ADVANCED BIOSTATISTICS ABSTAT17

Bayesian Inference

Lisete Sousa

Department of Statistics and Operations Research, CEAUL
Faculty of Sciences of Lisbon University

IGC, April 10th - 13th, 2017

Contents

- Introduction
- Bayes Theorem: Discrete Case
- Bayesian Updating
- Bayes Theorem: Continuous Case
- Choosing Prior Distributions
- Features on Posterior Distributions

Introduction

Under the Bayesian point of view to statistical inference, all unknown quantities in a statistical system are treated as random variables, reflecting (typically) subjective uncertainty measured by a probability distribution.

The Bayesian approach allows one to combine information from different sources to estimate unknown parameters.

- Both data and external information (prior) are used.
- Computations are based on the Bayes theorem.
- Parameters are defined as random variables.

- Direct probabilistic interpretation of a confidence interval for a parameter is possible, ie, given the observed data $\mathbf{x} = (x_1, \ldots, x_n)$ we can find an interval (or region in a multiparametric situation), say $[\theta_1, \theta_2]$, for a parameter θ of the hypothesized model $f_X(x|\theta)$ such that

  $$P(\theta \in [\theta_1, \theta_2] | \mathbf{x}) \geq 1 - \alpha,$$

  for a given α.
- In a hypothesis testing problem we can compute the probability that a specific hypothesis is true, given the data.
- Prior knowledge and reasonable prior concepts can be built into the analysis.

Why Bayesian methods

- Allow incorporation of (prior) scientific information.
- Appropriateness of methods does not depend on having large sample sizes.
- Direct probability interpretations.

Bayes Theorem: Discrete Case

- A scientist has M disjoint hypotheses $(H_1, H_2, \dots, H_M)$ about some random mechanism. These hypotheses are mutually exclusive and exhaustive.
- The "true" hypothesis cannot be observed, but the scientist may assign probabilities $p(H_i)$ to the events "hypothesis H_i is true". These are called prior probabilities. They should obey the axioms of probability, namely

$$0 \leq p(H_i) \leq 1, \quad i = 1, 2, \dots, M,$$

$$p(H_i \cap H_j) = 0, \quad i \neq j,$$
$$\sum_{i=1}^{M} p(H_i) = 1.$$

- An experiment can be performed with N observable effects $E_1, E_2, \ldots, E_N$. Given that hypothesis H_i holds, one expects to observe effects with conditional probabilities

$$0 \leq p(E_j|H_i) \leq 1, \quad i = 1, 2, \ldots, M, \quad j = 1, 2, \ldots, N,$$

$$p(E_j \cap E_k|H_i) = 0, \quad j \neq k,$$

$$\sum_{j=1}^{N} p(E_j|H_i) = 1.$$

- E is a random variable taking one of the states $E_j, j = 1, 2, \ldots, N$ and H a random variable taking one of the states $H_i, i = 1, \ldots, M$. The joint distribution of H and E is

$$p(H = H_i, E = E_j) = p(E_j|H_i)p(H_i),$$

$$i = 1, 2, \ldots, M; j = 1, 2, \ldots, N.$$

- Given that effect E_j is observed the conditional probability that H_i holds is

$$
\begin{aligned}
p(H_i|E_j) &= \frac{p(H = H_i, E = E_j)}{p(E_j)} \\
&= \frac{p(E_j|H_i)p(H_i)}{p(E_j)} \\
&= \frac{p(E_j|H_i)p(H_i)}{\sum_{i=1}^{M} p(E_j|H_i)p(H_i)} \\
&\propto p(E_j|H_i)p(H_i).
\end{aligned}
$$

This standard result of conditional probability is known as Bayes theorem. The probabilities $p(H_i|E_j)$ are called posterior probabilities. Given that E_j is observed we obtain the posterior distribution for H.

The last expression illustrates the concept of ”Bayesian learning”. This is a process by which a prior opinion is modified by the evidence to become a posterior opinion.

Bayesian Updating

Suppose now that there is additional evidence E_j'. Using Bayes theorem we have

$$\begin{aligned}
p(H_i|E_j', E_j) &= \frac{p(E_j', E_j|H_i)p(H_i)}{\sum_{i=1}^{M} p(E_j', E_j|H_i)p(H_i)} \\
&\propto p(E_j', E_j|H_i)p(H_i) \\
&= p(E_j'|E_j, H_i)p(E_j|H_i)p(H_i) \\
&\propto p(E_j'|E_j, H_i)p(H_i|E_j).
\end{aligned}$$

This indicates that the posterior distribution after evidence E_j conveys the prior opinion before E_j' is observed. It also describes how opinions are revised sequentially or, equivalently, how knowledge is modified by evidence.

If E_j and E_j' are conditional independent given H_i, then

$$p(E_j', E_j|H_i) = p(E_j'|H_i)p(E_j|H_i)$$

and

$$p(H_i|E_j', E_j) \propto p(E_j'|H_i)p(E_j|H_i)p(H_i).$$

More general, for n pieces of evidence, $\mathbf{E} = (E_{j_1}, \ldots, E_{j_n})$ and assuming conditional independence, we have

$$p(H_i|\mathbf{E}) \propto \prod_{\ell=1}^{n} p(E_{j_\ell}|H_i)p(H_i).$$

Example 1: Inheritance of Hemophilia

Suppose there is a non-hemophiliac woman whose father and mother are not affected by the disease but who has a hemophiliac brother. The woman can be a carrier or not. Let H_1 indicate that the woman is a carrier and H_2 indicate that she is not a carrier. Then, we can establish a priori that

$$P(H_1) = P(H_2) = \frac{1}{2}.$$

Suppose now that the woman has a non-hemophiliac son. Let E_1 represent this evidence. Hence we have:

$$P(E_1|H_1) = \frac{1}{2} \quad P(E_1|H_2) = 1.$$

By Bayes theorem we have the following posterior probabilities:

$$\begin{aligned} P(H_1|E_1) &= \frac{P(E_1|H_1)P(H_1)}{P(E_1|H_1)P(H_1)+P(E_1|H_2)P(H_2)} \\ &= \frac{1/4}{1/4+1/2} = \frac{1}{3} \\ P(H_2|E_1) &= \frac{2}{3} \end{aligned}$$

Suppose further that she has another son and he is also non-hemophiliac. Let E_2 represent this new evidence. To compute the posterior probabilities for H_1 and H_2 we can use as prior the posterior obtained before (that result from evidence E_1).

In this way our prior knowledge about H_1 and H_2 is now

$$P(H_1) = \frac{1}{3} \quad P(H_2) = \frac{2}{3}.$$

Assuming independence

$$P(E_2|H_1) = \frac{1}{2} \quad P(E_2|H_2) = 1.$$

Hence, applying again Bayes theorem we have

$$\begin{aligned} P(H_1|E_2) &= \frac{P(E_2|H_1)P(H_1)}{P(E_2|H_1)P(H_1) + P(E_2|H_2)P(H_2)} \\ &= \frac{1/6}{1/6 + 2/3} = \frac{1}{5} \\ P(H_2|E_2) &= \frac{4}{5} \end{aligned}$$

We can see that the same result can be obtained if the initial evidence (call it E) is that the woman has two non-hemophiliac sons.

Assuming independence

$$P(E|H_1) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \quad P(E|H_2) = 1.$$

Using as prior information

$$P(H_1) = P(H_2) = \frac{1}{2},$$

we have, by Bayes theorem,

$$\begin{aligned}
P(H_1|E) &= \frac{P(E|H_1)P(H_1)}{P(E|H_1)P(H_1)+P(E|H_2)P(H_2)} \\
&= \frac{1/8}{1/8+1/2} = \frac{1}{5} \\
P(H_2|E) &= \frac{4}{5}
\end{aligned}$$

as before.

Bayes Theorem: Continuous Case

As well as using Bayes' theorem for comparing models, we can use Bayes'theorem to estimate parameters of models.

Consider the situation where the role of the evidence is played by a vector of observations $x_1, \ldots, x_n$ and that we formulate a probability model for the correspondent random vector $X_1, \ldots, X_n$.

Usually this model depends on a parameter or a set of parameters θ with parameter space Θ.

- Let θ be a parameter or vector of parameters with prior probability density function $p(\theta)$. Let Θ be the support of θ, ie $\Theta = \{\theta : p(\theta) > 0\}$.

- Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the observed value of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ with joint probability distribution depending on θ, $f(x_1, \ldots, x_n|\theta)$.

- The marginal distribution of $\mathbf{X} = (X_1, \ldots, X_n)$ is

$$p(x_1, \ldots, x_n) = \int_\Theta f(x_1, \ldots, x_n|\theta)p(\theta)d\theta.$$

- Then the posterior distribution of θ, given the observed data $\mathbf{x} = (x_1, \ldots, x_n)$, is

$$\begin{aligned} p(\theta|x_1, \ldots, x_n) &= \frac{f(x_1, \ldots, x_n|\theta)p(\theta)}{p(x_1, \ldots, x_n)} \\ &= \frac{f(x_1, \ldots, x_n|\theta)p(\theta)}{\int_\Theta f(x_1, \ldots, x_n|\theta)p(\theta)d\theta} \quad \theta \in \Theta. \end{aligned}$$

This is the continuous version of Bayes theorem.

Choosing Prior Distributions

- Identify appropriate class of distributions, e.g.,
  - Data in $[0, 1]$: uniform distribution, Beta distribution
  - Data in $[0, \infty)$: gamma distribution, lognormal distribution, normal distribution (with $\mu \gg 0$)
  - Data in $(-\infty, \infty)$: normal distribution, t distribution
- Decide on informative *versus* non-informative or vague priors
  - non-informative (Jeffrey's prior) $p(\theta) \propto \sqrt{E(-\partial^2 f(x|\theta)/\partial\theta^2)}$
  - Vague: large variance
  - Informative: specify mean only *or* specify mean and variance *or* specify all parameters of distribution (if more than two)
  - We'll use vague priors for most of our examples.

Bernoulli Model and its Conjugate Prior - Beta Distribution

- We can consider a Bernoulli model for X with probability of success θ,ie,

$$P(X = x|\theta) = f(x|\theta) = \theta^x(1-\theta)^{1-x} \quad x = 0, 1$$

- Since θ is unknown we can assume that it is random and assign a probability distribution to it. This distribution will represent our prior opinion about the plausibility of values that θ takes in $\Theta = [0, 1]$

- For instance we can assume that θ has a Beta distribution with parameters (a, b) and hence write

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \quad \theta \in [0,1].$$

NOTE: *We say that a continuous random variable Y taking values in [0,1] has a beta distribution with parameters (a, b) if the probability density function is*

$$f_Y(y|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}y^{a-1}(1-y)^{b-1} \quad y \in [0,1]$$

- Assuming that X_i are iid Bernoulli with probability of success θ we have

  $$f(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

  If we assume a Beta prior for θ the posterior distribution is

  $$p(\theta|x_1, \dots, x_n) \propto \theta^{\sum_{i=1}^n x_i + a - 1}(1-\theta)^{n-\sum_{i=1}^n x_i + b - 1}$$

- Hence the posterior distribution is again beta with parameters

  $$(\sum_{i=1}^{n} x_i + a, n - \sum_{i=1}^{n} x_i + b).$$

  Note how the data transformed the prior opinion about θ.

	prior	posterior
hyperparameters	a, b	$\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b$
expected value	$\frac{a}{a+b}$	$\frac{\sum_{i=1}^n x_i + a}{n+a+b}$
variance	$\frac{ab}{(a+b)^2(a+b+1)}$	$\frac{(\sum_{i=1}^n x_i + a)(n - \sum_{i=1}^n x_i + b)}{(n+a+b)^2(n+a+b+1)}$
mode	$\frac{a-1}{a+b-2}$	$\frac{(\sum_{i=1}^n x_i + a - 1)}{n+a+b-2}$

The parameters of the prior are called hyperparameters. They are usually assumed to be known. Their values can be elicited using expert opinion.

Features of the posterior distribution

Posterior distributions give a complete description of the state of knowledge about an unknown, whether the unknown is a parameter, an hypothesis, a model, etc.They are the key for inferences in a Bayesian context.

NOTE: In the previous example we assumed we had a probabilistic model with a single parameter (the unknown), but as it was observed before θ *can be a vector of parameters (or anything which is unknown to us).*

The posterior distribution is used to draw inferences for θ. These inferences can be made in terms of features of the posterior distribution such has

- measures of location (mean, median, mode);
- quantiles;
- credible intervals;
- probabilities of sets, etc.

We will go through these measures using the previous example (Occurrence of Nucleotide A in a Sequence)

Example 2: Occurrence of Nucleotide A in a Sequence

Suppose that we have a sequence of nucleotides and we are interest in a specific nucleotide, say A.

- Let X_i ($i = 1\ldots, n$) be equal to 1 if A is in the position i and 0 otherwise. Suppose that the probability of occurrence of A, $P(A) = \theta$, is unknown but remains constant. Thus,

$$X_i \frown Bernoulli(\theta)$$

- Suppose that our prior opinion about the probability of occurrence of A in a long DNA sequence is $Beta(1,1)$, that is

$$\theta \frown U(0,1) \equiv Beta(1,1)$$

```
# Beta(a,b) with a=b=1
> a<-1
> b<-1
```

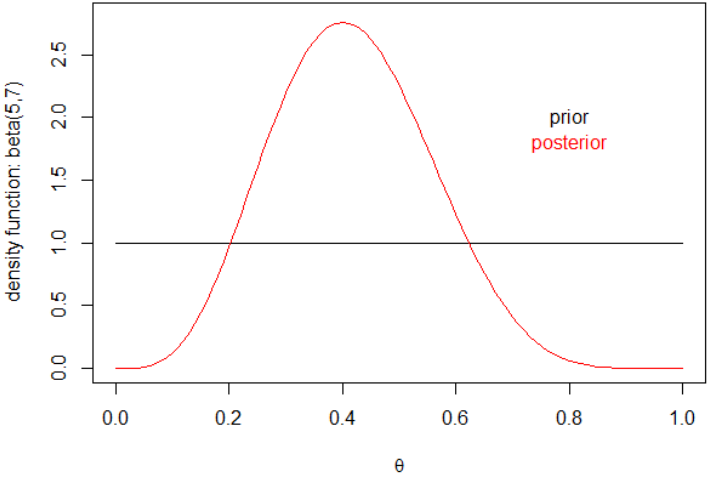
- We observe a subsequence of size $n = 100$ and obtain $t = \sum x_i = 43$. The posterior is $Beta(44, 58)$, that is

$$\theta|\mathbf{x} \frown Beta(44, 58)$$

```
> n<-100; x<-43
> a.star<-x+a; a.star
[1] 44
> b.star<-n-x+b; b.star
[1] 58
```

We can understand how our opinion was changed by the experiment, by plotting both densities:

```
> s<-seq(0,1,by=0.01)
> y<-dbeta(s,a,b)
> y.star<-dbeta(s,a.star,b.star)
> plot(s,y,type="n",xlim=c(0,1),ylim=c(0,2.8),
xlab=expression(theta),ylab="density > function: beta(5,7)")
> lines(s,y)
> lines(s,y.star,col="red")
> text(0.8,2,"prior")
> text(0.8,1.8,"posterior",col="red")
```


density function: beta(5,7)
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
θ
prior
posterior

Posterior Mean

The posterior mean of a parameter can be used as a point estimate for the parameter.

Since

$$p(\theta|\mathbf{x}) \equiv Beta(44, 58),$$

a Bayes estimate for θ, the probability of obtaining an A, is

$$\theta_{mean} = E(\theta|\mathbf{x}) = \frac{44}{102} = 0.4313.$$

```
> a.star/(a.star+b.star)
[1] 0.4313725
```

Posterior Median

The posterior median can also be used as a Bayes estimate for θ. The posterior median is the value of θ, θ_{median}, such that

$$P(\theta \leq \theta_{median}|\mathbf{x}) = \frac{1}{2}.$$

The function qbeta from R can be used to obtain the median)

```
> qbeta(0.5,a.star,b.star)
[1] 0.4309223
```

and we have

$$\theta_{median} = 0.4309.$$

Posterior Mode

This is the mode of the posterior distribution, ie, the value θ_{mode} such that $p(\theta|\mathbf{x})$ attains its maximum, ie

$$p(\theta_{mode}|\mathbf{x}) = \max_{\theta \in \Theta} p(\theta|\mathbf{x}).$$

Again θ_{mode} can be used as a Bayes estimate for θ. In the example

$$\theta_{mode} = \frac{\sum x_i + a - 1}{n + a + b - 2} = \frac{43}{100} = 0.43.$$

```
> (a.star-1)/(a.star+b.star-2)
[1] 0.43
```

Credible Interval

We can compute a credible interval with equal probability tails, ie, by considering for the lower interval the quantile $\alpha/2 = 0.025$ of the Beta distribution and for the upper interval the quantile $(1 - \alpha/2 = 0.975)$. We can do this using function `qbeta` from `R`.

```
> qbeta(c(0.025,0.975),44,58)
[1] 0.3372088 0.5280864
```

Hence our 95% credible interval for θ would be the interval $[0.337, 0.528]$ and we could say that

$P(0.337 < \theta < 0.528| \sum_{i=1}^{100} x_i = 43) = 0.95).$

In summary

Given data $\mathbf{x} = (x_1, \ldots, x_n)$ obtained under a parametric model $f(\mathbf{x}|\theta)$ indexed by finite-dimensional parameter $\theta \in \Theta \subset \mathcal{R}^k$,

Bayesian inferences on the parameter are based on the posterior distribution

$$p(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)p(\theta)}{\int f(\mathbf{x}|\xi)p(\xi)d\xi}$$

which is obtained via the familiar form of the Bayes theorem, relating the

- posterior distribution $p(\theta|\mathbf{x})$
- to the likelihood $f(\mathbf{x}|\theta)$
- and the prior distribution $p(\theta)$.

Acknowledgements:

Antónia Turkman (DEIO – FCUL) and Carina Silva-Fortes (ESTeSL – IPL), for allowing the use of some material produced by us in previous courses.

Bibliography:

Carlin, B.P. and Louis, T.A. (2002). *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd edition. London: Chapman & Hall.

Congdon, P. (2001). *Bayesian Statistical Modelling*. Chichester: John Wiley & Sons.