z_{t-1}\sqrt{2\pi})}\Big) \\ \text{Step 2:} & z_t \mid x_t & \sim \mathrm{U}\Big(0, \frac{1}{\sqrt{2\pi\sigma}} \exp(-0.5(x_t-\mu)^2/\sigma^2)\Big). \end{array}$$

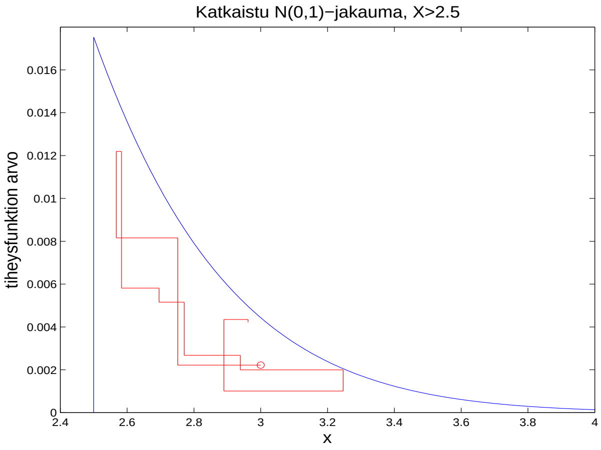
After some iterations, $(t = 1, 2, 3, \ldots)$, we obtain simulated draws $x_1, x_2, x_3, \ldots$. In these iterations, variable z is purely *instrumental variable*. In fact, slice sampling relies on *completion* of the target density $\pi(x)$ into $\pi(x, z)$, so that

$$\int \pi(x, z)\, \mathbf{d}z = \pi(x).$$

In this example:

$$\pi(x, z) \propto 1_{\{x>L\}}1_{\{0<z<\frac{1}{\sigma\sqrt{2\pi}}\exp(-0.5(x-\mu)^2/\sigma^2)\}},$$

which is a 2-dimensional uniform density over the area in x, z-plane below the density function $\pi(x)$, in the region where $x > L$.



7.7.2 Gibbs sampling

Gibbs sampling is also known as *alternating sampling.* Slice sampling is a special case of Gibbs sampling. To demonstrate Gibbs sampling, consider a simple 2D normal density:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\Big)$$

Recall that the 2D normal density function (mean zero, unit variance) is

$$\pi(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\Big).$$

It can be written in the form $\pi(x \mid y)\pi(y)$ or $\pi(y \mid x)\pi(x)$ since the marginal, and conditional, densities can be solved from the joint density:

$$\pi(x) = \int_{-\infty}^{\infty} \pi(x,y)\mathbf{d}y = N(0,1)$$

$$\pi(y) = \int_{-\infty}^{\infty} \pi(x,y)\mathbf{d}x = N(0,1)$$

and

$$\begin{aligned} \pi(y \mid x) &= \frac{\pi(x,y)}{\pi(x)} \\ &= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\Big)}{\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}(\rho x - y)^2) \quad = \quad \mathrm{N}(\rho x, 1-\rho^2) \end{aligned}$$

and similarly: $\pi(x \mid y) = \mathrm{N}(\rho y, 1-\rho^2)$.

Monte Carlo sampling from $\pi(x,y)$:

Sample x from $\pi(x) = \mathrm{N}(0,1)$, then sample y from $\pi(y \mid x) = N(\rho x, 1-\rho^2)$. Continue until required sample is collected. Each sampled pair of (x,y) is an independent draw from the joint distribution. (Same can be done by sampling first y from $\pi(y) = \mathrm{N}(0,1)$, and then x, given y).

Gibbs sampling from $\pi(x,y)$:

Pick some initial value for both $(x,y) = (x_0, y_0)$. Then, sample x_1 from $\pi(x \mid y_0)$. After that, sample y_1 from $\pi(y \mid x_1)$. Continue alternating in this way, sampling the next value of x (or y) given the previous sampled value of y (or x), until required sample is collected. Each sampled pair of (x,y) is NOT independent of the previous sampled values. In the long run, the empirical distribution of the large sample will approximate the target density. If the initial values (x_0, y_0) were in the fringe of the target density, it may take many steps before the sampled values represent the target density.

Gibbs sampling for diagnostics model

Assume we have a sample of N individuals which are all tested for some disease. We observe X of them positive. We wish to estimate the unknown disease prevalence p, but knowing that the diagnostic test has sensitivity q, which also is uncertain. The model for the observation is then

$$P(X \mid p, q) \propto (pq)^X (1 - pq)^{N-X}$$

Assume for simplicity $\mathrm{U}(0, 1)$ priors for both p and q independently. The posterior for p, q is then proportional to the above expression, but it would not be possible to extract a simple distribution for p, given fixed q, or for q, given fixed p. This does not seem to lead to simple conditional distributions needed for Gibbs. However, if we introduce latent variables

T_1 = unknown number of truly positives among X test positives
T_2 = unknown number of false negatives among $N - X$ test negatives

Note: $T_1 + T_2$ is the number of all truly positives among all tested $N = X + (N - X)$. Using these latent variables, we can write the results of the single test as a 2×2 table, in which only the row sums are uniquely identified from data:

	True		
Test	+	−	
+	T_1	$X - T_1$	X
−	T_2	$N - X - T_2$	$N - X$.

For this table, we can write the cell probabilities according to the model parameters p, q:

	True		
Test	+	−	
+	pq	0	
−	$p(1-q)$	$(1-p)$	.

Note: bayesian model is the joint distribution $\pi(\text{parameters}, \text{data})$, from which the posterior is obtained as $\pi(\text{parameters} \mid \text{data})$. In this example, the joint distribution of all unknown parameters, latent variables, and data can be written using the cell probabilities as:

$$\begin{aligned} \pi(p, q, T_1, T_2, X, N - X) &= \pi(T_1, T_2, X, N - X \mid p, q)\pi(p, q) \\ &\propto (pq)^{T_1} 0^{X-T_1} [p(1-q)]^{T_2} (1-p)^{N-X-T_2} \underbrace{\pi(p)}_{=1} \underbrace{\pi(q)}_{=1} \end{aligned}$$

By re-arranging the terms, we can write it as:

$$\underbrace{p^{(T_1+T_2+1)-1}(1-p)^{(N-X-T_2+1)-1}}_{\mathrm{Beta}(T_1+T_2+1,\, N-X-T_2+1)} \times \underbrace{q^{T_1+1-1}(1-q)^{T_2+1-1}}_{\mathrm{Beta}(T_1+1,\, T_2+1)}$$

It is now easy to recognize beta-distributions (up to normalizing constant) for parameters p, q. By re-arranging the terms slightly differently, the same joint probability can be written also in the form:

$$\Big[\frac{pq}{pq}\Big]^{T_1}\Big[\frac{0}{pq}\Big]^{X-T_1} \times \underbrace{[pq+0]^X}_{\text{constant with respect to } T_1 \text{ and } T_2}$$

$$\times\Big[\frac{p(1-q)}{p(1-q)+(1-p)}\Big]^{T_2}\Big[\frac{(1-p)}{p(1-q)+(1-p)}\Big]^{N-X-T_2} \times \underbrace{[p(1-q)+(1-p)]^{N-X}}_{\text{constant with respect to } T_1 \text{ and } T_2}$$

from which the binomial distributions can be recognized for T_1 and T_2. For T_1 we have trivially $T_1 = X$ as a constant, but for T_2 we get a binomial distribution. Gibbs sampling can be based on alternating sampling of all these unknown quantities.

Gibbs sampling algorithm for d-dimensional density:

Assume we have a d-dimensional unknown vector $\theta_1, \ldots, \theta_d$, with the target density $\pi(\theta_1, \ldots, \theta_d \mid X)$. Gibbs sampling is based on solving the **full conditional** densities (fcd) so that these can be easily sampled at each iteration step t:

$$\pi(\theta_j \mid \theta_{-j}^{t-1}, X),$$

where θ_{-j}^{t-1} represents all elements of vector θ at iterations step $t-1$, apart from the single element θ_j.

$$\theta_{-1}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1}).$$

Each iteration step consists of a cycle, in which every element θ_j is sampled in turn, from the full conditional distribution, that is conditional to all other elements θ_k, $k \neq j$ at their current values. Of course, the distributions are all conditional to the data X, (if our aim is to sample from a posterior).

7.7.3 Metropolis-Hastings algorithm

This is a very general tool that can be used for simulating from complicated distributions, as long as we check it's working properly. For the algorithm, we need to choose a *proposal distribution* Q which is used to generate a proposed value x^* for the next round of iteration, that depends on the values at the previous step. The proposal density is thus of the form $Q(x^* \mid x_{i-1})$. The proposed value becomes accepted with probability

$$r = \min\Big(\frac{\pi(x^* \mid \text{data})Q(x_{i-1} \mid x^*)}{\pi(x_{i-1} \mid \text{data})Q(x^* \mid x_{i-1})}, 1\Big).$$

If it becomes accepted, it will be the new generated value for x at this current iteration step. Otherwise, the previous value remains also for the current step.

As seen from the acceptance probability formula, it is sufficient that we are able to compute the posterior density without normalizing constant. This is the innovation!

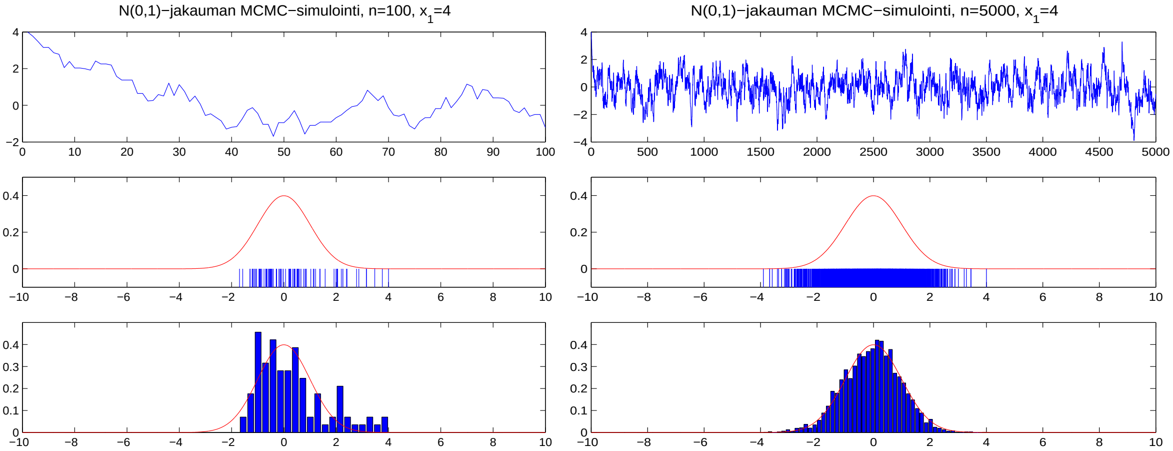
All terms that are constants with respect to x will cancel out from the ratio. The algorithm can be slow if the proposal density is badly chosen, either too narrow or too wide. In practice, the acceptance probability is often computed in logarithmic form which makes fractions and products into a summations that are easier to compute. As a special case, we obtain Gibbs-sampler where the acceptance probability is always one.

7.7.4 Metropolis algorithm

Metropolis algorithm is a special case of M-H algorithm, in which the proposal density Q is symmetric. Therefore, it cancels out from the acceptance probability.

Example: MCMC simulation of N(0,1)

Assume the target distribution is $\mathrm{N}(0,1)$ and choose the proposal density as uniform distribution centered around the previous value $\mathrm{U}(x_{i-1}-L, x_{i-1}+L)$. The value of the proposal density is now simply constant and we only need to compute the ratio of two normal densities $\mathrm{N}(x^* \mid 0,1)/\mathrm{N}(x_{i-1} \mid 0,1)$.



7.7.5 Convergence diagnostics

Although it is never possible to exactly prove from the obtained MCMC sequence that convergence has been achieved, it is reasonably possible to detect poor convergence. The convergence diagnostics can be done in various ways, although each of them is imperfect.

1. Common sense! Based on knowing your problem, you should have an idea what are reasonable estimates to expect from any analysis. If the MCMC sampler produces something very different, be suspicious! It may not have converged, (or there may be an error in the model definition).

2. See the history plots of the MCMC sample paths. Do they look stable? Or is there a drifting trend still after 10^5 iterations? Do the parameter values move at all along the simulation?

3. Check the autocorrelations in your MCMC chain. (In WinBUGS: check the 'auto cor' button). If the autocorrelations are high, this indicates that your MCMC chain is slowly mixing. Therefore, longer chain is needed to get more representative sample of the posterior. To achieve more like independent Monte Carlo sampling, thinning of the original MCMC iterations can be used. (In WinBUGS: this can be done either by not saving all iterations while running, e.g. save only every tenth, or by afterwards using only every tenth value for final results).

4. Check the results after increasing the number of iterations for some 10^k extra iterations. Are they still the same? (In WinBUGS: check also the automatic estimate of Monte Carlo error).

5. Run the MCMC with some different initial values. Do you still get the same results?

6. Compute some formal convergence diagnostics. (Editors/referees of journals may request them!). These are based on some aspects of the long run frequentist properties of the MCMC sample. (Some people think this is the only use of frequentist techniques in bayesian analysis!). The common diagnostics of BGR is based on simulating at least $K \geq 2$ independent MCMC chains each with different (overdispersed) starting values, and then simulating n iterations from each. Then, within sequence and between sequence variances are computed. The modified method of BGR is available in WinBUGS/OpenBUGS. You just need to generate at least two chains with o. The outline of the method is grossly the following.

Let's say the variance of the posterior distribution (our target) is $\sigma^2 = \int (x-\mu)^2 \pi(x \mid \text{data}) \mathbf{d}x$, where μ is the true posterior mean. Let x_{ki} be the sampled value in chain k at iteration i, and $\bar{x}$ be the overall mean, and $\bar{x}_k$ the mean of the k:th chain. The sample variance in each of the chains is $s_k = \frac{1}{n-1}\sum_i (x_{ki} - \bar{x}_k)^2$. Denote $W = s_k/K$, the average within chain sample variance. This may be used as an estimate of σ^2. Also, note that the sample variance, σ^2_M, of the means $\bar{x}_k$ is σ^2/n. Hence, $\sigma^2 = n\sigma^2_M$. Now, the sample variance of the means is $\sigma^2_M = \frac{1}{K-1}\sum_k (\bar{x}_k - \bar{x})^2$. Hence, we also get the estimate for σ^2 as $B = n\frac{1}{K-1}\sum_k (\bar{x}_k - \bar{x})^2$. Finally, define the weighted estimate of σ^2 as

$$\hat{\sigma}^2_+ = (1 - \frac{1}{n})W + \frac{1}{n}B$$

This would be an unbiased estimate of σ^2 if the starting values were already obtained from the target distribution, but is an overestimate if the starting values are from an overdispersed distribution. In the book of Gelman et al [?] they describe the 'potential scale reduction factor' as $\hat{R} = \sqrt{\hat{\sigma}^2_+/W}$ and Brooks et al refined the original method further. The modified method is implemented in BUGS. In WinBUGS/OpenBUGS: check the automatic BGR diagnostic tool for graphical output. The factor should approach 1 as $n \rightarrow \infty$.

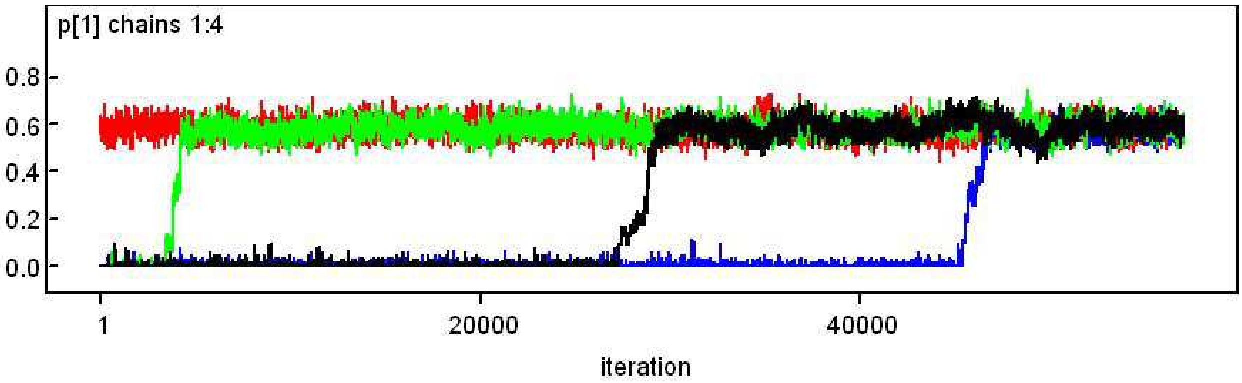


Figure 7: Amazing: convergence finally achieved in all MCMC chains!

7.8 Exercises

1. Using the inverse cumulative distribution method, generate Monte Carlo simulations of the exponential density $\pi(x) = \lambda \exp(-\lambda x)$.

2. Verify the result of Monty Hall problem by simulation.

3. Assume the prevalences of a disease in two different populations are p_1 and p_2. A small sample of individuals is observed from both, (N_1, N_2). In these observed samples, X_1 and X_2 individuals have the disease. Choose some value for p_1 and p_2. Then choose the sample sizes N_1, N_2 and simulate the observed data X_1 and X_2. Then simulate the posterior (beta)density of p_1 and p_2 and the difference $q = p_1 - p_2$. Study if there was evidence for $p_1 < p_2$ by computing $P(q < 0 \mid X_1, N_1, X_2, N_2)$.

4. Write out the principle of a simulation algorithm which generates predicted values of some X from a distribution $\pi(X \mid \theta)$ accounting for uncertain θ. How the prediction uncertainty could be described if we were not allowed to use probability density for θ? (i.e. if using non-bayesian methods where θ is unknown constant so that there is no distribution for it).

5. Implement the Gibbs sampler for the simple 2D normal density $N(0, \Sigma)$ with unit variances and some correlation ρ. You can use any software available. How quickly you get a representative sample, if ρ is nearly ± 1, and if the starting value is 'in a bad corner'.

6. Assume 2D-normal model with mean (μ_1, μ_2) and covariance matrix

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Solve the full conditional densities for a Gibbs sampler.

7. Continue Example 3.5. Compute by simulation the posterior probability that μ_1 (Ahonen) is larger than μ_2 (Janda). Using only the results of the first 7 competitions, simulate the posterior of (μ_1, σ_1) and (μ_2, σ_2). Then, simulate predicted values X_1, X_2 for the last competition. From this predictive distribution (based on 7 competitions), compute the probability that the difference of Tournament total points is less than 1, given the total sum of the previous 3 games of the Tournament. Was the actual final result surprising at that point?

8. Simulate posterior density of λ, the mean concentration of virus per litre, based on the previous example. Simulate also the posterior of proportion of servings with infective dose. What is the posterior probability of infection?

9. Simulate X from beta-binomial distribution, by using Gibbs sampler. (step 1): $X \sim \text{Bin}(N, r)$, (step 2): $r \sim \text{Beta}(X + \alpha, N - X + \beta)$. You can implement this in R, or any suitable software.

10. Simulate X from beta-binomial, but using Monte Carlo sampling in two steps at each iteration: $r \sim \text{Beta}(\alpha, \beta)$, $x \sim \text{Bin}(N, r)$. You can implement this in R, or any suitable software.

11. Simulate from $\pi(p_1, p_2)$ defined as uniform over the plane $[0, 1] \times [0, 1]$ restricted in the area where $p_1 < p_2$. Hence, $\pi(p_1, p_2) = 2 \times 1_{\{p_1 < p_2\}}(p_1, p_2)$. Implement (1) a rejection sampler and (2) alternating

sampler, and compare the results.

12. Assume the model $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$ and the prior $\pi(\mu, \sigma^2) \propto 1/\sigma^2$. Verify that the posterior is then $\pi(\mu, \sigma^2 \mid Y_{1,\ldots,n}) \propto (\sigma^2)^{-(n/2+1)} \exp(-0.5 \sum (Y_i - \mu)^2/\sigma^2)$. If σ^2 is assumed as known, this simplifies to the case where $\pi(\mu \mid \sigma^2, Y_{1,\ldots,n}) = \mathrm{N}(\bar{Y}, \sigma^2/n)$. Verify from the table of distributions that: if μ is assumed as known, we get $\pi(\sigma^2 \mid \mu, Y_{1,\ldots,n}) = \text{Inv-Gamma}(n/2, 0.5 \sum (Y_i - \mu)^2)$. In other words, for $\tau = 1/\sigma^2$: $\pi(\tau \mid \mu, Y_{1,\ldots,n}) = \text{Gamma}(n/2, 0.5 \sum (Y_i - \mu)^2)$. Construct a Gibbs sampler (e.g. in R) and simulate joint posterior for μ, σ^2, with some data $Y_{1,\ldots,n}$ you have first generated with some selected 'true values' μ_0, σ_0^2.