STA5904: Bayesian Modeling and Inference **Topic 1**

Introduction to Bayesian Inference

1 History of Bayesian Inference

Bayesian inference has a long history, beginning with reverend Thomas Bayes' discovery of *Bayes rule*. In modern notation, Bayes' rule takes the form

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \propto p(\theta)p(y \mid \theta),$$

where θ and y are random variables with joint density $p(\theta)\, p(y \mid \theta)$. Bayes lived in the 18^{th} century, and his results on probability were published post-humously. Bayes focused, in particular, on the particular problem of inferring a binomial probability from data, i.e., he studied the case where $Y \sim \text{Binomial}(n, \theta)$ when $\theta \sim \text{Uniform}(0, 1)$.

In the late 18^{th} century, the mathematician Pierre-Simon Laplace rediscovered and greatly generalized Bayes rule to the form which we now use.

Validity of Bayes Rule While the story of Bayesian inference begins with the developement of Bayes rule, it is important not to conflate the use of Bayes rule with Bayesian inference in general. Bayes rule is a theorem of probability theory, and nobody (Frequentist or Bayesian) doubts its validity. The tension between Frequentists and Bayesians is centered instead around when it is valid to *use* Bayes rule.

1.1 Performing Bayesian Inference

At this stage in the history of statistics, there was no philosophical debate about "Frequentist" and "Bayesian" statistics. Bayes intended his rule to be a principle for taking prior beliefs (or probability) and updating it to reflect new evidence.

In order to perform Bayesian inference, then, we will need to quantify our prior beliefs about the parameter θ of interest. This stands in contrast to Frequentist methods of inference, wherein θ is thought of as being a *fixed, unknown, quantity*. Whereas uncertainty in θ is quantified by the Frequentist using tools such as confidence intervals and error rates, Bayes and Laplace treated θ in much the same way they treated the data y, and quantified uncertainty in θ using probability distributions over θ.

Hence, to perform inference we must first specify

- a *likelihood* $L(\theta) = p(y \mid \theta)$ describing the distribution of the data Y given the parameter θ; and
- a *prior* $\pi(\theta)$, encoding our prior beliefs about the likely values of θ

To illustrate the principles, we will consider one of Laplace's first applications of Bayes' rule.

Example 1. It had been noticed by the time Laplace developed his theory that birth records indicated that there were more males being born than females. Laplace was interested in establishing whether or not this

was likely to be either (i) a coincidence or (ii) a general rule. Laplace had access to birth records in Paris between 1745 and 1770, and found that there were $n = 493,472$ births, of which $y = 241,945$ of them were females.

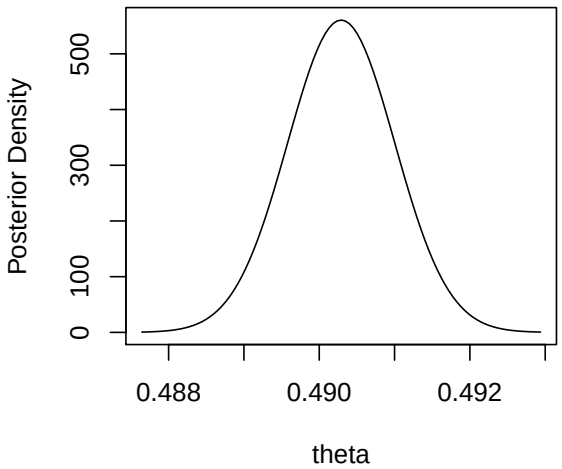
A reasonable likelihood for this problem is to model y as the realization of $Y \sim \text{Binomial}(n, \theta)$ conditional on the parameter θ (the probability of a female birth). To specify a prior for θ, Laplace reasoned that *if he were a-priori ignorant about the likely values of* θ, then $\theta \sim \text{Uniform}(0, 1)$ would make sense.

Taking these as given, we can now compute the posterior $\pi(\theta \mid y)$ as

$$\pi(\theta \mid y) \propto \pi(\theta)\, p(y \mid \theta) \propto \pi^y (1-\pi)^{n-y} I(0 \leq \theta \leq 1).$$

This is the kernel of a $\text{Beta}(y+1, n-y+1)$ random variable. In Laplace's case, this leads to the posterior $\theta \sim \text{Beta}(241946, 251528)$. A picture of the posterior density of θ is given below.

```
## Laplace's Example ------------------------------------------------------

num_female <- 241945
num_total  <- 493472
num_male   <- num_total - num_female

## Make posterior
alpha <- num_female + 1
beta  <- num_male + 1

## Plot posterior ---------------------------------------------------------

xlim <- qbeta(c(0,1) + c(1,-1) * 1E-4, alpha, beta)

plot(function(x) dbeta(x, alpha, beta), xlim = xlim,
     xlab = "theta", ylab = "Posterior Density")
```

Computations for this example are carried out in the `R` code for this topic. We find that, under the Uniform$(0, 1)$ prior, the probability that more boys than girls are born is

$$\Pr(\theta < 0.5 \mid Y = y) = 1 - 1.14 \times 10^{-42}.$$

Laplace concluded that he was "morally certain" that male births are more common than female births.

Based on this prior, we can also find endpoints (L, U) such that we are (say) 99% certain that $\theta \in (L, U)$. In this case, we have $\Pr(0.4885 \leq \theta \leq 0.4921) \approx 99\%$. This is very close to $\theta = 0.5$, but evidently we have enough data to rule this out.

2 The Principle of Indifference

The argument above is based on the uniform prior for θ. This encodes the prior belief that all values of θ are, in some sense, equally likely. The choice $\theta \sim \text{Uniform}(0, 1)$ motivated by what is called the *principle of indifference*, which we state below.

> *Given a set Θ consisting of n elements, exactly one of which describes nature θ_0 and are indistinguishable except for their names, then each $\theta \in \Theta$ is equally likely sense that $\Pr(\theta = \theta_0) = 1/n$.*

Of course, we note that this principle requires essentially no prior knowledge to apply; this is not typical of most real-world problems. Nevertheless, it may often be reasonable to ask what inference would be obtained by someone who truly is ignorant.

Now, if we consider a grid $\mathcal{G}_n$ of points evenly-spaced on $[0, 1]$ then, as $n \to \infty$, we arrive at the Uniform$(0, 1)$ prior. In deriving the Uniform$(0, 1)$ prior we have made a somewhat unjustified leap: we have assumed that the grid on $[0, 1]$ is equally-spaced! Had we not made this assumption, but instead assumed an equal spacing on $\psi = \log\{\theta/(1-\theta)\}$, we in fact would have arrived at a completely different prior! Moreover, when applied to ψ, an evenly spaced grid would have to cover all of $(-\infty, \infty)$, leading to a distribution $\pi(\psi) \propto 1$, which cannot be normalized to a probability distribution.

These problems with the principle of indifference were used by later statisticians to criticize Bayesian methods and promote Frequentist methods. The parameterization-invariance would be addressed much later by Harold Jefferys with the development of the Jeffreys prior.

3 From Laplace to Gauss and Least Squares

Laplace developed and applied his theory of statistical inference to many problems, leading ultimately to him establishing the *central limit theorem*. He was particularly interested in performing inference in astronomy.

Around the same time, the great mathematician Carl Friedrich Gauss was tackling the problem of computing the orbits of celestial bodies. Ultimately, this problem can be cast in the form

$$Y = X\beta + \epsilon, \qquad \epsilon \sim G,$$

where X is a known $N \times P$ design matrix, $\beta \in \mathbb{R}^P$ is a parameter of interest, $Y \in \mathbb{R}^N$ is a vector of responses, and ϵ is an error vector with independent components.

At the time, the fundamental role of the Gaussian distribution had not been established, and it was not clear what the natural distribution of the errors would be. Laplace considered a double-exponential error

distribution

$$g(\epsilon_i) = \frac{1}{2\sigma} e^{-|\epsilon_i|/\sigma},$$

which is now sometimes called the *Laplace distribution*. Gauss, instead, realized the importance of the Gaussian distribution for the errors

$$g(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\epsilon_i^2/(2\sigma^2)}.$$

Gauss then appealed to the principle of indifference to essentially specify a prior distribution for β of the form $\pi(\beta) = 1$. As noted above, this $\pi(\beta)$ has the unfortunate problem of not being a probability density because $\int \pi(\beta)\, d\beta = \infty$. Such a prior distribution is referred to as an *improper prior*.

Despite the use of an improper prior, we can still appeal *formally* to Bayes rule and compute the posterior distribution anyway. A formal application of Bayes rule gives

$$\begin{aligned}\pi(\beta \mid y) &\propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - x_i^\top \beta)^2/(2\sigma^2)} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right\}.\end{aligned}$$

Based on this, Gauss took as an estimate of β the maximizer of this expression, or equivalently

$$\widehat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2 = (X^\top X)^{-1} X^\top Y.$$

That is, we have arrived at the least-squares estimate of β! While, of course, the reader will know that $\widehat{\beta}$ can be motivated as a maximum likelihood estimator, the original derivation was Bayesian.

We can do more here. Letting $\widehat{\beta} = (X^\top X)^{-1} X^\top Y$, more manipulations with the posterior distribution show that

$$\pi(\beta \mid y) \propto \exp\left\{-\frac{1}{2}(\widehat{\beta} - \beta)^\top [X^\top X/\sigma^2]^{-1} (\widehat{\beta} - \beta)\right\}. \tag{1}$$

Hence, we have the posterior distribution $\beta \sim \text{Normal}(\widehat{\beta}, \sigma^2[X^\top X]^{-1})$. Note the similarity with the usual sampling distribution $\widehat{\beta} \sim \text{Normal}(\beta, \sigma^2[X^\top X]^{-1})$! Evidently, Bayesian inference with a $\text{Uniform}(-\infty, \infty)$ prior on β reproduces the classical results.