APPENDIX B

# Decision theory

## B.1 Introduction

As discussed in Chapters 1, 3, and 4, the Bayesian approach with low-information priors strikes an effective balance between frequentist robustness and Bayesian efficiency. It occupies a middle ground between the two and, when properly tuned, can generate procedures with excellent frequentist properties, often "beating the frequentist approach at its own game." As an added benefit, the Bayesian approach more easily structures complicated models and inferential goals.

To make comparisons among approaches, we need a method of keeping score, commonly referred to as a loss *structure.* Loss structures may be explicit or implicit; here we discuss the structure adopted in a *decision-theoretic* framework. We outline the general decision framework, highlighting those features that suit the present purpose. We also provide examples of how the theory can help the practitioner. Textbooks such as Ferguson (1967) and DeGroot (1970) and the recent essay by Brown (2000) provide a more comprehensive treatment of the theory.

Setting up the general decision problem requires a prior distribution, a sampling distribution, a class of allowable actions and decision rules, and a loss function. Specific loss function forms correspond to point estimation, interval estimation, and hypothesis testing. We use the notation

$$\begin{aligned}
\textit{prior distribution:} &\quad G(\theta),\ \theta \in \Theta \\
\textit{sampling distribution:} &\quad f(\mathbf{x}|\theta) \\
\textit{allowable actions:} &\quad a \in \mathcal{A} \\
\textit{decision rules:} &\quad d \in \mathcal{D} : \mathcal{X} \rightarrow \mathcal{A} \\
\textit{loss function:} &\quad l(\theta, a)\ .
\end{aligned} \tag{B.1}$$

The loss function $l(\ \theta\ \ a)$ computes the loss incurred when $\theta$ is the true state of nature and we take action *a.* Thus for point estimation of a parameter $\theta$, we might use *squared error loss* (SEL),

$$l(\theta, a) = (\theta - a)^2$$

or *weighted squared error loss* (WSEL),

$$l(\theta, a) = w(\theta)(\theta - a)^2 ,$$

or *absolute error loss,*

$$l(\theta, a) = |\theta - a| ,$$

or, for discrete parameter spaces, 0-1 loss,

$$l(\theta, a) = \begin{cases} 0 & \text{if } \theta = a \\ 1 & \text{if } \theta \neq a \end{cases} . \tag{B.2}$$

The decision rule $d$ maps the observed data $\mathbf{x}$ into an action $a$; in the case of point estimation, this action is a proposed value for $\theta$ (e.g., $d(x) = \bar{x}$

The Bayesian outlook on the problem of selecting a decision rule is as follows. In light of the data $\mathbf{x}$, our opinion as to the state of nature is summarized by the *posterior* distribution of $\theta$, which is given by Bayes' Theorem as

$$dG(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)dG(\theta)}{m_G(\mathbf{x})} ,$$

where

$$m_G(\mathbf{x}) = \int f(\mathbf{x}|u)dG(u) .$$

The Bayes rule minimizes the *posterior risk,*

$$\rho(G, d(\mathbf{x})) = E_{\theta|\mathbf{x}}[l(\theta, d(\mathbf{x}))] = \int l(\theta, d(\mathbf{x}))dG(\theta|\mathbf{x}) . \tag{B.3}$$

Note that posterior risk is a single number regardless of the dimension of $\theta$, so choosing $d$ to minimize $\rho(G, d(\mathbf{x}))$ is well defined.

### *B.1.1 Risk and admissibility*

In the frequentist approach, we have available neither a prior nor a posterior distribution. For a decision rule *d,* define its *frequentist risk* (or simply *risk*) for a given true value of $\theta$ as

$$R(\theta, d) = E_{\mathbf{x}|\theta}[l(\theta, d(\mathbf{x}))] = \int l(\theta, d(\mathbf{x}))f(\mathbf{x}|\theta)d\mathbf{x} , \tag{B.4}$$

the average loss, integrated over the distribution of X conditional on 0. Note that frequentist risk is a *function of* $\theta$ not a single number like the posterior risk (B.3).

**Example B.1** In the interval estimation setting of Example 1.1, consider the loss function

$$l(\theta, a) = \begin{cases} 0 & \text{if } \theta \in \delta(\mathbf{x}) \\ 1 & \text{if } \theta \notin \delta(\mathbf{x}) \end{cases} .$$

Then the risk function is

$$\begin{aligned} R(8, d) &= E_{\mathbf{X}|\theta,\sigma^2}[l(0,\delta(\mathbf{X}))] \\ &= P_{\mathbf{X}|\theta,\sigma^2}[\theta \quad 6(X)] \\ &= .05\,, \end{aligned}$$

which is constant over all possible values of $\theta$ and $\sigma^2$. ■

This controlled level of risk across all parameter values is one of the main selling points of frequentist confidence intervals. Indeed, plotting $R$ versus $\theta$ for various candidate rules can be very informative. For example, if $d_1$ and $d_2$ are two rules such that

$$R(\theta, d_1) \;\#\; R(\theta, d_2) \quad \text{for all } \theta\,,$$

with strict inequality for at least one value of $\theta$, then under the given loss function we would never choose $d_2$, since its risk is never smaller that $d_1$'s and can be larger. In such a case, the rule $d_2$ is said to be *inadmissible,* and *dominated* by $d_1$. Note that $d_1$ may itself be inadmissible, since there may be yet another rule which uniformly beats $d_1$. If no such rule exists, $d_1$ is called *admissible.*

Admissibility is a sensible and time-honored criterion for comparing decision rules, but its utility in practice is rather limited. While inadmissible rules often need not be considered further, for most problems there will be many admissible rules, creating the problem of which one to select. Moreover, admissibility is not a guarantee of sensible performance; the example in Subsection 4.2.1 provides an illustration. Therefore, additional criteria are required to select a frequentist rule. We consider the three most important such criteria in turn.

### *B.1.2 Unbiased rules*

*Unbiasedness* is a popular method for reducing the number of candidate decision rules and allowing selection of a "best" rule within the reduced class. A decision rule $d(x)$ *is* unbiased if

$$E_{x|\theta}[l(\theta', d(x))] \geq E_{x|\theta}[l(\theta, d(x))] \text{ for all } \theta \text{ and } \theta'\ . \qquad \text{(B.5)}$$

Under squared error loss, (B.5) is equivalent to requiring $E_{x|\theta}[d(x)] = \theta$. For interval estimation, unbiasedness requires that the interval have a greater chance of covering the true parameter than any other parameter. For hypothesis testing, unbiasedness is equivalent to the statistical power being greater under an alternative hypothesis than under the null.

Though unbiasedness has the attractive property of reducing the number of candidate rules and allowing the frequentist to find an optimal rule, it is often a very high price to pay - even for frequentist evaluations. Indeed, a principal thesis of this book is that a little bias can go a long way in improving frequentist performance. In Section B.2 below we provide the basic example based on squared error loss. Here we present three estimation examples of how unbiasedness can work against logical and effective inference.

**Example B.2 : Estimating P(no events)** Ferguson (1967) provides the following compelling example of a problem with unbiased estimates. Assume a Poisson distribution and consider the goal of estimating the probability that there will be no events in a time period of length $2t$ based on the observed number of events, $X$, in a time period of length $t$. Therefore, with $\lambda$ the Poisson parameter we want to estimate $e^{-2t\lambda}$, based on $X$ which is distributed Poisson$(t\lambda)$. The MLE of this probability is $e^{-2X}$, while the best (indeed, the only) unbiased estimate is $(-1)^X$. This latter result is derived by matching terms in two convergent power series.

This unbiased estimate is patently absurd; it is either +1 or -1 depending on whether an even or an odd number of events has occurred in the first t time units. Using almost any criterion other than unbiasedness, the MLE is preferred. For example, it will have a substantially smaller MSE. ■

**Example B.3: Constrained parameter space.** In many models the unbiased estimate can lie outside the allowable parameter space. A common example is the components of variance models (i.e., the Type II ANOVA model), where

$$Y_{ij} = \mu + b_i + \epsilon_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, n_i,$$

where $\mu \in \Re$, $b_i \overset{iid}{\sim} N(0, \tau^2)$, and $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$. The unbiased estimate of $\tau^2$, computed using the within and between mean squared errors, can sometimes take negative values (as can the MLE). Clearly, performance of these estimates is improved by constraining them to be nonnegative. ■

**Example B.4: Dependence on the stopping rule**. As with all frequentist evaluations, unbiasedness requires attention to the sampling plan and, if applicable, to the stopping rule used in collecting the data. Consider again Example 1.2, where now we wish to estimate the unknown success probability based on $x$ observed successes in $n$ trials. This information unambiguously determines the MLE as $x/n$, but the unbiased estimate depends on how the data were obtained. If $n$ is fixed, then the MLE is unbiased. But if the sampling were "inverse" with data gathered until failures were obtained, then the number of observed successes, $x$, is negative binomial, and the unbiased estimate is $\frac{x}{x+r-1} = \frac{x}{n-1}$. If both $n$ and $x$ are random, it may be impossible to identify an unbiased estimate. ■

### *B.1.3 Bayes rules*

While many frequentists would not even admit to the existence of a subjective prior distribution $G(\theta)$, a possibly attractive frequentist approach to choosing a best decision rule does depend on the formal use of a prior. For a prior $G$, define the *Bayes risk* as

$$r(G,d) = E_\theta E_{\mathbf{x}|\theta} l(\theta, d(\mathbf{x})) = E_\theta R(\theta, d) \ , \tag{B.6}$$

the expected frequentist risk with respect to the chosen prior $G$. Bayes risk is alternatively known as *empirical Bayes risk* because it averages over the variability in both $\theta$ and the data. Reversing the order of integration in (B.6), we obtain the alternate computational form

$$r(G,d) = E_{\mathbf{X}} E_{\theta|\mathbf{X}} l(\theta, d(\mathbf{x})) = E_{\mathbf{X}} \rho(G, d(\mathbf{x})) \ ,$$

the expected posterior risk with respect to the marginal distribution of the data $\mathbf{X}$. Since this is the posterior loss one expects *before* having seen the data, the Bayes risk is sometimes referred to as the *preposterior risk.* Thus $r(G, d)$ is directly relevant for Bayesian experimental design (see Section 4.5).

Since $r(G,d)$is a scalar quantity, we can choose the rule $d_G$ that minimizes the Bayes risk, i.e.,

$$d_G(\mathbf{x}) = \arg\min_{d \in \mathcal{D}} r(G,d) \ . \tag{B.7}$$

This minimizing rule is called the *Bayes rule.* This name is somewhat confusing, since a subjective Bayesian would not average over the data (as $r(G, d)$ does) when choosing a decision rule, but instead choose one that minimized the posterior risk (B.3) given the observed data. Fortunately, it turns out that these two operations are virtually equivalent: under very broad conditions, minimizing the Bayes risk (B.6) is equivalent to minimizing the posterior risk (B.3) for all $x$ such that $m_G(\mathbf{x}) > 0$. For further discussion of this result and related references, see Berger (1985, p.159).

**Example B.5: Point estimation**. Under SEL the Bayes rule is found by minimizing the posterior risk,

$$\rho(G,a) = \int (\theta - a)^2 g(\theta|\mathbf{x}) d\theta \ .$$

Taking the derivative with respect to a, we have

$$\frac{\partial}{\partial a}[\rho(G,a)] = \int 2(\theta - a)(-1) g(\theta|\mathbf{x}) d\theta \ .$$

Setting this expression equal to zero and solving for a, we obtain

$$a = \int \theta \, g(\theta|\mathbf{x}) d\theta = E(\theta|\mathbf{x}) \ ,$$

the posterior mean, as a solution. Since

$$\frac{\partial^2}{\partial a^2}[\rho(G,a)] = 2 \int g(\theta|\mathbf{x}) d\theta = 2 > 0 \ ,$$

the second derivative test implies our solution is indeed a minimum, confirming that the posterior mean is the Bayes estimate of $\theta$ under SEL.

We remark that under absolute error loss, the Bayes estimate is the posterior median, while under 0-1 loss, it is the posterior mode. ■

**Example B.6: Interval estimation**. Consider the loss function

$$l(\theta, a) = I_{\{\theta \notin a\}} + \text{c x volume(a)},$$

where *I* is the indicator function of the event in curly brackets and a is a subset of the parameter space $\Theta$. Then the Bayes rule will be the region (or regions) for $\theta$ having highest posterior density, with c controlling the tradeoff between the volume of *a* and the posterior probability of coverage. Subsection 2.3.2 gives a careful description of Bayesian confidence intervals; for our present purpose we note (and Section 4.3 demonstrates) that under noninformative prior distributions, these regions produce confidence intervals with excellent frequentist coverage and modest volume. ■

**Example B.7: Hypothesis testing**. Consider the comparison of two simple hypotheses, $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, so that we have a two-point parameter space $\Theta = \{\theta_0, \theta_1\}$. We use the "$0$–$l_i$" loss function

$$l(\theta, a) = al_0 + (1 - a)l_1,$$

where both $l_0$ and $1_1$ are nonnegative, and a $\in$ {0, 1} gives the index of the accepted hypothesis. Then it can be shown that the Bayes rule with respect to a prior on $\pi = P(\theta = \theta_0)$ is a likelihood ratio test; in fact, many authors use this formulation to prove the Neyman-Pearson lemma. This model is easily generalized to a countable parameter space. ■

### *B.1.4 Minimax rules*

A final (and very conservative) approach to choosing a best decision rule in a frequentist framework is to control the worst that can happen: that is, to minimize the maximum risk over the parameter space. Mathematically, we choose the rule $d*$ satisfying the relation

$$\sup_{\theta \in \Theta} R(\theta, d^*) = \inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, d) .$$

This $d*$ may sacrifice a great deal for this control, since its risk may be unacceptably high in certain regions of the parameter space. The minimax rule is attempting to produce controlled risk behavior for all $\theta$ *and x,* but is making no evaluation of the relative likelihood of the $\theta$ values.

The Bayesian approach is helpful in finding minimax rules via a theorem which states that a Bayes rule (or limit of Bayes rules) that has constant risk over the parameter space is minimax. Generally, the minimax rule is not unique (indeed, in Chapter 3 we find more than one), and failure to find a constant risk Bayes rule does not imply that the absence of a minimax rule. Of course, as with all other decision rules, the minimax rule depends on the loss function. For the Gaussian sampling distribution, the sample mean is minimax (it is constant risk and the limit of Bayes rules). Finding minimax rules for the binomial and the exponential distributions have been left as exercises.

## B.2 Procedure evaluation and other unifying concepts

### *B.2.1 Mean squared error*

To see the connection between the Bayesian and frequentist approaches, consider again the problem of estimating a parameter under squared error loss (SEL). With the sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, let $d(\mathbf{X})$ be our estimate of $\theta$. Its frequentist risk is

$$E_{\mathbf{x}|\theta}[l(\theta, d(\mathbf{x}))] = E_{\mathbf{x}|\theta}[(\theta - d(\mathbf{x}))^2] \equiv MSE_d(\theta) \ ,$$

the *mean squared error* of $d$ given the true $\theta$. Its posterior risk with respect to some prior distribution $G$ is

$$E_{\theta|\mathbf{x}}[l(\theta, d(\mathbf{x}))] = E_{\theta|\mathbf{x}}[(\theta - d(\mathbf{x}))^2] \equiv MSE_{d,G}(\mathbf{x}) \ ,$$

and its preposterior risk is

$$E_{\theta,\mathbf{x}}[l(\theta, d(\mathbf{x}))] = E_{\theta,\mathbf{x}}[(\theta - d(\mathbf{x}))^2] \equiv MSE_{d,G} \ ,$$

which equals $E_\theta[MSE_d(\theta)]$ and $E_\mathbf{x}[MSE_{d,G}(\mathbf{x})]$. We have not labeled $d$ as "frequentist" or "Bayesian;" it is simply a function of the data. Its properties can be evaluated using any frequentist or Bayesian criteria.

### *B.2.2 The variance-bias tradeoff*

We now indicate how choosing estimators having minimum MSE requires a tradeoff between reducing variance and bias. Assume that conditional on $\theta$, the data are i.i.d. with mean $\theta$ and variance $\sigma^2(\theta)$. Let $d(\mathbf{x})$ be the sample mean, $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, and consider estimators of $\theta$ of the form

$$d_c(\mathbf{x}) = cX \ . \tag{B.8}$$

For 0 < c < 1, this estimator shrinks the sample mean toward 0 and has the following risks:

$$\begin{aligned} \text{frequentist:} \quad & c^2\frac{\sigma^2(\theta)}{n} + (1-c)^2\theta^2 \\ \text{posterior:} \quad & Var(\theta|\mathbf{x}) + [c\bar{x} - E(\theta|\mathbf{x})]^2, \\ \text{preposterior (Bayes):} \quad & c^2\frac{E[\sigma^2(\theta)]}{n} + (1-c)^2\{Var(\theta) + [E(\theta)]^2\}, \end{aligned}$$

where in the last expression the expectations and variances are with respect to the prior distribution G. Notice that all three expressions are of the form "variance + bias squared."

Focusing on the frequentist risk, notice that the first term decreases like $\frac{1}{n}$ and the second term is constant in $n$. If $\theta$ = 0 or is close to 0, then a c near 0 will produce small risk. More generally, for relatively small $n$, it will be advantageous to set c to a value betweeen 0 and 1 (reducing the variance at the expense of increasing the bias), but, as $n$ increases, c should converge to 1 (for large $n$, bias reduction is most important). This phenomenon underlies virtually all of statistical modeling, where the standard way to reduce bias is to add terms (degrees of freedom, hence more variability) to the model.

Taking c = 1 produces the usual unbiased estimate with MSE equal to the sampling variance. Comparing c = 1 to the general case shows that if

$$\frac{n\theta^2 - \sigma^2(\theta)}{n\theta^2 + \sigma^2(\theta)} < c \leq 1,$$

then it will be advantageous to use the biased estimator. This relation can be turned around to find the interval of $\theta$s around 0 for which the estimator beats $\bar{X}$. This interval has endpoints that shrink to 0 like $\frac{1}{\sqrt{n}}$.

Turning to the Bayesian posterior risk, the first term also decreases like $\frac{1}{n}$, and the Bayes rule (not restricted to the current form) will be $E(\theta|\mathbf{x})$. The c that minimizes the preposterior risk is

$$c = c_n = \frac{n\{Var(\theta) + [E(\theta)]^2\}}{n\{Var(\theta) + [E(\theta)]^2\} + E[\sigma^2(\theta)]}\ .$$

As with the frequentist risk, for relatively small $n$ it is advantageous to use $c \neq 1$ and have $c_n$ increase to 1 as $n$ increases.

Similar results to the foregoing hold when an estimator (B.8) is augmented by an additive offset (i.e., $d_{a,c}(\mathbf{x}) = a + c\bar{X}$). We have considered SEL and MSE for analytic simplicity; qualitatively similar conclusions and relations hold for a broad class of convex loss functions.

*Shrinkage* estimators of the form (B.8) arise naturally in the Bayesian context. The foregoing discussion indicates that they can have attractive frequentist properties (i.e., lower MSE than the standard estimator, $\bar{X}$). Under squared error loss, one wants to strike a tradeoff between variance and bias, and rules that effectively do so will have good properties.

## B.3 Other loss functions

We now evaluate the Bayes rules for three loss functions that show the effectiveness of Bayesian structuring and how intuition must be aided by the rules of probability. Chapter 4 and Section 7.1 present additional examples of the importance of Bayesian structuring.

### *B.3.1 Generalized absolute loss*

Consider the loss function for $0 \leq p \leq 1$,

$$\begin{array}{rl} p|\theta - a| & \text{if} \quad a < \theta \\ (1-p)|\theta - a| & \text{if} \quad a \geq \theta \ . \end{array}$$

It is straightforward to show that the optimal estimate is the $p^{th}$ percentile of the posterior distribution. Therefore, $p$ = .5 gives the posterior median. Other values of $p$ are of more than theoretical interest, as illustrated below in Problem 3.

### *B.3.2 Testing with a distance penalty*

Consider a hypothesis testing situation wherein there is a penalty for making an incorrect decision and, if the null hypothesis is rejected, an additional penalty that depends on the distance of the parameter from its null hypothesis value. The loss function

$$l(\theta,\ a) = \begin{cases} 0 & , \quad a = 0, \theta < 0 \text{ or } a = 1, \theta > 0 \\ \theta^2 & , \quad a = 0, \theta > 0 \text{ or } a = 1, \theta < 0 \end{cases}$$

computes this score. Assume the posterior distribution is $N(\mu, \tau^2)$. Then

$$R(\theta, a = 1) - R(\theta, a = 0) = \tau^2[1 - 2\Phi(\mu/\tau)] + \frac{\mu}{\tau}\phi(\mu/\tau) \ ,$$

where the first term corresponds to 0-1 loss, while the second term corresponds to the distance adjustment. If $\mu$ > 0, 0-1 loss implies $a$ = 1, but with the distance adjustment, if $\tau$ < 1, then for small $\mu$ > 0, $a$ = 0. This decision rule is non-intuitive, but best.

### *B.3.3 A threshold loss function*

Through its Small Area Income and Poverty Estimates (SAIPE) project, the Census Bureau is improving estimated poverty counts and rates used in allocating Title I funds (National Research Council, 2000). Allocations must satisfy a "hold-harmless" condition (an area-specific limit on the reduction in funding from one year to the next) and a threshold rule (e.g., concentration grants kick in when estimated poverty exceeds 15%). ACS (2000) demonstrates that statistical uncertainty induces unintended consequences of a hold-harmless condition, because allocations cannot appropriately adjust to changing poverty estimates. Similarly, statistical uncertainty induces unintended consequences of a threshold rule. Unintended consequences can be ameliorated by optimizing loss functions that reflect societal goals associated with hold-harmless, threshold, and other constraints. However, amelioration will be limited and societal goals may be better met through

| condition | eligible for concentration funds? | loss |
|---|---|---|
| $\theta \geq T$ | yes | $(\theta - a)^2$ |
| $\theta < T$ | no | $a^2$ |

Table B.1 *A threshold loss function.*

replacing these constraints by ones designed to operate in the context of statistical uncertainty.

**Example B.8** To show the quite surprising influence of thresholds on optimal allocation numbers, we consider a greatly simplified, mathematically tractable example. Let $\theta$ be the true poverty rate for a single area (e.g., a county), a be the amount allocated to the area per child in the population base, and $\mathbf{Y}$ denote all data. For a threshold $T$, consider the societal loss function in Table B.8. The optimal per-child allocation value is then

$$a_T(\mathbf{Y}) = E(\theta \mid \theta \geq T, \mathbf{Y}) \times \mathrm{pr}(\theta \geq T \mid \mathbf{Y}) .$$

Though the allocation formula as a function of the true poverty rate (0) has a threshold, the optimal allocation value has no threshold. Furthermore, for $T > 0$, $a_T(\mathbf{Y})$ is not the center of the the posterior distribution (neither the mean, median, nor mode). It is computed by multiplying the posterior mean (conditional on $\theta > T$) by a posterior tail area. This computation always produces an allocation value smaller than the posterior mean; for example, if the posterior distribution is exponential with mean $\mu$, $a_T = \mu(1+r)e^{-r}$ where $r = T/\mu$. But this reduction is compensated by the lack of a threshold in $a_T$. It is hard to imagine coming up with an procedure in this setting without Bayesian structuring.

Actual societal loss should include multiplication by population size, incorporate a dollar cap on the total allocation over all candidate areas, and "keep score" over multiple years. Absolute error should replace squared-error in comparing a to $\theta$. ■

## B.4 Multiplicity

There are several opportunities for taking advantage of multiple analyses, including controlling error rates, multiple outcomes, multiple treatments, interim analyses, repeated outcomes, subgroups, and multiple studies. Each of these resides in a domain for control of statistical properties, including control for a single analysis, a set of endpoints, a single study, a collection of studies, or even an entire research career! It is commonly thought that the Bayesian approach pays no attention to multiplicity. For example, as

| $\alpha$ | $Z_{\alpha/2}$ | $Z_{\alpha_{bf}/2}$ | $\sqrt{-2\log\alpha}$ |
|---|---|---|---|
| .01 | 2.57 | 2.81 | 3.03 |
| .02 | 2.33 | 2.57 | 2.80 |
| .05 | 1.96 | 2.24 | 2.45 |
| .10 | 1.65 | 1.95 | 2.15 |

Table B.2 *Comparison of upper univariate confindence limits, multiple comparisons setting with $K = 2$.*

shown in Example 1.2, the posterior distribution does not depend on a data monitoring plan or other similar aspects of an experimental design. However, a properly constructed prior distribution and loss function can acknowledge and control for multiplicity.

**Example B.9** When a single parameter is of interest in a $K$-variate parameter structure, the Bayesian formulation automaticallly acknowledges multiplicity. For the $K$ = 2 case, consider producing a confidence interval for $\theta_1$ when: $\boldsymbol{\theta} = (\theta_1,\ \theta_2) \sim N_2(\mathbf{0}, I_2)$. The (1 - $\alpha$   HPD region is

$$|| \boldsymbol{\theta} ||^2 \leq -2\log(\alpha)\ .$$

Projecting this region to the first coordinate gives

$$-\sqrt{-2\log(\alpha)} \leq \theta_1 \leq \sqrt{-2\log(\alpha)}\ .$$

Table B.2 shows that for $K$ = 2 the Bayesian HPD interval is longer than either the unadjusted or the Bonferroni adjusted intervals ($\alpha_{bf}$ is the Bonferroni non-coverage probability). That is, the Bayesian interval is actually more conservative! However, as $K$ increases, the Bayesian HPD interval becomes narrower than the Bonferroni interval. ■

## B.5 Multiple testing

### *B.5.1 Additive loss*

Under additive, component-specific loss, Bayesian inference with components that are independent a priori separately optimizes inference for each component with no accounting for the number of comparisons. However, use of a hyperprior Bayes (or empirical Bayes) approach links the components, since the posterior "borrows information" among components. The $k$-ratio $t$-test is a notable example. With $F$ denoting the $F$-test for a one-way ANOVA,

$$\begin{aligned} F &= (\hat{\sigma}^2 + K\hat{\tau}^2)/\hat{\sigma}^2 \\ (1 - \hat{B}) &= (F - 1)/F \end{aligned}$$

$$\text{and } Z_{12} = \left(\frac{F-1}{F}\right)^{1/2} \frac{Y_1 - Y_2}{\sqrt{2}\sigma}$$

The magnitude of $F$ adjusts the test statistic. For large $K$, under the global null hypothesis $H_0 : P(\text{all } Z_{ij} = 0) \geq 0.5$, the rejection rate is much smaller than 0.5. Note that the procedure depends on the estimated prior mean $(\hat{\mu})$ and shrinkage $(\hat{B})$ and the number of candidate coordinates can influence these values. If the candidate coordinates indicate high heterogeneity, then $\hat{B}$ is small as is the Bayesian advantage. Therefore, it is important to consider what components to include in an analysis to control heterogeneity while retaining the opportunity for discovery.

### *B.5.2 Non-additive loss*

If one is concerned about multiplicity, the loss function should reflect this concern. Consider a testing problem with a loss function that penalizes for individual coordinate errors and adds an extra penalty for making two errors:

$$\begin{aligned} \text{parameters:} \quad & \theta_1, \theta_2 \in \{0, 1\} \\ \text{decisions:} \quad & a_1, a_2 \in \{0,\ 1\} \\ l(a,\ \boldsymbol{\theta}) \quad = \quad & a_1(1-\theta_1)\ell_0 + (1-a_1)\theta_1\ell_1 \\ & + a_2(1-\theta_2)\ell_0 + (1-a_2)\theta_2\ell_1 \\ & + \gamma(1-\theta_1)(1-\theta_2)a_1a_2. \end{aligned}$$

With $\ell_0 = \ell_1 = 1$ and $G_{ij} = P(\theta_1 = i,\ \theta_2 =$ j data), a straightforward computation shows that the risk is

$$a_1(1-2G_{1+}) + a_2(1-2G_{+1}) + \gamma a_1 a_2 G_{00} + \{G_{1+} + G_{+1}\}\ .$$

The four possible values of $(a_1,\ a_2)$ then produce the risks in Table B.3, and the the Bayes decision rule (the $(a_1, a_2)$ that minimize risk) is given in Table B.4. Note that in order to declare the second component "$a_2 = 1$," we require more compelling evidence than if $\gamma = 0$.

| $a_1$ | $a_2$ | risk $- \{G_{1+} + G_{+1}\}$ |
|---|---|---|
| 0 | 0 | $0$ |
| 1 | 0 | $1 - 2G_{1+}$ |
| 0 | 1 | $1 - 2G_{+1}$ |
| 1 | 1 | $(1 - 2G_{1+}) + (1 - 2G_{1+}) + \gamma G_{00}$ |

Table B.3 *Risks for the possible decision rules.*

| condition | optimal rule |
|---|---|
| $G_{1+} \le .5, G_{+1} \le .5$ | $a_1 = 0, a_2 = 0$ |
| $G_{1+} \le .5, G_{+1} > .5$ | $a_1 = 0, a_2 = 1$ |
| $G_{1+} > .5, G_{+1} \le .5$ | $a_1 = 1, a_2 = 0$ |
| $G_{1+} > G_{+1} > .5$ | $a_1 = 1,\ a_2 = \begin{cases} 0, & \text{if } (2G_{+1} - 1) < \gamma G_{00} \\ 1, & \text{if } (2G_{+1} - 1) \ge \gamma G_{00} \end{cases}$ |

Table B.4 *Optimal decision rule for non-additive loss.*

*Discussion*

Statistical decision rules can be generated by any philosophy under any set of assumptions. They can then be evaluated by any criteria - even those arising from an utterly different philosophy. We contend (and much of this book shows) that the Bayesian approach is an excellent "procedure generator," even if one's evaluation criteria are frequentist. This somewhat agnostic view considers the parameters of the prior, or perhaps the entire prior, as "tuning parameters" that can be used to produce a decision rule with broad validity. Of course, no approach automatically produces broadly valid inferences, even in the context of the Bayesian models. A procedure generated assuming a cocksure prior that turns out to be far from the truth will perform poorly in both the Bayesian and frequentist senses.

## B.6 Exercises

1. Show that the Bayes estimate of a parameter $\theta$ under weighted squared error loss, $l(\theta, a) = w(\theta)(\theta - a)^2$, is given by
$$d_\pi(\mathbf{x}) = \frac{E[\theta w(\theta)|\mathbf{x}]}{E[w(\theta)|\mathbf{x}]} .$$

2. Suppose that the posterior distribution of $\theta$, $p(\theta|\mathbf{x})$, is discrete with support points $\{\theta_1, \theta_2, \ldots\}$. Show that the Bayes rule under 0-1 loss (B.2) is the posterior mode.

3. Consider the following loss function:
$$l(\theta, a) = \begin{cases} p|\theta - a|, & \theta > a \\ (1-p)|\theta - a|, & \theta \le a \end{cases}$$
Show that the Bayes estimate is the 100 x $p^{th}$ percentile of the posterior distribution.

(Note: The posterior median is the Bayes estimate for $p$ = .5. This loss function does have practical application for other values of $p$. In many organizations, employees are allowed to set aside pre-tax dollars for health care expenses. Federal rules require that one cannot get back funds not used in a 12-month period, so one should not put away too much. However, putting away too little forces some health care expenses to come from post-tax dollars. If losses are linear in dollars (or other currency), then this loss function applies with $p$ equal to the individual's marginal tax rate.)

4. For the binomial distribution with $n$ trials, find the minimax rule for squared error loss and normalized squared error loss, $L(\theta, a) = \frac{(\theta-a)^2}{\theta(1-\theta)}$. (*Hint:* The rules are Bayes rules or limits of Bayes rules.)

   For each of these estimates, compute the frequentist risk (the expected loss with respect to $f(x|\theta)$), and plot this risk as a function of $\theta$ (the *risk plot).* Use these risk plots to choose and justify an estimator when $\theta$ *is*

   (a) the head probability of a randomly selected coin;

   (b) the failure probability for high-quality computer chips;

   (c) the probability that there is life on other planets.

5. For the exponential distribution based on a sample of size $n$, find the minimax rule for squared error loss.

6. Find the frequentist risk under squared error loss (i.e., the MSE) for the estimator $d_{a,c}(\mathbf{x}) = a + cX$. What true values of $\theta$ favor choosing a small value of $c$ (high degree of shrinkage)? Does your answer change if $n$ *is* large?

7. Let X be a random variable with mean $\mu$ and variance $\sigma^2$. You want to estimate $\mu$ under SEL, and propose an estimate of the form *(1 - b)X.*

   (a) Find $b$*, the $b$ that minimizes the MSE. (*Hint:* Use the "variance + (bias)$^2$" representation of MSE.)

   (b) Discuss the dependence of $b*$ on$\mu$ and $\sigma^2$ and its implications on the role of shrinkage in estimation.

8. Consider a linear regression having true model of the form $Y_i = \alpha + \beta x_i + \epsilon_i$, where the $\epsilon_i$ are i.i.d. with mean 0 and variance $\sigma^2$. Suppose you have a sample of size $n$, use least-squares estimates (LSEs), and want to minimize the *prediction error:*

$$\text{PE} = E[Y_{new} - \hat{Y}(X_{new})^2] .$$

   Here, $Y(x)$ is the prediction based on the estimated regression equation and $X_{new}$ is randomly selected from a distribution with mass $\frac{1}{n}$ at each of the $x_1, \ldots, x_n$.

(a) Assume that $\sum_i x_i = 0$ and show that even though you may *know* that $\beta \neq 0$, if $\sum_i x_i^2$ is sufficiently small, it is better to force $\hat{\beta} = 0$ than to use the LSE for it.
(Note: This result underlies the use of the $C_p$ statistic and PRESS residuals in selecting regression variables; see the documentation for SAS Proc REG.)

(b) *(More difficult)* As in the previous exercise, assume you plan to use a prediction of the form $\hat{Y}(x) = \hat{\alpha} + (1-b)\hat{\beta}x$. Find $b*$, the optimal value of *b.*

9. In an errors-in-variables simple regression model, the least squares estimate of the regression slope $(\beta)$ is biased toward 0, an example of *attenuation.* Specifically, if the true regression (through the origin) is $Y = x\beta + \epsilon$, but Y is regressed on X, with $X = x + \delta$, then the least squares estimate $(\hat{\beta})$ has expectation: $E[\hat{\beta}] \approx \rho\beta$, with $\rho = \frac{\sigma_x^2}{\sigma_x^2+\sigma_\delta^2} \leq 1$.

   If $\rho$ is known or well-estimated, one can correct for attenuation and produce an unbiased estimate by using $\hat{\beta}/\rho$ to estimate $\beta$. However, this estimate may have poor MSE properties. To minimize MSE, consider an estimator of the form: $\hat{\beta}_c = c(\hat{\beta}/\rho)$.

   (a) Let $\sigma^2 = Var[\hat{\beta}]$ and find the *c* that minimizes MSE.

   (b) Discuss the solution's implications on the variance/bias tradeoff and the role of shrinkage in estimation.

10. Compute the HPD confidence interval in Example B.9 for a general *K*, and compare its length to the unadjusted and Bonferrom adjusted intervals.