Bayes and Classical Approaches: A Comparative Overview

Econ 690

Purdue University

December 23, 2011
Outline

1. Review

2. Exchange Paradox

3. Example #1

4. exchangeability

5. The Likelihood Principle
   - Example #1
   - Example #2
For both the Bayesian and the Frequentist, the object of interest is denoted as $\theta \subseteq \Theta \subseteq R^k$, a parameter which characterizes some feature of the population at large (e.g., the average height of men in the U.S.)

A sample of data is collected, which must then be used to shed light on the value of $\theta$.

A variety of estimators - rules taking the data $y$ and mapping that data into a “guess” regarding the value of $\theta$ - are possible. The performance of an estimator is evaluated by the frequentist by analyzing its sampling distribution.
“Frequentist econometrics is tied to the notion of sample and population” (Lancaster, 2004, page 360). A “good” estimator is one which is “close” to $\theta$, where “closeness” is measured by its performance averaged over hypothetical repeated sampling from the population.

Therefore, although a given estimate obtained by a particular researcher may be far away from $\theta$, he or she could take comfort in the fact that other researchers, obtaining different samples from the population of interest, and making use of the same estimator, will obtain estimates that are “close” to $\theta$, at least on average.
To illustrate, consider the estimator:

For any sample of data, the realized estimate can never equal the true parameter of interest $\theta$. Thus, ex post, (i.e., after seeing the data), we know that the estimate is not “good” in the sense that, whenever the estimator it is applied, there is no chance of realizing a estimate equal to the true parameter $\theta$.

However, viewed according to the ex ante criterion of unbiasedness, the estimator is “close” to the true parameter $\theta$ in the sense that if we could apply our estimator over and over again, on average, we would get the “correct” $\theta$. 
The Bayesian Approach

- Unlike the frequentist point of view, the Bayesian perspective provides no role for data that could have been observed, but is not. That is, its focus is *ex post* - conditioning on the observed data $y$, rather than *ex ante* - which involves averaging over the sampling distribution of $y|\theta$.

- In fairness, the Bayesian approach requires more inputs, and must specify the full joint distribution of observables $y$ that become known under sampling and unobservables $\theta$. This joint distribution is typically decomposed as

  $$p(y, \theta) = p(y|\theta)p(\theta),$$

  into the likelihood $p(y|\theta)$ [which, viewed as a function of $\theta$ for given $y$ is denoted $L(\theta)$] and a prior $p(\theta)$. 
The Bayesian Approach

- The demands required of the model specification open the Bayesian up to criticism regarding sensitivity of results to the specification of the likelihood and/or the prior. In frequentist econometrics, far less is required - no prior is needed, and in many cases, and a complete specification of the sampling distribution $p(y|\theta)$ is not necessary and only moment conditions are assumed to hold.

- However, in such cases, approximations must be used to characterize the sampling distribution of the estimator since the model is not rich enough to allow for analytic calculation of this distribution. Thus, flexibility comes at a cost.
Bayesians regard the necessary additional structure as a worthy investment: By construction, Bayesian inference provides \textit{exact finite sample results}.

In many cases, flexible representations of the likelihood (allowing for fat tails, skewness or multimodality) can go a long way toward easing concerns regarding its specification.

In sufficiently large samples, the influence of the prior generally vanishes, and it is typically wise to conduct a sensitivity analysis over alternate prior choices. (For testing purposes, the prior is far more important).
The Bayesian Approach

- Rather than having a variety of competing estimators for the same problem, the Bayesian “estimator” is unique - all information regarding $\theta$ is summarized in the posterior distribution $p(\theta | y)$. Different features of this posterior can be used for point estimates, e.g., the posterior mean or the posterior mode, depending on the relevant *Loss function*.

- Nuisance parameters can cause headaches for frequentist econometrics. In the Bayesian framework, they are handled intuitively and consistently across various models, by simply integrating them out of the posterior distribution.
In finite dimensional models with standard regularity conditions, the following are known:

- **Asymptotic Interval Agreement** There exist regions $R_n$ such that
  \[ \int_{R_n} p(\theta | y_n) d\theta = 1 - \alpha \text{ and } R_n \text{ has frequentist coverage } 1 - \alpha + O(n^{-1}). \]
Exponential Family Example

Consider a set of scalar independent and identically distributed (iid) observations, denoted here by $y_1, y_2, \ldots, y_n$, from the natural exponential family of distributions:

$$p(y_i|\theta) = h(y_i) \exp [\theta y_i - g(\theta)], \quad i = 1, 2, \ldots, n,$$

where the functions $h$ and $g$ are known and $y_n$ will be used to denote the full vector of $n$ observations.
Exponential Family Example: Sampling Theory

Provided one can interchange the order of integration and differentiation:

where \( p_\theta(y|\theta) \equiv \frac{\partial p(y|\theta)}{\partial \theta} \) and \( g_\theta(\theta) \) is defined similarly.

This immediately implies

\[
E(y) = g_\theta(\theta_0)
\]
By extension:

\[ \text{Var}(y) = g_{\theta\theta}(\theta_0), \]

where \( g_{\theta\theta}(\cdot) \) is analogously defined as the second derivative.

It is straightforward to show that the MLE is determined by solving

\[ g_{\theta}(\hat{\theta}_n) = \bar{y}_n, \]

where \( \bar{y}_n \equiv n^{-1} \sum_{i=1}^{n} y_i \) and \( \hat{\theta}_n \) represents the MLE.
Under suitable regularity conditions, we obtain the following asymptotic approximation to the sampling distribution of $\hat{\psi}_n$:

from which large-sample confidence intervals can be constructed.
Consider the prior

\[ p(\theta|a, b) \propto \exp[a\theta - bg(\theta)], \]

Using the same type of argument, it is straightforward to show, provided the moment exists:

\[ E(\psi) \equiv \mu_{\psi} = a \cdot b^{-1}, \]

with \( \psi \equiv g_{\theta}(\theta). \)
Bayesian Inference

Combining the likelihood and prior, we obtain via Bayes Theorem:

\[ p(\theta|y_n) \propto \exp \left[ (n\overline{y}_n + a)\theta - (n + b)g(\theta) \right]. \]

It therefore follows:

- 

which we write equivalently as

- 


Justin L. Tobias (Purdue)  Comparative Overview  December 23, 2011  16 / 46
Bayesian Inference

The previous equation immediately reveals

Since $\omega_n \to 1$, we see that the Bayesian posterior mean and the frequentist MLE in this case are asymptotically equivalent.
In order to conduct inference, for which finite sample results are seldom available, the frequentist typically relies on asymptotic approximations to the sampling distribution of the estimator.

But under this large sample metric, the sampling distribution of the Bayes rule is identical, suggesting that, according to his or her own recipe for inference, the prior does not matter.

To question the role of the prior, then, from a sampling theory perspective is to introduce the importance of finite sample results; otherwise concerns about the prior must be misplaced and unwarranted, as the Bayes rule simply represents a different estimator with identical large sample sampling properties.
Another reason why one may pursue a Bayesian approach is that prior information is available (and should be used) while not using such information can lead to undesirable results.

An example of this is the Exchange Paradox (or “two envelope problem”): see, e.g. Christensen and Utts (1992 American Statistician)

An amount of money \( m \) is placed in one envelope and \( 2m \) placed in another. One of the envelopes is randomly handed to a player and then opened. The player is permitted to keep the prize or, if desired, choose the other envelope.
The agent opens the envelope, revealing an amount $x$.
He reasons that the other envelope either contains $x/2$ or $2x$.
Since the envelope was distributed randomly, the chance of each is 1/2.
His expected return from switching is

so he switches.
However, your “opponent” thinks the same way and switches as well. How can this be reasonable?
A solution to this problem notes that the agent probably has some beliefs over the distribution of prize money, and this should be (and probably is) taken into account when making a decision.

Formally, let $M$ be the prize placed in the first envelope, and $X$ be the (ex ante) random amount placed in your envelope. Upon seeing $X = x$ in your envelope, $M = x$ or $M = x/2$.

The sampling distribution is
Bayes Theorem gives:

Likewise,
Exchange Paradox

If the agent keeps the envelope, he gets $x$ with certainty. If he switches, he gets:

- 
- 

Therefore, the expected return from switching is:
It is straightforward to show that, when \(2p(x) = p(x/2)\), the expected return from switching is \(x\). So, when

it is optimal for the agent to switch. Perhaps most importantly, when will the expected return equal \(5x/4\)?
This example is constructed to illustrate the differences between the ex ante and ex post perspectives. (It also helps to clarify the between a Bayesian posterior interval and a frequentist confidence interval).
Given an unknown parameter $\theta$, $-\infty < \theta < \infty$, suppose $Y_i \ (i = 1, 2)$ are iid binary random variables with probabilities of occurrence equally distributed over points of support $\theta - 1$ and $\theta + 1$. That is,

1. $\Pr(Y_i = \theta - 1|\theta) = 1/2$ and 
2. $\Pr(Y_i = \theta + 1|\theta) = 1/2$.

Suppose the prior for $\theta$ is constant.
Ex Ante v.s. Ex Post: An Example

\[ \Pr(Y_i = \theta - 1|\theta) = \frac{1}{2}, \quad \text{and} \quad \Pr(Y_i = \theta + 1|\theta) = \frac{1}{2}. \]

(a) Suppose we observe \( y_1 \neq y_2 \). Find the posterior distribution of \( \theta \).

(b) Suppose we observe \( y_1 = y_2 \). Find the posterior distribution of \( \theta \).

(c) Consider

\[ \hat{\theta} = \begin{cases} 
\frac{1}{2}(Y_1 + Y_2) & \text{if } Y_1 \neq Y_2 \\
Y_1 - 1 & \text{if } Y_1 = Y_2.
\end{cases} \]

What is the ex ante probability that this estimator contains (i.e., equals) \( \theta \)?
Pr(\(Y_i = \theta - 1|\theta\)) = 1/2, \text{ and } Pr(\(Y_i = \theta + 1|\theta\)) = 1/2.

(a) If \(y_1 \neq y_2\), then one of the values must equal \(\theta - 1\) and the other equals \(\theta + 1\). \textit{Ex post}, averaging these two values it is \textit{absolutely certain} that \(\theta = (1/2)(y_1 + y_2)\), i.e.,

(b) If \(y_1 = y_2 = y\), say, then the common value \(y\) is either \(\theta - 1\) or \(\theta + 1\). Since the prior does not distinguish among values of \(\theta\), \textit{ex post}, it is \textit{equally uncertain} whether \(\theta = y - 1\) or \(\theta = y + 1\), i.e.,
Pr\((Y_i = \theta - 1|\theta) = 1/2\), and \(Pr(Y_i = \theta + 1|\theta) = 1/2\).

(c) Parts (a) and (b) suggest the \textit{ex post} (i.e., posterior) probability that our estimator equals \(\theta\) is either 1 or 1/2, depending on whether \(y_1 \neq y_2\) or \(y_1 = y_2\).

From the \textit{ex ante} perspective, however,

The difficult question for the pure frequentist is: why use the realized data to estimate \(\theta\) and then report the \textit{ex ante} confidence level 75% instead of the appropriate \textit{ex post} measure of uncertainty?
Consider a sequence of Bernoulli random variables $x_1, x_2, \ldots, x_n$ and the joint probability

$$p(x_1, x_2, \ldots, x_n).$$

Suppose, additionally, that you feel the labels $1, 2, \ldots, n$ are uniformative in the sense that all of the marginal distributions of the random quantities (and marginals of pairs, triples, etc.) are the same. Such a belief implies that

$$p(x_1, \ldots, x_n) = p(x_{\pi(1)}, \ldots x_{\pi(n)}),$$

where $\pi$ is an operator that permutes the labels $i = 1, \ldots, n$. If the equation above holds for any permutation operator $\pi$, then the random quantities $x_1, x_2, \ldots, x_n$ are regarded as exchangeable.
What would NOT be an exchangeable sequence?

Consider the case of “streaks” in sports, where a team who has just won its previous game is more likely to win the next, and conversely, a team who has just lost a game is more likely to loose the next. In this case,

\[ p(1, 1, 1, 0, 0, 0) \]

would not be believed to equal

\[ p(1, 0, 1, 0, 1, 0) \]

and thus the joint probability is not preserved under permutation. Thus, the sequence would not be regarded as exchangeable.
What would NOT be an exchangeable sequence?

[Bernardo and Smith (2000)] Suppose that $x_1, \ldots, x_n$ denote measurements of a particular substance, all made by the same individual using the same process and equipment. In this case, it may be reasonable to assume that the sequence of measurements is exchangeable.

If, however, the measurements are taken by a series of different labs, then one might be skeptical of assuming exchangeability of the entire sequence. One might, however, still be comfortable with assuming that exchangeability is appropriate for the sequences within each lab. Such beliefs are regarded as partially (or conditionally) exchangeable.
Our common assumption of iid data is a stronger condition than assuming the data are exchangeable.

To see this, we need to show the equivalence of $p(x_1, x_2, \ldots, x_n)$ and $p(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)})$ when the data are iid:

\[
p(x_1, x_2, \ldots, x_n) = \prod_{j=1}^{n} \Pr(X_j = x_j)
\]

\[
= \prod_{j=1}^{n} \Pr(X = x_j)
\]

\[
= \Pr(X = x_{\pi(1)})\Pr(X = x_{\pi(2)}) \cdots \Pr(X = x_{\pi(n)})
\]

\[
= p(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)})
\]
Suppose

\[ Y = [Y_1 \ Y_2 \ \cdots \ Y_T]' \sim N(0_T, \Sigma), \]

where

\[ \iota \]

and \( \iota \) is a \( T \times 1 \) of ones.

Let \( \pi(t) \ (t = 1, 2, \cdots, T) \) be a permutation of \( \{1, 2, \cdots, T\} \) and write

\[ [Y_{\pi(1)} \ Y_{\pi(2)} \ \cdots \ Y_{\pi(T)}]' = AY, \]

where \( A \) is a \( T \times T \) selection matrix such that for \( t = 1, 2, \cdots, T \), row \( t \) in \( A \) consists of all zeros except column \( \pi(t) \) which is unity.

Show that these beliefs are exchangeable but the events are not necessarily independent.
exchangeability and de Finetti: Example

Note that

Then,

\[ AY \sim N(0_T, \Omega) \]

where

Hence, beliefs regarding \( Y_t(t = 1, 2, \cdots T) \) are exchangeable. Despite this exchangeability, if \( \alpha \neq 0 \), \( Y_t(t = 1, 2, \cdots T) \) are not independent.
Suppose that $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of Bernoulli random variables (every finite subsequence is exchangeable). Then, the joint probability distribution $p(x_1, x_2, \cdots, x_n)$ must take the form:

$$
p(x_1, x_2, \ldots, x_n) = \int_0^1 \left[ \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{(1-x_i)} \right] dQ(\theta)
$$

for some distribution function $Q(\theta)$. 
de Finetti’s Representation Theorem

\[ p(x_1, x_2, \ldots, x_n) = \int_0^1 \left[ \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{(1-x_i)} \right] dQ(\theta). \]

This result is profound from the viewpoint of subjectivist probability modeling. It:

1. “justifies” thinking about the data as being conditionally independent Bernoulli random variables (conditioned on a parameter \( \theta \)).
2. provides an explicit justification for the “prior” \( Q(\theta) \).
3. The assumption of exchangeability implies the Bayesian viewpoint of likelihood times prior.
The Likelihood Principle

Definition

The likelihood principle (loosely) asserts that if two experiments yield proportional likelihood functions for realized data $y$, i.e., $L_1(\theta) = cL_2(\theta)$, where $c$ does not depend on $\theta$, then (under the same prior), identical inferences should be obtained about $\theta$.

- This is a controversial statement (see, e.g. Robins and Wasserman, *JASA*, 2000), although it follows from two, seemingly reasonable conditions: Conditionality and (weak) sufficiency.
- In the Bayesian paradigm, this is not an imposed “principle,” but rather, follows as a consequence of Bayes Theorem.
To see this, consider two experiments with realized data $y_1$ and $y_2$, which yield proportional likelihood functions $L_1(\theta)$ and $L_2(\theta)$, respectively.

The posterior under experiment 1 is:

Thus, upon normalization, it is seen that the posteriors under each experiment are identical, and thus identical inferences are obtained.
In frequentist calculations, proportional likelihoods may not yield identical inferences, since the Classical paradigm provides an explicit role for data that could have been observed, but were not.

The following simple example illustrates this point.
The Likelihood Principle

- Consider two researchers, A and B.
- Researcher A observes a random variable $X$ having the Gamma distribution $G(3, \theta^{-1})$ with density
- Researcher B observes a random variable $Y$ having the Poisson distribution $Po(2\theta)$ with mass function
- In collecting one observation each, Researcher A observes $X = 2$ and Researcher B observes $Y = 3$. 
The respective likelihood functions are

$$L_A(\theta; x = 2) = \left( \frac{\theta^3}{\Gamma(3)} \right) 2^2 \exp(-2\theta) \propto \theta^3 \exp(-2\theta),$$

and

$$L_B(\theta; y = 3) = \frac{(2\theta)^3 \exp(-2\theta)}{3!} \propto \theta^3 \exp(-2\theta).$$
Since these likelihood functions are proportional, the evidence in the data about $\theta$ is the same in both data sets.

Because the priors and loss functions are the same, from the Bayesian standpoint the two researchers would obtain the same posterior distributions and therefore, make identical inferences.

From the standpoint of maximum likelihood estimation, since the likelihoods are proportional, the maximum likelihood point estimates would be the same, $\hat{\theta} = 1.5$.

However, the maximum likelihood estimators would be very different:

In the case of Researcher A, the sampling distribution would be continuous, whereas in the case of Researcher B the sampling distribution would be discrete.
The Likelihood Principle

- A second (and more widely cited) example also emphasizes this point.
- Suppose researcher $A$ decides to flip a coin $n$ times, and observes $l$ successes (heads) during these $n$ trials.
- The likelihood function in this case is the standard Bernoulli pdf:

$$L(\theta) = \binom{n}{l} \theta^l (1 - \theta)^{n-l}.$$
Now, suppose that researcher $B$ decides to flip a coin until she observes $l$ heads. It happens that it takes her exactly $n$ trials to observe $l$ heads.

In this case, the appropriate likelihood is the negative binomial distribution, which describes the distribution over the number of trials ($n$) necessary to achieve $l$ successes.

Since these two likelihoods are proportional, from the Bayesian point of view, these two researchers will obtain identical posterior inferences provided they employ the same prior.
However, inference in the classical case regarding $\theta$ will differ across the two experiments. Specifically, suppose $n = 12, l = 9$. Then, in the Bernoulli case:

$$\Pr(X \geq 9|H_0 : \theta = 1/2) = 0.073.$$ 

while in the Negative Binomial case

$$\Pr(X \geq 9|H_0 : \theta = 1/2) = 0.0327.$$ 

In fact, the classical econometrician cannot make inferences about $\theta$ unless he or she knew what the researcher intended to do before conducting the trials!

Depending on the intention of the researcher, you either reject (negative binomial) or fail to reject.