Basic Bayesian Ideas

Petter Mostad

Bayesian paradigm

- Probability distributions are used to model uncertain *knowledge*
- The goal of statistics is to investigate how this model changes when new information (”data”) is received
- Examples:
  - Is a hypothesis true or not? Hypotheses that are initially unlikely needs more evidence
  - Knowledge about a value before and after it has been measured

Updating knowledge

- Knowledge is updated when new information limits the possible ”states of the world”
- The prior probability is rescaled to a smaller set.
- Often: We formulate our model so that θ represents what we want to know about, and y the ”data” we will observe. Knowledge is represented as a joint probability distribution on both.
- The use of Bayes formula is a consequence of this setup.

Bayes formula: Bayesian modelling

- Bayes formula: $$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)}$$
  - θ : The parameter of interest (may have two possible values, many possible values, or be a continuous variable with one or more dimensions)
  - y : The observed data
  - $\pi(y|\theta)$: The probability of data y for a value of θ (likelihood)
  - $\pi(\theta)$: The prior, initial distribution for θ.
  - $\pi(\theta|y)$: The posterior distribution for θ, given the data
  - $\pi(y)$: The total probability for the given data y.
- Example, when θ is either 0 or 1:

$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y \mid \theta = 1)\pi(\theta = 1) + \pi(y \mid \theta = 0)\pi(\theta = 0)}$$

Example: Hemofiliac disease

- $\Theta = 1$ or 0: A mother (M) is a carrier of hemofiliac disease or not.
- Known: Mother of M is carrier, husband of M is healthy, son of M is healthy.
- Bayesian formulation:
  - Prior, given that mother of M is carrier: $p(\theta = 1) = 0.5$
  - Likelihood: $p(\text{healthy son} \mid \theta = 1) = 0.5$, $p(\text{healthy son} \mid \theta = 0) = 1$.
  - Find posterior for θ
  - Find predictions of health for second child.

Example with two possible values for θ

- Assume about half of all people have trait A, but it is hard to observe directly. We know that 99% of all people with trait A also have trait B, whereas only 90% of all people without trait A have trait B. Given a person with trait B, what is the probability he has trait A?
- We code the information as
  - $\pi(A=1) = \pi(A=0) = 0.5$
  - $\pi(B=1|A=1) = 0.99$
  - $\pi(B=1|A=0) = 0.90$
- Solution:

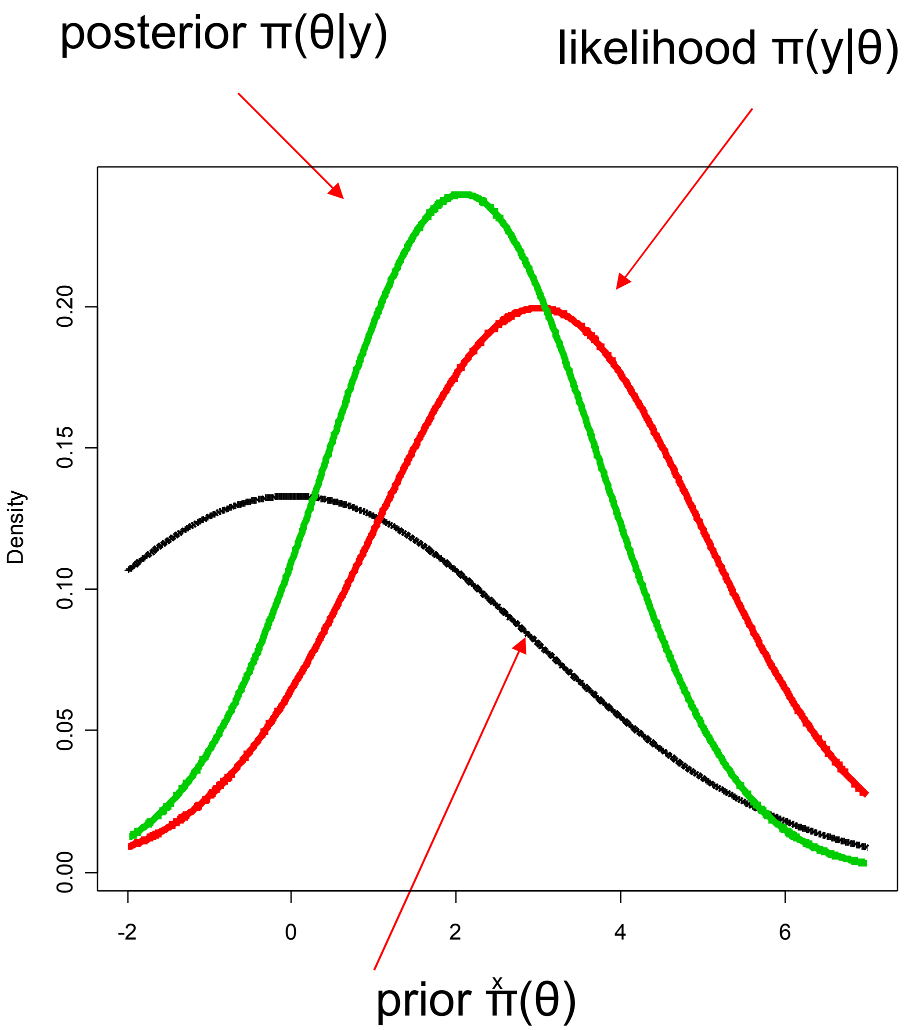
$$\pi(A=1 \mid B=1) = \frac{\pi(B=1 \mid A=1)\pi(A=1)}{\pi(B=1 \mid A=1)\pi(A=1) + \pi(B=1 \mid A=0)\pi(A=0)} = \frac{0.99 \cdot 0.5}{0.99 \cdot 0.5 + 0.90 \cdot 0.5} = 0.52$$

Bayes formula with continuous θ

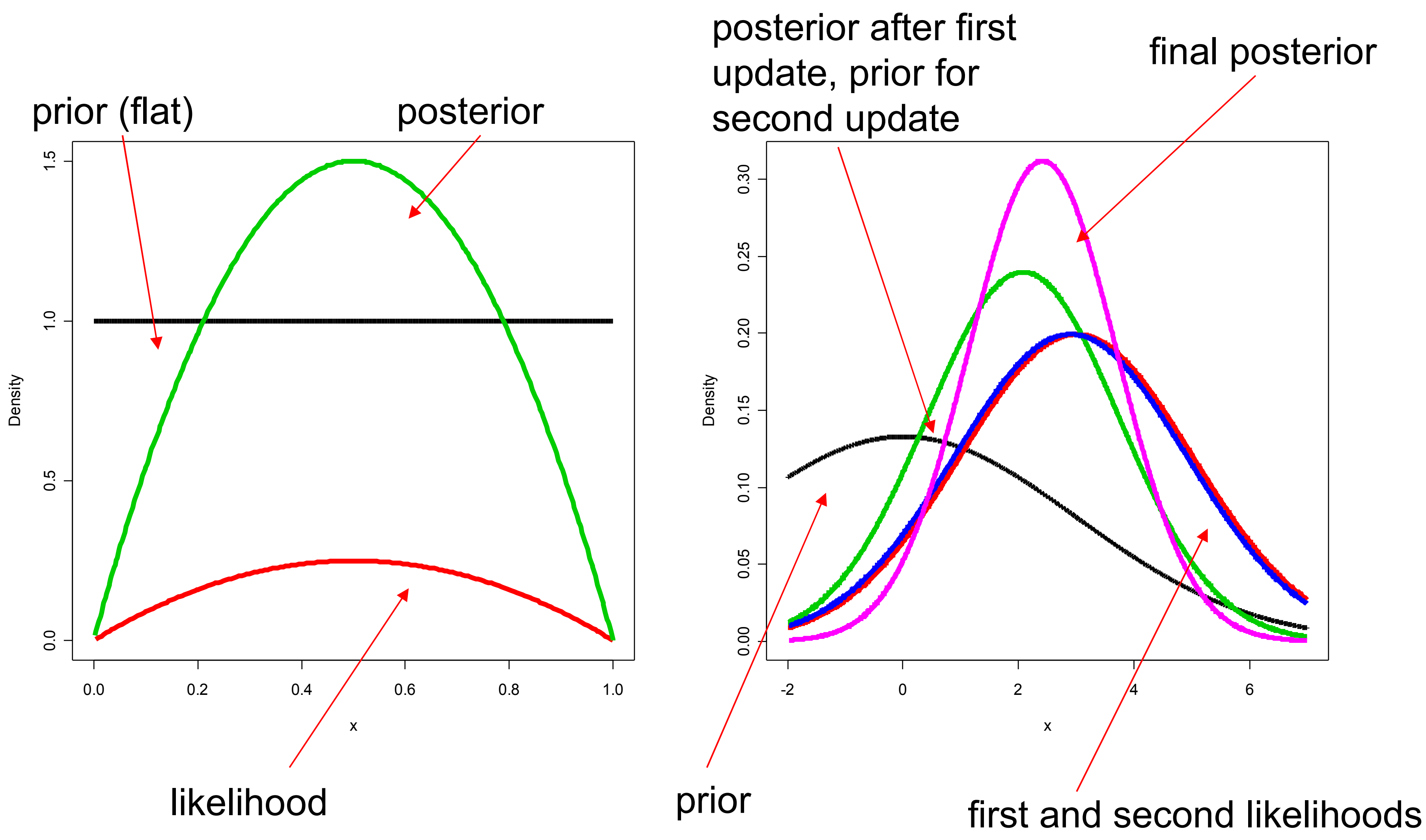
$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)}$$

- Now, $\pi(\theta \mid y)$ and $\pi(\theta)$ are continuous probability distributions for θ, and $\pi(y \mid \theta)$ is the likelihood function

- Fixing y we get
  $\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta)$
  meaning that the two sides are proportional as functions of θ.



Examples



prior (flat)
posterior
likelihood
Density
x
posterior after first update, prior for second update
final posterior
prior
first and second likelihoods

Combining information from different sources

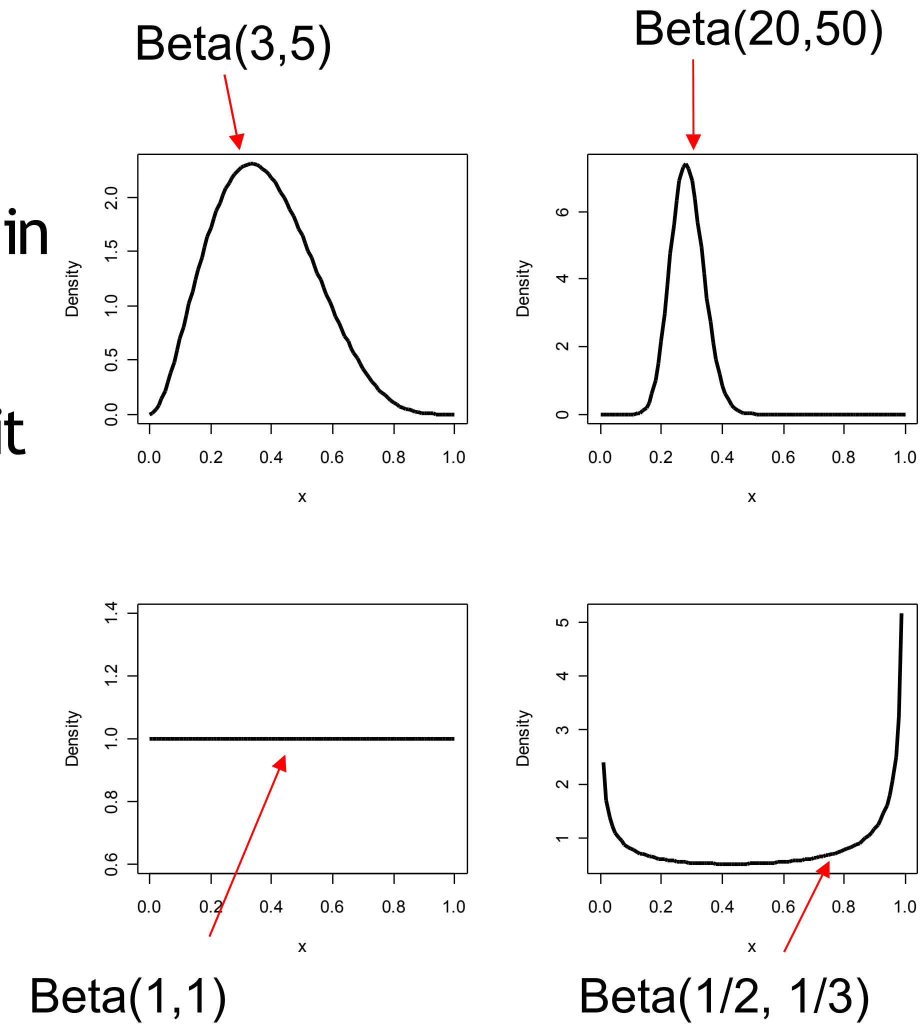
- In general the knowledge about θ is held in a probability distribution.
- This probability distribution is updated when new data is considered, by multiplying it with the likelihood function, and scaling it so that it integrates to 1.
- The resulting posterior distribution can then be used as a new prior distribution which can be updated with further data.

Example: The Beta distribution

- The Beta distribution Beta(α , β) is a distribution for numbers in the interval [0,1].
- The probability distribution for θ when it is Beta(α , β) is proportional to

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

- Expectation α /(α +β), variance α β /((α +β)²(α +β +1))
- R-functions dbeta and rbeta

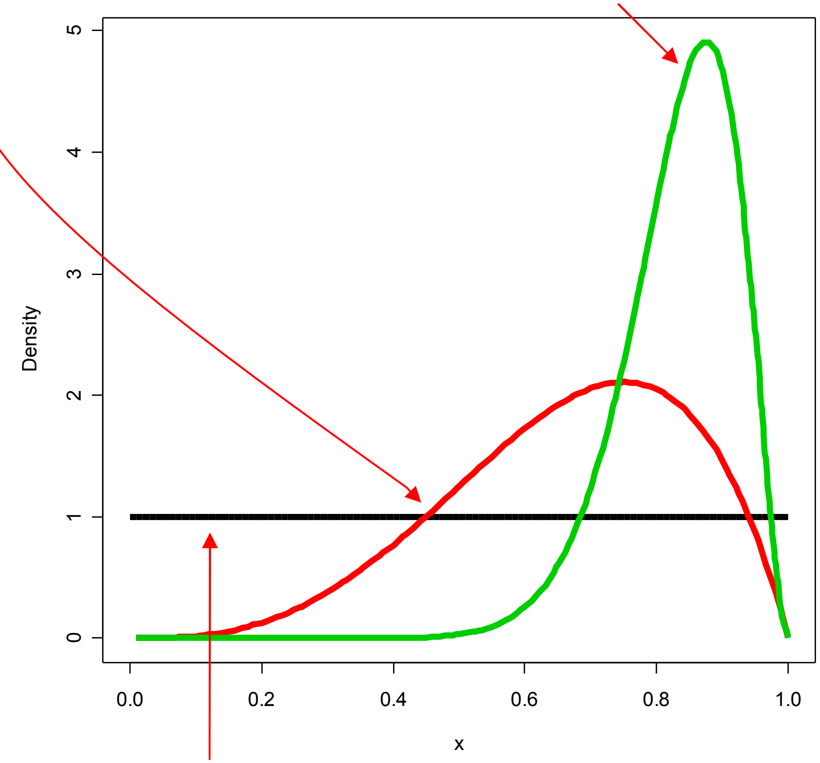


Using the Beta distribution for probabilities

- If p has a Beta distribution and you make a new observation whose probability for "success" is p, then the posterior distribution for p|y is also Beta.
- More precisely, if the prior for p is Beta(α,β) and y is "success", then the posterior for p is Beta(α+1,β), if y is a "failure", then the posterior is Beta(α,β+1)
- More generally, observing x successes and y failures gives the posterior Beta(α+x,β+y).

posterior after 3 successes, 1 failure

posterior after 14 successes, 2 failures



possible prior

Example: Estimating a probability

- Assume you want to learn about the probability p for individuals from your species to have a certain contition. You investigate 3 individuals, and they all have the condition. What can you say about p?
- If you believe that, apriori, any value of p is as likely as any other, use a Beta(1,1) prior. The posterior for p becomes Beta(4,1). The expected value for p is now 4/5 = 0.8.
- Another approach is to make an estimate p0 for p, as p0 = 3/3 = 1. When is this a reasonable guess?
- Several other priors for p are reasonable, and used in many situations.

Example: Allele database

- DNA "fingerprinting" is dependent on establishing a database of the frequencies of different alleles at different genetical loci. This should be done separately for separate populations.
- It is then customary to say estimate the probability of observing an allele in this population is equal to its frequency in the database.
- If an allele has not been observed yet in the population, is it reasonable to say that the probability of observing it is zero?
- One solution: Use a prior for the probabilities assigning some low probability for each possible allele, and then update it with information from database

Comparison with classical statistics

- In classical statistics, we have
  - An unknown parameter of interest
  - A model for how data depends on the parameter
  - A way to estimate the parameter from data
  - "Confidence intervals" for estimates, p-values for hypotheses
- Example: Getting an estimate, with confidence interval, from measurements.
- Classical methods can generally be described in the Bayesian setting, and vice versa
- "Bayesian methods": Applying Bayesian paradigm to situations where combination of information from different sources is central.