Lecture 1: Bayesian Inference

Gordon J. Ross

Bayesian Statistics

The statistics in this course will be taught from the Bayesian perspective. This has several important differences to the classical (frequentist) perspective that most people learn when they first study statistics,

Key difference - in frequentist statistics, probability statements are interpreted to be statements about the properties of repeated samples drawn from a distribution. In Bayesian statistics, probability statements simply express degrees of belief.

To a Bayesian a statement like ”the probability of there being aliens in our universe is 70%” is perfectly **meaningful** - it expresses a degree of belief about the world. However to a frequentist, these statements lie outside the realm of mathematical statistics unless they can be rephrased in terms of repeated sampling (”consider a process which creates universes at random, where each universe has a given probability of containing aliens...”).

Why Bayesian Statistics - Three Advantages

1. It allows prior information (e.g. from 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 there maybe only be 4 or 5 earthquakes to have ever occurred on some particular fault.
2. Bayesian probability statements are easy to interpret which is important when communicating with non-statisticians. Remember: a frequentist 95% confidence interval for a parameter θ does **not** mean that we are 95% sure that θ lies in the interval. However, Bayesian credible intervals **do** have this interpretation!
3. It makes everything a lot easier - no need to worry about coverage, unbiasedness, etc. 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.

Bayes Theorem

Bayes theorem is the core engine of Bayesian statistics, and allows us to update our beliefs about the world given new information.

Let A and B be statements. We start with an initial prior probability (= belief) $p(A)$ for A being true. If we then find out that B is true, then we update our beliefs about A given this new information. If $p(A|B)$ denotes our posterior probability for A being true given that B is true, then:

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

i.e. knowing that B is true alters our beliefs about A.

Bayes Theorem - Proof

Given statements A and B, the probability of both being true is (obviously) equal to the probability of A being true, multiplied by the probability of B being true given that A is true, i.e:

$$p(A \text{ and } B) = p(A)p(B|A)$$

On the other hand, the probability of both statements being true is also equal to the probability of B being true, multiplied by the probability of A being true given that B is true:

$$p(A \text{ and } B) = p(B)p(A|B)$$

Equating the two gives:

$$p(A)p(B|A) = p(B)p(A|B)$$

Rearranging:

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

Example 1

Suppose we toss a coin three times. There are eight equally likely outcomes: $\{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$. Define:

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

Then $p(A) = 3/8$ (since three of the eight outcomes have two heads) and $p(B) = 1/2$

We can compute $p(A|B)$ directly. Given that B it true, the only possible outcomes are $\{HHH, HHT, HTH, HTT\}$. Of these, two outcomes have two heads, so $p(A|B) = 2/4 = 1/2$.

Similarly if A occurs, the possible events are $\{HHT, HTH, THH\}$ and so $p(B|A) = 2/3$.

So we can verify Bayes Theorem here:
$p(A|B) = p(B|A)p(A)/p(B) = (2/3 * 3/8)/(1/2) = 1/2$

Bayes Theorem

Sometimes the denominator of Bayes Theorem $p(B)$ will not be given explicitly, and must be derived.

In the simplest case, let A' denote the statement "A is false" (i.e. the 'opposite' of A). Then by the theorem of total probability

$$p(B) = p(B|A)p(A) + p(B|A')p(A')$$

So Bayes theorem becomes:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|A')p(A')}$$

Example 2

A new medical screening test is developed to assess whether a patient has a particular disease. The test is advertised to have the following 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 chance that a person testing positive on the test really has the disease?

We first define the statements:

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

We need $p(A|B)$

Example 2 (cont)

Representing the given information mathematically, we have:

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

So:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|A')p(A')}$$

$$= \frac{0.95 * 0.001}{0.95 * 0.001 + 0.02 * 0.999} = 0.045$$

So 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 probabilty $p(A)$ of the person having the disease is very low.

Bayes Theorem

Although Bayes theorem is integral to Bayesian statistics, use of Bayes theorem does NOT make an analysis Bayesian!

Bayes theorem is simply a mathematical statement about how the probabilities of events and statements relate to each other. It features heavily in all types of statistics.

Again, the main difference is that in a Bayesian analysis, **probabilistic statements reflect degrees of beliefs**. We always start with prior beliefs before seeing the data, and use Bayes theorem to update these to give our posterior beliefs after incorporating the information from the data.

In frequentist statistics, we **never** make probabilistic statements about parameters. Remember: a confidence interval for θ does not mean that we are 95% sure that θ lies in some interval.

Bayesian Inference for Parameters

In a typical inference problem we have an unknown parameter θ which we wish to estimate. For example, θ may be the mean of a Normal distribution, or the probability of a particular coin landing heads when tossed. We also have data Y, such as the outcome of tossing the coin multiple times. We wish to use the data Y to learn about θ.

- The **prior distribution** $p(\theta)$ represents our beliefs about θ before incorporating the information from the data.
- The **posterior distribution** $p(\theta|Y)$ represents our beliefs about θ after incorporating the information from the data.

Bayes theorem tells us how to move from $p(\theta)$ to $p(\theta|Y)$. I.e. given we have some beliefs about θ before seeing the data, it tells us the beliefs $p(\theta|Y)$ we should have about θ after seeing the data.

Bayes Theorem for Parameters

The form of Bayes theorem used here is almost identical to before:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)}$$

Here $p(Y|\theta)$ denotes the likelihood function, and $p(Y)$ is the marginal probability of observing Y, which by the theorem of total probability is:

$$p(Y) = \int p(Y|\theta)p(\theta)d\theta$$

Example

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, and a confidence interval can be constructed around this.

Is this reasonable? Well, say 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. The vast majority of coins in the world are not biased, and observing 48 heads in 100 tosses is a normal outcome from tossing an unbiased coin.

In other words, rather than concluding that $\theta = 0.48$, we may wish to include prior information to make a more informed judgement.

Example (cont)

In a Bayesian analysis, we first need to represent our prior beliefs about θ. Specifically, we construct a probability distribution $p(\theta)$ which encapsulates our beliefs.

There is no one way to do this! $p(\theta)$ represents the beliefs of one particular person based on their assessment of the prior evidence – it will not be the same for different people if they 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 since θ here represents the probability of the coin landing heads, it must lie between 0 and 1. So the function we use to represent our beliefs should only have mass in the interval $[0, 1]$, which rules out (e.g.) the Normal distribution.

Example (cont)

In this situation it is usual to represent our prior beliefs as a Beta distribution (I will explain why in more detail later). The Beta distribution only has mass in $[0, 1]$ so it is a sensible choice when θ is a probability.

Recall the Beta distribution has the form:

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

where $B(\alpha, \beta)$ is the Beta function. α and β are **parameters** which control the shape of the distribution. We choose these to reflect our prior beliefs about θ. How do we do this?

Example (cont)

It can be shown (see any statistics textbook or wikipedia) that the mean and variance of the Beta distribution is given by:

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

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

So, 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 lets us express α and β in terms of the mean/variance:

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

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

Example (cont)

We can write an R function for converting beliefs about the mean/variance of θ into beliefs about α and β:

```
#returns params of beta distribtuion in terms of mean/variance
betaParams <- function(mu, sigma2) {
  alpha <- ( (1-mu)/sigma2 - 1/mu)*mu^2
  beta <- alpha*(1/mu - 1)
  return(c(alpha,beta))
}
> mu <- 0.5
> sigma <- 0.03
> var <- sigma^2
> betaParams(mu,var)
[1] 138.3889 138.3889
```

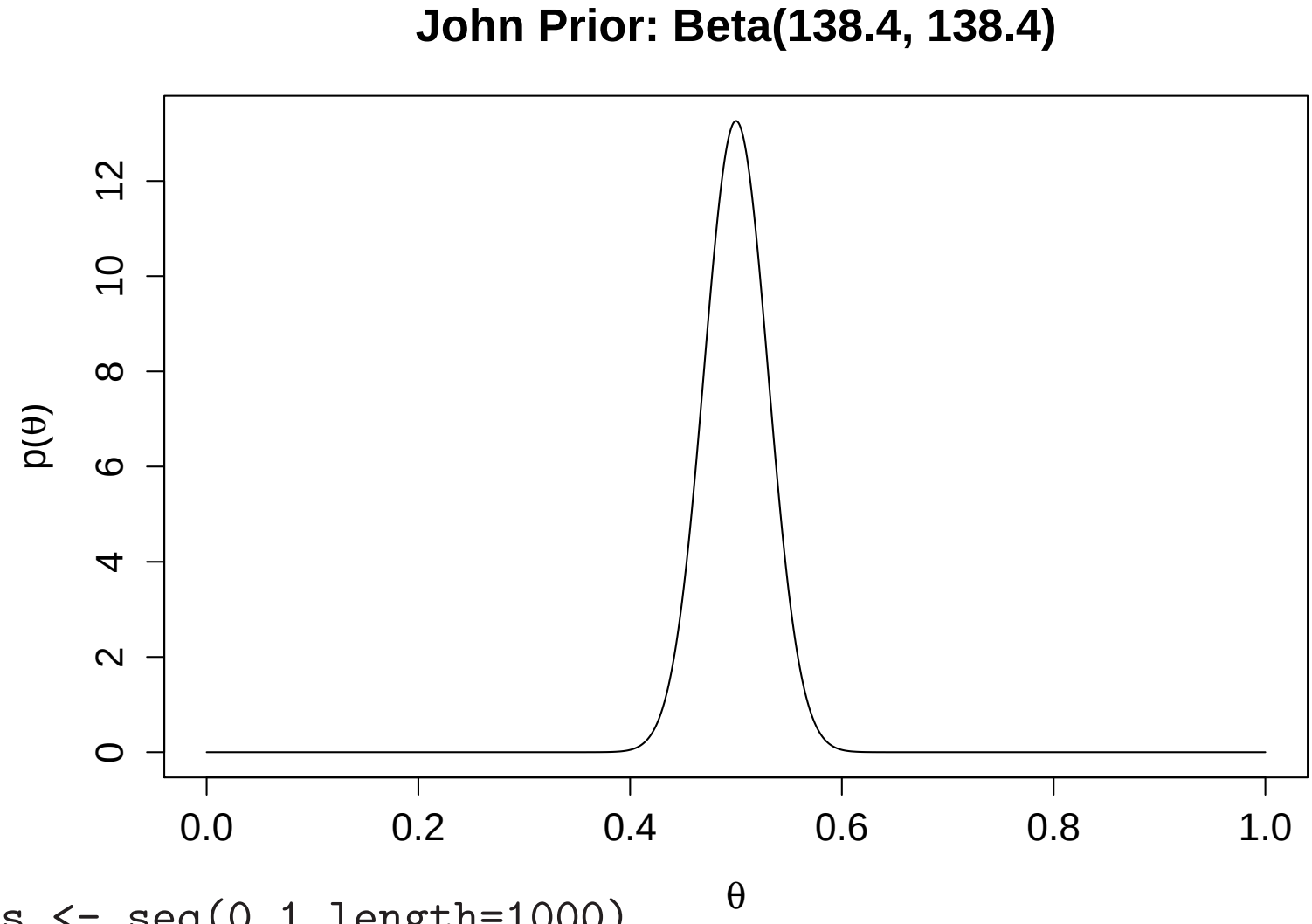
Example (cont)

So for example, lets say that John has prior belief that $E(\theta) = 0.5$. He doesnt expect the coin to be based (most coins are not biased) so he takes the standard deviation to be low, say 0.03. Based on the previous equations, his prior is hence: Beta$(138.4, 138.4)$

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

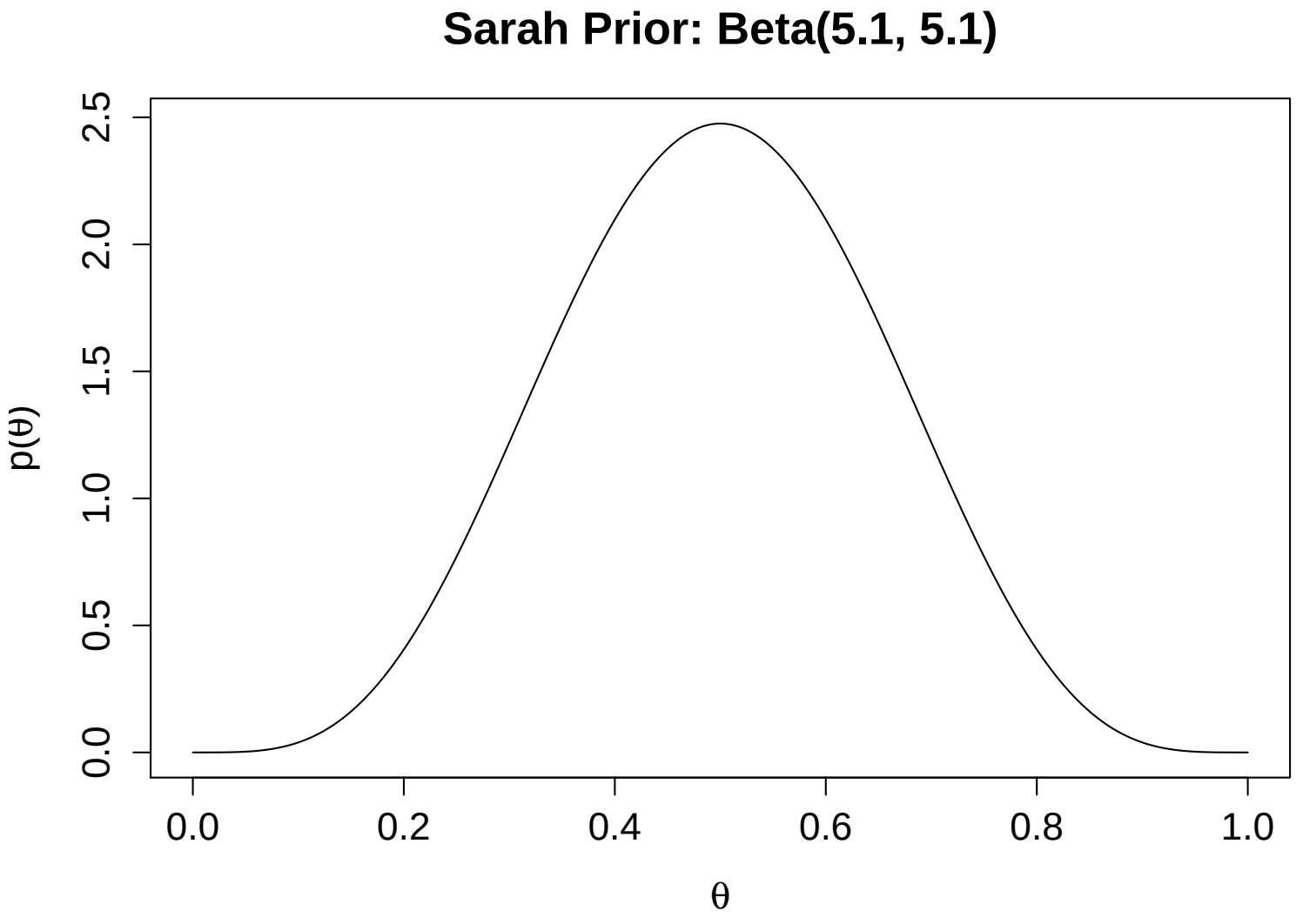
We can plot these Beta distributions in a language like R, to see their shape

Example (cont) - John's Prior



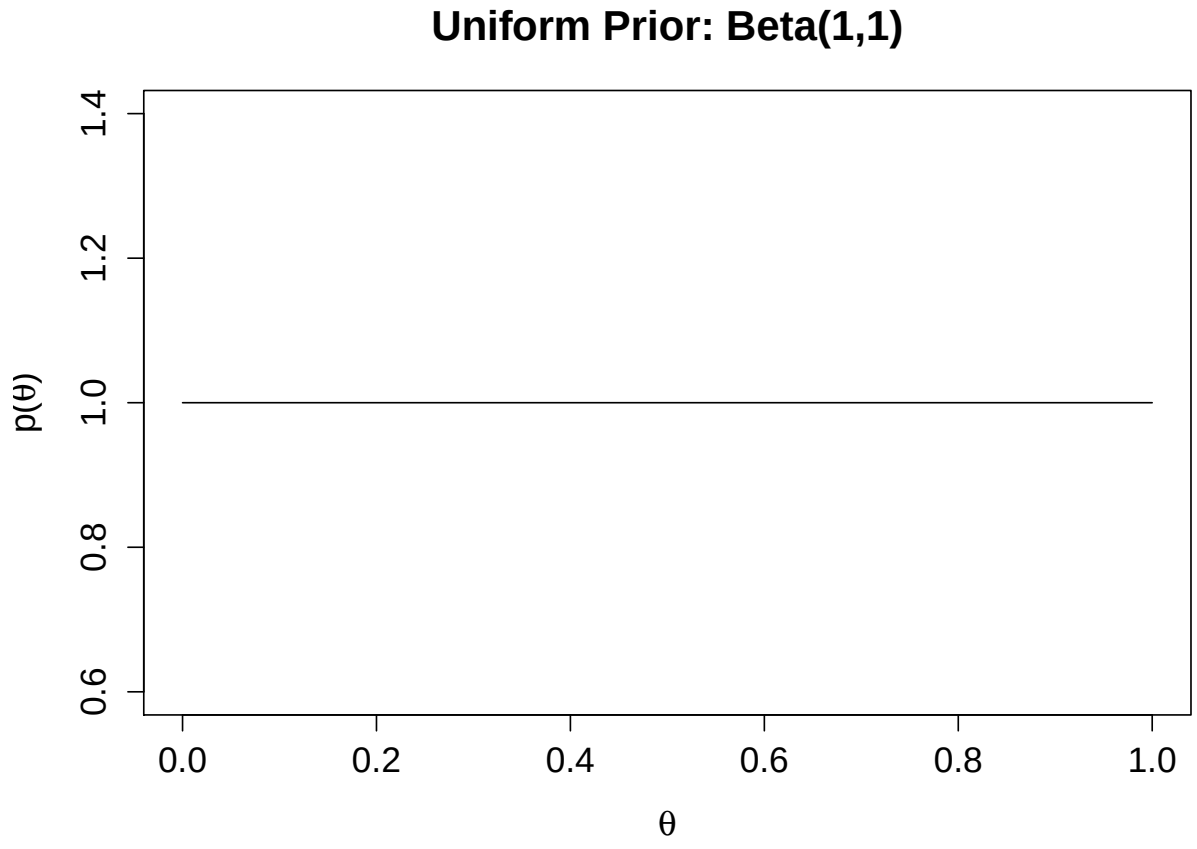

```
xaxis <- seq(0,1,length=1000)
plot(xaxis, dbeta(xaxis, 138.4, 138.4), type='l')
```

Example (cont) - Sarah's Prior



Example (cont) - Uniform Prior

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:



Example (cont) - Analysis

The prior distribution is hence $Beta(\alpha, \beta)$. Since this is coin tossing, each of the N tosses has probability θ to be heads. Each individual toss follows a Bernoulli(θ) distribution, and so the likelihood $p(Y|\theta)$ for the number of heads Y is hence a Binomial(N, θ) distribution.

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

Example (cont) - Analysis

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

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

The numerator here is:

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

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

Example (cont) - Analysis)

The denominator is:

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

This looks horrible, but there is a standard trick we can use here (and in many other situations – so learn it!). First, take everything that doesnt depend on θ outside the integral:

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

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

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

Example (cont) - Analysis

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

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

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:

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

So:

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

Example (cont) - Analysis

Combining this with the numerator gives:

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

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(α, β) distribution. And after seeing the data, we must revise our beliefs about θ to be a Beta$(\alpha + Y, \beta + N - Y)$ distribution.

Example (cont) - Analysis

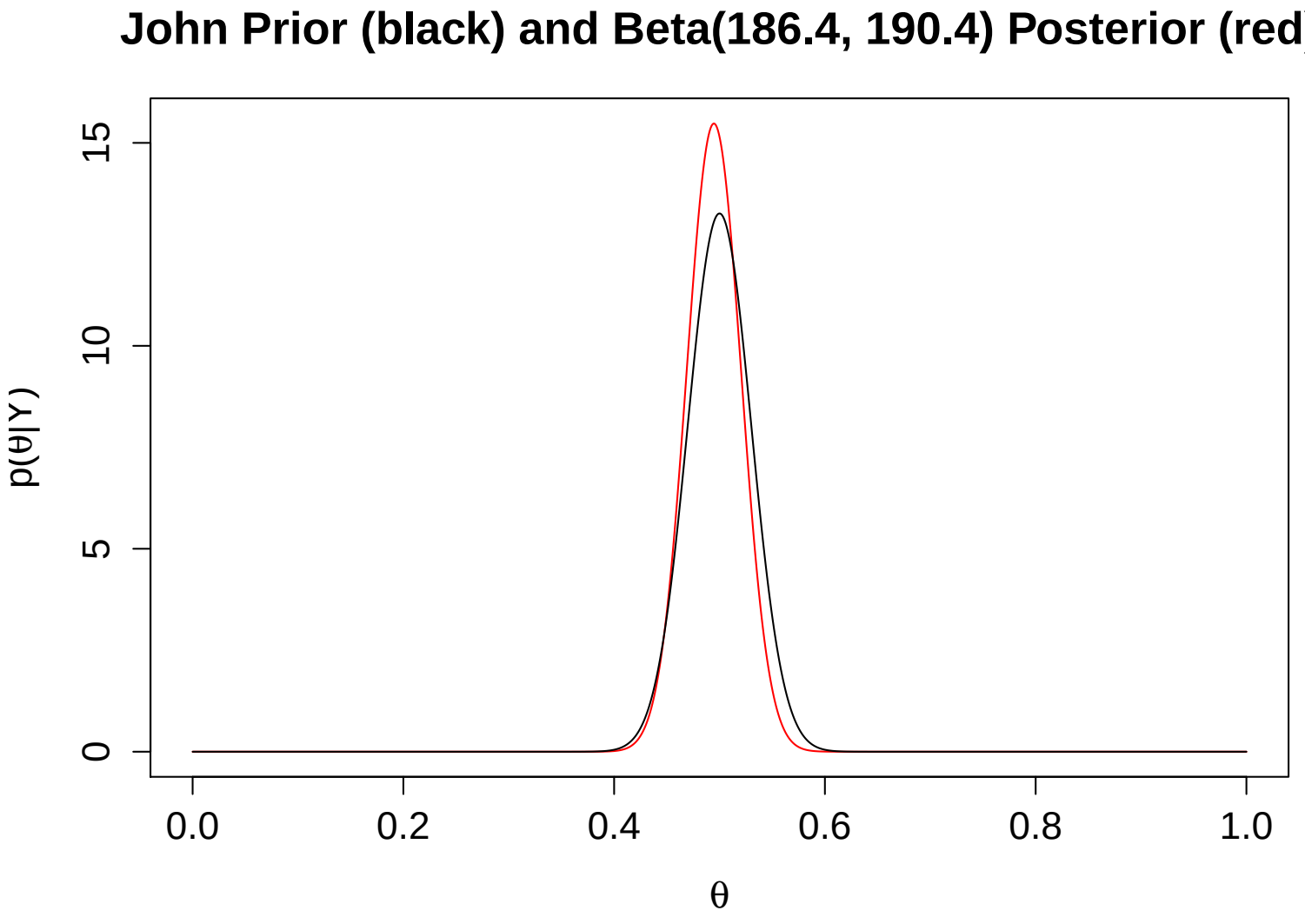
The coin was tossed 100 times, and 48 tosses were heads.

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.

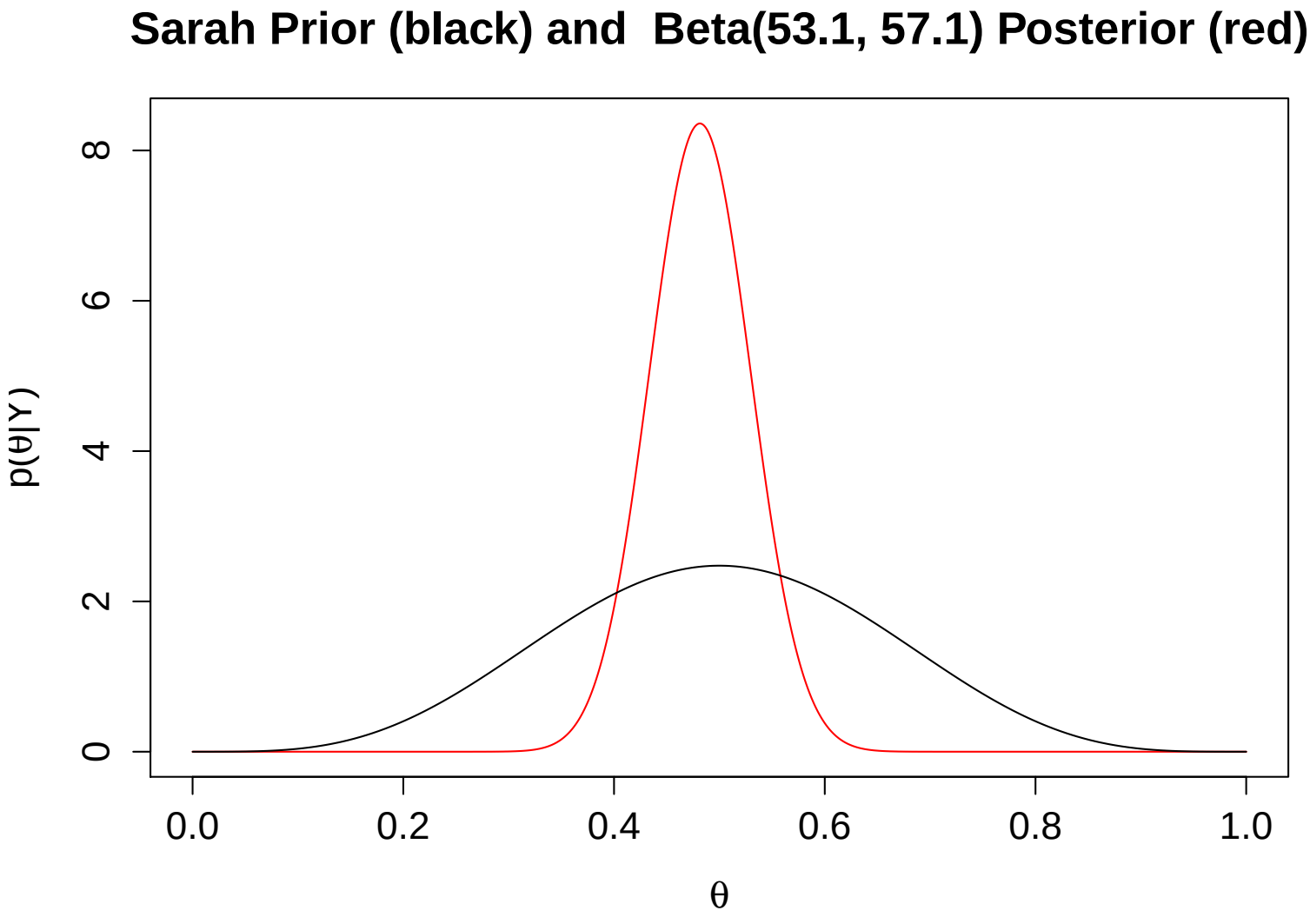
Similarly, Sarah's prior beliefs had a , and her posterior is Beta(53.1, 57.1)

And if we had used the uniform Beta(1,1) prior, the posterior would be Beta(49, 53)

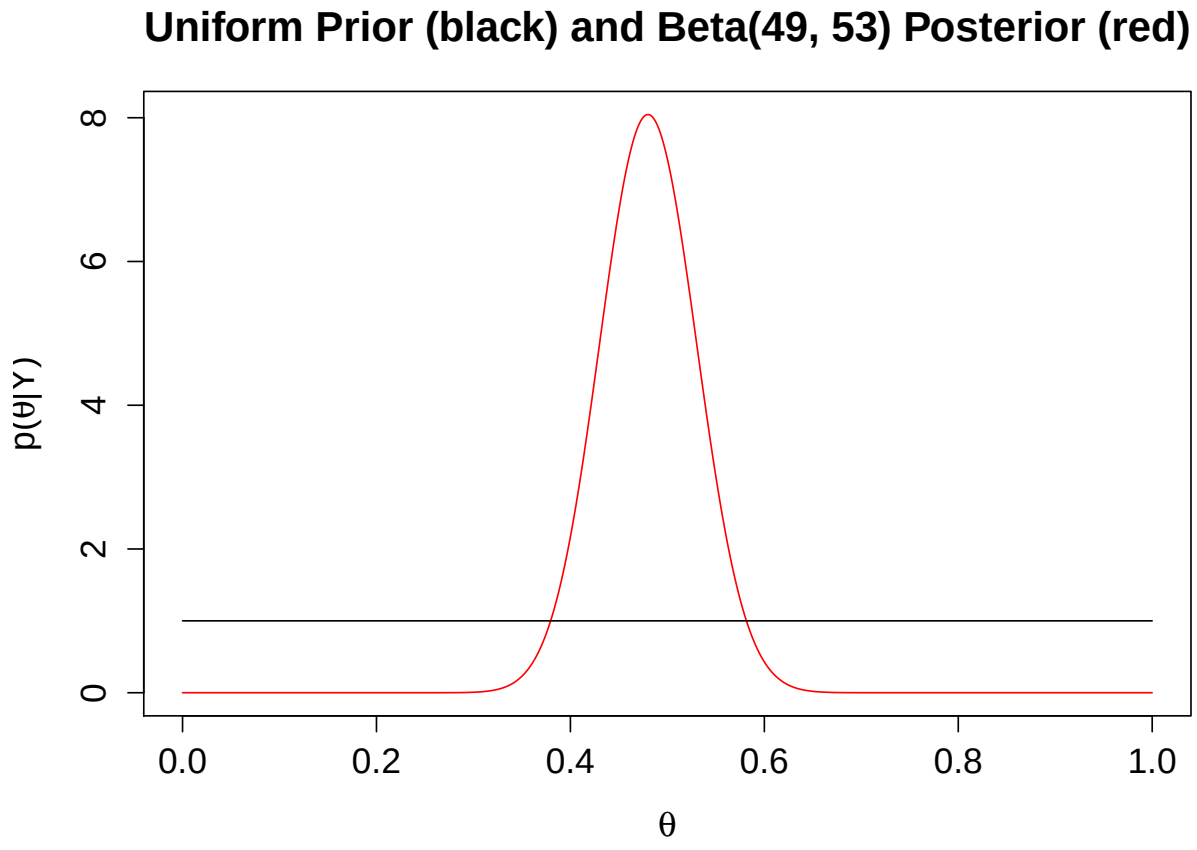
Example (cont) - John's Posterior



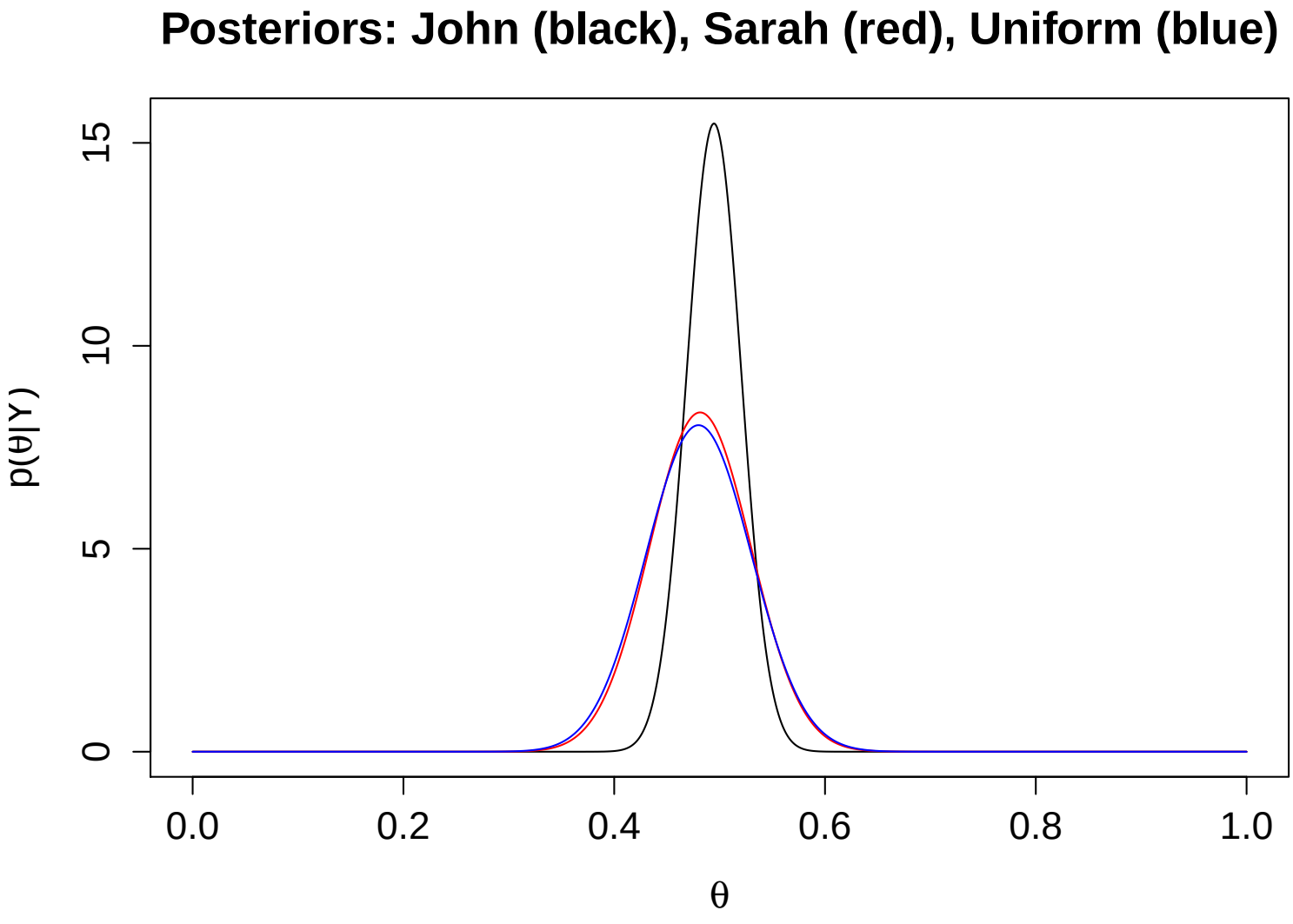
Example (cont) - Sarah's Posterior



Example (cont) - Posterior with Uniform Prior



Example (cont) - All Posteriors



Posterior Summaries

Key point: **The posterior distribtion $p(\theta|Y)$ represents all our knowledge about θ after observing Y.** In other words, any statements we make about θ should be based on the posterior and nothing else.

In many situations we will want to give a point estimate of θ (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 - next week we will be more precise about this. But for now, suppose we choose to use the posterior mean.

Posterior Summaries - Example

John's posterior $p(\theta|Y)$ was Beta(186.4, 190.4). Recall from earlier that the mean of a Beta distribution is given by $\alpha/(\alpha+\beta)$. So the mean of John's posterior is $186.4/(186.4+190.4) = 0.49$

Similarly, the mean of Sarah's posterior is $53.1/(53.1+57.1) = 0.48$, and 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 (his prior had less variance).

Credible Intervals

We can also use the posterior distribution to construct an interval estimate for θ to represent our uncertainty. A frequentist would express uncertainty about θ using a confidence interval. The Bayesian equivalent is a credible interval.

Recall: a 95% confidence interval for θ is an interval $[a, b]$ of the **sampling distribution** of θ which contains 95% of the total area (i.e. which integrates to 0.95).

Similarly, a 95% credible interval for θ is an interval $[a, b]$ of the **posteror 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]". This is **not** (not, not, not!) the case for confidence intervals - a fact that can

Example (cont) - An Additional Remark

The previous prior was Beta(α, β) and the posterior was Beta($Y + \alpha$, $N - Y - \beta$). Looking closely, we see the posterior depends on the data through the number of heads (Y) and the number of tails (N-Y).

The prior parameters α and β seem to feature in the posterior as **additional heads and tails**. I.e. our prior beliefs in this particular situation seem to be adding extra heads and tails to the data we have observed. This is (in this particular situation) how our prior beliefs get incorporated mathematically.

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.

Example (cont) - An Additional Remark

The same applies 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?

Well, 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. So it becomes his prior for the next round of tosses.

So his eventual posterior after all 300 tosses is Beta$(186.4 + 103, 190.4 + 200 - 103)$ = Beta(289.4, 287.4), which has a mean of 0.502.

This fact that new information can be easily incorporated in this way is a key feature of Bayesian inference.

Choice of Prior Distribution

The posterior distributiion 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.

The key point here is the integral in the denominator of Bayes theorem: $p(Y) = \int p(Y|\theta)p(\theta)d\theta$. Typically if we can solve this integral analytically then the posterior will be easy to analyse. But in many cases it will be impossible to do this integral (remember: outside of A-Level mathematics classes, "most" integrals cannot be solved analytically!).

In this case the integral must be solved numerically instead. We will discuss this at length in a future lecture. But for now, lets focus on the cases where this integral can be solved analytically.

Note that we solved it here by **recognising that it had the same form as a Beta distribution**. This was not a coincidence!

Choice of Prior Distribution

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.

When using conjugate priors, we can solve the integral $p(Y) = \int p(Y|\theta)p(\theta)d\theta$ analytically by using the trick we used earlier – i.e "recognising that it has the same form as the prior" and must integrate to 1 due to being a probability distribution., This makes the posterior easy to analyse. So, choosing a conjugate prior makes things much simpler. How do we find conjugate priors?

Conjugate Priors

Definiton: If $p(\theta)$ belongs to the same family of probability distributions as $p(\theta|Y)$ then $p(\theta)$ is a conjugate prior for θ

From Bayes theorem, focusing only on the numerator, this will be true if $p(\theta|Y)$ has the same general form as $p(\theta)p(Y|\theta)$. By 'general form' I mean "the parts which depend on θ.

In our example the likelihood was Binomial:

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

In order for the prior to keep its same form when multiplied by this, the part depending on θ must be proportional to:

$$p(\theta|Y) \propto \theta^{r}(1-\theta)^{s}$$

for some constants r and s. But this is just the form of the Beta distribution! So the Beta distribution is the conjugate prior for the Binomial distribution.

Conjugate Priors

In general to find a conjugate prior, either use the argument from the previous slide to derive one, or consult a standard table (wikipedia has a very good list on the "Conjugate Prior" page)

n.wikipedia.org/wiki/Conjugate_prior — wikipedia conjugate priors

assimilation.

Table of conjugate distributions [edit]

Let *n* denote the number of observations. In all cases below, the data is assumed to consist of *n* points $x_1, \ldots, x_n$ (which will be random vectors in the multivariate cases).

If the likelihood function belongs to the exponential family, then a conjugate prior exists, often also in the exponential family; see Exponential family: Conjugate distributions.

Discrete distributions [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]
Bernoulli	*p* (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + n - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]
Binomial	*p* (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]
Negative Binomial with known failure number *r*	*p* (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + rn$	$\alpha - 1$ total successes, $\beta - 1$ failures[note 1] (i.e. $\frac{\beta - 1}{r}$ experiments, assuming r stays fixed)
Poisson	λ (rate)	Gamma	$k,\ \theta$	$k + \sum_{i=1}^{n} x_i,\ \frac{\theta}{n\theta + 1}$	k total occurrences in $1/\theta$ intervals
Poisson	λ (rate)	Gamma	$\alpha,\ \beta$ [note 3]	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + n$	α total occurrences in β intervals
Categorical	$\boldsymbol{p}$ (probability vector), *k* (number of categories, i.e.	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + (c_1, \ldots, c_k)$, where c_i is the number of observations in category *i*	$\alpha_i - 1$ occurrences of category i[note 1]

New Example

Suppose that in a particular region of the world, N earthquakes have occurred over the last 2000 years. Their occurrence times are $t_1, t_2, \ldots, t_N$. Under the most simple model of seismicity, these earthquakes are assumed to follow a Poisson process, in which case the time-between-events $\tau_i = t_i - t_{i-1}$ follow an Exponential distribution with parameter λ.

For the purpose of predicting the occurrence of future earthquakes, we wish to learn about λ. I.e. given the independent and identically distribtuted observations $\tau_1, \ldots, \tau_{N-1}$ where $\tau_i \sim \mathit{Exponential}(\lambda)$, we wish to infer λ.

As before, we start with a prior distribution which represents our knowledge about λ before analysing the data. Lets try to find a conjugate prior.

New Example - Conjugate Prior

Since the data is Exponential, the likelihood is:

$$p(\lambda|Y) = \lambda e^{-\lambda Y}$$

For the prior distribution to keep its same form after being multiplied by the likelihood, it must be proportional to:

$$p(\lambda) \propto \lambda^r e^{-s\lambda}$$

for some r and s. If we consult list of probability distributions, we find that distribution with this form is the Gamma distribution:

$$p(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

So this is the conjugate prior for the Exponential distribution.

Note: if you do not recognise these forms (or have never seen the Gamma distribution before) then don't worry – again, the conjugate priors for most common probability distributions can just be looked up in a table.

New Example

As before, we choose the parameters of the prior α and β to reflect our prior beliefs about λ. These will be based on either seismological theory, or evidence from other similar earthquake regions.

Your task: given a particular choice of α and β, compute the posterior distribution using a similar argument to that which we used for the previous coin tossing example. If you get stuck, the Exercise sheet for this week will walk you through the process.

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 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.

Additional Remark

If you are observant then you may have noticed that when doing the Binomial example, **we didnt actually need to calculate the integral in the denominator at all!** Recall that the numerator was:

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

As soon as we see this expression, we can notice that the θ dependent part has the same form as the Beta distribution. Therefore the posterior must be Beta. Everything that does not depend on θ is simply part of the normalising constant which ensures the distribution integrates to 1. As such, in Bayesian analysis we often focus only on the part of the posterior which is proportional to the parameter of interest:

$$p(\theta|Y) \propto \theta^{Y+\alpha-1}(1-\theta)^{N-Y+\beta-1}$$

We will discuss this at length in a future lecture.