1 Statistical rules, loss and risk

We saw that a major focus of classical statistics is comparing various inference methods, e.g., comparing power and efficiency of testing rules. A formal study of this was initiated by Abraham Wald (Ann. Math. Stat. 1939), later dubbed as Decision Theory by Eric Lehman in the 1950s. Decision theory provides a synthesis of many concepts in statistics, including frequentist properties, loss functions, Bayesian procedures and so on. These notes provide a brief overview of the concepts and results.

2 Decisions, loss and risk

Consider data $X \in S$ modeled as $X \sim f(x|\theta), \theta \in \Theta$. Decision theory deals with the case where based on observed data $X = x$, one has to choose an “action” $a$ from a pre-specified action space $A$. A typical example is hypothesis testing, where $A = \{\text{reject } H_0, \text{accept } H_0\}$. A statistical decision method or rule is a map $\delta$ from $S$ to $A$. A comparison between various rules is relevant only when the consequences of such decision making have been well laid out in the form of a loss function $L(\theta, a)$ which measures the loss incurred when the true parameter value is $\theta$ and an action $a$ is taken. We will work with non-negative loss functions. For any decision rule $\delta$, its risk function is defined as the expected loss:

$$R(\theta, \delta) := E\{L(\theta, \delta(X))|\theta\}.$$

As we have seen with testing rules, it is not meaningful to seek the rule with uniformly smallest risk at every $\theta$. Decision theory deals with other more reasonable modes of comparison and offers some interesting results on various modes of rule generation processes, such as the Bayesian approach, MLE, shrinkage.

Keeping with convention, we will restrict our discussion to the task of “parameter estimation”, where one is trying to provide a point estimate of quantity $\eta = \eta(\theta) \in \mathcal{E}$. In this case $A = \mathcal{E}$ and the loss function is some sort of distance measure between $\eta(\theta)$ and the reported estimate $a$. Common loss functions are:

$$L(\theta, a) = \begin{cases} 
(\eta(\theta) - a)^2 & \text{squared-error loss} \\
|\eta(\theta) - a|^p & \text{$L_p$-loss for some } p > 0 \\
I\{|\eta(\theta) - a| > c\} & \text{0-1 loss for some } c > 0.
\end{cases}$$

I do not view point estimation as a statistical inference task, but it is often the first step toward hypothesis testing of interval estimation.
Figure 1: Risk functions under square error loss for 5 rules: \( \delta_1(X) = X \) (black), \( \delta_2(X) \equiv 1 \) (red), \( \delta_3(X) = X I(|X| > 2) \) (green), \( \delta_4(X) = 0.5X + 1 \) (blue) and \( \delta_5(X) = 1.1X \) (cyan) for scalar data \( X \) modeled as \( X \sim N(\theta, 1) \) and the quantity of interest is \( \theta \) itself.

3 Risk comparison

An obvious difficulty with comparing rules by their risk functions is that the latter are not scalars, and hence not necessarily well ordered. Figure 1 shows this for the scalar data \( X \) modeled as \( X \sim N(\theta, 1) \), \( \theta \in \mathbb{R} \). Various notions have been proposed to overcome this difficulty:

3.1 Admissibility

A rule is called inadmissible if there exists another rule with a risk function that is everywhere smaller or equal and strictly smaller at some \( \theta \). In Figure 1, \( \delta_5 \) is inadmissible because \( \delta_1 \) dominates it everywhere. A rule is admissible if it is not inadmissible. Admissibility does not rule out rules like \( \delta_2 \), which wagers all its money on \( \theta = 1 \), and cannot be beaten at that \( \theta \) value.

3.2 Best among a restricted class

One might seek the rule that has uniformly the lowest risk within a smaller class of rules. For point estimation, there are two popular choices for such restrictions:

1. Unbiased estimators. Restrict to estimators \( \delta(X) \) such that \( \mathbb{E}\{\delta(X)|\theta\} = \eta(\theta) \) for all \( \theta \in \Theta \). For squared error loss:

   \[
   R(\theta, \delta) = \text{Var}\{\delta(X)|\theta\} + [\mathbb{E}\{\delta(X)|\theta\} - \eta(\theta)]^2
   \]
and hence the risk of an unbiased estimator at \( \theta \) equals its variance at \( \theta \). The minimum risk estimator within this class, if it exists, is called the uniformly minimum variance unbiased (UMVU) estimator. We will later see that \( \delta_1(X) = X \) is the UMVU estimator for the problem in Figure 1.

2. **Equivariant estimators** A number of statistical problems come with some notion of “invariance under certain transformation”. Consider for example measuring the mean population height of 6 year old girls. It is desirable that an estimation rule does not lead to conflicting estimates based on whether measurements are taken in inches or centimeters. Rules with such invariant properties are called equivariant rules. For simple geometric transformations, such as location shift, or scaling or rotation, one can often find a rule with uniformly smallest risk among the class of equivariant rules. For the problem in Figure 1, the rule \( \delta(X) = X \) is also the minimum risk equivariant estimator under location shifts and/or scale changes.

3.3 **Scalar risk summaries**

One approach to avoid the ordering problem is to summarize the entire risk function of a rule into a single number. Two popular summaries are:

1. **Weighted average & Bayes rules.** One could use a weighting function \( w(\theta) \) on \( \Theta \) to obtain a single number summary: \( r(w, \delta) = \int R(\theta, \delta) w(\theta) d\theta \). If \( w \) has a finite integral over \( \Theta \), then it is equivalent to consider only the normalized version \( \pi \) of \( w \). If \( \pi \) is a pdf, then any rule \( \delta \) that minimizes \( r(\pi, \delta) \) is called a Bayes rule w.r.t. \( \pi \). A Bayes rule always exists and can be found as:

\[
\delta_\pi(x) = \arg\min_{a \in A} \int L(\theta, a) \pi(\theta|x) d\theta.
\]

For squared error loss, the (unique) Bayes rule is simply the posterior mean estimate: \( \delta(x) = E(\eta(\theta)|x) = \int \eta(\theta(\pi|\theta|x) d\theta \). The rules \( \delta_2(x) \equiv 1 \) and \( \delta_4(x) = 0.5x + 1 \) in Figure 1 are both Bayes rules, with \( \pi(\theta) = N(1, 0) \) (i.e., the Dirac measure at 1) and \( \pi(\theta) = N(2, 1) \) respectively. The rule \( \delta_1(x) = x \) is a formal Bayes rule under the improper prior \( \pi(\theta) \propto 1 \). It is also a generalized Bayes rule, i.e., the limit of the sequence of Bayes rules generated from a sequence of proper priors.

2. **Maximum risk & minimax rules.** The weighting approach requires some background knowledge of which \( \theta \) weigh more heavily than others to define the average risk. A more omni-purpose strategy is to guard against the worst case scenario, given by the maximum-risk \( \sup_{\theta \in \Theta} R(\theta, \delta) \) associated with a rule \( \delta \). A rule with the minimum maximum-risk is called a minimax rule. In Figure 1, \( \delta_1 \) has the lower maximum-risk among the five rules. It turns out that \( \delta_1 \) is indeed the minimax rule for this problem. We will later see, however, that a minimax rule need not necessarily be admissible.
4 More on UMVU

4.1 Sufficiency & Rao-Blackwellization

A statistic $T(X)$ is called sufficient for $\theta$ if the conditional distribution of $X$ given $T(X)$ does not depend on $\theta$, i.e., for any $t \in T$ the set of values $T(X)$ can assume,

$$P(X \in A | \theta = \theta_1, T(X) = t) = P(X \in A | \theta = \theta_2, T(X) = t)$$

for every $\theta_1, \theta_2 \in \Theta$ and $A \subset S$. Any unbiased estimator $\hat{\delta}(X)$ that is not a function of a sufficient statistic $T(X)$ can be improved by considering its Rao-Blackwellized version $\tilde{\delta}(X) = E\{\delta(X)|T(X)\}$, which also is unbiased. Note that this expectation could be carried out under any value of $\theta$. This holds because of the usual "total variance identity":

$$\text{Var}\{\delta(X)|\theta\} = E[\text{Var}\{\delta(X)|\theta, T(X)\}|\theta] + \text{Var}[E\{\delta(X)|\theta, T(X)\}|\theta]$$

and hence $\text{Var}\{\tilde{\delta}(X)|\theta\} < \text{Var}\{\delta(X)|\theta\}$.

Sufficient statistics are easiest to identify by the factorization theorem: $T(X)$ is sufficient if and only if $f(x|\theta) = h(x)g(T(x), \theta)$ for some function $g(t, \theta)$. This technique could be used to argue that for $X_i \overset{\text{IID}}{\sim} \text{Poisson}(\mu)$, $\mu \in \mathbb{R}_+$, $T(X) = \sum_{i=1}^n X_i$ is a sufficient statistic for $\mu$. More generally if $X \sim f(x|\theta) = h(x) \exp\{\theta T(x) - A(\theta)\}$ is an exponential family model, then $T(X)$ is a sufficient statistic for $\theta$. As a non-exponential family example, for the model $X_i \overset{\text{IID}}{\sim} \text{Uniform}(0, \theta), \theta \in \mathbb{R}_+, T(X) = \max(X_1, \ldots, X_n)$ is a sufficient statistic.

4.2 The information inequality

There is a limit to how much an unbiased estimate could be improved, essentially because there is a limit to how compressed a sufficient statistic could be. Another approach that gives such a limit is the information inequality. If $\delta(X)$ is unbiased for $\eta(\theta)$ then,

$$\text{Var}\{\delta(X)|\theta\} \geq \frac{\dot{\eta}(\theta)^2}{I_F(\theta)}$$

where $I_F(\theta)$ is the Fisher information at $\theta^1$. To prove this at a fixed $\theta$, let

$$S(X) = \dot{\ell}_X(\theta) = \frac{\partial}{\partial \theta} \log f(X|\theta)$$

denote the score function. Then, under regularity conditions, $E\{S(X)|\theta\} = 0$, $\text{Var}\{S(X)|\theta\} = I_F(\theta)$ and

$$E\{\delta(X)S(X)|\theta\} = \int \delta(x) \frac{\partial}{\partial \theta} \log f(x|\theta) f(x|\theta) dx = \frac{\partial}{\partial \theta} \int \delta(x) f(x|\theta) dx = \dot{\eta}(\theta)$$

\(^1\)There is multivariate version of the information inequality. Suppose $\text{dim}(\theta) = k$, and, $\text{dim}(\delta(X)) = \text{dim}(\eta(\theta)) = m$. Then,

$$\text{Var}\{\delta(X)|\theta\} \geq \dot{\eta}(\theta)^T I_F^{-1}(\theta) \dot{\eta}(\theta)$$

where, $\dot{\eta}(\theta)$ the $k \times m$ matrix of first order partial derivatives $\frac{\partial}{\partial \theta} \eta_j(\theta)$, and, for two square matrices $A, B$, by $A \geq B$ we mean $A - B$ is non-negative definite
from which the information inequality follows because of the inequality: \( \text{Cov}(W,V)^2 \leq \text{Var}(W)\text{Var}(V) \).

The proof above indicates that for a canonical exponential family model \( f(x|\theta) = h(x) \exp\{\theta T(x) - A(\theta)\} \), the lower bound is indeed achieved by \( \delta(X) = T(X) \) for estimating \( \eta(\theta) = \dot{A}(\theta) \). That is \( T(X) \) is UMVU estimator of \( \dot{A}(\theta) \). It can be proved that an UMVU estimator exists for every \( \eta(\theta) \) which admits an unbiased estimators.

### 4.3 Non-existence of UMVU estimators

Consider data \( X \in \mathbb{Z} \), where \( \mathbb{Z} \) denotes the set of all integers (positive, negative or zero). modeled as: \( X \sim f(x|\theta), \theta \in \mathbb{Z} \), where \( f(x|\theta) = 1/3 \) if \( x \in \{\theta - 1, \theta, \theta + 1\} \) and \( f(x|\theta) = 0 \) otherwise. Then, there does not exist an UMVU estimator for any non-constant \( \eta(\theta) \).

### 4.4 Bayes rules are biased

Interestingly, a (square-error loss) Bayes estimate is never unbiased! Because, if otherwise, \( \delta(x) \) was the Bayes rule w.r.t. \( \pi \) and satisfied \( \mathbb{E}\{\delta(X)|\theta\} = \eta(\theta) \) for all \( \theta \) and so

\[
\mathbb{E}\{\delta(X)\eta(\theta)\} = \begin{cases} 
\mathbb{E}[\mathbb{E}\{\delta(X)|\theta\}\eta(\theta)] = \mathbb{E}\{\eta^2(\theta)\} \\
\mathbb{E}[\delta(X)\mathbb{E}\{\eta(\theta)|X\}] = \mathbb{E}\{\delta^2(X)\}
\end{cases}
\]

and therefore

\[
\mathbb{E}\{\delta(X) - \eta(\theta)\}^2 = \mathbb{E}\{\delta^2(X)\} + \mathbb{E}\{\eta^2(\theta)\} - 2\mathbb{E}[\delta(X)\mathbb{E}\{\eta(\theta)|X\}] = 0,
\]

which cannot happen unless \( \eta(\theta) \equiv \text{const.} \).

### 5 Minimax and Bayes

Interestingly the most systematic way to find minimax rules comes through considerations of Bayes rules that have constant risk functions.

For any prior \( \pi \), its Bayes risk is defined as \( r_\pi = \inf_\delta r(\pi, \delta) \), the average risk associated with the corresponding Bayes rule. A prior \( \pi \) with maximum Bayes risk is called a least favorable prior. It turns out that Bayes rules of least favorable priors have close relations to minimaxity.

**Theorem 1.** Suppose a prior \( \pi \) with associated Bayes rule \( \delta_\pi \) satisfies

\[
r_\pi = \sup_{\theta \in \Theta} R(\theta, \delta_\pi)
\]

then \( \delta_\pi \) is minimax and \( \pi \) is least favorable.

The condition of the theorem is same as saying that the risk function \( R(\theta, \delta_\pi) \) must be constant over the support of \( \pi \).
Example (Binomial model.). Consider the model \( X \sim \text{Binomial}(n, p) \), \( p \in (0, 1) \) where we want to estimate \( p \) under the squared-error loss \( (\delta(X) - p)^2 \). For \( \pi = \text{Beta}(a, b) \), the Bayes estimate is \( \delta_{a,b}(X) = \frac{a + X}{a + b + n} \) whose risk function is given by

\[
R(p, \delta_{a,b}) = \frac{1}{(a + b + n)^2} \left\{ np(1 - p) + [a(1 - p) - bp]^2 \right\}.
\]

A straightforward calculation shows that for \( a = b = \sqrt{n}/2 \), the corresponding Bayes rule has a constant risk function. Hence \( \text{Beta}(\sqrt{n}/2, \sqrt{n}/2) \) is a least favorable prior and the corresponding Bayes rule \( \delta(X) = (X + \sqrt{n}/2)/(n + \sqrt{n}) \) is a minimax estimator. This is not a very interesting estimator (unless you really feel strongly that \( \delta \) corresponds Bayes rule has a constant risk function. Hence \( \pi \) should put most of its mass around \( p = 1/2 \)). For moderately large \( n \), this minimax rule has a higher risk than the conventional estimator \( X/n \) at most \( p \) outside of a tiny interval around 1/2.

If however we used the scaled squared-error loss \( (\delta(X) - p)^2 / \{p(1 - p)\} \) then \( \delta(X) = X/n \) is the Bayes rule associated with \( \pi = \text{Uniform}(0, 1) \) and has a constant risk function. So \( X/n \) is minimax under this loss.

For dealing with unbounded \( \Theta \), the above theorem could be relaxed to the following.

**Theorem 2.** Suppose \( \pi_n \) is a sequence of priors such that \( r := \lim_{n \to \infty} r_{\pi_n} \) is finite. If any rule \( \delta \) satisfies \( \sup_{\theta \in \Theta} R(\theta, \delta) = r \) then \( \delta \) is minimax and \( r \geq r_{\pi'} \) for every \( \pi' \).

Example (Normal mean.). Suppose \( X \sim N_p(\theta, \sigma^2 I_p) \), \( \theta \in \mathbb{R}^p \). Let \( \pi_n = N_p(0, nI_p) \). Then, under the squared-error loss \( L(\theta, a) = ||\theta - a||^2 \),

\[
\delta_{\pi_n}(X) = (1 + \sigma^2/n)^{-1}X, \quad r_{\pi_n} = p\sigma^2(1 + \sigma^2/n)^{-1}.
\]

So, \( r = p\sigma^2 \). Now \( \delta(X) = X \) has risk \( R(\theta, X) = p\sigma^2 = r \). So \( X \) is minimax! Of course, \( X \) is also UMVU estimate, the MLE and the Bayes estimate under the reference Bayes analysis. But below we will see that it’s not admissible when \( p \geq 3! \)

6 Normal means, James-Stein & shrinkage

Consider again the problem of estimating \( \theta \in \mathbb{R}^p \) from data \( X \sim N_p(\theta, \sigma^2 I_p) \) where \( \sigma \) is known. The Bayes estimator and Bayes risk under \( \pi = N_p(0, \tau^2 I_p) \) are:

\[
\delta^\tau = \frac{\tau^2}{\tau^2 + \sigma^2}X, \quad r^\tau = \frac{p\sigma^2\tau^2}{\tau^2 + \sigma^2}.
\]

Under this prior, the marginal pdf of \( X \) is \( N_p(0, (\sigma^2 + \tau^2)I_k) \), under which an unbiased estimator of \( \rho = \tau^2/(\tau^2 + \sigma^2) \) can be obtained as \( \hat{\rho} = 1 - \frac{(p-2)\sigma^2}{\|X\|^2} \). Plugging in this to \( \delta^\tau \) we get the famous James-Stein estimator:

\[
\delta^{JS}(X) = \left(1 - \frac{(p-2)\sigma^2}{\|X\|^2}\right)X
\]

with risk function

\[
R(\theta, \delta^{JS}) = p\sigma^2 - (p-2)^2\sigma^4E\left\{\frac{1}{\|X\|^2} | \theta \right\}
\]
which is uniformly smaller than the risk function of $X$, $R(\theta, X) = p\sigma^2$. That is, the UMVU, ML, reference-Bayes estimator $X$ is not admissible!

It turns out that the James-Stein estimator is itself inadmissible, and is dominated by the positive part estimator:

$$\delta^+(X) = \left(1 - \frac{(p - 2)\sigma^2}{\|X\|^2}\right)_+ X$$

where $z_+ = \max(z, 0)$. The James-Stein and the positive part estimators triggered great interest in shrinkage (the factor in front of $X$ shrinks the ML estimator $X$ toward zero when $\theta$ is close to zero) and empirical Bayes (plugging in an estimate of $\tau$ in the Bayes estimators).

7 Bayes & Admissibility

It is easy to see that all Bayes rules are essentially admissible. A more precise statement is as follows.

**Theorem 3.** Suppose $\delta_\pi$ is a Bayes estimator having finite Bayes risk with respect to a prior density $\pi$ which is positive for all $\theta \in \Theta$, and the risk function of every estimator $\delta$ is continuous in $\theta$. Then $\delta_\pi$ is admissible.

The result is easy to prove, but a crucial assumption is continuity of the risk function of every estimator $\delta$. This, however, is not a very strong condition. For example, for a canonical exponential family model $f(x|\theta) = h(x) \exp\{\theta^T T(x) - A(\theta)\}$ and any loss function $L(\theta, \delta)$ for which $R(\eta, \delta)$ is finite, $(\theta, \delta)$ is also continuous in $\theta$.

We conclude this discussion with a partial converse of the above result.

**Theorem 4.** Suppose $f(x|\theta) > 0$ for every $x \in S, \theta \in \Theta$. Let $L(\theta, a)$ be continuous and strictly convex in $a$ for every $\theta$ and satisfy,

$$\lim_{|a| \to \infty} L(\theta, a) = \infty, \quad \text{for all } \theta \in \Theta.$$

Then every admissible rule $\delta(X)$ is a generalized Bayes rule, i.e., there is a sequence $\pi_n$ of prior distributions with support on a finite subset of $\Theta$ such that $\delta_{\pi_n}(X) \to \delta(X)$ [almost surely].

The two theorems together say that under some regularity conditions on the model and the loss function, every Bayes rule is admissible and every admissible rule is generalized-Bayes.