Advanced Probabilistic Modelling

Unidimensional Data

Irantzu Barrio Beraza

Contents

1. Probability as a tool for quantifying uncertainty

2. The Bayesian learning process

3. Statistics from a Bayesian perspective

    Introduction

    Likelihood information

    Prior distributions

    Posterior distributions

    Predictive distributions

    Bayesian Inference

    Bayesian modeling of uniparametric distributions

    Bayesian modeling of multiparametric distributions

Probability as a tool for quantifying uncertainty

Probalibilty

Probability allows us to quantify the uncertainty of things.

- What is the probability of getting two heads in 10 throws?
- What is the probability that the average height of a person is greater than 1.75m? less than 1.75m?
- What is the probability that we reject the null hypothesis that a person's mean height is greater or equal than 1.75m?
- What is the probability that he/she will pass this subject on the first exam?

It is a fundamental tool for Inference and Statistical Prediction, but also for Statistical Modeling.

Interpretation of probability in the **classical** context (I)

- The results are equally likely to be (equiprobability)

$$P(E) = \frac{\#\ \text{favorable cases}}{\#\ \text{posible cases}}$$

- It helps us to answer questions such as what is the probability of getting heads when flipping a coin:

  Let E be the event "To get heads when flipping a coin"

$$P(E) = \frac{\#\ \text{favorable cases}}{\#\ \text{posible cases}} = \frac{1}{2}$$

Interpretation of probability in the **classical** context (II)

- In a similar way we can obtain the probability of getting two faces in 10 throws:

  Let X be $\#$ (number) of heads in 10 throws:

$$P(X=2) = \frac{\#\text{ favorable cases}}{\#\text{ posible cases}} = \frac{\binom{10}{2}}{2^{10}} \overset{\text{Binamial distribution}}{=} 0.043945$$

Interpretation of probability in the **classical** context (III)

- Knowing the probability of one head, we can calculate the probability of two heads in 10 throws using also that the number of successes X in a fixed number of independent experiments is distributed as a binomial:

  $X \sim \text{Binomial}(n = 10, p = 1/2)$

$$P(X = 2) = \binom{10}{2}(1/2)^2(1/2)^8 = 0.043945$$

```r
dbinom(2,10,0.5)
```

```
## [1] 0.04394531
```

Interpretation of **frequentist** probability (I)

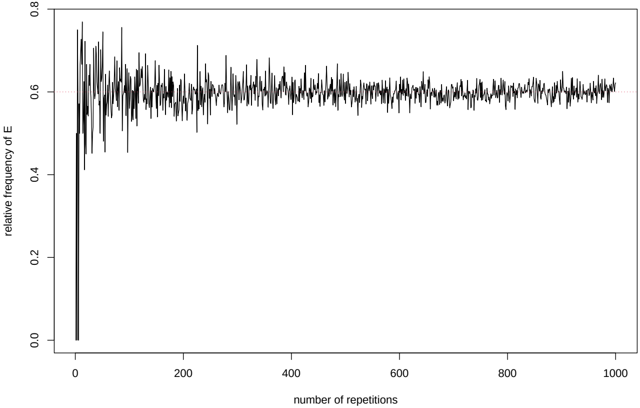
- Experimental context: successive repetitions of the same experiment with random results.

$$P(E) = \frac{\#\text{ times that occurs } E}{\#\text{ number of repetitions}}$$

- Probability of a remote control falling face up? upside down?
  - We cannot use the ratio of favorable and possible cases because there is no equiprobability.
  - But we can repeat many times and approximate that probability by the number of times the thing we want happens divided by the number of times the experiment has been performed.
- We can approximate the probability of getting heads if the coin is tricked and we do not know what the probability is.
- We can even assess whether a coin, a casino, etc., are tricked.

Interpretation of **frequentist** probability (II)

```
E <-  vector(l=1000)
for(s in 1:1000){
  x <- rbinom(1,s, prob=0.6)
  E[s] <- x/s
}
```

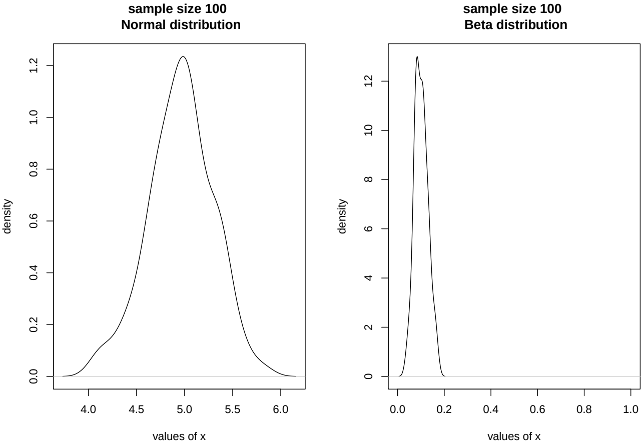


Estimating probability from observed data

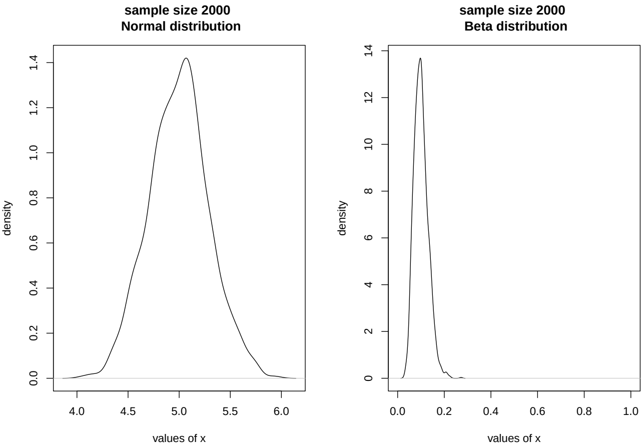
```
par(mfrow=c(1,2))
x1 <- rnorm(100, mean=5, sd=0.3)
plot(density(x1),  xlab="values of x", ylab = "density",
     main="sample size 100 \n Normal distribution")

x2 <- rbeta(100,10,90)
plot(density(x2),  xlab="values of x", ylab = "density",
     main="sample size 100 \n Beta distribution", xlim=c(0,1))
```

Estimating probability from observed data



Estimating probability from observed data



Subjective interpretation of probability (I)

- What to do when it is not possible to repeat the experiment many times?
- For example, how to assign probabilities in management contexts when decisions must be made about single-occurrence events? What is the probability ...
  - ... of passing this course in the first exam?
  - ... that there will be an earthquake tomorrow?
  - ... of a sports team winning a competition?
  - ... of a bank merger between Santander and BBVA?
- $P(E) =$ Subjective valuation of E occurring.

Subjective interpretation of probability (II)

- How to carry out the subjective assessment?
- The probability of an event is the personal belief about the possibility of occurrence of that event. It can be assigned by introspection.
- In reality we all understand the concept “probable event”.
- However, some think that this option is not applicable, since it is not possible to have a sequence of independent repetitions, except in an imaginary sense (subjective).
- We are going to use this type of probability a lot.

The Bayesian learning process

Fundamentals of Bayesian Methodology

- Interpretation of the Bayes Theorem:

$$P(H_i|B) = \frac{P(H_i)P(B|H_i)}{\sum_{j=1}^{k} P(H_j)P(B|H_j)} \quad \text{= P(B)}$$

  - H_i possible event
  - $P(H_i)$ prior probability of such event
  - $P(B|H_i)$ sample information about what happened
  - $P(H_i|B)$ h_i's updated knowledge with information on B

- Bayesian Statistics:

$$\text{Posterior Information} = \left\{ \begin{array}{c} \text{Prior information} \\ + \\ \text{Sample information.} \end{array} \right.$$

Bayesian learning process (I)

- Reasoning in terms of probability
  - about the observable and variable in the sampling,
  - but also about the unknown and unobservable.
- In other words, we measure uncertainty with probability: what we know we express with *probability distributions*.
- Learning process:
  - Construction of the joint distribution of the uncertain elements of the problem:
    - first the observable, the **likelihood**,
    - and then the unknown through the prior (or initial) distribution over the unknown.
  - Use of Bayes' Theorem to update information about the unknown using what is known.

Bayesian learning process (II)

- Constructing the joint distribution of the uncertain elements of the problem:
  - the sampling information expressed through the **likelihood** $l(\theta) = P(\mathbf{x}|\theta)$

    in pur example, theta=p, in the normal distribution: mu and sigma
    x is my sampling data
  - the **prior** (initial) **distribution** of the parameter θ which we denote by $P(\theta)$,

    chosing the prior will be the difficult part
  - all together:

$$P(\mathbf{x}, \theta) = P(\theta)P(\mathbf{x}|\theta).$$

- Obtaining the posterior distribution of the parameter via Bayes' Theorem:

$$P(\theta|\mathbf{x}) = \frac{P(\mathbf{x}, \theta)}{P(\mathbf{x})} = \frac{P(\theta)P(\mathbf{x}|\theta)}{P(\mathbf{x})} = \frac{P(\theta)P(\mathbf{x}|\theta)}{\int P(\theta)P(\mathbf{x}|\theta)d\theta}$$

  Since $P(\mathbf{x})$ does not depend on θ:

$$P(\theta|\mathbf{x}) \overset{\text{proportional}}{\propto} P(\theta) \times P(\mathbf{x}|\theta).$$

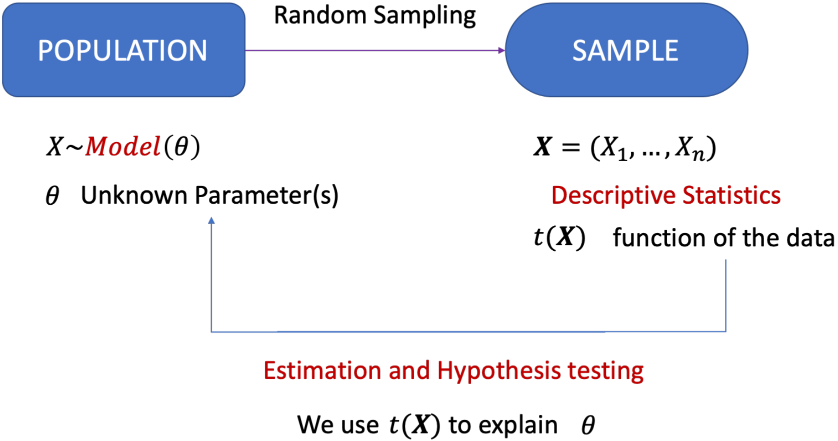
Statistics from a Bayesian perspective

Introduction

Statistics from a Bayesian perspective

- The concepts and methods of statistics allow us to:
  - describe variability, model the situation to be analyzed,
  - to plan the research taking into account this variability,
  - analyze the data to extract the maximum information from them,
  - as well as determine the reliability of the conclusions.
- Three main areas:
  - Statistical modeling,
  - Descriptive statistics, and
  - Statistical inference:
    - estimation, point or interval estimation,
    - and hypothesis testing.

Areas of Statistics



POPULATION

Random Sampling

SAMPLE

$X \sim Model(\theta)$

θ Unknown Parameter(s)

$\boldsymbol{X} = (X_1, \ldots, X_n)$

Descriptive Statistics

$t(\boldsymbol{X})$ function of the data

Estimation and Hypothesis testing

We use $t(\boldsymbol{X})$ to explain θ

Elements of a statistical analysis

- **Parameter**: population characteristic of interest, unknown, on which to infer from observed data.
  - Proportion (p) of people with liver disease, $Z \sim Bin(p)$
  - mean value (μ) of the price of a car, $X \sim N(\mu, \sigma)$
- **Data**: observations of the characteristic of interest on a sector (sample) of the population (realizations of the random variable)
  - whether or not you have a liver condition, $z_1 = 1$, $z_2 = 0$, ...
  - prices of cars obtained at a car show, $x_1 = 12500$, $x_2 = 11000$, ...
- **Likelihood**: obtained from the probabilistic model assumed on the data, it gives an idea about the most (and least) plausible values of the parameters.

Limitations of the frequentist approach (I)

- In the evaluation of the quality of a product, groups of people to whom the product is shown are often used. We want to test whether the proportion of people who rate the product positively p is greater than 0.8.
- We will model the problem as a bernoulli model:
  - Y represents whether a person values positively the product $(Y = 1)$ or not $(Y = 0)$.
  - $Y \sim Ber(p)$ a Bernoulli random variable.
  - p: Bernoulli parameter, the proportion of people who rate the product positively.
  - The data would be the realization of a random sample $(Y_1, \ldots, Y_n)$
  - If, for example, we analyze 10 people we can obtain $y_1 = 0, y_2 = 1, y_3 = 1, y_4 = 1, y_5 = 1, y_6 = 0, y_7 = 1, y_8 = 1, y_9 = 1, y_{10} = 0$

Limitations of the frequentist approach (II)

- To test whether the proportion of people who rate the product positively is greater than 0.8 we perform a hypothesis test (statistical inference):

$$H_0 : p \leq 0.8$$
$$H_1 : p > 0.8.$$

- If 36 out of 40 people rate the product positively (this is the data), using the classic test based on the sampling distribution of the maximum likelihood statistic $\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n} = 36/40 = 0.9$, we obtain a p-value of 0.083, which **does NOT allow us to conclude whether or not the proportion of people is greater than 0.8**, although the confidence interval (at 95%) that we constructed for p is (0.78;1.00).

Limitations of the frequentist approach (III)

```r
prop.test( x = 36, # number of people who value positively
           n = 40, # total number of respondents
           p = 0.8,
           alternative = "greater", # Alternative hipotesis
           conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  36 out of 40, null probability 0.8
## X-squared = 1.9141, df = 1, p-value = 0.08326
## alternative hypothesis: true p is greater than 0.8
## 95 percent confidence interval:
##  0.7797152 1.0000000
## sample estimates:
##   p
## 0.9
```

Limitations of the frequentist approach (IV)

- From the frequentist approach we can calculate certain probabilities such as:
  - What is the probability that the MSE will take a value higher than the true value of the parameter? $P(\hat{p} > p)$
  - What is the percentage of times (in the sample) that the confidence interval will "get" the true value of p ?
  - What is the probability that we are wrong in rejecting H_0?

Limitations of the frequentist approach (V)

- But not the ones we are interested in:
  - With what certainty can we guarantee that p is greater than 0.8? The p-value is NOT the probability of H_0.
  - What certainty do we have that p has a value between 0.7 and 0.9?
  - What value do we expect p to have and what certainty in it?
  - If we decide to repeat the experiment with another 50 people, what is the probability that more than 40 will like the product?

Bayesian Approach

- The Bayesian approach is another way of understanding and performing statistics.
- Therefore, when we use Statistics to solve a problem (inference and prediction about the unknown parameters of the model we have proposed) we can do it with the Bayesian or the frequentist approach (or with both and evaluate the differences).
- Much better known is the frequentist approach to perform inference (both estimation and hypothesis testing) and parameter prediction.
- In the first part of this course we are going to see how to model and, above all, once we have the model that describes our situation, how we can estimate and predict its parameters, using Bayesian statistics.

Likelihood information

Proportion of people with Diabetes at UPV/EHU (I)

- Suppose we have the objective of analyzing the percentage of people diagnosed with Diabetes at the University of the Basque Country.
- How do we pose the problem?
- We will model the problem as a bernoulli model:
  > X represents whether a person has $(X = 1)$ or not Diabetes $(X = 0)$.

  > $X \sim Ber(p)$, a Bernoulli random variable.

  > $p:$ parameter of the Bernoulli, the proportion of people with Diabetes

  > The data would be the realization of a random sample $(X_1, ..., X_n)$

  > If, for example, we analyze 50 people, we can obtain

$$x_1 = 0, x_2 = 0, x_3 = 0, \dots, x_{34} = 1, x_{35} = 0, x_{50} = 0$$

Proportion of people with Diabetes at UPV/EHU (II)

- Let's assume that we find 2 diagnosed people out of the total of $n = 50$.

- The **likelihood function** in this case is:

$$P((X_1, ..., X_n) = 2 \text{ diagnosed out of } 50|p) = l(p) \propto p^2(1-p)^{48}$$

- And the maximum likelihood estimator would be:

$$\frac{\partial \text{ log } l(p)}{\partial p} = 0 \rightarrow \hat{p} = \frac{2}{50}$$

- As we have done before, we could perform (classical) inference by intervals and also by means of a hypothesis test on the parameter.

The importance of the Likelihood

- The ultimate goal of inference is to gain information about the unknown parameter(s) based on the available information.
- Although Bayesian inference is best known for incorporating prior information (as we will see below), we must not forget the importance of the other source of information.
- The data are that other source, and the information they contain is expressed through the **likelihood function**.
- The likelihood function gives a measure of how probable each parameter value is once the data are known and fixed.

Prior distributions

Proportion of people with Diabetes at UPV/EHU (III)

- As we have seen before, if we find 2 diagnosed people out of the total of $n = 50$, the likelihood function in this case is:

$$l(p) \propto p^2(1-p)^{48}$$

- From this information we can make inference:
  - point estimate: $\hat{p} = 2/50$
  - confidence interval for p
  - hypothesis testing for p
- But do we know anything about such a parameter before we start?
- Bayesian inference allows us to incorporate (in addition to sample information) information about the parameters before we start.

The Bayesian learning process

- Before taking data, one has beliefs about the value of the proportion (mean, variance...) and one models his or her beliefs in terms of a **prior distribution**
- Different functional forms for a prior distribution can be used
- After data has been observed, one update's one's beliefs about the proportion (mean, variance...) by computing posterior distribution
- We summarize this probability distribution to perform inferences.

The Bayesian learning process

- Constructing the joint distribution of the uncertain elements of the problem
  - > the sampling information expressed through the likelihood $l(\theta) = P(\mathbf{x}|\theta)$,
  - > the **prior distribution** of the parameter that we denote by $P(\theta)$
  - > all together $P(\mathbf{x}, \theta) = P(\mathbf{x}|\theta)P(\theta)$
- Obtaining the **posterior distribution** of the parameter via Bayes' Theorem:

$$P(\theta|\mathbf{x}) = \frac{P(\mathbf{x}, \theta)}{P(\mathbf{x})} = \frac{P(\mathbf{x}|\theta)P(\theta)}{P(\mathbf{x})} = \frac{P(\mathbf{x}|\theta)P(\theta)}{\int P(\mathbf{x}|\theta)P(\theta)d\theta}$$

Since $P(\mathbf{x})$ does not depend on θ:

$$P(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta) \times P(\theta)$$

Proportion of people with Diabetes at UPV/EHU (IV)

- The likelihood function in this case is:

$$l(p) \propto p^2(1-p)^{48}$$

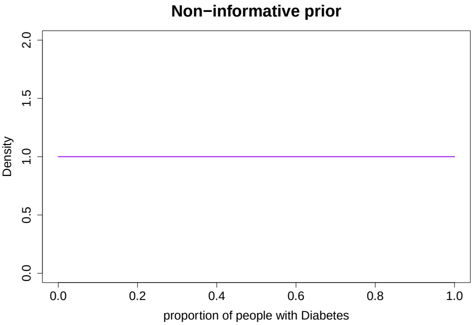
- In order to apply the Bayesian learning process we need to incorporate initial information about the parameter of interest, the proportion p of people with Diabetes in the UPV/EHU.
- What proportion of people have Diabetes in general?
- Do we use what is said on the Internet, official media, up-to-date information? Based on clinical tests? Information based on other diseases?
- How do we incorporate such information?
- What if I don't know anything?

Prior distributions

- Bayesian methods require the establishment of an a priori distribution over the unknown parameters. But how do we get that prior distribution?
- Possibilities:
  - **Objective** prior information, i.e., a priori nothing is known about the parameter of interest:
    - we will need a probability distribution for the parameter that indicates no knowledge or at least little knowledge.
    - An **objective** analysis is expected to provide results that are as objective as (and similar to) those produced by a frequentist analysis.
  - A priori **subjective** information, specified in terms of a probability distribution obtained from probability distribution obtained from:
    - Information from experts in the field.
    - Information from previous experiments.

Proportion of people with Diabetes at UPV/EHU (V)

- If we don't have a clear idea of what the proportion of people with Diabetes at UPV/EHU would be (for whatever reason), we can use a non-informative prior that implies ignorance.
- Since it is a probability (it takes real values between 0 and 1), we can use a distribution that indicates that any interval of values is equally likely: Uniform(0,1).



Non-informative prior distributions

- Very useful in those situations in which previous information is very difficult to quantify or it is not convenient to use.
- Constant distribution in all values of the parameter space (even if improper) [1]
- It is possible to use the Bayesian learning process even when the prior is improper, but it is necessary to check that the resulting posterior is proper (i.e. integrates 1 the density function).

[1] A distribution is improper if it integrates infinity in parameter space.

$$\int_{-\infty}^{\infty} f_X(x)dx = \infty$$

Non-informative a priori Jeffreys distributions

- It is possible to automatically find distributions that are not very informative.
- Jeffreys prior distributions, invariant to transformations

$$P(\theta) \propto [I(\theta)]^{1/2},$$

where $I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\text{ln }P(\mathbf{x}|\theta)\right]$, represents θ's Fisher information and $P(\mathbf{x}|\theta)$ the likelihood function.

Proportion of people with Diabetes at UPV/EHU (VI)

- What if we choose an initial probability distribution that is "similar" to the structure of the likelihood?
- In the Diabetes example, the likelihood is:

$$P(\mathbf{x}|p) = l(p) = \binom{n}{r} p^r (1-p)^{n-r} \propto p^2(1-p)^{48}$$

- What if we choose a prior Beta distribution of parameters a and b?

$$P(p) \propto p^{a-1}(1-p)^{b-1}$$

- By multiplying likelihood and prior, the calculations are simplified

Conjugate prior distributions (I)

- *Definition*: a class $\mathcal{T}$ of a priori distributions is said to be a *conjugate family for* $\mathcal{F}$ (the class of all density functions $P(x|\theta)$ of parameter θ), if the posterior distribution $P(\theta|x)$ is in the class $\mathcal{T}$ for all x of the parameter space and for all a priori distributions of $\mathcal{T}$.
- In other words: the family $\mathcal{T}$ of distributions for θ, $P(\theta)$, is conjugate with respect to the family $\mathcal{F} = \{P(x|\theta)\}$, if the posterior $P(\theta|x) \in \Gamma, \quad \forall P(\cdot) \in \mathcal{T}$ and for all $P(\cdot|\cdot) \in \mathcal{F}$.

Therefore, we obtain a posterior distribution with the same parametric form as the a priori distribution.

Conjugate prior distributions (II)

- *Natural* conjugate families, appear by considering $\mathcal{T}$ as the family with the same functional form as the likelihood.

- Justifications for their use:

  > they simplify calculations in obtaining the posterior,
  > they facilitate the description of the results,
  > they are of great utility in the construction of more complicated models.

- Disadvantages:

  > except in the simplest cases, they are often too rigid to represent the information.

Construction of priors with information about the parameter

- It is possible to incorporate the knowledge we have of the parameter using the parameters of the prior distribution.
- It is sufficient to equate:
  - the value we think the parameter has with the mean of the distribution used as a prior (beta, gamma, etc.);
  - the uncertainty with the variance of this distribution.

Proportion of people with Diabetes at UPV/EHU (VII)

- The mean and variance of a Beta(a, b) distribution are respectively:

si tengo una uestra de 10 con 3 positivos, podemos usar una Beta con a=3, b=7

$$E(X) = \frac{a}{a+b} \quad Var(X) = \frac{ab}{(a+b+1)(a+b)^2}$$

- We match the value that we think the proportion of people with Diabetes at the UPV/EHU is (e.g. 1 out of 4, 0.25) and its uncertainty (e.g. a variance around 0.25 of 0.1).
- That is:

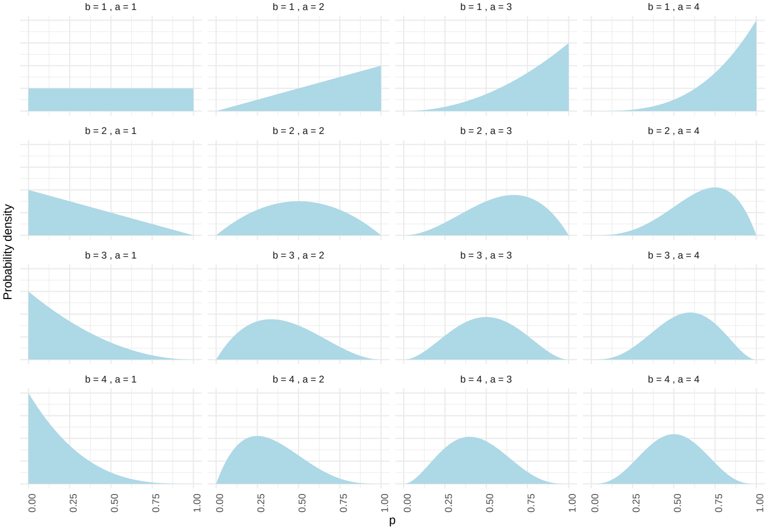
$$0.25 = \frac{a}{a+b} \quad 0.1 = \frac{ab}{(a+b+1)(a+b)^2},$$

resulting a Beta(0.22, 0.66) prior.

Construction of priors based on previous experiments

- Beta conjugate prior distributions allow an interpretation of the prior as an equivalent experiment in which a parameter of a Bernoulli is studied.
- In fact, we can consider a and b as if they were the previously observed data, such that we would have obtained a successes and b failures in a total of $a + b$ trials.
- In the example of the proportion of people with Diabetes at UPV/EHU, we could incorporate as possible information that in another previous experiment we have obtained 10 patients with Diabetes out of a total of 100 people with a Beta(10,90).

Beta distribution



Other possible priors

- Any prior distribution is valid, but in practice they can:
    - complicate calculations,
    - make the interpretation of the posterior less easy.
- Triangular prior distribution: impractical (complex posterior).

Posterior distributions

Posterior distributions

- The application of the *Bayesian learning process* results in the **posterior distribution of the parameters of interest**.

$$P(\theta|\mathbf{x}) = \frac{P(\mathbf{x},\theta)}{P(\mathbf{x})} = \frac{P(\theta)P(\mathbf{x}|\theta)}{P(\mathbf{x})} = \frac{P(\theta)P(\mathbf{x}|\theta)}{\int P(\theta)P(\mathbf{x}|\theta)d\theta}$$

- Depending on the initial distribution used the result will vary, so the prior distribution should accurately reflect our knowledge of the parameter (or our lack of knowledge about it when it is the case).

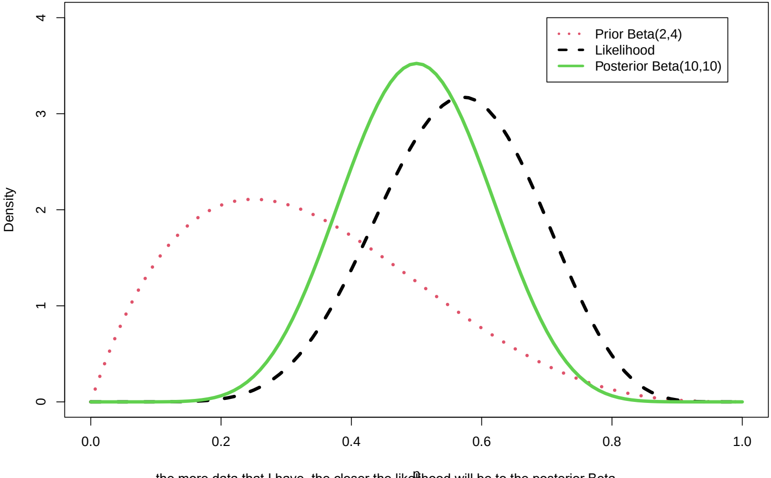
Proportion of people with Diabetes at UPV/EHU (VIII)

- In the Diabetes example, in which we want to estimate p, we are going to analyze different possibilities of both initial knowledge and sample information.

- If our previous knowledge is that 2 out of 6 people have the disease and if we have an experiment that tells us that 8 out of 14 people have it:
  - The likelihood: $P(\mathbf{x}|p) = l(p) \propto p^r(1-p)^{n-r} = p^8(1-p)^6$,
  - Previous knowledge is expressed by a Beta(2,4): $P(p) \propto p^{2-1}(1-p)^{4-1}$,
  - which results in a posterior distribution:

$$\begin{aligned} P(p|\mathbf{x}) &\propto P(\mathbf{x}|p) \times P(p) \propto p^{8+2-1}(1-p)^{6+4-1} \\ p|\mathbf{x} &\sim Beta(10,10) \end{aligned}$$

Proportion of people with Diabetes at UPV/EHU (IX)

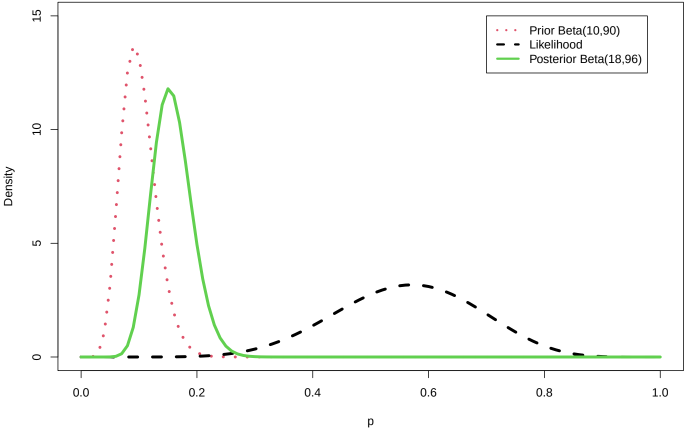


the more data that I have, the closer the likelihood will be to the posterior Beta

Proportion of people with Diabetes at UPV/EHU (X)

We now observe how it changes if we are much more sure of our prior knowledge. If we can say that out of 100 people 10 have the disease, with the same experiment as before:

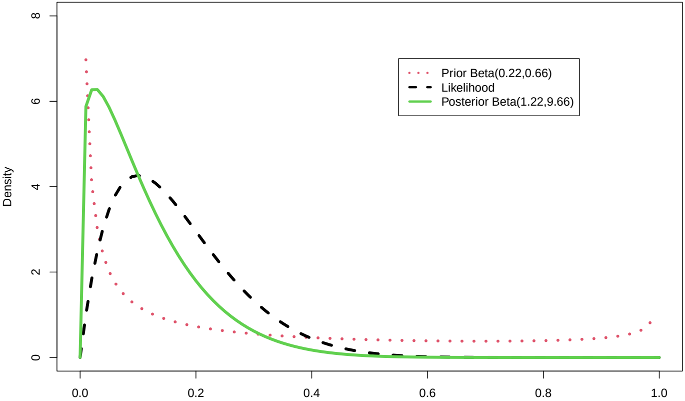
- Prior Beta(10,90) → posterior Beta(18,96).



Proportion of people with Diabetes at UPV/EHU (XI)

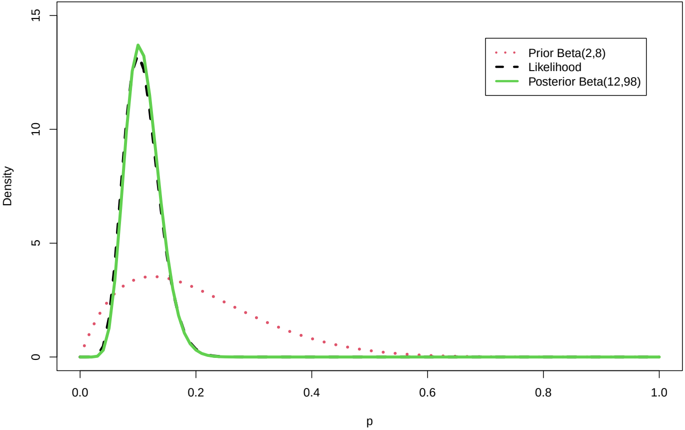
Conversely, the initial knowledge may be very small and the data will command.

- If we have a sample of 10 people of which 1 has Diabetes and our prior knowledge is that the proportion of sick people is 0.25 with a variance of about 0.1:



Proportion of people with Diabetes at UPV/EHU (XII)

- Similarly, if we have a sample of 100 people of which 10 have Diabetes and our previous knowledge is that approximately 2 out of 10 have it:



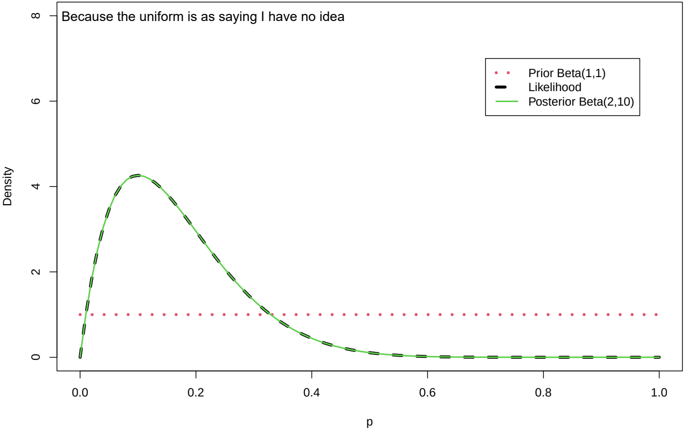
OBJECTIVE Bayesian analysis of a Bernoulli model

In a Bernoulli/Binomial model we can address the lack of knowledge about the parameter p by using:

- an uninformative prior distribution (thinking that any probability interval is equally likely) such as the Uniform(0,1) which is equivalent to a Beta(1,1);
- the Jeffreys minimum informative one, which is a Beta(0.5,0.5)

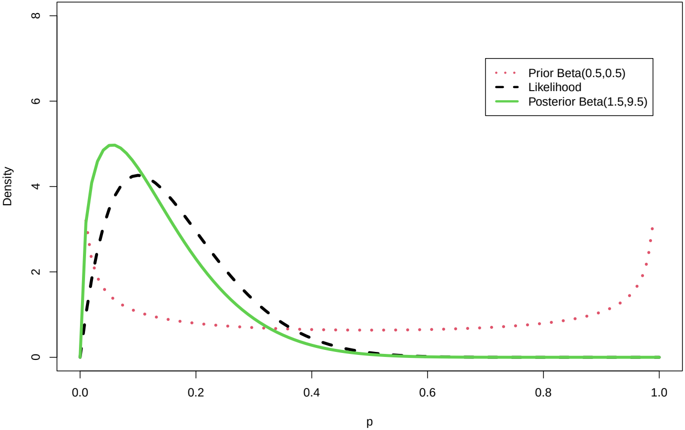
Proportion of people with Diabetes at UPV/EHU (XIII)

- If we use a non-informative one such as Uniform(0,1), which is equivalent to Beta(1,1), note that it would be like saying that we don't know anything about p.
- Then, if we have a sample of 10 people of which 1 has Diabetes :



Proportion of people with Diabetes at UPV/EHU (XIV)

- If we use Jeffreys' prior, Beta(0.5,0.5), which would be like saying that we think we would have 0.5 sick of each, note that we give less information but we give more probability to the extremes.
- Thus, if we have a sample of 10 people of whom 1 has Diabetes



Relation of the posterior with the prior and sample information (I)

- Note that the posterior distribution is a compromise between the sampling information and the prior distribution.
- In fact, the expectation (the mean) of the posterior distribution is a weighted average of:
  - the expectation (mean) of the prior ($E(p) = \frac{a}{a+b}$)
  - the maximum likelihood estimator, $\hat{p} = \frac{r}{n}$.
- Note that if the prior is a $Beta(a, b)$ and in the experiment we have r events in n trials, then the posterior is a $Beta(r + a, n - r + b)$, thus:

$$E(p|\mathbf{x}) = \frac{r+a}{n+a+b} = \frac{n}{n+a+b} \times \hat{p} + \frac{a+b}{n+a+b} \times E(p)$$

Relation of the posterior with the prior and sample information (II)

- And the posterior variance:

$$Var(p|\mathbf{x}) = \frac{(r+a)(n-r+b)}{(n+a+b)^2(n+a+b+1)} = \frac{E(p|\mathbf{x})(1-E(p|\mathbf{x}))}{n+a+b+1}$$

- Note that prior parameters lose influence as the sample size increases. This is consistent with the foundation of statistical inference.

- In fact, if n increases when a and b are fixed, the posterior mean and variance no longer have the influence of the prior:

$$E(p|\mathbf{x}) \sim \hat{p}$$

$$Var(p|\mathbf{x}) \sim \frac{\hat{p}}{n}(1-\hat{p})$$

Exercise

Suppose we are interested in studying the sleep habits of students at a certain school. It seems that doctors recommend a minimum of 8 hours of sleep for an adult person, so the study is posed in terms of finding out the proportion of students who sleep at least 8 hours. We will call this proportion p.

A sample of 27 students is taken so that 11 of them sleep at least 8 hours and the rest do not. We consider to make inferences about the proportion p taking into account the previous information that we have. Suppose we consider an a prior beta distribution. What is the posterior distribution for p?

Sequential incorporation of the information (I)

- One of the most important features of Bayesian inference is the ease with which information can be incorporated **sequentially**.
- It is also a consequence of the way in which information is incorporated in the Bayesian learning process.
- If the posterior distribution $P(\theta|\mathbf{x})$ is constructed from data $\mathbf{x}$, and a second set of data $\mathbf{y}$ distributed **independently** of the first is observed, then:

$$P(\theta|\mathbf{x},\mathbf{y}) \propto P(\theta|\mathbf{x}) \times P(\mathbf{y}|\theta,\mathbf{x}) = P(\theta|\mathbf{x}) \times P(\mathbf{y}|\theta),$$

- then the old posterior distribution $P(\theta|\mathbf{x})$ is now the new prior distribution.

Sequential incorporation of the information (II)

- But more importantly, the resulting new posterior distribution **is the same as if the two data sets had been obtained at the same time** and we used the prior we used first:

$$P(\theta|\mathbf{x},\mathbf{y}) \propto P(\theta) \times P(\mathbf{x},\mathbf{y}|\theta)$$

,

- and it works in the same way if the order of observation is reversed.

Predictive distributions

Prediction in the statistical context

- One of the major uses of a model is to be able to use it for *prediction*
- In a simple linear regresion model, for example, the aim is to predict the response for a certain value of the covariate. Recall that this prediction is:

$$\hat{y}_{pred} = \hat{\beta}_0 + \hat{\beta}_1 x_{pred} \ ,$$

- and that it is possible to evaluate the uncertainty of the prediction by with

$$\hat{y}_{pred} \mp t_{n-1,\alpha/2} \sqrt{\hat{\sigma}\left(1+\frac{1}{n}\right)}.$$

### Proportion of people with Diabetes at UPV/EHU (XV)

- Could we transfer this idea of prediction to the Diabetes problem?

Predictive distributions

- Bayesian statistics considers everything unknown as a random variable, so when predicting we look for the probability distribution of a new realization (under the same conditions) of the variable of interest conditioned to the knowledge we have about the model parameters.
- This distribution will allow us to know which is the most (and least) probable value, and it is called **predictive distribution**
- Predictive distributions are the essential tool for the design and planning of a future experiment.
- We can distinguish two predictive distributions depending on:
  - if we use information about the parameter before the experiment, *a priori predictive distribution*
  - or, if we also use the information provided by the experiment, *a posterior predictive distribution*

Prior predictive distribution

- The *prior predictive distribution* is the distribution of a new realization of the variable of interest (or of any transformation of it expressed in terms of the model parameters) before the experiment is conducted using only the undated information (observed data) on the parameters:
  - > if $P(\theta)$ is the prior distribution that collects the prior information that we have on the parameter that governs the variable of interest, then,
  - > the **prior predictive distribution** of a new X (X_{new}) is

$$m(X_{pred} = \int P(X_{new}|\theta)P(\theta)d\theta.$$

Posterior predictive distribution

- The *posterior predictive distribution* is the distribution of a new realization of the variable of interest (or of any transformation of it expressed in terms of the model parameters) after the experiment is conducted using the already updated information (with the observed data) on the parameters:
  - if $\mathbf{x} = (x_1, \dots, x_n)$ is a realization of a random sample $\mathbf{X} = (X_1, \dots, X_n)$ of a random variable $X \sim F(x|\theta)$, with unknown θ;
  - $P(\theta|\mathbf{x})$ is the posterior distribution

$$P(\theta|\mathbf{x}) = \frac{P(\theta)P(\mathbf{x}|\theta)}{\int P(\theta)P(\mathbf{x}|\theta)d\theta}$$

  - then, the **posterior predictive distribution** of a new X (X_{new}) is

$$m(X_{pred|\mathbf{x}} = \int f(X_{new}|\theta)P(\theta|\mathbf{x})d\theta.$$

Proportion of people with Diabetes at UPV/EHU (XVI)

- Supose we decided to conduct a new experiment in which we observe what is the proportion of people with Diabetes at UPV/EHU, in a new sample of m people.
- If our previous knowledge is that 5 out of 100 people have the disease and if we have an experiment that tells us that 15 out of 700 people have it:
  - the likelihood is: $P(\mathbf{x}|p) = p^{15}(1-p)^{685}$
  - the initial knowledge is expressed by a Beta(5,95), $P(p) p$^{\{5\text{-}1\}(1\text{-}p)}$\{95-1\}, which results in
  - posterior distribution:

$$P(p|\mathbf{x}) \propto P(\mathbf{x}|p) \times P(p) \propto p^{15+5-1}(1-p)^{685+95-1}$$

$$p|\mathbf{x} \sim Beta(20, 780)$$

Proportion of people with Diabetes at UPV/EHU (XVII)

- The distribution that **predicts a priori** the number of people with Diabetes, Z, in a sample of m people is given by[2] :

$$
\begin{aligned}
P(Z=z) &= \int_0^1 P(Z=z|Z\sim Bin(m,p)) \times Beta(5,95)dp = \\
&= \int_0^1 \binom{m}{z} p^z(1-p)^{m-z} \times \frac{p^4(1-p)^{94}}{B(5,95)}dp = \\
&= \binom{m}{z}\frac{1}{B(5,95)}\int_0^1 p^{z+4}(1-p)^{m-z+94}dp = \\
&= \binom{m}{z}\frac{B(z+4,m-z+94)}{B(5,95)}.
\end{aligned}
$$

[2] The Beta function has the form $B(a,b)=\int_o^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, where $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$.

Beta-Binomial distribution

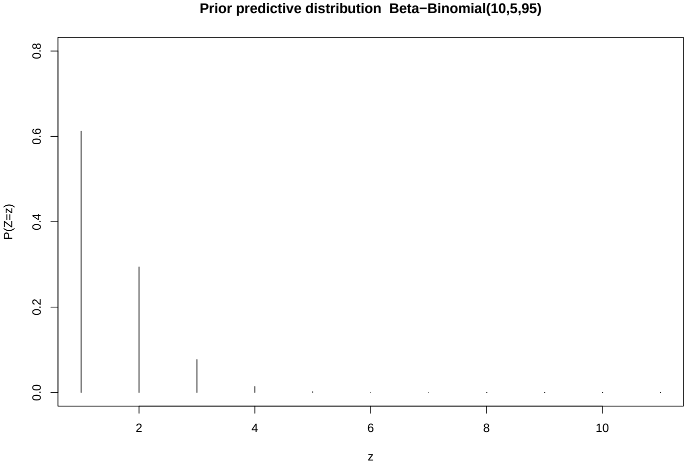
- The previous probability distribution is known as Beta-Binomial:

$$P(Z = z) = \binom{m}{z} \frac{B(z+4, m-z+94)}{B(5,95)} = \\ \sim \text{Beta-Binomial}(m, 5, 95).$$

- The Beta-binomial distribution is implemented in the following R packages (among others): `TailRank`,`extraDistr`, `VGAM`.

Proportion of people with Diabetes at UPV/EHU (XVIII)

- Before we begin with the experiment, the prior predictive distribution of the number of people with Diabetes in a sample of size 10, Z, is distributed as a Beta-Binomial(10,5,95):



## Proportion of people with Diabetes at UPV/EHU (XIX)

- The distribution that a posteriori predicts the number of people with Diaberes, Z, in a sample of m people is given by:

$$\begin{aligned}
P(Z=z) &= \int_0^1 P(Z=z|Z\sim \text{Bin}(m,p)) \times \text{Beta}(20,780)dp = \\
&= \int_0^1 \binom{m}{z} p^z(1-p)^{m-z} \times \frac{p^{20-1}(1-p)^{780-1}}{B(20,780)}dp = \\
&= \binom{m}{z}\frac{1}{B(20,780)}\int_0^1 p^{z+20-1}(1-p)^{m-z+780-1}dp = \\
&= \binom{m}{z}\frac{B(z+19,m-z+779)}{B(20,780)} \\
&\sim \text{Beta-Binomial}(m,20,780).
\end{aligned}$$

### Proportion of people with Diabetes at UPV/EHU (XX)

- After conducting the experiment in which it is observed that out

Bayesian Inference

Inference

- Making inference is equivalent to making a detailed description of the posterior distribution:
    - through a graphical representation of its density (or of its probability function): not very effective in large dimensions;
    - finding a representative value of the posterior distribution such as the generalized maximum likelihood (the mode of the posterior), the posterior expectation, the median of the posterior, can be considered as a way to perform point estimation;
    - finding regions (credible intervals and regions of maximum density) of the posterior is a way of interval estimation;
    - calculating the probabilities of the parameter taking values from a set is a way of hypothesis testing.

Point estimation

- To perform the point estimate we look for the most common location measures of the posterior distribution
  - Generalized maximum likelihood: **mode of the posterior distribution**.
    - Simplicity of calculation.
    - It does not require the normalization constant.
    - It coincides with the frequentist maximum likelihood estimator if a constant prior is used.
    - But it completely ignores distribution queues.
  - Posterior Expectation: **mean of the posterior distribution**.
    - Bayesian estimator par excellence, the most widely used.
    - Good properties.
    - But it depends heavily on the distribution queues.
  - **Posterior Median**
    - Avoids the problem of queues.
    - More complicated calculation.

Precision of the point estimation

- To indicate the **precision** of the estimation made we can use measures of **dispersion** of the posterior distribution such as variance, standard deviation, etc.
- To measure the distance between a particular estimate δ and the parameter to be estimated:
  - Expected cuadratic error: $E[(\theta - \delta)^2]$ coincides with $Var(\theta|\mathbf{x})$ when $\delta = E[\theta]$
  - Its square root is used as a measure of the standard error of the estimate.
- The posterior expectation $E(\theta|\mathbf{x})$ minimizes $E[(\theta - \delta)^2]$ for all possible δ, and is therefore the estimate with the smallest standard error.

$$E[(\theta - \delta)^2] = Var(\theta|\mathbf{x}) + (E(\theta|\mathbf{x}) - \delta)^2$$

- The usual practice is to use $E(\theta|\mathbf{x})$ to estimate θ and to use $\sqrt{Var(\theta|\mathbf{x})}$ as the standard error of the estimation.

Credible intervals

- The description of the posterior distribution through the probability that the parameter falls in a certain region of interest is the equivalence (but with a real probability interpretation) of the classical confidence intervals.
- A **Credible Interval** is defined as the $100(1-\alpha)\%$ for the parameter θ to the subset C of the parameter space that satisfies:

$$1-\alpha \leq P(C|\mathbf{x}) = \int_C P(\theta|\mathbf{x})d\theta$$

in the continuous case and $1-\alpha \leq \sum_{\theta \in C} P(\theta|\mathbf{x})$ in the discrete case.

- Note that, in practice, for one-dimensional distributions this translates into using intervals properly defined by quantiles of the posterior distribution easily calculable with R.

Hypothesis contrast

- To test hypotheses we use the posterior distribution to calculate the probabilities of both hypotheses.
- Specifically, in uniparametric models, hypotheses are set to contrast

$$
\begin{aligned}
H_0 &: \theta \in \Theta_0 \\
H_1 &: \theta \in \Theta_1
\end{aligned}
$$

- and the objective being to find out the correct hypothesis, the solution is marked by
  - Calculate $\alpha_0 = P(\Theta_0|\mathbf{x})$ $\alpha_1 = P(\Theta_1|\mathbf{x})$
  - Solve according to the rule: reject H_0 when $\alpha_0 < \alpha_1$.

Hypothesis contrast

- Other alternatives:
  - Prior Odds $= f_1/f_0$ (where f_i is the prior probability of Θ_i). It gives us an idea of how the relative plausibility of the hypotheses BEFORE observing the data.
  - Posterior Odds $= \alpha_1/\alpha_0$, which informs us of the relative plausibility of the hypotheses after observing the data.

Bayes Factor

- Conjugating both odds, the **Bayes Factor** is defined as a measure of the evidence for a hypothesis based on the data:

$$B_{10} = \frac{\alpha_1/\alpha_0}{f_1/f_0}$$

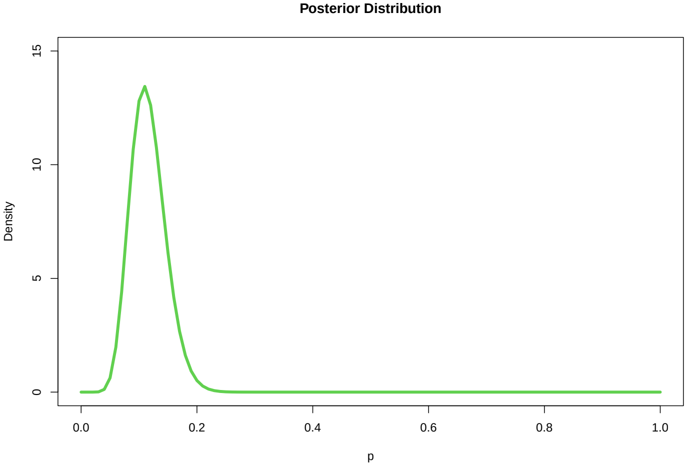
- B_{10} gives us an idea of the evidence for H_1 with respect to H_0. Large values of B_{10} indicate greater support for H_1. Similarly $B_{01} = 1/B_{10}$ gives us an idea of more evidence for H_0 with respect to H_1.
- So for example, if $B_{10} = 5$, what it indicates is that the data are 5 times more likely under H_1 than under H_0.

Bayes Factor (II)

- Although there are no thresholds for accepting or rejecting hypotheses, there have been several attempts that may allow an automatic interpretation.
- The best known is Jeffreys:
  - starting from the value 1 which would be something similar to both hypotheses are equally supported,
  - Bayes Factor between 1 and 3 provide weak evidence of H_1, while if they are between 3 and 10 we speak of moderate evidence of H_1. Above 10 there is already strong evidence of H_1.
  - Clearly their inverses give us similar information about H_0: between 1 and $1/3$ weak evidence of H_0, between $1/3$ and $1/10$ moderate evidence of H_0, and from $1/10$ strong evidence of H_0.

Proportion of people with Diabetes at UPV/EHU (XXI)

If we have that out of 100 people 10 have Diabetes and our prior knowledge is that 2 out of 10 have it, the posterior is a Beta(13,99).



Proportion of people with Diabetes at UPV/EHU (XXII)

- We can get the estimation of p by means of:
  - the posterior expectation $E(p|\mathbf{x}) = \frac{13}{13+99} = 0.1161$;
  - the posterior median `qbeta(0.5,13,99)`= 0.1138;
  - the posterior mode (we maximize the kernel of a beta(13,99), which is $p^{12}(1-p)^{98}$) = 0.1091

```
kernel_beta <- function(x){x^12*(1-x)^98}
optimize(kernel_beta, interval=c(0, 1),
         maximum=TRUE)$maximum
```

- A 95% credible region:

$C =$ (`qbeta(0.025, 13, 99)`, `qbeta(0.975, 13, 99)`) $=$ (0.0639,0.1812).

Proportion of people with Diabetes at UPV/EHU (XXIII)

To check if the proportion is greater than 0.15 (more than 15% of people with Diabetes),

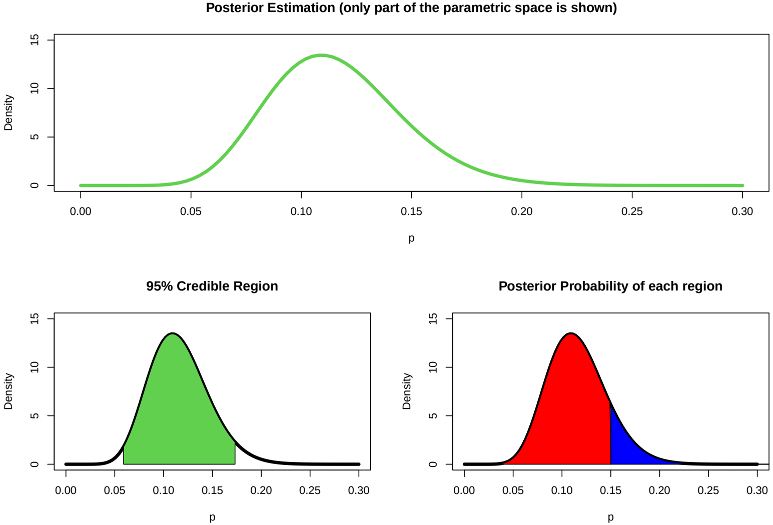
$$\begin{aligned} H_0: &\quad p \geq 0.15 \\ H_1: &\quad p < 0.15 \end{aligned}$$

Since $\alpha_0 = P(p > 0.15|\mathbf{x}) = 1-$ `pbeta(0.15, 13, 99)` $=$ 0.1330 and $\alpha_1 = P(p < 0.15|\mathbf{x}) =$ `pbeta(0.15, 13, 99)` $=$ 0.8670, we reject H_0.

Proportion of people with Diabetes at UPV/EHU (XXIV)

- The posterior odds are 6.52, we are therefore 6.52 times more in favour of the assumption that the percentage of people with Diabetes is less than 0.15 after having observed the data.
- The prior odds are marked by the a priori probabilities of both hypotheses: $f_0 = P(p \geq 0.15) = 1-$ `pbeta (0.15, 3, 9)` $=$ 0.7788 and $f_1 = P(p < 0.15) =$ `pbeta(0.15, 3, 9)` $=$ 0.2212.
- Thus: $f_1/f_0 = 0.2840$, we are more sure that the percentage of people with Diabetes is less than 0.15 after looking at the data.
- The Bayes Factor is: $BF_{10} = \frac{\alpha_1/\alpha_0}{f_1/f_0} = 22.96$. So we have strong evidence for H_1 with respect to H_0.

Proportion of people with Diabetes at UPV/EHU (XXV)



Posterior Estimation (only part of the parametric space is shown)

95% Credible Region

Posterior Probability of each region

Exercise (I)

If our previous knowledge is that out of 100 people 10 have the disease. And we have conducted an experiment in which 8 out of 14 people have the disease (check in previous example which is the posterior distribution).

- Estimate the probability of having the disease
- Obtain a 95% credible region
- Check if the probability is greater or equal than 0.15
- Plot the results obtained.

Exercise (II)

Now consider a prior distribution different from those seen so far (but providing an a posterior beta). Think of a mean and variance for the proportion and obtain the parameters of the beta distribution. Repeat the previous exercise.

Help:

```
estBetaParams <- function(mu, var) {
    alpha <- ((1 - mu) / var - 1 / mu) * mu ^ 2
    beta <- alpha * (1 / mu - 1)
    return(params = list(alpha = alpha, beta = beta))
}
```

Bayesian modeling of uniparametric distributions

The Bayesian Learning Process

- Let $\mathbf{x} = (x_1, \dots, x_n)$ be a realization of a random sample $\mathbf{X} = (X_1, \dots, X_n)$ of a random variable X.
- Consider a model over X $X \sim F(x, \theta)$ with unknown θ.
- To make inference about the parameter θ
  - Data information via the Likelihood: $P(\mathbf{x}|\theta)$.
  - The prior distribution of the parameter: $P(\theta)$
  - The posterior distribution of the parameter via Bayes' Theorem

Bayesian analysis of a Poisson model (I)

- Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a realization of a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ of a random variable $Po(\lambda)$.
- The likelihood function:

$$p(\mathbf{x}|\lambda) = \prod_{i=1}^{n} \left[\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right] = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} [x_i!]} \propto \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}$$

Bayesian analysis of a Poisson model (I)

- Observing its shape, we can consider as a prior distribution the conjugate [3] distribution $\lambda \sim Gamma(a, b)$:

  $$P(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-\lambda b} \propto \lambda^{a-1} e^{-\lambda b},$$

  then, the posterior distribution:

  $$\begin{aligned} P(\lambda|\mathbf{x}) &\propto P(\mathbf{x}|\lambda) \times P(\lambda) \\ &\propto \lambda^{\sum_{i=1}^n x_i} e^{-n\lambda} \lambda^{a-1} e^{-\lambda b} \\ &\propto \lambda^{\sum_{i=1}^n x_i + a - 1} e^{-(n+b)\lambda} \end{aligned}$$

  $$\lambda|\mathbf{x} \sim Gamma(\sum_{i=1}^n x_i + a, n + b).$$

[3] When the functions $p(\mathbf{x}|\theta)$ and $p(\theta)$ are combined in such a way that the posterior distribution belongs to the same family (has the same shape) as the initial distribution, then we say that $p(\theta)$ is conjugate for $p(\mathbf{x}|\theta)$

Bayesian analysis for the mean of a normal model with known variance (I)

Suppose we have some data following a normal distribution of **unknown mean** μ. That is, $\mathbf{x} = (x_1, \ldots, x_n)$ a realization of a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ with $X_i \sim N(\mu, \sigma)$, for $i = 1, \ldots, n$, where σ^2 is the **known variance**.

- The likelihood function:

$$P(\mathbf{x}|\mu) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right) \right] =$$

$$= \left(2\pi\sigma^2\right)^{-n/2} exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2 \right)$$

Bayesian analysis for the mean of a normal model with known variance (II)

- Observing its form, we can consider as the prior distribution the normal conjugate distribution $\mu \sim N(\nu, \tau)$, with mean ν and variance τ^2 fixed and known.

$$P(\mu) = \frac{1}{\sqrt{2\pi}\tau} exp\left(-\frac{(\mu-\nu)^2}{2\tau^2}\right)$$

- For the prior distribution above, its corresponding posterior distribution of μ is (after doing several calculations):

$$\mu|\mathbf{x} \sim N(w\nu + (1-w)\bar{x}, \sqrt{\frac{\sigma^2}{n}(1-w)}), \text{ where } w = \frac{\sigma^2/n}{\sigma^2/n + \tau^2}.$$

- Note that the posterior mean is a compromise between the prior mean ν and the sample mean $\bar{x}$.

Bayesian analysis for the mean of a normal model with known variance (III)

- When $n \to \infty$:

$$\mu|\mathbf{x} \to N(\bar{x}, \frac{\sigma}{\sqrt{n}})$$

- If we consider a non-informative prior, such as, $p(\mu) \propto 1$, then

$$\mu|\mathbf{x} \to N(\bar{x}, \frac{\sigma}{\sqrt{n}})$$

Bayesian modeling of multiparametric distributions

The idea

- In multiparametric models, there are usually several parameters that are of major interest.
- The objective is to find the posterior distribution of these parameters:

$$P(\theta_1|\mathbf{x}) = \int P(\theta_1, \theta_2|\mathbf{x})d\theta_2 = \int P(\theta_1|\theta_2, \mathbf{x})P(\theta_2|\mathbf{x})d\theta_2.$$

- The above integral is not usually evaluated since it usually does not have a simple expression.
- But it gives us already a first very good idea of how to approximate the posterior distribution by marginalizing and simulating conditionally.

Bayesian analysis for a normal model with **unknown** mean and variance (I)

- Suppose we have data following a normal distribution with mean μ and variance σ^2 both unknown. That is, $\mathbf{x} = (x_1, \dots, x_n)$ a realization of a random sample $\mathbf{X} = (X_1, \dots, X_n)$ with $X_i : N(\mu, \sigma)$ for $i = 1, \dots, n$, where μ and σ are both **unknown**.
- Then the **likelihood** function:

$$P(\mathbf{x}|\mu, \sigma) \propto \sigma^{-n} exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{x} - \mu)^2]\right),$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sample variance.

Bayesian analysis for a normal model with **unknown** mean and variance (II)

- The choice of prior distributions for multiparametric models is often not a simple task.
- We can consider a **prior non-informative** for the two parameters that assumes independence between the two parameters and is uniform on the scale of μ and σ^2

$$P(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

- Then the joint posterior distribution is:

$$P(\mu, \sigma^2 | \mathbf{x}) \propto \sigma^{-n-2} exp \left(-\frac{1}{2\sigma^2} [(n-1)s^2 + n(\bar{x} - \mu)^2] \right).$$

Bayesian analysis for a normal model with **unknown** mean and variance (III)

- To get the marginal of greatest interest we factorize as before:

$$P(\mu|\mathbf{x}) = \int P(\mu, \sigma^2|\mathbf{x}) d\sigma^2 = \int P(\mu|\sigma^2, \mathbf{x}) P(\sigma^2|\mathbf{x}) d\sigma^2.$$

- the posterior of the mean of a normal with known variance:

$$\mu|\sigma^2, \mathbf{x} = N(\bar{x}, \frac{\sigma}{\sqrt{n}})$$

Bayesian analysis for a normal model with **unknown** mean and variance (IV)

- To obtain $P(\sigma^2|\mathbf{x})$ we integrate the joint distribution over μ (marginalize)

$$P(\sigma^2|\mathbf{x}) = \int P(\mu, \sigma^2|\mathbf{x}) d\mu \propto (\sigma^2)^{-(n+1)/2} exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right),$$

$$\sigma^2|\mathbf{x} \sim Inv - \chi^2(n-1, s^2)$$

.

Marginal posterior for μ

- With the above two expressions we factor the joint distribution

$$\begin{array}{rcl} P(\mu, \sigma^2|\mathbf{x}) & = & P(\mu|\sigma^2, \mathbf{x}) \times P(\sigma^2|\mathbf{x}) \\ \mu, \sigma^2|\mathbf{x} & = & N(\bar{x}, \frac{\sigma}{\sqrt{n}}) \times Inv - \chi^2(n-1, s^2) \end{array}$$

- and we can obtain the marginal $P(\mu|\mathbf{x}) = \int P(\mu, \sigma^2|\mathbf{x}) d\sigma^2$, which turns out to

$$\mu|\mathbf{x} \sim t_{n-1}(\bar{x}, s^2/n)$$

where $t_n(a, b)$ is a t-student with n degrees of freedom centered at a and with location parameter b[4].

[4] In R it is in the library `LaplacesDemon` (`dist.Student.t`)
https://www.rdocumentation.org/packages/LaplacesDemon/versions/16.1.4/topics/dist

Acknowledgement

Note: this document is based on material kindly provided by Professor David Conesa of the University of Valencia and the Valencia Bayesian Research Group (http://vabar.es/)