MS-A0503 First course in probability and statistics

5A Bayesian inference

Jukka Kohonen

Deparment of mathematics and systems analysis
Aalto SCI

Academic year 2022–2023
Period III

Contents

Motivational example: Coin tossing

A coin, presumed fair, was tossed 5 times. We obtained data ($0 =$ tails, $1 =$ heads)

$$\vec{x} = (0, 0, 0, 0, 0).$$

The data set is small, individual data are not normal. Traditional "CLT" statistics might not be useful (or easy, or valid).

Yet it seems the data has something to say. The ML-estimator $\hat{p} = 0.0$ seems pretty low. What should we do?

- Have absolute faith that $p = 0$ exactly? (Not wise.)
- Obtain more data? (Might be difficult or impossible.)
- Disregard the result, continue believing "the coin is fair"? (Then we are not learning from the data at all.)
- Try to combine the data with our prior knowledge of the heads probability? — Then we need to define what we mean by (prior) knowledge.

Modelling knowledge and uncertainty

- We are interested in the value of an unknown quantity, for example, a parameter θ of a data-generating model.
- If we have *some* information about the quantity, model it by assigning probabilities for its possible values.
- We then say that the quantity is *a random variable* Θ.
  "Random" should be understood in a broad sense.
- $\mathbb{P}(a \leq \Theta \leq b) = 95\%$ means that based on the available information, the quantity is within $[a, b]$ with probability 95%.

The information could be from e.g.

- symmetry assumptions or measurements (homogeneous coin)
- understanding the physical process (mechanics of coin tossing)
- empirical observations (coin toss statistics)

Frequentist and Bayesian statistics

Frequentist statistics assigns probabilities only to quantities that result from a stochastic **process**, usually repeatable.

- Eg. the **results of coin tossing** X_1, X_2, X_3, X_4, X_5 and their *statistics* such as $(X_1 + X_2 + X_3 + X_4 + X_5)/5$.

Bayesian statistics extends the notion of probabilities and distributions to quantities whose **value is unknown**, even if traditionally speaking they do not result from "random processes", or are not repeatable.

- Eg. the **parameter** p that describes the coin tossing process.

(Then one usually applies Bayes' formula, hence name "Bayesian".)

Compare **Fr**equentist and **Ba**yesian approaches — Polling

In the population, the proportion of party-A supporters is p.

- **(Fr)** p is unknown, but *fixed*, so not random, no distribution, no probabilities on it.
- **(Ba)** p is unknown, so we treat it as a *random variable* that has a distribution.

Choose 1000 persons randomly. Out of them, K support party A.

- **(Fr&Ba agree)** K is a random variable, whose distribution depends on p.

Both approaches use exactly the same rules of probability. The difference is which quantities we treat as "random".

Compare **Fr**equentist and **Ba**yesian approaches — Coin

We have a coin that turns up “heads” with probability p.

- **(Fr)** p is unknown, but fixed quantity, so not random, no distribution, no probabilities on it.
- **(Ba)** p is unknown, so we treat it as a *random variable* that has a distribution.

Toss the coin 5 times and obtain K heads.

- **(Fr&Ba agree)** K is a random variable, whose distribution depends on p.

Both approaches use exactly the same rules of probability. The difference is which quantities we treat as “random”.

Contents

Example: Unknown coin

In a box, there are 10 coins: six fair (heads probability $\theta = 0.5$), two biased ($\theta = 0.25$ and 0.75) and two complete scams ($\theta = 0$ and 1). A coin is chosen from the box *at random*. Then its type Θ has distribution

θ	0	0.25	0.5	0.75	1
$\mathbb{P}(\Theta = \theta)$	0.1	0.1	0.6	0.1	0.1

E.g. the down-biased coin ($\theta = 0.25$) is chosen with probability $1/10$.

The coin we chose is tossed **once**, and we observe **tails**.
What is now the probability that we chose the down-biased coin?
Applying the law of total probability

$$\mathbb{P}(\text{tails}) \;=\; (0.1 \cdot 1) + (0.1 \cdot 0.75) + (0.6 \cdot 0.5) + (0.1 \cdot 0.25) + (0 \cdot 1) \;=\; 0.5$$

and applying Bayes's rule

$$\mathbb{P}(\Theta = 0.25 \,|\, \text{tails}) \;=\; \frac{\mathbb{P}(\Theta = 0.25) \cdot \mathbb{P}(\text{tails} \,|\, \Theta = 0.25)}{\mathbb{P}(\text{tails})} \;=\; \frac{0.1 \cdot 0.75}{0.5} \;=\; 0.15$$

Unknown coin, continued

We tossed the coin once, and observed **tails**.

For *each* possible coin type θ, compute the probability that our coin is of *that* type. (Similar calculation in each case.)

θ	0	0.25	0.5	0.75	1
$\mathbb{P}(\Theta = \theta \mid \text{tails})$	0.20	0.15	0.60	0.05	0.00

This the posterior distribution of the coin type, after one toss.

How did the one observation affect our probabilities?

- Types that often result "tails" had probabilities *increased*.
- Types that seldom result "tails" had probabilities *decreased*.
- Type $\Theta = 1$ went to probability zero. Makes sense: The coin does not result only heads, because we already saw tails.

About notation that we will use

Full notation	Shorthand	Explanation
$f_X(x)$	$f(x)$	density of X
$f_\Theta(\theta)$	$f(\theta)$	density of Θ ("prior")
$f_{X\,|\,\Theta}(x\,|\,\theta)$	$f(x\,|\,\theta)$	density of X, if $\Theta = \theta$ ("likelihood")
$f_{\Theta\,|\,X}(\theta\,|\,x)$	$f(\theta\,|\,x)$	density of Θ, if $X = x$ ("posterior")

In full notation, the subscripts indicate which random variable(s) we are talking about.

In shorthand, we drop subscripts, so f can refer to different functions. It should be understood from the argument inside the parentheses.

Note that $f(x\,|\,\theta)$ and $f(\theta\,|\,x)$ are both conditional densities (in opposite ways).

Caveat: What does $f(5\,|\,3)$ mean? Use the subscripts to clarify if needed!

Knowledge update formula — Discrete model

Knowledge about Θ before observing the data:

- Prior density $f_\Theta(\theta) \;=\; \mathbb{P}(\Theta = \theta)$

Knowledge about Θ after observing data $X = x$:

- Posterior density $f_{\Theta|X}(\theta \,|\, x) \;=\; \mathbb{P}(\Theta = \theta \,|\, X = x)$

Stochastic model of the data source:

- Likelihood function $f_{X|\Theta}(x \,|\, \theta) \;=\; \mathbb{P}(X = x \,|\, \Theta = \theta)$

Fact

The posterior density is obtained by multiplying pointwise (= for each value of θ) prior times likelihood, then

$$f(\theta \,|\, x) \;=\; \frac{f(\theta)\, f(x \,|\, \theta)}{\sum_{\theta'} f(\theta')\, f(x \,|\, \theta')}.$$

Knowledge update formula — Proof

The law of total probability (or addition law) says

$$\begin{aligned}\mathbb{P}(X = x) &= \sum_{\theta'} \mathbb{P}(\Theta = \theta')\,\mathbb{P}(X = x \,|\, \Theta = \theta') \\ &= \sum_{\theta'} f(\theta')\, f(x \,|\, \theta'),\end{aligned}$$

and then applying Bayes' formula

$$\begin{aligned}f(\theta \,|\, x) &= \mathbb{P}(\Theta = \theta \,|\, X = x) \\ &= \frac{\mathbb{P}(\Theta = \theta)\mathbb{P}(X = x \,|\, \Theta = \theta)}{\mathbb{P}(X = x)} \\ &= \frac{f(\theta) f(x \,|\, \theta)}{\mathbb{P}(X = x)} \\ &= \frac{f(\theta) f(x \,|\, \theta)}{\sum_{\theta'} f(\theta')\, f(x \,|\, \theta')}.\end{aligned}$$

Example: Unknown coin

Unknown parameter: Type Θ of the coin

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

Data $x = 0$ (result was heads)

Likelihood $f_{X\,|\,\Theta}(x \,|\, \theta) = \mathbb{P}(X = x \,|\, \Theta = \theta)$

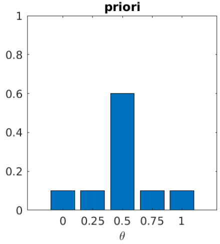
θ	Prior $f(\theta)$	Likelihood $f(0 \,\|\, \theta)$	Product	Posterior $f(\theta \,\|\, 0)$
0	0.1	1.00	0.100	0.20
0.25	0.1	0.75	0.075	0.15
0.5	0.6	0.50	0.300	0.60
0.75	0.1	0.25	0.025	0.05
1	0.1	0.00	0.000	0.00
sum	1.0		0.5	1.00

"Product" is the product of prior and likelihood; also called *unnormalized posterior*. By normalizing (so that sum=1), it becomes the posterior distribution.

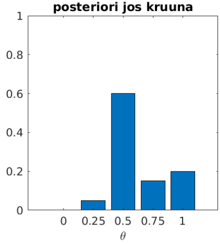
Note. The values of the likelihood function do not form any probability distribution.

Prior vs. posterior distributions in coin tossing

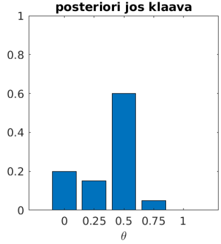
Prior



Posterior after heads



Posterior after tails



Unknown coin: Many observations

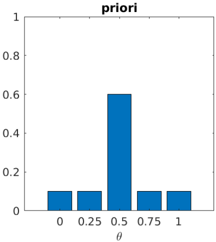
Box contains 10 coins as before. One coin chosen randomly. *That* coin tossed twice. We observe 2 tails, that is, sequence $(0, 0)$.
What is now the posterior distribution of the coin type?
Likelihood for data $(x_1, x_2) = (0, 0)$ is

θ	0	0.25	0.5	0.75	1
$f(0, 0 \mid \theta)$	1.0000	0.5625	0.2500	0.0625	0.0000

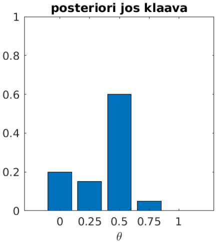
θ	Prior $f(\theta)$	Likelihood $f(0, 0 \mid \theta)$	Product	Posterior $f(\theta \mid 0, 0)$
0	0.1	1.0000	0.1000	0.32
0.25	0.1	0.5625	0.0563	0.18
0.5	0.6	0.2500	0.1500	0.48
0.75	0.1	0.0625	0.0063	0.02
1	0.1	0.0000	0.0000	0.00

"Product" is the product of prior and likelihood; after normalization (so that sum=1), it becomes the posterior distribution.

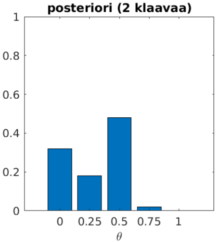
Prior vs. various posteriors



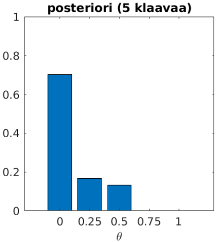
$f(\theta)$



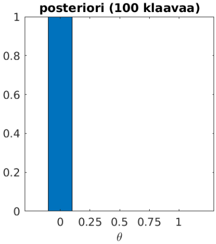
$f(\theta\,|\,0)$



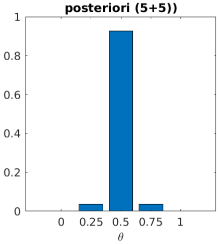
$f(\theta\,|\,00)$



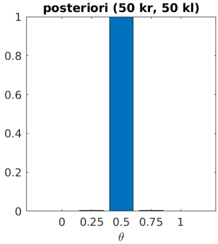
$f(\theta\,|\,00000)$



$f(\theta\,|\,0000\ldots)$



$f(\theta\,|\,0000011111)$



$f(\theta\,|\,00\ldots11\ldots)$

Variation: Updating in two phases

Knowledge of Θ before any data:

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

Knowledge of Θ after seeing some data:

- Posterior $f(\theta \,|\, x_1) \;=\; \mathbb{P}(\Theta = \theta \,|\, X_1 = x_1)$.
- Posterior $f(\theta \,|\, x_1, x_2) \;=\; \mathbb{P}(\Theta = \theta \,|\, X_1 = x_1, X_2 = x_2)$.

Stochastic model of the data source:

- Likelihood $f(x_1 \,|\, \theta) \;=\; \mathbb{P}(X_1 = x_1 \,|\, \Theta = \theta)$
- Likelihood $f(x_2 \,|\, \theta, x_1) \;=\; \mathbb{P}(X_2 = x_2 \,|\, \Theta = \theta, X_1 = x_1)$

Fact

The posterior $f(\theta \,|\, x_1, x_2)$ can be obtained by taking $f(\theta \,|\, x_1)$ as the prior, multiplying by the likelihood $f(x_2 \,|\, \theta, x_1$, and normalizing:

$$f(\theta \,|\, x_1, x_2) \;=\; \frac{f(\theta \,|\, x_1)\, f(x_2 \,|\, \theta, x_1)}{\sum_{\theta'} f(\theta' \,|\, x_1)\, f(x_2 \,|\, \theta', x_1)}.$$

Updating in two phases — Proof

If $D_1 = \{X_1 = x_1\}$ and $D_2 = \{X_2 = x_2\}$, applying the product rule

$$\begin{aligned} \mathbb{P}(\Theta = \theta, D_1, D_2) &= \mathbb{P}(D_1)\mathbb{P}(\Theta = \theta \mid D_1)\mathbb{P}(D_2 \mid \Theta = \theta, D_1) \\ &= \mathbb{P}(D_1)\, f(\theta \mid x_1)\, f(x_2 \mid \theta, x_1), \end{aligned}$$

and the law of total probability

$$\begin{aligned} \mathbb{P}(D_1, D_2) &= \sum_{\theta'} \mathbb{P}(\Theta = \theta', D_1, D_2) \\ &= \sum_{\theta'} \mathbb{P}(D_1)\, f(\theta' \mid x_1)\, f(x_2 \mid \theta', x_1), \end{aligned}$$

and combining them,

$$\begin{aligned} f(\theta \mid x_1, x_2) &= \frac{\mathbb{P}(\Theta = \theta, X_1 = x_1, X_2 = x_2)}{\mathbb{P}(X_1 = x_1, X_2 = x_2)} \\ &= \frac{\mathbb{P}(D_1)\, f(\theta \mid x_1)\, f(x_2 \mid \theta, x_1)}{\sum_{\theta'} \mathbb{P}(D_1)\, f(\theta' \mid x_1)\, f(x_2 \mid \theta', x_1)} \\ &= \frac{f(\theta \mid x_1)\, f(x_2 \mid \theta, x_1)}{\sum_{\theta'} f(\theta' \mid x_1)\, f(x_2 \mid \theta', x_1)}. \end{aligned}$$

Updating knowledge — Summary

- Prior distribution $f_\Theta(\theta)$ models the knowledge of the unknown parameter Θ, before observations
- Likelihood function $f(x \mid \theta)$ models the stochastic model — how the data is generated
- Posterior distribution models the knowledge obtained by combining prior and data
- Posterior distribution $f(\theta \mid x)$ computed by multiplying prior and likelihood, and then normalizing (so it becomes a distribution)
- This approach is called Bayesian inference

Contents

Unknown coin

We tossed an unknown coin, and observed data $\vec{x} = (0,0,0,0,0,0,1,0,1,0)$, where 0=tails, 1=heads. We assume the coin has a parameter Θ (heads probability), but have no prior reasons to think some values are more probable than others. Find the posterior distribution of Θ.

Let the prior distribution be uniform over interval $[0,1]$, with density

$$f(\theta) \;=\; \begin{cases} 1, & \theta \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$$

Likelihood $f(\vec{x}\,|\,\theta) \;=\; \theta^2(1-\theta)^8$
Posterior density

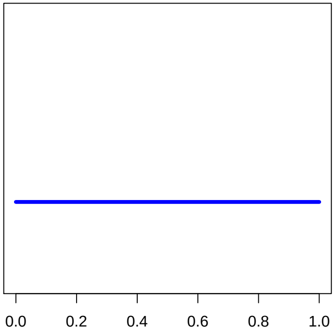
$$f(\theta\,|\,\vec{x}) \;=\; c\, f(\theta) f(\vec{x}\,|\,\theta) \;=\; \begin{cases} c\,\theta^2(1-\theta)^8, & \theta \in [0,1], \\ 0, & \text{otherwise,} \end{cases}$$

where the normalizing constant is $c = \left(\int_0^1 t^2(1-t)^8\,dt\right)^{-1}$

Unknown coin

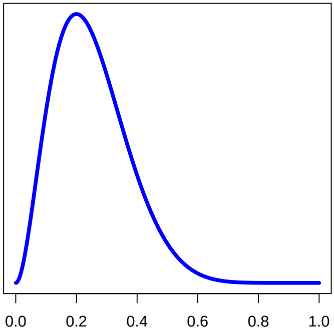
Data: $\vec{x} = (0, 0, 0, 0, 0, 0, 1, 0, 1, 0)$

Prior



$f(\theta) = 1$

Posterior



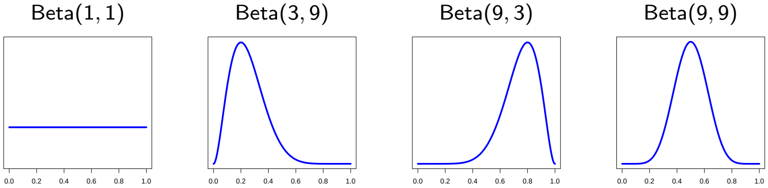
$f(\theta|\vec{x}) = c\,\theta^2(1-\theta)^8$

Beta distribution

The distribution $\text{Beta}(a, b)$, called the *beta distribution with parameters $a > 0$ and $b > 0$*, has density

$$f(\theta) = \begin{cases} c\,\theta^{a-1}(1-\theta)^{b-1}, & \text{when } \theta \in [0,1], \\ 0, & \text{otherwise,} \end{cases}$$

where the normalizing constant $c = \frac{(a+b-1)!}{(a-1)!(b-1)!}$.



- Possible values are in the interval $[0, 1]$
- Expected value $\mu = \frac{a}{a+b}$ and standard deviation $\sigma = \sqrt{\frac{\mu(1-\mu)}{a+b+1}}$

R: `dbeta(theta,a,b); pbeta(theta,a,b)`

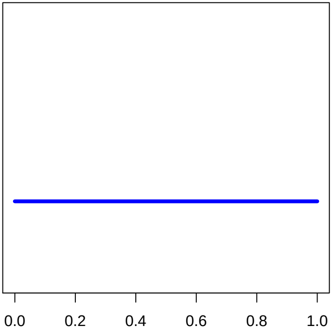
Matlab: `betapdf(theta,a,b); betacdf(theta,a,b)`

Unknown coin

Data: $\vec{x} = (0, 0, 0, 0, 0, 0, 1, 0, 1, 0)$

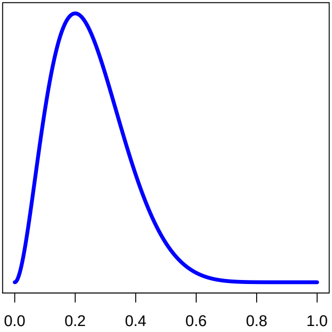
Prior: Uniform, that is Beta$(1, 1)$
Posterior: Beta$(3, 9)$

Prior



$f(\theta) = 1$

Posterior



$f(\theta|x) = c\,\theta^2(1-\theta)^8$

Variation: Posterior from number of heads

Coin tossed n times, observed x heads. Prior assumed uniform. Find posterior density of Θ (heads probability).

Prior density: $f(\theta) = 1$, $\theta \in [0, 1]$
Likelihood for x is according to $\text{Bin}(n, \theta)$

$$f(x \mid \theta) \;=\; \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

This leads to posterior density

$$f(\theta \mid x) \;=\; \frac{f(\theta) f(x \mid \theta)}{\int f(t) f(x \mid \theta')\, d\theta'} \;=\; c\, \theta^x (1-\theta)^y$$

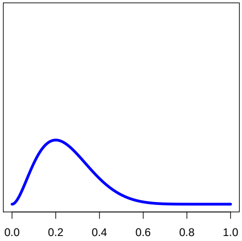
which is $\text{Beta}(x+1, y+1)$, where $y = n - x$ is number of tails.

Note

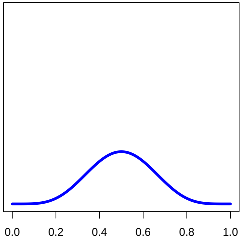
- From $n = 10$ and $x = 2$, we get exactly the same posterior $\text{Beta}(3, 9)$, as if we used the detailed data vector $\vec{x} = (0, 0, 0, 0, 0, 0, 1, 0, 1, 0)$.
- The normalizing constant c is determined from $\int_0^1 f(\theta \mid x) d\theta = 1$. It can be shown that $c = \frac{(x+y+1)!}{x!y!}$

Unknown coin: Posteriors after x heads, y tails

$n = 10$

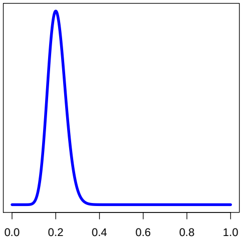


$x = 2,\ y = 8 \Rightarrow \text{Beta}(3, 9)$

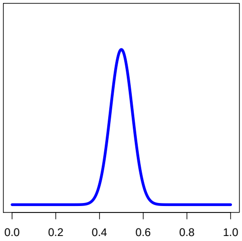


$x = 5,\ y = 5 \Rightarrow \text{Beta}(6, 6)$

$n = 100$



$x = 20,\ y = 80 \Rightarrow \text{Beta}(21, 81)$



$x = 50,\ y = 50 \Rightarrow \text{Beta}(51, 51)$

Contents

Calculating the posterior in nutshell

For each possible value of the parameter θ,

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

or

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

where ... on vakio (same for all θ), it is needed to so that $\sum_\theta f(\theta \mid x) = 1$.

Easiest (usually): First calculate the *unnormalized posterior*

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

- It is not a probability distribution, but ...
- it already shows in *relative* terms which θ are more probable,
- to find its maximum point, normalizing is not needed,
- to get a genuine probability distribution, just normalize it.

Normalizing the posterior

The unnormalized posterior

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

is *almost* a distribution for Θ, except that its sum

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

could be $\neq 1$. So just divide its values by c, $\Rightarrow$ you get the genuine posterior distribution for Θ.

If Θ is continuous, replace sums with integrals.

Coin from box (discrete parameter)

Parameter Θ indicates what kind of coin we have

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

Data $x = 0$ (one heads)

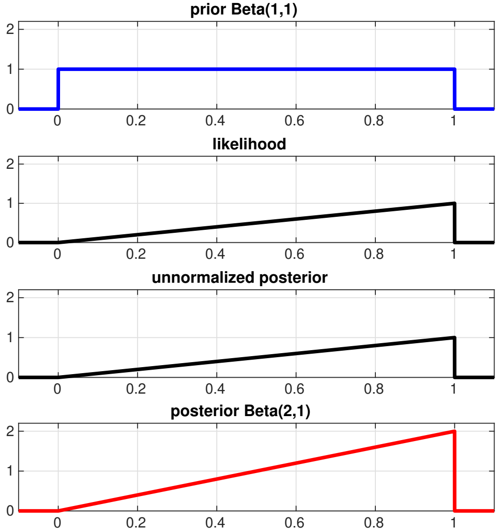
Likelihood $f(0 \mid \theta) = \mathbb{P}(X = 0 \mid \Theta = \theta) = 1 - \theta$

θ	Prior $f(\theta)$	Likelihood $f(0 \mid \theta)$	Product	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

Already from the unnormalized posterior we see (in relative terms) which parameter values are more probable than others. other).

Genuine posterior obtained by dividing by the sum (here 0.5).

Coin, continuous parameter — unif prior, first result 1



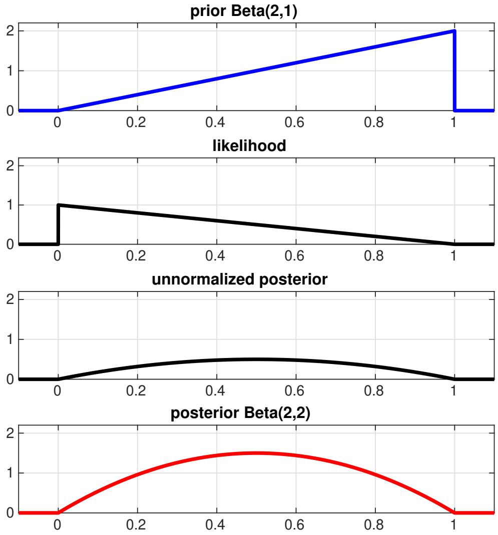
$$1$$

$$\theta$$

$$\theta$$

$$2\theta$$

Update with second observation which is 0



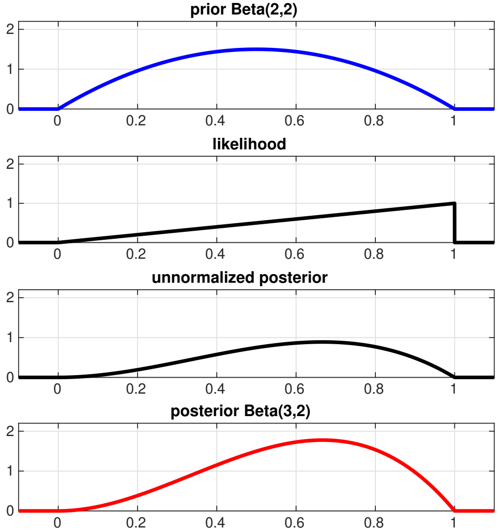
$$2\theta$$

$$(1-\theta)$$

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

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

Update with third observation which is 1



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

$$\theta$$

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

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

What about the prior?

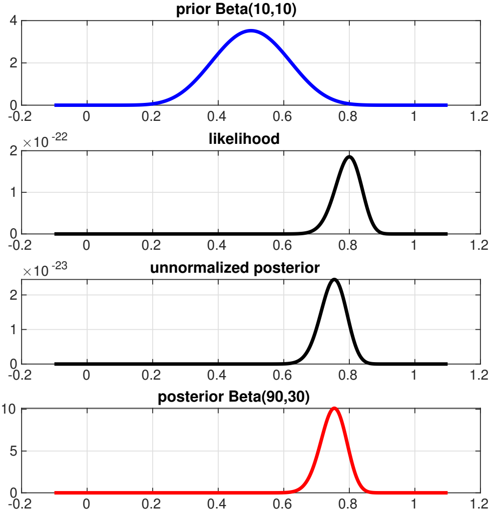
The uniform prior Beta(1,1) says that (before observing anything) we thought all parameter values within $[0, 1]$ equally probable. This is one way of saying “unknown kind of coin”.

In the real world, when we encounter a coin, we might already have prior beliefs that they are probably approximately fair. How could we state such prior belief?

What if we make our prior Beta(10,10), which is also symmetric like Beta(1,1) but more concentrated?

Let’s try. Start with “probably approximately fair” and then observe 80 heads, 20 tails.

Prior “prob. app. fair” + Observations 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

Example. Noisy measurement

- Star (A) has absolute brightness θ
- Scientist (B) measures distorted brightness $X \sim \mathrm{N}(\theta, \sigma^2)$, with $\sigma = 2$ known

A sends light with same brightness θ, measured three times. B measures values $\vec{x} = (3, 8, 7)$ and tries to estimate θ.

- The ML-estimate is simply the average $m(\vec{x}) = (3 + 8 + 7)/3 = 6$

However, B has prior knowledge that the brightness Θ has normal distribution with parameters $\mu_0 = 5$ and $\sigma_0 = 1$.

Bayesian normal model

Prior: $\Theta \sim \mathrm{N}(\mu_0, {\sigma_0}^2)$,

$$f(\theta) = (2\pi\sigma_0^2)^{-1/2} e^{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}}$$

Likelihood: $(X_i \,|\, \theta) \sim \mathrm{N}(\theta, \sigma^2)$,

$$\begin{aligned} f(x_i \,|\, \theta) &= (2\pi\sigma^2)^{-1/2} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}} \\ f(x_1, \ldots, x_n \,|\, \theta) &= f(x_1 \,|\, \theta) \cdots f(x_n \,|\, \theta) \end{aligned}$$

Example. Noisy channel:

- Prior for brightness: $\Theta \sim \mathrm{N}(\mu_0, \sigma_0^2)$, $\mu_0 = 5$, $\sigma_0 = 1$
- Measured brightness: $(X_i \,|\, \theta) \sim \mathrm{N}(\theta, \sigma^2)$, $\sigma = 2$

Bayesian normal model: Posterior distribution

Prior: $\Theta \sim \mathrm{N}(\mu_0, \sigma_0^2)$
Likelihood: $(X_i \mid \theta) \sim \mathrm{N}(\theta, \sigma^2)$

Fact

In the Bayesian normal model, with data $\vec{x} = (x_1, \ldots, x_n)$, the posterior is also normal $\mathrm{N}(\mu_1, \sigma_1^2)$, where

$$\mu_1 = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2} m(\vec{x})}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \qquad \sigma_1 = \frac{1}{\sqrt{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}},$$

and $m(\vec{x}) = \frac{1}{n}\sum_{i=1}^n x_i$ is the average of the observed data.

Example: Noisy measurement

Brightness measured by B has normal distribution, mean θ, standard deviation $\sigma = 2$.
Furthermore, we have prior information that the brightness Θ is normal with mean $\mu_0 = 5$, standard deviation $\sigma_0 = 1$.

Then the posterior for brightness, after observing three data points $\vec{x} = (3, 8, 7)$, is $\mathrm{N}(\mu_1, \sigma_1^2)$, where

$$\mu_1 = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}m(\vec{x})}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} = \frac{\frac{1}{1^2} \times 5 + \frac{3}{2^2} \times 6}{\frac{1}{1^2} + \frac{3}{2^2}} \approx 5.43$$

and

$$\sigma_1 = \frac{1}{\sqrt{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}} = \frac{1}{\sqrt{\frac{1}{1^2} + \frac{3}{2^2}}} \approx 0.756$$

Noisy measurement: Point and interval estimates

After data $\vec{x} = (3, 8, 7)$, the posterior distribution for Θ is $\mathrm{N}(\mu_1, \sigma_1^2)$, where $\mu_1 = 5.43$ ja $\sigma_1 = 0.756$.
Since we have a genuine *distribution* for Θ, we can find for example

- Posterior mean: $\mu_1 = 5.43$ (mean of the distribution)
- Posterior mode: $\mu_1 = 5.43$ (maximum point of the distribution)

Task. Find an interval that contains the star brightness, with probability 90%.
Solution Solve c from equation

$$0.90 \; = \; \mathbb{P}(\Theta = \mu_1 \pm c) \; = \; \mathbb{P}\left(\frac{\Theta - \mu_1}{\sigma_1} = 0 \pm c/\sigma_1\right) \; = \; \mathbb{P}(|Z| \leq c/\sigma_1)$$

From tables: $\mathbb{P}(|Z| \leq 1.64) = 0.90$, thus $c = 1.64 \times 0.756 = 1.24$.
Interval $5.43 \pm 1.24 = [4.19, 6.67]$ **contains the brightness with probability 90%.**

Contents

Some benefits of the Bayesian approach

If you are willing to treat Θ as a random variable (and assign a prior distribution to it), you gain:

- **Mathematical unification.** "Parameters" and "observations" are unified as "quantities" that follow the same mathematical laws of probability.
- **General applicability.** With the same framework, you can calculate posteriors
  - for small data, even $n = 1$, where e.g. "normal approximations" would not apply at all
  - for big data
  - for non-normal data, e.g. exponential observations
  - for more complicated models
- **Full posterior distribution** of Θ. It gives you a richer understanding of Θ's possible values than just a single point estimate or interval estimate. You can inspect it visually, and ask and answer any questions like mean, mode, median, probability of this interval ...

Next lecture: How to use the Bayesian posterior distributions for various tasks