# Inference Tutorial 8

1. Consider again the grasshopper survival data from tutorial 7, but this time using a Bayesian approach. The data are observations of, $x$, the length of time in excess of 14 days that it took for a random sample of 10 grasshoppers infected with a fungal disease to die, where the times are measured from infection.

   `1.6 12.1 5.2 6.2 6.8 9.6 17.8 24.4 4.1 7.8`

   It is believed that a gamma distribution may be a good model here, so that the p.d.f. of $x$ is

   $$f(x) = \frac{\beta^\alpha}{(\alpha-1)!}x^{\alpha-1}e^{-x\beta} \quad x > 0$$

   (the parameterization here is slightly different to that used in tutorial 7). In this case the parameter $\beta$ could be any positive real number, but $\alpha$ can only take positive integer values (because of the requirements of the simulation model in which the estimated parameters will eventually be used). Note that in functions `rgamma` and `dgamma`, $\alpha$ is the `shape` parameter and $\beta$ is the `rate` parameter.

   (a) Show that a gamma prior for $\beta$ is conjugate. i.e. that if $\beta$ has a gamma prior then the posterior for $\beta$ is also a gamma distribution. Also find the expressions for the parameters of this distribution in terms of the data, $x_i$.

   (b) Write an MCMC loop to sample from the posterior distribution of $\alpha$ and $\beta$, assuming a gamma prior for $\beta$ (`shape`=1 and `rate`=.01) and an improper discrete uniform distribution on the positive integers for $\alpha$. You should use a Gibbs step to update $\beta$ and a Metropolis Hastings step to update $\alpha$. For the MH step, you can use a random walk proposal with with equal probability of increasing or decreasing $\alpha$ by one (but don't forget that $\alpha \leq 0$ is impossible according to the prior).

   (c) Plot your simulated $\alpha$ values against MCMC iteration, and do the same for the $\beta$. Hence check that your chain seems to have converged, and decide on how many of the samples should be discarded as 'burn-in' at the start of the simulation.

   (d) Also use the `acf` function to examine auto-correlation of the samples from the chain. If you wanted to sub-sample the chain by retaining every $m^{\text{th}}$ sample, what value of $m$ would you take in order to be able to treat the samples as approximately independent?

   (e) Use the `quantile` function to find an approximate 95% credible interval for $\beta$.

   (f) Why can you not (sensibly) find a 95% credible interval for $\alpha$? Use the `tabulate` function to evaluate the approximate posterior probability mass function for $\alpha$. How might you sensibly report confidence intervals for $\alpha$?

   **Solution**

   (a) The p.d.f. of prior on $\beta$ is $\pi(\beta) = b^a\beta^{a-1}e^{-\beta b}/(a-1)!$. The likelihood is $f(\mathbf{x}|\beta,\alpha) = \beta^{\alpha n}\prod x_i^{\alpha-1}e^{-\sum x_i\beta}/(\alpha-1)!^n$ (where $n$ is the number of data). Hence

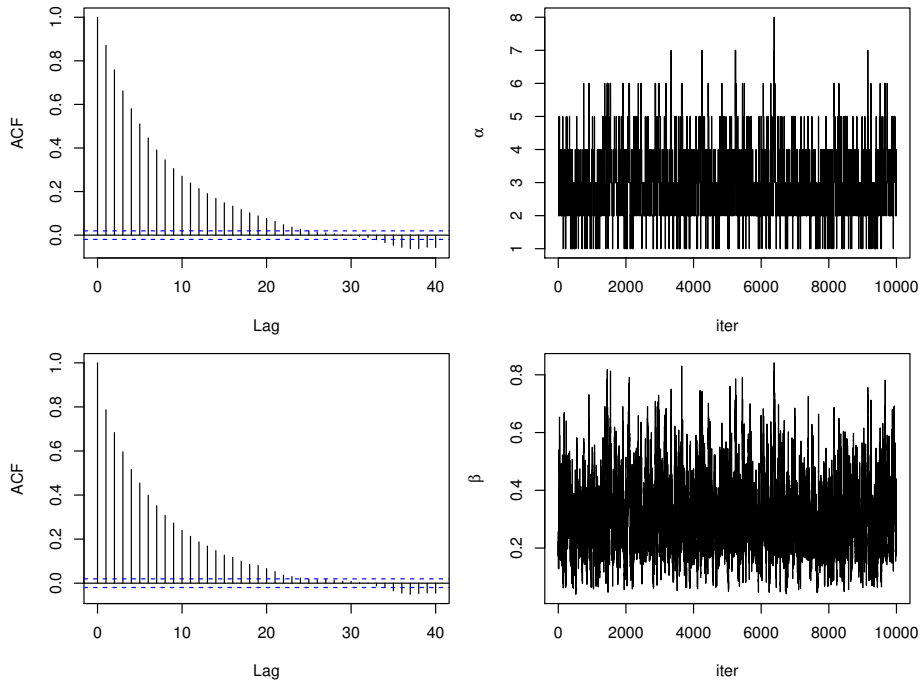
   $$\begin{aligned} f(\beta|\mathbf{x}) &\propto f(\mathbf{x}|\beta,\alpha)\pi(\beta) \\ &\propto \beta^{\alpha n}\prod x_i^{\alpha-1}e^{-\sum x_i\beta}b^a\beta^{a-1}e^{-\beta b}/\{(a-1)!(\alpha-1)!^n\} \\ &\propto \beta^{n\alpha+a-1}e^{-\beta(\sum x_i+b)} \end{aligned}$$

   which we can recognise as a gamma p.d.f. with shape parameter $n\alpha + a$ and scale parameter $\sum x_i + b$.

   (b)
   ```
   x <- c(1.6,12.1,5.2,6.2,6.8,9.6,17.8,24.4,4.1,7.8)
   sx <- sum(x)
   n <- length(x)
   a <- 1;b <- 0.01 ## gamma prior parameters for beta
   accept <- 0
   n.rep <- 10000
   post <- matrix(0,2,n.rep) ## storage for simulated alpha/beta
   alpha <- 1
   for (i in 1:n.rep) {
     beta <- rgamma(1,shape=n*alpha+a,rate=b + sx) ## Gibbs step
     alpha.prop <- alpha + sample(c(-1,1),1) ## random walk alpha proposal
   ```

```
  if (alpha.prop>0) lfp <- sum(dgamma(x,shape=alpha.prop,rate=beta,log=TRUE))
  lf <- sum(dgamma(x,shape=alpha,rate=beta,log=TRUE))
  if (alpha.prop>0 && exp(lfp-lf)>runif(1)) { ## MH accept/reject
    accept <- accept + 1
    alpha <- alpha.prop
  }
  post[,i] <- c(alpha,beta) ## store simulated values
}
accept/n.rep ## acceptance rate (around .3)
```

(c)
```
par(mfrow=c(2,2),mar=c(4,4,1,1)) ## c) and d) plots together
acf(post[1,])
plot(post[1,],type="l",ylab=expression(alpha),xlab="iter")
acf(post[2,])
plot(post[2,],type="l",ylab=expression(beta),xlab="iter")
```




From the trace plots (right), convergence looks fairly rapid (almost immediate). Dropping the first 1000 as burn-in should be more than adequate.

(d) The ACFs suggest that independence would be a reasonable assumption by about lag 30, so $m = 30$ would be OK (we don't want to make $m$ higher than necessary as this leads to a reduced sample size). Really the ACFs should have been produced after discarding burn-in, but it makes no difference to the answer here.

(e)
```
burn <- 1:1000
quantile(post[2,-burn],c(0.025,.975))
     2.5%     97.5%
0.1035392 0.5896429
```

(f) Because the parameter takes discrete values we can not find an interval that has exactly a 95% probability of including the truth.

```
pa <- tabulate(post[1,-burn])/(10000-length(burn))
pa
[1] 0.0612222222 0.3577777778 0.3457777778 0.1655555556 0.0551111111
[6] 0.0134444444 0.0010000000 0.0001111111
```

It makes more sense to construct intervals such that values inside the interval are more probably than those outside it, and to give the exact probabilities associated with those. For example,

```
sum(pa[2:4]);sum(pa[1:5])
[1] 0.8691111
[1] 0.9854444
```

implies that $[2, 4]$ is an 87% CI for $\alpha$, while $[1 : 5]$ is a 98.5% CI.

2. Suppose you are in your first job as a statistician. Depending on your preference, you can imagine that this is working for the department of health advising on the sugar tax, or at an investment bank, deciding whether to buy shares in alcoholic drinks producers, or at a brewers deciding whether or not to invest in a massive new lager brewery. Your boss would like you to assess the following paper, which she has read about in the media...

   `http://jech.bmj.com/content/early/2018/01/11/jech-2017-209791#DC1` is a recent paper on the possible impact of tax increases on sugary drinks. The link to `jech-2017-209791-SP1.pdf` gets you to the description of the model actually used (it's basically a linear model with some methods being applied that avoid problems with having an excess of zeros for some of the data). The results are stated to be

   > An increase in the price of high-sugar drinks leads to an increase in the purchase of lager, an increase in the price of medium-sugar drinks reduces purchases of alcoholic drinks, while an increase in the price of diet/low-sugar drinks increases purchases of beer, cider and wines. Overall, the effects of price rises are greatest in the low-income group.

   which is what received media attention. By examining the paper, comment on the paper statistically, and decide whether you believe the stated results.

   **Solution** There's no absolutely right or wrong answer here, but there are several issues that would make me worry about taking the results at face value.

   (a) The paper basically looks at the how the price of one group of products effects the consumers tendency to buy that product or other products. This is done using a very specific set of linear modelling assumptions, without much in the paper to indicate how well these assumptions are met. Parts of the description are not clear enough for me to be able to repeat the analysis even given the data. My impression is that this combines quite alot of quite restrictive assumptions into the model, without sufficient checking, so I fear that the results may be driven as much by the assumptions as by the data. Hence, I'm not completely confident that the associations reported are real.

   (b) Obviously this is an observational study and even if the associations are real, establishing causality here is unlikely to be possible. So you really would not want to set policy or make investment decisions based on this.

   (c) A large number of elasticities are reported, and many appear 'significant' , because the sample size is so high (have they corrected for multiple testing?). However, some do not seem very plausible from a causal perspective: why on earth should high income groups greatly reduce their consumption of cider in response to increasing the price of high sugar drinks, for example? This sort of thing tends to support the sceptical view in a) above.