1.5: Bayesian Statistical Inference

Frequentists and Bayesians

There field of statistical inference is roughly divided into two schools of practice: frequentism and Bayesianism. You can find several articles on the web discussing the differences in these two approaches. In both cases we would like to model observable data $\mathbf{x}$ by a distribution. The distribution in general depends on one or more parameters θ. The basic worldviews of the two approaches are as follows:

> ***Frequentism:*** *Data $\mathbf{x}$ are a random sample drawn from an infinite pool at some frequency. The underlying parameters θ, which are to be estimated, remain fixed during this process. There is no information prior to the model specification. The experimental conditions under which the data are collected are presumed to be controlled and repeatable. Results are generally expressed in terms of confidence intervals and confidence levels, obtained via statistical hypothesis testing. Probabilities have meaning only for data yet to be collected. Calculations generally are computationally straightforward.*

> ***Bayesianism:*** *The only data $\mathbf{x}$ which matter are those which have been observed. The parameters θ are unknown and described probabilistically using a prior distribution, which is generally based on some available information but which also may be at least partially subjective. The priors are then to be updated based on observed data $\mathbf{x}$. Results are expressed in terms of posterior distributions and credible intervals. Calculations can be computationally intensive.*

In essence, frequentists say *the data are random and the parameters are fixed*. while Bayesians say *the data are fixed and the parameters are random*[11]. Overall, frequentism has dominated over the past several hundred years, but Bayesianism has been coming on strong of late, and many physicists seem naturally drawn to the Bayesian perspective.

Updating Bayesian priors

Given *data* D and a *hypothesis* H, Bayes' theorem tells us

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}\ . \tag{1.5.1}$$

Typically the data is in the form of a set of values $\mathbf{x} = \{x_1, \ldots, x_N\}$, and the hypothesis in the form of a set of parameters $\theta = \{\theta_1, \ldots, \theta_K\}$. It is notationally helpful to express distributions of $\mathbf{x}$ and distributions of $\mathbf{x}$ conditioned on θ using the symbol f, and distributions of θ and distributions of θ conditioned on $\mathbf{x}$ using the symbol π, rather than using the symbol P everywhere. We then have

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\,\pi(\theta)}{\int\limits_\Theta d\theta'\ f(\mathbf{x}|\theta')\,\pi(\theta')}\ , \tag{1.5.2}$$

where $\Theta \ni \theta$ is the space of parameters. Note that $\int_\Theta d\theta\,\pi(\theta|\mathbf{x}) = 1$. The denominator of the RHS is simply $f(\mathbf{x})$, which is independent of θ, hence $\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\,\pi(\theta)$. We call $\pi(\theta)$ the *prior* for θ, $f(\mathbf{x}|\theta)$ the *likelihood* of $\mathbf{x}$ given θ, and $\pi(\theta|\mathbf{x})$ the *posterior* for θ given $\mathbf{x}$. The idea here is that while our initial guess at the θ distribution is given by the prior $\pi(\theta)$, after taking data, we should *update* this distribution to the posterior $\pi(\theta|\mathbf{x})$. The likelihood $f(\mathbf{x}|\theta)$ is entailed by our model for the phenomenon which produces the data. We can use the posterior to find the distribution of new data points $\mathbf{y}$, called the *posterior predictive distribution*,

$$f(\mathbf{y}|\mathbf{x}) = \int\limits_\Theta d\theta\, f(\mathbf{y}|\theta)\,\pi(\theta|\mathbf{x})\ . \tag{1.5.3}$$

This is the update of the *prior predictive distribution*,

$$f(\mathbf{x}) = \int_\Theta d\theta\, f(\mathbf{x}|\theta)\,\pi(\theta)\,. \tag{1.5.4}$$

Example 1.5.1: Coin Flipping

Consider a model of coin flipping based on a standard Bernoulli distribution, where $\theta \in [0,1]$ is the probability for heads ($x=1$) and $1-\theta$ the probability for tails ($x=0$). That is,

$$\begin{aligned} f(x_1,\ldots,x_N|\theta) &= \prod_{j=1}^N \left[(1-\theta)\,\delta_{x_j,0} + \theta\,\delta_{x_j,1}\right] \\ &= \theta^X (1-\theta)^{N-X}\,, \end{aligned}$$

where $X = \sum_{j=1}^N x_j$ is the observed total number of heads, and $N-X$ the corresponding number of tails. We now need a prior $\pi(\theta)$. We choose the Beta distribution,

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathsf{B}(\alpha,\beta)}\,, \tag{1.5.5}$$

where $\mathsf{B}(\alpha,\beta) = \Gamma(\alpha)\,\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the Beta function. One can check that $\pi(\theta)$ is normalized on the unit interval: $\int_0^1 d\theta\,\pi(\theta) = 1$ for all positive α,β. Even if we limit ourselves to this form of the prior, different Bayesians might bring different assumptions about the values of α and β. Note that if we choose $\alpha=\beta=1$, the prior distribution for θ is flat, with $\pi(\theta)=1$.

We now compute the posterior distribution for θ:

$$\pi(\theta|x_1,\ldots,x_N) = \frac{f(x_1,\ldots,x_N|\theta)\,\pi(\theta)}{\int_0^1 d\theta'\, f(x_1,\ldots,x_N|\theta')\,\pi(\theta')} = \frac{\theta^{X+\alpha-1}(1-\theta)^{N-X+\beta-1}}{\mathsf{B}(X+\alpha,N-X+\beta)}\,. \tag{1.5.6}$$

Thus, we retain the form of the Beta distribution, but with updated parameters,

$$\begin{aligned} \alpha' &= X+\alpha \\ \beta' &= N-X+\beta\,. \end{aligned}$$

The fact that the functional form of the prior is retained by the posterior is generally *not* the case in Bayesian updating. We can also compute the prior predictive,

$$\begin{aligned} f(x_1,\ldots,x_N) &= \int_0^1 d\theta\, f(x_1,\ldots,x_N|\theta)\,\pi(\theta) \\ &= \frac{1}{\mathsf{B}(\alpha,\beta)} \int_0^1 d\theta\,\theta^{X+\alpha-1}(1-\theta)^{N-X+\beta-1} = \frac{\mathsf{B}(X+\alpha,N-X+\beta)}{\mathsf{B}(\alpha,\beta)}\,. \end{aligned}$$

The posterior predictive is then

$$\begin{aligned} f(y_1,\ldots,y_M|x_1,\ldots,x_N) &= \int_0^1 d\theta\, f(y_1,\ldots,y_M|\theta)\,\pi(\theta|x_1,\ldots,x_N) \\ &= \frac{1}{\mathsf{B}(X+\alpha,N-X+\beta)} \int_0^1 d\theta\,\theta^{X+Y+\alpha-1}(1-\theta)^{N-X+M-Y+\beta-1} \\ &= \frac{\mathsf{B}(X+Y+\alpha,N-X+M-Y+\beta)}{\mathsf{B}(X+\alpha,N-X+\beta)}\,. \end{aligned}$$

Hyperparameters and conjugate priors

In Example 1.5.1, θ is a *parameter* of the Bernoulli distribution, the likelihood, while quantities α and β are *hyperparameters* which enter the prior $\pi(\theta)$. Accordingly, we could have written $\pi(\theta|\alpha,\beta)$ for the prior. We then have for the posterior

$$\pi(\theta|\mathbf{x},\alpha) = \frac{f(\mathbf{x}|\theta)\,\pi(\theta|\alpha)}{\int\limits_\Theta d\theta'\; f(\mathbf{x}|\theta')\,\pi(\theta'|\alpha)}\;, \tag{1.5.7}$$

replacing Equation [BayesPost], , where $\alpha \in A$ is the vector of hyperparameters. The hyperparameters can also be distributed, according to a *hyperprior* $\rho(\alpha)$, and the hyperpriors can further be parameterized by *hyperhyperparameters*, which can have their own distributions, *ad nauseum*.

What use is all this? We've already seen a compelling example: when the posterior is of the same form as the prior, the Bayesian update can be viewed as an automorphism of the hyperparameter space A, one set of hyperparameters α is mapped to a new set of hyperparameters $\widetilde{\alpha}$.

> ***Definition:*** *A parametric family of distributions* $\mathcal{P} = \big\{\pi(\theta|\alpha)\,|\,\theta\in\Theta,\,\alpha\in A\big\}$ *is called a conjugate family for a family of distributions* $\big\{f(\mathbf{x}|\theta)\,|\,\mathbf{x}\in\mathcal{X},\,\theta\in\Theta\big\}$ *if, for all* $\mathbf{x}\in\mathcal{X}$ *and* $\alpha\in A$,
>
> $$\pi(\theta|\mathbf{x},\alpha) \equiv \frac{f(\mathbf{x}|\theta)\,\pi(\theta|\alpha)}{\int\limits_\Theta d\theta'\; f(\mathbf{x}|\theta')\,\pi(\theta'|\alpha)} \in \mathcal{P}\;. \tag{1.5.8}$$
>
> *That is,* $\pi(\theta|\mathbf{x},\alpha) = \pi(\theta|\widetilde{\alpha})$ *for some* $\widetilde{\alpha}\in A$, *with* $\widetilde{\alpha} = \widetilde{\alpha}(\alpha,\mathbf{x})$.

As an example, consider the conjugate Bayesian analysis of the Gaussian distribution. We assume a likelihood

$$f(\mathbf{x}|u,s) = (2\pi s^2)^{-N/2}\exp\left\{-\frac{1}{2s^2}\sum_{j=1}^N (x_j-u)^2\right\}. \tag{1.5.9}$$

The parameters here are $\theta = \{u,s\}$. Now consider the prior distribution

$$\pi(u,s|\mu_0,\sigma_0) = (2\pi\sigma_0^2)^{-1/2}\,\exp\left\{-\frac{(u-\mu_0)^2}{2\sigma_0^2}\right\}. \tag{1.5.10}$$

Note that the prior distribution is independent of the parameter s and only depends on u and the hyperparameters $\alpha = (\mu_0,\sigma_0)$. We now compute the posterior:

$$\begin{aligned}\pi(u,s|\mathbf{x},\mu_0,\sigma_0) &\propto f(\mathbf{x}|u,s)\,\pi(u,s|\mu_0,\sigma_0)\\ &= \exp\left\{-\left(\frac{1}{2\sigma_0^2}+\frac{N}{2s^2}\right)u^2 + \left(\frac{\mu_0}{\sigma_0^2}+\frac{N\langle x\rangle}{s^2}\right)u - \left(\frac{\mu_0^2}{2\sigma_0^2}+\frac{N\langle x^2\rangle}{2s^2}\right)\right\},\end{aligned}$$

with $\langle x\rangle = \frac{1}{N}\sum_{j=1}^N x_j$ and $\langle x^2\rangle = \frac{1}{N}\sum_{j=1}^N x_j^2$. This is also a Gaussian distribution for u, and after supplying the appropriate normalization one finds

$$\pi(u,s|\mathbf{x},\mu_0,\sigma_0) = (2\pi\sigma_1^2)^{-1/2}\exp\left\{-\frac{(u-\mu_1)^2}{2\sigma_1^2}\right\}\quad, \tag{1.5.11}$$

with

$$\begin{aligned}\mu_1 &= \mu_0 + \frac{N\big(\langle x\rangle-\mu_0\big)\sigma_0^2}{s^2+N\sigma_0^2}\\ \sigma_1^2 &= \frac{s^2\sigma_0^2}{s^2+N\sigma_0^2}\quad.\end{aligned}$$

Thus, the posterior is among the same family as the prior, and we have derived the update rule for the hyperparameters $(\mu_0, \sigma_0) \to (\mu_1, \sigma_1)$. Note that $\sigma_1 < \sigma_0$, so the updated Gaussian prior is sharper than the original. The updated mean μ_1 shifts in the direction of $\langle x \rangle$ obtained from the data set.

The problem with priors

We might think that the for the coin flipping problem, the flat prior $\pi(\theta) = 1$ is an appropriate initial one, since it does not privilege any value of θ. This prior therefore seems 'objective' or 'unbiased', also called 'uninformative'. But suppose we make a change of variables, mapping the interval $\theta \in [0, 1]$ to the entire real line according to $\zeta = \ln\big[\theta/(1-\theta)\big]$. In terms of the new parameter ζ, we write the prior as $\tilde\pi(\zeta)$. Clearly $\pi(\theta)\,d\theta = \tilde\pi(\zeta)\,d\zeta$, so $\tilde\pi(\zeta) = \pi(\theta)\,d\theta/d\zeta$. For our example, find $\tilde\pi(\zeta) = \frac{1}{4}sech^2(\zeta/2)$, which is not flat. Thus what was uninformative in terms of θ has become very informative in terms of the new parameter ζ. Is there any truly unbiased way of selecting a Bayesian prior?

One approach, advocated by E. T. Jaynes, is to choose the prior distribution $\pi(\theta)$ according to the principle of maximum entropy. For continuous parameter spaces, we must first define a parameter space metric so as to be able to 'count' the number of different parameter states. The entropy of a distribution $\pi(\theta)$ is then dependent on this metric: $S = -\int d\mu(\theta)\,\pi(\theta)\,\ln\pi(\theta)$.

Another approach, due to Jeffreys, is to derive a parameterization-independent prior from the likelihood $f(\mathbf{x}|\theta)$ using the so-called *Fisher information matrix*,

$$\begin{aligned} I_{ij}(\theta) &= -\mathbb{E}_\theta\left(\frac{\partial^2 \ln f(\mathbf{x}|\theta)}{\partial\theta_i\,\partial\theta_j}\right) \\ &= -\int d\mathbf{x}\,f(\mathbf{x}|\theta)\,\frac{\partial^2 \ln f(\mathbf{x}|\theta)}{\partial\theta_i\,\partial\theta_j}\ . \end{aligned}$$

The *Jeffreys prior* \pi\ns_\ssr{J}(\Btheta) is defined as

\pi\ns_\ssr{J}(\Btheta)\propto\sqrt{\det\! I(\Btheta)}\ .

One can check that the Jeffries prior is invariant under reparameterization. As an example, consider the Bernoulli process, for which $\ln f(\mathbf{x}|\theta) = X\ln\theta + (N-X)\ln(1-\theta)$, where $X = \sum_{j=1}^N x_j$. Then

$$-\frac{d^2\ln p(\mathbf{x}|\theta)}{d\theta^2} = \frac{X}{\theta^2} + \frac{N-X}{(1-\theta)^2}\ , \tag{1.5.12}$$

and since $\mathbb{E}_\theta\,X = N\theta$, we have

I(\theta)={N\over\theta(1-\theta)}\qquad\Rightarrow\qquad \pi\ns_\ssr{J}(\theta)={1\over\pi}\,{1\over\sqrt{\theta(1-\theta)}}\ ,

which felicitously corresponds to a Beta distribution with $\alpha = \beta = \frac{1}{2}$. In this example the Jeffries prior turned out to be a conjugate prior, but in general this is not the case.

We can try to implement the Jeffreys procedure for a two-parameter family where each x_j is normally distributed with mean μ and standard deviation σ. Let the parameters be $(\theta_1, \theta_2) = (\mu, \sigma)$. Then

$$-\ln f(\mathbf{x}|\theta) = N\ln\sqrt{2\pi} + N\ln\sigma + \frac{1}{2\sigma^2}\sum_{j=1}^N (x_j-\mu)^2\ , \tag{1.5.13}$$

and the Fisher information matrix is

$$I(\theta) = -\frac{\partial^2 \ln f(\mathbf{x}|\theta)}{\partial\theta_i\,\partial\theta_j} = \begin{pmatrix} N\sigma^{-2} & \sigma^{-3}\sum_j(x_j-\mu) \\ \sigma^{-3}\sum_j(x_j-\mu) & -N\sigma^{-2} + 3\sigma^{-4}\sum_j(x_j-\mu)^2 \end{pmatrix}\ . \tag{1.5.14}$$

Taking the expectation value, we have $\mathbb{E}\,(x_j-\mu) = 0$ and $\mathbb{E}\,(x_j-\mu)^2 = \sigma^2$, hence

$$\mathbb{E}\,I(\theta) = \begin{pmatrix} N\sigma^{-2} & 0 \\ 0 & 2N\sigma^{-2} \end{pmatrix} \tag{1.5.15}$$

and the Jeffries prior is \pi\ns_\ssr{J}(\mu,\sigma)\propto \sigma^{-2}. This is problematic because if we choose a flat metric on the (μ, σ) upper half plane, the Jeffries prior is not normalizable. Note also that the Jeffreys prior no longer resembles a Gaussian, and hence is not a conjugate prior.

---

1. The exception is the unbiased case $p = q = \frac{1}{2}$, where $\langle X \rangle = 0$.↩
2. The origin of $\mathcal{C}$ lies in the $\mathcal{O}(\ln N)$ and $\mathcal{O}(N^0)$ terms in the asymptotic expansion of $\ln N!$. We have ignored these terms here. Accounting for them carefully reproduces the correct value of $\mathcal{C}$ in Equation [normC].↩
3. The function $s(x)$ is the *specific entropy*.↩
4. A function $g(x)$ is *concave* if $-g(x)$ is convex.↩
5. See 'An Introduction to Information Theory and Entropy' by T. Carter, Santa Fe Complex Systems Summer School, June 2011. Available online at astarte.csustan.edu/$\sim$tom/SFI-CSSS/info-theory/info-lec.pdf.↩
6. My colleague John McGreevy refers to $I(p)$ as the *surprise* of observing an event which occurs with probability p. I like this very much.↩
7. See §10 of *An Enquiry Concerning Human Understanding* by David Hume (1748).↩
8. Such a measure is invariant with respect to canonical transformations, which are the broad class of transformations among coordinates and momenta which leave Hamilton's equations of motion invariant, and which preserve phase space volumes under Hamiltonian evolution. For this reason $d\mu$ is called an *invariant phase space measure*.↩
9. Memorize this!↩
10. Jean Baptiste Joseph Fourier (1768-1830) had an illustrious career. The son of a tailor, and orphaned at age eight, Fourier's ignoble status rendered him ineligible to receive a commission in the scientific corps of the French army. A Benedictine minister at the École Royale Militaire of Auxerre remarked, "Fourier, not being noble, could not enter the artillery, although he were a second Newton." Fourier prepared for the priesthood but his affinity for mathematics proved overwhelming, and so he left the abbey and soon thereafter accepted a military lectureship position. Despite his initial support for revolution in France, in 1794 Fourier ran afoul of a rival sect while on a trip to Orleans and was arrested and very nearly guillotined. Fortunately the Reign of Terror ended soon after the death of Robespierre, and Fourier was released. He went on Napoleon Bonaparte's 1798 expedition to Egypt, where he was appointed governor of Lower Egypt. His organizational skills impressed Napoleon, and upon return to France he was appointed to a position of prefect in Grenoble. It was in Grenoble that Fourier performed his landmark studies of heat, and his famous work on partial differential equations and Fourier series. It seems that Fourier's fascination with heat began in Egypt, where he developed an appreciation of desert climate. His fascination developed into an obsession, and he became convinced that heat could promote a healthy body. He would cover himself in blankets, like a mummy, in his heated apartment, even during the middle of summer. On May 4, 1830, Fourier, so arrayed, tripped and fell down a flight of stairs. This aggravated a developing heart condition, which he refused to treat with anything other than more heat. Two weeks later, he died. Fourier's is one of the 72 names of scientists, engineers and other luminaries which are engraved on the Eiffel Tower.↩
11. "A frequentist is a person whose long-run ambition is to be wrong 5% of the time. A Bayesian is one who, vaguely expecting a horse, and catching glimpse of a donkey, strongly believes he has seen a mule." – Charles Annis.↩
