Bayesian modelling

Léo Belzile

Table of contents

Welcome

This book is a web complement to MATH 80601A *Bayesian modelling*, a graduate course offered at HEC Montréal.

The objective of the course is to provide a hands on introduction to Bayesian data analysis. The course will cover the formulation, evaluation and comparison of Bayesian models through examples and real-data applications.

1 Introduction

1.1 Bayes theorem

Denote by $p(X) \equiv \Pr(X)$ denotes the marginal density of X, $p(X \mid Y)$ the conditional of X given Y and $p(X, Y)$ the joint density. Bayes' theorem states that

$$p(X = x \mid Y = y) = \frac{p(Y = y \mid X = x)p(X = x)}{p(Y = y)}$$

In the case of discrete random variable X with support $\mathcal{X}$, the denominator can be evaluated using the law of total probability as

$$\Pr(Y = y) = \sum_{x \in \mathcal{X}} \Pr(Y = y \mid X = x) \Pr(X = x).$$

Example 1.1. Back in January 2021, the Quebec government was debating whether or not to distribute antigen rapid test, with strong reluctance from authorities given the paucity of available resources and the poor sensitivity.

A Swiss study analyse the efficiency of rapid antigen tests, comparing them to repeated polymerase chain reaction (PCR) test output, taken as benchmark (Jegerlehner et al. 2021). The results are presented in Table 1.1

Table 1.1: Confusion matrix of Covid test results for PCR tests versus rapid antigen tests, from Jegerlehner et al. (2021).

	PCR +	PCR −
rapid +	92	2
rapid −	49	1319
total	141	1321

Estimated seropositivity at the end of January 2021 according to projections of the Institute for Health Metrics and Evaluation (IHME) of 8.18M out of 38M inhabitants (Mathieu et al.

2020), a prevalence of 21.4%. Assuming the latter holds uniformly over the country, what is the probability of having Covid if I get a negative result to a rapid test?

Let R^- (R^+) denote a negative (positive) rapid test result and C^+ (C^-) Covid positivity (negativity). Bayes' formula gives

$$\begin{aligned}\Pr(C^+ \mid R^-) &= \frac{\Pr(R^- \mid C^+)\Pr(C^+)}{\Pr(R^- \mid C^+)\Pr(C^+) + \Pr(R^- \mid C^-)\Pr(C^-)} \\ &= \frac{49/141 \cdot 0.214}{49/141 \cdot 0.214 + 1319/1321 \cdot 0.786}\end{aligned}$$

so there is a small, but non-negligible probability of 8.66% that the rapid test result is misleading. Jegerlehner et al. (2021) indeed found that the sensitivity was 65.3% among symptomatic individuals, but dropped down to 44% for asymptomatic cases. This may have fueled government experts skepticism.

1.2 Probability and frequency

In classical (frequentist) parametric statistic, we treat observations $\boldsymbol{Y}$ as realizations of a distribution whose parameters $\boldsymbol{\theta}$ are unknown. All of the information about parameters is encoded by the likelihood function, which is optimized numerically or analytically to find the maximum likelihood estimator. Large-sample theory shows that the resulting estimator is asymptotically normal under regularity conditions.

The interpretation of probability in the classical statistic is understood in terms of long run frequency, which is why we call this approach frequentist statistic. Think of a fair die: when we state that values $\{1,\dots,6\}$ are equiprobable, we mean that repeatedly tossing the die should result, in large sample, in each outcome being realized roughly $1/6$ of the time (the symmetry of the object also implies that each facet should be equally likely to lie face up). This interpretation also carries over to confidence intervals: a $(1-\alpha)$ confidence interval either contains the true parameter value or it doesn't, so the probability level $(1-\alpha)$ is only the long-run proportion of intervals created by the procedure that should contain the true fixed value, not the probability that a single interval contains the true value. This is counter-intuitive to most.

In practice, the true value of the parameter $\boldsymbol{\theta}$ vector is unknown to the practitioner, thus uncertain: Bayesians would argue that we should treat the latter as a random quantity rather than a fixed constant. Since different people may have different knowledge about these potential values, the prior knowledge is a form of subjective probability. For example, if you play cards, one person may have recorded the previous cards that were played, whereas other may not. They thus assign different probability of certain cards being played. In

Bayesian inference, we consider $\boldsymbol{\theta}$ as random variables to reflect our lack of knowledge about potential values taken. Italian scientist Bruno de Finetti, who is famous for the claim "Probability does not exist' ', stated in the preface of Finetti (1974):

> Probabilistic reasoning — always to be understood as subjective — merely stems from our being uncertain about something. It makes no difference whether the uncertainty relates to an unforseeable future, or to an unnoticed past, or to a past doubtfully reported or forgotten: it may even relate to something more or less knowable (by means of a computation, a logical deduction, etc.) but for which we are not willing or able tho make the effort; and so on [...] The only relevant thing is uncertainty — the extent of our knowledge and ignorance. The actual fact of whether or not the events considered are in some sense *determined*, or known by other people, and so on, is of no consequence.

On page 3, de Finetti continues (Finetti 1974)

> only subjective probabilities exist — i.e., the degree of belief in the occurrence of an event attributed by a given person at a given instant and with a given set of information.

The likelihood $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) \equiv p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the starting point for Bayesian inference. However, we adjoin to it a **prior** distribution $p(\boldsymbol{\theta})$ that reflects the prior knowledge about potential values taken by the p-dimensional parameter vector, before observing the data $\boldsymbol{y}$. We thus seek $p(\boldsymbol{\theta} \mid \boldsymbol{y})$: the observations are random variables but inference is performed conditional on the observed sample. By Bayes' theorem, the posterior distribution $p(\boldsymbol{\Theta} \mid \boldsymbol{Y})$ is

$$p(\boldsymbol{\Theta} \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{\Theta}) p(\boldsymbol{\Theta})}{\int p(\boldsymbol{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}, \tag{1.1}$$

so the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is proportional, as a function of θ, to the product of the likelihood and the prior function. The integral in the denominator, termed marginal likelihood and denoted $p(\boldsymbol{Y}) = \mathsf{E}_{\boldsymbol{\theta}}\{p(\boldsymbol{Y} \mid \boldsymbol{\theta})\}$, is a normalizing constant that makes the right hand side integrate to unity.

For the posterior to be **proper**, we need the product on the right hand side to be integrable. The denominator of Equation 1.1 is a normalizing constant so that the posterior is a distribution. If $\boldsymbol{\theta}$ is low dimensional, numerical integration such as quadrature methods can be used to compute the latter. To obtain the marginal posterior,

$$p(\theta_j \mid \boldsymbol{y}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}_{-j},$$

additional integration is needed.

Consider a binomial likelihood with probability of success p and n trials, $Y \sim \mathsf{Bin}(n,p)$. If we take a beta prior, $p \sim \mathsf{Be}(\alpha,\beta)$ and observe k successes, the posterior is

$$\begin{aligned} p(\theta \mid y=k) &\propto \binom{n}{k} p^k(1-p)^{n-k} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\ &\overset{p}{\propto} p^{k+\alpha-1}(1-p)^{n-k+\beta-1} \end{aligned}$$

and the normalizing constant is

$$\int_0^1 p^{k+\alpha-1}(1-p)^{n-k+\beta-1}\mathrm{d}p = \frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta)},$$

a Beta function. Since we need only to keep track of the terms that are function of the parameter p, we could recognize directly that the posterior distribution is $\mathsf{Be}(k+\alpha, n-k+\beta)$.

If the sample size n grows, then the number of success should be roughly np and the number of failures $n(1-p)$ and so the likelihood contribution, relative to the prior, will dominate. The Beta distribution, whose density is $f(x;\alpha,\beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$, has expectation $\alpha/(\alpha+\beta)$ and variance $\alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$. An alternative parametrization takes $\alpha = \mu\kappa$, $\beta = (1-\mu)\kappa$ for $\mu \in (0,1)$ and $\kappa > 0$, so that the model is parametrized directly in terms of mean μ.

While a density integrates to 1 over the range of possible outcomes, the likelihood function does not when we integrate over the range of the parameters.

We call a prior *proper* if it's integral is finite: the best example is priors that arise from probability density function. We can still employ this rule for improper priors: for example, taking $\alpha, \beta \to 0$ in the Beta prior leads to a prior proportional to $x^{-1}(1-x)^{-1}$, the integral of which diverges on the unit interval $[0,1]$. However, as long as the number of success and the number of failures is larger than 1, meaning $k \geq 1, n-k \geq 1$, the posterior distribution would be proper, i.e., integrable. To find the posterior, normalizing constants are also superfluous.

The beta-binomial model is an example of conjugate model, meaning the posterior distribution is from the same family as the prior.[1] While we could calculate analytically the value of the normalizing constant, we could also in more complicated models use numerical integration in the event the parameter vector $\boldsymbol{\theta}$ is low-dimensional. For a single scalar p on the unit interval, numerical integration or Monte Carlo integration yield nearly identical results.

[1] This is a property of exponential families that will be revisited in the next chapter.

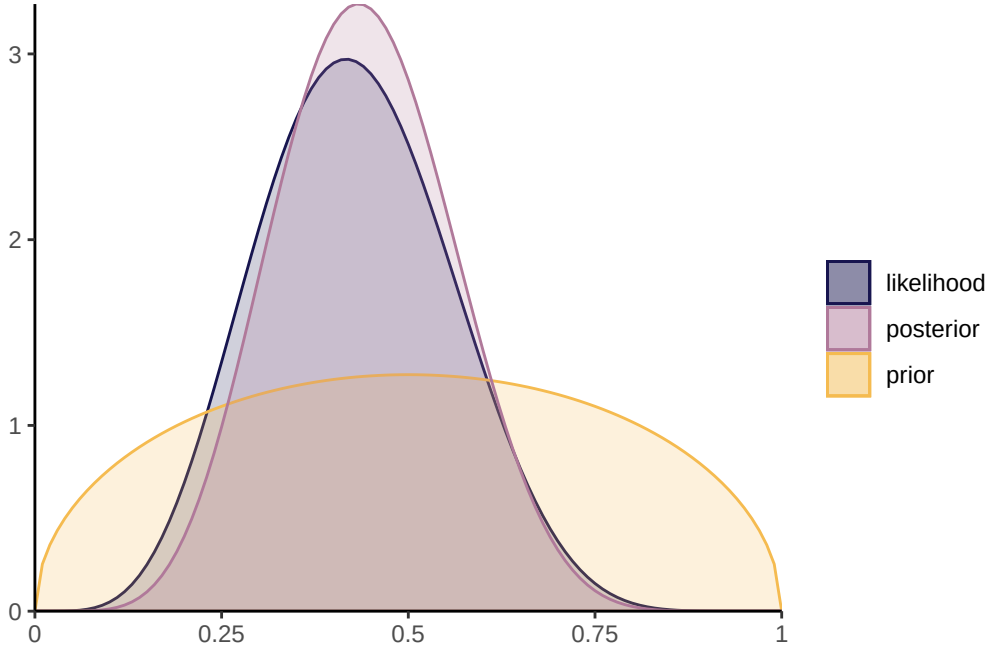

Figure 1.1: Binomial likelihood for six successes out of 14 trials, $\text{Beta}(3/2, 3/2)$ prior and posterior distribution from a beta-binomial model. The posterior curve is much closer to the likelihood than it is to the prior, even with a relatively small sample size.

```
k <- 6L # number of successes
n <- 14L # number of trials
alpha <- beta <- 1.5 # prior parameters
unnormalized_posterior <- function(p){
  p^(k+alpha-1) * (1-p)^(n-k + beta - 1)
}
integrate(f = unnormalized_posterior,
          lower = 0,
          upper = 1)
```

```
1.066906e-05 with absolute error < 1e-12
```

```
# Compare with known constant
beta(k + alpha, n - k + beta)
```

```
[1] 1.066906e-05
```

```
# Monte Carlo integration
mean(unnormalized_posterior(runif(1e5)))
```

```
[1] 1.064055e-05
```

When $\boldsymbol{\theta}$ is high-dimensional, the marginal likelihood is untractable. This is one of the main challenges of Bayesian statistics and the popularity and applicability has grown drastically with the development and popularity of numerical algorithms Gelfand and Smith (1990). Markov chain Monte Carlo methods circumvent the calculation of the denominator by drawing approximate samples from the posterior.

1.2.1 Bayesian updating

Subjective probabilities imply that different people with different prior beliefs would arrive at different conclusions. However, as more data are gathered, we can use Bayes theorem to update these prior beliefs and update the posterior. In most instances, the relative weight of the prior relative to the likelihood becomes negligible: if we consider independent data $\boldsymbol{y}_1, \boldsymbol{y}_n$ observed sequentially, then

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) &\overset{\boldsymbol{\theta}}{\propto} p(\boldsymbol{y}_k \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k-1}) \\ &\overset{\boldsymbol{\theta}}{\propto} \prod_{i=1}^{k} p(\boldsymbol{y}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \end{aligned}$$

If data are exchangeable, the order in which observations are collected and the order of the belief updating is irrelevant to the full posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots \boldsymbol{y}_n)$.

2 Priors

References

Finetti, Bruno de. 1974. *Theory of Probability: A Critical Introductory Treatment*. Vol. 1. New York: Wiley.

Gelfand, Alan E., and Adrian F. M. Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85 (410): 398–409. https://doi.org/10.1080/01621459.1990.10476213.

Geman, Stuart, and Donald Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6 (6): 721–41. https://doi.org/10.1109/TPAMI.1984.4767596.

Jegerlehner, Sabrina, Franziska Suter-Riniker, Philipp Jent, Pascal Bittel, and Michael Nagler. 2021. "Diagnostic Accuracy of a SARS-CoV-2 Rapid Antigen Test in Real-Life Clinical Settings." *International Journal of Infectious Diseases* 109 (August): 118–22. https://doi.org/10.1016/j.ijid.2021.07.010.

Mathieu, Edouard, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Joe Hasell, Bobbie Macdonald, et al. 2020. "Coronavirus Pandemic (COVID-19)." *Our World in Data*.