Hypothesis Testing
Appendix

James J. Heckman
University of Chicago

Econ 312
This draft, April 17, 2006
A  Further results on Bayesian vs. Classical Testing

(The phenomenon shows up even when we don’t have \( T \to \infty \).)

Point mass placed at \( \theta_0 \); \( \pi_0 \) For rest of parameter space, we have

\[
1 - \pi_0 = \pi_1; \quad g_1 \text{ is density of } \theta
\]

\( f (X \mid \theta) \) is model.
Marginal density of $X =$

$$m (X) = \int f (X | \theta) df (\theta)$$

$$= f (X | \theta_0) \pi_0 + (1 - \pi_0) m_1 (X)$$

$$m_1 (X) = \int f (X | \theta) g_1 (\theta) d\theta$$
Posterior probability that $\theta = \theta_0$ is

$$\pi (\theta_0 \mid X) = \frac{f (X \mid \theta_0) \pi_0}{m (X)}$$

$$= \frac{f (X \mid \theta_0) \pi_0}{f (X \mid \theta_0) \pi_0 + (1 - \pi_0) m_1 (X)}$$

$$= \frac{1}{1 + \frac{(1 - \pi_0)}{\pi_0} \frac{m_1 (X)}{f (X \mid \theta_0)}}$$
The posterior probability that for $H_0$

$$\text{Posterior Odds Ratio} = \frac{1}{\left(\frac{1 - \pi_0}{\pi_0}\right) \frac{m_1(X)}{f(X | \theta_0)}} = \left(\frac{\pi_0}{1 - \pi_0}\right) \left(\frac{f(X | \theta_0)}{m_1(X)}\right)$$

Bayes factor: $\frac{f(X | \theta_0)}{m_1(X)}$
Example: Let \( X \sim N(\theta, \sigma^2) \); \( \sigma^2 \) known.

\[
\mathcal{L} \propto N(\bar{X}, \sigma^2/T) = \frac{1}{\sqrt{2\pi\sigma^2/T}} \exp \left[-\frac{T}{2\sigma^2} (\theta - \bar{X})^2 \right]
\]

Let \( g_1 = N(\mu_1, \tau^2) \) be prior; then \( m_1 = N(\mu, \tau^2 + \sigma^2/T) \)

\[
\pi(\theta_0 \mid X) = \left[1 + \frac{1 - \pi_0}{\pi_0} \cdot \frac{\left[2\pi \left(\tau^2 + \frac{\sigma^2}{T}\right)\right]^{-\frac{1}{2}} \exp \left[-\frac{\left(\bar{X} - \mu\right)^2}{2\left(\tau^2 + \frac{\sigma^2}{T}\right)}\right]}{\left(2\pi \frac{\sigma^2}{T}\right)^{-\frac{1}{2}} \exp \left[-\frac{\left(\bar{X} - \theta_0\right)^2}{2\left(\frac{\sigma^2}{T}\right)}\right]} \right]^{-1}
\]

center at \( \mu = \theta_0 \). (This is a judgment about priors.)
\[
\pi(\theta_0 \mid X) = \left[ 1 + \frac{(1 - \pi_0)}{\pi_0} \left( \exp \left[ -\frac{t^2}{2[1+\sigma^2/(T\tau^2)]} \right] \right)^{-1} \right]^{-1}
\]

As \( T \to \infty \) we get “Lindley Paradox”

\[
t = \frac{|\bar{X} - \theta_0|}{\left( \frac{\sigma}{\sqrt{T}} \right)}
\]

usual normal statistic (2 tail test).
Look at the following table of values:
\[ \pi (\theta_0 \mid X), \mu = \theta_0, \pi_0 = 1/2, T = \sigma. \]

<table>
<thead>
<tr>
<th>( t ) score</th>
<th>( p ) value</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.645</td>
<td>.1</td>
<td>.42</td>
<td>.44</td>
<td>.49</td>
<td>.56</td>
<td>.65</td>
<td>.72</td>
<td>.89</td>
</tr>
<tr>
<td>1.960</td>
<td>.05</td>
<td>.35</td>
<td>.33</td>
<td>.37</td>
<td>.42</td>
<td>.52</td>
<td>.60</td>
<td>.80</td>
</tr>
<tr>
<td>2.576</td>
<td>.01</td>
<td>.21</td>
<td>.13</td>
<td>.14</td>
<td>.16</td>
<td>.22</td>
<td>.27</td>
<td>.53</td>
</tr>
<tr>
<td>3.291</td>
<td>.001</td>
<td>.086</td>
<td>.026</td>
<td>.024</td>
<td>.026</td>
<td>.034</td>
<td>.045</td>
<td>.124</td>
</tr>
</tbody>
</table>

Observe: Not monotone in \( T \):
Now, consider the minimum of $\pi (\theta_0 \mid X)$ over all $\tau$; and $g_1 = \text{(marginal on alternative)}$

Thm (Dickey and Savage):

$$\pi (\theta_0 \mid X) \geq \left[ 1 + \frac{(1 - \pi_0)}{\pi_0} \frac{r(X)}{f(X \mid \theta_0)} \right]^{-1}$$

where

$$r(X) = \sup_{\theta \neq \theta_0} f(X \mid \theta)$$

usually obtained by substituting in maximum likelihood estimate.
The bound on the Bayes factor is

\[ B = \frac{f(X | \theta_0)}{m_1(X)} \geq \frac{f(X | \theta_0)}{r(X)}. \]

The proof of this is really trivial.
Now, look at the example (MLE):

\[ r(X) = f(\bar{X} | X) \]

\[
\therefore \quad \pi(\theta_0 | X) = \left[1 + \frac{1 - \pi_0}{\pi_0} \exp \left( \frac{1}{2} t^2 \right) \right]^{-1} \\
\geq \left[1 + \frac{1 - \pi_0}{\pi_0} \frac{(2\pi \sigma^2 / T)^{-\frac{1}{2}}}{(2\pi \sigma^2 / T)^{-\frac{1}{2}} \exp - (\bar{X} - \theta_0)^2 / (2\sigma^2 / T)} \right]^{-1}
\]

For \( \pi_0 = \frac{1}{2} \) we get that for \( t > 1.68 \) (Berger and Selke; *JASA*, 1987).
\[ \pi (\theta_0 \mid x) \geq [P \text{ (value)}] \times [1.25t]. \]

<table>
<thead>
<tr>
<th>( t )</th>
<th>( P ) values</th>
<th>Lower bound on ( \pi (\theta_0 \mid x) )</th>
<th>Bound on Bayes factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.645</td>
<td>.1</td>
<td>.205</td>
<td>1/(3.87)</td>
</tr>
<tr>
<td>1.960</td>
<td>.05</td>
<td>.127</td>
<td>1/(6.83)</td>
</tr>
<tr>
<td>2.576</td>
<td>.01</td>
<td>.035</td>
<td>1/(27.60)</td>
</tr>
<tr>
<td>3.291</td>
<td>.001</td>
<td>.0044</td>
<td>1/(224.83)</td>
</tr>
</tbody>
</table>
Thm. Berger and Selke.

\[ G = \left\{ g_1 : \begin{array}{l}
g_1 \text{ is symmetric about } \theta_0 \\
\text{and non-increasing in } |\theta - \theta_0| 
\end{array} \right\} \]

| $t$      | $P$ values | Lower bound on $\pi (\theta_0 | x)$ | Bound on Bayes factor |
|----------|------------|-------------------------------------|-----------------------|
| 1.645    | .1         | .390                                | 1/(1.56)              |
| 1.960    | .05        | .290                                | 1/(2.45)              |
| 2.576    | .01        | .109                                | 1/(8.17)              |
| 3.291    | .001       | .018                                | 1/(54.55)             |
B Bayesian Credibility Sets

Def. A 100 \((1 - \alpha)\)% credible set for \(\theta\) is a subset \(C\) of \(\theta\) such that

\[
1 - \alpha \leq P(C \mid x) = \int_C dF^{\pi(\theta \mid x)}(\theta)
\]

\[
= \int_C \pi(\theta \mid x) \, d\theta.
\]
What is “best” ratio? A 100 $(1 - \alpha)\%$ highest posterior density credible set is a subset $C$ of $\theta$ such that

$$C = \{ \theta \mid \pi (\theta \mid x) \geq k (\alpha) \}$$

where $k (\alpha)$ is the largest constant such that $P (C \mid x) \geq 1 - \alpha$. 
Example. Let $\pi (\theta \mid x) = N (\mu (x), \sigma^2)$ be a posterior $100 (1 - \alpha) \%$ credible set is given by

$$
\left( \mu (x) + \Phi^{-1} \left( \frac{\alpha}{2} \right) \sigma, \mu (x) - \Phi^{-1} \left( \frac{\alpha}{2} \right) \sigma \right),
$$

identical to the classical case for normal posteriors. Consider, however, what would happen in the Dickey Leamer $T$ case. A possibility of disconnected credibility sets (multimodel). This can happen in the classical case.
Problems. The forms of credible sets are not subject to simple ends.

Example. Suppose the posterior is

\[
\pi (\theta \mid x) = \frac{ce^{\theta}}{1 + \theta^2} \quad \theta \geq 0
\]

\[
\ln \pi (\theta \mid x) = \ln c + \theta - \ln (1 + \theta^2)
\]

\[
\frac{\partial \ln \pi (\theta \mid x)}{\partial \theta} = 1 - \frac{2\theta}{1 + \theta^2}
\]
\[
\frac{\partial^2 \ln \pi(\theta \mid x)}{\partial \theta^2} = -\frac{2}{1 + \theta^2} + \frac{(2\theta)(2\theta)}{(1 + \theta^2)^2}
\]

\[
= \frac{2}{1 + \theta^2} \left[ \frac{2\theta^2}{1 + \theta^2} - 1 \right]
\]

\[
= \frac{2}{1 + \theta^2} \left[ \frac{\theta^2 - 1}{1 + \theta^2} \right] > 0,
\]

where \(\theta > 1\). This is increasing in \(\theta\). \(\therefore\) credible set is \((b(\alpha), x)\).
Use a transformation:

\[ \eta = \exp(\theta) \]

\[ \pi(\eta) = c \left(1 + (\log \eta)^2\right)^{-1} \]

decreasing on \((1, \exp(x))\). \therefore we have a credible set given by \((1, d(\alpha))\), and in terms of the original coordinate system, by \((0, \ln d(\alpha))\)

(upper tail in original parametrization, lower tail in new parametrization).
Model Selection in Bayesian Way

The common classical approach is to use a pretest procedure. We seek $P(Y \mid H), Y = X \beta + U, U \sim N(0, h^{-1}I_T)$. Predictive density under the hypothesis:

$$f(Y \mid H) = \int \int f(Y \mid \beta, h) f(\beta, h) \, d\beta \, dh,$$

$$P(H \mid Y) = f(Y \mid H) P(H)$$

Prior $P(H) : \beta \sim N(b^*; (H^*)^{-1})$

$$Y \sim N(Xb^*, h^{-1}I + X(H^*)^{-1} X')$$

where we assume that $h$ is known.
\[
V(Y) = h^{-1}I_t + X(H^*)^{-1}X'
\]

\[
f(y) = (2\pi)^{-\frac{1}{2}}|V(Y)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}hQ\right)
\]

\[
hQ = (Y - Xb^*)'V^{-1}(Y - Xb^*)
\]
Let $N^* = h^{-1} H^*.$

$$V^{-1}(y) = \left[ X (H^*)^{-1} X' + h^{-1} I_T \right]^{-1}$$

$$= h \left[ I_T - X (X'X + N^*)^{-1} X' \right]$$

$$\left| V(y) \right|^{-1} = h^T \left| N^* + X'X \right|^{-1}$$

$N = X'X$

$$\hat{b} = (X'X)^{-1} X'Y$$
Exercise. Prove that

\[
Q = \left( Y - X\hat{b} \right)' (Y - X\hat{b})
+ \left( \hat{b} - b^* \right) \left( (N^*)^{-1} + N^{-1} \right)^{-1} \left( \hat{b} - b^* \right)
= (Y - Xb^*)' (Y - Xb^*)
- \left( \hat{b} - b^* \right) N (N^* + N)^{-1} N \left( \hat{b} - b^* \right).
\]

For the case when \( h^{-1} \) is unknown (gamma-normal case),

\[
(\beta, h) \sim f_N (\beta \mid b^*, h^{-1} (N^*)^{-1}) \\
f_\nu (h \mid S^2_1, \nu_1).
\]

22
Then, as shown, we have predictive density:

\[
\begin{align*}
f (Y) &= \int \cdots \int f (Y \mid \beta, h) f (\beta, h) \, d\beta \, dh \\
&= k (\nu_1, T) \left| \frac{M}{S_1^2} \right|^{\frac{1}{2}} \left( \nu_1 + \frac{Q}{s^2} \right)^{-\frac{\nu_1 + T}{2}} \\
M &= I_T - X (N^* + X'X)^{-1} X' \\
|M| &= |N^*| |N^* + X'X|^{-1} \\
k (\nu_1, T) &= \frac{(\nu_1)^{\nu_1/2} \left( \frac{\nu_1}{2} + \frac{T}{2} - 1 \right)!}{\pi^{T/2} \left( \frac{\nu_1}{2} - 1 \right)!}
\end{align*}
\]
The Bayes factor for $H_i$ relative to $H_j$ is

$$\frac{k (\nu_{1i}, T)}{k (\nu_{1j}, T)} \frac{|N_i^*|^{\frac{1}{2}}}{|N_j^*|^{\frac{1}{2}}} \frac{|N_i^* + X_i'X_i|^{-\frac{1}{2}}}{|N_j^* + X_j'X_j|^{-\frac{1}{2}}} \cdot \frac{S_{1i}^{-T}}{S_{1j}^{-T}} \left( \nu_{1i} + \frac{Q_i}{S_{1i}^2} \right)^{-\frac{(\nu_{1i} + T)}{2}} \cdot \left( \nu_{ij} + \frac{Q_j}{S_{ij}^2} \right)^{-\frac{(\nu_{1j} + T)}{2}}.$$
D  Diffuse Prior Approach

Let $|N^*_\ell|^{\frac{1}{2}}$, $\ell = i, j$, get small and set $\nu_1 = 0$ (in which case we don’t get $\square$). What types of diffuse priors? (Many exist and they do not all lead to the same inference — a problem.)

\[
N^* = \delta I_K \quad \text{or}
\]
\[
N^* = (\delta)^{\frac{k}{2}} I_K \quad \text{or}
\]
\[
N^* = (\delta)^{\frac{k}{2}} X'X,
\]

and set $\delta \rightarrow 0$. This is like saying we don’t have many \textit{a priori} observations!
\[
\cdot \frac{f (Y \mid H_i)}{f (Y \mid H_j)} \rightarrow \delta \frac{\kappa_i}{2} - \frac{\kappa_j}{2} \frac{|X'_j X_j|^{1/2}}{|X'_i X_i|^{1/2}} \left( \frac{\text{ESS}_j}{\text{ESS}_i} \right)^{T/2},
\]

where ESS is the OLS regression sum of squares.
This strongly favors a small parameter model,

\[
= \frac{|X'_j X_j|^{1/2}}{|X'_i X_i|^{1/2}} \left( \frac{\text{ESS}_j}{\text{ESS}_i} \right)^{T/2}
\]

or

\[
= \left( \frac{\text{ESS}_j}{\text{ESS}_i} \right)^{T/2}.
\]

There is no unique diffuse prior.
E Dominated Priors and $BIC$

As $T$, $Q$, and $X'X$ grow,

$$
\frac{f (Y \mid H_i)}{f (Y \mid H_j)} = C \frac{|X_j'X_j|^{1/2}}{|X_i'X_i|^{1/2}} \left( \frac{\text{ESS}_j}{\text{ESS}_i} \right)^{T/2}.
$$

$T \to \infty$, to derive $BIC$, assume,:

$$
\frac{X'X}{T} \to \sum
$$
You acquire the term from the determinant
\[
\frac{N_i^* + \frac{X_i'X_i}{T_i^k}}{N_j^* + \frac{X_j'X_j}{T_j^k}}^{-\frac{1}{2}} \times \\
\left( \frac{S_{1j}}{S_{1i}} \right)^T \frac{(\nu_{1j} + Q_j/S_{1j}^2)^{\nu_{1j}+T}}{(\nu_{1i} + Q_i/S_{1i}^2)^{\nu_{1i}+T}} \frac{\nu_{1j}+T}{2} \frac{(\nu_{1j} + Q_j/S_{1j}^2)^{\nu_{1j}+T}}{(\nu_{1i} + Q_i/S_{1i}^2)^{\nu_{1i}+T}} \frac{\nu_{1j}+T}{2}
\]
\[
= \frac{S_{1j}^{-\nu_{1j}}}{S_{2j}^{-\nu_{2j}}} \frac{(\nu_{1j} + Q_j)^{\nu_{1j}+T}}{(\nu_{1i} + Q_i)^{\nu_{1i}+T}}
\]
\[
= \left( C^* T^{\frac{k_j-k_i}{2}} \right) \frac{(\nu_{1j} s_{1j} + Q_j)^{\nu_{1j}}}{(\nu_{1i} s_{1i} + Q_i)^{\nu_{1i}}} \frac{\left( \frac{\nu_{1j} s_{1j}}{T} + Q_j \right)^{\frac{T}{2}}}{\left( \frac{\nu_{1i} s_{1i}}{T} + Q_i \right)^{\frac{T}{2}}}
\]
\[
= CT^{\frac{k_j-k_i}{2}} \left( \frac{\text{ESS}_j}{\text{ESS}_i} \right)^{\frac{T}{2}}.
\]
Now observe $BIC$. For nested models, Bayes factor $B > 1 \iff$

$$F > \frac{T - K}{P} \left( T_{i}^{P} - 1 \right),$$

where $P = K_{j} - K_{i}$ is the number of restrictions. These methods easily handle non-nested setups (this is lacking in classical approaches), among other merits. Measures of location:

$$f (\beta_{i} \mid Y) = \sum_{j} f (\beta_{i} \mid Y, H_{j}) f (H_{j} \mid Y)$$

This allows for other model parameters to generate information about $i^{s}$ parameters. It also gives a measure of variability.
Issues:

1. Bayesian confidence intervals.
2. Bayesian $P$ values?
3. Multiple hypothesis testing.

Classical model selection, etc.