Outline

Finding Bayes Rules[1]

Theorem: For each $x \in \Omega$ and $a \in \mathcal{A}$, define

$$r(x, a) = \int_{\Theta} L(\theta, a)\pi(\theta|x)d\theta \tag{1}$$

For each $x \in \mathcal{X}$, suppose that there exists an $a_x \in \mathcal{A}$ such that

$$r(x, a_x) = \inf_{a \in \mathcal{A}} r(x, a) \tag{2}$$

Let δ^π be a function from $\mathcal{X}$ into $\mathcal{A}$ defined by $\delta^\pi(x) = a_x$.

If $\delta^\pi \in \mathcal{D}$, then δ^π is the Bayes rule with respect to π.

In Casella and Berger's own words:

> *The theorem tells us exactly what the Bayes rule should do for each $x \in \mathcal{X}$. In fact, having observed $X = x$, we need to solve (1) ONLY for this particular x.*

[1] Casella and Berger, Chapter 10, Section 10.3.2

Proof of the theorem

$$
\begin{aligned}
B(\pi,\delta) &= \int_\Theta R(\theta,\delta)\pi(\theta)d\theta = \int_\Theta \left[\int_{\mathcal{X}} L(\theta,\delta(x))f(x|\theta)dx\right]\pi(\theta)d\theta \\
&= \int_\Theta \int_{\mathcal{X}} L(\theta,\delta(x))f(x|\theta)\pi(\theta)dxd\theta \\
&= \int_\Theta \int_{\mathcal{X}} L(\theta,\delta(x))\pi(\theta|x)m(x)dxd\theta \\
&= \int_{\mathcal{X}} \int_\Theta L(\theta,\delta(x))\pi(\theta|x)m(x)d\theta dx \\
&= \int_{\mathcal{X}} \left[\int_\Theta L(\theta,\delta(x))\pi(\theta|x)d\theta\right] m(x)dx = \int_{\mathcal{X}} r(x,\delta(x))m(x)dx
\end{aligned}
$$

But by (2) and the definition of $\delta^\pi(x)$, for every $x \in \mathcal{X}$, $r(x,\delta^\pi(x)) = r(x,a_x)$ is the smallest possible value.

Thus δ^π minimizes the last integral and, hence, the Bayes risk.

Finding Admissible Rules[2]

Theorem: Consider a decision problem in which the parameter space Θ is a subset of the real line.

Suppose that for every decision rule $\delta \in \mathcal{D}$, the risk function $R(\theta, \delta)$ is a continuous function of θ.

Let $\pi(\theta)$ be a prior distribution on θ with the property that for any $\epsilon > 0$ and any $\theta \in \Theta$, the interval $(\theta - \epsilon, \theta + \epsilon)$ has positive probability under π.

Let δ^π be a Bayes rule with respect to π.

If $-\infty < B(\pi, \delta^\pi) < \infty$, then δ^π is an admissible decision rule.

Although not true in every instance, the general idea is that Bayes rules are admissible and, hence, Bayes rules are reasonable rules to consider.

Proof of the theorem

Suppose that δ^π is inadmissible. Then there exists a rule $\delta \in \mathcal{D}$ such that $R(\theta, \delta) \leq R(\theta, \delta^\pi)$ for all $\theta \in \Theta$ and for some θ, say θ' $R(\theta', \delta) \leq R(\theta', \delta^\pi)$.

Let $R(\theta', \delta^\pi) - R(\theta', \delta) = \nu > 0$. Since $R(\theta, \delta^\pi)$ and $R(\theta, \delta)$ are both continuous, so is $R(\theta, \delta^\pi) - R(\theta, \delta)$. Thus there exists an $\epsilon > 0$ such that

$$R(\theta, \delta^\pi) - R(\theta, \delta) > \frac{\nu}{2} \text{for all } \theta \in (\theta' - \epsilon, \theta' + \epsilon)$$

Since $-\infty < B(\pi, \delta^\pi) < \infty$, the following expression is well defined (not of the form $\infty - \infty$):

$$\begin{aligned} B(\pi, \delta^\pi) - B(\pi, \delta) &= \int_{-\infty}^{\infty} [R(\theta, \delta^\pi) - R(\theta, \delta)]\pi(\theta)d\theta \\ &\geq \int_{-\theta'-\epsilon}^{\theta'+\epsilon} [R(\theta, \delta^\pi) - R(\theta, \delta)]\pi(\theta)d\theta \geq \frac{\nu}{2} \int_{-\theta'-\epsilon}^{\theta'+\epsilon} \pi(\theta)d\theta > 0 \end{aligned}$$

This strict inequality contradicts the fact that δ^π is Bayes with respect to π. Hence δ^π is admissible.

Admissibility is not really a positive property, but rather the absence of a negative property.

An admissible estimator is not necessarily uniformly good, but it is not uniformly bad!

Knowing that a decision rule is admissible does not mean that this decision rule is obviously the rule to use since, in most cases, there are many admissible decision rules.

Some of these rules will be reasonable, some will be difficult to find or compute, and some may not be intuitively appealing.

Example

$X \sim binomial(n, p)$ with n known. Under the absolute error loss

$$\delta(x) = \frac{1}{3} \quad x = 0, 1, \ldots, n$$

is an admissible estimator since $R(\frac{1}{3}, \delta) = 0$.

Bias

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from $p(x|\boldsymbol{\theta})$ and $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{X})$ an estimator of $\boldsymbol{h}(\boldsymbol{\theta})$, for any given function $\boldsymbol{h}$.

$\boldsymbol{\delta}$ is an unbiased estimator of $\boldsymbol{h}(\boldsymbol{\theta})$ if $E[\boldsymbol{\delta}|\boldsymbol{\theta}] = \boldsymbol{h}(\boldsymbol{\theta}), \ \forall \boldsymbol{\theta}$.

The estimator $\boldsymbol{\delta}$ is said to be biased otherwise. In this case, the bias is denoted by $\boldsymbol{b}(\boldsymbol{\theta})$ and defined as

$$\boldsymbol{b}(\boldsymbol{\theta}) = E[\boldsymbol{\delta}|\boldsymbol{\theta}] - \boldsymbol{h}(\boldsymbol{\theta}).$$

Frequentist interpretation: After repeating sampling of $\boldsymbol{X}$ from $p(x|\boldsymbol{\theta})$ many times, averaging the corresponding values of $\boldsymbol{\delta}$ will produce $\boldsymbol{h}(\boldsymbol{\theta})$ as a result.

This is a desirable property because one formulates an estimator $\boldsymbol{\delta}$ in an effort to obtain the value of $\boldsymbol{h}(\boldsymbol{\theta})$.

Difficulties

In most cases only a single sample $\boldsymbol{X}$ is observed for time and/or financial restrictions.

Note also that unbiased estimation is always related to a given parametric function; an estimator can be biased with respect to a given function but unbiased with respect to another one.

Risk and Mean Square Error

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from $p(x|\boldsymbol{\theta})$ and denote now by $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{X})$, an estimator of $\boldsymbol{h}(\boldsymbol{\theta})$.

The frequentist risk of the estimator $\boldsymbol{\delta}$ is defined as

$$R_{\boldsymbol{\delta}}(\boldsymbol{\theta}) = E_{\boldsymbol{X}|\boldsymbol{\theta}}[L(\boldsymbol{\delta}(\boldsymbol{X}), \boldsymbol{\theta})]$$

In the case of a quadratic loss function L, the risk is given by

$$R_{\boldsymbol{\delta}}(\boldsymbol{\theta}) = E_{\boldsymbol{X}|\boldsymbol{\theta}}[(\boldsymbol{\delta} - \boldsymbol{h}(\boldsymbol{\theta}))'(\boldsymbol{\delta} - \boldsymbol{h}(\boldsymbol{\theta}))]$$

and is also called quadratic mean squared error (MSE, in short).

In the scalar case, the quadratic MSE reduces to

$$E_{\boldsymbol{X}|\theta}[\delta - h(\theta)]^2.$$

Best unbiased estimator

An estimator W^* is a best unbiased estimator of $\tau(\theta)$ if it satisfies

$$E_\theta W^* = \tau(\theta)$$

for all θ and, for any other estimator W with $E_\theta W = \tau(\theta)$, we have

$$Var_\theta W^* \leq Var_\theta W$$

for all θ.

W^* is also called a uniform minimum variance unbiased estimator(UMVUE) of $\tau(\theta)$.

Theorem 7.3.3 (C&B): If W is a best unbiased estimator of $\tau(\theta)$, then W is unique.

Example

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the $N(\theta, \sigma^2)$ distribution with σ^2 known and $h(\theta) = \theta$.

Taking $\delta_1(\boldsymbol{X}) = \overline{X}$ and $\delta_2(\boldsymbol{X}) = X_1$ gives

$$E[\delta_1(\boldsymbol{X})|\theta] = E[\delta_2(\boldsymbol{X})|\theta] = \theta$$

Also,

$$\begin{aligned} R_{\delta_1}(\theta) &= V(\overline{X}|\theta) = \frac{\sigma^2}{n} \\ R_{\delta_2}(\theta) &= V(X_1) = \sigma^2 \end{aligned}$$

For $n > 1$, $R(\delta_1) < R(\delta_2)$ for all values of θ.

Rao-Blackwell Theorem

When

- $\boldsymbol{X} = (X_1, \ldots, X_n)$ iid from $p(x|\boldsymbol{\theta})$
- $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{X})$ an unbiased estimator of $\boldsymbol{h}(\boldsymbol{\theta})$, for some function $\boldsymbol{h}$
- $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ a sufficient statistic for $\boldsymbol{\theta}$

then,

$$\boldsymbol{\delta}^* = \boldsymbol{\delta}^*(\boldsymbol{X}) = E(\boldsymbol{\delta}|\mathbf{T})$$

is an unbiased estimator of $\boldsymbol{h}(\boldsymbol{\theta})$ with

$$\mathbf{V}(\boldsymbol{\delta}^*|\boldsymbol{\theta}) \leq \mathbf{V}(\boldsymbol{\delta}|\boldsymbol{\theta})$$

for all $\boldsymbol{\theta}$.

Proof

Initially note that $\boldsymbol{\delta}^*$ is unbiased because

$$\begin{aligned} E[\delta^*(\boldsymbol{X})|\boldsymbol{\theta}] &= E\{E[\boldsymbol{\delta}(\boldsymbol{X})|\boldsymbol{T}(\boldsymbol{X})]\,|\boldsymbol{\theta}] \\ &= E[\boldsymbol{\delta}(\boldsymbol{X})\mid\boldsymbol{\theta}] = \boldsymbol{h}(\boldsymbol{\theta}). \end{aligned}$$

Finally note that

$$\mathbf{V}(\boldsymbol{\delta}^*|\boldsymbol{\theta}) = E[\,\mathbf{V}(\boldsymbol{\delta}|T)\,|\theta] + \mathbf{V}(\boldsymbol{\delta}^*|\boldsymbol{\theta})$$

As $\mathbf{V}(\boldsymbol{\delta}|\mathbf{T}) \geq 0$, its expectation is also non-negative positive and therefore

$$\mathbf{V}(\boldsymbol{\delta}|\boldsymbol{\theta}) \geq \mathbf{V}(\boldsymbol{\delta}^*\,|\boldsymbol{\theta})$$

□

Example 7.3.7 (C&B)

X_1, X_2 iid $N(\theta, 1)$. Then $\bar{X}$ is such that

$$E_\theta(\bar{X}) = \theta \quad \text{and} \quad V_\theta(\bar{X}) = 0.5$$

Now, let $\phi(X_1) = E_\theta(\bar{X}|X_1)$. Then

$$E_\theta(\phi(X_1)) = \theta \quad \text{and} \quad V_\theta(\phi(X_1)) \leq V_\theta(\bar{x})$$

However, $\phi(X_1) = 0.5(X_1 + \theta)$ is not an estimator!

Complete families of distributions

Estimators have their risks reduced if they are functions of sufficient statistics.

Maximal improvement in terms of risk is achieved if minimal sufficient statistics are used.

A related interesting question is to know if the reduction in risk was the smallest possible.

The search for maximal reduction is helped in a sense by the concept of complete families of distributions.

Definition Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from $p(x|\boldsymbol{\theta})$ and $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ any statistic. The family of distributions of $\boldsymbol{T}$ is complete if $\forall \boldsymbol{\theta}$,

$$E(\boldsymbol{g}(\boldsymbol{T})|\boldsymbol{\theta}) = \mathbf{0} \Rightarrow \boldsymbol{g}(\boldsymbol{T}) = \mathbf{0}$$

with probability 1.

Theorem

Let $X_1, \ldots, X_n$ be iid observations from an exponential family with pdf of the form

$$f(x|\theta) = h(x)c(\theta)\exp\{\omega(\theta)t(x)\}$$

Then the statistic

$$T(X) = \sum_{i=1}^{n} t(X_i)$$

is complete.

Cramer-Rao Theorem

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from $p(X|\boldsymbol{\theta})$ and $\boldsymbol{\delta}$ an unbiased estimator of $\boldsymbol{h}(\boldsymbol{\theta})$, for some function $\boldsymbol{h}$. Assume further that

- $\{\boldsymbol{X} : p(\boldsymbol{X}|\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$
- $\partial p(\boldsymbol{x}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\partial \boldsymbol{h}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ exist,
- $E(\boldsymbol{\delta}|\boldsymbol{\theta})$ is differentiable inside the integral
- the Fisher information $\boldsymbol{I}(\boldsymbol{\theta})$ is finite.

Then

$$\mathbf{V}(\boldsymbol{\delta}|\boldsymbol{\theta}) \geq \frac{\partial \boldsymbol{h}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} [\boldsymbol{I}(\boldsymbol{\theta})]^{-1} \left(\frac{\partial \boldsymbol{h}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)'.$$

In the case of a scalar θ, the inequality reduces to

$$V[\delta|\theta] \geq \frac{[h'(\theta)]^2}{I(\theta)}$$

Proof (of the scalar case)

Let

$$h(\theta) = \int \delta(\boldsymbol{x}) p(\boldsymbol{x} \mid \theta) d\boldsymbol{x}.$$

Differentiating both sides with respect to θ gives

$$\begin{aligned}
\frac{\partial h(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} \int \delta(\boldsymbol{x}) p(\boldsymbol{x}|\theta)\, d\boldsymbol{x} = \int \delta(\boldsymbol{x}) \frac{\partial p(\boldsymbol{x}|\theta)}{\partial \theta}\, d\boldsymbol{x} \\
&= \int \delta(\boldsymbol{x}) \frac{1}{p(\boldsymbol{x} \mid \theta)} \frac{\partial p(\boldsymbol{x} \mid \theta)}{\partial \theta} p(\boldsymbol{x} \mid \theta)\, d\boldsymbol{x} \\
&= E\left[\left(\delta(\boldsymbol{X}) \frac{\partial log\, p(\boldsymbol{X}|\theta)}{\partial \theta}\right) |\theta\right] = E[\delta(\boldsymbol{X})\, U(\boldsymbol{X}; \theta)|\theta].
\end{aligned}$$

As previously seen, $E[U(\boldsymbol{X}; \theta)] = 0$ and

$$\begin{aligned}
\frac{\partial h(\theta)}{\partial \theta} &= E\left[(\delta(\boldsymbol{X}) - h(\theta))\, U(\boldsymbol{X}; \theta)|\theta\right] \\
&= Cov\left[(\delta(\boldsymbol{X}),\, U(\boldsymbol{X}; \theta))\, |\theta\right]
\end{aligned}$$

Since $\rho^2(X, Y) \leq 1 \Rightarrow C^2(X, Y) \leq V(X)V(Y)$, it follows that

$$\left(\frac{\partial h(\theta)}{\partial \theta}\right)^2 \leq V[\delta(\boldsymbol{X}) \mid \theta]\, V\left[U(\boldsymbol{X}; \theta)|\theta\right].$$

But

$$V\left[U(\boldsymbol{X}; \theta)|\theta\right] = E\left[U^2(\boldsymbol{X}; \theta)|\theta\right] = I(\theta),$$

completing the proof. □

The unbiased estimator attains the lower bound when it has maximal correlation with the score function. In other words, when there are functions $\mathbf{c}$ and $\mathbf{d}$ of $\boldsymbol{\theta}$ such that

$$\boldsymbol{\delta}(\boldsymbol{X}) = \mathbf{c}(\boldsymbol{\theta})\,\mathbf{U}(\boldsymbol{X};\boldsymbol{\theta}) + \mathbf{d}(\boldsymbol{\theta}),$$

with probability 1.

Taking expectation of both sides with respect to $\boldsymbol{X}|\boldsymbol{\theta}$ gives that $\boldsymbol{\delta}(\boldsymbol{X})$ is an unbiased estimator of $\mathbf{d}(\boldsymbol{\theta})$ and therefore $\mathbf{d} = \boldsymbol{h}$.

When the MLE is unbiased, it attains the Cramer-Rao lower bound. This can be seen by solving the above equation for $\boldsymbol{\theta}$.

$$\mathbf{U}(\boldsymbol{X};\boldsymbol{\theta}) = \frac{\partial \log\, p(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\boldsymbol{\delta}(\boldsymbol{X}) - \mathbf{d}(\boldsymbol{\theta})}{\mathbf{c}(\boldsymbol{\theta})}.$$

Equating to 0 implies that $\boldsymbol{\delta}(\boldsymbol{X})$ is the MLE of $\mathbf{d}(\boldsymbol{\theta})$.

But we have already seen that $\mathbf{d} = \boldsymbol{h}$ and, by hypothesis, $\boldsymbol{\delta}$ is unbiased for $\boldsymbol{h}(\boldsymbol{\theta})$.

Hence, it attains the Cramer-Rao lower bound.

Efficiency

The estimator $\boldsymbol{\delta}$ of $\boldsymbol{h}(\boldsymbol{\theta})$ is said to be efficient if it is unbiased and its variance attains the Cramer-Rao lower bound, $\forall \boldsymbol{\theta}$.

The efficiency of an unbiased scalar estimator is given by the ratio between the Cramer-Rao bound and its variance.

There is no guarantee that UMVU estimators will attain the Cramer-Rao bound.

Efficient estimators are necessarily UMVU.

The Cramér-Rao lower bound may be strictly smaller than the variance of any unbiased estimator

Example 7.3.6 (C&B)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the $Pois(\theta)$ distribution. Then

$$\log p(\boldsymbol{x} \mid \theta) = -n\theta + \sum_{i=1}^{n} x_i \, log\theta - \sum_{i=1}^{n} \log x_i!$$

and therefore $U(\boldsymbol{X}; \theta) = -n + \sum_{i=1}^{n} X_i/\theta$

As estimators cannot possibly depend on the parameter they are supposed to estimate, define $c(\theta) = \theta/n$ and $d(\theta) = \theta$.

$\bar{X}$ is an efficient estimator of its mean θ.

Any linear function of $\bar{X}$ is an efficient estimator of the respective linear function of θ.

More than that, these are the unique efficient estimators that can be found in the presence of a random sample from the $Pois(\theta)$ distribution.