Bayesian Inference

(An essential introduction)

R. Bellio & N. Torelli

Spring 2018

University of Udine & University of Trieste

Bayes Theorem (basic)

- Bayes' theorem is a rule to compute **conditional probabilities**.
- In other words, it links probability measures on different spaces of events: given two events E and H, the **probability of H conditional on E** is the probability given to H *knowing that E is true* (i.e. E is the new sample space (Ω)).
- More precisely,
  - I have given a probability measure on E and H,
  - I am told that E has occurred,
  - how do I change (if I change) my opinion on H:

$$P(H) \to P(H|E) = ?$$

Theorem of Bayes (for events) Let E and H be two events, assume $P(E) \neq 0$, then

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{P(H)P(E|H)}{P(E)}$$

Bayes Theorem (an example: a crime case)

In an island a murder is committed

- there is no clue on who is the murderer;
- **but for** the DNA found on the victim (who had a fight with the murderer);
- within the population of the island 1000 persons may have committed the crime, it is a certain thing that the murderer is among them, with equal probability.

The police compares the DNA of all the 1000 suspects with that of the murderer.

The DNA test used by the island police force is not perfect, there is

- probability of a false positive 1%;
- probability of a false negative 2%.

Bayes Theorem (an example: a crime case)

Formally, if

- T is the 'event "the person is positive at the test",
- C is the event "the person is guilty",

the two assumption are then

- probability of a false positive 1% : $P(T|\bar{C}) = 0.01$,
- probability of a false negative 2% : $P(\bar{T}|C) = 0.02$.

The experimental observation is: the police starts testing the 1000 suspects and the 130-th is positive at testing.

Crime case: the investigation (likelihood inference)

The sheriff says that the experimental evidence, represented by the ratio between the likelihoods

$$\frac{P(T|C)}{P(T|\bar{C})} = \frac{0.98}{0.01} = 98$$

is overwhelmingly in favour of that person being guilty, thus constituting decisive evidence, so he asks the judge to arrest and condemn the guy.

Being more formal, who are model and likelihood?

- **model**: set of probability distributions which may have generated the sample T, there are two alternatives, represented by $\{C, \bar{C}\}$:

$$P(T|C) = 0.98 \quad P(T|\bar{C}) = 0.01$$

- **likelihood** the parameter space is $\{C, \bar{C}\}$, the likelihood takes two values

$$L_C = P(T|C) = 0.98 \quad L_{\bar{C}} = P(T|\bar{C}) = 0.01$$

- the maximum likelihood estimate is then C.

Crime case: the investigation (likelihood inference)

Even more formally,

- **parameter**: θ is 1 if the man is guilty, 0 otherwise (the parameter space is then $\Theta = \{0, 1\}$)
- **sample**: y is 1 if the test is positive, 0 otherwise (sample space $\{0, 1\}$)
- **model**:

$$f(y;\theta) = P(T|C)^{y\theta} P(\bar{T}|C)^{(1-y)\theta} P(T|\bar{C})^{y(1-\theta)} P(\bar{T}|\bar{C})^{(1-y)(1-\theta)}$$

- **likelihood**:

$$L(\theta; y) \quad \propto \quad f(y;\theta) \propto \left(98^y \left(\frac{2}{99}\right)^{1-y}\right)^{\theta} 0.99 \left(\frac{1}{99}\right)^{y} \propto \left(98^y \left(\frac{2}{99}\right)^{1-y}\right)$$

- **likelihood with** $y = 1$:

$$L(\theta; 1) \propto 98^{\theta}$$

- the maximum likelihood estimate is then $\theta = 1$.

Crime case: the (bayesian) defence

A Bayesian lawyer argues against the sheriff and notes that *despite the fact that the experimental evidence is much more compatible with the man being guilty, the verdict should be based on the probability of the man being guilty*: **the sheriff ignores prior probabilities**

- a priori: before the test, the suspect was only one among 1000 suspects, the probability of him being guilty is $P(C) = 0.001$.
- data and likelihood: having observed T and knowing that $P(T|C) = 0.98$

we obtain that

$$P(C|T) = \frac{P(C)P(T|C)}{P(T)} = 0.089$$

Crime case: the (bayesian) defence

To understand the result, consider what would happen, on average, if all 1000 suspects were tested:

- 999 are innocent, among them
  - $999P(T|\bar{C}) = 9.99$ test positive;
  - the other 989.01 test negative;
- 1 is guilty,

1. he tests positive with probability 0.98
2. he tests negative with probability 0.02

then, on average, the 1000 suspects partition as follows

	Pos	Neg
1 guilty	0.98	0.02
999 innocent	9.99	989.01
Tot	10.97	989.03

likelihood ratio

Another way of interpreting the process is driven by the formula

$$\begin{aligned}\frac{P(C|T)}{P(\bar{C}|T)} &= \frac{P(C)}{P(\bar{C})}\frac{L(C|T)}{L(\bar{C}|T} \\ &= \frac{1/1000}{999/1000}98 \\ &= 0.0981\end{aligned}$$

before the experiment, the individual is 999 times less likely of being guilty than not guilty.

Given the experiment we update this ratio by multiplying it for the likelihood ratios: the individual is $1/0.098 \approx 10$ times less likely of being guilty than not guilty.

Bayes' theorem (more than two hypotheses)

We now consider a more general version of Bayes' theorem where more than two events are involved,

Bayes (with n hypotheses)

`If`

1. $P(E) \neq 0$
2. $\{H_i | i = 1, \ldots, n\}$ partition of Ω
   $\cup_{i=1}^{n} H_i = \Omega$; $H_i \cap H_j = \phi$ if $i \neq j$ then

$$P(H_j|E) = \frac{P(H_j)P(E|H_j)}{P(E)} \propto P(H_j)P(E|H_j)$$

A second example: Box of seeds

The problem

A factory sells boxes of seeds. It produces 4 tipes of boxes and each box mixes 2 type of seeds: High Quality seeds and Normal seeds.

The boxed are labelled as Platinum (P), Gold (G), Extra (E) and Standard (S). Platinum has 90% of High Quality seeds, Gold 80%, Extra 70%, Premium 50%.

Assume you have an unlabelled box and you want to decide which kind of box it is by selecting (with replacement) a sample of 30 seeds. Now assume that in your sample the number of High quality seeds is 23.

Which kind of box is it?

Maximum likelihood estimation

We can rephrase the problem as follows:

- let p be the proportion of High quality seeds in a box.
- we observe a sample $y_1, y_2, \ldots, y_{30}$ from a rv $Y_i \sim Be(p)$
- we want to use these data to estimate the parameter p where $p = \{0.9, 0.8, 0.7, 0.5\}$

We can write the likelihood function, *i.e.*, the probability of observing x (23 in our case) high quality seeds when the box is P, G, E or S and p can take on one of the 4 values $p_1 = 0.9, p_2 = 0.8, p_3 = 0.7, p_4 = 0.5$

$$L(p_i) = \binom{30}{x} p_i^x \left(1 - p_i\right)^{20-x}$$

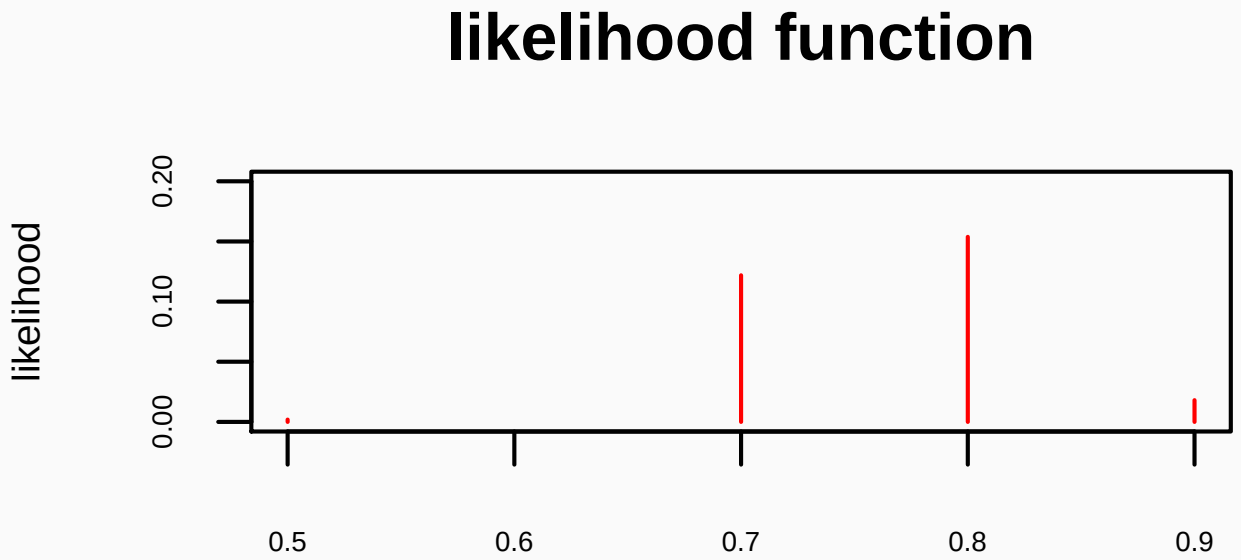
and calculate it. Note that p_i is our parameter and the parameter space contains only four elements.

Maximum likelihood estimation

```
p <- c(.9, .8, .7, .5); n <- 30; x=23;
L <- choose(30,x)*p^x*(1-p)^{30-x}
L
```

```
## [1] 0.018043169 0.153820699 0.121853726 0.001895986
```

```
plot(p,L,type="h",main="likelihood function", cex.lab=0.7,
     , cex.axis=0.5, ylab="likelihood", ylim=c(0,0.2), col=2)
```



A different perspective: toward bayesian inference

Assume that we know that in the factory the proportions of Platinum, Gold, Extra and Standard boxes are as follows

prop(Platinum)	prop(Gold)	prop(Extra)	prop(Standard)
0.1	0.2	0.3	0.4

One can then assume, even before seeing the sample of 30 seeds, that the probability of getting one specific type of boxes, *i.e*, of getting a specific value for p_i is:

p_i	0.9	0.8	0.7	0.5
$P(p_i)$	0.1	0.2	0.3	0.4

Bayesian solution

In Bayesian inference we want to express our uncertainty about the parameter p by giving a probability distribution on it. The quantity p is now random.

Note that we have:

- a probability distribution on the possible values of p before observing the sample $P(p_i)$. This is called the **prior distribution**
- the **likelihod function** $L(p_i)$, but since now p is a rv, then we can rewrite it as the conditional probability $P(x|p_i)$, where x is evidence from the sample.
- Bayes theorem can then be applied to get the so called **posterior distribution**

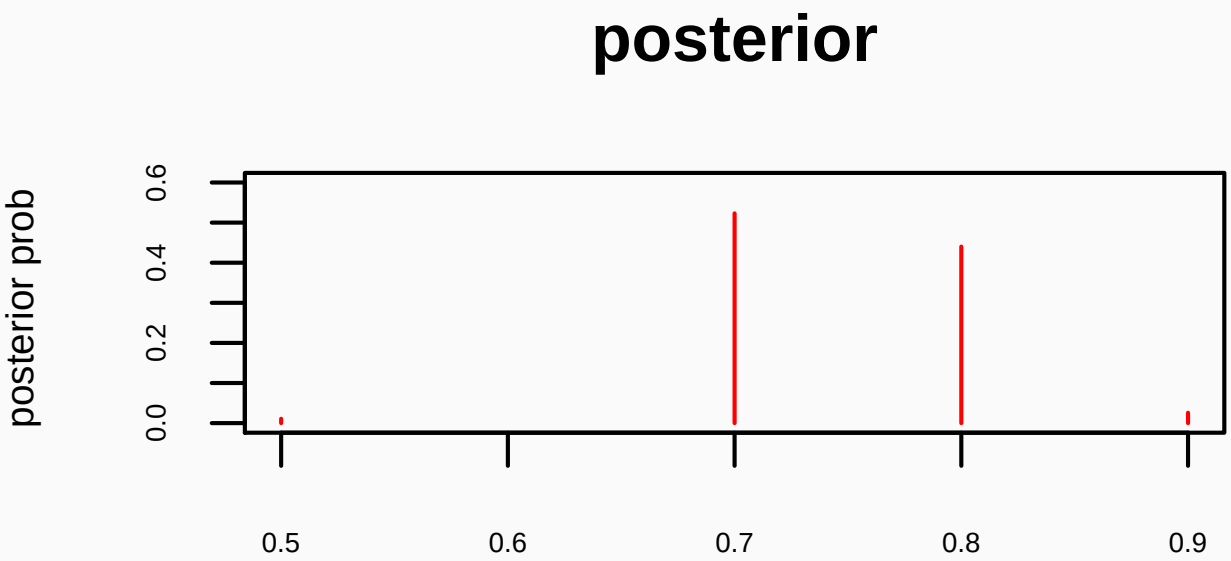
$$P(p_i|x) = \frac{P(p_i)L(p_i)}{\sum_i^4 P(p_i)L(p_i)} = \frac{P(p_i)P(x|p_i)}{\sum_i^4 P(p_i)P(x|p_i)} \propto P(p_i)P(x|p_i)$$

Bayesian solution

```
prior <- c(.1, .2, .3, .4)
like <- choose(30,x)*p^x*(1-p)^{30-x}
posterior <- prior*like/sum(prior*like); posterior
```

```
## [1] 0.02581912 0.44022371 0.52310482 0.01085235
```

```
plot(p,posterior,type="h", main="posterior", cex.lab=0.7,
   cex.axis=0.5, ylab="posterior prob ", ylim=c(0,0.6), col=2)
```



Likelihood vs Bayesian

In the example:

- The likelihood estimate was $p = 0.8$, Gold, since this is value of the parameter with the highest value of the likelihood.
- In Bayesian inference we have a probability distribution over the parameter space. We can say that the value is $p = 0.7$ with probability ≈ 0.52, Extra. This is a probability statement.
- Bayesian approach allows us to update our prior information with experimental data

$$\begin{array}{ccccc} \text{information} & & \text{information} & & \text{information} \\ \text{post} & \propto & \text{from} & \times & \text{prior to} \\ \text{experiment} & & \text{experiment} & & \text{experiment.} \end{array}$$

posterior $\propto$ prior$\times$likelihood

Bayes theorem: continuos variables

Bayes Theorem

If

(i) $\pi(\theta)$ density function
(ii) $f(y|\theta)$ density function of y given θ

then

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\int_\Theta f(y|\theta)\pi(\theta)d\theta} \propto \pi(\theta)f(y|\theta)$$

Note that $\int_\Theta \pi(\theta|y)d\theta = 1$ and that the quantity $\int_\Theta f(y|\theta)\pi(\theta)d\theta$ is called the normalization constant.

Bayesian paradigm: model and likelihood

Consider a **model**, a family of probability distributions indexed by a parameter θ among which we assume there is the distribution of y:

$$f(y|\theta), \quad \theta \in \Theta.$$

This is no different than the classical paradigm, but for the fact that the distributions are defined conditional on the value of the parameter (which is not a r.v. in the classical setting)

One defines then the likelihood

$$L(\theta; y) \propto f(y|\theta),$$

as in the classic paradigm.

Bayesian paradigm: prior distribution

A prior distribution is set on the parameter θ

$$\pi(\theta)$$

which is independent of observations (it is called prior since it comes before observation).

This is the new thing

Prior information and likelihood are combined in Bayes' theorem to give the posterior distribution

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\int_\Theta f(y|\theta)\pi(\theta)d\theta} \propto \pi(\theta)f(y|\theta) \propto \pi(\theta)L(\theta; y)$$

which sums up all the information we have on the parameter θ.

Inference on the quality of seeds

We are again interested in estimating the proportion of high quality seeds in the boxes. But now we assume that the proportion p can be any value in the interval $[0, 1]$. We still have a sample of n seeds drawn from a box (with replacement), and we count the number of high quality seeds x.

We want to infer on the value p. Data are i.i.d realizations from a $Be(p)$.

We can never know the real value of p, unless we can rely on a sample where $n \to \infty$.

We can design a procedure that selects values of p that are more supported by the data. We can then judge how uncertain is our procedure by looking at its behaviour in possible (not actual) replication of the sample under the same condition. **Classical inference**

We can try to give a probability distribution over possible values of the parameter p. And this probability distribution will summarize all the information we have about it: before and after observing the data **Bayesian inference**

Estimating p

1. Likelihood estimation is straightforward:

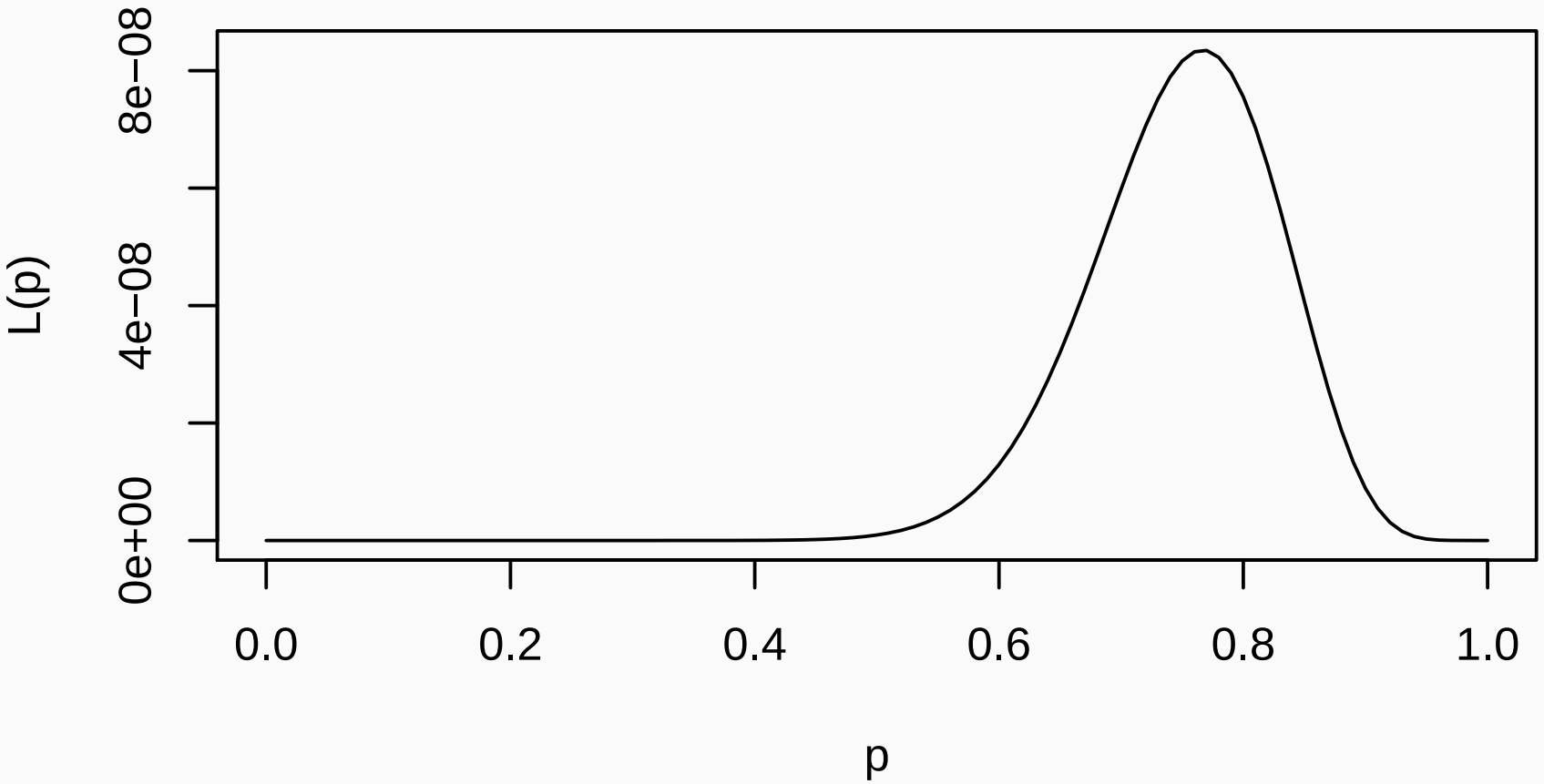
- $L(p) \propto p^x(1-p)^{n-x}$
- it is easy to show that ML estimate is $\hat{p} = x/n$. The observed proportion of high quality seeds in the sample.

2. Bayesian solution requires specification of the probability distribution $\pi(p)$.

- Since $p \in [0, 1]$ candidates are probability models whose support is the interval $[0, 1]$.
- Random variables belonging to the Beta family could be appropriate

The likelihood function

```
n <- 30; z=23;
curve(x^z*(1-x)^(30-z), xlim=c(0,1), xlab="p",ylab="L(p)")
```



The Beta distributions

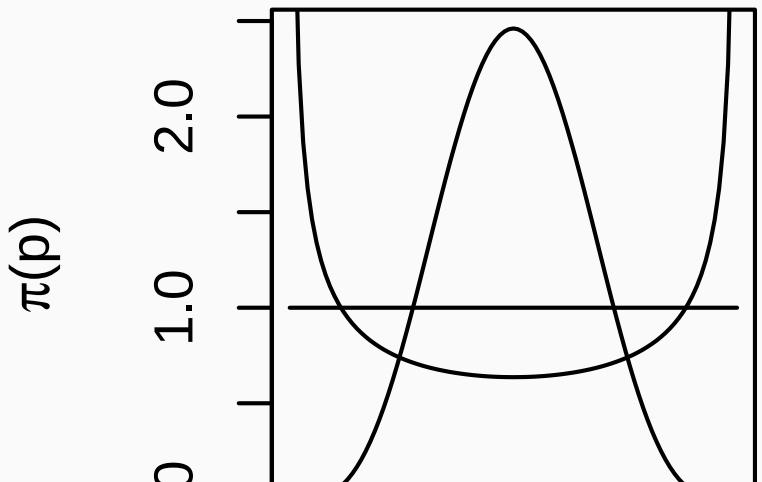
$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

dove $0 < \theta < 1$ e $\alpha, \beta > 0$,

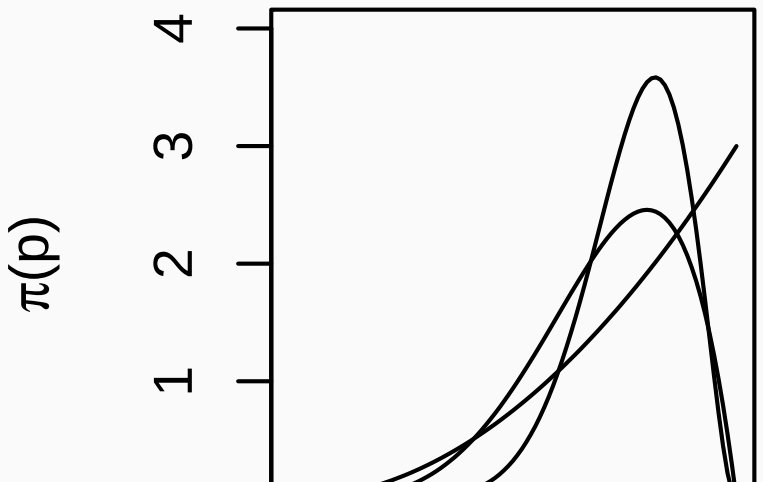
$$E(\theta) = \frac{\alpha}{\alpha+\beta} \quad V(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

remind that $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^t dt$ and if x is integer $\Gamma(x) = (x-1)!$ ##
The Beta distributions

α=β



α>β



The posterior distribution

Since

$$\begin{aligned}\pi(p|x) &\propto L(p)\pi(p) \propto p^x(1-p)^{n-x}p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{\alpha+x-1}(1-p)^{\beta+n-x-1}\end{aligned}$$

then

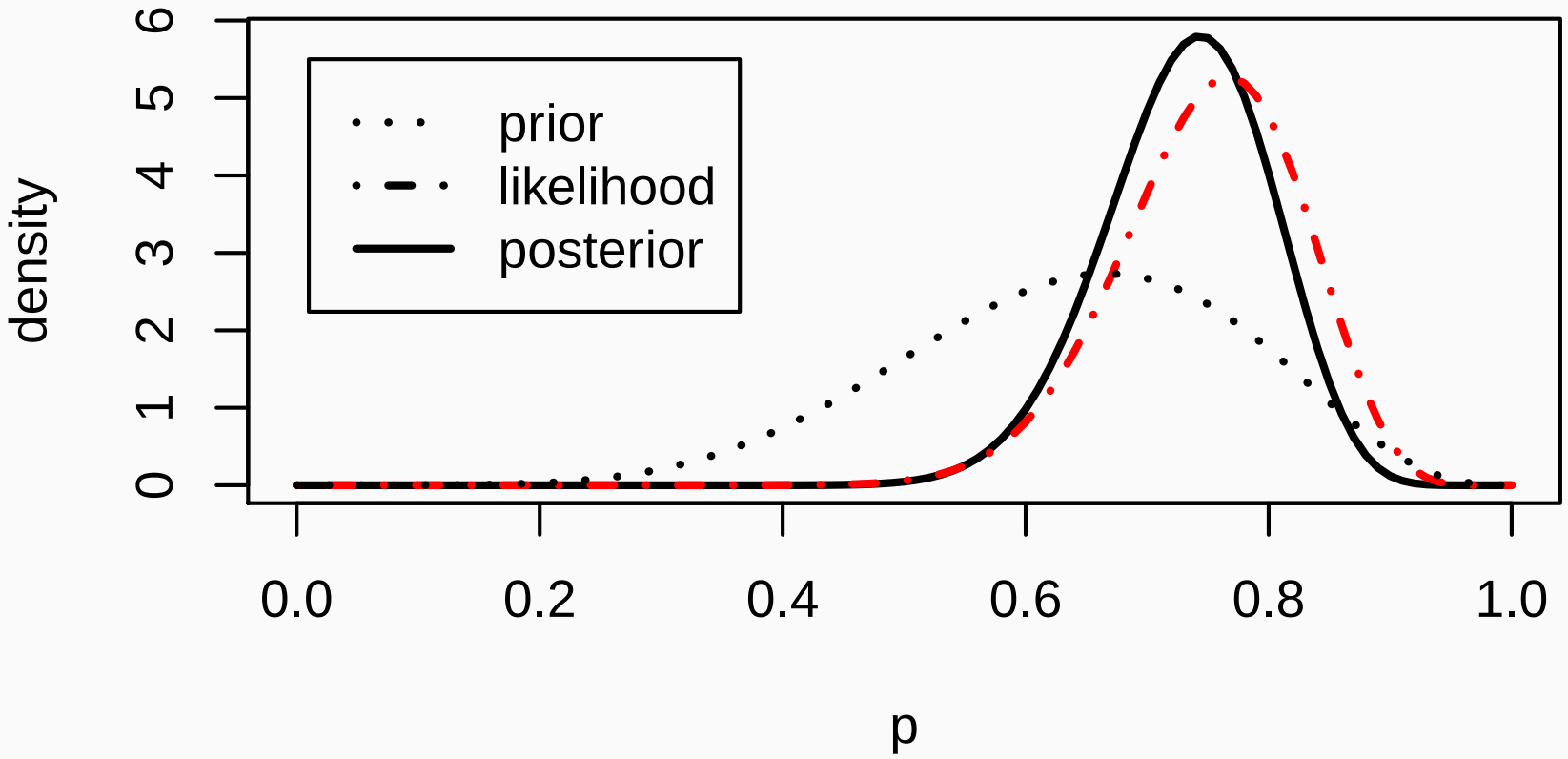
$$\pi(p|x) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-+x)}p^{\alpha+x-1}(1-p)^{\beta+n-x-1}$$

the posterior for θ is then a Beta with parameters $\alpha + x$ and $\beta + n - x$.

Likelihood, prior and posterior

Assume that in our case we believe, before getting the sample, that values above 0.5 are more likely. I could use as a prior a Beta(7,4) whose mean is 7/11=0.636. Note that the likelihood (normalized so that it integrate to 1) is also a Beta with parameters (24,8).

Then the posterior is still a Beta(31,10)



Likelihood and Bayesian inference

A statistical problem is faced when, given observations, we want to assess what random mechanism generated them

- In other words,
  - there are two or more probability distributions which may have generated the observations;
  - analyzing the data we want to infer on the actual distribution which generated the data (or on some property of it).
- How? Let us discriminate between the two approaches.
  - Based on the likelihood we compare $P(Data|Model)$ for the different models.
  - In Bayesian statistics comparing $P(Model|Data)$.

In the likelihood approach quality of the procedure is evaluated relying upon fictitious repetitions of the experiment.

Classical and Bayesian statistical inference, differences

In **CLASSICAL INFERENCE**

- the conclusion is not derived within probability calculus rules (these are used in fact, but the conclusion is not a direct consequence)
- the likelihood and the probability distribution of the sample are used;
- the parameter is a constant.

In **BAYESIAN INFERENCE**

- the reasoning and the conclusion is an immediate consequence of probability calculus rules (more specifically of Bayes' theorem);
- the likelihood and the prior distribution are used;
- the parameter is a random variable.

Bayesian vs classical inference

In the Bayesian approach the parameter is random: this is a fundamental difference between the two approaches, how can this be interpreted?

- In classical statistics, on the contrary, the parameter is a fixed quantity.
- the random character of θ represents our ignorance on it.
- random means, in this context, not known for lack of information. We measure our uncertainty about the model
- The randomness and the probability distribution on θ are subjective.
- The probability in Bayesian approach is a subjective probability, *i.e.*, the probability of a given event is defined as the "degree of belief of the subject on the event".

The role of subjective probability

- Consider events such as *tail is observed when a coin is thrown*,
  - everyone (presumably) would agree on the value of the probability;
  - the frequentist definition is intuitively applied;
  - $\rightarrow$ this is an 'objective' probability.
- For events such as *Juventus will be Italian champion next year* or *Right wing parties will win next elections*,
  - it is still possible to state a probability;
  - everyone would assign a different probability;
  - the probability given by someone will change in time depending on avalaable information.

One then accepts that the probability is not an objective property of a phenomenon but rather the opinion of a person and one defines

Subjective probability: definition (de Finetti)

The probability of an event is, for an individual, his degree of belief on the event.

Bayesian statistics and subjective probability

If the probability is a subjective degree of belief, it depends on the information which is subjectively available, and that by random we mean not known for lack of information.

The subjective definition of probability is most compatible with the Bayesian paradigm, in which:

- the parameter to be estimated is a well specified quantity but is not known for lack of information
- a probability distribution is (subjectively) specified for the parameter to be estimated, this is called **a priori**
- after seeing experimental results the probability distribution on the parameter is updated using Bayes' theorem to combine experimental results (likelihood) and the prior to obtain the posterior distribution.

Note that, starting in 1763 (the year Bayes' theorem were published), Bayesian statistics comes first, before the so–called classical statistics, initially developed by Galton and Pearson at the end of XIX century and then by Fisher in the twenties.