Decision and Risk

Course Overview

Purpose of This Course

The ultimate purpose of this course is to gain an understanding of how to use statistical techniques to quantity and reason about risks and uncertainty. We will discuss questions such as:

- What was the probability of a terrorist attack that killed as many people as the 9/11 New York event occurring, given what we know about other historical terrorist attacks? How long do we expect before a similarly sized attack occurs again?
- How can financial institutions make sensible decisions about the risks associated with the portfolios of stocks that they are invested in, so that they can try to control how much money they stand to lose on a day-to-day-basis?
- How can we predict how likely an earthquake is to occur in a certain geographical region, and how do we balance the costs of giving false alarms against the costs of not evacuating people in time?

Three Main Learning Outcomes

We will gain familiarity with three main areas:

- **Bayesian Inference** - how do we learn about unknown quantities?
- **Decision Theory** - how do we translate what we learn about unknown quantities into actual decision making?
- **Applied Data Analysis** - we will look at real data as we proceed. This can be done in a statistical programming language such as R.

Note – because we need to learn some techniques before we can perform any analysis, the mathematical content in this course is somewhat front-loaded. We will spend this week and next learning the basics of Bayesian statistics and decision theory, and then later see how they are applied to real problems.

Structure of This Course

- Nine 2 hour lectures, taking place each Friday at 9-11am. If time permits, I will devote the last part of selected lectures (10 minutes) to solve one exercise from problem sheets.
- Three workshops during weeks 21, 24, and 28 (i.e. the first is next week!)
- ICA in week 26, 22nd February 2019, followed by a final exam at the end of an academic year.
- All course material will be available on moodle.

Lecture 1: Bayesian Inference

Alex Donov

11 January 2019

Why Bayesian Statistics

Three Advantages

1. It allows prior information (e.g. expert judgement, or previous data) to be incorporated into the analysis, which is helpful in situations where there is not much data. For example, in earthquake modelling, it could be that only 4 or 5 earthquakes have ever occurred.
2. Bayesian probability statements are easy to interpret, which is important when communicating with non-statisticians. For example, interpret the frequentist 95% confidence interval for a parameter θ.
3. It makes analysis a lot easier - all inference is directly based on the posterior distribution of the parameters. This makes it easy to combine information from a variety of data sources.

The General Principles of Bayesian Methodology

Bayesian

- θ is a random variable (has an unknown distribution)
- inference uses both data and prior information

Frequentist

- θ is fixed (unknown)
- inference uses only data

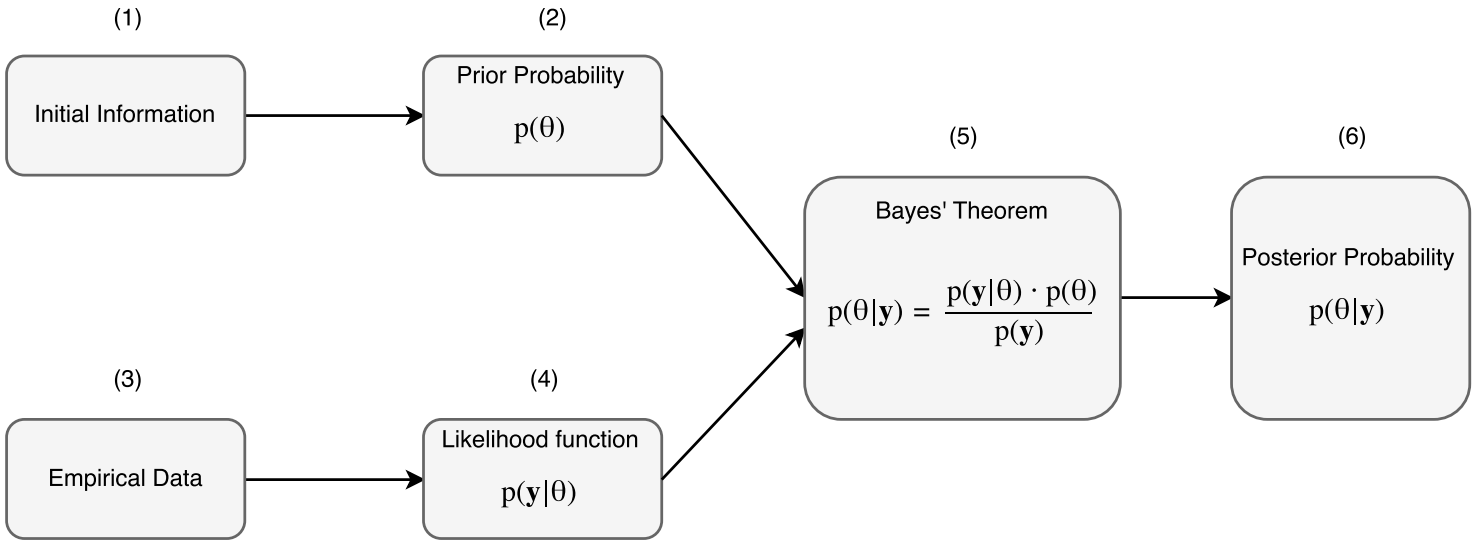
The General Principles of Bayesian Methodology

The modern understanding of Bayesian methodology contains the following statements:

Statement 1. The parameter of a stochastic system under study is random, and it is assigned prior distribution. The "randomness" is understood not only in a classical sense but also as "uncertainty".

Statement 2. The observed empirical data and the prior distribution are unified by the Bayes' theorem in order to obtain a posterior distribution of a parameter.
Bayes' theorem provides the foundation to transfer from prior to posterior information by incorporating empirical data.

Statement 3. A statistical conclusion or decision rule is accepted with a condition of maximal estimated utility, in particular, the minimization of loss related to this rule.



The process of revising probabilities

Bayes' theorem

- *Bayes' theorem*, also known as *Bayes' rule* or *Bayes' law*, is the cornerstone of the Bayesian framework.
- Bayes' theorem is a simple mathematical result that follows from the axioms of conditional probability. Nevertheless, it has profound implications.
- The importance of Bayes' theorem comes from its use for updating probabilities in light of new information (e.g. observed data).

Bayes' theorem

- Bayes' theorem applies to both discrete and to continuously distributed random variables, and is usually stated in terms of probabilities for observable events.
- Let A and B be events, then:

$$p(A|B) = \frac{p(A) \cdot p(B|A)}{p(B)} \tag{1}$$

Bayesian Inference for Parameters

- Previous result can be used to make probability statements about θ given the known value of the data $\mathbf{y} = (y_1, \ldots, y_n)$.
- In order to to do that, specify a model providing a joint probability distribution for θ and $\mathbf{y}$:
- The joint probability mass or density function can be written as:

$$p(\theta, \mathbf{y}) = \underbrace{p(\theta)}_{prior} \cdot p(\mathbf{y}|\theta). \tag{2}$$

Conditioning on the known value of the data $\mathbf{y}$, and using Bayes' theorem, yields the *posterior density*:

$$p(\theta|\mathbf{y}) = \frac{p(\theta, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\theta) \cdot p(\mathbf{y}|\theta)}{p(\mathbf{y})}. \tag{3}$$

where:

$p(\theta|\mathbf{y})$ - is the *posterior density* for θ and represents the uncertainty about θ after conditioning on the data $\mathbf{y}$;

$p(\theta)$ - is the *prior density* for θ that expresses our uncertainty about the values of θ before taking into account sample information (i.e. observed data).

$p(\mathbf{y}|\theta)$ - when regarded as a function of θ, for fixed $\mathbf{y}$, is the well-known *likelihood function*;

$p(\mathbf{y})$ - is the marginal density of the data $\mathbf{y}$.

- $p(\mathbf{y})$ is normally written as:

$$p(\mathbf{y}) = \underbrace{\sum_\theta p(\theta)p(\mathbf{y}|\theta)}_{\text{discrete } \theta} \qquad p(\mathbf{y}) = \underbrace{\int p(\theta)p(\mathbf{y}|\theta)d\theta}_{\text{continuous } \theta}$$

- It acts as a normalising constant to ensure that the value of $p(\theta|\mathbf{y})$ is a valid probability, i.e. a number between 0 and 1.

A note on proportionality symbol $\propto$

The symbol of proportionality, $\propto$, denotes cases where terms constant with respect to the random variable have been dropped from the *pdf* of that random variable. For example, suppose that the random variable, X, has a *pdf* as follows:

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{4}$$

Then, using the symbol of proportionality, we can write:

$$p(x) \propto e^{-\frac{x^2}{2}} \tag{5}$$

- It is often customary to retain in the expressions for the posterior *pdf* only the terms that contain the unknown parameter(s), and dropping the terms that are constant with respect to the parameter(s).
- An equivalent form of posterior distribution can be written omitting the factor $p(\mathbf{y})$, which does not depend on θ, and with $\mathbf{y}$ being fixed, it can thus be considered a constant, which yields the *unnormalized posterior density*, which is the right-hand side of (6):

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto p(\theta) \cdot p(\mathbf{y}|\theta) \\ &\propto \text{prior } \textit{pdf} \cdot \text{likelihood function} \end{aligned} \tag{6}$$

Example 1: Coin Tossing

Suppose we toss a coin three times. There are eight equally likely outcomes:

$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$.

Define the events:

- A: "There are two heads in three tosses"
- B: "The first toss was heads"

In this case, we can compute the quantities $p(A)$, $p(B)$, $p(A|B)$, $p(B|A)$ directly by enumerating the outcomes, and hence verify Bayes' theorem:

$$p(A) : \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\} = 3/8$$

$$p(B) : \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\} = 1/2$$

$$p(A|B) = \{HHH, HHT, HTH, HTT\} = 2/4 = 1/2$$

$$p(B|A) = \{HHT, HTH, THH\} = 2/3$$

So :

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{2/3 * 3/8}{1/2} = 1/2$$

Sometimes $p(B)$ will not be given explicitly, and must be derived.

Example 2: Medical Diagnosis

- A new medical screening test is developed to assess whether a patient has a particular disease.
- Advertised degrees of accuracy:

  "if the patient truly has the disease, then the test will correctly detect this and return a positive result with probability 0.95. If the patient truly does not have the disease, the test will correctly detect this and return a negative result with probability 0.98".
- Given that 1 in 1000 people in the population have the disease, what is the probability that a person testing positive on the test really has the disease?

We first define the events:

- A: The person truly has the disease
- A': The person truly does not have the disease
- B: The test comes back positive

We need to compute $p(A|B)$.

Representing the given information mathematically, we have:

- $P(A) = 1/1000 = 0.001$
- $P(B|A) = 0.95$
- $P(B|A') = 0.02$

Therefore:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|A')p(A')}$$
$$= \frac{0.95 * 0.001}{0.95 * 0.001 + 0.02 * 0.999} = 0.045$$

- Hence, the person has a 4.5% probability of having the disease if the test comes back positive.
- This is lower than we might expect given that the test had accuracies of 95% and 98%!
- The reason for this is that the prior probability $p(A)$ of the person having the disease is very low.

Example 3

Define:

- G: The suspect is guilty
- L: The suspect is left-handed

We need to compute $p(G|L)$.

We have:

$$p(G) = 0.6, \quad p(L|G) = 1, \quad p(L|G') = 0.1$$

Hence, using Bayes' Theorem:

$$p(G|L) = \frac{p(L|G)p(G)}{p(L|G)p(G) + p(L|G')p(G')} = \frac{1 * 0.6}{1 * 0.6 + 0.1 * 0.4} = 0.938$$

Example 4

- Suppose we are given a coin and told that it could be biased, so the probability of landing heads is not necessarily 0.5.
- Let θ denote the probability of it landing heads. We wish to learn about θ.
- We toss the coin N times and obtain Y heads. In frequentist statistics, the point estimate of θ would be Y/N.

- Suppose we performed 100 tosses and got 48 heads. The point estimate would be $\theta = 0.48$.
- However, in this situation it may be more reasonable to conclude that the coin isn't biased.
- In other words, rather than concluding that $\theta = 0.48$, we may wish to include prior information to make a more informed judgement.

- In a Bayesian analysis, we first need to represent our prior beliefs about θ, by constructing a probability distribution $p(\theta)$ which encapsulates our beliefs.
- $p(\theta)$ will not be the same for different people as they may have different knowledge about what proportion of coins are biased.
- In some cases, $p(\theta)$ may be based on subjective judgement, while in others it may be based on objective evidence. This is the essence of Bayesian statistics - probabilities express degrees of beliefs.

- However, θ must lie between 0 and 1 since it represents the probability of the coin landing heads.
- So the function we use to represent our beliefs should only have mass in the interval $[0, 1]$, which rules out, for example, the Normal distribution.

- In this situation it is usual to represent our prior beliefs as a Beta distribution, because Beta distribution has mass only in $[0, 1]$ so it is a sensible choice when θ is a probability.
- Recall the Beta distribution has the form:

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} \tag{7}$$

  where $B(\alpha, \beta)$ is the Beta function;
- α and β are commonly called **hyperparameters**, because they are parameters that control parameters.
- We choose these to reflect our prior beliefs about θ. How do we do this?

- It can be shown that the mean and variance of the Beta distribution is given by:

$$E(\theta) = \mu = \frac{\alpha}{\alpha + \beta} \tag{8}$$

$$Var(\theta) = \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{9}$$

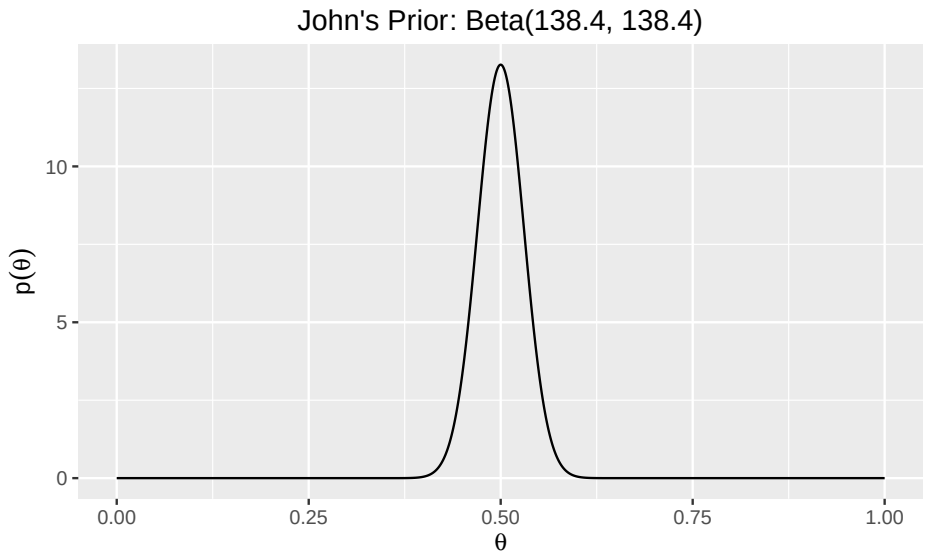
- Therefore, if we have a prior belief about the most likely value of θ (e.g. 0.5) then we choose $p(\theta)$ to have this as the expected value.

- Then, we express how uncertain we are about this value by the variance.
- Rearranging the above equations we can express α and β in terms of the mean and variance:

$$\alpha = \left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu} \right) \mu^2 \tag{10}$$

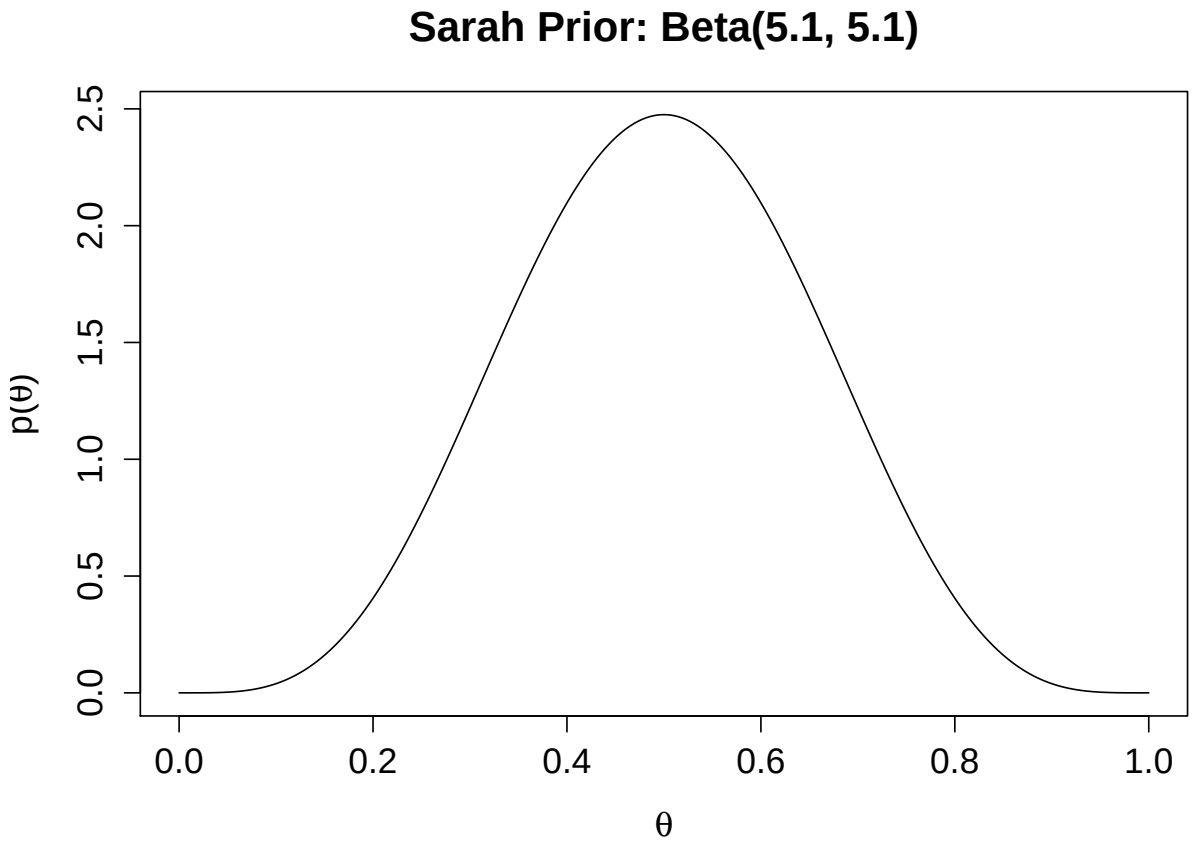
$$\beta = \alpha \left(\frac{1}{\mu} - 1 \right) \tag{11}$$

- For example, let's say that John has prior belief that $E(\theta) = 0.5$.
- He doesn't expect the coin to be biased, so he takes the standard deviation to be low, say 0.03.
- Based on the previous equations, his prior is $Beta(138.4, 138.4)$.

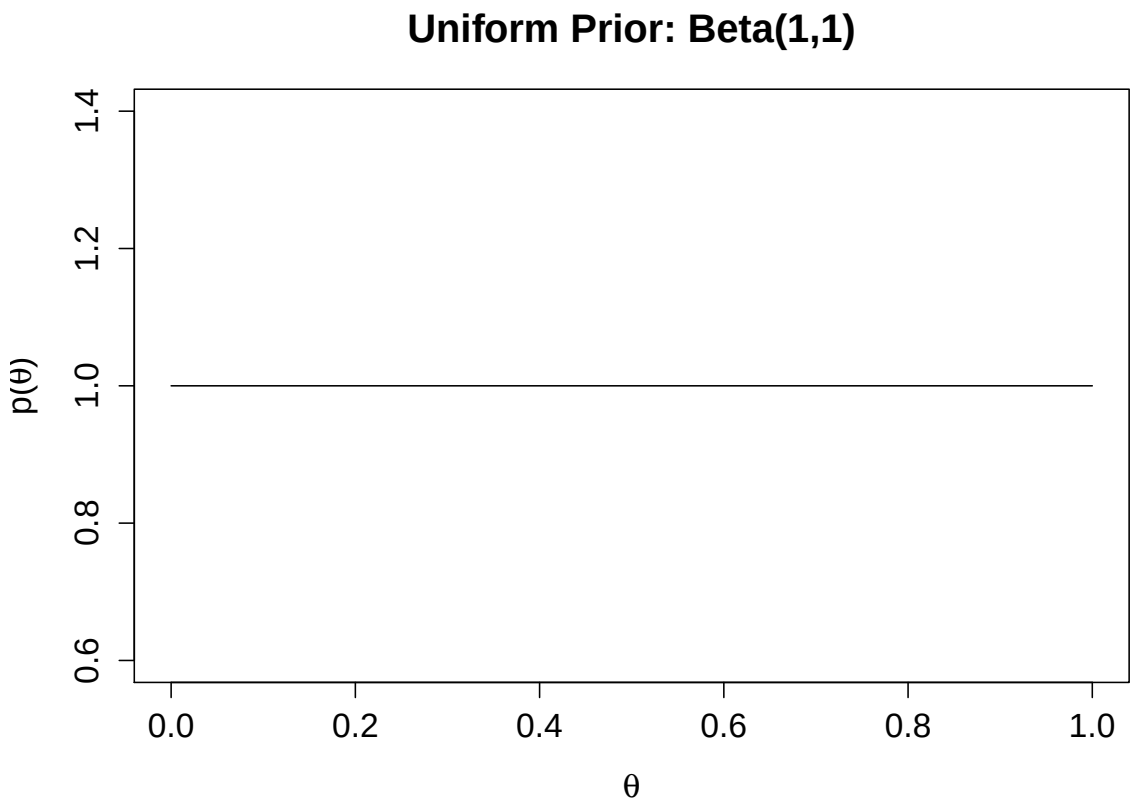


John's Prior

- Sarah, on the other hand, is more sceptical. She also assumes that $E(\theta) = 0.5$, but thinks the coin might be biased.
- She takes the standard deviation to be higher, say 0.15. Her prior distribution is hence $Beta(5.1, 5.1)$.



- Note there is also a special case of the Beta distribution when $\alpha = 1$ and $\beta = 1$ where it is flat, and equal to the Uniform distribution.
- This represents complete uncertainty, where any value of θ is assumed to be equally likely.



- Each individual toss follows a $Bernoulli(\theta)$ distribution, and each of the N tosses has probability θ to be heads.
- Therefore, the likelihood $p(Y|\theta)$ for the number of heads Y is a $Binomial(N, \theta)$ distribution:

$$p(Y|\theta) = \binom{N}{Y}\theta^{Y}(1-\theta)^{N-Y} \tag{12}$$

To learn about θ from the data, we need the posterior $p(\theta|Y)$, which by Bayes' theorem is:

$$p(\theta|Y) = \frac{p(\theta) \cdot p(Y|\theta)}{\int p(\theta)p(Y|\theta)d\theta} \tag{13}$$

The numerator here is:

$$p(\theta) \cdot p(Y|\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \cdot \binom{N}{Y}\theta^{Y}(1-\theta)^{N-Y} \tag{14}$$

$$= \binom{N}{Y}\frac{\theta^{\alpha+Y-1}(1-\theta)^{\beta+N-Y-1}}{B(\alpha,\beta)} \tag{15}$$

The denominator is:

$$\int p(\theta)p(Y|\theta)d\theta = \int \binom{N}{Y} \frac{\theta^{\alpha+Y-1}(1-\theta)^{\beta+N-Y-1}}{B(\alpha,\beta)} d\theta \qquad (16)$$

This looks horrible, but there is a standard trick we can use here.

First, take everything that doesn't depend on θ outside the integral:

$$\int p(\theta)p(Y|\theta)d\theta = \frac{\binom{N}{Y}}{B(\alpha,\beta)} \int \theta^{\alpha+Y-1}(1-\theta)^{\beta+N-Y-1} d\theta \qquad (17)$$

Now, recognise that the θ-dependent part inside the integral has the same form as the Beta distribution, which recall was:

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \tag{18}$$

So we make the substitution $\gamma = \alpha + Y$ and $\lambda = \beta + N - Y$, giving:

$$\int p(\theta)p(Y|\theta)d\theta = \frac{\binom{N}{Y}}{B(\alpha,\beta)} \int \theta^{\gamma-1}(1-\theta)^{\lambda-1}d\theta \tag{19}$$

Now, using the fact that this resembles the Beta distribution, and that the Beta distribution, like all probability distributions, must integrate to 1, we have:

$$\int \theta^{\gamma-1}(1-\theta)^{\lambda-1}d\theta = B(\gamma,\lambda) \tag{20}$$

So:

$$\int p(\theta)p(Y|\theta)d\theta = \binom{N}{Y}\frac{B(\gamma,\lambda)}{B(\alpha,\beta)}$$

$$= \binom{N}{Y}\frac{B(\alpha+Y,\beta+N-Y)}{B(\alpha,\beta)} \tag{21}$$

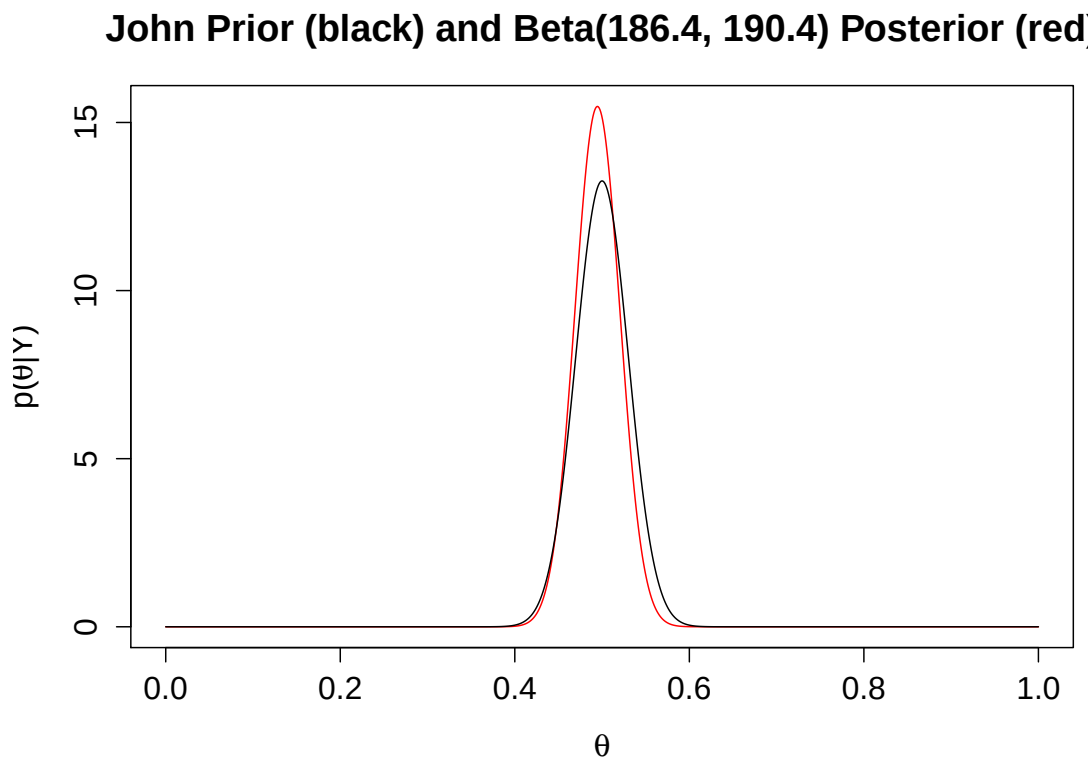
Combining this with the numerator gives:

$$p(\theta|Y) = \frac{\binom{N}{Y}\frac{\theta^{\alpha+Y-1}(1-\theta)^{\beta+N-Y-1}}{B(\alpha,\beta)}}{\binom{N}{Y}\frac{B(\alpha+Y,\beta+N-Y)}{B(\alpha,\beta)}} = \frac{\theta^{\alpha+Y-1}(1-\theta)^{\beta+N-Y-1}}{B(\alpha+Y,\beta+N-Y)} \tag{22}$$

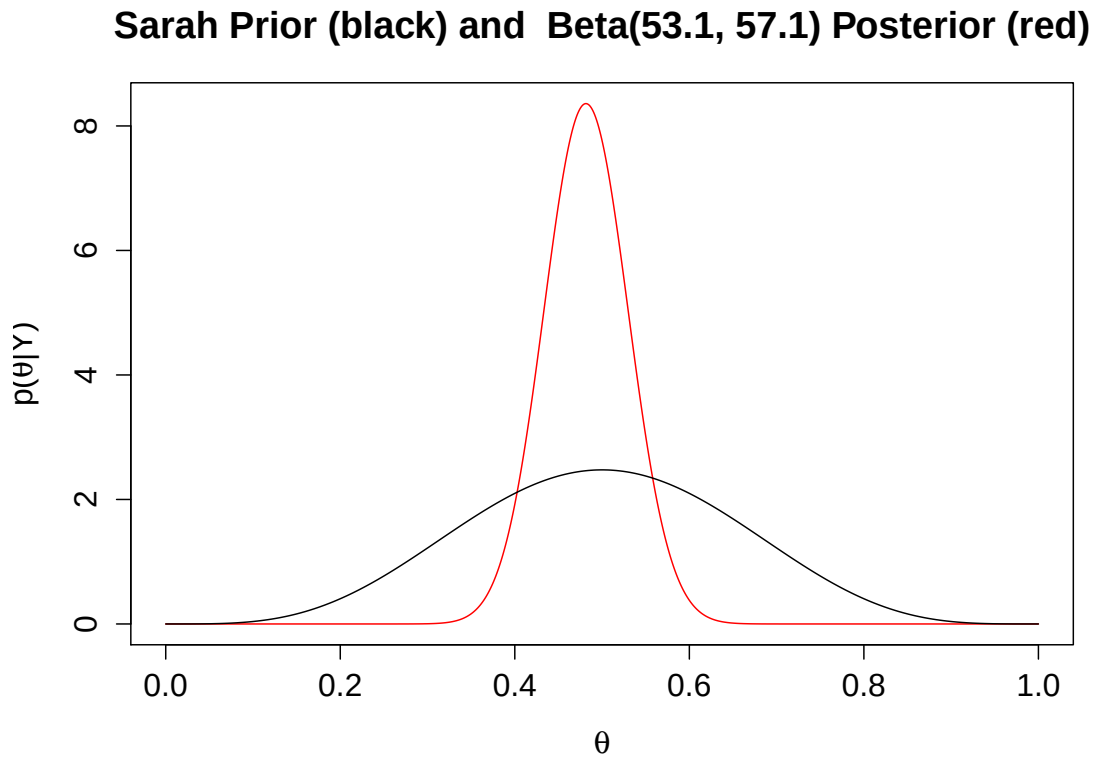
So $p(\theta|Y)$ is a $Beta(\alpha+Y,\beta+N-Y)$ distribution.

- In other words, our prior beliefs were that θ had a $Beta(\alpha,\beta)$ distribution.
- After seeing the data, we revised our beliefs about θ to be a $Beta(\alpha+Y,\beta+N-Y)$ distribution.

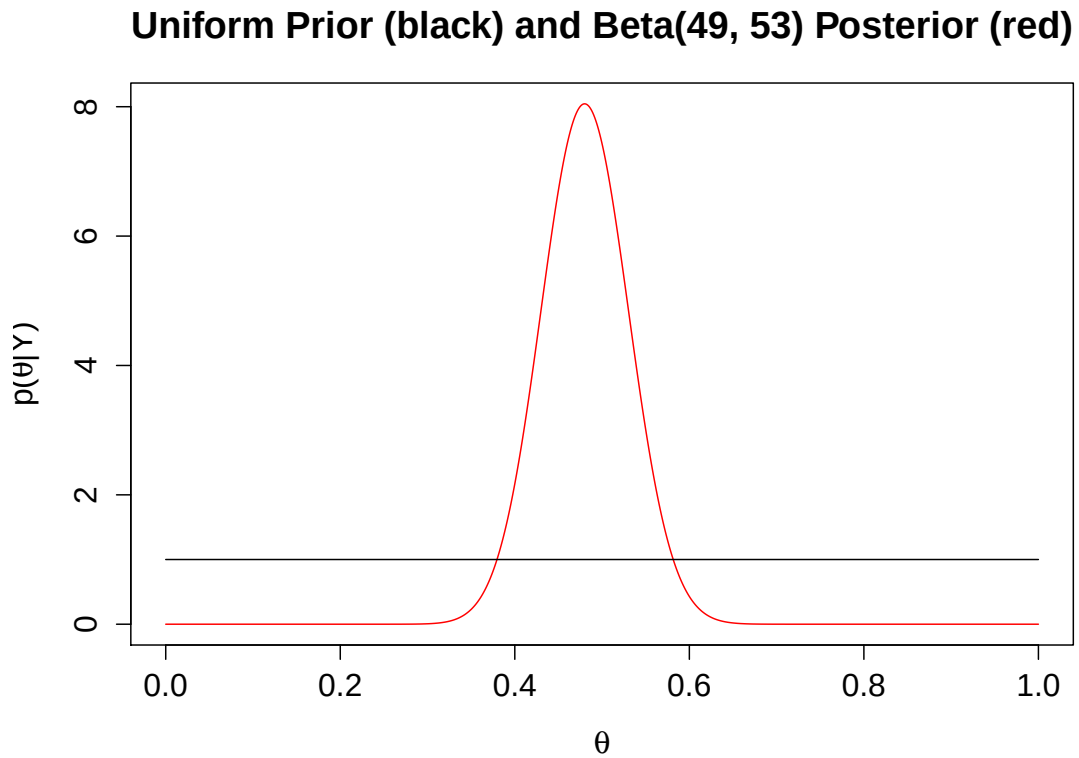
- The coin was tossed 100 times, and 48 tosses were heads.
- John's prior beliefs about θ were initially represented by a $Beta(138.4, 138.4)$.
- After seeing the data, his beliefs are updated to be a $Beta(138.4 + 48, 138.4 + 100 - 48) = Beta(186.4, 190.4)$ distribution.



- Similarly, Sarah's prior beliefs about θ were represented by $Beta(5.1, 5.1)$, and after seeing the data her posterior is $Beta(5.1 + 48, 5.1 + 100 - 48) = Beta(53.1, 57.1)$.



If the prior beliefs were represented by the uniform $Beta(1,1)$ prior, the posterior would be $Beta(49,53)$.



Key point: The *posterior distribution* $p(\theta|Y)$ represents all our knowledge about θ after observing Y.

In many situations we will want to give a point estimate of $p(\theta|Y)$ (e.g. similar to the frequentist maximum likelihood estimate). We have several choices, for example:

- We could estimate θ using the posterior **mean**
- We could estimate θ using the posterior **median**
- We could estimate θ using the posterior **mode**

All may be useful in different situations - in subsequent lecture we will discuss situations under which they arise.

Example 4 - continued

- But for now, suppose we choose to use the posterior **mean**.
- John's posterior $p(\theta|Y)$ was $Beta(186.4, 190.4)$. Recall that the mean of a Beta distribution is given by $\frac{\alpha}{\alpha+\beta}$. Hence, the mean of John's posterior is $186.4/(186.4 + 190.4) = 0.49$
- Similarly, the mean of Sarah's posterior is $53.1/(53.1 + 57.1) = 0.48$,
- The mean of the posterior based on the uniform prior is $49/(49+53) = 0.48$.
- Each person had a prior with a mean of 0.5. John has been less influenced by the data than Sarah because his prior beliefs that the coin was unbiased were stronger (i.e. his prior had less variance).

- Looking closely at the prior $Beta(\alpha, \beta)$ and the posterior $Beta(\alpha + Y, \beta + N - Y)$ we see that the posterior depends on the data through the number of heads Y, and the number of tails $N - Y$.
- Also, we can see how our prior beliefs get incorporated mathematically in this particular situation. The prior parameters α and β seem to feature in the posterior as additional heads and tails.
- That is, our prior beliefs in this particular situation seem to be adding extra heads and tails to the data we have observed.

- This suggests ways in which priors can be set up using objective information rather than subjective beliefs.
- Suppose that prior to the current round of 100 tosses, we had previously seen the same coin be tossed 20 times, of which 3 were heads.
- Then a reasonable prior for the current round of tosses would be $Beta(3, 17)$.
- In most situations we will try to construct sensible priors by incorporating previous information in this way.

- Furthermore, we can easily incorporate additional data if we do more tosses in the future. Suppose we tossed the coin another 200 times, and got 103 heads. What is John's posterior for θ after taking into account this additional data?
- After the earlier 100 tosses, his posterior was $Beta(186.4, 190.4)$. This is his belief about θ after those 100 tosses, but before the next 200 tosses. Thus, it becomes his prior for the next round of tosses.
- John's eventual posterior after additional 200 tosses is $Beta(186.4+103, 190.4+200-103) = Beta(289.4, 287.4)$, which now has a mean of 0.502.
- This fact that new information can be easily incorporated in this way is an attractive feature of Bayesian inference.

Credible Intervals

- We can also use the posterior distribution to construct an interval estimate for θ to represent our uncertainty, **credible interval**.
- Similarly, a 95% credible interval for θ is an interval $[a, b]$ of the posterior distribution of θ which contains 95% of the total area (i.e. which integrates to 0.95).
- Key point: unlike confidence intervals, credible intervals express degrees of belief. If $[a, b]$ is a 95% credible interval for θ, this means we assign probability 0.95 to the statement "θ lies in the interval $[a, b]$".

Choice of Prior Distribution

- The posterior distribution in this example was easy to analyse since it had a standard form - a Beta distribution.
- However in many situations things will not be as simple, and the posterior might end up being an unknown distribution, or one which can't be solved analytically.
- To make the posterior distribution easy to analyse mathematically, we often choose priors which are *conjugate to the likelihood*.
- Conjugacy means that the posterior distribution has the same form as the prior distribution, for example, a Beta prior with a Beta posterior, or a Gamma prior leading to a Gamma posterior, etc.

Next Week

Next Week - From Probability Distributions to Decision Making

- We have learned that Bayesian inference allows us to represent uncertainty about unknown quantities as posterior probability distributions.
- In many real situations we are not just interested in learning about model parameters - we are also interested in making decisions:

  - Should we evacuate a village because of a likely earthquake?
  - Should a bank change its investment portfolio to reduce its risk exposure?
  - Should a particular drug be recommended for patients?

- Real decisions depend on unknown quantities, but they also depend on the costs associated with each decision.
- How does the cost of failing to evacuate the village if the earthquake does happen, compare to the cost of wrongly evacuating the village if no earthquake occurs?

- Next week we will introduce a framework for linking posterior distributions about unknown quantities to actual decision making.