University of Glasgow

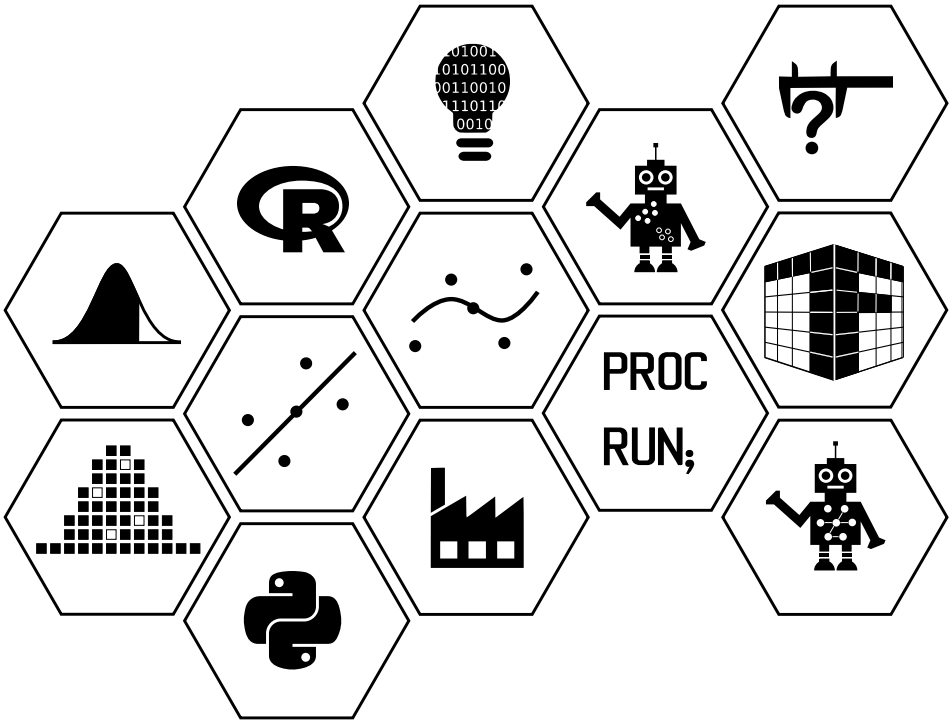
Learning from Data/Data Science Foundations

Claire Miller and Eilidh Jack

Academic Year 2021-22

Supplement - part 2:

Bayesian inference: deriving posterior priors



DATA ANALYTICS GLASGOW

Posterior distributions using conjugate priors

In the supplementary material part 1 for Bayesian inference we motivated the ideas for Bayesian inference, introduced the rules for formulating a posterior from a prior and likelihood, and defined the concept of a conjugate prior. In this supplementary material, we will use conjugate priors to derive posterior distributions for some simple examples, and briefly mention a few other important aspects of the Bayesian inference framework.

Deriving posterior distributions using conjugate priors

Example 1 (Estimating the proportion of success in a binomial distribution).

Let's return to the example of the music expert distinguishing between Mozart and Beethoven concertos by listening to the first 3 seconds of a piece of music, that we mentioned at the beginning of the supplementary part 1 material.

The expert does this correctly for 10 different pieces of music, and we assumed that the model for our data was $Y|\theta \sim \text{Bin}(10, \theta)$, where θ is the probability of success.

However, before the data were collected we had a prior assumption about θ. We might assume that the expert would do better than random chance and so we expect that θ might be around 0.7.

Let's derive the posterior distribution for $\theta|y$, using the information above.

We know that, for n trials and y successes, the binomial model states that:

$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

Considered as a function of θ, the likelihood is of the form:

$$L(\theta|y) \propto \theta^y(1-\theta)^{n-y}.$$

Suppose we assume that a priori $\theta \sim \text{Be}(\alpha, \beta)$, where Be is a beta distribution with parameters α and β:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

(See the video below for examples and discussion of the derivations here.)

The posterior probability distribution is then:

$$\begin{aligned} p(\theta|y) &\propto p(\theta)L(\theta|y) \\ &= \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^y(1-\theta)^{n-y} \\ &= \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}. \end{aligned}$$

The beta prior distribution is a conjugate family for the binomial likelihood. Therefore, the posterior distribution follows the same parametric form as the prior distribution. The conjugate family is mathematically convenient in that the posterior distribution follows a known parametric form.

Therefore, by identifying parameters for the posterior correctly from the expression above, we can write the posterior distribution for $\theta|y$ as:

$$\theta|y \sim \text{Be}(\alpha + y, \beta + n - y).$$

Since we know the form of the posterior distribution we can then find the full expression for the posterior distribution i.e. we can derive the normalising constant $p(y)$.

Therefore,

$$p(\theta|y) = \frac{p(\theta)L(\theta|y)}{p(y)},$$

i.e.

$$p(\theta|y) = \frac{\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}}{B(\alpha+y, \beta+n-y)}.$$

Notice that we have figured out the normalising constant $p(y)$ without actually solving the integral $\int p(\theta)L(\theta|y)d\theta$.

Now, let's consider the information that we have for the specific context in example 1.

From prior information we believe that the music expert will do better than simply guessing and so suppose we expect θ to lie around 0.7.

As illustrated algebraically above we are going to put a beta prior distribution on θ. We can use the `LearnBayes` package in `R` to help us establish the parameters α and β for this distribution which correspond to our prior beliefs.

In `R`, we can use the `LearnBayes` package to give us values for the parameters of the beta distribution that correspond to values from a particular quantile. Suppose a priori we believe that the median (0.5 quantile) for θ is around 0.7 and the 0.9 quantile is around 0.9.

The values for the parameters for the beta distribution for θ can be found in `R` using:

```
library(LearnBayes)
quantile2=list(p=.9, x=.9) ## p=quantile of interest, x=value of theta
quantile1=list(p=.5, x=.7)
beta.select(quantile1, quantile2)
```

```
## [1] 3.87 1.84
```

We see from this that an appropriate prior distribution for θ might be $\theta \sim \text{Be}(3.87, 1.84)$.

Combining this with the information from our data of 10 successes and 0 failures (i.e. $y = 10, n = 10$), we have:

$$p(\theta|y) = \frac{\theta^{3.87+10-1}(1-\theta)^{1.84+10-10-1}}{B(3.87+10, 1.84+10-10)},$$

i.e. our **posterior distribution** for example 1 is given by:

$$p(\theta|y) = \frac{\theta^{12.87}(1-\theta)^{0.84}}{B(13.87, 1.84)}.$$

We could now visualise the prior, likelihood and posterior distributions for example 1 in `R`.

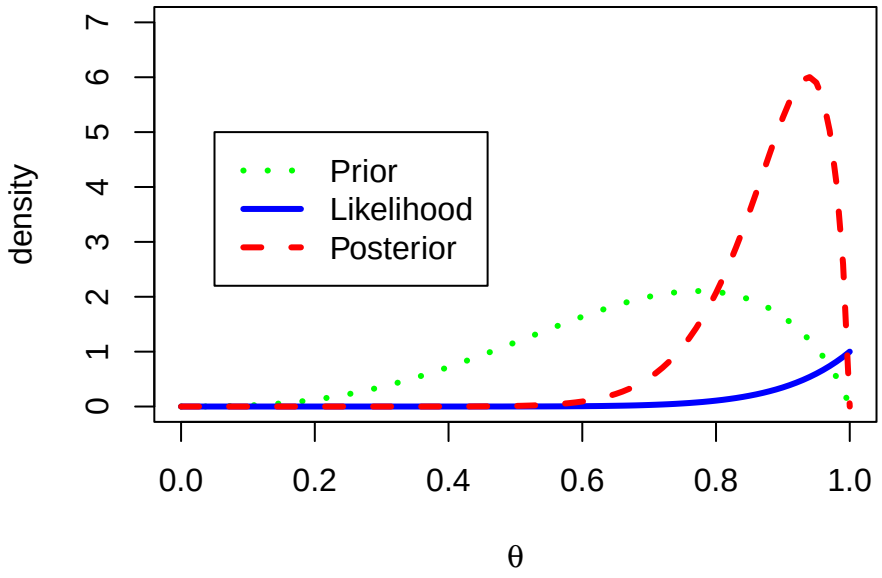
```
## The data
y <- 10  ## number of successes
n <- 10  ## number of trials


## The parameters
alpha <- 3.87  ## prior paramater 1
beta <- 1.84   ## prior parameter 2

A <- alpha+y   ## posterior parameter 1
B <- beta+n-y  ## posterior parameter 2


## Plotting prior, likelihood, posterior, where x denotes the parameter being estimated
curve(dbeta(x,alpha,beta),0,1,ylim=c(0,7),col="green", ylab="density", lwd=3, lty=3,
```

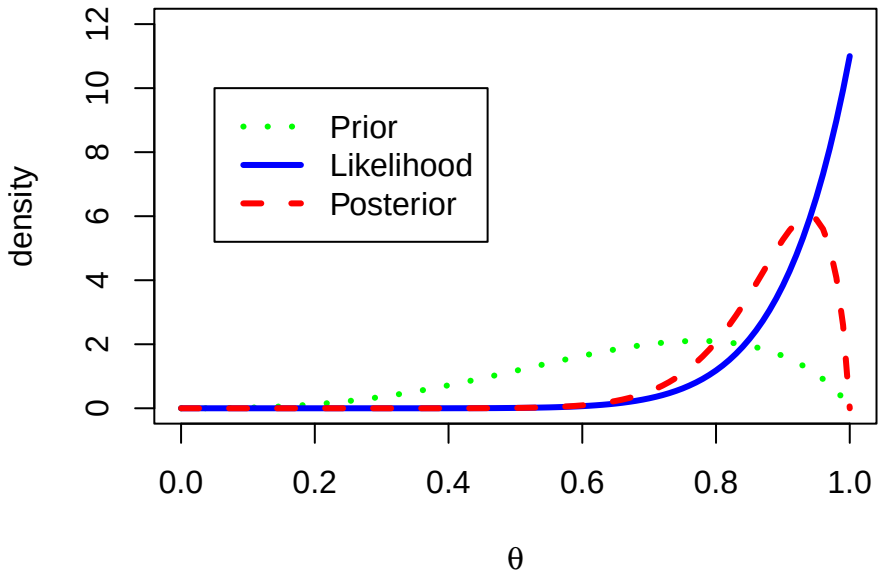
```
      xlab=expression(theta))#Prior
curve(dbinom(y,n,x),0,1,add=TRUE,col="blue", lwd=3, lty=1)#Likelihood
curve(dbeta(x,A,B),0,1,add=TRUE,col="red", lwd=3, lty=2)#Posterior
legend(0.05,5, c("Prior", "Likelihood", "Posterior"), lty=c(3,1,2), lwd=c(3,3,3),
       col=c("green", "blue", "red"))
```



We could scale the likelihood to make it easier to compare with the prior and posterior distributions. We can do this in R by using a beta distribution with parameters (number of successes+1, number of failures+1).

```
## Scaling the likelihood for easier comparison with prior and posterior

curve(dbeta(x,alpha,beta),0,1,ylim=c(0,12),col="green", ylab="density", lwd=3, lty=3,
      xlab=expression(theta))#Prior
curve(dbeta(x,y+1,n-y+1),0,1,add=TRUE,col="blue", lwd=3, lty=1)#Likelihood
curve(dbeta(x,A,B),0,1,add=TRUE,col="red", lwd=3, lty=2)#Posterior
legend(0.05,10, c("Prior", "Likelihood", "Posterior"), lty=c(3,1,2), lwd=c(3,3,3),
       col=c("green", "blue", "red"))
```



Task 1.

Continue example 1 above, but now suppose that you are in the context of the drunk person guessing the number of times a head will appear when tossing a fair coin Y. Again, $Y|\theta \sim \text{Bin}(10, \theta)$, and $y = 10$.

However, this time your prior belief is that θ should be around 0.5.

Derive the posterior distribution for $\theta|y$, using this information.

The second example below considers estimating the mean for a Poisson distribution. The first video for this week derives the posterior distribution of $\lambda|\mathbf{x}$ for a set of n observations $(x_1, \ldots, x_n)$ from a Poisson (λ) distribution.

Deriving the posterior distribution for a Poisson likelihood

https://youtu.be/LO3WvyAMcxU

Duration: 8m44s

Example 2 considers the case of one observation from a Poisson distribution.

Example 2 (Estimating the mean λ for a Poisson distribution).

Let's consider the car accidents example, which we first considered in week 5. Suppose that (Y) the number of car accidents at a fixed point on a road within a fixed time window (say one month) is 3. Suppose also that we know this road well and our prior impression of the number of car accidents is that there is on average 1 accident per month with variance 1.

Suppose Y is a sample from the Poisson distribution with mean λ, that is $Y \sim \text{Poi}(\lambda)$.

We know that a Ga(α, β) prior is a conjugate prior for the Poisson likelihood.

Derive the posterior distribution for $\lambda|y$.

The likelihood is:

$$L(\lambda|y) = p(y|\lambda) = \lambda^y \exp(-\lambda)/y!$$
$$L(\lambda|y) \propto \lambda^y \exp(-\lambda).$$

We're going to use a gamma distribution to state a prior for λ and hence our prior distribution has the form:

$$p(\lambda) \propto \lambda^{\alpha-1} \exp(-\lambda\beta).$$

The posterior is therefore,

$$p(\lambda|y) \propto p(\lambda)L(\lambda|y),$$

$$\begin{aligned} p(\lambda|y) &\propto \lambda^{\alpha-1} \exp(-\lambda\beta) \times \lambda^y \exp(-\lambda) \\ &= \lambda^{y+\alpha-1} \exp\left(-\lambda(\beta+1)\right). \end{aligned}$$

Thus the posterior distribution will be of the same general form as the prior distribution, that is to say a gamma distribution,

$$\lambda|y \sim \text{Ga}(y+\alpha, \beta+1).$$

A gamma prior is the conjugate prior for a Poisson likelihood resulting in a gamma distribution for the posterior distribution.

Therefore, the **posterior distribution** is:

$$p(\lambda|y) = \frac{p(\lambda)L(\lambda|y)}{p(y)},$$

i.e.

$$p(\lambda|y) = \frac{(\beta+1)^{y+\alpha}\, \lambda^{y+\alpha-1} \exp(-\lambda(\beta+1))}{\Gamma(y+\alpha)}.$$

Now from the context of the example we have that $y = 3$, and our prior impression of the number of car accidents is that there is on average 1 per month with variance 1.

Since λ had a gamma prior distribution we have that $\mathbb{E}(\lambda) = \frac{\alpha}{\beta} = 1$ and $\text{Var}(\lambda) = \frac{\alpha}{\beta^2} = 1$.

Therefore, we might take initial values for α and β to be 1, and we have that:

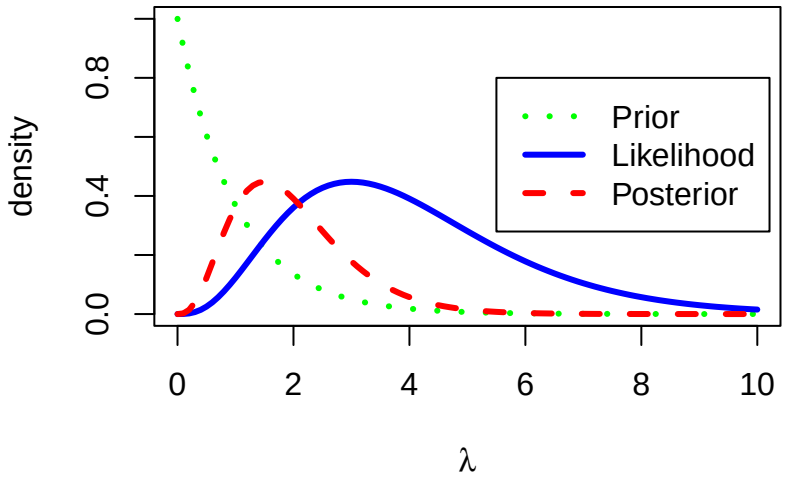
$$p(\lambda|y) = \frac{(1+1)^{3+1}\, \lambda^{3+1-1} \exp(-\lambda(1+1))}{\Gamma(3+1)} = \frac{(2)^4 \lambda^3 \exp(-\lambda(2))}{\Gamma(4)}.$$

Let's visualise the prior, likelihood and posterior in R:

```
y <- c(3)  ## the data
n <- length(y)  ## number of observations
alpha <- 1  ## prior parameter 1
beta <- 1   ## prior parameter 2
A <- alpha+sum(y)  ## posterior parameter 1
B <- beta+n        ## posterior parameter 2

## Plotting prior, likelihood and posterior, where x denotes the parameter being estimated
curve(dgamma(x,alpha,beta),0,10,col="green", ylab="density", lwd=3, lty=3,
```

```
      xlab=expression(lambda))#Prior
curve(( exp(-n*x)*x^(sum(y)) )/prod(y),0,10,add=TRUE,col="blue", lwd=3, lty=1)#Likelihood
curve(dgamma(x,A,B),0,10,add=TRUE,col="red", lwd=3, lty=2)#Posterior
legend(5.5,0.8, c("Prior", "Likelihood", "Posterior"), lty=c(3,1,2), lwd=c(3,3,3),
       col=c("green", "blue", "red"))
```



Now let's consider summarising the posterior distribution in the example here.

A 95% posterior credible interval for λ can be computed in R using:

```
qgamma(c(0.025, 0.975), A,B)
```

```
## [1] 0.5449327 4.3836365
```

```
## with A,B the posterior parameters found before, A <- alpha+sum(y), B <- beta+n
```

with posterior median and mean computed in R using:

```
## posterior median
qgamma(c(0.5), A,B)
```

```
## [1] 1.83603
```

```
## posterior mean
A/B
```

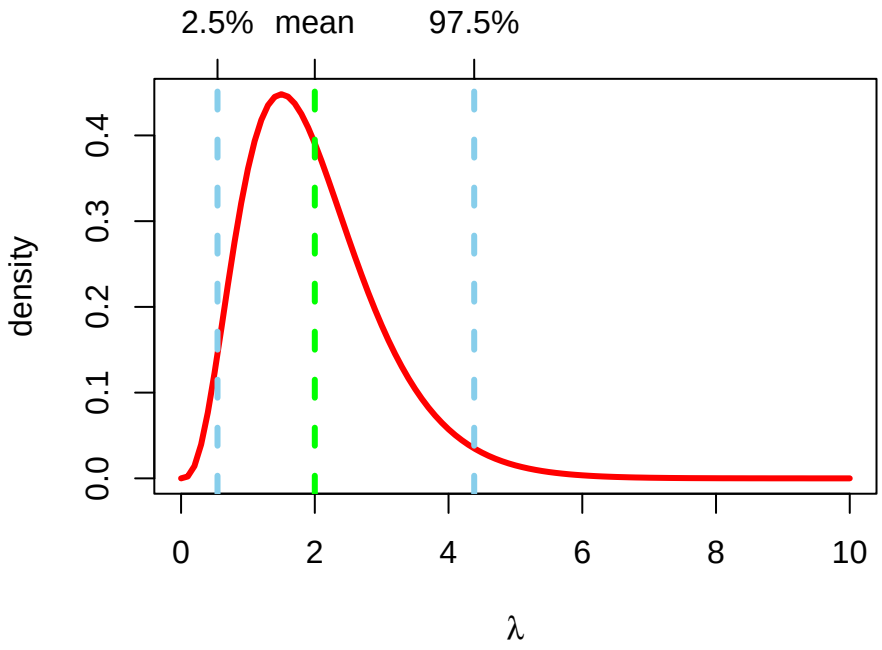
```
## [1] 2
```

Summary statistics and a 95% credible interval can also be obtained using the following package and command in R:

```
library(Bolstad)
poisgamp(3,1,1, plot=FALSE)  ##data, gamma prior parameter [1], gamma prior parameter[2]
```

```
## Summary statistics for data
## --------------------------
## Number of observations:   1
## Sum of observations:  3
##
## Summary statistics for posterior
## --------------------------------
## Shape parameter (r):  4
## Rate parameter (v):   2
## 95% credible interval for mu:    [0.54, 4.38]
```

We could also visualise these summaries on a plot of the posterior distribution in R:



Finally, let's derive the posterior distribution corresponding to a normal likelihood using a conjugate prior. We considered real-life examples of where such a distribution would be appropriate, e.g. investigating population mean IQ, population mean temperature, population mean house price, in earlier weeks, but let's just look at the theoretical derivation for this final example.

Example 3 (Estimating the mean θ of a normal distribution (assuming known variance ϕ)).

The mean of a normal distribution can take on any value between $-\infty$ and $+\infty$ and it turns out that a convenient conjugate prior distribution for the normal itself is the normal.

In the single observation case, if x is normally distributed with mean θ and known variance ϕ and the prior distribution of θ is normal with mean θ_0 and variance ϕ_0, where θ_0 and ϕ_0 are fixed known constants, let's derive the posterior distribution for $\theta|x$.

Then we have:

Model:

$$x|\theta \sim \mathrm{N}(\theta, \phi), \quad \phi \text{ known},$$

$$\theta \sim \mathrm{N}(\theta_0, \phi_0) \quad \theta_0, \phi_0 \text{ fixed constants},$$

$$p(\theta) = (2\pi\phi_0)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta-\theta_0)^2/\phi_0\right\},$$

$$L(\theta|x) = p(x|\theta) = (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\theta)^2/\phi\right\}.$$

The posterior distribution is given by:

$$\begin{aligned} p(\theta|x) &\propto p(\theta)L(\theta|x) \\ &= (2\pi\phi_0)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta-\theta_0)^2/\phi_0\right\} \times (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\theta)^2/\phi\right\} \\ &\propto \exp\left\{-\frac{1}{2}\theta^2(\phi_0^{-1}+\phi^{-1}) + \theta(\theta_0/\phi_0 + x/\phi)\right\}, \end{aligned}$$

since $-\frac{x^2}{2\phi}$ and $-\frac{\theta_0^2}{2\phi_0}$ are known constants.
We can now write:

$$\phi_1 = \frac{1}{\phi_0^{-1}+\phi^{-1}},$$

and let's take,

$$\theta_1 = \phi_1(\theta_0/\phi_0 + x/\phi),$$

so that,

$$\begin{aligned} \phi_0^{-1} + \phi^{-1} &= \phi_1^{-1}, \\ \theta_0/\phi_0 + x/\phi &= \theta_1/\phi_1, \end{aligned}$$

and hence,

$$p(\theta|x) \propto \exp\left\{-\frac{1}{2}\theta^2/\phi_1 + \theta\theta_1/\phi_1\right\}.$$

Adding into this:

$$-\frac{1}{2}\theta_1^2/\phi_1,$$

(which is a constant as far as θ is concerned) to aid with factorisation, we see that:

$$p(\theta|x) \propto \exp\left\{-\frac{1}{2}(\theta-\theta_1)^2/\phi_1\right\}.$$

That is, the posterior density is $\theta|x \sim \mathrm{N}(\theta_1, \phi_1)$.

Note:

We define the ***reciprocal of the variance*** of a particular random variable to be its ***precision***.

Task 2 (Several normal observations with a normal prior).

We can generalise the previous example by supposing that a priori,

$$\theta \sim \mathrm{N}(\theta_0, \phi_0),$$

where θ_0 and ϕ_0 are known constants, but that instead of having just one observation we have n independent observations $\mathbf{x} = (x_1, \ldots, x_n)$ such that

$$X_i \sim \mathrm{N}(\theta, \phi), \quad i = 1, \ldots, n$$

with ϕ known.

Derive the posterior distribution for $\theta|\mathbf{x}$.

Additional topics

Odds ratio to compare hypotheses

If we simply wish to consider the odds in favour of one simple hypothesis compared to another then a form of Bayes' theorem, the odds ratio form, can be found that does not require calculation of the ***normalising constant***, $p(x)$. Suppose we have two simple hypotheses, with θ_0 and θ_1 completely specified:

H_0: $\theta = \theta_0$

H_1: $\theta = \theta_1$

Then,

$$\frac{p(\theta_0|x)}{p(\theta_1|x)} = \frac{p(\theta_0)p(x|\theta_0)/p(x)}{p(\theta_1)p(x|\theta_1)/p(x)} = \frac{p(\theta_0)}{p(\theta_1)} \times \frac{p(x|\theta_0)}{p(x|\theta_1)},$$

i.e.

$$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio}.$$

This particular form of Bayes' theorem is useful for comparing two simple hypotheses. In the context of estimation these are hypotheses involving a single unknown value of a parameter for which the likelihood is completely specified. For more complex cases a modification of the *posterior odds* above is necessary, which will require the use of so-called Bayes' factors.

Predictive distribution

In classical (frequentist) statistics it is usual to fit a model to the past data, and then make predictions of future values on the assumption that this model is correct.

In making predictions about the future values on the basis of an estimated model there are two sources of uncertainty:

- uncertainty in the parameter values which have been estimated on the basis of past data, and
- uncertainty due to the fact that any future value is itself a random event.

Bayesian inference allows for both sources of uncertainty by simply averaging over the uncertainty in the parameter estimates.

After the data x have been observed, we can predict an unknown observable, $\tilde{x}$, from the same process. The distribution of $\tilde{x}$ is called the posterior predictive distribution.

Prediction of new data $\tilde{x}$:

$$p(\tilde{x}|x) = \int p(\tilde{x}|\theta)p(\theta|x)d\theta.$$

Sequential use of Bayes' theorem

Bayes' theorem provides the way by which our prior information is updated by data to give our posterior information. This then serves as our new prior information before more data become available. This gives rise to one question in particular: if we obtain a sequence of data, and we update our beliefs on the arrival of each data item, would we get a different result from waiting until all of the data had arrived, and then updating our prior?

The method to determine the posterior distribution can be sequentially applied as follows to update the posterior for new data. Suppose you have an initial sample of observations $\mathbf{x}$, and parameters $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{x}).$$

Then suppose you have a second set of observations $\mathbf{y}$ which are independent from the first sample. Then,

$$p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}),$$

and independence implies:

$$p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{x})L(\boldsymbol{\theta}, \mathbf{y}),$$

$$p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \propto p(\boldsymbol{\theta}|\mathbf{x})L(\boldsymbol{\theta}, \mathbf{y}).$$

The posterior for $\boldsymbol{\theta}$ given $\mathbf{x}$ and $\mathbf{y}$ can be found by treating your posterior given $\mathbf{x}$ as the prior for the observation $\mathbf{y}$. This is the same result that we would have obtained by updating on the basis of the entire information $(\mathbf{x}, \mathbf{y})$ directly.

Bayesian statistics in practice

Bayesian statistics proceeds smoothly and easily as long as we stick to well-known distributions and use conjugate priors. In principle, the posterior distribution is always available, although in realistically complex problems it cannot be represented analytically. Instead we often arrive at a posterior:

$$p(\theta|x) \propto p(\theta)L(\theta|x),$$

but have no easy way of finding the normalising constant $p(x)$, since the posterior distribution is not of a known distributional form. This presented a barrier to the implementation of the Bayesian approach until the development of numerical methods and powerful computers during the late 20th century. Now, posterior distributions can be constructed for highly complex problems using Markov chain Monte Carlo (MCMC) simulation. MCMC involves simulating a sample from the (joint) posterior distribution of the unknown parameters using one of three main algorithms: the Metropolis-Hastings algorithm, Gibbs sampling or slice sampling. With a sufficiently large sample, we are able to numerically generate the whole distribution from which we can make any inferences of interest. However, this can be computationally intensive and so there are also modern algorithms that approximate these results, which can be used to make the inference more efficient.

There are several different R packages that you can use to analyse complex problems: R2WinBUGS, R2OpenBUGS, BRugs, RStan, R-INLA, rjags, (to name a few) that mean you do not need to write your own sampling algorithms. However, the implementation of these packages and ideas are beyond the scope of this course.

Supplementary material:

Here are a couple of links to additional material on motivating Bayesian inference:

Understanding Bayes' - a look at the likelihood

Updating priors using the likelihood

Supplementary material:

Simulation demonstrating Bayesian updating on binomial after x heads out of 10 coin flips (resulting in a beta posterior, using a flat Be(1,1) prior). The code here is by David Robinson of DataCamp and was posted on Twitter.

```
library(tidyverse)
theme_set(theme_bw())
data_frame(probability = runif(5e6))%>%
  mutate(heads=rbinom(n(), 10, probability)) %>%
  ggplot(aes(probability, group=heads, color=heads)) +
  geom_density()+
  labs(x="Coin's probability of heads", y="Posterior",
       title="Posterior probability after x heads out of 10 flips")
```

Supplemetary exercises

Task 3.

For the binomial distribution with n trials and x successes, derive the full posterior distribution of θ, the proportion of success, using the conjugate prior for θ, $\theta \sim \text{Be}(1, 1)$ (make sure to state the normalising constant).

Task 4.

The data in the table below are annual fatalities on the roads in Scotland between 1999 and 2006.

Annual road fatalities

Year	1999	2000	2001	2002	2003	2004	2005	2006
No. of Fatalities	285	297	309	274	301	283	264	293

Assuming that the counts $x_1, \ldots, x_8$ follow a Poisson distribution with mean λ, use the following Gamma prior distribution for λ, $p(\lambda)$, to derive the posterior distribution for, λ, the average number of annual fatalities (make sure to state the normalising constant).

$$p(\lambda) \propto \lambda^{\alpha-1} \exp(-\lambda/\beta^2)$$

Task 5.

Consider the Pareto distribution:

$$p(y|\theta) = \frac{\theta k^\theta}{y^{\theta+1}}, \quad y > k$$

where $\theta > 0$ is an unknown parameter and $k > 0$ is a known constant.

- Write down the likelihood function, $L(\theta|\mathbf{y})$, for a set of data $\mathbf{y} = y_1, \ldots, y_n$ from this distribution.
- Prove that the prior for θ,

  $$\theta \sim \text{Ga}(\alpha, \beta)$$

  is a conjugate prior for the Pareto model above by showing that:

$$\theta|\mathbf{y} \sim \text{Ga}\left(\alpha + n, \beta + \sum_{i=1}^{n} \log\left(\frac{y_i}{k}\right)\right)$$

Answers to tasks

Answer to Task 1. Estimating the beta parameters, by taking the median to be 0.5 and the 0.9 quantile to be 0.6 (note: you could make different choices for these values) in R gives:

```
library(LearnBayes)
quantile2=list(p=.9, x=.6)
quantile1=list(p=.5, x=.5)
beta.select(quantile1, quantile2)
```

```
## [1] 20.37 20.37
```

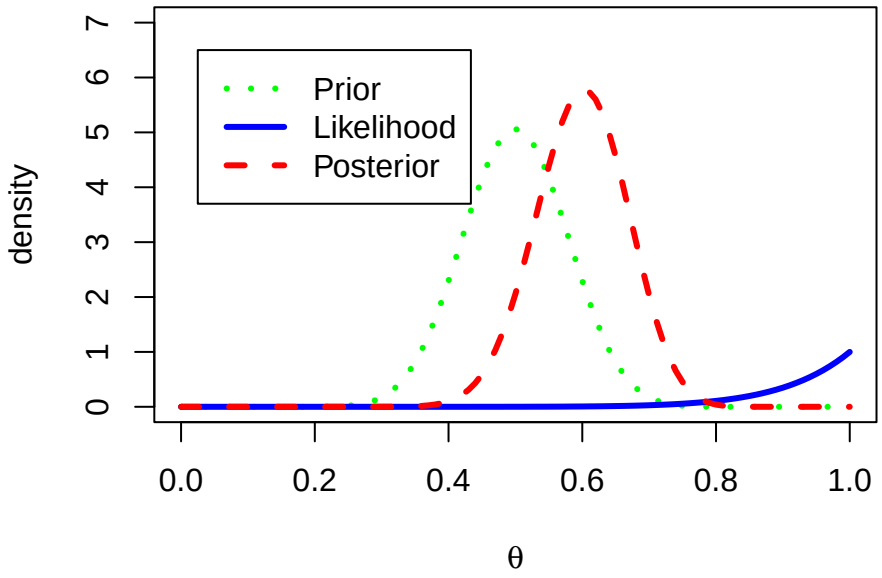
We see from this that an appropriate prior distribution for θ might be $\theta \sim \text{Be}(20.37, 20.37)$.

Combining this with the information from our data of 10 successes and 0 failures, we have:

$$p(\theta|x) = \frac{\theta^{20.37+10-1}(1-\theta)^{20.37+10-10-1}}{B(20.37+10, 20.37+10-10)}$$

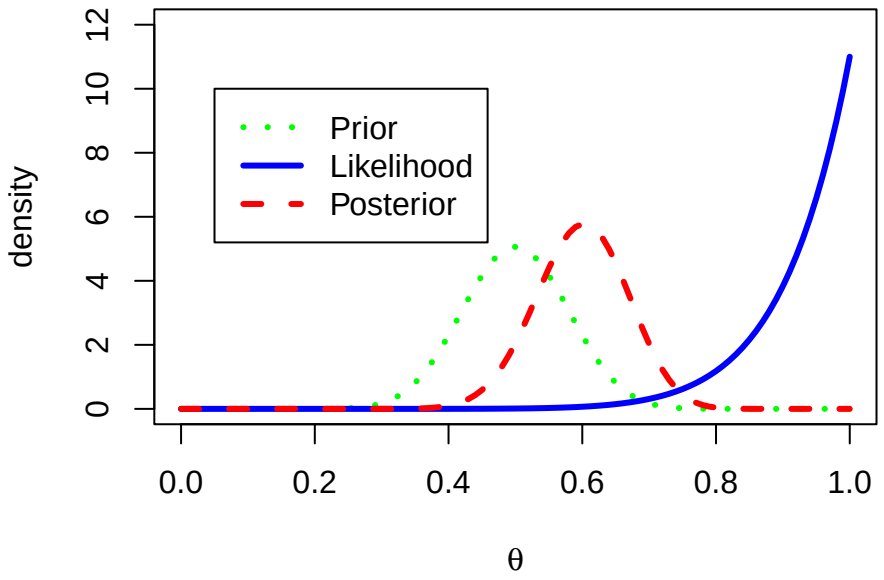
Plotting these distributions in R we get:

```
y <- 10
n <- 10
alpha <- 20.37
beta <- 20.37
A <- alpha+y
B <- beta+n-y
curve(dbeta(x,alpha,beta),0,1,ylim=c(0,7),col="green", lwd=3, ylab="density", lty=3,
      xlab=expression(theta))#Prior
curve(dbinom(y,n,x),0,1,add=TRUE,col="blue", lwd=3, lty=1)#Likelihood
curve(dbeta(x,A,B),0,1,add=TRUE,col="red", lwd=3, lty=2)#Posterior
legend(0.025,6.5, c("Prior", "Likelihood", "Posterior"), lty=c(3,1,2), lwd=c(3,3,3),
       col=c("green", "blue", "red"))
```



```
## scaling the likelihood

curve(dbeta(x,alpha,beta),0,1,ylim=c(0,12),col="green", lwd=3, ylab="density", lty=3,
      xlab=expression(theta))#Prior
curve(dbeta(x,y+1,n-y+1),0,1,add=TRUE,col="blue", lwd=3, lty=1)#Likelihood
curve(dbeta(x,A,B),0,1,add=TRUE,col="red", lwd=3, lty=2)#Posterior
legend(0.05,10, c("Prior", "Likelihood", "Posterior"), lty=c(3,1,2), lwd=c(3,3,3),
       col=c("green", "blue", "red"))
```



Answer to Task 2 (Several normal observations with a normal prior). Then,

$$\begin{aligned} p(\theta|x) \propto p(\theta)L(\theta|\mathbf{x}) &= (2\pi\phi_0)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta-\theta_0)^2/\phi_0\right\} \\ &\quad \times (2\pi\phi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2/\phi\right\} \\ &\propto \exp\left\{-\frac{1}{2}\theta^2(1/\phi_0+n/\phi)+\theta(\theta_0/\phi_0+\sum_{i=1}^{n}x_i/\phi)\right\} \end{aligned}$$

Therefore,

$$\theta|\mathbf{x} \sim \mathrm{N}(\theta_1, \phi_1)$$

where

$$\phi_1 = (1/\phi_0 + n/\phi)^{-1}$$

and taking

$$\theta_1 = \phi_1(\theta_0/\phi_0 + \sum_{i=1}^{n} x_i/\phi)$$

This follows from the equivalent results in example 3.

Answer to Task 3. For n trials and x successes, the binomial model states that:

$$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Considered as a function of θ, the likelihood is of the form:

$$L(\theta|x) \propto \theta^x(1-\theta)^{n-x}$$

Suppose we assume that a priori $\theta \sim \mathrm{Be}(1,1)$, where Be is a Beta distribution with parameters α and β:

$$p(\theta) \propto \theta^{1-1}(1-\theta)^{1-1}$$
$$p(\theta) \propto 1$$

Then,

$$\begin{aligned} p(\theta|x) &\propto p(\theta)L(\theta|x) \\ &= 1 \times \theta^x(1-\theta)^{n-x} \\ &= \theta^x(1-\theta)^{n-x} \\ \theta|x &\sim \mathrm{Be}(1+x, 1+n-x) \\ \theta|x &\sim \mathrm{Bi}(n,\theta) \end{aligned}$$

$$p(\theta|x) = \frac{p(\theta)L(\theta|x)}{p(x)}$$

i.e.

$$p(\theta|x) = \frac{\theta^x(1-\theta)^{n-x}}{B(1+x, 1+n-x)}$$

Answer to Task 4. $x_1, \ldots x_8 \sim \mathrm{Poi}(\lambda)$. Then the likelihood is:

$$L(\lambda|\mathbf{x}) = \prod_{i=1}^{8} p(x_i|\lambda) = \prod_{i=1}^{8} \lambda^{x_i}\exp(-\lambda)/x_i!$$
$$L(\lambda|\mathbf{x}) \propto \lambda^{\sum_{i=1}^{8} x_i}\exp(-8\lambda)$$

$$p(\lambda) \propto \lambda^{\alpha-1}\exp(-\lambda/\beta^2)$$

The posterior is therefore,

$$p(\lambda|\mathbf{x}) \propto p(\lambda)L(\lambda|\mathbf{x})$$

$$\begin{aligned} p(\lambda|\mathbf{x}) &\propto \lambda^{\alpha-1}\exp(-\lambda/\beta^2) \times \lambda^{\sum_{i=1}^{8} x_i}\exp(-8\lambda) \\ &= \lambda^{\sum_{i=1}^{8} x_i+\alpha-1}\exp\left(-\lambda\left(\frac{8\beta^2+1}{\beta^2}\right)\right) \end{aligned}$$

Thus the posterior distribution will just be of the same general form as the prior distribution, that is to say a gamma distribution,

$$\lambda|\mathbf{x} \sim \text{Ga}(2306 + \alpha, (8\beta^2 + 1)/\beta^2)$$

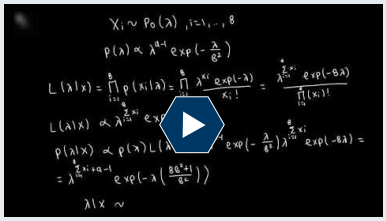
Therefore,

$$p(\lambda|\mathbf{x}) = \frac{p(\lambda)L(\lambda|\mathbf{x})}{p(\mathbf{x})}$$

i.e.

$$p(\lambda|\mathbf{x}) = \frac{\left(\frac{8\beta^2+1}{\beta^2}\right)^{2306+\alpha} \lambda^{2306+\alpha-1} \exp(-\lambda(\frac{8\beta^2+1}{\beta^2}))}{\Gamma(2306+\alpha)}$$

$$p(\lambda|\mathbf{x}) = \frac{\left(\frac{8\beta^2+1}{\beta^2}\right)^{2306+\alpha} \lambda^{2305+\alpha} \exp(-\lambda(\frac{8\beta^2+1}{\beta^2}))}{\Gamma(2306+\alpha)}$$



Video model answers

https://youtu.be/ZmeU_e3R0q4

Duration: 6m09s

Answer to Task 5. The Likelihood is:

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{n} \frac{\theta k^{\theta}}{y_i^{\theta+1}} = \theta^n k^{n\theta} \left\{\prod_{i=1}^{n} y_i\right\}^{-(\theta+1)}$$

$$\theta \sim \text{Ga}(\alpha, \beta)$$

$$\begin{aligned}
p(\theta|\mathbf{y}) &\propto p(\theta)L(\theta|\mathbf{y}) \\
&= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} . \theta^n k^{n\theta} \left\{\prod_{i=1}^{n} y_i\right\}^{-(\theta+1)} \\
&\propto \theta^{\alpha+n-1} e^{-\beta\theta} e\{\log(k^{n\theta}\{\prod_{i=1}^{n} y_i\}^{-\theta})\} \\
&\propto \theta^{\alpha+n-1} \exp\left\{-\beta\theta + n\theta \log k - \theta \sum_{i=1}^{n} \log(y_i)\right\} \\
&= \theta^{\alpha+n-1} \exp\left\{-\theta\left[\beta - n \log k + \sum_{i=1}^{n} \log(y_i)\right]\right\} \\
&= \theta^{\alpha+n-1} \exp\left\{-\theta\left[\beta + n \log k^{-1} + \sum_{i=1}^{n} \log(y_i)\right]\right\}
\end{aligned}$$

$$\theta|\mathbf{y} \sim \text{Ga}\left(\alpha + n, \beta + \sum_{i=1}^{n} \log \frac{y_i}{k}\right)$$