Bayesian Inference: Tutorial 2

1. Consider binary data where success is denoted with $X = 1$, and failure with $X = 0$. Suppose we observe x successes from n trials, so that $x_1 = 0, x_2 = 1, ..., x_n = 1$, $\sum_{i=1}^{n} x_i = x$. Given p, the observations are i.i.d. with, $P(X_i = 1|p) = p$.

   a) When the exact 0 or 1 observation for each one of the variables is modelled, the likelihood of the data is,

   $$P(\text{data}|p) = p(x_1, ..., x_n|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^x(1-p)^{n-x}.$$

   Assuming a Beta prior for p, $Beta(a, b)$, derive the marginal distribution for $\boldsymbol{x}$, $f(\boldsymbol{x})$. Check that this is a valid probability mass function for $n = 1$.

   b) When the total number of successes is modelled, the likelihood of the data is the Binomial distribution,

   $$P(\text{data}|p) = p(x_1, ..., x_n|p) = \binom{n}{x} p^x(1-p)^{n-x}.$$

   Assuming the same Beta prior for p, $Beta(a, b)$, derive the marginal distribution for $\boldsymbol{x}$, $f(\boldsymbol{x})$. Check that this is a valid probability mass function for $n = 1$. Are the two marginal densities the same?

   Hint: You could use directly Bayes Theorem, with the posterior distribution derived in the lecture notes.

2. Suppose that we specify a prior on the parameter θ such that,

   $$\begin{aligned} \theta|\lambda &\sim Po(\lambda) \\ \lambda &\sim \Gamma(\alpha, \beta), \end{aligned}$$

   where α and β are known values. By integrating out λ (i.e. by using $p(\theta) = \int_{-\infty}^{\infty} p(\theta, \lambda)d\lambda = \int_{-\infty}^{\infty} p(\theta|\lambda)p(\lambda)d\lambda$), show that this is equivalent to the prior,

   $$\theta \sim Neg - Bin(\alpha, \beta).$$

   where α and β are the negative binomial shape and inverse scale (sometimes called rate) parameters respectively. (Note that distributional information is given in one of the Appendices in the Lecture Notes section in Moodle.)

   Now suppose that we observe data x, such that, given θ and p,

   $$X \sim Bin(\theta, p).$$

   Derive an expression for the corresponding posterior distribution for θ. (Note - the distribution is not of a standard form)

3. (From December 2012 exam - number in brackets correspond to number of marks - the total exam is out of 50 marks.) A biologist designs an experiment in order to investigate the variability in the recorded observations from some laboratory equipment. It is assumed that the observations, denoted $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, independently follow a normal distribution $N(1, \sigma^2)$, where σ^2 is unknown.

a) Consider an inverse Gamma $\Gamma^{-1}(\alpha, \beta)$ prior for σ^2. Show that the posterior distribution of σ^2 given the data is given by,

$$\sigma^2|\boldsymbol{x} \sim \Gamma^{-1}\left(\frac{n}{2} + \alpha, 0.5\sum_{i=1}^{n}(x_i - 1)^2 + \beta\right).[4]$$

b) Consider the two following priors: (i) $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$; and (ii) $\sigma^2 \sim U(0, K)$, where α, β, K are known. Comment on the differences between the two priors and any impact they may have on the posterior distribution of σ^2. [2]

c) Consider data $x_1 = 5, x_2 = 7, x_3 = 7, x_4 = 9$ and the prior $\sigma^2 \sim \Gamma^{-1}(1, 2)$. State (i) the posterior distribution for σ^2 and (ii) the posterior mean for $\frac{1}{\sigma^2}$. [2]

d) For the same data as in (c) above, consider the prior $\sigma^2 \sim U(0, 100)$. Calculate the posterior distribution for σ^2. [4]

4. (From December 2012 exam - number in brackets correspond to number of marks - the total exam is out of 50 marks.)
Consider a sequence of independent Bernoulli trials with probability of success p. A Bayesian experiment is designed in order to obtain inference on p where the data correspond to the number of failures, y, before a pre-determined total number of m successes. Conditional on p, Y has a negative Binomial distribution with probability mass function,

$$p(y|p, m) = \binom{m + y - 1}{y} p^m (1 - p)^y,$$

for $y = 0, 1, \ldots, \infty$.

a) Calculate Jeffreys' prior for p. [*Note:* $E_Y(Y) = m(1 - p)/p$]. [4]

b) When the observed data are the total number of successes after a set number of independent Bernoulli trials, Jeffreys' prior for p is $Beta(0.5, 0.5)$. Discuss the implication of the given result in relation to your calculation in (a) and the Likelihood Principle. [2]

5. We observe data $\boldsymbol{x}$ from a Poisson distribution, with unknown mean μ. Calculate the Jeffreys' prior for μ. Then, what is the corresponding posterior distribution for μ, after observing data $\boldsymbol{x}$? Suppose that we observe data:

$$5,\ 6,\ 5,\ 6,\ 7,\ 5,\ 4,\ 5,\ 3,\ 6.$$

Calculate the posterior mean for μ.

6. (From 2009 exam - number in brackets correspond to number of marks - the total exam is out of 50 marks.) Let $X_1, \ldots, X_n$ be independent and identically distributed $Po(\lambda)$ random variables.

(a) Suggest **two** uninformative priors that could be specified for the parameter λ, explicitly giving the form of the prior probability density functions. Comment on the advantages and disadvantages of the two suggested priors. [6]

Bayesian Inference

Tutorial 2: Solutions

1. a) From Bayes Theorem,

$$f(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|p)p(p)}{\pi(p|\boldsymbol{x})} = \frac{p^x(1-p)^{n-x}\frac{1}{B(a,b)}p^{a-1}(1-p)^{b-1}}{\frac{1}{B(x+a,n-x+b)}p^{x+a-1}(1-p)^{n-x+b-1}}$$

$$= \frac{B(x+a,n-x+b)}{B(a,b)}$$

For $n = 1$, for this to be a valid probability mass function, we require $P(X = 0) + P(X = 1) = 1$. Now,

$$P(X=0) + P(X=1) = \frac{B(a,b+1)}{B(a,b)} + \frac{B(a+1,b)}{B(a,b)} = \frac{B(a,b+1) + B(a+1,b)}{B(a,b)}.$$

But,

$$B(a,b+1) + B(a+1,b) = \int_0^1 z^a(1-z)^{b-1}dz + \int_0^1 z^{a-1}(1-z)^b dz$$

$$= \int_0^1 z^{a-1}(1-z)^{b-1}(z+(1-z))dz = B(a,b).$$

Thus,

$$P(X=0) + P(X=1) = \frac{B(a,b)}{B(a,b)} = 1.$$

b) The derivation is almost identical to (a), with only the Binomial term $\binom{n}{x}$ appearing in the nominator,

$$f(\boldsymbol{x}) = \frac{\binom{n}{x}B(x+a,n-x+b)}{B(a,b)}.$$

Showing that this is a valid pmf is again an almost identical derivation, as $\binom{1}{0} = 1$ and $\binom{1}{1} = 1$.
The two marginal densities in (a) and (b) are different.

2. The prior for θ is given by,

$$\begin{aligned}
p(\theta) &= \int_0^\infty p(\theta|\lambda)p(\lambda)d\lambda \\
&= \int_0^\infty \frac{\lambda^\theta \exp(-\lambda)}{\theta!} \times \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\beta\lambda)d\lambda \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)\Gamma(\theta+1)}\int_0^\infty \lambda^{\theta+\alpha-1}\exp(-(\beta+1)\lambda)d\lambda \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)\Gamma(\theta+1)} \times \frac{\Gamma(\theta+\alpha)}{(\beta+1)^{\theta+\alpha}}\int_0^\infty \frac{(\beta+1)^{\theta+\alpha}}{\Gamma(\theta+\alpha)}\lambda^{\theta+\alpha-1}\exp(-(\beta+1)\lambda)d\lambda \\
&= \frac{\Gamma(\theta+\alpha)}{\Gamma(\alpha)\Gamma(\theta+1)}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^\theta
\end{aligned}$$

since the integral is of a $\Gamma(\alpha+\theta, \beta+1)$ pdf.

Note that $\frac{\Gamma(\theta+\alpha)}{\Gamma(\alpha)\Gamma(\theta+1)} = \frac{(\theta+\alpha-1)*(\theta+\alpha-2)*\cdots*(\alpha)}{\theta!} = \binom{\theta+\alpha-1}{\theta}$. Comparing this expression with the pmf of a Negative Binomial (See Appendix A in the lecture notes), we have that,

$$\theta \sim Neg - Bin(\alpha, \beta).$$

Now, we have that,

$$f(x|\theta) = \frac{\theta!}{x!(\theta-x)!}p^x(1-p)^{\theta-x}.$$

So that, by Bayes Theorem, for $\theta = x, x+1, ...$

$$\begin{aligned}
\pi(\theta|x) &\propto \frac{\theta!}{(\theta-x)!}(1-p)^\theta \frac{\Gamma(\theta+\alpha)}{\Gamma(\theta+1)}\left(\frac{1}{\beta+1}\right)^\theta \\
&= \frac{\Gamma(\theta+1)\Gamma(\theta+\alpha)}{\Gamma(\theta-x+1)\Gamma(\theta+1)}\left(\frac{1-p}{\beta+1}\right)^\theta \\
&= \frac{\Gamma(\theta+\alpha)}{\Gamma(\theta-x+1)}\left(\frac{1-p}{\beta+1}\right)^\theta.
\end{aligned}$$

The constant of proportionality is then,

$$\left(\sum_{\theta=x}^{\infty}\frac{\Gamma(\theta+\alpha)}{\Gamma(\theta-x+1)}\left(\frac{1-p}{\beta+1}\right)^\theta\right)^{-1}.$$

This defines the posterior distribution for θ.

3. a) We specify,

$$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta).$$

In other words $1/\sigma^2 \sim \Gamma(\alpha, \beta)$. Then, the corresponding posterior distribution is given by,

$$\begin{aligned}
\pi(\sigma^2|\boldsymbol{x}) &\propto \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-1)^2}{2\sigma^2}\right)\times(\sigma^2)^{-(\alpha+1)}\exp\left(-\frac{\beta}{\sigma^2}\right) \\
&\propto (\sigma^2)^{-(n/2+\alpha+1)}\exp\left(-\frac{\left(\frac{1}{2}\sum_{i=1}^{n}(x_i-1)^2+\beta\right)}{\sigma^2}\right) \\
\Rightarrow \sigma^2|\boldsymbol{x} &\sim \Gamma^{-1}\left(\frac{n}{2}+\alpha, \frac{1}{2}\sum_{i=1}^{n}(x_i-1)^2+\beta\right).
\end{aligned}$$

b) The shape of an inverse Gamma distribution can be quite different to the shape of a uniform distribution which is always flat. Importantly, the uniform distribution places an upper limit on the possible values of σ^2. Therefore, even if the data strongly support values greater than K, these values will have posterior probability zero. This is not an issue with the inverse Gamma density.

c) The posterior distribution for σ^2 is an inverse Gamma density $\Gamma^{-1}\left(\frac{4}{2}+1, 76+2\right) = \Gamma^{-1}(3, 78)$.
From the above, $\frac{1}{\sigma^2} \sim \Gamma(3, 78)$. So, the posterior mean for $\frac{1}{\sigma^2}$ is $\frac{3}{78} = 0.03$.

d) We specify,

$$\sigma^2 \sim U(0, 100).$$

Then, for $0 < \sigma^2 < 100$ the corresponding posterior distribution is given by,

$$\begin{aligned}
\pi(\sigma^2|\boldsymbol{x}) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-1)^2}{2\sigma^2}\right) \\
&\propto (\sigma^2)^{-(n/2)} \exp\left(-\frac{\left(\frac{1}{2}\sum_{i=1}^{n}(x_i-1)^2\right)}{\sigma^2}\right) \\
&\propto \text{ the pdf of a } \Gamma^{-1}\left(\frac{n}{2}+1, \frac{1}{2}\sum_{i=1}^{n}(x_i-1)^2\right).
\end{aligned}$$

Therefore, the posterior distribution for σ^2 is a truncated inverse Gamma density $\Gamma^{-1}\left(\frac{n}{2}+1, \frac{1}{2}\sum_{i=1}^{n}(x_i-1)^2\right)$, where $0 < \sigma^2 < 100$. For our data, the posterior distribution for σ^2 is an inverse Gamma density $\Gamma^{-1}\left(\frac{4}{2}+1, \frac{1}{2}\sum_{i=1}^{4}(x_i-1)^2\right) = \Gamma^{-1}(3, 76)$, $0 < \sigma^2 < 100$, truncated so that the pdf is zero for $\sigma^2 \geq 100$. For the truncated pdf of σ^2 to integrate to one, we need to divide the $\Gamma^{-1}(3, 76)$ pdf with $\int_0^{100} \frac{76^1}{\Gamma(1)} v^{-2} \exp\left(-\frac{76}{v}\right) dv = 76 \int_0^{100} v^{-2} \exp\left(-\frac{76}{v}\right) dv$.

4. a)

$$\begin{aligned}
\log(f(y|p,m)) &= m\log(p) + y\log(1-p) + \text{constant} \\
\frac{d\log(f(y|p,m))}{dp} &= \frac{m}{p} - \frac{y}{1-p} \\
\frac{d^2\log(f(y|p,m))}{dp^2} &= -\frac{m}{p^2} - \frac{y}{(1-p)^2}
\end{aligned}$$

Now, it is given that $E_Y(Y) = m(1-p)/p$. Therefore,

$$I(p|y) = -E_y\left[\frac{d^2\log(f(y|p,m))}{dp^2}\right] = \frac{m}{p^2} + \frac{m}{p(1-p)} = \frac{m}{p^2(1-p)}.$$

Jeffreys' rule implies that $p(p) \propto \sqrt{I(p|y)} \propto p^{-1}(1-p)^{-0.5}$.
[Note, this is not a proper distribution.]

b) The Likelihood principle states that probability models that lead to the same likelihood function will yield the same inference for the parameter of interest.
The two priors are different. Jeffreys' rule does not allow to decide which prior to use for the probability of success until we know the design of the experiment, i.e. if we have set a fixed number of trials, or if we wait until a fixed number of successes. This clearly violates the Likelihood principle, since Jeffreys prior, and hence the inferences made using it, may be different even if the likelihood functions for the two experiments are the same. Therefore, using Jeffreys' rule is not consistent with the Likelihood principle. The key problem is that to derive Jeffreys' prior, we must take an expectation over all possible values of the data, as determined by the experimental design; note that this includes data we will get, but also data we will not. Hence data we do not obtain plays a role in inference.

5. Consider a single data point, x. Then,

$$\log f(x|\mu) = x\log\mu - \mu + C,$$

where C is a constant. Then,

$$\frac{d^2\log f(x|\mu)}{d\mu^2} = -\frac{x}{\mu^2},$$

and hence Fisher's Information is,

$$I(\mu|x) = -\mathbb{E}\left(-\frac{x}{\mu^2}\right) = \frac{1}{\mu}.$$

Thus, Jeffreys' prior is,

$$p(\mu) \propto \mu^{-\frac{1}{2}}.$$

This is an improper prior, i.e. it cannot be made to integrate to one by multiplying with a constant. For the given prior, the corresponding posterior distribution is,

$$\begin{aligned}\pi(\mu|\boldsymbol{x}) &\propto \left(\prod_{i=1}^{n}\exp(-\mu)\mu^{x_i}\right) \times \mu^{-\frac{1}{2}} \\ &= \exp(-n\mu)\mu^{n\bar{x}-\frac{1}{2}}.\end{aligned}$$

Thus, we have that,

$$\mu|\boldsymbol{x} \sim \Gamma\left(n\bar{x}+\frac{1}{2}, n\right).$$

which is a proper distribution. Then, the posterior mean for μ is,

$$\mathbb{E}_\pi(\mu) = \frac{n\bar{x}+\frac{1}{2}}{n} = 5.25$$

6. (a) Possible priors include:

   (i) Jeffreys' prior - uninformative prior which is invariant to bijective transformations. To calculate Jeffreys' prior consider log-likelihood:

   $$\log f(x|\lambda) = x\log\lambda - \lambda + C,$$

   where C is a constant. Then,

   $$\frac{d^2\log f(x|\lambda)}{d\lambda^2} = -\frac{x}{\lambda^2},$$

   and hence Fisher's Information is,

   $$I(\lambda|x) = -\mathbb{E}\left(-\frac{x}{\lambda^2}\right) = \frac{1}{\lambda}.$$

   Thus, Jeffreys' prior is,

   $$p(\lambda) \propto \lambda^{-\frac{1}{2}}.$$

   So, Jeffreys prior is an improper prior (does not integrate to one).

   (ii) Uniform prior, $\lambda \sim U[0,T]$, where T is large, so that,

   $$p(\lambda) = \frac{1}{T}, \qquad 0 < \lambda < T.$$

   Thus, this is a flat prior, with all equal length intervals (within $[0,T]$) having the same probability. However, if we consider a (non-linear) transformation of λ, given by $f(\lambda)$, then the prior on $f(\lambda)$ is non-Uniform. Also need to specify T and should check that posterior distribution of parameters does not "reach" upper bound.

   (iii) Other possible priors include for example some truncated Normal density, $N^+(0,\sigma^2)$, for σ^2 large. This is approximately flat for very large σ^2 and defined on the positive real line. Same drawback of transformations as for previous Uniform case. A flat prior with support $[0,+\infty]$ is also possible (but will be improper) Or a $\Gamma(0.001, 0.001)$.