Basics of Bayesian Statistics

David V. Conesa Guillén

VAlencia BAyesian Research group

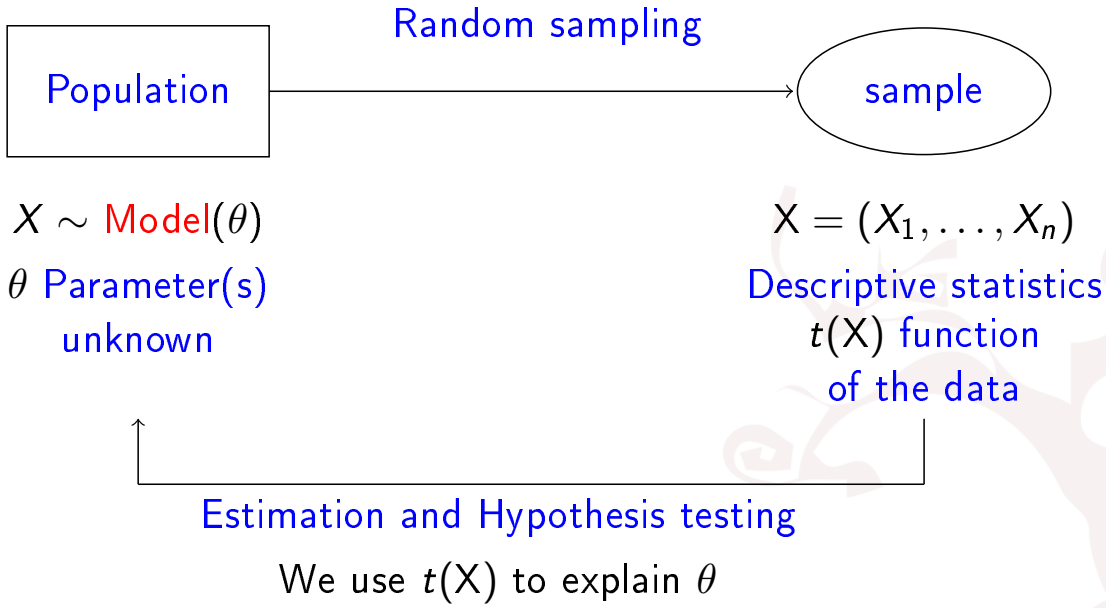
Dept. d'Estadística i Investigació Operativa

Universitat de València

1 The Bayesian way

Statistical Modelling (known stuff!)

- In general, a model is a small-scale representation of reality.
- Statistical models are models that allow us to incorporate the variability present in real life using randomness.



Statistical Modelling (2)

- But it is not easy to find the model that completely describes each real life situation...
- ... "essentially, all models are wrong, but some of them are useful" (Box, 1987).
- Dealing with a real problem, "it is important to set an adequate model to work with, and this decision implies some previous knowledge about the problem" (Martínez-Abraín et al., 2014).
- In A. Einstein's words: "the problem formulation is more essential than its own solution, which may simply be a mathematical or experimental skill".

Statistical Modelling (3)

- Let's play with a really basic example.
- Suppose we want to estimate the proportion of people with COVID in this room.
- Usual model is bernoulli-binomial ...

$$Y \sim \text{Ber}(\pi)$$

- ... after gathering data (e.g. checking if 5 people have simptoms)...

$$Y_i \sim \text{Ber}(\pi), i = 1, \ldots, n = 5; \quad \text{data are } \{y_i; i = 1, \ldots, n = 5\}$$

- ... we can make inference about that proportion using a "good" estimator like the sample proportion $\hat{\pi} = \frac{\sum_i y_i}{5}$.
- Is this the only way? We can also do it using the Bayesian approach.

Bayesian inference

- Inference and prediction about the unknown parameters usually done with the frequentist (a.k.a. classical) approach.
- Another option is to estimate and predict using Bayesian statistics:
  - Based on the fact that information and uncertainty about all the unknown can be better (and easily) expressed in terms of probability distributions.
  - Probability is then used (and understood) as a "subjective" measure of the uncertainty.
  - Consider as unknown elements of the problem not only data but also the parameters of the distribution that govern their behavior.
  - Data are used to update knowledge about the parameters.
  - Inference is made in terms of (posterior) probability distributions:

Why Bayesian is so popular now?

- The Bayesian point of view is not a technique in the field of Statistics.
- It is another way of understanding and performing Statistics.
- And so, when our data bring us to a any model ...
- ... we can solve it using both Bayesian and Frequentist methods.
- Bayesian statistical analysis has benefited from the explosion of cheap and powerful desktop computing over the last two decades: MCMC, INLA.
- Bayesian techniques can now be applied to complex modeling problems where they could not have been applied previously. SO FAMOUS IN MANY APPLIED FIELDS!!
- Bayesian perspective will probably continue to challenge, and perhaps sub-plant, traditional frequentist statistical methods which have dominated many disciplines of science for so long.

History of the "Theory that would not die"

- Beginning: Thomas Bayes 1702-1761, clergyman and vocational scientist/mathematician (inverse probability in insurance).

- "Middle" ages: Technical problems. Frequentist statistics becomes popular. Bayesian approach is even not well seen although it is used ("in the shadows").
- "Modern age": MCMC and the momentum of Bayesian Statistics.
- "Post-Modern" age: Applications everywhere.
- A good book to read is "The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy" (Sharon McGrayne).

2 Bayesian learning process

Elements of a statistical analysis

✓ Parameter: Population characteristic of interest (unknown) on which to infer from the data observed

Example

... Proportion of people with COVID in a region, glucosa level in blood, ...

✓ Data: observations of the characteristic of interest on a sector (sample) of the population

Example

... if someone has COVID, observed glucose, ...

✓ Likelihood: Obtained from the probabilistic model assumed on data, gives an idea about the most (and least) plausible (likely) parameter values.

Fundamentals of the Bayesian approach.

- Interpretation of Bayes theorem:

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

  - $P(H_i)$ previous knowledge about H_i
  - $P(B|H_i)$ sampling information
  - $P(H_i|B)$ updated knowledge about H_i

- Bayesian inference:

$$\text{Posterior information} = \begin{cases} \text{Prior information} \\ \text{sampling information} \end{cases} +$$

Fundamentals of the Bayesian approach (2)

Reasoning in terms of probability

- about the observed and variable in the sampling: data
- ... but ALSO about the unknown and unobserved: parameters

UNCERTAINTY $\equiv$ PROBABILITY

We express our knowledge about something through probability distributions

BAYESIAN LEARNING PROCESS

- Likelihood from the model assumed for the data, $l(\theta; x)$.
- Before observing data: Prior distribution about the parameters $p(\theta)$.
- Use Bayes theorem to update information about the parameters using data observed.

Fundamentals of the Bayesian approach (3)

- Construction of the joint distribution about the unknown elements of the problem:
  $l(\theta; x) = p(\mathrm{x}|\boldsymbol{\theta})$ is the likelihood function of the observed data,
  $p(\boldsymbol{\theta})$ is the prior distribution,

$$p(\mathrm{x}, \boldsymbol{\theta}) = p(\boldsymbol{\theta})p(\mathrm{x}|\boldsymbol{\theta})\,.$$

- Using Bayes theorem to obtain the posterior distribution:

$$p(\boldsymbol{\theta}|\mathrm{x}) = \frac{p(\mathrm{x}, \boldsymbol{\theta})}{p(\mathrm{x})} = \frac{p(\boldsymbol{\theta})p(\mathrm{x}|\boldsymbol{\theta})}{p(\mathrm{x})} = \frac{p(\boldsymbol{\theta})p(\mathrm{x}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathrm{x}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}$$

As $p(\mathrm{x})$ does not depend on $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\mathrm{x}) \propto p(\boldsymbol{\theta}) \times p(\mathrm{x}|\boldsymbol{\theta})\,.$$

Likelihood

Example

1. Suppose an interest in estimating the proportion of people with COVID in this room, the resulting likelihood for the parameter of interest (π) is based on the number of people that with it (r) and the total number of people sampled (n):
   Likelihood:
   $$p(x|\pi) = \ell(\pi) = \binom{n}{r} \pi^r (1-\pi)^{n-r}$$
2. But, if we were counting the number of people arriving to a hospital in one hour (Poisson distributed data) and we would like to regress them in terms of some covariates (hot day or not), we would model the data with a Poisson regression being the data with mean μ_i and a log link $g(\mu_i) = \log(\mu_i)$. The resulting likelihood is:
   $$\ell(\boldsymbol{\beta}|y, X) = \prod_{i=1}^{n} \left(\frac{1}{y_i!}\right) \exp\{y_i X_i{}^t \boldsymbol{\beta} - \exp(X_i{}^t \boldsymbol{\beta})\}$$

3 Prior distributions

Bayesian methodology requires establishing a prior distribution about the unknown parameters, but

how can we get that prior?

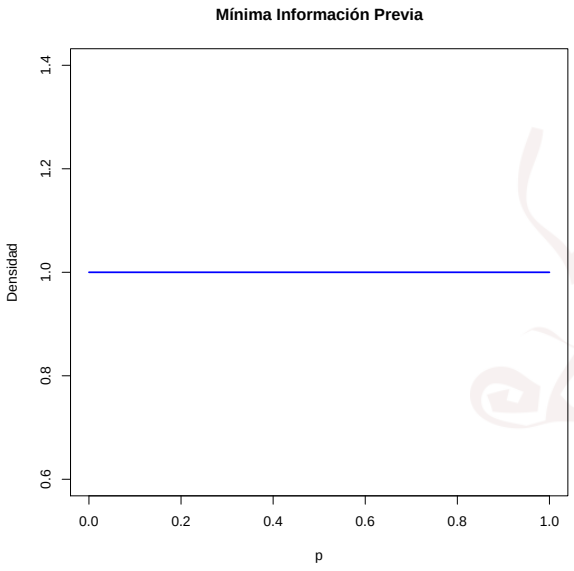
Possibilities:

- Objective prior information, that is, there is no prior knowledge about the parameter of interest.
  From an objective analysis, it is expected that it provides results as objective as those produced by a frequentist analysis.
- Subjective prior information, expressed in terms of a prior distribution obtained from:
  - Information gathered from experts on the subject.
  - Information obtained from previous experiments.

Example

Ignorance: non informative prior

$$\pi \text{ is a probability: } \pi \in [0,1] \quad \Rightarrow \quad \pi \sim Un(0,1) \equiv Be(1,1).$$



Non informative priors

Useful in those situations in which prior information is very difficult to measure or not convenient to use.

- A distribution constant in all the values of the parametric space (even improper distributions!).
  It is possible to use the Bayesian learning process even when the prior is improper, but it is necesary to check that the resulting posterior is proper.
- Jeffreys prior distributions, invariant in front of transformations:

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

  where $I(\theta) = -\mathrm{E}^{(\mathrm{x}|\theta)}\left[\frac{\partial^2 \ln p(\mathrm{x}|\theta)}{\partial \theta^2}\right]$, represents the Fisher information of θ.

Conjugate prior distributions

- A family P of prior distributions is said to be a conjugate family for F (the class of all density functions $p(\mathrm{x}|\boldsymbol{\theta})$ of parameter θ), if the posterior distribution $p(\boldsymbol{\theta}|\mathrm{x})$ belongs to the family P for all x of the parametric space and for all prior distributions of P.
- The family Γ of distributions for $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$, is *conjugate* respect de family $F = \{p(\mathrm{x}|\boldsymbol{\theta})\}$, if the posterior $p(\boldsymbol{\theta}|\mathrm{x}) \in \Gamma$, $\forall p(\cdot) \in \Gamma$ and for all $p(\cdot|\cdot) \in F$
  $\Rightarrow$ Posterior distribution with the same parametric form of the prior.

Conjugate prior distributions

- *Natural* conjugate prior distributions appear when considering P as the family with the same functional form than the likelihood.
- Advantages of their use:
  - simplify the calculations when obtaining the posterior,
  - facilitate the description of the results,
  - allow an interpretation of the prior as an equivalent experiment,
  - are useful in the construction of more complicated models.
- disadvantages:
  - except in the simplest cases, are often too rigid to represent the available information.

Example

1. In the previous example of estimating π, the probability of people with covid:
   - Likelihood:
   $$p(x|\pi) = \ell(\pi) = \binom{n}{r}\pi^r(1-\pi)^{n-r}$$
   - Conjugate prior distribution, in this case a Beta distribution with parameters a and b:
   $$p(\pi) \propto \pi^{a-1}(1-\pi)^{b-1}$$
2. A non-informative (and improper) prior for the Poisson regression example:
   $$p(\boldsymbol{\beta}) \propto 1$$

4 Posterior distributions

The application of Bayesian learning process results in the posterior distribution of the parameters of interest:

Example

1. In the previous example of estimating π, the probability of people with covid:
   - Likelihood: $p(x|\pi) = \ell(\pi) = \binom{n}{r}\pi^r(1-\pi)^{n-r}$
   - A Beta Prior distribution with parameters a and b: $p(\pi) \propto \pi^{a-1}(1-\pi)^{b-1}$
   - Posterior distribution:

$$\begin{aligned} p(\pi|x) &\propto p(x|\pi) \times p(\pi) \propto \pi^{r+a-1}(1-\pi)^{n-r+b-1} \\ \pi|x &\sim \text{Beta}(r+a, n-r+b) \end{aligned}$$

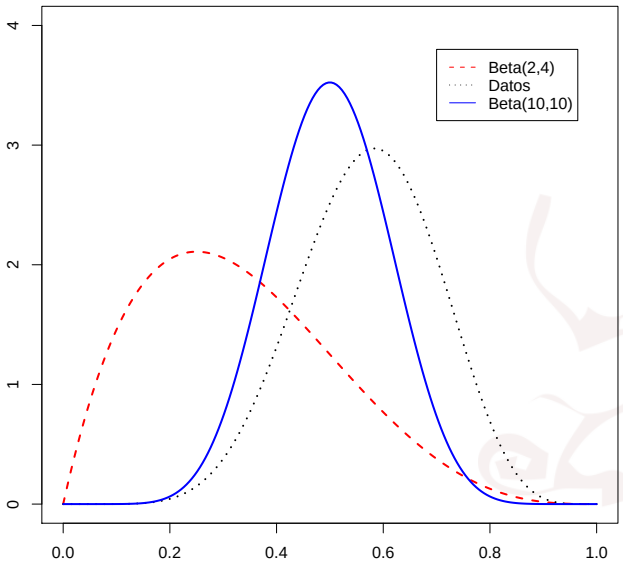
2. In the Poisson regression example:

$$p(\beta|X) \propto 1 \times \prod_{i=1}^{n} \left(\frac{1}{y_i!}\right) \exp\{y_i X_i{}^t \beta - \exp(X_i{}^t \beta)\}$$

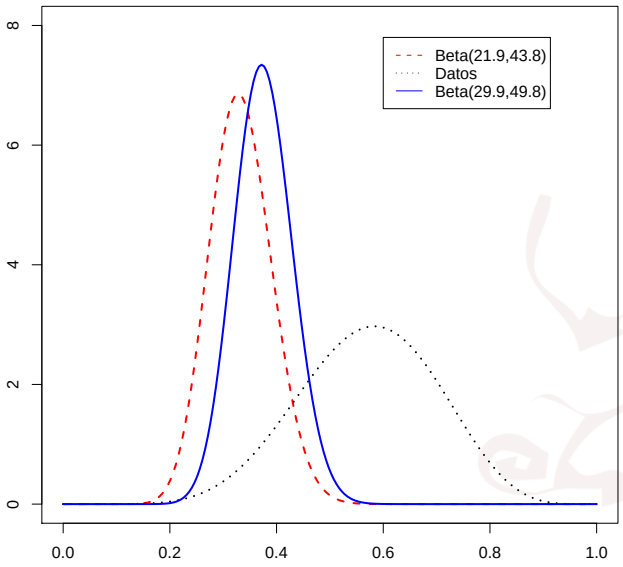
Note that the posterior does not have a known analytical expression.

Posterior distributions are influenced by the prior knowledge:

- If our previous knowledge about the people with covid is 1/3 (33.33 %) with a standard deviation of 0.2, we can use it to assign the values of a and b:
  $\Rightarrow$ Prior Beta(2,4) $\Rightarrow$ Posterior Beta(10,10), being $n = 14$ and $r = 8$.



- If our previous knowledge about the people with covid is 1/3 (33.33 %) with a standard deviation of 0.01 (that is, we are more convinced than before), we can use it again to assign the values for a and b:
  $\Rightarrow$ Prior Beta(21.9,43.8) $\Rightarrow$ Posterior Beta(29.9,48.8), being $n = 14$ and $r = 8$.



- The posterior distribution is a compromise between the sample information and the prior distribution:

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

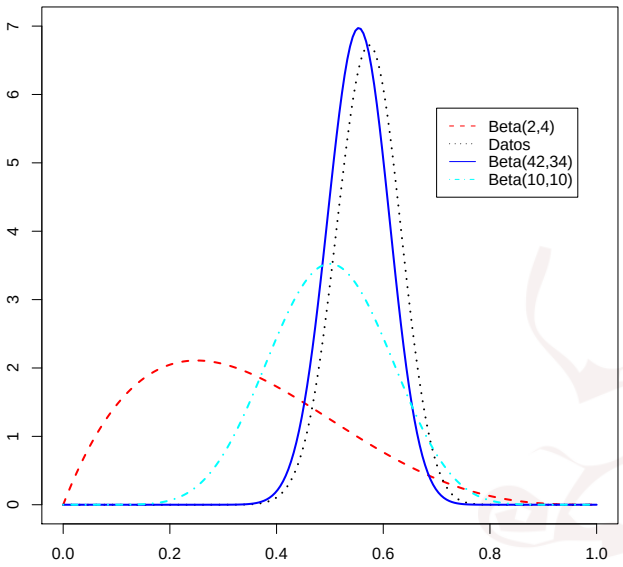
- Posterior variance:

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

- Parameters of the prior lose influence as the sample size increases. In both cases, when a and b are fixed, if n increases:

$$\begin{aligned} \mathrm{E}(\theta|\mathrm{x}) &\approx \hat{p} \\ \mathrm{Var}(\theta|\mathrm{x}) &\approx \frac{\hat{p}}{n}(1-\hat{p}) \end{aligned}$$

With a Beta(2,4) prior, if we get an information that the 40 people out of 70 has covid $\Rightarrow$ Beta(42,34) posterior.



Sequential incorporation of information

- If we get a posterior distribution $p(\boldsymbol{\theta}|\mathrm{x})$ from the data x, and we observe a second dataset y independently distributed from the previous one:

$$p(\boldsymbol{\theta}|\mathrm{x},\mathrm{y}) \propto p(\boldsymbol{\theta}|\mathrm{x}) \times p(\mathrm{y}|\boldsymbol{\theta},\mathrm{x}) = p(\boldsymbol{\theta}|\mathrm{x}) \times p(\mathrm{y}|\boldsymbol{\theta}),$$

  $p(\boldsymbol{\theta}|\mathrm{x})$ is now the prior distribution.

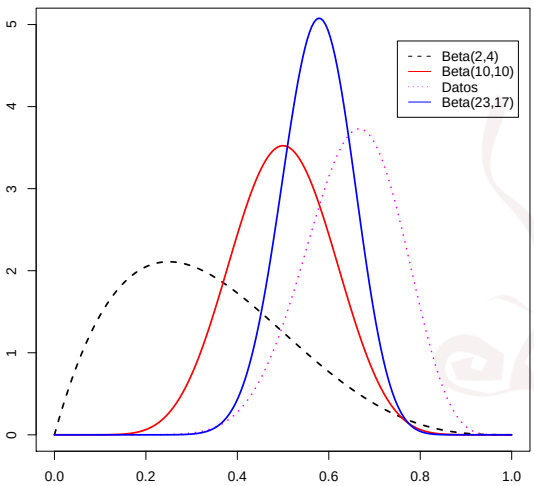
- Moreover, the resulting posterior is the same that it would have been obtained using all the data together:

$$p(\boldsymbol{\theta}|\mathrm{x},\mathrm{y}) \propto p(\boldsymbol{\theta}) \times p(\mathrm{y},\mathrm{x}|\boldsymbol{\theta}),$$

  and it works the same if the observation orden of the data is the opposite one.

Example

It is possible to add the information from a new experiment about the proportion of people with covid considering the posterior previously obtained, the Beta(10,10), as the new prior. If 13 out of 20 people has covid $\Rightarrow$ posterior Beta(23,17).



5 Predictive distributions

Predictive distributions

- Bayesian Statistics considers everything unknown as a random variable, so when predicting we look for the probability distribution of a new realization (in the same conditions) of the variable of interest conditioned to the knowledge we have about the parameters of the model.
- 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 designing and planning a future experiment.
- We can distinguish two predictive distributions depending on:
  - if we use the information about the parameter before the experiment, prior predictive distribution;
  - or if we also use the information provided by the experiment, posterior predictive distribution.

Prior predictive distributions

- The prior predictive distribution is the distribution of a new realization of the variable of interest (or any transformation of it expressed as a function of the model parameters) before performing the experiment, using only the not updated information (no observed data) about the parameters:
  - if $p(\theta)$ is the prior distribution that collects the previous information we have about the parameter that governs the variable of interest; then,
  - the prior predictive distribution of a new X is

$$m(X_{pred}) = \int p(X|\theta)p(\theta)d\theta\,.$$

- The predictive prior distribution of a model when the prior has no information (such as Jeffreys) does not make sense because there is no prior information or data and so, there is no information for predicting.

Posterior predictive distributions

- The posterior predictive distribution is the distribution of a new realization of the variable of interest (or any transformation of it expressed as a function of the model parameters) after performing the experiment using the already updated information (with the observed data) about the parameters:
  - if $\mathrm{x} = (x_1, \ldots, x_n)$ is the realization of a m.a. $\mathrm{X} = (X_1, \ldots, X_n)$ of a random variable of interest $X \sim F(x|\theta)$ where θ is unknown;
  - $p(\theta|\mathrm{x})$ is the posterior distribution of the parameter that governs said variable of interest

$$p(\theta|\mathrm{x}) = \frac{p(\theta)p(\mathrm{x}|\theta)}{\int p(\theta)p(\mathrm{x}|\theta)\mathrm{d}\theta\,;}$$

  - then, the posterior predictive distribution of a new X is

$$m(X_{pred}|\mathrm{x}) = \int f(X|\theta)p(\theta|\mathrm{x})d\theta\,.$$

Example

1. If we would decide to run a new experiment and check in how many people out of m have covid, the distribution that predicts (knowing what has happened before) the number of people with covid, Z, is:

$$p(z|\text{datos}) = \int \text{Binomial}(m, \theta)\text{Beta}(r + a, n - r + b)\text{d}\theta\,.$$

   In other words, Z is Beta-Binomial distributed.

2. In the case of the Poisson regression, if we were interested on predict the value of the analyzed variable (for instance, the number of patients arriving in one hour to the hospital in terms of some covariates of interest), we would have to use the posterior predictive distribution of that variable, in particular:

$$p(z|\text{data}) = \int \text{Poisson}(\mu|\boldsymbol{\beta})p(\boldsymbol{\beta}|\text{data})\text{d}\boldsymbol{\beta}\,.$$

   Note that the relationship between the mean and the covariates is $\log(\mu) = \text{X}_\text{i}{}^t\boldsymbol{\beta}$. This results in a posterior distribution with not known analytical expression.

6 Inference and prediction

Describing posteriors and predictives

- Posterior distributions and posterior (and prior) predictive distributions can be considered as the very last unique result of the inferential process and the predictive process.
- To perform inference is equivalent to perform a detailed description of the posterior distribution.
  - Graphical representation of the density: ineffective in large dimensions.
  - Point estimation: generalized maximum likelihood estimator (the mode of the posterior), the posterior mean, the posterior median.
  - credible sets and regions of high density.
  - Hypothesis testing
- Prediction: detailed description of the predictive distribution.

Point estimation

Use of standard measures of location:

- Generalized maximum likelihood estimator: the mode of the posterior.
  - Easy to obtain
  - Does not requiere normalization constant
  - It's equivalent (the same) than the frequentist m.l.e. if a constant prior is used
  - But totally ignores the distribution tails
- Posterior expectation
  - The most used Bayesian estimator
  - Good propierties
  - It largely depends on tails distribution
- Posterior median.
  - Avoids the tails problem
  - Hard to obtain (calculations more complicated)

Credible sets. Maximum density regions.

Description of the posterior distribution via the probabily that the parametric vector falls inside a given region of interest.

- *$100(1-\alpha)\,\%$ credible set* for the vector $\boldsymbol{\theta}$, subset C of the parametric space such as:

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

  in the continuous case and $1-\alpha \leq \sum_{\boldsymbol{\theta}\in C} p(\boldsymbol{\theta}|\mathrm{x})$ in the discrete.
- *Maximum density region at $100(1-\alpha)\,\%$* for the vector $\boldsymbol{\theta}$:

$$C = \{\boldsymbol{\theta} \in \Theta : p(\boldsymbol{\theta}|\mathrm{x}) \geq k(\alpha)\},$$

  where $k(\alpha)$ is the large constant such as $\mathrm{P}(C|\mathrm{x}) \geq 1-\alpha$.
  - These regions minimize the meauser of the region for a given probabilistic level, and
  - the value of the density inside the region is always higher than the ones outside.
- In one dimensional distributions, the intervals are usually defined by quantiles of the posterior distribution.

Hypothesis testing

Let $\begin{cases} \mathrm{H}_0: & \boldsymbol{\theta} \in \Theta_0 \\ \mathrm{H}_1: & \boldsymbol{\theta} \in \Theta_1 \end{cases}$ be two particular hypotheses to be tested:

- Objective: find the correct hypothesis.
- Answer:
  - Obtain $\alpha_0 = \mathrm{P}(\Theta_0|\mathrm{x})$ $\alpha_1 = \mathrm{P}(\Theta_1|\mathrm{x})$.
  - Resolve according to the rule: reject H_0 when $\alpha_0 < \alpha_1$.
- Other alternatives:
  - *Prior odds* $= \pi_0/\pi_1$ (where π_i is the prior probability of Θ_i).
  - *Posterior odds* $= \alpha_0/\alpha_1$.
  - *Bayes Factor*, $\frac{\pi_0/\pi_1}{\alpha_0/\alpha_1}$. A measure of how the data are in favor of H_0.

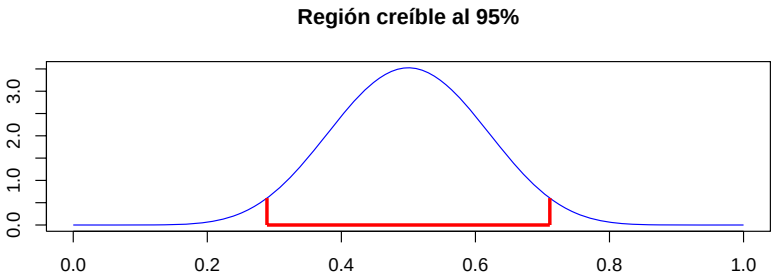
Example

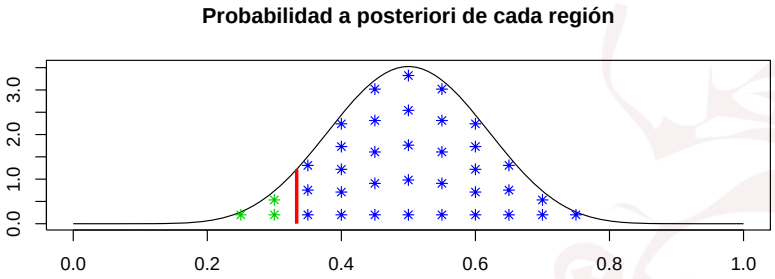
- In our study about the proportion of covid, suppose thatn $r = 8$ (out of $n = 14$) people has covid, the posterior is Beta(10,10) when the prior is a Beta(2,4):
  - Estimation for π: E(π|x) =0.5 equivalent to median and posterior mode.
  - A 95 % credible region (equivalent to the maximum density region):

$$C = [q_{0,025}\text{Beta}(10,10), q_{0,975}\text{Beta}(10,10)] = (0.289,0.711)$$

  - To test if the proportion is less than one third, $\begin{cases} H_0: & \theta \geq \frac{1}{3} \\ H_1: & \theta < \frac{1}{3} \end{cases}$

    As $\alpha_0 = P(\theta \geq \frac{1}{3}|x)$ =0.935 and $\alpha_1 = P(\theta < \frac{1}{3}|x) = 0.065$, We accept H_0.
    "Posterior odds" = 0.935/0.065, "Prior odds"= 0.461/0.539.
    Bayes Factor = $\frac{0.461/0.539}{0.935/0.065} = 0.059$.
    Conclussion: we are 14 times more in favor of the hypothesis that the proportion is higher than one third after observing the result.
- But, what should we do in the case of Poisson regression when we do not have analytical expression for the posterior distributions?





7 Differences between Classical and Bayesian

Analyzing the covid example again

- Suppose again we want to estimate the proportion of people with COVID in this room.
- Usual model is bernoulli-binomial ...

$$Y \sim \text{Ber}(\pi)$$

- ... after gathering data (e.g. checking if 5 people have simptoms)...

$$Y_i \sim \text{Ber}(\pi), i = 1, \ldots, n = 5; \quad \text{data are } \{y_i; i = 1, \ldots, n = 5\}$$

- ... we can make inference about that proportion using a "good" estimator like the sample proportion $\hat{\pi} = \frac{\sum_i y_i}{5}$.
- Is this the only way? We can also do it using the Bayesian approach.

Limitations of the frequentist approach

- Suppose that we would like to know whether π (proportion of people with covid) is greater than 0.08, that is,

$$\begin{aligned} H_0 &: \pi \leq 0.08 \\ H_1 &: \pi > 0.08, \end{aligned}$$

  based on the fact that we find 36 out of 400 people with COVID.
- Using a classical test, based on the m.l.e. $\hat{\pi} = 36/350$, we obtain a p–value of 0.05749, which does not allow to conclude about the contrast.
- The (95 %) confidence interval for π is (0.079;1).

Limitations of the frequentist approach (2).

From a frequentist approach we can obtain certain probabilities such as:

- What's the probability that the m.l.e. takes values above the real value of the parameter? $Pr(\hat{\pi} > \pi)$
- What's the percentage of times (in the sampling) in which the confidence interval "will catch up" the real value of π?
- What's the probability of being wrong rejecting H_0?

... But not the most interesting ones:

- What's the probability that π is higher than 0.08? The p–value is NOT the probability of H_0
- What's the probability that π is between 0.05 and 0.25?
- What's the expected value of π and what's the uncertainty about it?
- If we decide to repeat the experiment with another 50 people, what's the probability of 20 with covid?

The Bayesian learning process

- Construction of the joint posterior distribution of unknowns:
  $l(\theta; x) = p(\mathrm{x}|\boldsymbol{\theta})$ is the likelihood,
  $p(\boldsymbol{\theta})$ is the prior distribution,

$$p(\mathrm{x}, \boldsymbol{\theta}) = p(\boldsymbol{\theta})p(\mathrm{x}|\boldsymbol{\theta})\,.$$

- Obtaining the posterior distribution via Bayes theorem:

$$p(\boldsymbol{\theta}|\mathrm{x}) = \frac{p(\mathrm{x}, \boldsymbol{\theta})}{p(\mathrm{x})} = \frac{p(\boldsymbol{\theta})p(\mathrm{x}|\boldsymbol{\theta})}{p(\mathrm{x})} = \frac{p(\boldsymbol{\theta})p(\mathrm{x}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathrm{x}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}$$

  As $p(\mathrm{x})$ does not depend of $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\mathrm{x}) \propto p(\boldsymbol{\theta}) \times p(\mathrm{x}|\boldsymbol{\theta})\,.$$

Example: Interest of estimating π, the proportion of COVID

- Our model: $Y \sim \mathrm{Ber}(\pi)$
- Data $Y_i \sim \mathrm{Ber}(\pi), i = 1, \ldots, n$
- Observed data $\mathrm{y} = \{y_i; i = 1, \ldots, n\}$
- Likelihood: $p(\mathrm{x}|\pi) = \ell(\pi) = \binom{n}{r}\pi^r(1-\pi)^{n-r}$
- A Beta Conjugate Prior distribution with parameters a and b:

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

- Posterior distribution:

$$\begin{aligned} p(\pi|\mathrm{y}) &\propto p(\mathrm{y}|\pi) \times p(\pi) \propto \pi^{r+a-1}(1-\pi)^{n-r+b-1} \\ \pi|\mathrm{y} &\sim \mathrm{Beta}(r+a, n-r+b) \end{aligned}$$

- The posterior predictive distribution that predicts (knowing what has happened before) the number of persons with COVID, Z, out of m is:

$$\begin{aligned} p(Z|\mathrm{y}) &= \int \mathrm{Binomial}(m, \pi)\mathrm{Beta}(r+a, n-r+b)\mathrm{d}\pi \\ &= \text{Beta-Binomial}(m, r+a, n-r+b) \end{aligned}$$

Estimating the proportion of COVID depending on age

- Our previous simple model only had one parameter. Life is usually more complex.
- Let's now include a dependence with a covariate (AGE): Logistic regression (GLM).
- Model in two pieces (being X_i, the age of person i):

$$\begin{aligned} Y_i &\sim \text{Ber}(\pi_i), \quad \forall i = 1, \ldots, n \\ \text{logit}(\pi_i) &= \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i\,, \quad \forall i = 1, \ldots, n \\ \pi_i &= \text{logit}^{-1}(\beta_0 + \beta_1 X_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\,, \quad \forall i = 1, \ldots, n \end{aligned}$$

- Likelihood:

$$\begin{aligned} \ell(\beta_0, \beta_1) = p(\mathbf{x}, \mathbf{y}|\beta_0, \beta_1) &= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{y_i} \left(1 - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{1-y_i}. \end{aligned}$$

- Classical estimation not easy: but so easy to use `glm` function in R. Do we really know what happens behind?

Estimating the proportion of COVID depending on age (2)

- To make inference in a GLM within the Bayesian framework, we first assign priors to the parameters:
  - independent Gaussian prior for the coefficients of the regressors (centered at zero with large variance), e.g. N(0,10000), that is, $p(\beta_0, \beta_1) = N(\beta_0|0, 10000) \times N(\beta_1|0, 10000)$, o
  - Improper flat prior $p(\beta_0, \beta_1) \propto 1 \times 1 \propto 1$.
- The posterior distribution does not have an analytical expression, neither the predictives:

$$\begin{aligned} p(\beta_0, \beta_1|\mathbf{y}, \mathbf{x}) &\propto p(\beta_0, \beta_1) \times \ell(\beta_0, \beta_1) \\ &\propto 1 \times \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{r} \left(1 - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{n-r}. \end{aligned}$$

- Numerical tools are needed.
- But life is even more complicated, as we would like to have more covariates and our friends the random effects.

Estimating proportion of COVID depending on age and city

- Imagine now that we have data from a random sample of c cities in which we still analyze whether having COVID is related to age (GLMM).
- Model in three pieces (being X_i, the age of person i):

$$\begin{aligned} Y_i &\sim \text{Ber}(\pi_i), \quad \forall i = 1, \ldots, n \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 X_i + \text{City}_{j(i)}, \quad \forall i = 1, \ldots, n; \forall j = 1, \ldots, c \\ \text{City}_{j(i)} &\sim N(0, \sigma_C^2) \end{aligned}$$

- To make inference in a GLMM within the Bayesian framework, we first assign priors to the parameters and hyperparameters:

$$p(\beta_0, \beta_1, \sigma_C^2 | \mathbf{y}, \mathbf{x}) \propto p(\beta_0, \beta_1, \sigma_C^2) \times \ell(\beta_0, \beta_1, \sigma_C^2)$$

- The posterior distribution does not have an analytical expression, neither the predictives, and again Numerical tools are needed.