A Beginner's Notes on Bayesian Econometrics

Pierre-Carl Michaud
CentER, Tilburg University

September 6, 2002

1 Preliminaries

It is most probable that you have encountered in your studies the term Bayesian. The term Bayesian, as applied in statistical inference, recognized the contribution of the seventeenth-century English clergyman Thomas Bayes. Later, econometricians like Arnold Zellner have adapted Bayesian inference to econometrics. Nowadays, Bayesian econometrics is still not widely used but yet very promising. Thus I propose in these brief notes to give an introduction to Bayesian econometrics keeping the notation as simple as possible. One of the major obstacle of Bayesian econometrics is that it can become very messy and thus requires an advance technical background. Since I do not have this background, I will rather concentrate on defining as well as possible the basic concept of bayesian econometrics.

These notes were taken mainly from Arthur Van Soest's lectures on Bayesian Econometrics in Tilburg University, Netherlands, in the fall semester of 2001. Also, I have made used of Part 9 of the book *Foundations of Econometrics* by Mittlehammer, Judge and Miller (2000). At times, I will also give references to classical papers (the name may be misleading) on this topic and provide the reader with various examples drawn from the lectures and from the book.

2 Illustrations of the Use of Bayes' Rule

Bayesian statistical inference is mainly based on solving one problem that we call the inverse problem in order to contrast it from classical statistical inference. According to Mittlehammer et al. (now MJM) p646: "In the problem posed by Bayes, we observe data, and thereby know the values of the data outcomes, and wish to know what probabilities are consistent with those outcomes." This is different from the classical perspective where given the data we observe the likelihood that observations have been generated from some parameter to find. In the classical framework we thus give probabilities to sample observations that they have been drawn from a known distribution with parameter θ. In the bayesian framework, the give probabilities to plausible values of θ that are plausible with the data that has been observed. In this sense, these probabilities may change once the data is observe and from these probabilities we may be able to choose one particular value of θ as being the most plausible value, thus the true value. We verify this intuition by introducing the cornerstone of Bayesian Analysis, that is not suprisingly, Bayes' rule or theorem:

Given two outcomes A and B, we have that

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} \tag{1}$$

When dealing with probabily densities we may have accordingly,

$$f(x \mid y) = \frac{f(y \mid x)f_x(x)}{f_y(y)} \tag{2}$$

where $f_y(y) = \int_x f(t,y)dt$ and $f_x(x) = \int_y f(x,t)dt$ are the marginal densities and $f(x,y)$ is the joint density of the variables. Remember that since $f(x \mid y) = \frac{f(x,y)}{f_y(y)}$ and $\frac{f(x,y)}{f_x(x)} f_x(x) = f(y \mid x)f_x(x)$ we get the expression in (2). Enough abstract concepts you will say. Let's consider an example that uses Bayes' rule.

2.1 Example 1 Types of Female Workers

We have three types of workers $S_i \subset \mathbb{S}$ $i = 1,2,3$ with $S_1 \cup S_2 \cup S_3 = \mathbb{S}$ and $S_i \cap S_j = \{\oslash\} \forall i \neq j$. We are given that the probability that we observe w, a female who is working given that she is of type 1 is $p(w \mid S_1) = 0.5$, $p(w \mid S_2) = 0.1$, $p(w \mid S_3) = 0.2$. Further we know that the probability of observing a type i worker is $p(S_1) = 0.4$, $p(S_2) = 0.4$, $p(S_3) = 0.2$.

Now what is the probability of observing a worker of the first type given we observe w. We have the inverse problem and using Bayes' rule. Carefully,

$$\begin{aligned} p(S_1 \quad \mid \quad w) &= \frac{p(S_1)p(w \mid S_1)}{p(S_1)p(w \mid S_1) + p(S_2)p(w \mid S_2) + p(S_3)p(w \mid S_3)} \\ &= \frac{0.2}{0.2 + 0.04 + 0.04} = \frac{5}{7} \end{aligned}$$

Thus the probability of observing a first type worker given that this worker is a female is $\frac{5}{7}$. We have solved the inverse problem using Bayes' rule.

3 Basis Structure of Bayesian Inference

Remember, the philosophy here is that we work with inverse probabilities. Once the drawn sample is given and observed, what is the probability that we observe θ given the data. The Bayesian problem format is the following according to MJM.

1. Available sample $x = (x_1, ..., x_n)$ with $f(x \mid \theta)$ and $l(\theta \mid x)$, $\theta \in \Theta$

2. Prior information in the form of a **prior distribution** or a **prior probability density** $p(\theta)$ for the parameter vector $\theta \in \Theta$ in the sampling probability model $f(x \mid \theta)$.

3. The likelihood function $l(\theta \mid x)$ and prior density combined by Bayes theorem to yield the posterior density of θ $p(\theta \mid x)$.

The general Structure of Bayesian Inference is shown in the Table 1 (MJM, p648).

3.1 Prior Distribution

You can possibly argue, what is a Prior? Is it objective? Subjective? We will show later the impact of the choice of a Prior on the estimation and thus characterize different kinds of priors. For the moment we will stick to the general interpretation of a prior. It is presumed that if $p(\theta)$ represents subjective information, then the analyst has adhered to the axioms of probability (whatever that is!) in defining $p(\theta)$ so that the function is indeed a legitimate probability measure on θ values. This is pretty loose guidance on the choice of a prior. In fact there as not been many attemps to improve guidance on the choices of priors. The title of Kass and Wasserman's (1996) paper is instructive on the dilemma posed to researchers: The Selection of Prior Distributions by Formal Rules. The question may well be Where do prior's come from? Later we will show that in fact the choice of a prior does not matter that much after all in large samples!

MJM p651 further add: "By formalizing uncertainty regarding model parameters in the form of prior probability distributions, the Bayesian approach allows differing beliefs about the plausible values of these parameters to be incorporated explicitely into inverse problem solutions."

3.2 Posterior Distribution

We can define the joint PDF of (Y, Θ) as,

$$f(x, \theta) = f(x \mid \theta)p(\theta) = p(\theta \mid x)f_x(x)$$

thus,

$$p(\theta \mid x) = \frac{f(x \mid \theta)p(\theta)}{f_x(x)}$$

and since $f_x(x)$ is a constant, we can can write

$$\begin{aligned} p(\theta \mid x) &\propto f(x \mid \theta)p(\theta) \\ &\text{or} \\ p(\theta \mid x) &\propto l(x \mid \theta)p(\theta) \end{aligned}$$

The sign $\propto$ means proportinal to and since we are talking of distributions, proportion can be neglected for estimation (remember in the Maximum likelihood framework the constant can be dropped without affecting the optimization problem and it's solution). We can always retreive this proportion scalar by using the fact that probabilities have to some up to 1. We call $p(\theta \mid x)$ the posterior distribution. Why? Look at the last expression. The usual likelihood that assigns probabilities to observations that they are drawn from a distribution with parameter value θ is present. However each possible value of θ is weighted by our beliefs about the values it can take.

Right, now we know that given the data we will update our beliefs and define a possible different posterior distribution about the values of the parameter. Thus we updated our beliefs using the data. The distribution itself may change or only the parameter values. Nothing is restricted in this process of updating. This process of updating is what will permit latter comparisons between the classical estimators and the bayesian estimators. It naturally follows that we consider some examples so that we really get a grasp of the meaning of all those concepts.

3.3 Example 2 Firms and Quality Control

We have the following Prior distribution about the probability of default of the firms:

$$p(\theta) = \left\{ \begin{array}{c} 0.5 \text{ for } \theta = 0.25 \\ 0.5 \text{ for } \theta = 0.5 \end{array} \right\}$$

Implied is that $\Theta = \{0.25, 0.5\}$. We have $x_1, x_2, ..., x_n$ a random sample of a quality of product measure from a binomial distribution:

$$x_i \sim B(1, \theta)$$

So that $x_i = 1$ with probablity θ and 0 with probability $(1 - \theta)$. We must compute the posterior distribution of θ given that $\theta = \theta_0$, the true value of the parameter vector.

$$\begin{aligned} p(\theta_0 \quad \mid \quad x) &= \frac{p(\theta_0)p(x \mid \theta_0)}{p(x)} \\ &= \frac{p(\theta_0)p(x \mid \theta_0)}{\sum_{\theta \in \Theta} p(\theta)p(x \mid \theta)} \end{aligned}$$

Now we have some calculations to do in order to compute this distribution. We first consider the case $\theta_0 = 0.25$:

$$\begin{aligned} p(0.25 \quad \mid \quad x) &= \frac{0.5\left(0.25^{\Sigma x_i} 0.75^{n - \Sigma x_i}\right)}{0.5\left(0.25^{\Sigma x_i} 0.75^{n\Sigma x_i}\right) + 0.5\left(0.5^{\Sigma x_i} 0.5^{n - \Sigma x_i}\right)} \\ &= \frac{0.5^{\Sigma x_i} 1.5^{n - \Sigma x_i}}{0.5^{\Sigma x_i} 1.5^{n\Sigma x_i} + 1} \end{aligned}$$

and consequently $p(0.25 \mid x) = 1 - p(0.25 \mid x)$.

$$\begin{aligned} p(0.5 \quad \mid \quad x) &= 1 - \frac{0.5^{\Sigma x_i} 1.5^{n-\Sigma x_i}}{0.5^{\Sigma x_i} 1.5^{n\Sigma x_i} + 1} \\ &= \frac{1}{0.5^{\Sigma x_i} 1.5^{n\Sigma x_i} + 1} \end{aligned}$$

Then we see that the probability that the firm is a bad type ($\theta = 0.5$) is small when the data reveals that there is not many defaults. More interestingly we can interpret that given our beliefs that this value should be given a probability of 0.5, if the data reveals few defaults, then this probability will decrease, again as a consequence of the updating. For the sake of completness, the posterior distribution is

$$p(\theta \mid x) = \left\{ \begin{array}{l} \frac{0.5^{\Sigma x_i} 1.5^{n-\Sigma x_i}}{0.5^{\Sigma x_i} 1.5^{n\Sigma x_i} + 1} \text{ for } \theta = 0.25 \\ \frac{1}{0.5^{\Sigma x_i} 1.5^{n\Sigma x_i} + 1} \text{ for } \theta = 0.5 \end{array} \right\}$$

which now depends on the data, so the data was used in a good way after all...

3.4 Example 3 An Uninformative Prior

We have a prior distribution on θ $U(0,1)$, a uniform distribution on the interval 0,1. This kind of prior is **uninformative** because each values of θ are assigned equal probabilities (we could say the same thing in the preceding example). We have data on n independent draws from a binomial distribution $B(1,\theta)$ and thus $x_1, x_2..., x_J \sim B(n,\theta)$. (You could think of how many defects by firms.) Recall:

$$p(x = k \mid \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

Now, the prior density is,

$$p(\theta) = \left\{ \begin{array}{ll} 1 & 0 \leq \theta \leq 1 \\ 0 & \theta < 0 \text{ or } \theta > 1 \end{array} \right\}$$

and thus we can compute the posterior density.

$$p(\theta_0 \mid x) = \frac{p(\theta_0) p(x \mid \theta_0)}{\int_{\theta \in \Theta} p(\theta) p(x \mid \theta) d\theta}$$

since $\int_{\theta \in \Theta} p(\theta) p(x \mid \theta) d\theta$ is a constant that does not depend on θ_0, we have for $0 \leq \theta_0 \leq 1$

$$
\begin{aligned}
p(\theta_0 \quad | \quad x) &\propto p(\theta_0) p(x \mid \theta_0) \\
p(\theta_0 \quad | \quad x) &\propto 1 \binom{n}{x} \theta^x (1-\theta)^{n-x}
\end{aligned}
$$

and again $1\binom{n}{x}$ is some scalar so we forget about it. Therefore,

$$
p(0 \leq \theta_0 \leq 1 \mid x) \propto \theta^x (1-\theta)^{n-x}
$$

and

$$
p(\theta < 0 \ \cup \ \theta > 1 \mid x) = 0
$$

Thus the Posterior distribution is

$$
p(\theta_0 \mid x) = \begin{array}{c} \theta^x (1-\theta)^{n-x} \text{ if } 0 \leq \theta_0 \leq 1 \\ 0 \text{ if } \ \theta < 0 \text{ or } \theta > 1 \end{array}
$$

the first part is a Beta distribution if you recall. For such a distribution we have,

$$
\begin{aligned}
E\{\ \theta\} &= \frac{p}{p+q} \\
V\{\ \theta\} &= \frac{pq}{(p+q)^2(p+q+1)}
\end{aligned}
$$

with $p = x+1$ and $q = n-x+1$ we then have that

$$
\theta_0 \sim Beta\left(\frac{x+1}{n+2}, \frac{(x+1)(n-x+1)}{(n+2)^2 (n+3} \right)
$$

Now recall the prior distribution. We had $E\{\ \theta\} = V\{\ \theta\} = \frac{1}{12}$. Notice that if we observe no data, the expected value of not updated. As soon as data is used and that we observe some defects then the expected value of θ will change. If we have few defects, then the probability of defects is updated downward. If there are many defects, this probability is increased. We then update our beliefs given the data.

3.5 Example 4 Informative Prior

We have data $x_1, x_2 ..., x_n \sim N(\beta, \sigma^2)$ with σ^2 known. Prior distribution of β is $N(\mu, \tau^2)$ with μ, τ^2 known. Working out the posterior density for β we have,

$$\begin{aligned} p(\beta \mid x) &\propto p(\beta)p(x \mid \beta) \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{(\beta-\mu)^2}{\tau^2} + \sum_{i=1}\frac{(x_i-\beta)^2}{\sigma^2}\right]\right\} \end{aligned}$$

Working the awful expression in the exponential,

$$\begin{aligned} \sum_{i=1}\frac{(x_i-\beta)^2}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}(x_i-\beta)^2 \\ &= \sum(x_i-\overline{x})^2 + n(\overline{x}-\beta)^2 \end{aligned}$$

and since $\sum(x_i-\overline{x})^2$ is not dependent on β we take it out into the proportionality constant (we have the exponential of a sum). We obtain by replacing,

$$\begin{aligned} p(\beta \mid x) &\propto \exp\left\{-\frac{1}{2}\left[\frac{(\beta-\mu)^2}{\tau^2} + \frac{n(\overline{x}-\beta)^2}{\sigma^2}\right]\right\} \end{aligned}$$

Dividing by n in the second term and then expressing in a common denominator,

$$\propto \exp\left\{-\frac{1}{2}\left[\frac{(\beta-\mu)^2\left(\frac{\sigma^2}{n}\right) + (\tau^2)(\overline{x}-\beta)^2}{\tau^2\frac{\sigma^2}{n}}\right]\right\}$$

since only the term in beta should be keept, the others being again sended to the proportionality condition, we have by working out the polynomials,

$$\propto \exp\left\{-\frac{1}{2\tau^2\frac{\sigma^2}{n}}\left[\beta^2\left(\frac{\sigma^2}{n}+\tau^2\right) + \left(-2\mu\frac{\sigma^2}{n} - 2\overline{x}\tau^2\right)\beta\right]\right\}$$

The attentive reader will see that the term in bracket is nothing else then,

$$\propto \exp\left\{-\frac{1}{2\tau^2\frac{\sigma^2}{n}}\left(\frac{\sigma^2}{n}+\tau^2\right)\left(\beta - \frac{\mu\frac{\sigma^2}{n} + \overline{x}\tau^2}{\frac{\sigma^2}{n}+\tau^2}\right)^2\right\}$$

which can be rewritten as,

$$p(\beta \mid x) \propto \exp\left\{ -\frac{1}{\left(\frac{1}{\frac{\sigma^2}{n}} + \frac{1}{\tau^2}\right)} \left(\beta - \frac{\mu\frac{1}{\tau^2} + \overline{x}\frac{1}{\frac{\sigma^2}{n}}}{\frac{1}{\frac{\sigma^2}{n}} + \frac{1}{\tau^2}} \right)^2 \right\}$$

$$p(\beta \mid x) \propto \exp\left\{ \frac{(\beta - \widetilde{\mu})^2}{\widetilde{\sigma}^2} \right\}$$

with $\widetilde{\mu} = \frac{\mu\frac{1}{\tau^2} + \overline{x}\frac{1}{\frac{\sigma^2}{n}}}{\frac{1}{\frac{\sigma^2}{n}} + \frac{1}{\tau^2}}$ and $\widetilde{\sigma}^2 = \left(\frac{1}{\frac{\sigma^2}{n}} + \frac{1}{\tau^2}\right)$. The Posterior distribution is

$$p(\beta \mid x) \sim N(\widetilde{\mu}, \widetilde{\sigma}^2)$$

We may observe several remarks before going further.

Remark 1 — The prior is normal and so is the posterior. Then the prior is called a conjugate prior. We say that $N(\mu, \tau^2)$ is conjugate to the data $N(\beta, \sigma^2)$ for σ^2 fixed. Because data is also normal, then the prior is natural conjugate prior. We can use the posterior as a prior for the next iteration and are garantied that this prior is also conjugate to the data.

Remark 2 — Posterior only depends on the data through $\overline{x}$, the sample mean, which in this case is a sufficient statistic for β. We have,

$$\begin{aligned} p(\beta \quad \mid \quad x) &= g(\beta, \overline{x}) h(x) \\ &= g(\beta, \overline{x}) \cdot 1 \end{aligned}$$

The posterior mean is a natural estimator for β (minimizes posterior expected quadratic loss). We call it the Bayes estimator under quadratic loss. (We will come back to loss functions later.) What is the Bayes estimator?

$$\widetilde{\mu} = \underbrace{\frac{\frac{1}{\tau^2}}{\frac{1}{\frac{\sigma^2}{n}} + \frac{1}{\tau^2}}\mu}_{W.\text{Prior Mean}} + \underbrace{\frac{\frac{1}{\frac{\sigma^2}{n}}}{\frac{1}{\frac{\sigma^2}{n}} + \frac{1}{\tau^2}}\overline{x}}_{W\text{Sample Mean}}$$

We see that $\widetilde{\mu} \underset{n\to\infty}{\to} \overline{x}$ and $\widetilde{\sigma}^2 \underset{n\to\infty}{\to} 0$, So that the estimator is the classical estimator in large sample as the prior is given little weight (the likelihood is predominant in the posterior). We can write,

$$\sqrt{n}(\beta - \widetilde{\mu}) \sim N(0, n\widetilde{\sigma}^2) \underset{n\to\infty}{\to} N(0, \sigma^2)$$

Remark 3 — In general $\widetilde{\sigma}^2 \underset{n\to\infty}{\to} 0$, then the posterior becomes approximately normal if $n \to \infty$. Posterior does not depend on prior if $n \to \infty$. The likelihood always dominates in the posterior and thus the prior has no effect on the distribution (take logs of the general form of the posterior to see it)

Remark 4 — What if prior information is lousy? Given a prior $N(\mu, \tau^2)$, suppose that τ^2 is large and therefore that our beliefs are diffused (we call such a prior not suprisingly a diffuse prior). Most info in this case comes from the data as the weight on the prior mean decreases. We can see it from the density where when $\tau^2 \to \infty$ $\exp(a)$ goes to 1 and thus the prior is uninformative in the posterior. The proportionality factor will be 0 to accomodate for probability one on each beta (which have to sum to 1). We call such a prior an improper prior. The most often used is the following,

$$p(\beta) = \begin{array}{c} \frac{1}{2M} \text{ if -}M \leq \beta_0 \leq M \\ 0 \text{ if } \beta_0 < 0 \text{ or } \beta_0 > 1 \end{array}$$

Then the posterior is given by

$$p(\beta \mid x) = \begin{array}{c} \frac{1}{2M} \exp\left\{-\frac{1}{2}\sum_{i=1} \frac{(x_i-\beta)^2}{\sigma^2}\right\} \text{ if -}M \leq \beta_0 \leq M \\ 0 \text{ if } \beta_0 < -M \text{ or } \beta_0 > M \end{array}$$

If M is large then the posterior is the likelihood. We call these priors, improper priors.

In the last example we have said that the Posterior mean was the Bayesian estimator under the quadratic loss function. We now look at loss functions, a common tool used both in classical and bayesian estimation.

4 Bayesian Estimator and Loss Functions

We saw in the last section that the Bayesian estimator is a weighted average of the sample mean and the prior mean. However we did not establish why such an estimator was the best. Why not the median or the mode? It turns out that the optimal choice between the three first moment estimators depends on the loss function that we use. Let's define the best estimator as one that minimizes a loss function $l(\theta, m)$ where θ is the true value of the parameter and m is our choice variable in order to minimize the function. In the Bayesian context however we must condition on the data since the bayesian estimator will be a post-data estimator. We have that

$$\widehat{\theta} = \arg\min_m \mathbf{E}\{l(\theta, m) \mid x\}$$

Surely, the choice of the estimator will depend on the specification of the loss function. We typically encounter three types of loss functions:

$$\begin{aligned} l(\theta, m) &= (\theta - m)^2 \\ l(\theta, m) &= |\theta - m| \\ l(\theta, m) &= 1\{|\theta - m| > \varepsilon\} \end{aligned}$$

The first one is called the quadratic loss function, the second one, the absolute value loss function and the turn one as no name but you can probably think of one for yourself. We will call it the discrete loss function. We look at these three cases alternatively.

4.1 The Quadratic Loss Function

We must find an estimator for θ that minimizes the quadratic loss function. The problem is then,

$$\widehat{\theta} = \arg\min_m \int_{-\infty}^{+\infty} (\theta - m)^2 p(\theta \mid x) d\theta$$

Under certain regularity conditions, FOC will be given by

$$\int_{-\infty}^{+\infty} \frac{\partial}{\partial m'} (\theta - m)^2 p(\theta \mid x) d\theta = 0$$

which yields,

$$\begin{aligned} -\int_{-\infty}^{+\infty} 2(\theta - m) p(\theta \quad \mid \quad x) d\theta &= 0 \\ 2m(1) &= 2\mathbf{E}\{\theta \mid x\} \end{aligned}$$

which implies that $m = \mathbf{E}\{\theta \mid x\}$, the posterior mean. Then we say that under the quadratic loss function, the posterior mean is the Bayes Estimator.

4.2 The Absolute Value Loss Function

We must find an estimator for θ that minimizes the Absolute value loss function. The problem is then,

$$\widehat{\theta} = \arg\min_m \int_{-\infty}^{+\infty} |\theta - m| \ p(\theta \mid x) d\theta$$

We first partition the integral at $\theta = m$, since we know that this is the value of m for which the function is not differentiable.

$$\widehat{\theta} = \arg\min_{m} \int_{-\infty}^{m} (m-\theta)\ \ p(\theta \mid x)d\theta + \int_{m}^{+\infty} (\theta - m)\ \ p(\theta \mid x)d\theta$$

and furthermore we expand integrals since $\int (a+b)\,dt = \int a dt + \int b dt$,

$$\begin{aligned} \arg\min_{m} m \int_{-\infty}^{m} p(\theta \quad \mid \quad x)d\theta - \int_{-\infty}^{m} \theta\ p(\theta \mid x)d\theta \\ + \int_{m}^{+\infty} \theta\, p(\theta \quad \mid \quad x)d\theta - m \int_{m}^{+\infty} p(\theta \mid x)d\theta = \widehat{\theta} \end{aligned}$$

Using Leibniz's rule for the first order condition,

$$\begin{aligned} 1 \cdot p(\theta \quad < \quad m \mid x) + mp(m \mid x) - mp(m \mid x) \\ -mp(m \quad \mid \quad x) - 1 \cdot p(\theta > m \mid x) + mp(m \mid x) = 0 \end{aligned}$$

which implies,

$$p(\theta < m \mid x) = p(\theta > m \mid x) = \frac{1}{2}$$

which is only possible if m is the median of $\theta \mid x$. Thus we have that under the absolute value loss function, the Bayes estimator is the median of the posterior distribution.

4.3 The Discrete Loss Function

We must find an estimator for θ that minimizes the Discrete loss function. The problem is then,

$$\widehat{\theta} = \arg\min_{m} \int_{-\infty}^{+\infty} 1\left\{|\theta - m| > \varepsilon\right\}\ \ p(\theta \mid x)d\theta$$

We only need take the integral when the indicator function is 1. Thus we have

$$\widehat{\theta} = \arg\min_{m} \int_{-\infty}^{m-\varepsilon} p(\theta \mid x)d\theta + \int_{m+\varepsilon}^{+\infty} p(\theta \mid x)d\theta$$

The Objective function is then nothing else than $[1 - p\,(m - \varepsilon < \theta < m + \varepsilon)]$ and thus the problem can be put as to maximize $[1 - p\,(m - \varepsilon < \theta < m + \varepsilon)]$ by choosing m such that this probability is the highest. Examining a distribution like the normal density yields

us to conclude that this area is maximized if we choose the mode of the distribution, the highest point of the density. Obsiously $\varepsilon \to 0$ in order for this to be true. Otherwise we can always find a distribution where this is not true (+ some regularity conditions that we don't cover here).

We will mostly use the mean as the Bayes estimator, however the reader should note that the estimator is the same if the posterior density is a normal symmetric density, the most encountered density (Asymptotic relies on convergence to normal distribution so we should not worry to much about the choice of loss functions.).

5 The Linear Model

We now get our hands dirty with real econometrics, if we can call it that way. Suppose the linear model,

$$y_i = x_i'\beta + \varepsilon_i$$

with $\varepsilon_i \sim N(0, \sigma^2)$ and x_i fixed. We formulate an uninformative prior. Denote $\theta = (\beta, \sigma^2)$ and assume,

$$p(\beta) \propto 1$$

An improper prior and

$$p(\sigma) \propto 1\{\sigma > 0\}\frac{1}{\sigma}$$

You might ask why this prior for σ? We follow here the explanation of MJM p654-655 for this specification of the prior. Regarding the choice of a prior for sigma, note that the purpose of the value of sigma in the regression model is to parametrize or determine the standard deviation σ of the y's. We need that

$$p(\sigma \in A) = p(\tau\sigma \in A).$$

Since $\tau\sigma \in A$ iff $\sigma \in \tau^{-1}A$ then the set $\tau^{-1}A$ denotes the element in A each divided by the positive constant τ, such that

$$\int_A p(\sigma)d\sigma = \int_{\tau^{-1}A} p(\sigma)d\sigma = \int_A p(\tau^{-1}\sigma)\tau^{-1}d\sigma$$

This implies that the prior PDF must satisfy $p(\sigma) = p(\tau^{-1}\sigma)\tau^{-1}\ \forall\sigma$. This is satisfied by the PDF family $p(z) = z^{-1}$, then $p(\sigma) = \frac{1}{\sigma}1$.

5.1 The Joint Prior Distribution (Uninformative)

In the previous examples we only had one unknown parameter on which we had a prior. This time however we have two. A natural thing to do is to find the Joint Prior Distribution. Since both are independent then $p(\beta,\sigma) = p(\beta)p(\sigma)$. Thus,

$$p(\beta,\sigma) \propto I(\sigma > 0)\frac{1}{\sigma}$$

In fact we see that because $p(\beta)$ is totally uninformative we get that $p(\beta,\sigma) = p(\sigma) \propto 1\{\sigma > 0\}\frac{1}{\sigma}$. If is also an improper prior since $p(\sigma) = \int_0^\infty \frac{1}{\sigma} d\sigma = \ln\sigma \mid_0^\infty = \infty$. Notice also that we could use the uninformative prior $p(\beta,\sigma) = I(\sigma > 0)$. As MJM note, the choice between these two priors is negligible even in small sample ($n \simeq 20$) as vanishes as the sample size increases. Thus the choice that we make here is purely of convenience and of convention.

5.1.1 The Posterior Distribution

We now are familiar with posteriors and thus we have that since $p(x) = 1$, $p(y) = p(y \mid x)$. Thus,

$$p(\beta,\sigma \mid y) = p(\beta,\sigma) \times p(y \mid \beta,\sigma).$$

Now for $\sigma > 0$,

$$p(\beta,\sigma \mid y) \propto \frac{1}{\sigma}\left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - x_i'\beta)^2}{\sigma^2}\right\}.$$

We can rewrite

$$\begin{aligned}(y - X\beta)'(y - X\beta) &= \\ & y'y - y'(X\beta) - (X\beta)'y - (X\beta)'X\beta.\end{aligned}$$

Now repleace $y = Xb + e$ and we obtain (don't forget that $X'e = 0$),

$$\begin{aligned}(y - X\beta)'(y - X\beta) &= \\ & (Xb+e)'(Xb+e) - (Xb+e)'X\beta \\ & -(X\beta)'Xb - (X\beta)'X\beta \\ &= (Xb - X\beta)'(Xb - X\beta) + e'e \\ &= (b-\beta)'X'X(b-\beta) + e'e\end{aligned}$$

Now replacing into the posterior, we obtain (using the proportionality argument)

$$p(\beta, \sigma \mid y) \propto \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{1}{2\sigma^2}\left((b-\beta)'X'X(b-\beta)+e'e\right)\right\}$$

and using the classical estimator of σ^2, s^2 yields

$$p(\beta, \sigma \mid y) \propto \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{1}{2\sigma^2}\left((b-\beta)'X'X(b-\beta)+(n-k)s^2\right)\right\}$$

Now usually, Bayesians try to find the posterior of both parameters to find their estimator and their distribution. Using the Bayesian rule again we have that

$$p(\beta \mid y, \sigma) = \frac{p(\beta, \sigma \mid y)}{\int p(\beta, \sigma \mid y) d\beta}$$

and thus the denominator does not depend on beta anymore (it is the marginal of σ). Thus,

$$p(\beta \mid y, \sigma) \propto p(\beta, \sigma \mid y)$$

and since we have the exponential of a term that does not involve β in the posterior, we use again the proportionality trick and finally,

$$p(\beta \mid y, \sigma) \propto \exp\left\{-\frac{1}{2\sigma^2}(b-\beta)'X'X(b-\beta)\right\}.$$

Then we conclude that $\beta \mid y, \sigma \sim N_k(b, \sigma^2(X'X)^{-1})$. It is a conditional density. We find similarities with the classical estimator but this is not the same,

Classical	$b \sim N_k(\beta, \sigma^2(X'X)^{-1})$
Bayesian	$\beta \mid y, \sigma \sim N_k(b, \sigma^2(X'X)^{-1})$

5.1.2 Marginal Posterior

Now what is the marginal posterior for β (with estimated σ)? We have

$$\begin{aligned} p(\beta \mid y) &= \int_0^\infty p(\beta, \sigma \mid y) d\sigma \\ &= \int_0^\infty \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{1}{2\sigma^2} a\right\} d\sigma \\ a &= \left((b-\beta)'X'X(b-\beta)+(n-k)s^2\right) \end{aligned}$$

now let $z = \frac{1}{2\sigma^2}a$, then $\sigma^2 = \frac{a}{2z} \rightarrow \sigma = \sqrt{\frac{a}{2}}\sqrt{\frac{1}{z}} \rightarrow d\sigma = \sqrt{\frac{a}{2}} \times -\frac{1}{2}z^{-3/2}dz$. Then

$$\begin{aligned} &= \int_0^\infty \frac{1}{\left(\sqrt{\frac{a}{2}}\sqrt{\frac{1}{z}}\right)^{n+1}} \exp(-z)\sqrt{\frac{a}{2}} \times -\frac{1}{2}\sqrt{\frac{a}{2}}z^{-3/2}dz \\ &= \left(\frac{a}{2}\right)^{-n/2}\frac{1}{2}\int_0^\infty z^{\frac{n-2}{2}}e^{-z}dz \end{aligned}$$

Now the integral is over z and thus once the integral calculated, it does not depend anymore on the parameters. Thus this term goes again..... in the proportionality constant as for $\frac{1}{2}^{1-n/2}$ and therefore,

$$p(\beta \mid y) \propto a^{-n/2}$$

and if we replace this expression we have,

$$p(\beta \mid y) \propto \left((b-\beta)'X'X(b-\beta) + (n-k)s^2\right)^{-n/2}$$

which we can rewrite as

$$p(\beta \mid y) \propto \left(1 + \frac{(b-\beta)'X'X(b-\beta)}{(n-k)s^2}\right)^{-n/2}.$$

Now this is not evident but if you look into a statistic's book you will find that this expression is the expression of a Multivariate Student distribution (We have a multivariate normal on the numerator and a chi-square at the denominator.) Thus if $x \in \mathbb{R}$, $\beta \in \mathbb{R}$ then

$$p(\beta \mid y) \propto \left(1 + \frac{(b-\beta)^2x'x}{(n-k)s^2}\right)^{-n/2}$$

with $z = \frac{\beta - b}{s^2(x'x)^{-1/2}}$, z has density $p(z \mid y, \sigma) \propto \left(1 + \frac{z^2}{n-1}\right)^{-n/2}$ and $z \sim t_{n-1}$. Again we can interpret the result by comparing the classical estimator and the bayesian estimator,

Classical	$\frac{b-\beta}{s(x'x)^{-1/2}} \sim t_{n-1}$
Bayesian	$\frac{\beta-b}{s(x'x)^{-1/2}} \mid y \sim t_{n-1}$

As MJM note: The marginal posterior can be used to make posterior inferences about subsets or functions of the parameter vector β without having to consider σ which is a nuisance parameter in this context. On the impact of the choice of the prior on this distribution, MJM note that the only difference is that the exponent in the expression of the posterior marginal distribution is $-(n-1)/2$ instead of$-n/2$. Thus the difference is negligeable when n is large.

5.2 An Informative Joint Prior Distribution

Assume the following Joint Prior Distribution,

$$p(\beta,\sigma) \propto \sigma^{-m} \exp\left\{-\frac{1}{2\sigma^2}\left(\eta + (\beta-\mu)'\Psi^{-1}(\beta-\mu)\right)\right\}$$

with $\eta > 0$, Ψ non singular symetric positive definite matrix. We call such a prior if you recall, a conjugate prior. According to MJM we have the following definition: A familiy of prior distributions, that when combined with the likelihood function via Bayes' theorem, result in a posterior distribution that is of the same parametric family of distributions as the prior distribution.

In practice, as MJM note: The analyst must specify the parameters of the prior distribution. This involves setting $m, \eta, \mu, \sigma^2, \Psi$. In this case, we use empirical Bayes methods which estimates those parameters from the data.

5.2.1 The Joint Posterior

Combining through Bayes' theorem the likelihood and the prior,

$$\begin{aligned} p(\beta,\sigma \quad | \quad y) \propto \sigma^{-(n+m)} \exp\left\{-\frac{1}{2\sigma^2}\left(\eta + (\beta-\mu)'\Psi^{-1}(\beta-\mu)\right)\right\} \times \\ \exp\left\{-\frac{1}{2\sigma^2}\left((b-\beta)'X'X(b-\beta) + (n-k)s^2\right)\right\} \end{aligned}$$

and after some manipulation can be rewritten as (see MJM because I could'nt do it myself!):

$$\begin{aligned} p(\beta,\sigma \quad | \quad y) &\propto \sigma^{-(n+m)} \exp\left\{-\frac{1}{2\sigma^2}\left((\beta-\beta^*)'\left(\Psi^{-1}+X'X\right)(\beta-\beta^*)+\xi\right)\right\} \\ & \textit{with} \\ \beta^* &= \left(\Psi^{-1}+X'X\right)^{-1}\left(\Psi^{-1}\mu + X'Xb\right) \\ \xi &= \eta + (n-k)s^2 + \mu'\Psi^{-1}\mu + b'X'Xb - \beta^{*\prime}\left(\Psi^{-1}+X'X\right)\beta^* \end{aligned}$$

We don't go further here because it gets very messy.

6 Asymptotics of Bayesian Estimators

We look at the asymptotic properties of the posterior which is treated in detail at page 673 in MJM. We remember in a classical context that $\sqrt{n}(\theta - \widehat{\theta}_{ML}) \rightarrow^d N(0, I(\theta_0)^{-1})$. In a Bayesian Context, we reverse everything,

$$\sqrt{n}(\underset{\text{random}}{\theta} - \widehat{\theta}_{ML}) \rightarrow^{d} N(0, p\lim I(\widehat{\theta}_{ML})^{-1})$$

In order to make the proof we need some notation. First denote $z = \sqrt{n}(\theta - \widehat{\theta}_{ML})$ and then $\theta = \frac{z}{\sqrt{n}} + \widehat{\theta}_{ML}$ and thus $p_z(z \mid x) = \frac{1}{\sqrt{n}} p(\frac{z}{\sqrt{n}} + \widehat{\theta}_{ML})$. Then $p(\theta \mid x) = \frac{1}{\sqrt{n}} p(\frac{z}{\sqrt{n}} + \widehat{\theta}_{ML} \mid x)$ and

$$p(\theta \mid x) \propto p(\frac{z}{\sqrt{n}} + \widehat{\theta}_{ML}) p(x \mid \frac{z}{\sqrt{n}} + \widehat{\theta}_{ML})$$

and if $x_1, x_2, ..., x_n$ is i.i.d.

$$p(\frac{z}{\sqrt{n}} + \widehat{\theta}_{ML}) p(x \quad \mid \quad \frac{z}{\sqrt{n}} + \widehat{\theta}_{ML}) =$$
$$p(\frac{z}{\sqrt{n}} + \widehat{\theta}_{ML}) \prod_{i=1}^{n} p(x_i \quad \mid \quad \frac{z}{\sqrt{n}} + \widehat{\theta}_{ML})$$

Now notice that if we take logs, we can probably use the mean value theorem (the instrument in asymptotics!) and apply the Central Limit theorem on the first term. Thus we concentrate on this expression. We have that

$$\ln \prod_{i=1}^{n} p(x_i \mid \frac{z}{\sqrt{n}} + \widehat{\theta}_{ML}) = \sum_{i=1}^{n} \ln p(x_i \mid \frac{z}{\sqrt{n}} + \widehat{\theta}_{ML}).$$

Now apply MVT (in fact Taylor approximation of degree 2 so we work with $\approx$) around $\widehat{\theta}_{ML}$,

$$\sum_{i=1}^{n} \ln p(x_i \quad \mid \quad \frac{z}{\sqrt{n}} + \widehat{\theta}_{ML}) \approx \underbrace{\sum_{i=1}^{n} \ln p(x_i \mid \widehat{\theta}_{ML})}_{\text{does not depend on } \theta}$$
$$\underbrace{+\frac{1}{\sqrt{n}} z' \sum_{i=1}^{n} \ln p_{\theta=\widehat{\theta}_{ML}}(x_i \mid \frac{z}{\sqrt{n}} + \widehat{\theta}_{ML})}_{\text{Score of ML evaluated at } \widehat{\theta}_{ML} \text{ thus } =0} + \frac{1}{2} z' \left(\frac{1}{n} \sum_{i=1}^{n} \ln p_{\theta\theta=\widehat{\theta}_{ML}}(x_i \mid \frac{z}{\sqrt{n}} + \widehat{\theta}_{ML}) \right) z$$

The last term converges to the population expectation of the gradient of the ML (applying consistency and continuity) thus $z' \left(\frac{1}{n} \sum_{i=1}^{n} \ln p_{\theta\theta=\widehat{\theta}_{ML}}(x_i \mid \frac{z}{\sqrt{n}} + \widehat{\theta}_{ML}) \right) z \rightarrow I(\widehat{\theta}_{ML})^{-1}$. Thus coming back to the posterior,

$$p(z \mid x) \underset{\approx}{\propto} p(\frac{z}{\sqrt{n}} + \widehat{\theta}_{ML}) \exp\left\{ \frac{1}{2} z' \left(\frac{1}{n} \sum_{i=1}^{n} \ln p_{\theta\theta=\widehat{\theta}_{ML}}(x_i \mid \frac{z}{\sqrt{n}} + \widehat{\theta}_{ML}) \right) z \right\}$$

now the first part converges to $\widehat{\theta}_{ML}$ since$plim\,(z) = \widehat{\theta}_{ML}$ is a consistent estimator. Thus this expression does not depend anymore on θ and thus again goes into the proportionality condition.. We have finally,

$$p(z \mid x) \underset{\approx}{\propto} \exp\left\{\frac{1}{2}z'\left(\frac{1}{n}\sum_{i=1}^{n}\ln p_{\theta\theta=\widehat{\theta}_{ML}}(x_i \mid \frac{z}{\sqrt{n}} + \widehat{\theta}_{ML})\right)z\right\}$$

under certain regularity conditions,

$$z \mid x \sim N(0, p\lim I(\widehat{\theta}_{ML})^{-1})$$

Again the interpretation is completely different. Regularity conditions are similar to those for ML and are listed in MJM p674.

7 Relation between the MSE and the Bayes Estimator

In a classical context, $\widehat{\theta}$ is some estimator of θ. We have that

$$MSE_{\widehat{\theta}}(\theta) = \mathbf{E}_\theta\left\{(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)'\right\} \text{ for } \theta \in \Theta.$$

for all θ the estimator has to be small. We choose the one that minimizes this difference.

$$MSE_{\widehat{\theta}}(\theta) = \int p(x \mid \theta)(\widehat{\theta}(x) - \theta)(\widehat{\theta}(x) - \theta)'dx$$

Expected loss over all possible outcomes of the data and the estimator. For the Bayes estimator with loss function $l(\theta, m) = (\theta - m)(\theta - m)'$. Let's consider the univariate case setting $\theta \in \mathbb{R}$. Then,

$$MSE_{\widehat{\theta}}(\theta) = \int p(x \mid \theta)(\widehat{\theta}(x) - \theta)^2 dx$$

Then $l(\theta, m) = (\theta - m)^2$ and we have the Bayes estimator,

$$\widehat{\theta} = \arg\min_m \int_\Theta p(\theta \mid x)(\widehat{\theta} - \theta)^2 d\theta.$$

Here the bayes estimator can always be computed. In the classical case, uniqueness of the estimator is not garantied. For some subset of the parameter space, one estimator may be superior while it is not the case for another subset where another estimator minimizes the MSE. Then instead of minimizing the MSE of each value of θ we can also find an estimator that minimizes the Expected MSE weighting each θ using the prior distribution.

We then find $\widehat{\theta}$ that minimizes $\int_\Theta MSE_{\widehat{\theta}}(\theta)p(\theta)d\theta$,

$$\begin{aligned}\int_\Theta MSE_{\widehat{\theta}}(\theta)p(\theta)d\theta &= \\ &\int_\Theta \int_X p_x(x \quad | \quad \theta)p(\theta)(\widehat{\theta}-\theta)^2 dxd\theta\end{aligned}$$

We see that the posterior is used up to a proportionality constant. Then minimizing Expected MSE in the classical sense gives the Bayes Estimator,

$$= \int_X c(x) \int_\Theta p(\theta \mid x)(\widehat{\theta}-\theta)^2 d\theta dx$$

where $c(x) = \left(\int_\Theta p(\theta)p(x \mid \theta)d\theta\right)^{-1}$ (Recall Bayes's theorem, divide by the marginal and goes into the proportionality constant). How to minimize? We see immidiately that the Bayes Estimator solve this problem. Thus the Bayes estimator is a powerful estimator when the classical estimators do not uniformaly minimize the MSE over the whole parameter space. Then minimizing the Expected MSE amounts to using the Bayes' estimator under a quadratic loss (the posterior mean).

8 Bayesian Inference

With Bayesian estimation we have the explicit form of the posterior and so we can proceed without test statistics to make inference. Then we don't talk of confidence interval but of credible region in the bayesian language. Indeed, we have that the θ is random and follows a distribution given the data. Thus using the CDF of this distribution we can specify credibility region.

8.1 Credible Region

We may want to define three types of credible region. First is an upperbound on some parameter while the second is defining a lower bound. Obviously we may also want to have both. We look at these three types sequentially.

Upper Bound — Let $\theta \in \Theta \subset \mathbb{R}$ for simplicity. Then a credibility region $(-\infty, u)$ with u an upper bound is chosen by setting a credibility level $1-\alpha$.

$$p\left\{\theta \in (-\infty, u) \mid x\right\} \geq 1-\alpha$$

Since θ follows a distribution (the posterior!), then we choose u such that we get the smallest interval with a probability $1-\alpha$. Let's consider an example. Suppose $\theta \mid x \sim$

$N(\mu_p, \sigma_p^2)$, then $\frac{\theta-\mu_p}{\sigma_p} \mid x \sim N(0,1)$ and thus using the CDF of the standard normal choose u such that $p\left\{\frac{\theta-\mu_p}{\sigma_p} < u \mid x\right\} = 1-\alpha$. Then the credible region is $(-\infty, \mu_p + \sigma_p u)$.

Lower Bound — Let $\theta \in \Theta \subset \mathbb{R}$ for simplicity. Then a credibility region $(l, +\infty)$ with l an lowerbound is chosen by setting a credibility level $1-\alpha$.

$$p\{\theta \in (l, +\infty) \mid x\} \geq 1-\alpha$$

Since θ follows a distribution (the posterior!), then we choose u such that we get the smallest interval with a probability $1-\alpha$. Let's consider an example. Suppose $\theta \mid x \sim N(\mu_p, \sigma_p^2)$, then $\frac{\theta-\mu_p}{\sigma_p} \mid x \sim N(0,1)$ and thus using the CDF of the standard normal choose u such that $p\left\{\frac{\theta-\mu_p}{\sigma_p} > l \mid x\right\} = 1-\alpha$. Then the credible region is $(\mu_p - \sigma_p l, +\infty)$.

Bounded Credible region — We want a credible region of the form (l, b). Bayesian don't do it the classical way. They choose the interval such that

1. $p\{\theta \in (l, b) \mid x\} \geq 1-\alpha$
2. $b-l$ is as small as possible.

For a symmetric distribution the second criteria is of no use since no amelioration can be made by displacing the interval to the right or to the left on the normal density. However, for an asymetric distribution, this second criteria tells us that the usual symetric interval may not be the one that is the smallest. Thus Bayesian credible region may be asymetric as we will see.

The sufficient condition for these two criteria to be respected is that the height of the probability distribution at the boundaries must be the same. Thus we choose the region sich that $f(b) = f(l) = c$ and thus we define the region as,

$$CR = \{\theta \in \Theta \mid p(\theta \mid x) \geq c\}$$

for the values of c giving $p\{p(\theta \mid x) \geq c\} = 1-\alpha$.

8.2 Hypothesis Testing

We have $H_0 : \theta \geq 0$, $H_a : \theta < 0$ with $\theta \in \Theta \subset \mathbb{R}$ for simplificty. Thus H_0 is true with probability

$$p\{\theta \in H_0 \mid x\} = p\{\theta \geq 0 \mid x\}$$

Then reject H_0 if $p\{\theta \geq 0 \mid x\} < \alpha$. Bayesian however like minimizing loss functions. Define

$$\begin{aligned} L_I &: = \text{loss from type 1 error} \\ L_{II} &: = \text{loss from type 2 error} \end{aligned}$$

What is the test procedure then? We want to minimize expected loss. We compute,

$$\begin{array}{l|l} \text{Reject } H_0 & L_I \times p\{\theta \geq 0 \mid x\} \\ \hline \text{Don't reject } H_0 & L_{II} \times p\{\theta < 0 \mid x\} \end{array}$$

Then we reject $H_0 \Leftrightarrow L_I \times p\{\theta \geq 0 \mid x\} < L_{II} \times p\{\theta < 0 \mid x\}$

$$\text{reject } H_0 \Leftrightarrow \frac{p\{\theta \geq 0 \mid x\}}{p\{\theta < 0 \mid x\}} < \frac{L_{II}}{L_I}$$

We call the ratio of type 1 to type 11 errors as the Posterior odds ratio. This is typicall of Decision theory. In the symetric case, $L_{II} = L_I$ then reject if $p\{\theta \geq 0 \mid x\} < \frac{1}{2}$. We can rewrite

$$\begin{aligned} p\{\theta \geq 0 \mid x\} &< \frac{L_{II}}{L_I} p\{\theta < 0 \mid x\} \\ &\Leftrightarrow \\ p\{\theta \geq 0 \mid x\} &< \frac{L_{II}}{L_I}(1 - p\{\theta \geq 0 \mid x\}) \\ p\{\theta \geq 0 \mid x\} &< \frac{\frac{L_{II}}{L_I}}{1 + \frac{L_{II}}{L_I}} \end{aligned}$$

we can think of $\frac{\frac{L_{II}}{L_I}}{1+\frac{L_{II}}{L_I}}$ as the α which implies for $\alpha = 0.05$ a ratio $\frac{L_{II}}{L_I} = 1/19$. Thus we see that this type of testing is not very different but still implies another philosophy about the testing of hypothesis.

9 Conclusion

These notes are very sketchy but still present the main ideas of Bayesian econometrics. The main disadvantage is that analytically it become quite messy and requires a torough knowledge of statistical theory. Before it becomes widely used by economist and practitionners, there is a long way to go. The classical paradigm is still dominent. We have seen that computing posterior is quite cumbersome. However, with the recent development of computers, many algorythms have been proposed to at least find the posterior distribution numerically. It remains however in the domain of the unknown for many practitioners with limited knowledge of statistical theory.