18
Optimal Fixed Sample Size

In this lecture we will face the problem of choosing the sample size of an experiment. From the Bayesian viewpoint this task is quite natural in the way that we just need to obtain the sample size that maximizes the expected utility. We will see this through a couple of examples in this lecture.

Recalling last lectures, $X_1, X_2, \cdots, X_n$ is a sample of size n from $f(x|\theta)$, where θ is the underlying parameters in the model. The prior information about θ will be summarized in $\pi(\theta)$. Suppose also that conditional on the chosen sample size, n, the terminal or final decision, among $\mathbf{a}$ in $\mathcal{A}$, will be that one that minimizes the posterior expected loss, $E(L(\theta, \delta_n^\pi, n))$, with respect to $\mathbf{a}$. Therefore, the combined decision rule is to choose n that minimizes

$$r(\pi, n) = \int_\Theta \int_\mathcal{X} L(\theta, \mathbf{a}, n) f(\mathbf{x}^n|\theta)\pi(\theta) dx d\theta \qquad (18.0.1)$$

where $\mathbf{x}^n = (x_1, \cdots, x_n)$.

Notice that this is not a sequential problem because we do not observe, say x_1 before making the decision about taking or not another observation. In the next lecture we will extend this problem to cases of sequential decision applied to the choice of the optimal sample sizes. Now, let us explore (18.0.1) in a couple of examples.

Example 18.0.1 Suppose that $X_1, X_2, \cdots, X_n$ is sample from a $N(\theta, \sigma^2)$, with σ^2 known and the our terminal decision is to estimate θ, where the loss function is $L(\theta, a, n) = (\theta - a)^2 + C(n)$. The prior information about θ is a $N(\mu, \tau^2)$, with μ and τ^2 known.

We already know that $\bar{X}_n \sim N(\theta, \sigma^2/n)$ and the the posterior distribution for θ is,

$$\pi(\theta|x_1, \cdots, x_n) \sim N(\mu_n, tau_n^2) \quad (18.0.2)$$

where

$$\begin{aligned} \mu_n &= \frac{\sigma^2/n}{\sigma^2/n + \tau^2}\mu + \frac{\tau^2}{\sigma^2/n + \tau^2}\bar{x}_n \\ \tau_n^2 &= \tau^2\sigma^2/(\sigma^2 + n\tau^2) \end{aligned}$$

The Bayes risk is τ_n^2 and the final risk is

$$r(\pi, n) = \tau_n^2 + C(n) \quad (18.0.3)$$

If $C(n) = cn$ for some positive constant c, then it is easy to show that

$$n^* = \sigma/\sqrt{c} - \sigma^2/\tau^2 \quad (18.0.4)$$

is the optimal sample size. Notice that, as the prior variance increases the optimal sample size has to increase as well, which is quite reasonable because we are loosing information from the prior. The reverse is also true, that is the sample size decrease as we increase our prior information. Therefore, this example tells us that the prior information plays an important role in finding out optimal sample sizes.

Example 18.0.2 **Information and Sample Size : The Exponential case** (Berry and Parmigiani [**?**]). A common design problem in clinical and other applications is determining the best fixed sample size for the experiment. To make the notation more standard, let $D = n$. If the data to be collected are exchangeable, then $\mathcal{I}(n)$ is increasing and convex in n. Therefore, if the cost of observation increases linearly or faster in n and if information and cost are additive, there are at most two (adjacent) solutions.

We will consider here the exponential survival data with unknown failure rate Λ and conjugate prior gamma(α, β). Let $t = \sum_i^n x_i$. The expected information in a sample of size n is:

$$\begin{aligned} \mathcal{I}(n) &= \int\int \log\left[\frac{\Gamma(\alpha)(\beta+t)^{\alpha+n}}{\Gamma(\alpha+n)\beta^\alpha}\lambda^n e^{-\lambda t}\right]\frac{\beta^\alpha}{\Gamma(\alpha)\Gamma(n)}\lambda^{\alpha+n-1}t^{n+1}e^{-\lambda(b+t)}dtd\lambda \\ &= \log\frac{\Gamma(a)}{\Gamma(a+n)} + (a+n)\psi(a+n) - a\psi(a) - n \quad (18.0.5) \end{aligned}$$

where ψ is the digamma function and the integral are from zero to infinity.

The value of $\mathcal{I}(n)$ does not depend on β. The amount one expects to learn depends solely on the number of prior failures and not on the prior total survival.

A useful corollary of (18.0.5) is the expected information from the next observation, which is

$$\mathcal{I}(1) = \frac{1}{\alpha} + \psi(\alpha) - \log\alpha.$$

If $\alpha = 1$, then $\mathcal{I}(1) = 1 - C = .4288$, where C is Euler's constant.

The following large-n approximation brings out further interesting features of $\mathcal{I}(n)$. Using

$$\log\Gamma(n) \approx \frac{1}{2}\log 2\pi - n + (n - \frac{1}{2})\log n$$

and

$$\psi(n) \approx \log n - \frac{1}{2n},$$

we get:

$$\mathcal{I}(n) \approx K(\alpha) + \frac{1}{2}\log(\alpha + n)$$

where $K(\alpha) = \log\Gamma(\alpha) - \alpha(\psi(\alpha) - 1) - \frac{1}{2}\log 2\pi - \frac{1}{2}$. Using this approximation one can solve analytically for the optimal fixed sample size with sampling cost c (in information units) per observation. In particular, taking n to be continuous, the optimum is:

$$n^* = \frac{1}{2c} - a.$$