Principled Bayesian Hypothesis Testing

Sudipto Banerjee

Division of Biostatistics
School of Public Health
University of Minnesota

April 29, 2008
Statistical decision theory lends a mathematically formal approach to study hypothesis testing and credible intervals.
Statistical decision theory lends a mathematically formal approach to study hypothesis testing and credible intervals.

This approach often tends to divorce the methods from real-life connotations.
• Statistical decision theory lends a mathematically formal approach to study hypothesis testing and credible intervals.

• This approach often tends to divorce the methods from real-life connotations.

• But it provides a unified conceptual framework for formulating principled solutions to very diverse problems.
• Statistical decision theory lends a mathematically formal approach to study hypothesis testing and credible intervals.

• This approach often tends to divorce the methods from real-life connotations.

• But it provides a unified conceptual framework for formulating principled solutions to very diverse problems.

• A decision problem has the following ingredients:
  • Observed data \( y \) following a probability model \( f(y | \theta) \), where the parameter lies in some set \( \Theta \).
- Statistical decision theory lends a mathematically formal approach to study hypothesis testing and credible intervals.

- This approach often tends to divorce the methods from real-life connotations.

- But it provides a unified conceptual framework for formulating principled solutions to very diverse problems.

- A decision problem has the following ingredients:
  - Observed data $y$ following a probability model $f(y | \theta)$, where the parameter lies in some set $\Theta$.
  - A space $\mathcal{A}$ of actions, whose elements we denote by $a$ (often results from a decision function $d(y)$).
Elements of Bayesian Decision Theory

- Statistical decision theory lends a mathematically formal approach to study hypothesis testing and credible intervals.

- This approach often tends to divorce the methods from real-life connotations.

- But it provides a unified conceptual framework for formulating principled solutions to very diverse problems.

A decision problem has the following ingredients:

- Observed data $y$ following a probability model $f(y | \theta)$, where the parameter lies in some set $\Theta$.

- A space $\mathcal{A}$ of actions, whose elements we denote by $a$ (often results from a decision function $d(y)$).

- A Loss function $L(\theta, a) : \Theta \times \mathcal{A} \rightarrow \mathbb{R}$, which quantifies the loss incurred when the parameter is $\theta$ and action (decision) $a$ is taken.
Both estimation and testing are special cases of statistical decision theory.
Both estimation and testing are special cases of statistical decision theory.

For estimating a real-valued function $\tau(\theta)$, we could take:

\[ A = \mathbb{R} \]

For testing $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$:

\[ A = \{a_0, a_1\} \]

where $a_j$ is the decision to accept $H_j$.

\[ L(\theta, a_j) = 0 \text{ if } \theta \text{ satisfies } H_j \text{ and } L(\theta, a_j) = 1 \text{ otherwise.} \]

\[ d(y) = a_1 \text{ if } y \text{ falls in the critical region and equals } a_0 \text{ otherwise.} \]
Both estimation and testing are special cases of statistical decision theory.

For estimating a real-valued function $\tau(\theta)$, we could take:

- $\mathcal{A} = \mathbb{R}$
- $L(\theta, a) = (a - \tau(\theta))^2$. 

The decision function $d(y)$ will be the estimate of $\tau(\theta)$. 

For testing $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$:

- $A = \{a_0, a_1\}$ where $a_j$ is the decision to accept $H_j$.
- $L(\theta, a_j) = 0$ if $\theta$ satisfies $H_j$ and $L(\theta, a_j) = 1$ otherwise.

$d(y) = a_1$ if $y$ falls in the critical region and equals $a_0$ otherwise.
Both estimation and testing are special cases of statistical decision theory.

For estimating a real-valued function \( \tau(\theta) \), we could take:

- \( \mathcal{A} = \mathbb{R} \)
- \( L(\theta, a) = (a - \tau(\theta))^2 \).

The decision function \( d(y) \) will be the estimate of \( \tau(\theta) \).

For testing \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta \neq \theta_0 \):

- \( \mathcal{A} = \{a_0, a_1\} \) where \( a_j \) is the decision to accept \( H_j \).
Both estimation and testing are special cases of statistical decision theory.

For estimating a real-valued function $\tau(\theta)$, we could take:

- $\mathcal{A} = \mathbb{R}$
- $L(\theta, a) = (a - \tau(\theta))^2$.
- The decision function $d(y)$ will be the estimate of $\tau(\theta)$.

For testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$:

- $\mathcal{A} = \{a_0, a_1\}$ where $a_j$ is the decision to accept $H_j$.
- $L(\theta, a_j) = 0$ if $\theta$ satisfies $H_j$ and $L(\theta, a_j) = 1$ otherwise.
Both estimation and testing are special cases of statistical decision theory.

For estimating a real-valued function \( \tau(\theta) \), we could take:

- \( \mathcal{A} = \mathbb{R} \)
- \( L(\theta, a) = (a - \tau(\theta))^2 \).
- The decision function \( d(y) \) will be the estimate of \( \tau(\theta) \).

For testing \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta \neq \theta_0 \):

- \( \mathcal{A} = \{a_0, a_1\} \) where \( a_j \) is the decision to accept \( H_j \).
- \( L(\theta, a_j) = 0 \) if \( \theta \) satisfies \( H_j \) and \( L(\theta, a_j) = 1 \) otherwise.
- \( d(y) = a_1 \) if \( y \) falls in the critical region and equals \( a_0 \) otherwise.
To make optimal decisions, we introduce a **risk function**: 

\[ R(\theta, d) = E_y[\theta[L(\theta, \delta(y))]]. \]

A Bayesian further accounts for uncertainty in \( \theta \). Let \( \pi(\theta) \) be the prior. Then the **pre-posterior risk** is defined as:

\[ \int_{\Theta} R(\theta, d)\pi(\theta)d\theta = R(\pi, d). \]

After observing \( y \), the relevant distribution of \( \theta \) is given by the posterior \( p(\theta | y) \). This leads to the **posterior risk**:

\[ \int_{\Theta} L(\theta, a)p(\theta | y)d\theta = \psi(y, a). \]

In principle, there are two Bayesian decision problems:

- At the planning stage, choose an optimal \( d(y) \), say \( d_\pi \), to minimize \( d_\pi \). This leads to Bayesian designs.
- Given the data, choose \( a \) to minimize \( \psi(y, a) \).
For any $d$, we have $R(\pi, d) = E_y[\psi(y), d(y)]$.

Proof:

$$R(\pi, d) = E_{\theta}[E_y | \theta[R(\theta, d)]]$$

$$= E_y[E_{\theta | y}[R(\theta, d(y))]] = E_y[\psi(y, d(y))] .$$

Suppose $d_0 = a(y)$ minimizes $\psi(y, a)$. That is,

$$\psi(y, d_0) = \inf_{a \in \mathcal{A}} \psi(y, a).$$

Then $d_0$ minimizes $R(\pi, d)$. Proof:

$$R(\pi, d_0) = E_y[\psi(y, a(y))]$$

$$\leq \sup_d E[\psi(y, d(y))] = R(\pi, d).$$

Hence, $R(\pi, d_0) = \inf_d R(\pi, d)$. 
Consider the following loss function for hypothesis tests:

<table>
<thead>
<tr>
<th></th>
<th>( H_0 ) is TRUE</th>
<th>( H_1 ) is TRUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do not Reject ( H_0 )</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reject ( H_0 )</td>
<td>( K )</td>
<td>0</td>
</tr>
</tbody>
</table>

Suppose we wish to test \( H_0 : \theta = \theta_0 \) versus \( H_1 : \theta = \theta_1 \), where \( \theta_1 > \theta_0 \), for data \( y_1, \ldots, y_n \sim N(\theta, \sigma^2) \); \( \sigma^2 \) is known.

Assume apriori that \( P(\theta_0) = 1 - P(\theta_1) = \pi \).

A Bayesian decision between \( H_0 \) and \( H_1 \) will be based upon their posterior probabilities. The null hypothesis is *not rejected* when

\[
\frac{P(H_0 \mid y)}{P(H_1 \mid y)} > \frac{1}{K} \implies P(H_0 \mid y) \geq \frac{1}{1 + K}.
\]

This rule *minimizes* the posterior expected loss computed from the above loss function.
HOMWORK: Prove that this rule implies that the null hypothesis is not rejected if:

\[ \bar{y} \leq \frac{\sigma^2 \log \left( K \frac{\pi}{1-\pi} \right)}{n\delta} + \frac{\theta_1 + \theta_0}{2}, \]

where \( \bar{y} \) is the sample mean and \( \delta = \theta_1 - \theta_0 \).
A Bayesian “utility function” is given by:

\[
U(y, n, \delta, \pi, K) = KP(H_0)P(\text{correct decision} \mid H_0) + P(H_1)P(\text{correct decision} \mid H_1)
\]

\[
= K\pi P_{\theta_0} \left( \bar{y} \leq \frac{\sigma^2 \log \left( K \frac{\pi}{1-\pi} \right)}{n\delta} + \frac{\theta_1 + \theta_0}{2} \right)
\]

\[
+ (1 - \pi) P_{\theta_1} \left( \bar{y} > \frac{\sigma^2 \log \left( K \frac{\pi}{1-\pi} \right)}{n\delta} + \frac{\theta_1 + \theta_0}{2} \right)
\]

\[
= K\pi \Phi \left( \frac{\sigma \log \left( K \frac{\pi}{1-\pi} \right)}{\sqrt{n}\delta} + \frac{\delta \sqrt{n}}{2\sigma} \right)
\]

\[
+ (1 - \pi) \left[ 1 - \Phi \left( \frac{\sigma \log \left( K \frac{\pi}{1-\pi} \right)}{\sqrt{n}\delta} - \frac{\delta \sqrt{n}}{2\sigma} \right) \right].
\]
Let us revisit the sample size problem. A Bayesian finds the sample size to ensure a minimum rate $r$ of correct classification.

Let $\sigma^2 = 1$ and $\delta = 0.10$.

Assuming a Type-I error of $\alpha = 0.05$ and a power of 0.90, the frequentist obtains $n = 857$.

A Bayesian, assuming $\pi = 0.5$, $K = 1$ (a $0 - 1$ loss function) and $r = 0.9283$ also obtains $n = 857$.

Computationally, the Bayesian solves $U(y, n, \delta, \pi, K) = r$. Easiest to feed in different values of $n$ until the utility exceeds $r$. 

---

**Sample size problem**

- Let us revisit the sample size problem. A Bayesian finds the sample size to ensure a minimum rate $r$ of correct classification.

- Let $\sigma^2 = 1$ and $\delta = 0.10$.

- Assuming a Type-I error of $\alpha = 0.05$ and a power of 0.90, the frequentist obtains $n = 857$.

- A Bayesian, assuming $\pi = 0.5$, $K = 1$ (a $0 - 1$ loss function) and $r = 0.9283$ also obtains $n = 857$.

- Computationally, the Bayesian solves $U(y, n, \delta, \pi, K) = r$. Easiest to feed in different values of $n$ until the utility exceeds $r$. 

---

**9**
Consider a blood test conducted for determining the sugar level of a person with diabetes two hours after he had his breakfast. Doctors want to see if his medication has controlled his blood sugar levels. Assume that the blood-test result $Y \sim N(\theta, 100)$. In the appropriate population (diabetic, but under the same treatment), $\theta \sim N(100, 900)$. Then, the posterior distribution $p(\theta \mid Y = y)$ is:

$$
N \left( \theta \mid \frac{900}{1000}y + \frac{100}{1000}100, \frac{100 \times 900}{1000} = 90 \right).
$$
Suppose that the observed blood test shows \( y = 130 \).
Then the posterior is \( N(127, 90) \). Consequently, we have:

\[
P(\theta \leq 130 \mid Y = 130) = \Phi \left( \frac{130 - 127}{\sqrt{90}} \right) = \Phi(0.316) = 0.624.
\]

Thus the posterior odds ratio is \( 0.624/(1 - 0.624) = 1.66 \).

The prior odds: \( \Phi(1)/(1 - \Phi(1)) = 0.8413/0.1587 = 5.3 \).

The Bayes Factor is \( 1.66/5.3 = 0.313 \).
For testing \( H_0 : \theta = \theta_0 \) versus \( H_1 : \theta \neq \theta_0 \), we have

\[
m(y) = \pi_0 f(y \mid \theta_0) + (1 - \pi_0)m_1(y),
\]

where

\[
m_1(y) = \int_{\theta \neq \theta_0} f(y \mid \theta)\pi_1(\theta)\,d\theta.
\]

Therefore,

\[
\pi(\theta_0 \mid y) = \frac{\pi_0 f(y \mid \theta_0)}{m(y)} = \left\{ 1 + \frac{1 - \pi_0}{\pi_0} \frac{m_1(y)}{f(y \mid \theta_0)} \right\}^{-1}.
\]

It then follows that the Bayes Factor is \( f(y \mid \theta_0) / m_1(y) \).