University of Glasgow

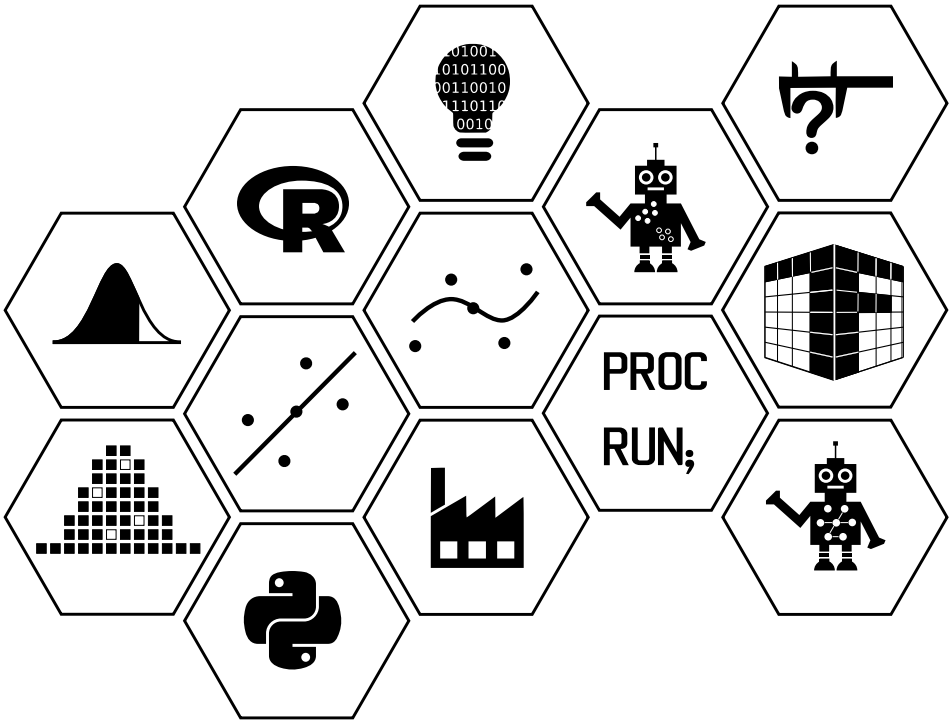
Learning from Data/Data Science Foundations

Claire Miller and Eilidh Jack

Academic Year 2021-22

Supplement - part 1:

Introduction to Bayesian inference



DATA ANALYTICS GLASGOW

Introduction to Bayesian inference

This supplement will provide a short introduction to the other major philosophy within statistics known as Bayesian inference. We will only give an introduction to the main ideas and principles here which will then be expanded upon in courses to follow, particularly the course on *uncertainty assessment and Bayesian computation*.

Motivation

Consider the scenarios in the following example.

Example 1 (A music expert, a drunk person and a tea drinker).

In each of the following cases our data model is $Y|\theta \sim \text{Bin}(10;\theta)$ and we observe $y = 10$.

- A music expert claims that they can distinguish between Mozart and Beethoven concertos by listening to the first 3 seconds of a piece of music. The expert does so correctly for 10 different concertos.
- A person, who has drunk a lot of alcohol, claims that they can predict the outcome of tossing a fair coin, and does so correctly for 10 tosses.
- A tea drinker claims that they can detect from a cup of tea whether the milk was added before or after the tea. The tea drinker does so correctly for 10 cups.

Ask yourself the following questions about the above scenarios.

- Do any of the above scenarios and outcomes surprise you?
- If you had been asked to predict in each case how many of the 10 trials each person would get correct, would you have guessed 10 out of 10 in any of these cases?
- Before obtaining the data, would you have guessed that there would be a different proportion correct in each of the scenarios? If so, why?
- Using a likelihood approach, would would be the maximum likelihood estimate (MLE) in each case?

For the scenarios above that involved the experts, i.e. a musician and tea drinker, before the experiment you might have had a prior impression that these two people have prior experience that they will use to inform their decision (instead of it being purely random chance). For both of these scenarios, you might have guessed that the proportion correct would be quite high.

However, for the second scenario with the tossing a coin example, your initial impression (before collecting data) might have been that the proportion correct here may be closer to 0.5, since with a fair coin this is what we would expect, on average, over many trials. Whether or not the person has been drinking alcohol is irrelevant here.

Taking just a likelihood approach, the MLE in all cases would be 1.

The point in these scenarios is that experiments are sometimes not abstract events.

We sometimes have knowledge about the process being investigated before obtaining the data.

It is, therefore, sensible that inferences should be based on the combined information that this prior knowledge and the data represent.

Bayesian inference is the mechanism for drawing inference from this combined knowledge.

The first video provides a motivating example for the use of Bayesian inference. A real example of lake water quality is described in which Bayesian inference has been used to combine data obtained from two different sources (in-situ samples and satellite remote sensing) and to incorporate prior information on the data sources. This inference provides information on how the water quality changes over space within the lake.

Motivating example for Bayesian inference

https://youtu.be/_6HXdvEAJRA

Duration: 7m44s

Bayesian inference

Bayesian inference is a method of statistical inference in which evidence is used to estimate parameters and predictions. Additional available information can be accounted for besides that provided by the data e.g. from previous studies. It is a statistical method for using data to update prior beliefs.

The arguments that are made in favour of the Bayesian approach are that it offers more intuitive and meaningful inferences, that it gives the ability to tackle more complex problems and that it allows the incorporation of prior information in addition to the data. However, the subjectivity surrounding incorporating the prior information is one of the main objections to the method.

As stated in week 3, non-Bayesian statistics is often called frequentist statistics and the two forms of statistics have different definitions for probability:

- Bayesian: the degree to which a person believes in a proposition;
- Frequentist: the *limit* of the relative frequency of an event in a large number of trials.

In Bayesian statistics probabilities can be assigned to anything with or without data. However, in frequentist statistics probabilities can only be assigned to *random experiments* or *events*. In Bayesian statistics more information can be accounted for besides that in the data (e.g. from previous studies). However, the way in which that information is incorporated has to be carefully considered.

Probability statements about parameters or even more fundamentally about predictions based on parameters become the goal of inference.

For any experiment that we undertake we can issue a probability statement before we undertake it. This is then referred to as a prior probability. This prior probability can then be updated using the information provided by the data from our experiment to produce a posterior probability. This updating is done using Bayes' theorem, which is why the approach is called **Bayesian**.

A general framework

Suppose that you are interested in the values of k unknown parameters:

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k),$$

and you have some prior beliefs about the values of the parameters which can be expressed in terms of a p.d.f., $p(\boldsymbol{\theta})$. Suppose also that we have n observations $X_1, ..., X_n$ which have a probability distribution that depends on the k unknown parameters, so that the p.d.f. of the data $X_1, ..., X_n$ depends on the parameters $\boldsymbol{\theta}$. Assuming $X_1, ..., X_n$ are random variables; the dependence of $X_1, ..., X_n$ on $\boldsymbol{\theta}$ can be expressed in terms of a p.d.f., $p(\mathbf{X}|\boldsymbol{\theta})$, this is the data distribution.

To simplify the notation here we will use $p()$ for a p.m.f. or p.d.f.

Bayes' theorem for random variables can be used to express your beliefs about $\boldsymbol{\theta}$ taking into account both your prior beliefs and the data.

Bayes' Theorem - a reminder

Consider two events A and B. We can write their joint probability as $P(A \cap B)$. This can be calculated in one of two ways so that we can write

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B),$$

where $P(A|B)$ is the probability of event A conditional on event B.

From this we can obtain an expression for $P(A|B)$ as

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}.$$

This expression is a simple form of what is known as Bayes' theorem, named after a posthumous paper of 1763 by Thomas Bayes (1701-1761). The theorem is, of itself, an uncontroversial and elementary result in conditional probability[1]. However, it seems to offer the possibility of providing a means by which probability statements may be updated using evidence.

Task 1 (Clinical test for HIV).

Revise the use of Bayes' theorem through the following example.

A clinical test for HIV is applied to a population which is at high risk for HIV; 10% of this population are believed to be HIV positive. The clinical test is positive for 90% of people who are genuinely HIV positive, and negative for 85% of people who are not HIV positive.

What are the probabilities of false positive and false negative results?

To be more precise, let:

D = { disease is present }

D^c = { disease is not present }

T = { clinical test is positive }

T^c = { clinical test is negative }

find Pr$(D^c|T)$ and Pr$(D|T^c)$.

Computing posterior distributions

Bayes' theorem works for probability density functions too giving us (in the case of continuous $\boldsymbol{\theta}$):

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{X}) &= \frac{p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})}{p(\mathbf{X})} \\ &= \frac{p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &\propto p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta}), \end{aligned}$$

since the denominator is a number (see page 5 for more detail).

[1] See the course of Probability and Stochastic Models or Probability and Sampling Fundamentals for more detail

In the case of discrete $\boldsymbol{\theta}$ the above becomes:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{X}) = & \quad \frac{p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})}{p(\mathbf{X})} \\ = & \quad \frac{p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})}{\sum p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})} \\ \propto & \quad p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta}). \end{aligned}$$

In previous weeks we have introduced the fact that, for x, $p(x|\theta) = L(\theta|x)$ or, for $x_1, \ldots, x_n$, $L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_1, \ldots, x_n|\theta)$ i.e.

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{X})}{p(\mathbf{X})}.$$

Definitions

For the case of one data point x and a parameter θ:

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

- θ, x: are the parameter and the data respectively;
- $p(\theta|x)$: is the ***posterior distribution*** for a parameter given the data (and prior);
- $L(\theta|x)$: is the ***likelihood*** for a parameter given the data;
- $p(\theta)$: is the ***prior distribution*** of the parameter θ;
- $p(x)$: is the unconditional (marginal) distribution of the data.

The prior distribution $p(\theta)$ can be interpreted as the probability distribution before we see the data, or, our belief in the parameters before we see the data.

The posterior distribution $p(\theta|x)$ can be interpreted as the probability distribution after we see the data (and have incorporated our prior beliefs for the parameters).

The unconditional (marginal) distribution of the data can be written (for the continuous case) as:

$$p(x) = \int p(\theta)p(x|\theta)d\theta,$$

or

$$p(x) = \int p(\theta)L(\theta|x)d\theta,$$

i.e. integrating out θ for the numerator. Hence $p(x)$ is a **normalising constant**. It is sometimes called the **marginal distribution** of the data or the prior predictive distribution of the data.

We can state that:

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

i.e. ignoring the normalising constant.

Therefore, **the posterior probability is proportional to the prior probability times the likelihood,**

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

The second video below looks at visualising the prior, likelihood and posterior distributions and their definitions.

Visualising prior, likelihood and posterior

https://youtu.be/WMAVrGchrfw

Duration: 3m37s

Examples

Example 2.

We toss a coin and the outcome is either a head or a tail. Suppose that the outcome of a head is considered a success and hence X is the number of heads observed. The proportion of successes θ can be modelled using a binomial distribution with sample size n and parameter θ.

Let's look at likelihood, prior and posterior distributions.

The binomial model states that:

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

where in this example n is the number of trials and x is the number of successes or the number of heads. We could use Bayesian inference to estimate the proportion of success, θ.

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

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

such that,

$$L(\theta|x) = p(x|\theta)$$

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

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

To perform Bayesian inference here we must specify a prior distribution for the parameter θ.

The question is, how, in general do we approach the problem of constructing prior distributions?

Typically, the prior distribution should include all plausible values of θ, but the distribution need not be realistically concentrated around the true value, because often the information about θ contained in the data will far outweigh any reasonable prior probability specification. We will use a beta prior distribution here. A beta distribution is a **conjugate family** for the binomial likelihood and this means that the prior has the same parametric distributional form as the posterior distribution. We'll return to more details on conjugate priors later.

The beta distribution is defined on the [0,1] interval and so is suitable for proportions and probabilities.

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

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

This prior density is equivalent to $\alpha - 1$ prior successes and $\beta - 1$ prior failures.

A special case of this is that:

$$\alpha = \beta = 1$$

yields a Uniform $(0, 1)$ distribution.

The parameters of the prior distribution are often referred to as ***hyperparameters***.

If x is distributed as a binomial distribution with known sample size n and unknown parameter θ and the prior distribution for θ is a beta distribution with parameters α and β, then the posterior distribution for θ is also a beta distribution (since the beta distribution is a conjugate prior here). See the link to supplementary material on page 10 for a derivation of this result.

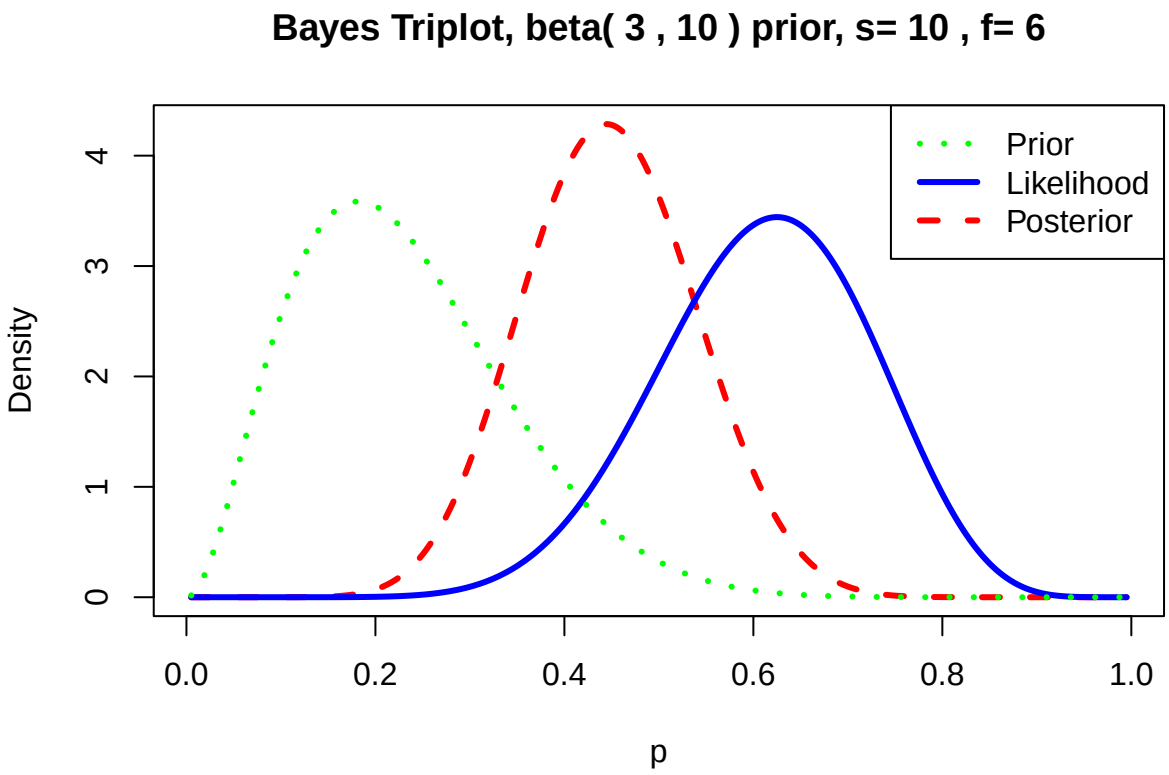
The following examples illustrate the relationship between the prior, posterior and likelihood distributions where we have simple binomial data and a prior that has a beta distribution.

Example 3 (Binomial data with 10 successes and 6 failures).

Suppose we use a conjugate beta prior for θ with parameters $\alpha = 3$ and $\beta = 10$, i.e. the prior is Be(3,10).

Let's visualise the prior, likelihood and posterior distributions using R.

```
library(LearnBayes)
prior <- c(3,10)      # proportion has a Be(3, 10) prior
data <- c(10,6)       # observe 10 successes and 6 failures
triplot(prior,data)
```



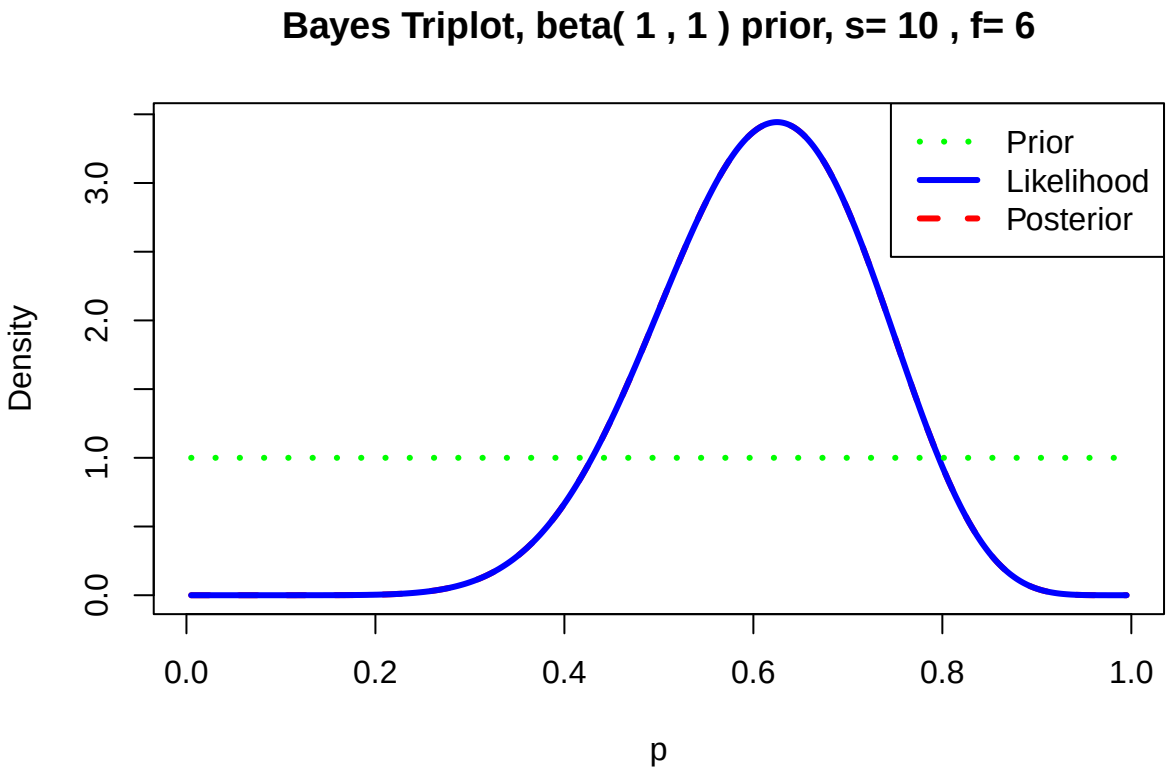
- The likelihood here provides an MLE for θ of 0.625 (where, on the x-axis, p=θ).
- The prior distribution has pulled the posterior to the left of the likelihood;
- The posterior distribution appears centred between the prior and likelihood.

Note: the likelihood has been scaled in the `LearnBayes` package to make it easier to compare with the prior and posterior.

Example 4 (Binomial data with 10 successes and 6 failures).

Now suppose we change the parameters of the prior distribution to $\alpha = \beta = 1$.

```
library(LearnBayes)
prior <- c(1,1)     # proportion has a Be(1, 1) prior
data <- c(10,6)      # observe 10 successes and 6 failures
triplot(prior,data)
```



- The likelihood here provides an MLE for θ of 0.625;
- The prior is flat;
- The likelihood and the posterior are indistinguishable.

Task 2.

Use the `LearnBayes` package in `R` to visualise the prior, likelihood and posterior distributions for the following:

- Binomial data with 10 successes and 10 failures and a Beta(1, 1) prior
- Binomial data with 10 successes and 10 failures and a Beta(0.5, 2) prior.

Note: in the examples above the beta parameters α and β have been selected in an arbitrary way to illustrate the effect of different parameters for the priors. There is no specific reason for the choice of these values, and, in fact, selecting an appropriate prior can be challenging.

Prior distributions

This issue is fundamental to the Bayesian framework.

The fundamental difference between Bayesian and classical statistics is that in Bayesian statistics unknown parameters are treated as random variables. The use of Bayes' theorem requires the specification of prior distributions for these parameters. The choice of prior distribution cannot be made blindly; considerable care is needed and there are some very substantial issues involved.

The prior represents your beliefs about θ before observing the data. Someone else's priors would lead to a different posterior analysis, as we've seen in the examples and task above.

As long as the prior is not *completely unreasonable* then the effect of the prior becomes less influential as more and more data become available.

Often we might have a *rough idea* what the prior should look like (perhaps we could give its mean and variance), but cannot be more precise than that.

Sometimes we might feel that we have no prior information about a parameter. In such situations we might wish to use a prior which reflects our ignorance about the parameter.

The computational difficulties arise in using Bayes' theorem when it is necessary to evaluate the normalisation constant in the denominator i.e. evaluating:

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

In this course we will focus on what are known as **conjugate priors** (as we've introduced in example 2).

Conjugate priors have the same parametric distributional form as the posterior distribution. They are easy to update into posteriors and are easy to display and summarize. This is because all we have to compute is:

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

and the normalising constant is then known from the distributional form of the posterior.

For conjugate priors of low-dimensional statistical models it is usually computationally straight forward to also compute these algebraically, see the link to the supplementary material on page 10 for examples here.

There are clear advantages in using a conjugate prior because of the greater simplicity of the computations (i.e. the distributional form of the posterior distribution is known). However, when no member of the conjugate family is appropriate then you may well have to proceed using numerical integration to investigate the properties of the posterior.

Here is a summary of the conjugate priors for some known distributions:

Likelihood	Conjugate prior	Posterior
Binomial	Beta	Beta
Geometric	Beta	Beta
Poisson	Gamma	Gamma
Exponential	Gamma	Gamma
Normal (σ^2 known)	Normal	Normal

Note: there are many other options for prior distributions but further discussion of this is beyond the scope of the course. In this course we are simply introducing the ideas and considerations involved here. A much fuller treatment on types of priors and how to select priors will be provided in the course on *uncertainty assessment and Bayesian computation*.

Task 3.

Suppose that we have data $x_i, i = 1, \ldots, n$, that we have assumed have arisen from a geometric distribution with parameter λ, and we have a conjugate prior for λ which is a $\mathrm{Be}(\alpha, \beta)$ distribution.

What is the distributional form for the posterior distribution?

Summarising posterior inference

Once we have obtained our posteriors, the posterior probability distribution contains all the current information about the parameter of interest θ. For many practical purposes various numerical summaries of the distribution are desirable. Commonly used summaries of location are the mean, median and mode of the distribution. The standard deviation, inter-quartile range and other quantiles are used to summarise the variation. Typically, the posterior mean will be taken as the Bayesian point estimate of θ.

In addition to point summaries, it is nearly always important to report posterior uncertainty. Sometimes the probability that the parameter lies in a particular interval may be of interest. It would appear sensible to look for an interval in which *most of the distribution* lies. That is an interval which is such that the density at any point inside the interval is greater than the density at any point outside it. It would also appear sensible to seek an interval that is as short as possible. We shall refer to such an interval as a highest (posterior) density region or HDR.

If the posterior distribution is unimodal and symmetric (for example, the normal distribution), an HDR is a central posterior interval. This is defined to be the posterior $\alpha/2$ and $1-\alpha/2$ quantiles i.e. 2.5% to 97.5% for a 95% interval (i.e. $\alpha = 0.05$).

Other terms in common use for these intervals are Bayesian confidence intervals or **credible intervals**.

Relation to classical statistics

Frequentists cannot assign probabilities to parameters, hence a 95% confidence interval is interpreted thus:

- under repeated sampling and recalculation, 95% of confidence intervals would contain the true population value.

Bayesian confidence intervals or credible intervals have what is usually considered a more *natural* interpretation. A 95% credible interval is interpreted thus:

- there is a probability of 0.95 that the true population value lies within the credible interval.

Supplementary material:

An additional document is also available at Supplementary material - part 2 which illustrates how to derive posterior distributions using a prior and likelihood.

Answers to tasks

Answer to Task 1 (Clinical test for HIV). **Bayes' theorem**

We have Pr(D) = 0.1 and Pr(D^c) = 0.9

We have Pr($T|D$) = 0.9 and Pr($T^c|D$) = 0.1

We have Pr($T^c|D^c$) = 0.85 and Pr($T|D^c$) = 0.15

Pr(T) = Pr($T|D$)Pr(D) + Pr($T|D^c$)Pr(D^c) = 0.225 and Pr(T^c) = 0.775

So

Pr($D^c|T$) = $\frac{\Pr(T|D^c)\Pr(D^c)}{\Pr(T)}$ = 0.6

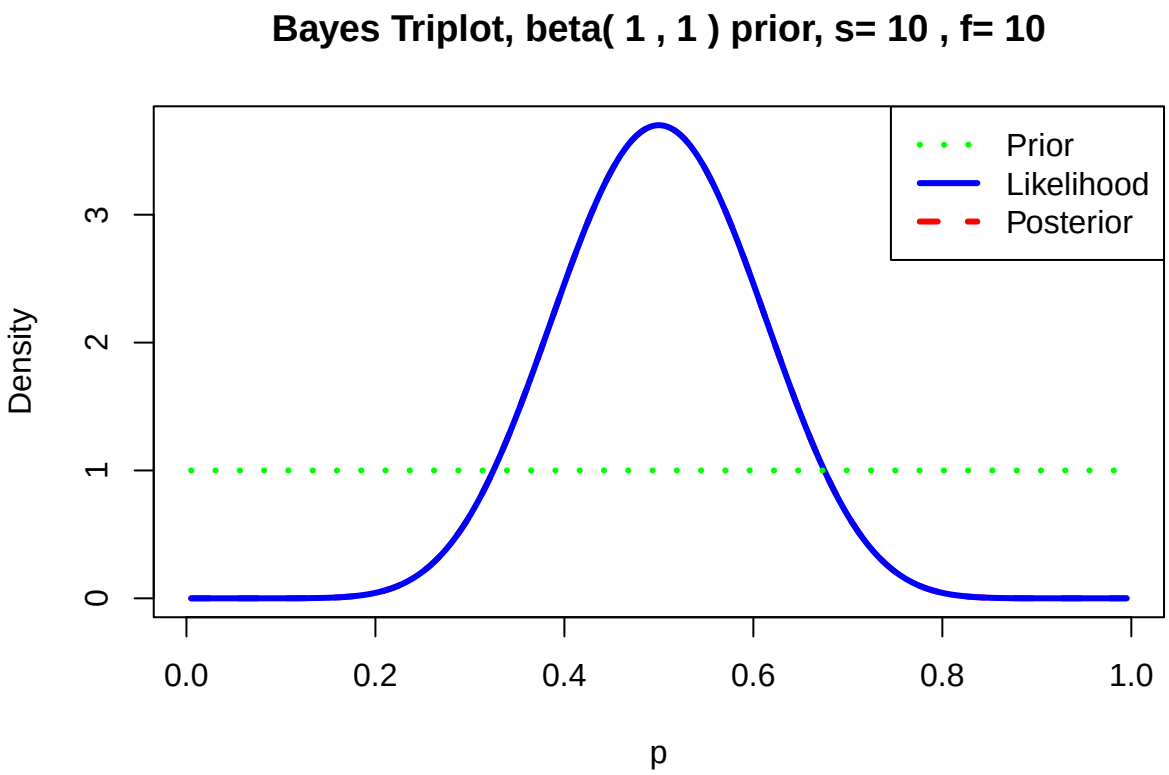
and

Pr($D|T^c$) = $\frac{\Pr(T^c|D)\Pr(D)}{\Pr(T^c)}$ = 0.013.

Answer to Task 2. **Binomial data with 10 successes and 10 failures and a Be(1, 1) prior**

In this scenario, we use a Be(1,1) prior with $n = 20$ and $x = 10$.

```
library(LearnBayes)
prior <- c(1,1)      # proportion has a Be(1, 1) prior
data <- c(10,10)       # observe 10 successes and 10 failures
triplot(prior,data)
```

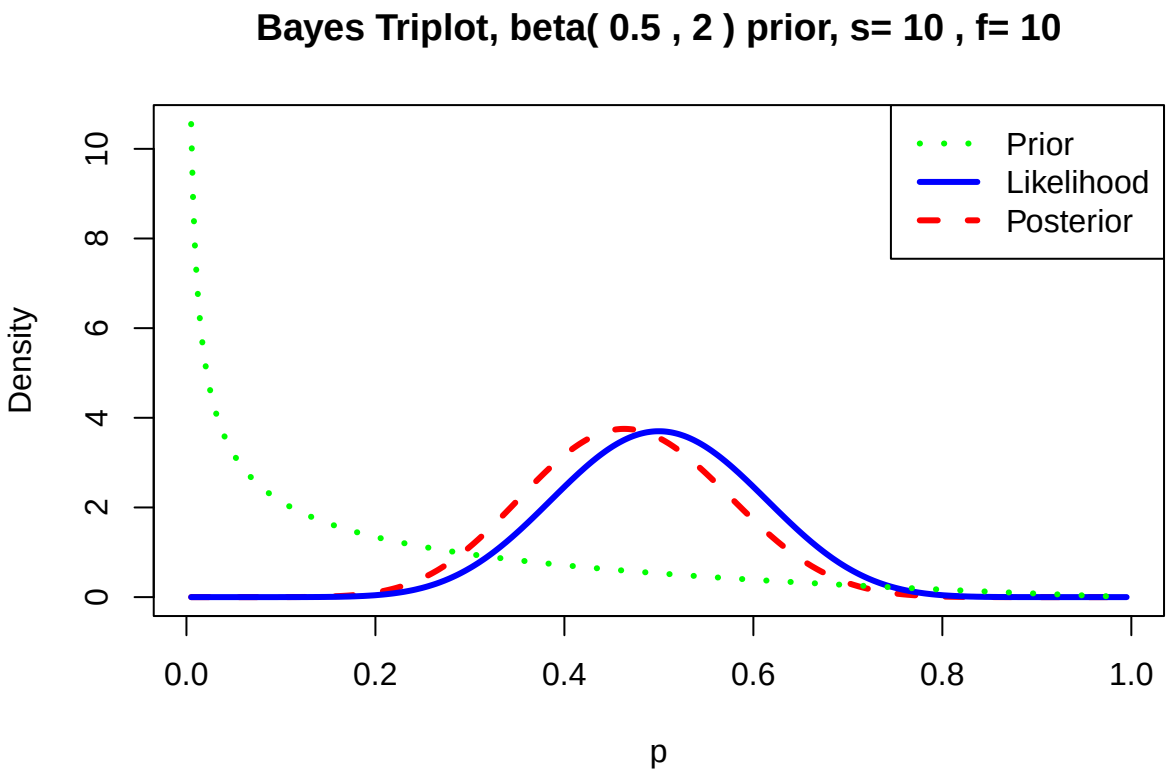


- The likelihood here provides an MLE for θ of 0.5;
- The prior is flat;
- The likelihood and the posterior are indistinguishable.

Binomial data with 10 successes and 10 failures and a Be(0.5, 2) prior

The prior is Be(0.5,2) for $n = 20$ and $x = 10$.

```
library(LearnBayes)
prior <- c(0.5,2)      # proportion has a Be(0.5, 2) prior
data <- c(10,10)       # observe 10 successes and 10 failures
triplot(prior,data)
```



- The likelihood here provides an MLE for θ of 0.5;
- The prior is strong but the data over powers;
- The prior just pulls the posterior slightly to the left of the data, but it retains the shape of the data.

Answer to Task 3. Since we have been told that the beta prior here is a conjugate prior for the geometric distribution then we know that the posterior will have the same distributional form as the prior distribution. The posterior distribution will, therefore, be a beta distribution.