1 Philosophy of Bayesian vs. Frequentist Statistics

Broadly, statistics can be divided into two schools of thought: frequentist statistics and Bayesian statistics.

Up to this point in the course, we have been learning frequentist statistics, which defines a probability as the long-term relative frequency of an event. Here are a few examples:

- **Bernoulli RV:** If we flip a coin, then the probability of heads ($X = 1$) is .5; this is because if we were to keep flipping the coin repeatedly, the number of heads divided by the number of flips would approach .5. In fact, Karl Pearson famously flipped a coin 24,000 time to record the number of heads. The expectation $\mathbb{E}(X) = p = .5$ is the long-term average value of X over all these flips. Similarly, the variance is the long-term average squared deviation of X from $\mathbb{E}(X)$, etc.
- **Gamma RV:** The amount of rain in a day (assuming it rains) Y_t has often be modeled as a Gamma distribution with shape parameter α and scale parameter β. In the frequentist view, this is saying that the probability of the event that $Y_t \leq y_t$ is $\mathbb{P}(Y_t \leq y_t) = F_{\alpha,\beta}(y_t)$, where F is the CDF. The meaning of probability here is that if we were to repeatedly observe rainfall amounts over time, the long-term pattern is that $\mathbb{P}(Y_t \leq y_t) = F_{\alpha,\beta}(y_t) = \int_0^\infty f_{\alpha,\beta}(y_t)dy_t$. The expectation of Y_t is $\alpha\beta$, and the variance is $\alpha\beta^2$.
- **CI RV** A $(1-\alpha)100\%$ confidence interval $C(X)$ was defined as an RV such that $\mathbb{P}(\theta \in C(X)) \geq 1-\alpha$. This probability refers to the long-term relative frequency that θ would be captured in the interval, i.e., the number of intervals $C(X)$ containing θ divided by the number of intervals considered.

Because of this definition of probability, parameters like p, α, and β are fixed, non-random values that are unknown in practice. As a result, we must estimate them by collecting a random sample from the population and calculating an estimator based on it. Other parameters like $\mathbb{E}(X)$ might be of interest too, which we would try to similarly estimate. And of course we can also provide a range of values for our estimates through confidence intervals.

In Bayesian statistics, probability is viewed as a subjective measure of personal belief. In this framework, one must describe/quantify their personal belief by specifying a prior distribution $f(\theta)$ for the distributional parameter θ; therefore, θ is treated as random. We will also have a distribution that describes the behavior of the RV(s) X. In Bayesian statistics, we call this the likelihood and write it as $f(x|\theta)$. Notice that we are conditioning on θ, since it's considered random here. (In contrast, when we were covering frequentist statistics in this course, we liked to write $f_\theta(x)$ or $f(x;\theta)$ because θ was considered non-random). Finally, we will have the posterior distribution $f(\theta|x)$, which describes our updated view of θ, i.e.,

what we believe about θ in light of the data we observed. In theory, if we have the posterior distribution, then we have all the information we could possibly want about θ. For example, we could calculate the posterior mean $\mathbb{E}(\theta|X)$, the maximum a posteriori probability (MAP) estimator, and credible intervals (the Bayesian analogue to frequentist confidence intervals).

How do we calculate the posterior distribution? Bayes' Rule tells us that

$$f(\theta|X) = \frac{f(x|\theta)f(\theta)}{f(x)}.$$

Note that $f(x)$ is the marginal distribution of X and in practice may be difficult to calculate, since $f(x) = \int f(x|\theta)f(\theta)d\theta$. But since $f(x)$ doesn't depend on θ, we don't actually need to calculate it – proportionality is enough:

$$f(\theta|X) \; \alpha_\theta \; f(x|\theta)f(\theta).$$

How do Bayesian and frequentist approaches compare? Here is a theorem about estimation:

Theorem 1.1. Let $\hat{\theta}_n$ be the MLE with $\hat{se} = \frac{1}{\sqrt{I_n(\theta)}}$. Then the posterior is asymptotically normal with mean $\hat{\theta}$ and standard deviation $\hat{se}$.

As a consequence, as the sample size n increases, the posterior mean converges to the MLE. This makes sense intuitively, since as the amount of data increases, one's prior belief about θ should become less important. But in general, Bayesian and frequentist approaches can have quite different results and interpretations.

2 Choosing a Prior

There are various approaches, giving rise to different types of Bayesianism:

(a) **Subjective Bayesianism:.** The researcher specifies their prior knowledge of and uncertainties about the problem

(b) **Objective Bayesianism:.** The prior should be computed according to a system that anyone can follow, i.e., it should not contain information from the researcher's subjective beliefs

- A few examples are uniform (flat) priors, Jeffreys' prior, and reference priors. We allow these priors to be improper, i.e., not integrating to 1

(c) **Robust Bayesianism:.** The prior should be chosen by exploring how it will affect the final results

2.1 Jeffreys' Prior

Not all methods for objective Bayesian are invariant to transformations. That is, if we have a parameter θ and $\phi = g(\theta)$, then invariance would mean that the change-of-variables formula holds:

$$f_\phi(\phi) = f_\theta(g^{-1}(\phi))\left|\frac{dg^{-1}(\phi)}{d\phi}\right|.$$

Intuitively, wanting invariance is wanting the prior for a transformation to be mapped "correctly" (in some sense) from the original prior. [Of course, some people may have varying views of "correctness" for a prior/transformation of a prior – but invariance is one notion of "correctness".]

A benefit of Jeffreys' prior is that it satisfies invariance.

Definition 2.1. Jeffreys' prior. If θ is 1-dimensional, then Jeffreys' prior is

$$f(\theta) \; \alpha \; I_n(\theta)^{1/2}.$$

3 Calculation of Posterior Distribution

For conjugate prior distributions, the calculations can be done in closed form.

Definition 3.1. Conjugate Prior Distribution. A conjugate prior distribution for θ is one for which the prior $f(\theta)$ and the posterior $f(\theta|x)$ belong to the same parametric family.

In practice, we can rarely calculate the posterior distribution in closed form. However, we can approximate the posterior distribution numerically, thanks to advances in computing.

In such a case, we can perform a Monte Carlo approximation to the posterior mean. The challenge is that we cannot necessarily sample from the posterior distribution. There are various algorithms that handle this, including rejection sampling, importance sampling, and Markov Chain Monte Carlo (MCMC).

Problem 1. Conjugate Priors

Exercises 1 and 2 from the Lecture 5 notes.

(a) Suppose $X_1, \ldots, X_n|\lambda \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$, and the prior is $\lambda \sim \text{Gamma}(a, b)$. Find the posterior distribution for λ.

(b) Suppose $X_1, \ldots, X_n|\sigma^2 \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$, where θ is known. Let the prior distribution for σ^2 be inverse gamma with parameters a and b. The prior PDF is

$$f(\sigma^2; a, b) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}\exp(-\beta/\sigma^2).$$

Find the posterior distribution for σ^2.

Solution

Part (a)

The joint likelihood is

$$\begin{aligned} f(x_1, \ldots, x_n|\lambda) &= \prod_{i=1}^{n} \frac{\lambda^{x_i}e^{-\lambda}}{x_i!} \text{ (since the data are iid)} \\ &= \frac{\lambda^{\sum x_i}e^{-n\lambda}}{\prod x_i!}, \end{aligned}$$

and the prior density is

$$f(\lambda; a, b) = \frac{1}{\Gamma(a)b^a}\lambda^{a-1}e^{-\lambda/b}.$$

So the posterior is

$$\begin{aligned} f(\lambda|X_1, \ldots, X_n) &\propto f(x_1, \ldots, x_n|\lambda)f(\lambda) \\ &= \frac{\lambda^{\sum x_i}e^{-n\lambda}}{\prod x_i!} \cdot \frac{1}{\Gamma(a)b^a}\lambda^{a-1}e^{-\lambda/b} \\ &\propto \lambda^{\sum x_i}e^{-n\lambda}\lambda^{a-1}e^{-\lambda/b} \\ &= \lambda^{\sum x_i + a - 1}e^{-\lambda(n+\frac{1}{b})}. \end{aligned}$$

The exponential term can be re-expressed as

$$\begin{aligned} e^{-\lambda(n+\frac{1}{b})} &= e^{-\lambda(\frac{nb+1}{b})} \\ &= e^{-\lambda/(\frac{b}{nb+1})}. \end{aligned}$$

Thus, the posterior is

$$f(\lambda|X_1,\ldots,X_n) \;\alpha\; \lambda^{\sum x_i+a-1}e^{-\lambda/(\frac{b}{nb+1})},$$

and so $\lambda|X_1,\ldots,X_n \sim \text{Gamma}\left(a+\sum X_i, \frac{b}{nb+1}\right)$.

Part (b)

The joint likelihood is

$$\begin{aligned} f(x_1,\ldots,x_n|\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(x_i-\theta)^2\right) \text{ (since the data are iid)} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(\frac{-1}{2\sigma^2}\sum(x_i-\theta)^2\right). \end{aligned}$$

The posterior then is

$$\begin{aligned} f(\sigma^2|X_1,\ldots,X_n) \;\alpha\; \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(\frac{-1}{2\sigma^2}\sum(x_i-\theta)^2\right) \cdot f(\sigma^2;a,b) &= \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}\exp(-\beta/\sigma^2) \\ \alpha\; \sigma^{-n} \exp\left(\frac{-1}{2\sigma^2}\sum(x_i-\theta)^2\right)(\sigma^2)^{-a-1}\exp(-\beta/\sigma^2) \\ = (\sigma^2)^{-(a+n/2)-1} \exp\left(\frac{-1}{\sigma^2}\left(\frac{1}{2}\sum(x_i-\theta)^2+b\right)\right), \end{aligned}$$

so we have

$$\sigma^2|X_1,\ldots,X_n \sim \text{InvGamma}\left(a+\frac{n}{2}, \frac{1}{2}\sum(x_i-\theta)^2+b\right).$$

Problem 2. Jeffreys' Prior

Exercises 5-8 in the Lecture 5 notes.

(a) Find the Jeffreys' prior for λ when $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$.

(b) Is the Jeffreys' prior proper?

(c) Find the implied prior distribution for $\phi = \log(\lambda)$.

(d) Show that the prior in part (c) is the same as the Jeffreys' prior for ϕ.

Solution

Part (a)

The individual likelihood is

$$f(x_i|\lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

so the corresponding log-likelihood is

$$l_i(\lambda) = x_i \log(\lambda) - \lambda - \log(x_i!)$$

Its first and second derivatives are

$$l_i'(\lambda) = \frac{x_i}{\lambda} - 1$$
$$l_i''(\lambda) = \frac{-x_i}{\lambda^2}.$$

Thus, the Fisher Information for one observation is

$$I(\lambda) = -\mathbb{E}(l_i''(\lambda)) = \frac{\mathbb{E}(X_i)}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}.$$

Finally, the Jeffreys' prior is

$$f(\lambda) \; \alpha \; I(\lambda)^{1/2} = \frac{1}{\lambda^{1/2}}.$$

Part (b)

The Jeffreys' prior would be proper if it integrates to 1 like standard PDFs are supposed to.

Integrating, we have

$$\begin{aligned}\int_0^\infty f(\lambda)d\lambda &= \int_0^\infty \frac{1}{\lambda^{1/2}} \\ &= 2\sqrt{\lambda}\Big|_0^\infty \\ &= 2\left(\lim_{\lambda\to\infty}\sqrt{\lambda} - 0\right) \\ &= \infty,\end{aligned}$$

so this prior is improper.

Part (c)

Let us apply the change-of-variables formula

$$f_\phi(\phi) = f_\theta(g^{-1}(\phi))\left|\frac{dg^{-1}(\phi)}{d\phi}\right|,$$

where $\phi = \log(\lambda) = g(\lambda)$ and $\theta = \lambda$.

We have $g^{-1}(\phi) = e^\phi$, so

$$\begin{aligned}&f_\lambda(e^\phi)\ \alpha\ \frac{1}{(e^\phi)^{1/2}} = \frac{1}{e^{\phi/2}} \text{ and} \\ &\frac{d}{d\phi}e^\phi = e^\phi.\end{aligned}$$

Plugging it all in, the Jeffrey's prior for ϕ is

$$f_\phi(\phi)\ \alpha\ \frac{1}{e^{\phi/2}}e^\phi = e^{\phi/2}.$$

Part (d)

Now we shall check that the change-of-variables formula gave us the same result as if we "from scratch" computed the Jeffreys' prior for ϕ.

The likelihood for one observation can be written as

$$f(x_i|\lambda = e^\phi) = \frac{e^{\phi x_i}e^{-e^\phi}}{x_i!},$$

and the corresponding log-likelihood is

$$l_i(\phi) = \phi x_i - e^{\phi} - \log(x_i!).$$

The first and second derivatives with respect to ϕ are

$$l_i'(\phi) = x_i - e^{\phi}$$
$$l_i''(\phi) = -e^{\phi},$$

so the Fisher Information for one observation is

$$I(\phi) = -\mathbb{E}(l_i''(\phi)) = e^{\phi}.$$

Thus, the Jeffreys' prior is

$$f(\phi) = I(\phi)^{1/2} = e^{\phi/2},$$

which is exactly what we got in the previous part. This works because the Jeffreys' prior is invariant to parametrizations.

Problem 3. Numerical Example

Assume the same set-up as Problem 1(a). Suppose $n = 1$ (for simplicity). That is, we have $X \sim \text{Poisson}(\lambda)$ and $\lambda \sim \text{Gamma}(a, b)$. Find the approximate posterior through rejection sampling. Compare this approximate posterior to the closed-form posterior we calculated in 1(a). Compare the posterior means as well.

Solution

See R code.