CLASSICAL STATISTIC IS BASED ON ASSUMPTION
BUT THAT IS NOT VERY "SCIENTIFIC"

Bayesian Inference

(An essential introduction)

N. Torelli, G. Di Credico, V. Gioia

2024

University of Trieste

Introduction

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)). IT LINKS E TO H
- 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:

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

I BELIEVE/GET MY PROBABILITIES ON SOMETHING THAT HAPPENS

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

I KEEP THE PROBABILITY OF E ON H

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

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 assumptions 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 (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

PRIOR LIKELIHOOD

$$P(C|T) = \frac{P(C)P(T|C)}{P(T)} = \frac{0.001 \times 0.98}{0.001 \times 0.98 + 0.999 \times 0.01} = 0.0893$$

P(BEING GUILTY AND POSITIVE)

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

and the probability of being gilty given you teste positive is simply $\frac{0.98}{10.97} = 0.893$

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)

$\{H_i | i = 1, \ldots, n\}$ is a partition of the sample space Ω such that $\cup_{i=1}^{n} H_i = \Omega$; $H_i \cap H_j = \phi$ if $i \neq j$ then

$$P(H_j|E) = \frac{P(H_j)P(E|H_j)}{\sum_{j=1}^{n} P(H_j)P(E|H_j) = P(E)} \propto P(H_j)P(E|H_j)$$

A second example: which 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 boxes are labelled as Standard (S), Extra (E) and Platinum (P). 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 $x_1, x_2, \ldots, x_{30}$ from a rv $X_i \sim Be(p)$ and let $x = \sum_i^{30} x_i$
- we want to use these data to estimate the parameter p where $p = \{0.5, 0.7, 0.8, 0.9\}$

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 $p_1 = 0.5, p_2 = 0.7, p_3 = 0.8, p_4 = 0.9$

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

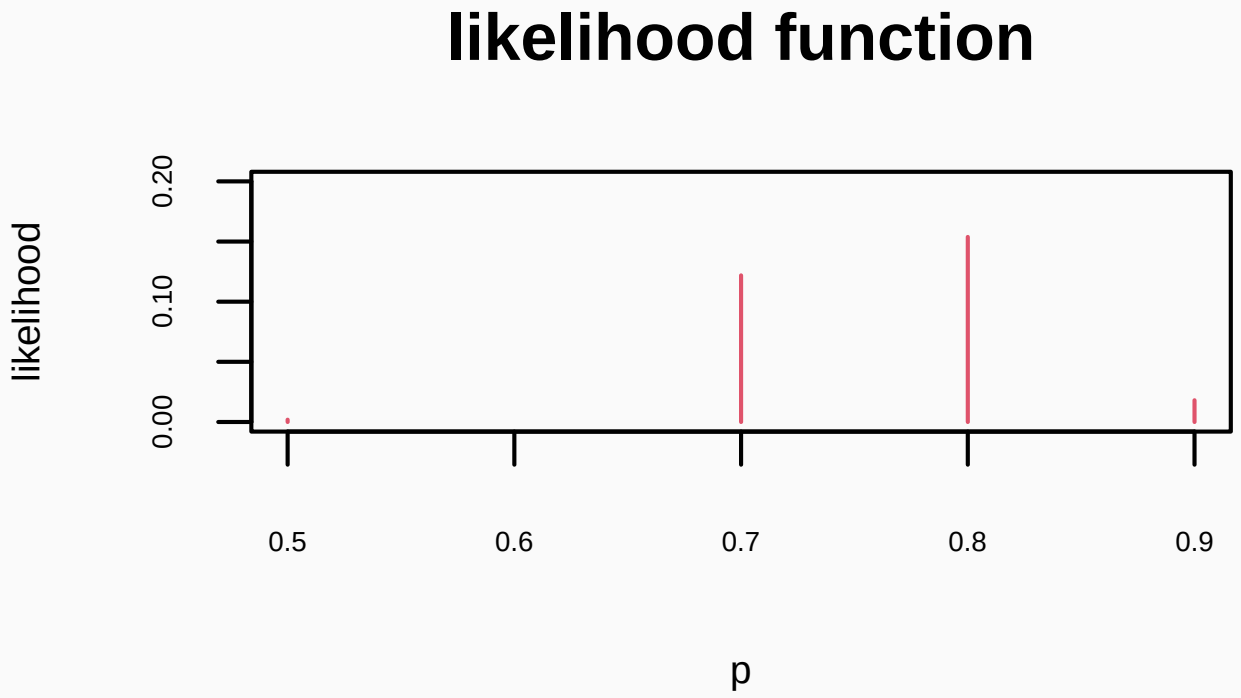
and calculate it. Note that p_i is our parameter and the parameter space contains only four elements.

Maximum likelihood estimation

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

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

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

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

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 **likelihood 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**

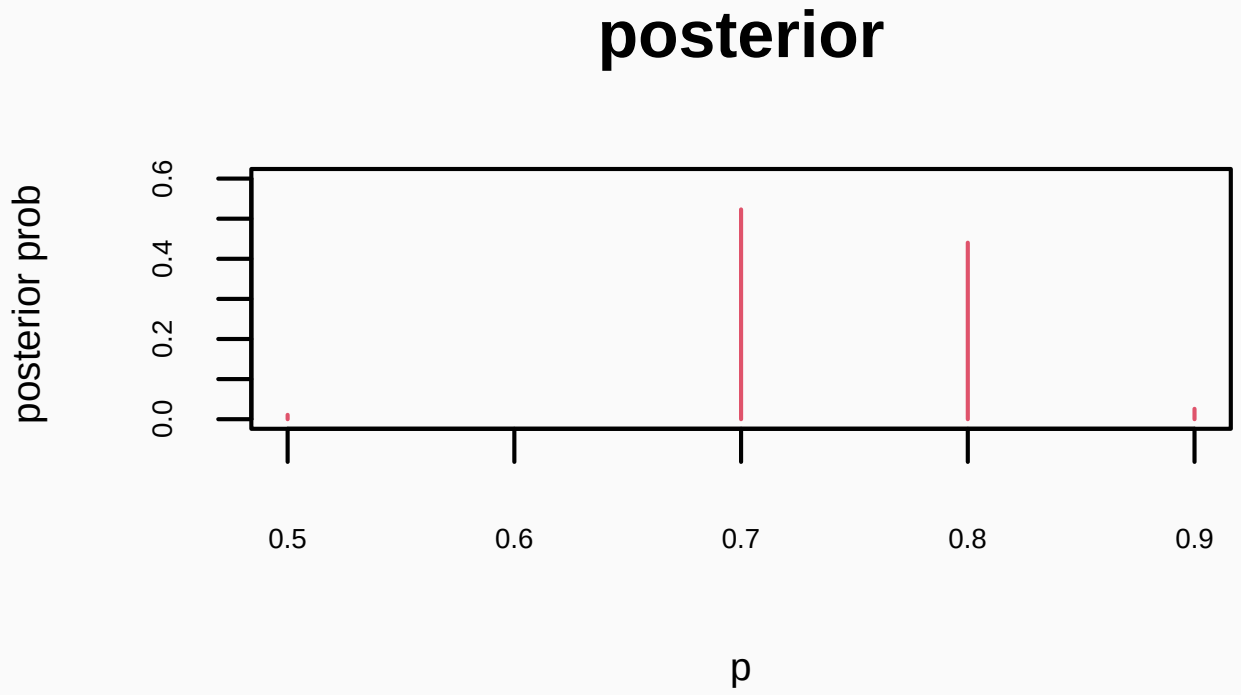
$$P(p_i|x) = \frac{P(p_i)L(p_i)}{\sum_i^4 P(p_i)L(p_i)} = \frac{P(p_i)P(x|p_i)}{\sum_i^4 P(p_i)P(x|p_i)} \propto P(p_i)P(x|p_i)$$

Bayesian solution

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

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

```
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

information post experiment $\propto$ information from experiment $\times$ information prior to experiment

posterior $\propto$ prior$\times$likelihood

Introducing Bayesian inference

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**

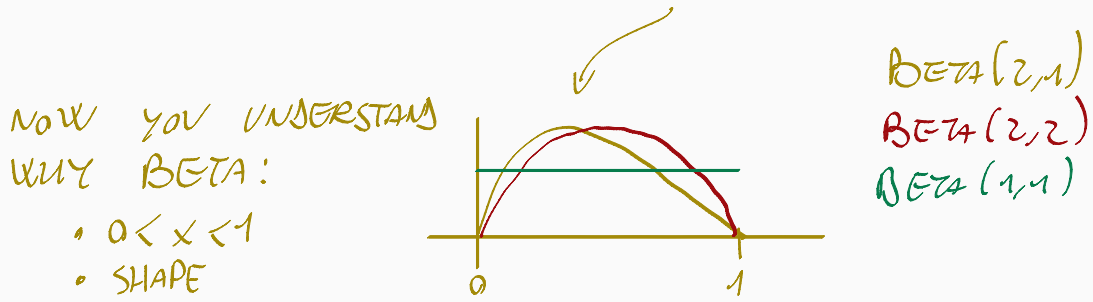
Inference on p

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 $\pi(p)$.

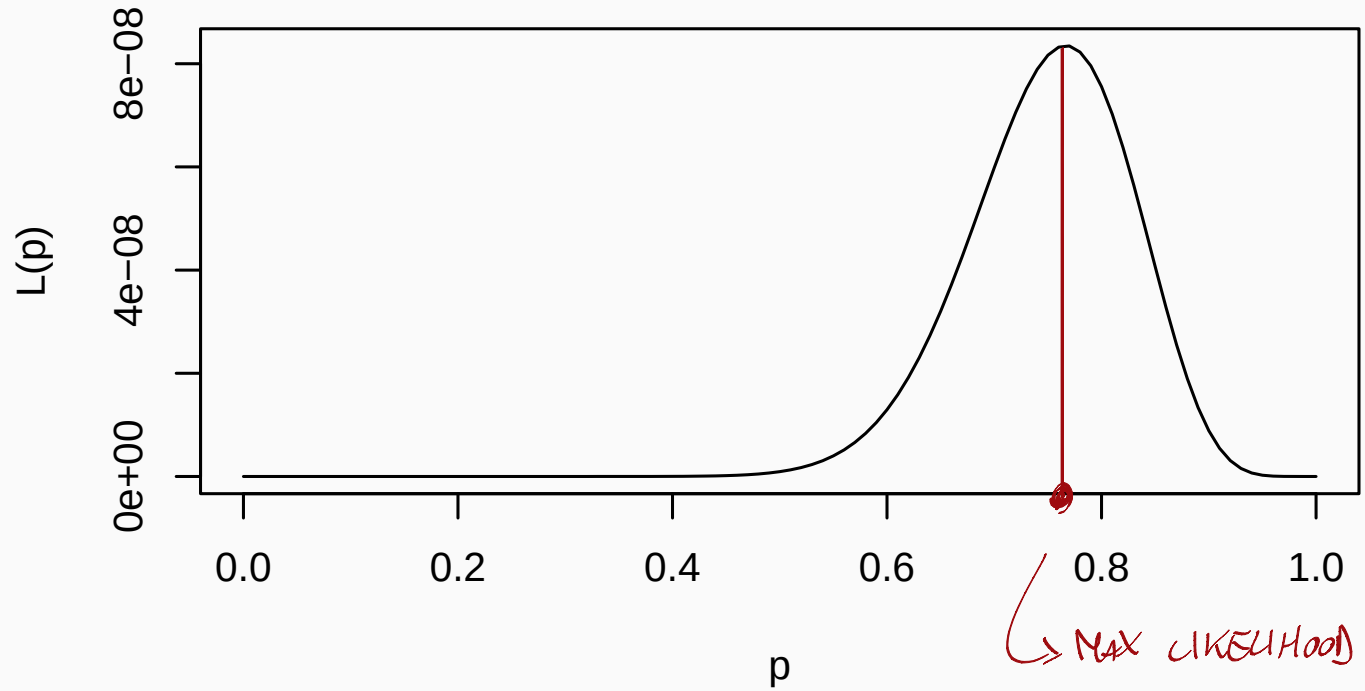
- Since $p \in [0,1]$ candidates are probability models whose support is the interval $[0,1]$.
- Random variables belonging to the Beta family could be appropriate



The likelihood function

BINOMIAL FORMULA : $\pi(p) \cdot p^{Y}(1-p)^{n-Y}$

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



The Beta distributions

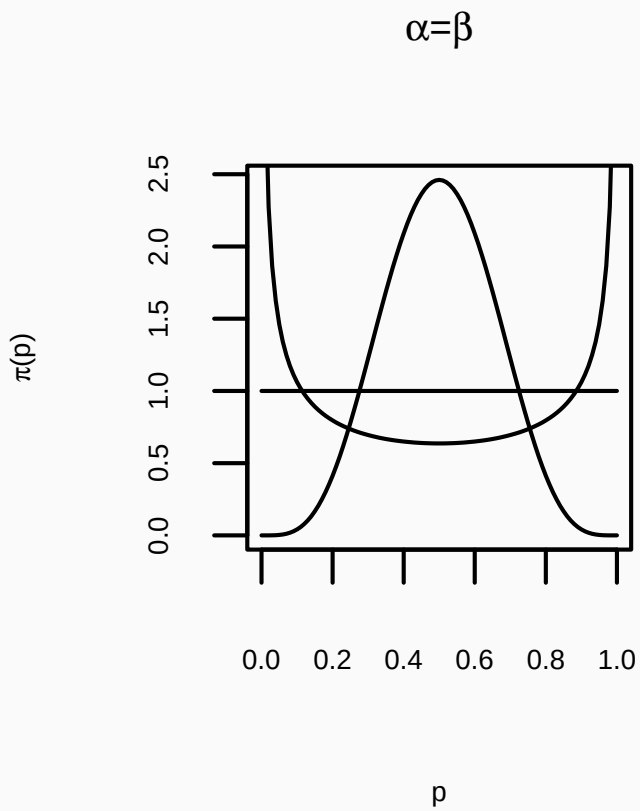
$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

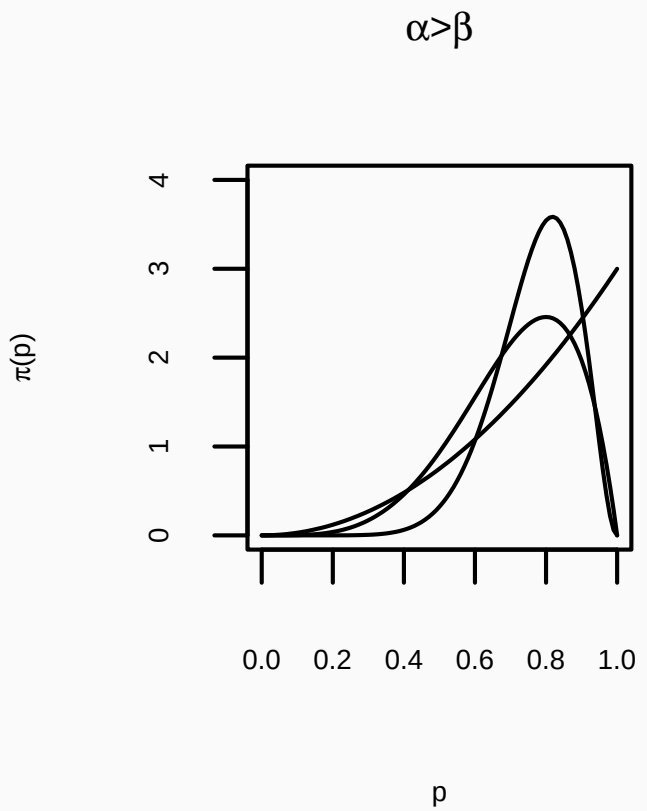
dove $0 < \theta < 1$ e $\alpha, \beta > 0$,

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

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





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 $\beta + n - x$.

NEW VALUE

IS DIFFERENT FROM THE PRIOR DISTRIBUTION

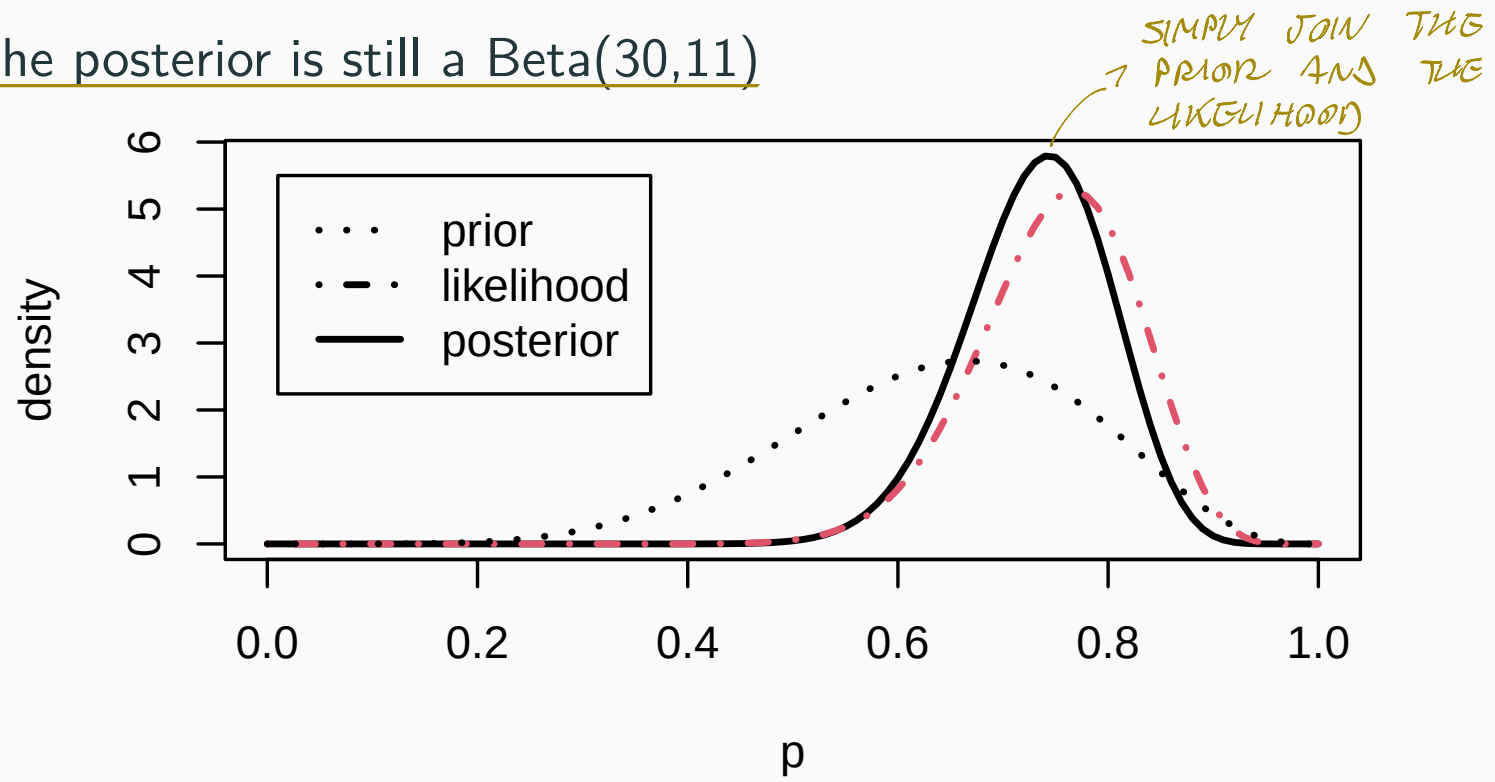
BUT NOW:

$$\bar{E} = \frac{\alpha+x}{\alpha+\beta+n}$$

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(30,11)



Classical and Bayesian Inference

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 avalaible 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 was 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.

Bayesian models

Bayesian models

- A **prior distribution** is defined on the parameter θ

$$\pi(\theta)$$

- we assume an i.i.d. sample $y = (y_1; y_2, \ldots, y_n)$ is obtained from a distribution belonging to a family indexed by θ, and

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

  is then proportional to the likelihood function.

- Bayes theorem allows us to combine prior information and likelihood to give the **posterior distribution**

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

The posterior distribution then sums up all the information (and the uncertainty) about the parameter θ. Inference on the parameter amount at studying and appropriately summarizing the posterior.

Again on the proportion of seeds

Consider again the example of the estimation the proportion p of High quality seeds in a box. Recall that for inference on p we assumed that:

- $\pi(p)$ is $Beta(\alpha, \beta)$, the prior
- $f(x_i|p)$ are $Be(p)$ and $f(x|p) \propto Bi(n, p)$ is the likelihood

(where $x = \sum_i^n x_i$ and x_i are i.i.d).

- $\pi(p|x)$, the posterior, is $Beta(x + \alpha, \beta + n - x)$.

In our example, $n = 30, x = 23$, the prior is $Beta(7, 4)$

Then the posterior is a $Beta(30, 11)$

As stated the posterior distribution summarizes what we know about the parameter combining prior knowledge and experimental data.

So inference on p derives from the analysis of this distribution. And we will use it to illustrate the procedures for point and interval estimation

Point estimation of the the proportion of seeds

If we want to select a single value as a point estimate of p, let say $\hat{p}$, we are back to a classical problem: how to select a single number to summarize a distribution $\pi(p)$.

Classical solutions are:

- the expected value $E(P)$ of the posterior distribution of the rv P, $E(p) = \int_0^1 p\pi(p|x)dp$
- the median Me of the posterior distribution, $Me : \int_0^{Me} \pi(p|x)dp = 0.5$
- the mode Mo of the posterior distribution,$i.e.$, the value of p for which $\pi(p|x)$ is maximum.

One can choose one of these as point estimate and, provided that the posterior is unimodal, they provide an appropriate synthesis.

Obviously the three values are equivalent if the posterior is symmetric and unimodal.

Bayes risk

More formally, Bayes estimators can be defined as the quantity that minimizes the posterior expected value of a loss function $L(\theta, \widehat{\theta})$

The quantity $E_{\pi|x}(L(\theta, \widehat{\theta}))$ is called **Bayes risk**.

1. When $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$,

i.e. a quadratic loss is used, the quantity that minimizes the Bayes risk is the posterior mean.

2. Posterior median can be also justified as the quantity that minimizes a "linear" loss function, with $a > 0$, defined as $L(\theta, \widehat{\theta}) = a|\theta - \widehat{\theta}|$.

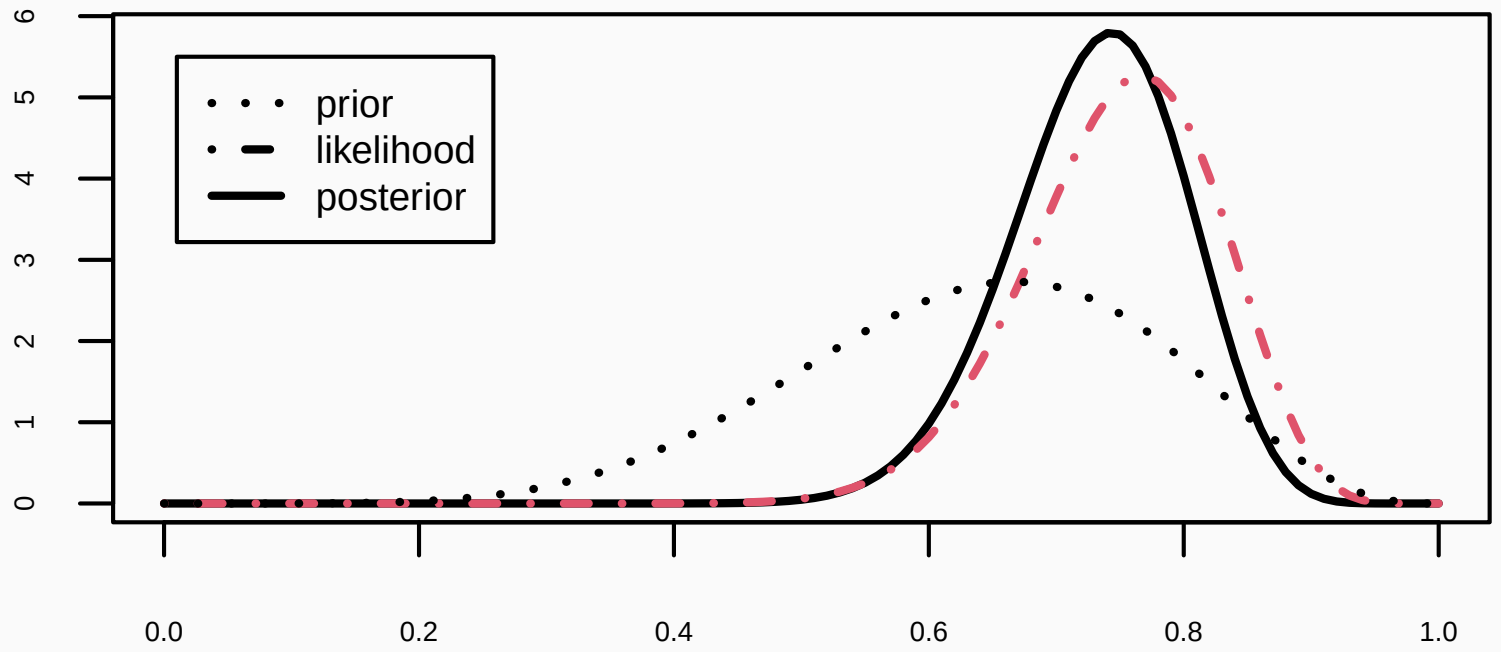
3. Posterior mode (MAP: Maximum A-posteriori Probability) can be justified as the minimizer of the trickier loss function of the form

$$L(\theta, \widehat{\theta}) = \begin{cases} 0 & \text{if } |\widehat{\theta} - \theta| < c, \\ 1 & \text{otherwise,} \end{cases}$$

as c goes to 0

Estimating the quantity of seeds by the posterior mean

```
curve(dbeta(x,7+23,4+7),xlab="p", ylab="density",lty=1,lwd=2,
      cex.axis=.5, cex.lab=.6, ann=F)
curve(dbeta(x,23+1,7+1),add=TRUE,lty=4,lwd=2, col=2)
curve(dbeta(x,7,4),add=TRUE,lty=3,lwd=2)
legend(.01,5.5,c("prior","likelihood","posterior"), lty=c(3,4,1))
```



Using Posterior mean as an estimator of p

Recall that for if $X \sim Beta(\alpha, \beta)$ then $E(X) = \frac{\alpha}{\alpha+\beta}$

being $\pi(\theta|y)$ a Beta distribution, the posterior mean is

$$
\begin{aligned}
&= \frac{\alpha + x}{\alpha + \beta + n} \\
&= \frac{\alpha + \beta}{\alpha + \beta + n} \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \frac{x}{n} \\
&= \frac{\alpha + \beta}{\alpha + \beta + n} E_\pi(p) + \frac{n}{\alpha + \beta + n} \hat{p}_{ML}
\end{aligned}
$$

IF $n \to 0$ GO TO 0 => FOR LARGE n, PRIOR IS NOT RELEVANT AND THE MLE DOMINATES

$$
E(P|x) = \frac{\alpha + \beta}{\alpha + \beta + n} \underbrace{E_\pi(P)}_{\text{prior expectation}} + \frac{n}{\alpha + \beta + n} \underbrace{\hat{p}}_{MLE}
$$

A closer look to posterior distribution

We have seen that the posterior mean is a weighted average of the prior expectation and the ML estimate, where

- ML estimate prevails if n is large;
- ML estimate prevails if α and β are small (the variance of the prior distribution is large). It is worth noting that $\alpha + \beta$ can be interpreted as the equivalent number of observation of the prior distribution.

The posterior distribution as a whole is a compromise between the prior and the likelihood, and the likelihood prevails if

- n is large;
- α and β are both close to 1 (the prior is diffuse)

To appreciate the quality of the posterior mean as an estimate we can look at the posterior variance (or at standard deviation)

$V(\theta) = \frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$ that for large n is $\approx \frac{1}{n}\frac{x}{n}(1-\frac{x}{n})$

A model for gaussian data

Assume that observations come from a gaussian distribution (variance known)

- $Y_1, \ldots, Y_n \sim iid \quad N(\mu, \sigma^2)$ conditional to parameter(s) value(s)

μ is the parameter, σ^2 is known;

the likelihood L $(= L(\mu))$ is

$$L \propto \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \right)$$

Gaussian model; σ^2 known

Likelihood:

$$\begin{aligned} L(\mu) &\propto \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \right) \\ &\propto e^{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2} \end{aligned}$$

Assume a gaussian prior on μ,

$$\mu \sim N(\mu_0, \sigma_0^2)$$

MLE ON THIS PARAMETERS

The posterior distribution is then

$$\pi(\mu|y) \propto L(\mu)\pi(\mu)$$

Gaussian model; σ^2 known

$$
\begin{aligned}
\pi(\mu|y) &\propto e^{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2} e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} \\
&\propto e^{-\frac{n}{2\sigma^2}\mu^2-\frac{1}{2\sigma_0^2}\mu^2+\frac{\mu\bar{y}n}{\sigma^2}+\frac{\mu\mu_0}{\sigma_0^2}} \\
&\propto e^{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2+\mu\left(\frac{n}{\sigma^2}\bar{y}+\frac{1}{\sigma_0^2}\mu_0\right)} \\
&\propto e^{-\frac{1}{2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}}\left(\mu^2-2\mu\frac{\frac{n}{\sigma^2}\bar{y}+\frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right)}
\end{aligned}
$$

$$
\begin{aligned}
\pi(\mu|y) &\propto L(\mu)\pi(\mu) \\
&\propto e^{-\frac{1}{2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}}\left(\mu-\frac{\frac{n}{\sigma^2}\bar{y}+\frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right)^2}
\end{aligned}
$$

Gaussian model; σ^2 known

$$\begin{aligned} \pi(\mu|y) &\propto L(\mu)\pi(\mu) \\ &\propto e^{-\frac{1}{2(\sigma^*)^2}(\mu-\mu^*)^2} \qquad N(\mu^*, (\sigma^*)^2) \end{aligned}$$

that is, we obtain a gaussian posterior distribution with parameters μ^* and σ^* which are a function of prior distribution's parameters and of the data:

$$\mu^* = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\mu_0\sigma^2 + \bar{y}n\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$(\sigma^*)^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

Gaussian model; σ^2 known

The **posterior mean** is a weighted average of the prior mean and of the ML estimate, where the weights are the reciprocal of the respective variances

$$\mu^* = \mu^*_{n,\sigma_0} = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\frac{1}{V(\bar{y})}\bar{y} + \frac{1}{V(\mu)}\mu_0}{\frac{1}{V(\bar{y})} + \frac{1}{V(\mu)}}$$

- $\mu^*_{n,\sigma_0} \underset{n\to\infty}{\longrightarrow} \bar{y}$ as n grows, the ML estimates weights more
- $\mu^*_{n,\sigma_0} \underset{\sigma_0\to 0}{\longrightarrow} \mu_0$ the more concentrated is the prior distribution, the more the prior mean weights.

It is interesting to write the posterior mean as

$$\mu^* = \mu^*_{n,\sigma_0} = \mu_0 + (\bar{y} - \mu_0)\frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

the posterior mean is the prior mean plus an adjustment toward the sample mean.

$$\mu^* = \mu^*_{n,\sigma_0} = \bar{y} - (\bar{y} - \mu_0)\frac{\sigma^2}{\sigma^2 + n\sigma_0^2}$$

Gaussian model; σ^2 known

The reciprocal of the **posterior variance** is the sum of the reciprocals of the prior variance and the variance of ML estimator

$$(\sigma^*)^2 = (\sigma^*_{n,\sigma_0})^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} = \left(\frac{1}{V(\bar{y})} + \frac{1}{V(\mu)}\right)^{-1}$$

- $\sigma^*_{n,\sigma_0} \underset{n\to\infty}{\longrightarrow} 0$ as n grows the variance of the posterior diminish
- $\sigma^*_{n,\sigma_0} \underset{\sigma_0\to 0}{\longrightarrow} 0$ also if the variance of the prior is reduced the posterior is more concentrated

Application of Bayesian mean principle: estimating the score for Tripadvisor ratings

Tripdavisor uses a formula for calculating and comparing the ratings of restaurants by its users. It derives from using a weighted mean that relies upon the Bayesian idea

The following formula was the base to calculate $W = \frac{Rv+Cm}{v+m}$

where:

- W = weighted rating
- R = average rating (stars) for the restaurant (1 to 5) - *the likelihood*
- v = number of votes/ratings for the restaurant = (votes)
- m = weight given to the prior estimate (in this case, the number of votes for a stable average rating)
- C = the mean vote across the whole pool - *the prior*

W is just the weighted arithmetic mean of R and C with weights v and m.

As the number of ratings surpasses m, the confidence of the average rating surpasses the confidence of the prior knowledge, and the weighted

Bayesian interval estimation and testing

Bayesian interval estimation

The posterior distribution can be summarized by posterior expectation and variance;

- these roughly correspond to point estimate and its standard error in classical inference (although the interpretation is a bit different).
- Given that θ is a random variable, it is natural to think at an analogue of confidence intervals;
- this analogue is called **credibility interval**.
- there is a big difference in interpretation where credibility interval are much more natural and close to common sense.
- most non statisticians actually interpret confidence intervals as if they were credibility intervals.

Classical confidence interval vs credibility interval

Classical interval estimate (confidence interval)

An interval is associated to the sample y such that with a confidence level $1-\alpha$, contains the true value of the parameter.

Interpretation: if N samples were observed and for each of them a $1-\alpha$ confidence interval were obtained, on average $100(1-\alpha)$ of them would contain the true value of the parameter.

An interval is associated to the sample y such that it **contains the true value of the parameter with probability** $1-\alpha$.

Bayesian interval estimate (credibility interval)

A credibility interval for θ is a pair of statistics $L(Y), U(Y) \in \Theta$ such that

$$P(L(Y) \leq \theta \leq U(Y)) \geq 1-\alpha$$

where the probability is with respect to the distribution of θ,

$$P(L(Y) \leq \theta \leq U(Y)) = \int_{L(Y)}^{U(Y)} \pi(\theta|y)d\theta$$

Credibility intervals

Given a distribution for θ, $\pi(\theta|y)$ there is not a unique interval satisfying the condition

$$P(L(Y) \leq \theta \leq U(Y)) = \int_L^U \pi(\theta|y)d\theta = 1 - \alpha$$

the easiest choice is to set L and U equal to the quantiles $\alpha/2$ and $1 - \alpha/2$ of $\pi(\theta|y)$, that is, such that

$$\int_{-\infty}^{L} \pi(\theta|y)d\theta = \int_U^{+\infty} \pi(\theta|y)d\theta = \alpha/2$$

this interval satisfies the condition but is not, generally, the smallest one.

HPD (High Posterior Density) region

A better (smaller) interval is defined as

High posterior density (HPD)

The high posterior density credibility region is a set $C \subset \Theta$ such that

$$P(\theta \in C) = 1 - \alpha$$

and

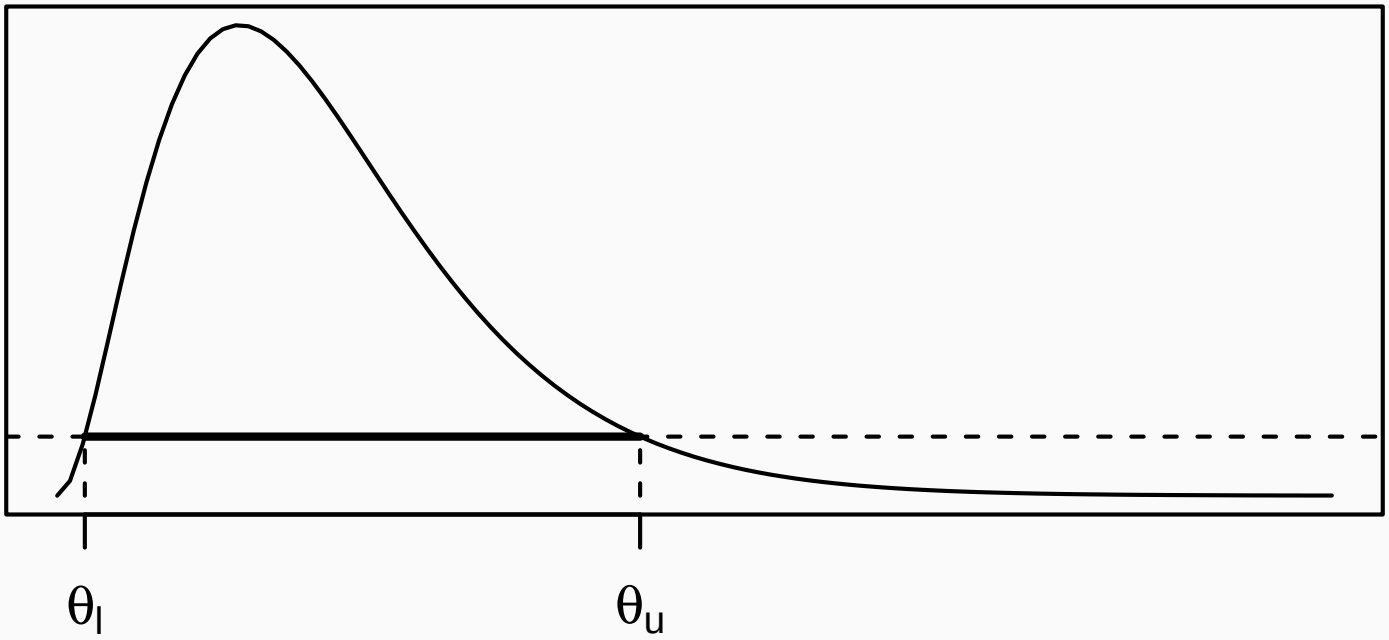
$$\pi(\theta_1|y) > \pi(\theta_2|y)$$

if $\theta_1 \in C$ and $\theta_2 \notin C$

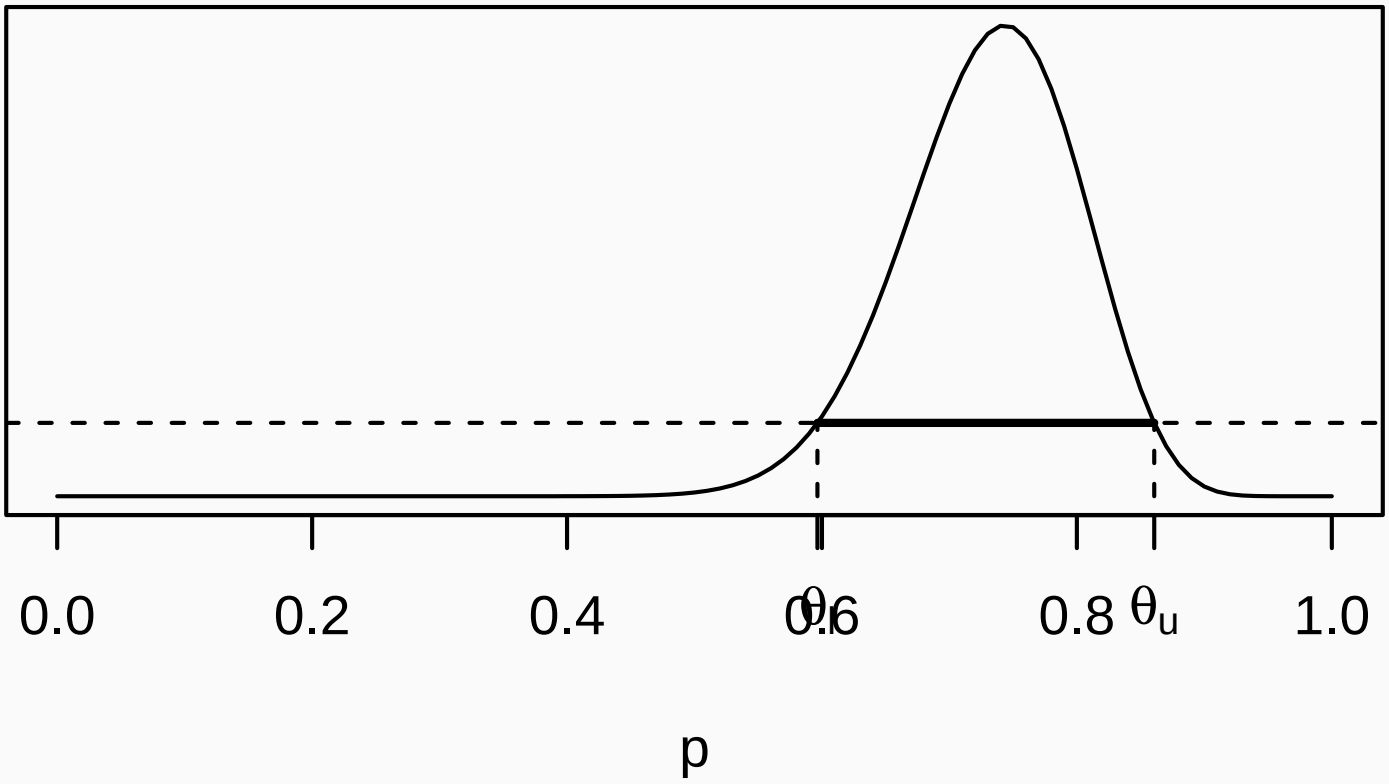
Given $\pi(\theta|y)$ the HPD interval C is obtained including the values of θ corresponding to a higher density

HPD region

HPD region

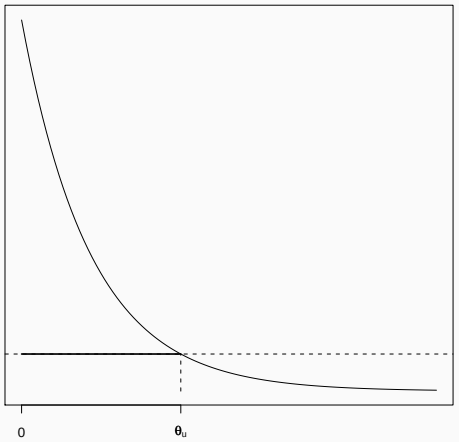


.95 HPD region for p in the seeds example

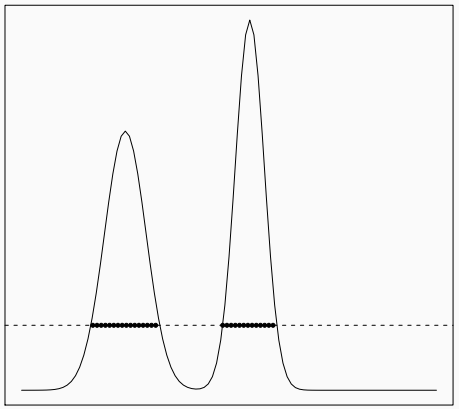


Special cases

monotone posterior



multimodal posterior the HPD region is not necessarily an interval but can be the union of disjoint intervals



Finding the HPD region

For a unimodal posterior (not necessarily symmetric) we may use an algorithm to find the interval:

start from $k_m = 0$, $k_M = \max_\theta \pi(\theta|y)$ then at step i

1. $k_i = (k_m + k_M)/2$
2. determine $C = \{\theta|\pi(\theta|y) > k_i\}$
3. compute $I = \int_C \pi(\theta|y)d\theta$

- if $I < 1 - \alpha$ $\quad$ $k_m \leftarrow k_i$ (shorter interval) return to 1
- if $I > 1 - \alpha$ $\quad$ $k_M \leftarrow k_i$ (longer interval), return to 1
- if $I = 1 - \alpha$ $\quad$ STOP $\quad$ C is the solution

Hypoteses testing

Suppose you want to test the Hypothesis

$H_0 : \theta \in \Theta_0; \qquad H_1 : \theta \in \Theta_1$

Θ_0 and Θ_1 form a partition of the parameter space

The beliefs about the two hypotheses are summarized by the posterior odds ratio

$$\frac{p_0}{p_1} = \frac{P(\theta \in \Theta_0 | y)}{P(\theta \in \Theta_1 | y)} = \frac{\int_{\Theta_0} \pi(\theta|y) d\theta}{\int_{\Theta_1} \pi(\theta|y) d\theta}$$

A measure of the evidence provided by the data in support of H_0 is the **Bayes factor**

$$BF = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{p_0/p_1}{\pi_0/\pi_1}$$

Where π_0 and π_1 are respectively $\int_{\Theta_0} \pi(\theta) d\theta$ and $\int_{\Theta_1} \pi(\theta) d\theta$ the probability of the two hypotheses prior observing the data. Note that you can then evaluate the posterior probability that the null hypothesis is true

$p_o = \frac{\pi_0 BF}{\pi_0 BF + 1 - \pi_0}$

Selecting the prior

The prior distribution

The idea of using prior knowledge in Bayesian statistics is a critical issue and it is a new element in comparison with classical statistics.

Formally this knowledge is introduced by specifying a prior distribution that includes information other than what is directly observed in the process of inference.

The most common concerns are about the use of subjective probability in deriving the prior (for instance, by using experts' opinions for eliciting the prior).

For this reasons some relevant topics in Bayesian statistics refer to:

1. the choice of prior distributions that are diffuse (non informative) in order to give more (or exclusively) weight to experimental data or to obtain results that are consistent with results from likelihood based inference
2. the analysis of the sensitivity of the inference to the alternative choice of the prior (Bayesian robustness)

Objections on the use of prior distributions

One (non-Bayesian statistician) could argue that if I specify a subjective prior distribution, since I can chose any distribution, I can also modify the result and obtain whatever conclusion I want. The result could then be manipulated it is subjective and hence not scientific.

Counter-objections include

- classical procedures are also subjective, for example in the specification of the model;
- the relevance of the prior distribution is limited and tends to vanish if the sample size increases;
- actually, the information conveyed by the data would outweigh the information in the prior for any reasonable specification;
- a possible compromise is to use standard priors which do not involve personal (subjective) opinions.

Coniugacy

Note that in the two examples considered above prior and posterior distribution have the same functional form

For the seeds example the posterior distribution is a Beta like the prior as well as for the Normal mean example

Likelihood $L(\theta; y)$	Prior $\pi(\theta)$	Posterior $\pi(\theta \mid y)$
Binomial	$Beta(\alpha, \beta)$	$Beta(\alpha + \sum_i y_i, \beta + n - \sum_i y_i)$
Normal	$N(\mu, \sigma^2)$	$N(\mu^*, (\sigma^*)^2)$

This property relates the family of the prior distribution with the likelihood and is called **conjugacy**. Example of conjugate families are:

- Beta prior and Binomial likelihood
- Normal and Normal
- Gamma prior and Poisson likelihood

Use of conjugacy would lead to select the prior in order to make computation of the posterior easy and strightforward. None the less, conjugacy could not be the best solution to reflect real prior knowledge about the parameter. Moreover for more complex models colud be impossible to find a conjugate model.

Non informative priors

The prior distribution is meant to reflect the opinion of the researcher prior to observing any data. What if there is no opinion? (Whether this is realistic is disputable.)

This is a relevant issue and a possible answer to the objection that the results of inference should not depend on subjective opinions.

It has then been proposed to use 'standard' distributions which, in some sense, bring no (ore very limited) information on the parameter.

An intuitive solution is to assume $\pi(\theta) \propto k$ so that no values of θ are privileged (principle of insufficient reason).

- Strictly speaking, this is admissible only if the parameter space is limited.
- If the parameter space is not limited a constant has an infinite integral and so is not a probability distribution.
- It is possible however, that a proper posterior distribution is obtained even starting from an improper prior. If this is the case, the inference is valid.

Jeffreys' prior

The non informative nature of the uniform distribution is disputable

- Let
$$\pi(\theta) \propto k$$
- consider the reparametrization $\psi = \psi(\theta)$, then
$$\pi(\psi) = \pi(\theta^{-1}(\psi)) \left| \frac{d\theta}{d\psi} \right|$$
which is not uniform in general.
- that is, assuming that uniform means non informative, by specifying a uniform distribution for the parameter θ, we are specifying instead an informative prior on its transform $\psi = \psi(\theta)$.

Jeffreys' prior

The above issue may be overcome by posing

$$\pi(\theta) = \sqrt{\det H(\theta)}$$

where H is the information matrix, that is, the matrix with (i, j) element

$$[H(\theta)]_{ij} = -E\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\theta; y)\right)$$

then, for any parametrization $\psi = \psi(\theta)$

$$\pi(\psi) = \sqrt{\det H(\psi)} = \sqrt{\det H(\theta)} \left|\det\left(\frac{d\theta}{d\psi}\right)\right|$$

Consider, for instance, a Binomial experiment, so the log-likelihood is

$$l(\theta) = y \log\theta + (n - y)\log(1 - \theta)$$

then

$$[H(\theta)] = -E\left(\frac{d^2}{d\theta^2} l(\theta; y)\right) = \frac{n}{\theta(1 - \theta)}$$

the Jeffreys' prior is then a Beta$(1/2, 1/2)$

$$\pi(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$$

Bayes computation

Bayes computation

To answer the basic questions of statistical inference we need to know the posterior

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

The principal practical challenge is that $\int_\Theta f(y|\theta)\pi(\theta)d\theta$ is usually intractable for many interesting models and also quantity related to $\pi(\theta|y)$ of direct interest for summarizing inference (such as mean, median, percentiles, probabilities) cannot be evaluated.

There are then two main strategies to overcome the problem:

- approximate the integrals by numerical methods
- find a way to get a (simulated) sample from $\pi(\theta|y)$ without requiring evaluation of the integrals.
- finding a function which provides a good approximation of the posterior

The second strategy is based on the fact that if we simulating from a density is as good as being able to evaluate the density, and sometimes better. This is achieved mainly by Monte Carlo Markov Chain methods.

Monte Carlo Markov Chain - MCMC

Monte Carlo Markov Chain (MCMC) methods simulate values from a Markov chain whose stationary distribution is exactly the posterior distribution of interest.

Once a sample of simulate values is given this can be used to evaluate all the quantities of interest for Bayesian inference

Two are the main algorithms to obtain this sample of simulate values

- Metropolis-Hastings algorithm
- Gibbs sampling

Variational inference

- **Variational inference** is a method from machine learning that approximates probability densities through optimization. The idea is to use this approach to approximate the posterior distribution
- It has been used in many applications with a complex parameters space (the most notable is topic modelling) and tends to be faster than methods based on sampling fromm the posterior distribution, such as Markov chain Monte Carlo.
- The idea behind variational inference is to first posit a family of densities over the parameter space and then to find the member of that family which is close to the target. Closeness can be measured, for instance, by Kullback-Leibler divergence.