19

Optimal Fixed Sample Size

In this lecture we will consider the problem of choosing the sample size of an experiment. From the Bayesian viewpoint the solution is quite natural because 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. Bayesian approaches are becoming increasingly common for sample size determination. For a review and a typical application see (Adc97; YBL96).

Recalling past lectures, $X^n = (X_1, X_2, \ldots, X_n)$ is a sample of size n from $p(x|\theta)$, where θ are 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 $\boldsymbol{a}$ in $\mathcal{A}$, will be that one that maximizes the posterior expected utility, $E(u(\theta, \boldsymbol{a}, n))$, with respect to $\boldsymbol{a}$. Let δ_n^π be the Bayes rule based on n observations, so

$$\delta_n^\pi = \arg\max_{\boldsymbol{a} \in \mathcal{A}} \int_\Theta u(\theta, \boldsymbol{a}, n)\pi(\theta|x^n)d\theta \tag{19.1}$$

Therefore, the combined decision rule is to choose n that maximizes

$$U(\pi, n) = \int_\mathcal{X} \int_\Theta u(\theta, \delta_n^\pi, n)\pi(\theta|x^n, \theta)p(x^n)d\theta dx^n \tag{19.2}$$

where $p(x^n) = p(x^n|\theta)\pi(\theta)/\pi(\theta|x^n)$ and $x^n = (x_1, \ldots, x_n)$. In this notation, U is the negative of the Bayes risk.

We will explore (19.2) in a couple of examples, but before let us explore some properties of this method when the utility function is of the form,

$$u(\theta, \boldsymbol{a}, n) = u(\theta, \boldsymbol{a}) - C(n) \tag{19.3}$$

where $C(n)$ is a cost function. In this case, δ_n^π does not depend on $C(n)$ if $C(n)$ is independent of θ. Therefore, the expression (19.2) can be written as,

$$U(\pi, n) = \int_\mathcal{X} \int_\Theta u(\theta, \delta_n^\pi)\pi(\theta|x^n, \theta)p(x^n)d\theta dx^n - C(n) \tag{19.4}$$

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.

19.1 Normal IID observations

Suppose that $X_1, X_2, \ldots$ are conditionally independent $N(\theta, \sigma^2)$, with σ^2 known, and the terminal decision is to estimate θ. The utility function is $u(\theta, a, n) = -(\theta - a)^2 - C(n)$. 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,

$$\pi(\theta|x_1, \cdots, x_n) \sim N(\mu_n, \tau_n^2) \tag{19.5}$$

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 posterior expected utility, not including the cost for observations, is τ_n^2. Because this is independent of $\bar{x}_n$, we have

$$U(\pi, n) = -\tau_n^2 - C(n). \tag{19.6}$$

Suppose that $C(n) = cn$ for some positive constant c, or, in words, that each additional observation costs a fixed amount. Approximating n with a continuous variable and taking derivatives, you can convince yourself that

$$n^* = \sigma/\sqrt{c} - \sigma^2/\tau^2 \tag{19.7}$$

is the approximately optimal sample size. Looking at the two integer values next to n^* and choosing the one with the largest posterior expected utility is the optimal solution, because this function is convex.

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 optimal sample sizes. Another fact, is that n^* is not a monotone function of σ.

Definition: η is a nuisance parameter if $u(\alpha, \boldsymbol{a})$ is independent of η. We partition $\alpha = (\eta, \theta)$ where η includes all nuisance parameters.

Theorem 23 *If η is a nuisance parameters, $a^*(\eta)$ is the Bayes solution assuming η is known, $\pi(\eta)$ is the prior distribution of η, and, finally, conditions are such that differentiation and integration can be interchanged, then the optimum a^* is*

$$a^* = \int_{\Xi} a^*(\eta)\pi(\eta)d\eta$$

The proof is left as an exercise (see problem 4 at the end).

19.2 Information and Sample Size : The Exponential case

This example is taken from PB94. A common design problem in clinical and other applications is determining the best fixed sample size for the experiment. If the data to be collected are exchangeable, then $\mathcal{I}(n)$ is increasing and convex in n (see the notation of Lectures 16 and 17). 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:

$$\begin{aligned} U(\pi, n) &= \int\int \log\left[\frac{\Gamma(\alpha)(\beta+t)^{\alpha+n}}{\Gamma(\alpha+n)\beta^\alpha}\lambda^n e^{-\lambda t}\right]\frac{\beta^\alpha}{\Gamma(\alpha)\Gamma(n)}\lambda^{\alpha+n-1}t^{n+1}e^{-\lambda(b+t)}dtd\lambda \\ &= \log\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} + (\alpha+n)\psi(\alpha+n) - \alpha\psi(\alpha) - n \qquad (19.8) \end{aligned}$$

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

The value of $U(\pi, n)$ 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 (19.8) is the expected information from the next observation, which is

$$U(\pi, 1) = \frac{1}{\alpha} + \psi(\alpha) - \log\alpha.$$

If $\alpha = 1$, then $U(\pi, 1) = 1 - C = .4288$, where C is Euler's constant.

The following large-n approximation brings out further interesting features of $U(\pi, n)$. 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:

$$U(\pi, n) \approx K(\alpha) + \frac{1}{2}\log(\alpha + n)$$

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:

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

19.3 Optimal design by curve fitting

This example is taken from MP95.

Consider the problem of choosing the optimal sample size n for a binomial experiment. The likelihood function is given by $p(y|\theta) = C_y^n \theta^y (1-\theta)^{n-y}$, where y is the number of successes. We take a prior distribution on θ given by

$$\pi(\theta) = 1.5\theta^2 + 15\theta^2(1-\theta)^2$$

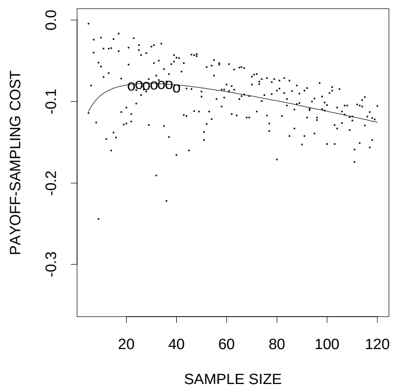
perhaps representing the mixture of two experts' opinions.

As terminal decision problem we take the estimation of θ under absolute error loss; therefore, as it is well known, $u(\theta, \delta(y)) = -|m_y - \theta|$, where $\delta(y)$ is the Bayes rule —in this case the posterior median m_y (problem 5). Let the sampling cost be 0.0008 per observation, so that $C(n) = 0.0008n$. The objective function is then (equation 19.2)

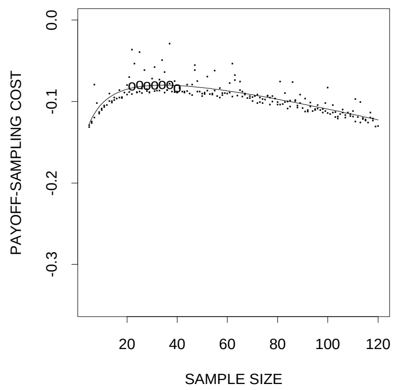
$$U(\pi, n) = \int_0^1 \sum_{y=0}^{n} [-|m_y - \theta| - .0008n] C_y^n \theta^y (1-\theta)^{n-y} \pi(\theta) d\theta \qquad (19.9)$$

There is no analytical solution to this decision problem and we explore the alternative of casting it as a standard statistical problem of fitting a curve for $U(n)$ based on only one Monte Carlo sample which is spread over the whole design space, $\mathcal{N}$, which will be restricted to all integers bounded by 120. Maximizing the fitted curve will then solve the optimization in a deterministic way. The following is the basic implementation strategy:

1. Select $n_i \in \mathcal{N}$, $i = 1, ..., I$;
2. Draw θ_i, y_i from the prior and the model respectively, for $i = 1, \ldots, I$.
3. Compute $u_i = u(\theta_i, \delta(y_i)) - C(n_i)$;
4. Fit a smooth curve $\tilde{U}(n)$ to the resulting points (n_i, u_i); It should be clear that $\tilde{U}$ is also a function of the prior π.
5. Evaluate deterministically the maximum of $\tilde{U}(n)$.



(a) (n_i, u_i)



(b) $(n_i, \bar{u}_i)$

Figure 42 Estimation of the expected utility curve. The points in Panel (a) have coordinates (n_i, u_i). The line shows $\tilde{U}(n)$. The points in Panel (b) plot the pairs $(n_i, \bar{u}_i)$. The line shows $\bar{U}(n)$. In both plots, the small circles represent "exact" evaluations of $U(\pi, n)$ by large scale Monte Carlo at a few selected designs, i.e. $\hat{U}(n) = 1/M \sum_1^M u(\theta_i, \delta(y_i))$ where (θ_i, y_i) are sampled from $\pi(\theta)$ and $p(y|\theta)$, respectively. We used a Monte Carlo sample size of $I = 400$.

The second step is usually straightforward and is best done by generating θ from the prior and then y from the likelihood. Evaluation of the maximum can be carried out with standard global optimization methods. Efficient implementation of the remaining two steps needs further discussion, which is out of our scope. This argument can be applied when U is also a function of n and when C is a function of n, y and θ.

The simulation points n_i are chosen uniformly in the range $(0, 120)$. The Monte Carlo sample size is $I = 200$. Figure **??** shows the simulated pairs (n_i, u_i). This problem is not particularly favorable to curve fitting optimization, as the low cost of observation results in a rather flat response. Also the absolute error loss makes the surface relatively noisy. Squared error loss and a higher observation cost would lead to a much more favorable situation.

Yet, even in such a small Monte-Carlo sample there is enough information to make a satisfactory design choice. We fitted the data by a nonlinear regression of the form:

$$\tilde{U}(n) = -.0008n - .1929(1 + bn)^{-a}, \tag{19.10}$$

where a and b are to be estimated based on the Monte Carlo sample. This curve incorporates information regarding (i) the sampling cost; (ii) the value of the expected payoff when no observations are taken —which can easily be computed and is .1929; and (iii) the rate of convergence of the expected absolute error to zero —assumed here to be polynomial. The results are illustrated in Figure 42a.

The functional form chosen for fitting is very simple, and more flexible model specifications may be preferable in other applications. One advantage of this form is that it lends itself very nicely to solving the optimization problem analytically. In particular, the approximate optimal sample size $\hat{n}^*$ is given by

$$\hat{n}^* = \frac{1}{\hat{b}}\left[\left(\frac{.1929\hat{a}\hat{b}}{.0008}\right)^{\frac{1}{\hat{a}+1}} - 1\right]. \tag{19.11}$$

The sample shown gives $\hat{n}^* = 29$.

The next developments shows that, if nothing else, spending some time trying to solve analytically as many problems as possible still a major advantage in science, despite the rapid increase in computational power. To see that, we rewrite the objective function as

$$U(\pi, n) = \sum_{y=0}^{n} \int_0^1 [-|m_y - \theta| - .0008n]\pi_y(\theta) d\theta \, p(y) \tag{19.12}$$

where m is the marginal distribution of the data and π_y is the posterior distribution of θ —again a mixture of Beta densities. The inner integral can be evaluated directly, as long as computation of the incomplete beta function is available. Let the evaluation of the inner integral corresponding to n_i, y_i be $\bar{u}_i$. In alternative to the procedure of Figure 42a, we can in this case interpolate directly the points $(n_i, \bar{u}_i)$, with substantial reduction of the variability around the fitted curve. This is clearly recommendable whenever possible. The resulting estimated curve $\bar{\mathcal{U}}(n)$ is shown in Figure 42b. The new estimate is $\bar{n}^* = 34$.

19.4 Hypothesis testing

(Wei97) Assume that $X_1, X_2, \cdots$ is a sequential sample from a $N(\theta, 1)$ density, and that it is desired to test $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$. The decision loss is zero for a correct decision and $|\theta|$ for an incorrect decision. The cost of a fixed sample of size n is $C(n) = 0.01n$. If the prior density of θ is $N(0, 4)$, approximate the optimal fixed sample size.

Table 14 shows the loss function associated to each decision.

Table 14 Utility function for the fixed sample size problem

	True value of θ	
$H_0 : \theta \leq 0$	0	$-\theta$
$H_1 : \theta > 0$	θ	0

From table (14) it can be shown that

$$u(\theta, H_0) = \begin{cases} 0 & \theta \leq 0 \\ -\theta & \theta > 0 \end{cases} \quad \text{and} \quad u(\theta, H_1) = \begin{cases} \theta & \theta \leq 0 \\ 0 & \theta > 0 \end{cases}$$

Therefore, the posterior expected utilities for H_0 and H_1 are, respectively,

$$\begin{aligned} U(\pi_x, H_0) &= -\int_0^{\infty} \theta\pi(\theta|\boldsymbol{x})d\theta \\ U(\pi_x, H_1) &= \int_{-\infty}^{0} \theta\pi(\theta|\boldsymbol{x})d\theta \end{aligned}$$

so that the Bayes rule will be H_0 if

$$\int_0^{\infty} \theta\pi(\theta|\boldsymbol{x})d\theta < -\int_{-\infty}^{0} \theta\pi(\theta|\boldsymbol{x})d\theta \tag{19.13}$$

and will be H_1, otherwise. Note that the last inequality can be rewritten in terms of the posterior mean, that is, we decide in favor of H_0 if

$$\int_0^{\infty} \theta\pi(\theta|\boldsymbol{x})d\theta + \int_{-\infty}^{0} \theta\pi(\theta|\boldsymbol{x})d\theta \equiv E(\theta|\boldsymbol{x}) < 0 \tag{19.14}$$

It can be shown that $E(\theta|\boldsymbol{x}) = 4n\bar{x}/(4n+1)$, so we decide in favor of $H_0(H_1)$ if $\bar{x} < (>)0$, or alternatively,

$$\delta_n^{\pi}(x^n) = \begin{cases} H_0 & \bar{x} \leq 0 \\ H_1 & \bar{x} > 0 \end{cases}$$

Basically, this defines the terminal decision (or final decision). We still need to address the first decision concerning the sample size, and we will, as usual, choose the one that maximizes the posterior expected utility function. The decision risk of the Bayes decision rule is,

$$v(\theta, \delta^{\pi}) = \begin{cases} -\theta P(\bar{X} < 0|\theta) & \theta > 0 \\ \theta P(\bar{x} \geq 0|\theta) & \theta < 0 \end{cases}$$

and the expected utility will be

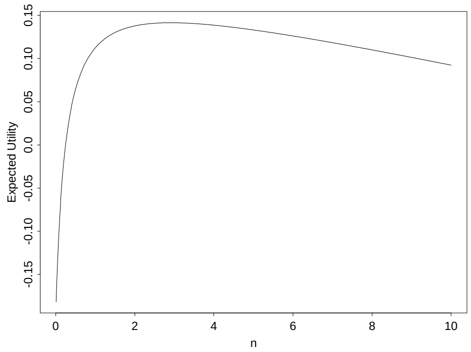


Figure 43 Bayes Risk (hypothesis testing problem)

$$\begin{aligned}
U(\pi, n) &= -\int_0^\infty \theta P(\bar{X} < 0|\theta)\pi(\theta)d\theta + \int_{-\infty}^0 \theta P(\bar{X} \geq 0|\theta)\pi(\theta)d\theta - C(n) \\
&= -\int_{-\infty}^\infty \theta P(\bar{X} < 0|\theta)\pi(\theta)d\theta + \int_{-\infty}^0 \theta\pi(\theta)d\theta - C(n) \\
&= \int_{-\infty}^0 \int_{-\infty}^\infty \theta p(\bar{x}|\theta)\pi(\theta)d\theta d\bar{x} - C(n) + K \\
&= \int_{-\infty}^0 p(\bar{x}) \underbrace{\int_{-\infty}^\infty \theta \frac{p(\bar{x}|\theta)\pi(\theta)}{p(\bar{x})} d\theta}_{E(\theta|\bar{x})} d\bar{x} - C(n) + K \\
&= \frac{4n}{4n+1} \int_{-\infty}^0 \bar{x} p(\bar{x}) d\bar{x} - C(n) + K
\end{aligned}$$

where K is the prior probability that θ is negative. The two integrals in the last expression are the expected values of normal distributions truncated at zero (O'H94, page 73 has some discussion about truncated normals and their expected values). Using such results and the fact that $C(n) = 0.01n$, last expression can be rewritten as (show that this is true):

$$U(\pi, n) = 1 - \frac{1}{\sqrt{2\pi}} \left[\frac{8n+3}{4n+1}\right] - 0.01n \qquad (19.15)$$

Figure 43 shows $U(\pi, n)$ as a function of n (real n's). For n integer, it is easy to check that the optimal sample size is three.

19.5 Exercises

1. Suppose that $x_1, x_2, \ldots$ are conditionally independent Bernoulli experiments, all with success probability θ. You interested in estimating θ and your terminal decision utility function is

$$u(\theta, \boldsymbol{a}) = -\frac{(\theta - \boldsymbol{a})^2}{\theta(1-\theta)}$$

This is the negative of the traditional squared error loss, standardized using the variance of a single observation. With this utility function, errors of the same absolute magnitude are considered worse if the are made near the extremes of the interval $(0, 1)$. You have a Beta(1.5,1.5) prior on θ. Each observation costs you c. What is the optimal fixed sample size? You may start using computers at any stage in the solution of this problem.

2. You are the mayor of Solvantville and you interested in estimating the difference in the concentration of a certain contaminant in the water supply of homes in two different sections of town, each with its own well. If you found that a difference you would further investigate whether there may be contamination due to industrial waste. How big a sample of homes do you need for your purposes? Specify a terminal loss function, a cost of observation, a likelihood and a prior. Keep it very simple, but justify your assumptions and point out limitations if necessary.

3. Can you find the optimal sample size in example 19.1 when σ is not known? You can use a conjugate prior, and standard results such as those of section 3.3 in "Bayesian Data Analysis" by Gelman et al.

4. Proof Theorem (23).

5. Show that the posterior median is the Bayes rule when the utility function is of the form $u(\theta, a) = -|\theta - a|$.

6. In the context of Section 19.4 determine the optimal n by maximizing

$$U(\pi, n) = \int U(\pi_x, \delta_n^\pi(x^n)) p(x^n) dx^n$$

19.6 Applications of Optimal Stopping

In this lecture we will present some examples of optimal stopping. One of the examples is applied to the problem of estimation, and the other one is applied to hypotheses testing (see also, DeG70 and Ber85).

Example: [From (Ber85)] Assume that $X_1, X_2, \ldots$ is a sequential sample from a Bernoulli with parameter θ, and that it is desired to estimate θ under utility $u(\theta, a, n) = -(\theta - a)^2 - nc$. Assume a priori that θ has uniform distribution on (0,1) and $c = 0.01$. We will find the Bayes 3–truncated procedure $\boldsymbol{d}^3$.

Solution: Let us first state some basic results:

i) The posterior distribution of θ given $\mathbf{x}^n$ is a $Beta(\sum_{i=1}^n x_i+1, n-\sum_{i=1}^n x_i+1)$.

ii) Let $s_n = \sum_{i=1}^n x_i$. The marginal distribution of $\mathbf{x}^n$ is given by:

$$p(\mathbf{x}^n) = \frac{\Gamma(s_n+1)\Gamma(n-s_n+1)}{\Gamma(n+2)}$$

iii) The predictive distribution of X_{n+1} given $\mathbf{x}^n$ is:

$$p^*(x_{n+1}) = \frac{p(\mathbf{x}^{n+1})}{p(\mathbf{x}^n)} = \begin{cases} \frac{n+1-s_n}{n+2} & \text{if } x_{n+1} = 0 \\ \frac{s_n+1}{n+2} & \text{if } x_{n+1} = 1 \end{cases}$$

Now, since the utility is $u(\theta, a) = -(\theta - a)^2$, the optimal action a^* is the posterior mean, and the posterior expected utility is minus the posterior variance. Therefore,

$$U_0(\pi^n, n) = \sup_a \left[\int_\Theta \left(u(\theta, a) - c(n)\right) \pi^n(\theta) d\theta \right] = -Var_{\pi^n}(\theta) - 0.01n$$

Using this result:

$$\begin{aligned} U_0(\pi^0, 0) &= -Var_\pi(\theta) = -1/12 \approx -0.0833 \\ U_0(\pi^1, 1) &= -Var_{\pi^1}(\theta) - 0.01 = -2/36 - 0.01 \approx -0.0655 \\ U_0(\pi^2, 2) &= \begin{cases} -0.0700, & \text{if } s_2 = 1 \\ -0.0575, & \text{if } s_2 = 0 \text{ or } s_2 = 2 \end{cases} \\ U_0(\pi^3, 3) &= \begin{cases} -0.0700, & \text{if } s_3 = 1 \text{ or } s_3 = 2 \\ -0.0567, & \text{if } s_3 = 0 \text{ or } s_3 = 3 \end{cases} \end{aligned}$$

Observe that:

$$\begin{aligned}
U_j(\pi^n, n) &= \max\left\{U_0(\pi^n, n), E^{p^*}(U_{j-1}(\pi(\theta|\mathbf{x}^n, X_{n+1}), n+1))\right\} \\
&= \max\left\{-Var^{\pi^n}(\theta) - nc, E^{p^*}(U_{j-1}(\pi(\theta|\mathbf{x}^n, X_{n+1}), n+1))\right\} \\
&= \max\left\{-Var^{\pi^n}(\theta) - nc, -E^{p^*}Var(\theta|\mathbf{x}^{n+j-1}, X_{n+1}) - nc - c\right\} \\
&= \max\left\{-Var^{\pi^n}(\theta), -E^{p^*}Var(\theta|\mathbf{x}^{n+j-1}, X_{n+1}) - c\right\} - nc
\end{aligned}$$

Since:

$$\begin{aligned}
E^{p^*}[Var(\theta|X)] &= P(X=0)Var(\theta|X=0) + P(X=1)Var(\theta|X=1) = 1/18 \\
E^{p^*}[Var(\theta|\mathbf{x}^1, X)] &= 1/24 \\
E^{p^*}[Var(\theta|\mathbf{x}^2, X)] &= \begin{cases} 18/600, \text{ if } s_2 = 0 \text{ or } s_2 = 2 \\ 6/150, \text{if } s_2 = 1 \end{cases}
\end{aligned}$$

we have:

$$\begin{aligned}
U_1(\pi^0, 0) &= \max\{-1/12, -(1/18 + 0.01)\} \approx -0.0655 \\
U_2(\pi^0, 0) &= \max\{-1/12, -(1/24 + 0.01)\} \approx -0.0517.
\end{aligned}$$

Following the same idea:

$$\begin{aligned}
U_3(\pi^0, 0) &= \begin{cases} -0.04, \text{ if } s_2 = 0 \text{ or } s_2 = 2 \\ -0.02, \text{ if } s_2 = 1 \end{cases} \\
U_1(\pi^1, 1) &= -0.0617 \\
U_2(\pi^1, 1) &= \begin{cases} -0.05, \text{ if } s_2 = 0 \text{ or } s_2 = 2 \\ -0.06, \text{ if } s_2 = 1 \end{cases} \\
U_1(\pi^2, 2) &= \begin{cases} -0.07, \text{ if } s_2 = 1 \\ -0.0575, \text{ if } s_2 = 0 \text{ or } s_2 = 2. \end{cases}
\end{aligned}$$

Let us now study the optimal 3–truncated Bayes procedure. Since at stage 0, $U_0(\pi^0, 0) = -0.0833 < U_3(\pi^0, 0)$, X_1 should be observed. After observing X_1, as $U_0(\pi^1, 1) < U_2(\pi^1, 1)$, we should observe X_2. After observing it, since we have $U_0(\pi^2, 2) = U_1(\pi^2, 2)$, it is optimal to stop.

The Bayes action is:

$$\boldsymbol{d}^{\pi^2}(\mathbf{x}^2) = \begin{cases} 1/4 \text{ if } s_2 = 0 \\ 3/4 \text{ if } s_2 = 2 \\ 1/2 \text{ if } s_2 = 1 \end{cases}$$

⋆

Example: [From Ber85 and DeG70] Assume that $X_1, X_2, \ldots$ is a sequential sample from a Bernoulli distribution with (unknown) probability of success θ. It is desired to test $H_0 : \theta = 1/3$ versus $H_1 : \theta = 2/3$. Let a_i denote accepting $H_i (i = 0, 1)$, and suppose that $u(\theta, a, n) = u(\theta, a) - nc$ with $c = 1$ and $u(\theta, a)$ being "0-20" utility, that is, no decision loss for a correct decision, and a decision loss of 20 for an incorrect decision (see table 19.6). Finally, let π_0 denote the prior probability that H_0 is true, that is, $\pi(\theta = 1/3) = \pi_0$ (and $\pi_1 = 1 - \pi_0$ the prior probability that H_1 is true).

Imagine that it is desired to find $\boldsymbol{d}^2$, the optimal procedure among all those taking at most two observations.

	True value of θ	
	$\theta = 1/3$	$\theta = 2/3$
a_0	0	-20
a_1	-20	0

Table 15 Utility function for the testing hypotheses problem

Let us first state some results. Let $s_n = \sum_{i=1}^n x_i$.

i) The marginal distribution of $\mathbf{X}^n$ is given by:

$$\begin{aligned} p(\mathbf{x}^n) &= p(\mathbf{x}|\theta = 1/3)\pi_0 + p(\mathbf{x}|\theta = 2/3)(1 - \pi_0) \\ &= (1/3)^{s_n}(2/3)^{n-s_n}\pi_0 + (2/3)^{s_n}(1/3)^{n-s_n}(1 - \pi_0) \\ &= (1/3)^n \left[2^{n-s_n}\pi_0 + 2^{s_n}(1 - \pi_0)\right] \end{aligned}$$

ii) and the posterior distribution of θ given $\mathbf{x}$ is such that:

$$\begin{aligned} \pi(\theta = 1/3|\mathbf{x}) &= \frac{\pi_0}{\pi_0 + 2^{2s_n - n}(1 - \pi_0)} \\ \pi(\theta = 2/3|\mathbf{x}) &= 1 - \pi(\theta = 1/3|\mathbf{x}) \end{aligned}$$

iii) The predictive distribution of X_{n+1} given $\mathbf{x}^n$ is:

$$\begin{aligned} p^*(x_{n+1}) &= \frac{p(\mathbf{x}^{n+1})}{p(\mathbf{x}^n)} \\ &= \begin{cases} \frac{2^{n+1-s_n}\pi_0 + 2^{s_n}(1-\pi_0)}{2^{n-s_n}\pi_0 + 2^{s_n}(1-\pi_0)} & \text{if } x_{n+1} = 0 \\ \frac{2^{n+2-s_n}\pi_0 + 2^{s_n+1}(1-\pi_0)}{2^{n-s_n}\pi_0 + 2^{s_n}(1-\pi_0)} & \text{if } x_{n+1} = 1 \end{cases} \end{aligned}$$

Before taking observations, the expected utility of immediately making a decision is given by:

$$\begin{aligned}
U_0(\pi^0, 0) = \rho_0(\pi) &= \max_a \left[\sum_\theta (u(\theta, a))\pi(\theta) \right] \\
&= \max\{E^\pi[u(\theta, a_0)], E^\pi[u(\theta, a_1)]\} \\
&= \max\{-20(1-\pi_0), -20\pi_0\} \\
&= \begin{cases} -20\pi_0 & \text{if } 0 \le \pi_0 \le 1/2 \\ -20(1-\pi_0) & \text{if } 1/2 \le \pi_0 \le 1. \end{cases}
\end{aligned}$$

Therefore, the Bayes decision is a_0 if $1/2 \le \pi \le 1$ and a_1 if $0 \le \pi \le 1/2$.

Suppose that x_1 has been observed. If $x_1 = 0$, the posterior expected utility of making a decision without taking another observation is:

$$\rho_0(\pi^1(\theta|0)) = \begin{cases} \frac{-40\pi_0}{1+\pi_0} & \text{if } 0 \le \pi_0 \le 1/3 \\ \frac{-20(1-\pi_0)}{1+\pi_0} & \text{if } 1/3 \le \pi_0 \le 1. \end{cases}$$

If $x_1 = 1$ the posterior expected utility is:

$$\rho_0(\pi^1(\theta|1)) = \begin{cases} \frac{-20\pi_0}{2-\pi_0} & \text{if } 0 \le \pi_0 \le 2/3 \\ \frac{-40(1-\pi_0)}{2-\pi_0} & \text{if } 2/3 \le \pi_0 \le 1 \end{cases}$$

Thus, the expected posterior utility is

$$\begin{aligned}
E^{p^*}[\rho_0(\pi(\theta|X))] &= \rho_0(\pi(\theta|0))p^*(0) + \rho_0(\pi(\theta|1))p^*(1) \\
&= \begin{cases} -20\pi_0 & \text{if } 0 \le \pi_0 \le 1/3 \\ -20/3 & \text{if } 1/3 \le \pi_0 \le 2/3 \\ -20(1-\pi_0) & \text{if } 2/3 \le \pi_0 \le 1 \end{cases}
\end{aligned}$$

The utility from the optimal procedure in which no more than one observation is taken is:

$$\begin{aligned}
\rho_1(\pi) &= \max\left\{\rho_0(\pi), E^{p^*}[\rho_0(\pi(\theta|X))] - 1\right\} \\
&= \begin{cases} -20\pi_0 & \text{if } 0 \le \pi_0 \le 23/60 \\ -23/3 & \text{if } 23/60 \le \pi_0 \le 37/60 \\ -20(1-\pi_0) & \text{if } 37/60 \le \pi_0 \le 1 \end{cases}
\end{aligned}$$

The calculation of ρ_2 follows the same idea. It can be shown that

$$E^{p^*}[\rho_1(\pi(\theta|X))] = \begin{cases} -20\pi_0 & \text{if } 0 \le \pi_0 \le 23/97 \\ -(83\pi_0 + 23)/9 & \text{if } 23/97 \le \pi_0 \le 37/83 \\ -20/3 & \text{if } 37/83 \le \pi_0 \le 46/83 \\ -(106 - 83\pi_0)/9 & \text{if } 46/83 \le \pi_0 \le 60/97 \\ -20(1-\pi_0) & \text{if } 60/97 \le \pi_0 \le 1 \end{cases}$$

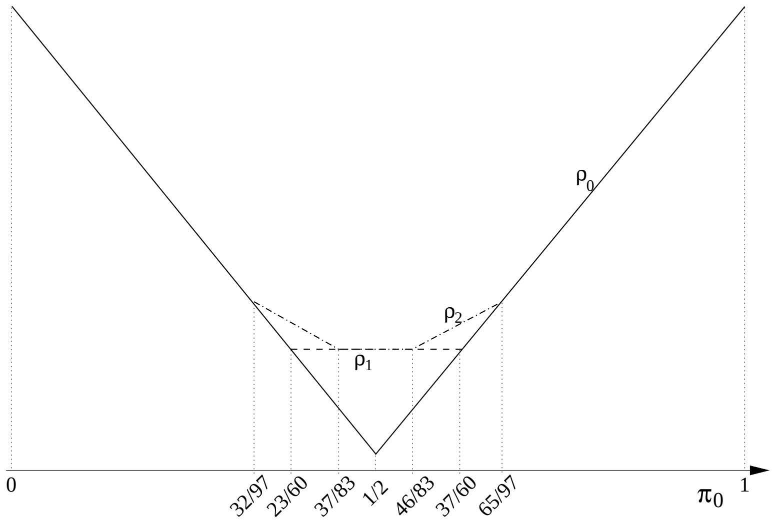


Figure 44 Expected utility functions

which leads to

$$\begin{aligned}\rho_2(\pi) &= \max\left\{\rho_0(\pi), E^{p^*}[\rho_1(\pi(\theta|X))] - 1\right\} \\ &= \begin{cases} -20\pi_0 & \text{if } 0 \leq \pi_0 \leq 32/97 \\ -(83\pi_0 + 32)/9 & \text{if } 32/97 \leq \pi_0 \leq 37/83 \\ -23/3 & \text{if } 37/83 \leq \pi_0 \leq 46/83 \\ -(115 - 83\pi_0)/9 & \text{if } 46/83 \leq \pi_0 \leq 65/97 \\ -20(1-\pi_0) & \text{if } 65/97 \leq \pi_0 \leq 1 \end{cases}\end{aligned}$$

Figure 44 shows the utility functions ρ_0, ρ_1 and ρ_2 as functions of the prior probability π_0.

To simplify the discussion, suppose that $\pi_0 = 2/5 = 0.4$. As $\rho_0(\pi) = -8 < \rho_2(\pi) \approx -7.24$, we should observe X_1. If $x_1 = 0$, we have $\rho_0(\pi^1) = \rho_1(\pi^1) = -60/7$. Therefore, it is optimal to stop and choose a_0. If $x_1 = 1$, $\rho_0(\pi^1) = \rho_1(\pi^1) = -5$ and, again, it is optimal to stop, but we choose a_1. ⋆

Example: [From DeG70] Suppose that $X_1, X_2, \ldots$ is a sequential sample from a normal distribution with unknown mean θ and specified precision ω (that is, with variance $1/\omega$). Assume that θ is estimated under utility $u(\theta, a) = -(\theta - a)^2$ and that there is a fixed cost c per observation. Also, suppose that the prior distribution π of θ is a normal distribution with mean μ and precision τ. Under these conditions, the posterior distribution of θ given

$\mathbf{X}_n = \mathbf{x}_n$ is normal with posterior precision $\tau + n\omega$. Therefore,

$$\begin{aligned} U_0(\pi^0, 0) &= -Var_\pi(\theta) = -\frac{1}{\tau} \\ U_0(\pi^n, n) &= -Var_{\pi^n}(\theta) - nc = -\frac{1}{\tau + n\omega} - nc \end{aligned}$$

and, so $\rho_0(\pi^n) = -\frac{1}{\tau+n\omega}$ for $n = 0, 1, 2, \ldots$ This result shows that the expected utility depends only on the number of observations taken and not on their observed values. This implies that the optimal sequential decision procedure is a procedure in which a fixed number of observations is taken. For a formalization of this result, see Exercise 2.

Exploring a bit more this example, we can observe that

$$\begin{aligned} \rho_1(\pi^n) &= \max\{\rho(\pi^n), E^{p^*}[\rho(\pi^n(\theta|X_{n+1}))] - c\} \\ &= \max\left\{-\frac{1}{\tau + n\omega}, -\frac{1}{\tau + (n+1)\omega} - c\right\} \\ &= -\frac{1}{\tau + n\omega} \\ &\quad \text{if } \frac{1}{\tau + n\omega} < \frac{1}{\tau + (n+1)\omega} + c \Longleftrightarrow \omega < c(\tau + n\omega)(\tau + (n+1)\omega). \end{aligned}$$

In fact, by using an induction argument (Exercise 3) one can show that $U_j(\pi^n) = -\frac{1}{\tau+n\omega} - nc$ for $\beta_n < \omega < \beta_{n+1}$ where β is defined as: $\beta_0 = 0, \beta_n = c(\tau + (n-1)\omega)(\tau + n\omega)$ for $n = 1, \ldots, j$ and $\beta_{j+1} = \infty$. This implies that the optimal sequential decision is to take exactly n^* observations, where n^* is an integer such that $\beta_{n^*} < \omega < \beta_{n^*+1}$. If the decision maker is not allowed to take more than m observations, then if $n^* > m$, he should take m observations. $\star$

19.7 Exercises

1. Sequential problems get very hard very quickly. This is one of the simplest you can try, but it will take some time.

Try to address the same question as worked exercise 19.6, except with Poisson observation and a Gamma$(1, 1)$ prior. Set $c = 1/12$.

2. [From DeG70] Let π be the prior distribution of θ in a sequential decision problem. (a) Prove by using an induction argument that, in a sequential decision problem with fixed cost per observation, if $\rho_0(\pi^n)$ is constant for all observed values $\mathbf{x}^n$, then every function ρ_j for $j = 1, 2, \ldots$ as well as ρ_0

has this property. (b) Conclude that the optimal sequential decision procedure is a procedure in which a fixed number of observations is taken.

3. [From DeG70] (a) In the context of Example 19.6, for $n = 0, 1, \ldots, j$, prove that $U_j(\pi^n) = -\frac{1}{\tau+n\omega} - nc$ for $\beta_n < \omega < \beta_{n+1}$ where β is defined as: $\beta_0 = 0, \beta_n = c(\tau + (n-1)\omega)(\tau + n\omega)$ for $n = 1, \ldots, j$ and $\beta_{j+1} = \infty$. (b) Conclude that if the decision maker is allowed to take no more than m observations, then the optimal procedure is to take $n^* \leq m$ observations where n^* is such that $\beta_{n^*} < \omega < \beta_{n^*+1}$.