MS-A0503 First course in probability and statistics

5B Bayesian estimates (point and interval)

Jukka Kohonen

Deparment of mathematics and systems analysis
Aalto SCI

Academic year 2019–2020
Period III

Contents

How to obtain the posterior of Θ

For each possible value of the parameter θ,

$$\text{posterior} = \frac{\text{prior} \cdot \text{likelihood}}{\text{normalizing constant}}$$

that is,

$$f(\theta \mid x) = \frac{f(\theta) \cdot f(x \mid \theta)}{\ldots}$$

where $\ldots$ is such a constant (not depending on θ) that it makes $\sum_\theta f(\theta \mid x) = 1$.

Often convenient: Start by calculating the (for each θ) the unnormalized posterior

$$f(\theta) \cdot f(x \mid \theta).$$

- it already shows the proportions of the parameter probabilities
- (for finding the MAP estimate, unnormalized is enough)

Normalizing the unnormalized posterior

Unnormalized posterior

$$f(\theta) \cdot f(x \mid \theta)$$

is almost a probability density for Θ, except that its sum

$$c = \sum_{\theta} \Big(f(\theta) \cdot f(x \mid \theta) \Big)$$

is usually $\neq 1$. Dividing all its values by c, you get a genuine probability density (the posterior density).

Coin from the box (discrete parameter)

Unknown parameter: Θ, indicating the type of the coin we got

Prior $f(\theta) = \mathbb{P}(\Theta = \theta)$

Data $x = 0$ (one tails)

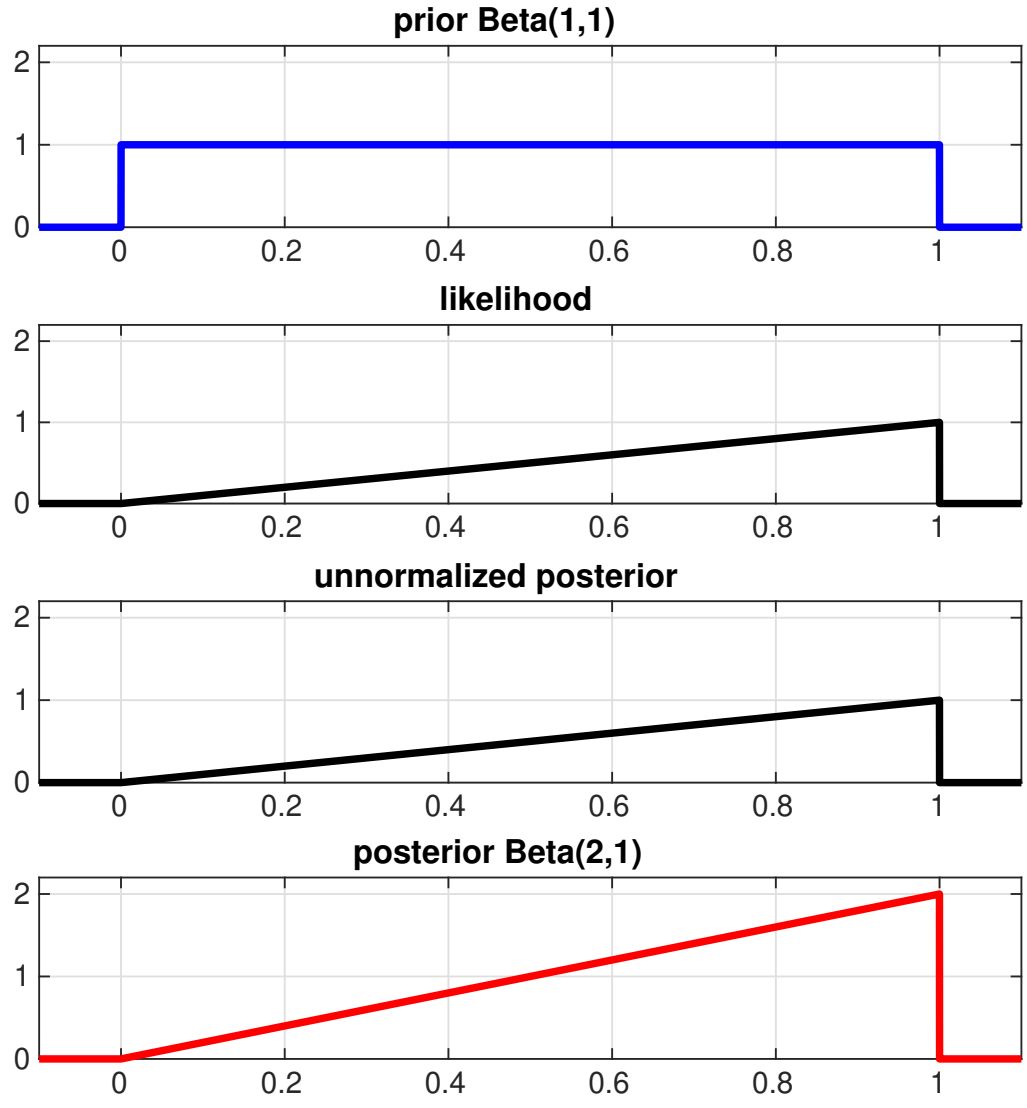
Likelihood $f(0 \mid \theta) = \mathbb{P}(X = 0 \mid \Theta = \theta) = 1 - \theta$

θ	Prior $f(\theta)$	Likelihood $f(0 \mid \theta)$	Unnormalized posterior	Posterior $f(\theta \mid 0)$
0	0.1	1.00	0.100	0.20
0.25	0.1	0.75	0.075	0.15
0.5	0.6	0.50	0.300	0.60
0.75	0.1	0.25	0.025	0.05
1	0.1	0.00	0.000	0.00
$\sum$	1.0		0.500	1.00

Unnormalized posterior already indicates which parameter values have high or low probabilities (compared to each other).

True posterior obtained by normalizing it = dividing the values by their sum (here 0.5).

Coin with continuous parameter — unif prior, first result 1



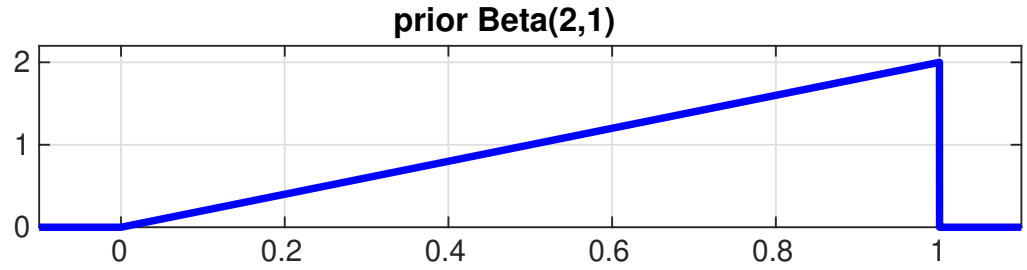
1

θ

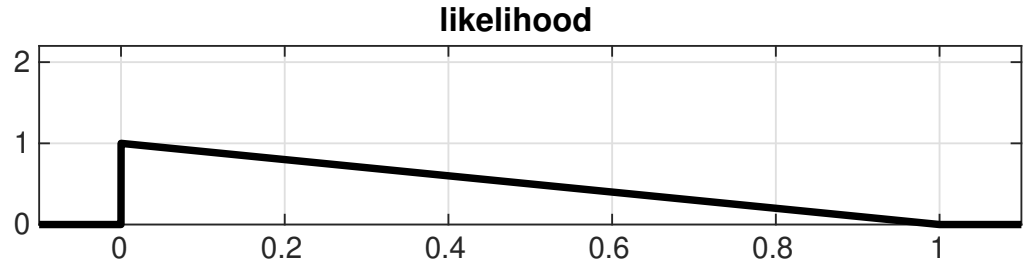
θ

2θ

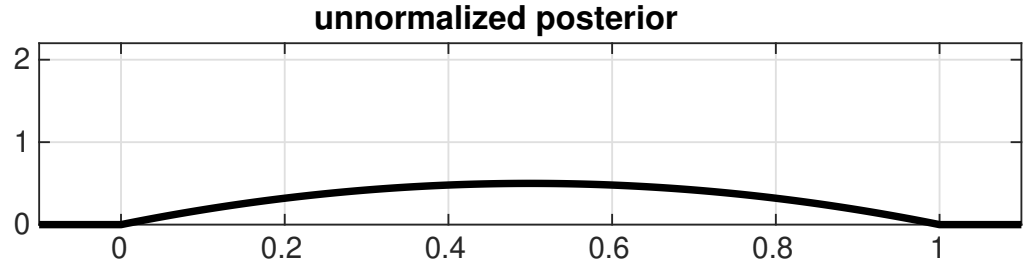
Incremental update — second result is 0



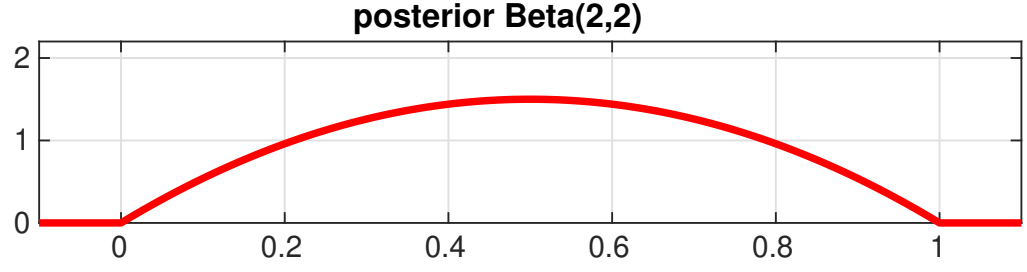
2θ



$(1-\theta)$

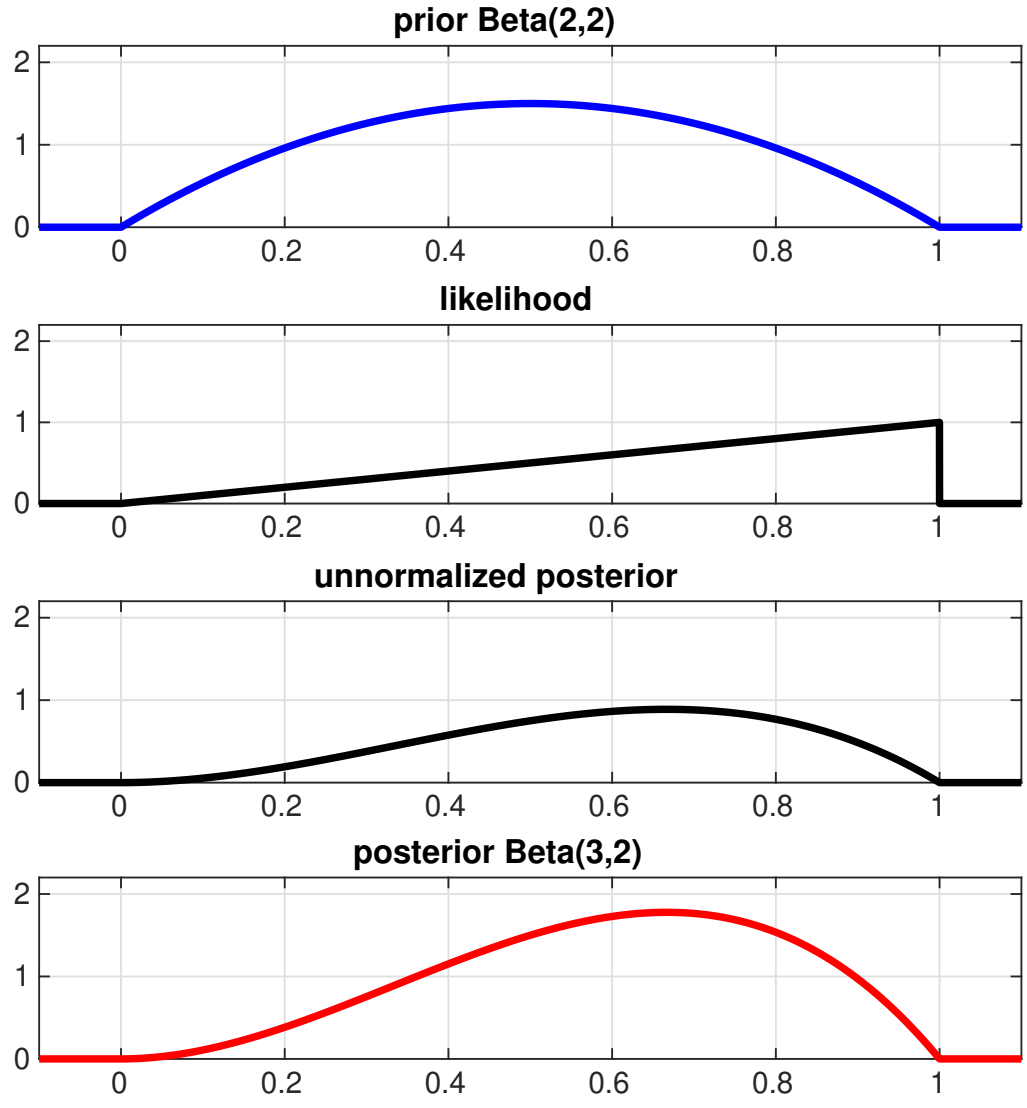


$2\theta(1-\theta)$



$6\theta(1-\theta)$

Incremental update — third result is 1



$6\theta(1-\theta)$

θ

$6\theta^2(1-\theta)$

$12\theta^2(1-\theta)$

What about the prior?

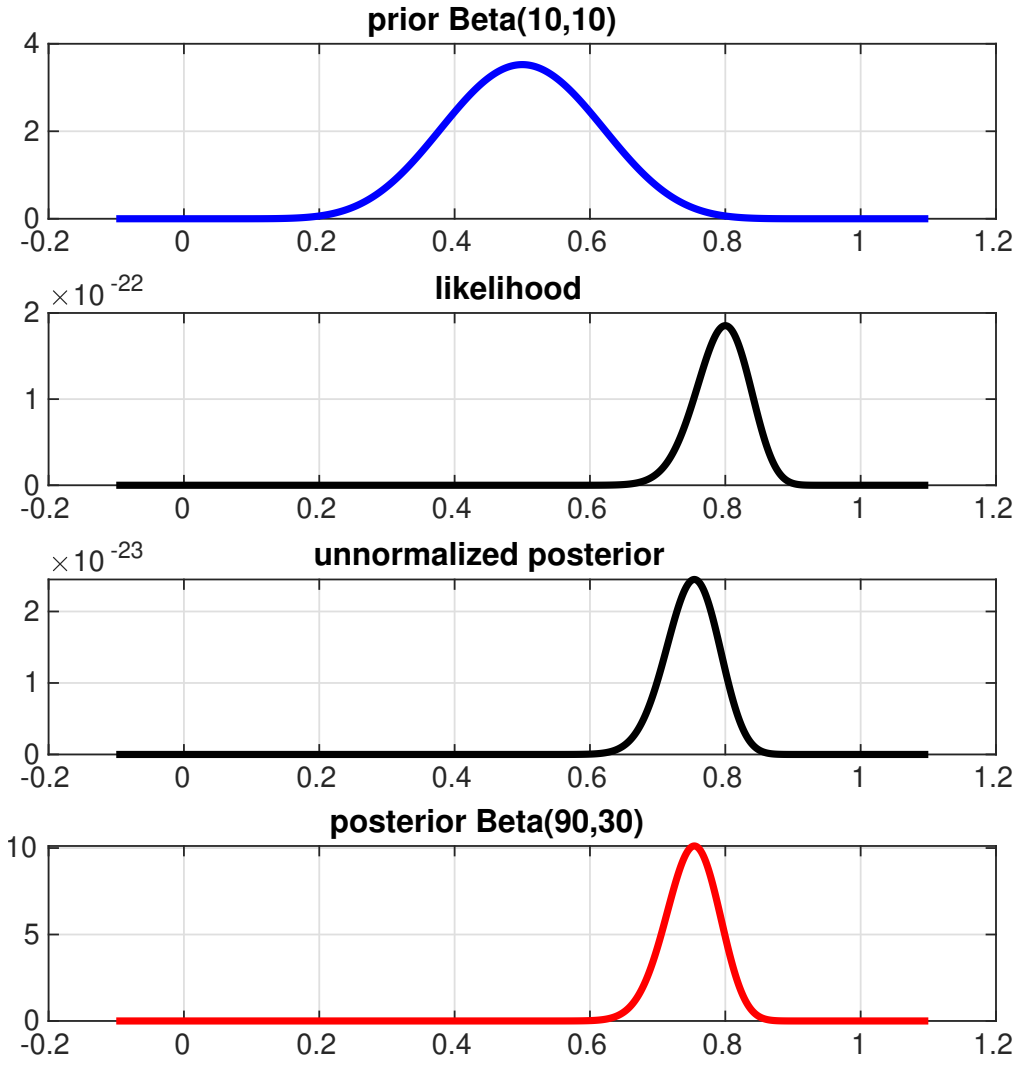
Uniform prior Beta(1,1) says that the heads probability could *a priori* be anything between 0 and 1, with equal densities. We could say this is an "unknown coin". But is this realistic for real coins?

For a real coin, we might want to incorporate a prior belief that the coin is probably approximately fair. We can do this by using a more peaked, symmetric prior such as Beta(10,10).

(Mathematically, it is as if we started with the "unknown coin" Beta(1,1), but then observed 9 heads and 9 tails.)

Let's see how it works out. We start with a symmetric "probably fair" belief, but then observe 80 heads and 20 tails.

Prior "probably fair", but asymmetric data (80+20)



$$\propto \theta^9(1-\theta)^9$$

$$\propto \theta^{80}(1-\theta)^{20}$$

$$\propto \theta^{89}(1-\theta)^{29}$$

$$\propto \theta^{89}(1-\theta)^{29}$$

Contents

How to use the posterior distribution

Congratulations, you have the posterior distribution of the unknown parameter Θ. How can you use the distribution?

Like any distribution, you can use it in many ways, depending on

1. what question you want to answer
2. what is convenient to calculate.

Some typical uses:

- mode of the posterior distribution = where it is maximized
- mean of the posterior distribution = probability-weighted average
- median of the posterior distribution = 50% probability below
- credible interval, containing e.g. 95% of posterior probability
- report/visualize the full posterior distribution
- predictions of future data, based on posterior

Next, we will look closer into each alternative.

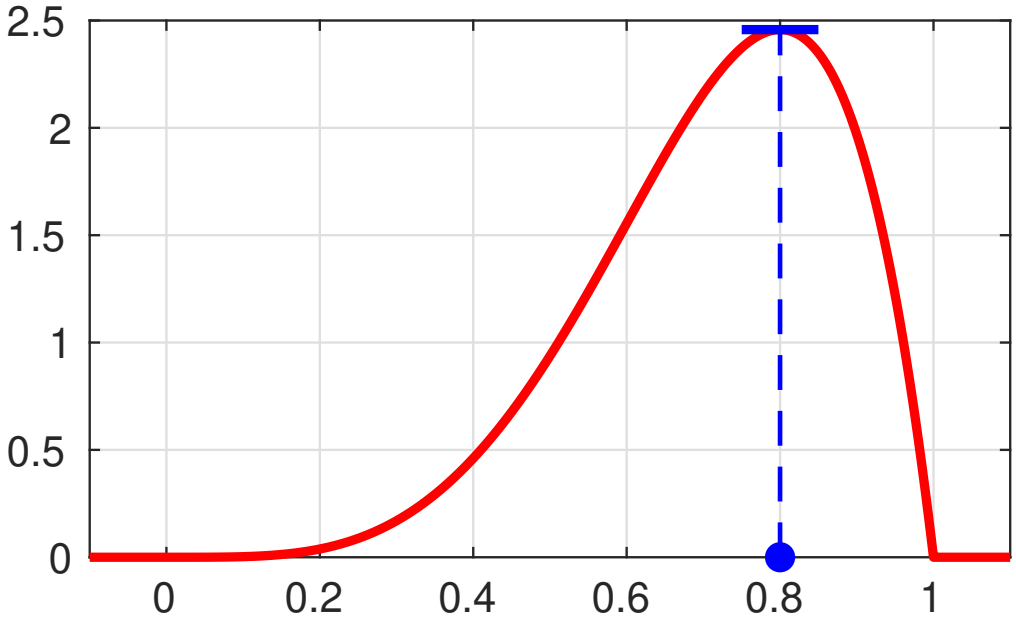
Posterior mode (MAP = **M**aximum **A** **P**osteriori estimate)

Unknown coin, uniform prior, observed 4 heads, 1 tails.
Posterior is Beta(5,2), with density (for $0 \leq \theta \leq 1$)

$$f(\theta \mid \vec{x}) = 30\theta^4(1-\theta).$$

Mode can be found by finding where derivative is zero.

(The normalizing constant 30 plays no role in the maximization, so we could as well use the unnormalized posterior. Also compare to ML estimate.)



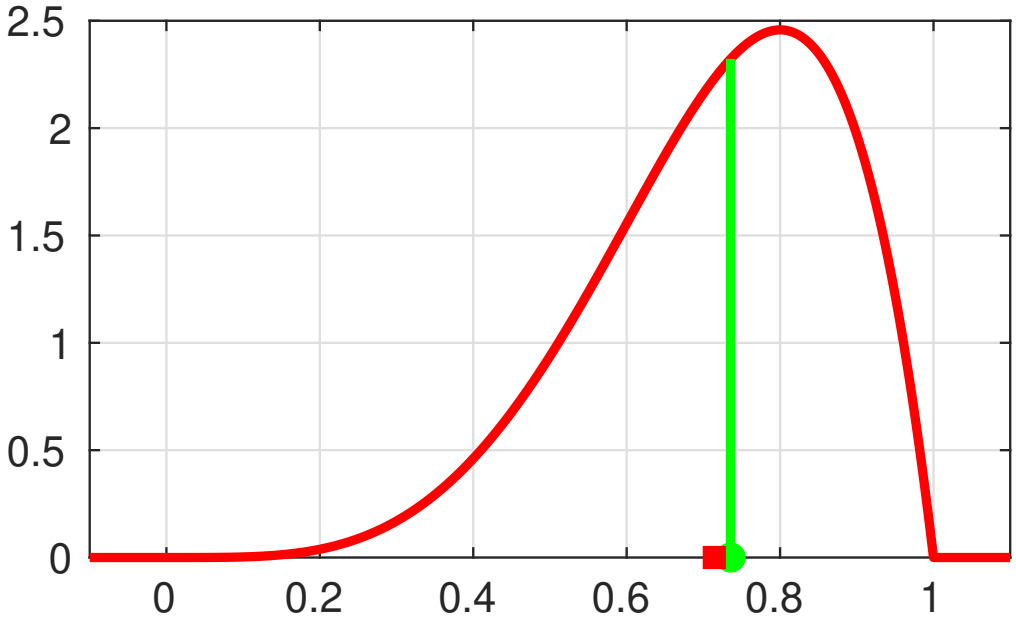
Mode = MAP estimate = 0.8

Posterior mean and median

Unknown coin, uniform prior, observed 4 heads, 1 tails.
Posterior is Beta(5,2), with density (for $0 \leq \theta \leq 1$)

$$f(\theta \mid \vec{x}) = 30\theta^4(1-\theta).$$

Mean can be found by integrating.
Median can be found by solving where CDF=0.5. (In R: `qbeta`)



Mean $= 5/7 \approx 0.7143$

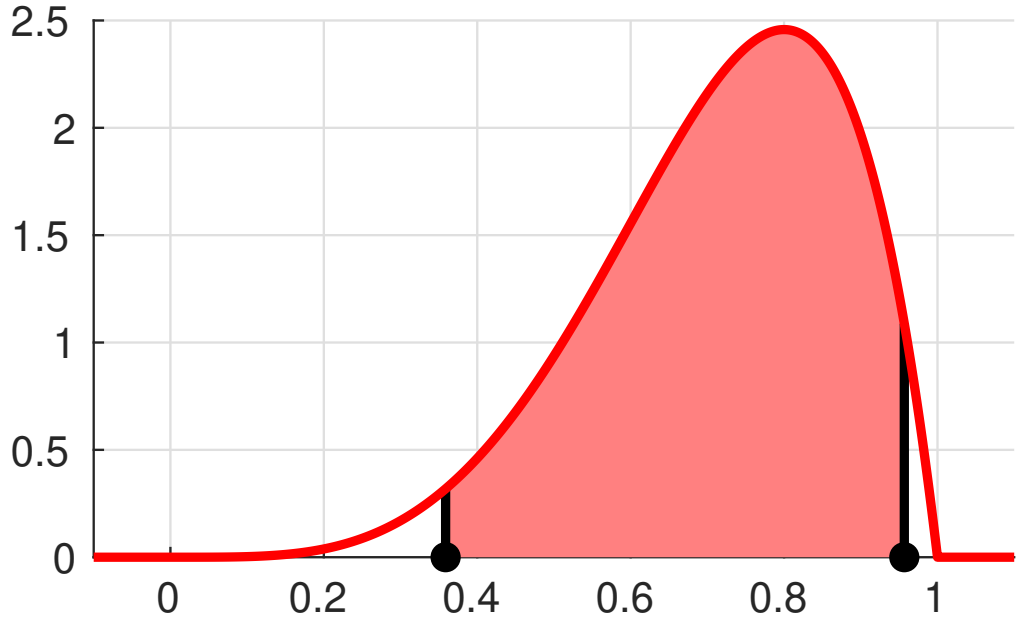
Median ≈ 0.7356

Credible interval

Unknown coin, uniform prior, observed 4 heads, 1 tails.
Posterior is Beta(5,2), with density (for $0 \leq \theta \leq 1$)

$$f(\theta \mid \vec{x}) = 30\theta^4(1-\theta).$$

Find points where CDF is 0.025 and 0.975.
$\Rightarrow \Theta$ is between those points with 95% probability.



$$\mathbb{P}(\mathbf{0.3588} \leq \Theta \leq \mathbf{0.9567} \mid \vec{X} = \vec{x}) = 95\%$$

Prediction of future data

- The posterior distribution of Θ is our best knowledge of what the parameter value can be (combining prior and data).
- Usually the posterior distribution is **not** a single point. This openly shows our uncertainty; we do not pretend that we know the parameter value exactly.
- But the more data we obtain, the more precise the posterior becomes.

Question. After seeing five observations $\vec{x} = (1, 1, 1, 1, 0)$, we have the posterior $\Theta \sim \text{Beta}(5, 2)$.
What can we say about the next observation?

Answer. We form the (posterior) predictive distribution for it, again applying basic rules of probability.

Prediction of future data (coin example)

We have five observations $\vec{x}$, and wish to predict next observation Y. From the law of total probability, we have

$$f_{Y|\vec{X}}(y \mid \vec{x}) = \int f(y \mid \theta) f(\theta \mid \vec{x}) d\theta.$$

Different values of θ give different predictions for Y. These predictions are averaged, weighted by the posterior density of Θ.

This gives our best understanding of Y, considering what we now know about Θ.

Note that we did **not** choose just one value of θ, perhaps the "most probable" one, and use that as the probability of $Y = 1$.

Prediction of future data (coin example)

We have five observations $\vec{x}$ (four heads and one tails), and wish to predict next observation Y.

- stochastic model as before: $\mathbb{P}(Y = 1 \,|\, \Theta = \theta) = \theta$.
- posterior for Θ is Beta(5,2).

So calculate:

$$\begin{aligned}
\mathbb{P}(Y = 1 \,|\, \vec{X} = \vec{x}) &= f_{Y|\vec{X}}(1 \,|\, \vec{x}) \\
&= \int f_{Y|\Theta}(1 \,|\, \theta)\; f_{\Theta|\vec{X}}(\theta \,|\, \vec{x}) d\theta \\
&= \int_0^1 \theta \cdot 30\theta^4(1-\theta) d\theta \\
&= 30 \int_0^1 (\theta^5 - \theta^6) d\theta \\
&= 30 \cdot \left(\frac{1}{6} - \frac{1}{7}\right) \approx \mathbf{0.7143}.
\end{aligned}$$

Prediction of future data — effect of uncertainty

Being honest about our uncertainty of Θ can have a huge effect on predictive distributions.

We could do this with the continuous-θ coin, but let us do a simpler **discrete** example.

Consider the following two models:

- Model A: We have a fair coin, $\Theta = 0.5$ certainly.
- Model B: We have a coin that might be unfair: Θ is either 0, 0.5 or 1, with probabilities 0.1, 0.8, 0.1 respectively.

For predicting one result, the models are equivalent. Each says that the next result is heads with 50% probability.

For predicting next 100 results, the models disagree strongly.
$\Longrightarrow$ Work out on blackboard

Contents

Multiple categories

We worked with the binary model: data were 0-1-valued (or their counts), and we had a single probability parameter, discrete or continuous.

Next we consider sequences of categorical (nominal) data that have several categories (more than two).

Examples:

- Rolls of a loaded die (3,6,6,2,6,1,3,4,6,6)
- Party stances in a sample (A,B,A,A,C,B,A,A,C,C)
- DNA sequence with four bases chosen randomly GTCTACCAG...
- Text, as a sequence of words, each word chosen randomly with some probabilities (the, quick, brown, fox, jumped, over, the, lazy, dog)

You can view the data either as a sequence of categorical variables, or as a vector of counts of the different values.

Multinomial model

- n independent observations $(X_1, X_2, \ldots, X_n)$.
- Each X_i from the same discrete distribution over k possibilies
- The distribution has k probability parameters $\vec{p} = (p_1, p_2, \ldots, p_k)$
- We can treat the probabilities as unknown, a random vector $\vec{P} = (P_1, P_2, \ldots, P_k)$

We can use the familiar methods:

- Assume a prior distribution $f(\vec{P})$
- Assume a stochastic model $f(X \mid \vec{P})$ (likelihood)
- After observations, work out posterior $f(\vec{P} \mid X)$

Stochastic model — Three-category example

A large population contains supporters of three parties A, B, C with proportions $\vec{p} = (p, q, r) = (0.5, 0.3, 0.2)$.

A random sample of $n = 10$ people is taken. Each person sampled has the probabilities $\vec{p}$ for the three parties.

Two questions:

- What kinds of (ordered) sequences are we likely to observe?
  example: AAAABBBBCC
- What kinds of count vectors are we likely to observe?
  example: $(4, 4, 2)$

For example,

- $\mathbb{P}(\text{AAAAAAAAAA}) = p^{10} \approx 0.000977$ Small
- $\mathbb{P}(\text{AAAABBBBCC}) = p^4 q^4 r^2 \approx 0.000020$ Smaller!?

Stochastic model — Three-category example

From elementary combinatorics, we know there are $3^{10} = 59049$ different 10-person strings from three letters. Let us list them, grouped by the counts of A,B,C. Recall $(p, q, r) = (0.5, 0.3, 0.2)$.

sequence	letter counts	$\mathbb{P}$(sequence)	
AAAAAAAAAA	$(10, 0, 0)$	$p^{10} = 0.000977$	} 1 sequence
...			
AAAABBBBCC	$(4, 4, 2)$	$p^4 q^4 r^2 = 0.000020$	} 3150 seq.
BBCAABBAAC	$(4, 4, 2)$	$p^4 q^4 r^2 = 0.000020$	
AABCCAABBB	$(4, 4, 2)$	$p^4 q^4 r^2 = 0.000020$	
...			
CCBBBBAAAA	$(4, 4, 2)$	$p^4 q^4 r^2 = 0.000020$	
...			
CCCCCCCCCC	$(0, 0, 10)$	$r^{10} = 0.0000001$	} 1 sequence

$$\mathbb{P}(\text{counts are 4,4,2}) = 3150 \times 0.000020 = 0.0638 \approx \mathbf{6.4\%}$$

$$\mathbb{P}(\text{counts are 10,0,0}) = 1 \times 0.000977 = \approx \mathbf{0.1\%}$$

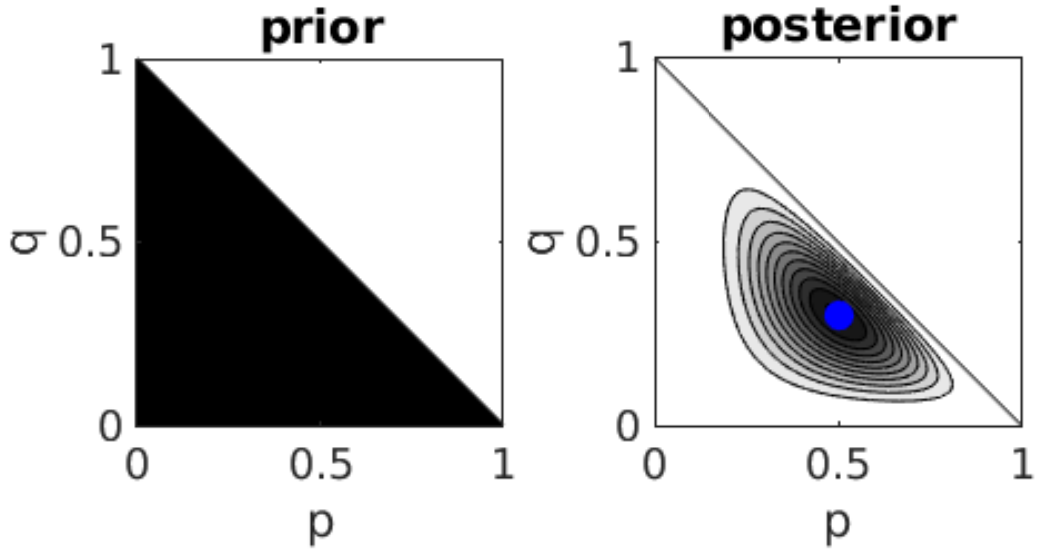
Multinomial model — Prior

- The three probabilities (p, q, r) are not three free parameters, because we must have $p + q + r = 1$.
- We can consider a two-element parameter vector (p, q), and then $r = 1 - (p + q)$.
- Furthermore, we need $p \geq 0$ and $q \geq 0$ and $p + q \leq 1$. So (p, q) is constrained to be in a triangular area. (Picture!)
- Let the prior be the uniform density over the triangle,

$$f_{P,Q}(p, q) = 2 \qquad \text{if } p, q \geq 0 \text{ and } p + q \leq 1.$$

- We now have the likelihood and the prior, so we can proceed with Bayesian inference.

Multinomial model — Inference



After observing counts $(5, 3, 2)$, the posterior density of (P, Q) is

$$f(p, q \mid \vec{x}) = c \cdot p^5 q^3 (1 - p - q)^2$$

in the triangle, and c is again normalizing constant.
We can use the posterior density to compute mode, mean, credible regions, predictions etc. Posterior mode here shown as blue dot.

Contents

Choice of prior

Sometimes people are worried about the apparent subjectivity of Bayesian inference. If you want to report a certain posterior distribution you like, you could choose your prior so that you get the posterior you wanted?

- You should be honest in making your prior to be a fairly good representation of what is known about Θ before the data.
- Uniform priors often work out nice. Not always, in complicated models.
- Beware of assigning zero density to some parameter values that might actually be true. Zero prior leads to zero posterior, whatever your data are.
- With lots of data, the effect of the prior diminishes as the "data speaks for itself".
- When reporting your results, report the model and prior you used. Then your results are completely objective: anyone using that prior will get the same posterior.

Next week: Significance tests...