2. Introduction to Bayesian inference

de Finetti

The Bayesian approach comes originally from the work of Bayes and Laplace and much of the modern theory comes from de Finetti in the 1930's.

Recommended reading

- Este artículo de José Bernardo es una buena introducción a la inferencia bayesiana.

- Lindley (1983) is a useful article to read.

Characteristics of Bayesian inference

Firstly, Bayesian inference depends directly on the subjective definition of probability.

We can all have our own probabilities for a given event:

$$P(\text{head}),\ P(\text{rain tomorrow}),\ P(\text{Mike was born in 1962}).$$

Our probabilities may be different as they are our own measures of the likelihood of given events.

Secondly, given a sample $\mathbf{x}$, and a prior distribution $p(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$, we can update our beliefs using $Bayes\ theorem$:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}) &= \frac{f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{f(\mathbf{x})} \\ &\propto f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\mathbf{x})p(\boldsymbol{\theta}) \end{aligned}$$

Bayesian inference satisfies the likelihood principle

Proof Suppose that $l(\boldsymbol{\theta}|\mathbf{x}_1) \propto l(\boldsymbol{\theta}|\mathbf{x}_2)$ and assume a prior distribution $p(\boldsymbol{\theta})$. Then

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}_1) &\propto p(\boldsymbol{\theta}) l(\boldsymbol{\theta}|\mathbf{x}_1) \\ &\propto p(\boldsymbol{\theta}) l(\boldsymbol{\theta}|\mathbf{x}_2) \\ &\propto p(\boldsymbol{\theta}|\mathbf{x}_2) \\ &= p(\boldsymbol{\theta}|\mathbf{x}_2). \end{aligned}$$

In the above, we are assuming that the prior distribution is subjective, i.e. that it is not just chosen via formal rules, e.g. always uniform. In this case, the likelihood principle can be violated. See chapter 5.

Estimation and credible intervals

For a Bayesian, estimation is treated as a decision problem. In a given situation, we should elect an estimator in order to minimize the loss that we expect to incur. *Utility theory* can be used to choose an optimal estimator. See chapter 7.

A 95% credible interval for θ is an interval $[a, b]$ such that our probability that θ lies in $[a, b]$ is 95%. See chapter 5 for a formal definition.

Nuisance parameters and prediction

There are no (theoretical) problems in dealing with with nuisance parameters.

If $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_2$ are nuisance parameters, then we can write the joint density as $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)p(\boldsymbol{\theta}_2)$ and therefore,

$$p(\boldsymbol{\theta}_1|\mathbf{x}) = \int p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{x})p(\boldsymbol{\theta}_2|\mathbf{x})\, d\boldsymbol{\theta}_2.$$

Note however that if $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{x})$ varies a lot given different values of $\boldsymbol{\theta}_2$, then this sensitivity should be taken into account. See Box and Tiao (1992), Section 1.6.

Prediction is also straightforward. If Y is a new observation, then the predictive distribution of Y is

$$f(y|\mathbf{x}) = \int f(y|(\mathbf{x}), \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}$$

Bayesian analysis of Example 9

Example 11

Suppose that our prior beliefs about θ are represented by a uniform distribution $\theta \sim \mathcal{U}(0,1)$.

The uniform distribution is an example of a beta distribution,

$$p(\phi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\phi^{\alpha-1}(1-\phi)^{\beta-1} \quad \text{for } 0 < \phi < 1$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function. Setting $\alpha = \beta = 1$ gives the uniform distribution. This is not a very realistic representation of typical prior knowledge. It would be more appropriate to use a symmetric beta distribution, e.g. $\mathcal{B}(5,5)$.

We can now calculate the posterior distribution via Bayes theorem.

Calculation of the posterior distribution

From Bayes theorem, the posterior distribution is

$$\begin{aligned} p(\theta|x) &\propto 1 \times \left(\begin{array}{c} 12 \\ 9 \end{array} \right) \theta^9(1-\theta)^3 \\ &\propto \theta^9(1-\theta)^3 \propto \theta^{10-1}(1-\theta)^{4-1} \\ &= \frac{1}{B(10,4)}\theta^{10-1}(1-\theta)^{4-1} \end{aligned}$$

which implies that $\theta|\mathbf{x} \sim \mathcal{B}(10,4)$.

It can now be demonstrated that $P(\theta \leq 1/2|\mathbf{x}) \approx .046$ and we might choose to reject the hypothesis that $\theta \leq 0.5$. Note however that this does not constitute a formal hypothesis test. These will be analyzed in chapter 7.

The posterior mean as a weighted average

From the properties of the beta distribution, we know that if $\phi \sim \mathcal{B}(\alpha, \beta)$, then $E[\phi] = \frac{\alpha}{\alpha+\beta}$.

Thus, in our case, we have

$$E[\theta|x] = \frac{10}{10+4} = \frac{5}{7} \quad \text{and moreover,}$$

$$\frac{5}{7} = \frac{1}{7} \times \frac{1}{2} + \frac{6}{7} \times \frac{9}{12}$$

which implies that

$$E[\theta|x] = \frac{1}{7}E[\theta] + \frac{6}{7}\hat{\theta}$$

where $E[\theta] = 1/(1+1) = 1/2$ is the prior mean and $\hat{\theta} = 9/12$ is the MLE of θ.

Thus, the posterior mean is a weighted average of the prior mean and the MLE.

Interpretation

One interpretation of this result is that the data have six times the weight of the prior distribution in determining the posterior distribution.

Equally, we might assume that the information represented in the prior distribution is equivalent to the information contained in an experiment where a coin is tossed twice and one head and one tail are observed. This interpretation will be formalized in the following chapter.

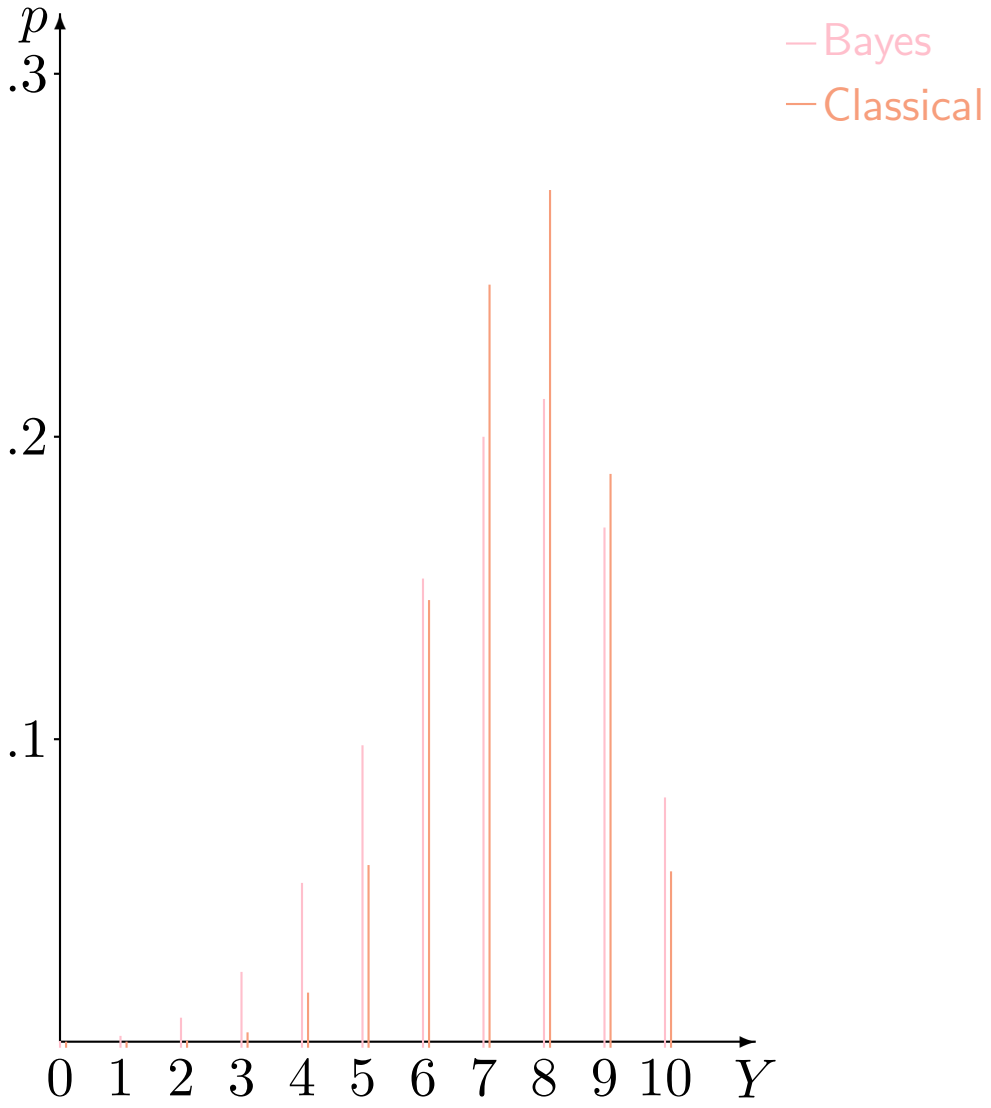
Prediction

Suppose that we wish to predict the number of heads, Y, in ten further tosses of the same coin. Thus, we have $Y|\theta \sim \mathcal{BI}(10, \theta)$ and therefore,

$$
\begin{aligned}
f(y|\mathbf{x}) &= \int f(y|\mathbf{x}, \theta) p(\theta|\mathbf{x})\, d\theta = \int f(y|\theta) p(\theta|\mathbf{x})\, d\theta \\
&= \int_0^1 \left(\begin{array}{c} 10 \\ y \end{array} \right) \theta^y (1-\theta)^{10-y} \times \frac{1}{B(10,4)} \theta^{10-1} (1-\theta)^{4-1}\, d\theta \\
&= \left(\begin{array}{c} 10 \\ y \end{array} \right) \frac{1}{B(10,4)} \times \int_0^1 \theta^{10+y-1} (1-\theta)^{14-y-1}\, d\theta \\
&= \left(\begin{array}{c} 10 \\ y \end{array} \right) \frac{B(10+y, 14-y)}{B(10,4)} \quad \text{for } y = 0, 1, \ldots, 10
\end{aligned}
$$

which is the so called *beta-binomial distribution.* The following diagram illustrates the predictive probability distribution of Y and the binomial predictive distribution ($\mathcal{BI}(10, .75)$) derived from substituting the MLE, $\hat{p} = 0.75$, for p.

The predictive distribution



The predictive mean and variance

We can calculate the predictive mean of $Y|\mathbf{x}$ without having to evaluate the whole predictive distribution. Thus, for variables Z and Y, we have

$$E[Z] = E[E[Z|Y]].$$

In our example, we have $E[Y|\theta] = 10\theta$ and $E[\theta|\mathbf{x}] = \frac{5}{7}$ and therefore

$$E[Y|\mathbf{x}] = 10 \times \frac{5}{7} \approx 7.141.$$

In order to calculate the predictive variance, we can use the formula

$$V[Z] = E[V[Z|Y]] + V[E[Z|Y]].$$

Beta prior distributions

As noted earlier, the uniform distribution is not a realistic prior distribution in many cases. In many cases we may have more knowledge that we can express in the form of a beta distribution.

Suppose that $\theta \sim \mathcal{B}(5,5)$. This prior might be used to express prior belief that the true probability is near to $1/2$.

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto \frac{1}{B(5,5)}\theta^{5-1}(1-\theta)^{5-1}\theta^{9}(1-\theta)^{3} \\ &\propto \theta^{14-1}(1-\theta)^{8-1} \\ \theta|\mathbf{x} &\sim \mathcal{B}(14,8) \end{aligned}$$

Suppose now that we think that the coin is likely to be biased in favour of heads. Then we might assume that $\theta \sim \mathcal{B}(5,1)$. In this case,

$$p(\theta|\mathbf{x}) \propto \theta^{5-1}(1-\theta)^{1-1}\theta^{9}(1-\theta)^{3}$$

and therefore, $\theta|\mathbf{x} \sim \mathcal{B}(14,4)$.

We may think the coin is biased in favour of tails. In this case, we may assume that $\theta \sim \mathcal{B}(1,5)$, when $\theta|\mathbf{x} \sim \mathcal{B}(10,8)$.

In all 4 cases we have considered, the prior and posterior distributions are beta distributions.

The beta prior distribution is $conjugate$ to the binomial sampling distribution. Similar situations will be considered in chapter 3.

The scaled likelihood function

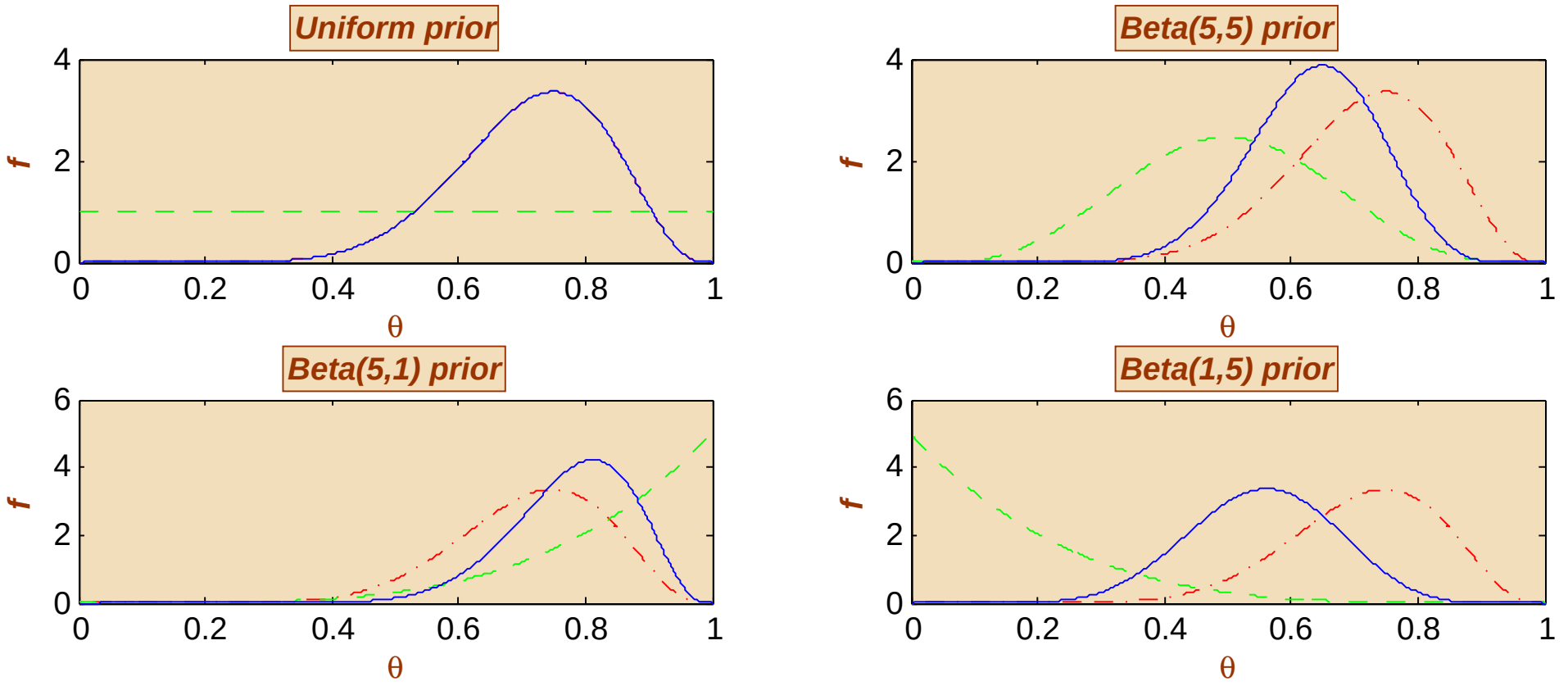
Sometimes, (if θ is one dimensional), it is possible to observe the influence of the prior distribution and the likelihood function on the posterior distribution. In order to do this, we define the scaled likelihood as

$$\frac{l(\theta|\mathbf{x})}{\int l(\theta|\mathbf{x})\ d\theta}$$

Note that the scaled likelihood does not always exist. However, when it can be defined, we can construct a diagram showing the prior, the scaled likelihood and the posterior together.

The following diagrams illustrate the effects of using the different prior distributions in this problem. In each graphic, the prior is given as a green dashed line, the scaled likelihood as a red dash-dot line and the posterior as a solid blue line.

The results of using different prior distributions



What if the prior distribution is not beta?

If we do not use a beta prior in the binomial sampling problem, then calculation of the posterior density and its properties is more complex.

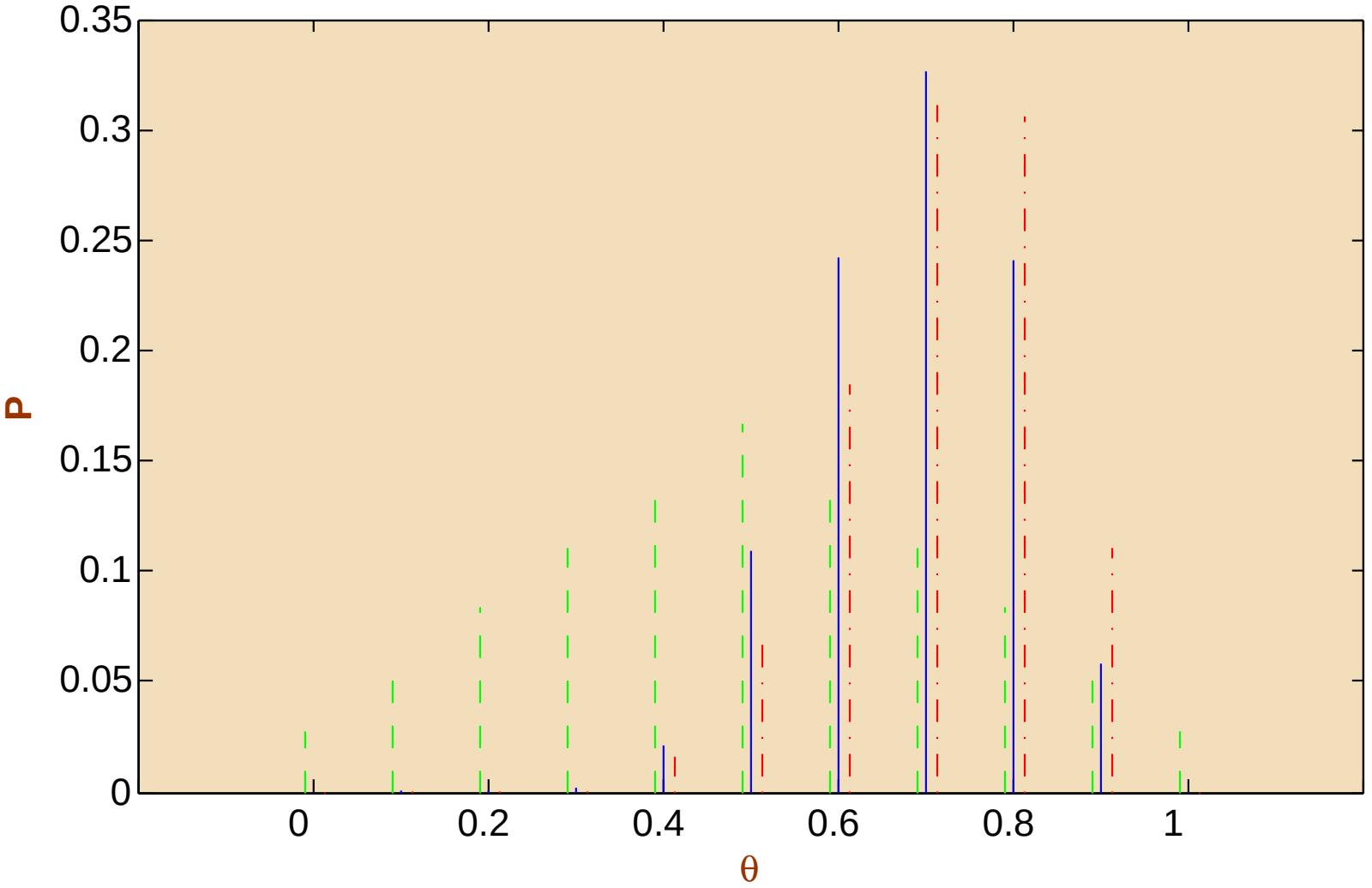
Suppose that we assume a discrete prior distribution

θ	$P(\theta)$
0	1/36
0.1	2/36
0.2	3/36
0.3	4/36
0.4	5/36
0.5	6/36
0.6	5/36
0.7	4/36
0.8	3/36
0.9	2/36
1.0	1/36

In this case, we have to use Bayes theorem to calculate all of the probabilities.

θ	$P(\theta)$	$\theta^9(1-\theta)^3$	$\frac{\theta^9(1-\theta)^3}{\sum \theta^9(1-\theta)^3}$	$P(\theta)\frac{\theta^9(1-\theta)^3}{\sum \theta^9(1-\theta)^3}$	$P(\theta\|\mathbf{x})$
0	$1/36$	0.0000	0.0000	0.0000	0.0000
0.1	$2/36$	0.0000	0.0000	0.0000	0.0000
0.2	$3/36$	0.0000	0.0001	0.0000	0.0001
0.3	$4/36$	0.0000	0.0019	0.0002	0.0020
0.4	$5/36$	0.0001	0.0162	0.0022	0.0212
0.5	$6/36$	0.0002	0.0697	0.0116	0.1097
0.6	$5/36$	0.0006	0.1841	0.0256	0.2415
0.7	$4/36$	0.0011	0.3110	0.0346	0.3263
0.8	$3/36$	0.0011	0.3065	0.0255	0.2412
0.9	$2/36$	0.0004	0.1106	0.0061	0.0580
1.0	$1/36$	0.0000	0.0000	0.0000	0.0000
Total	1	0.0035	1	0.1059	1.0000

The diagram shows the prior and posterior probabilities and scaled likelihood. The posterior mean can be calculated to be $E[\theta|\mathbf{x}] = 0.6824$.



Results with a continuous prior

Suppose that we use a continuous, non-beta prior, for example $\theta \sim \mathcal{N}\left(0.5, 0.2^2\right)$. In this case, the posterior distribution is

$$p(\theta|\mathbf{x}) \propto \exp\left(-\frac{(\theta-0.5)^2}{0.4}\right)\theta^9(1-\theta)^3$$

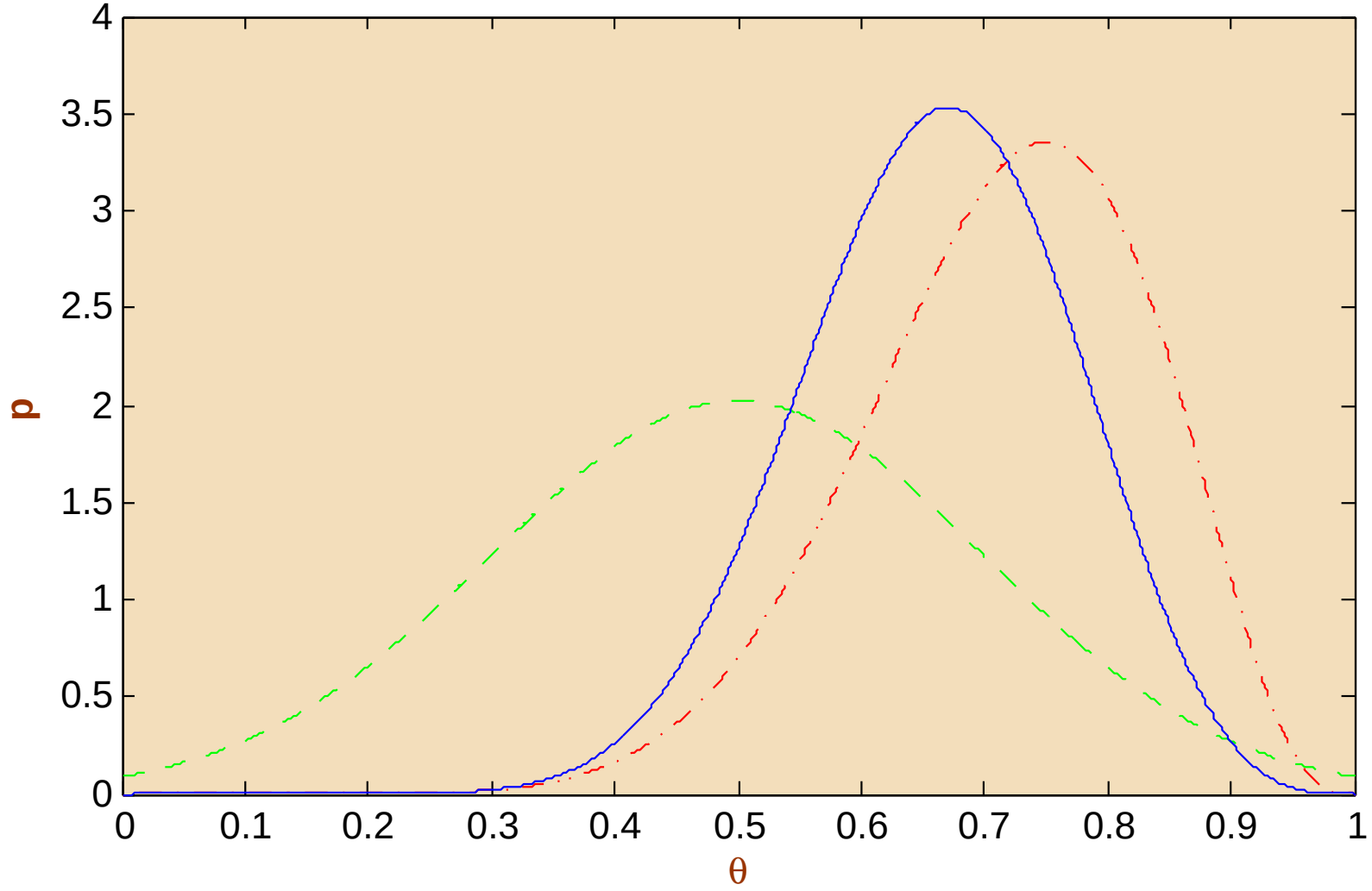
and the integration constant of this distribution must be calculated numerically.

We can do this in MATLAB by defining a function;

```
function ftheta=myfn(theta)
lik=betapdf(theta,10,4);
prior=normpdf(theta,0.5,0.2)/(normcdf(1,0.5,0.2)-normcdf(0,0.5,0.2));
ftheta=prior.*lik;
```

and using `intconst=quad(@myfn,0,1)`, which produces the result `intconst=1.1143`.

The diagram shows the prior and posterior densities and likelihood.



The posterior mean and variance etc. also have to be calculated numerically.

```
function meantheta=myfn2(theta)
lik=betapdf(theta,10,4);
prior=normpdf(theta,0.5,0.2)/(normcdf(1,0.5,0.2)-normcdf(0,0.5,0.2));
meantheta=theta.*prior.*lik;
```

Then the mean is given by `etheta=quad(@myfn2,0,1)/intconst` from which we find $E[\theta|\mathbf{x}] = 0.66$.

References

Box, G.E. and Tiao, G.C. (1992) *Bayesian Inference in Statistical Analysis* (Wiley classics library ed.). New York: Wiley.

Lindley, D.V. (1983). Theory and Practice of Bayesian Statistics. *The Statistician*, **32**, 1–11.