International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

An Informative Overview on Classical Inference and Bayesian Inference

Maitreya N. Acharya

Department of Statistics, Maharaja Krishnakumarsinhji Bhavnagar University, Bhavnagar, Gujarat, India.

Abstract: ***This paper is an informative article on the topic of Statistical Inference. Our main focus is on Classical Inference and Bayesian Inference.***

Keywords: Statistical Inference, Classical Inference, Bayesian Inference, Proper Priors, Improper Priors, Uninformative or Non-Informative or Diffuse Priors, Conjugate Priors, Loss Functions or Cost Functions, Symmetric Loss Functions, Asymmetric Loss Functions, Linex Loss Function, General Entropy Loss Function, Change Point.

1. Statistical Inference

First of all let us begin by thoroughly understanding the meaning of the word **"Statistical Inference"**. In simple words, we can say that **"Statistical Inference"** is the process in which the analysis of the data is carried out and on the basis of that statistical analysis, the properties of an underlying distribution are deduced or extrapolated or extended by inferring the unknown values from the trend values in the given data which is already known. This type of statistical analysis gives the inference about a population and its properties, which includes testing of hypotheses and deriving estimates on the basis of theory of estimation. Here, we note that the population to be considered for our case study should be assumed to be larger as compared to the observed set of data. In other words, we can say that the observed data is assumed to be sampled from population which is larger.

2. Classical Inference and Bayesian Inference

Let us now understand the concept of **"Classical Inference"** in broader sense. We consider a model $f(x,\theta)$ where 'θ' is a parameter and a fixed unknown quantity. We make inferences about 'θ' on the basis of information given in the sample under observation. This entire process and concept of making statistical inferences is known as **"Classical Inference"**. Now, a question arises in our mind that if this is classical inference, then what is **"Bayesian Inference"** and how **"Classical Inference"** is different from **"Bayesian Inference"**. For that, we need to understand the concept of **"Bayesian Inference"** in broader aspect. Most of the times, it is believed that subjective probabilities measuring degrees of belief are used to know about the value or values of unknown parameter 'θ'. Further, we note that these subjective probabilities are used to define what is called the **prior distribution** for the parameter 'θ', which is prior to sampling. In other words, we can say that the parameter 'θ' may be treated as a random variable with known prior distribution, say $g(\theta)$. In this case, the probability density (mass) function $f(x,\theta)$ can be denoted by a conditional probability notation $f(x|\theta)$ to stress the fact that *'f'* can only be used to compute probabilities for a given value of 'θ'. This clearly indicates that *'f'* conditionally depends on 'θ'. Here, we observe a random sample $\underline{x} = (x_1, x_2, ..., x_n)$. Also, the joint unconditional density for the sample $\underline{x}$ and the parameter 'θ' is $f(\underline{x}|\theta) \cdot g(\theta)$. Now, we shall use the famous Bayes Theorem to get the conditional density of 'θ' for the given sample $\underline{x}$ as under:

$$f(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\ \ g(\theta)}{\int_{\theta} f(\underline{x}|\theta)\ \ g(\theta)\ \ d\theta}$$

where, $\int_{\theta} f(\underline{x}|\theta)\ \ g(\theta)\ \ d\theta$ is the marginal density of *'x'*. Here, $f(\theta|\underline{x})$ is called **"Posterior Density"** of 'θ'. Thus, the foundation of Bayesian Inference is the posterior density $f(\theta|\underline{x})$ of 'θ'. It is quite clear and understandable that all the phases of the inference problem depend on this distribution. Further, we note that the posterior mean, the posterior median and the posterior mode are the **"Bayesian Point Estimates"** of 'θ'. The use of such point estimates entirely depends upon the nature of the posterior distribution, whether it is symmetric or asymmetric.

Now, let us understand the prior density and their nature and characteristics to understand the concept of Bayesian Inference in really well. The choice of prior density is most important as well as controversial in Bayesian Inference. Priors are generally classified as follows:

3. Proper Priors and Improper Priors

We need to focus on the concept of proper and improper prior distributions for better understanding. Let us first understand the definition of proper prior distribution. In general, a weight function that allocates positive weights to possible values of the parameters is termed as **"proper prior"**. The proper prior satisfies the definition of a probability mass function or a probability density function, whichever the case it may be, ie; (discrete or continuous). Now, we shall understand the concept of improper prior. Let us assume any weight function that sums or integrates over the possible values of the parameter to a value other than one say *'k'*, so if *f(x)* is uniform over the entire line from -∞

to ∞, $f(x) = k \;; -\infty < x < \infty \;, k > 0$. Then, it cannot be termed as a proper density. Since the integral $\int_{-\infty}^{\infty} f(x)\, dx = k \int_{-\infty}^{\infty} dx$ does not exist, we are least concerned how small *'k'* is. Such density functions are called **"Improper Prior Distributions"** or simply **"Improper Priors"**. A proper prior can be induced by an improper prior and thereby the function can be normalized if *'k'* is finite. Otherwise it continues to be improper and simply plays the role of a weighting function. In such cases, some adhoc method is used for determining a proper prior. The local behavior of the prior distribution in the region where likelihood is appreciable can be represented by these types of density functions, but their application is not possible in practice over its entire admissible range.

4. Uninformative or Non-Informative or Diffuse Priors

The words **"uninformative"** or **"non-informative"** or **"diffuse"** are enough to express the meaning of the nature of the prior. They also convey the vague or general information about the variable. The term **"uninformative prior"** clearly indicates that the prior might not be very informative and hence could be termed as a prior of objective nature, which is somewhat of a misnomer. In other words, the non-informative prior is the one that is not subjectively elicited. The uninformative priors can express only the **"objective"** information regarding the variable such as the nature whether the variable is positive or negative. It can also provide basic information about the variable's limitations if the variable is less than some limit. We can apply the rule of principle of indifference, which in practice is the simplest and oldest rule as far as determination of a non-informative prior is concerned. This rule assigns equal probabilities to all possibilities.

The non-informative priors are considered and referred in the cases when limited information or relatively little information is available apriori. There have been attempts to use the Bayesian approach even when no prior information or minimal prior information is available in practical situations. Such priors are often referred to as **"diffused priors"** or **"priors of ignorance"**. It is so because the priori information about the parameter is not considered substantial or enough to be provided by the sample of empirical data. It provides little information rather than a state of complete ignorance. In practical situations, when the experimenter is willing to accept the set of parameter values which are equally likely choices for the parameter, a non-informative prior is definitely considered. Here, we note that the experimenter is indifferent as far as choice of parameter and its values are concerned. This state of indifference may be expressed by taking a prior to be locally uniform. In general, an approximate prior is taken proportional to the square root of Fisher's Information. This rule is known as **"Jeffrey"s Rule"** and the related prior is known as **"Jeffrey"s Prior"**.

5. Conjugate Priors

In simple words, if the prior probability distribution $p(\theta)$ and the posterior distributions $p(\theta|x)$ belong to same family, then prior and posterior are then called **"conjugate distributions"** and the prior is called a **"conjugate prior"** for the given likelihood function. This is the general concept of conjugate prior in Bayesian Theory. Howard Raiffa and Robert Schlaifer were first to introduce the concept as well as the term "conjugate prior" in their work on Bayesian Decision Theory. George Alfred Barnard had also independently discovered a similar concept in his works.

Three properties have been put forth as desirable properties for conjugate families of distribution. These properties include mathematical tractability, richness and ease of interpretation. The prior distribution should reflect the statistician's prior information. This is how we can understand the property of richness. The prior should be possible to specify the conjugate family in such a way that it is readily interpreted by the person whose prior information is of interest. This is how we can understand the concept and property of interpretation.

Let us assume that the likelihood function is fixed. Usually, the likelihood function can be well-determined from a statement of the data-generating process. The algebraic forms of the prior and the posterior, generally with different parameter values, are same as far as certain choices of priors are concerned. Such a choice is a **"conjugate prior"**. There is specific quality about conjugate prior distributions is that they ease the computational burden when used as a prior distribution.

In other words, a conjugate prior is an algebraic convenience which gives a closed-form expression for the posterior. The numerical integration may become necessary in a case otherwise. Further, we can say that conjugate priors may give intuition showing how a likelihood function updates a prior distribution in a more transparent manner. Here, we shall note that all the members of the exponential family have conjugate priors.

6. Loss Functions

Let us now focus on the concept of **"Loss Functions"**. We need to understand the meaning, type and uses of loss functions to understand the concept well. In the field of Mathematics or Statistics, we come across this particular term **"loss function"**. **"Loss Functions"** are also referred to as **"Cost Functions"**. They are mostly used and applied in mathematical optimization, decision theory and machine learning. We can define a loss function or cost function as a function which is generally used for mapping an event or values of one or more variables onto a real number. It is quite obvious that the cost function intuitively representing some **"cost"** associated with that event. We need to minimize a loss function as far as optimization problem and the phenomenon of optimizing techniques is concerned. One thing is to be always kept in mind that a loss function or a cost function is always to be minimized, while an **"objective function"** which is either a loss function or its negative is to be

maximized. We note that the objective function can also be referred to as a **"reward function"** or a **"profit function"** or a **"utility function"** or a **"fitness function"** as far as the phenomenon of optimization theory is concerned in general.

6.1 Symmetric Loss Functions

Now, we shall understand the types of loss functions applicable in the Bayesian Theory of Estimation. Generally, loss functions are classified into two main categories, **"Symmetric"** and **"Asymmetric"**. Let us now understand the Squared Error Loss Function in broader sense. Following expression represents the Bayes estimator of a generic parameter or a function thereof 'α' under the squared error loss function **(SEL)**. We note that it is a continuous case, not discrete.

$$L_1(\alpha,d) \propto (\alpha-d)^2, \qquad \alpha,d \in \Re \quad (1)$$

We note that we can obtain the posterior mean when 'α' is a real valued parameter and 'd' is a decision rule to estimate 'α'. Consequently, the SEL function relates to an integer parameter given below

$$L_1'(m,v) \propto (m-v)^2, \qquad m,v = 0,1,2... \quad (2)$$

As a result, the Bayesian estimate of an integer-valued parameter under the SEL function (2) no longer remains the posterior mean and can be obtained by minimizing the corresponding posterior loss numerically. Thus, we can say that such a Bayesian estimate is equal to the nearest integer value to the posterior mean in general.

We get the other Bayes estimators of 'α' based on the loss functions using following expressions which represent the values of posterior median and posterior mode respectively.

$$L_2(\alpha,d) = |\alpha-d|,$$

$$L_3(\alpha,d) = \begin{cases} 0, & if\ |\alpha-d| < \varepsilon,\ \varepsilon > 0 \\ 1, & otherwise \end{cases}$$

6.2 Asymmetric Loss Functions

An unknown quantity or a generic parameter or a function thereof 'α' can be estimated using the loss function L(α,d). It can provide a measure of the financial consequences which arise on account of a wrong decision rule 'd'. The choice of the appropriate loss function is absolutely independent from the estimation procedure which is used. It entirely depends only on financial considerations. The economic considerations can be formulated through loss functions and the Bayes approach allows such considerations to be used in a rational manner. We can obtain the posterior expectation of the loss function by combining the selected loss function with the posterior density, and the 'd' value that minimizes the expected loss is defined as the **"Bayes Estimate"**, which is optimal relative to the loss function chosen. Expert studies have revealed that an overestimation of the reliability function is usually much more serious than an under estimation and this was the main reason why the use of symmetric loss functions was generally considered as inappropriate. That is why; **"Asymmetric Loss Functions"** have been used extensively and have been given large attention in recent times**.**

6.3 Linex Loss Function

In the year **1975**, a Statistician named **Varian** introduced a useful asymmetric loss function known as the "**Linex Loss Function"**. Linex Loss Function has a characteristic of rising approximately exponentially on one side of zero and approximately linearly on the other side. Linex Loss Function can be expressed as under as we assume that the minimal loss occurs at 'd':

$$L_4(\alpha,d) = \exp[q_1(d-\alpha)] - q_1(d-\alpha) - 1,\ q_1 \neq 0 \quad (3)$$

Here q_1 is the shape parameter. Its sign reflects the deviation of the asymmetry. It means that result shows $q_1>0$ if over estimation is more serious than under estimation and vice-a-versa. We note that the magnitude of q_1 clearly reflects the degree of asymmetry.

Varian also introduced the posterior expectation of Linex Loss Function which is given by $E_i[L_4(\alpha,d)] = \exp(q_1 d)E_i[\exp(-q_1\alpha)] - q_1(d-E_i(\alpha)) - 1$, i=1, 2.

Further, the Bayes estimates α_L^* is the value of 'd' that minimizes $E_i[L_4(\alpha,d)]$ and hence we can get the solution of the following equation:

$$\frac{\partial E_i[L_4(\alpha,d)]}{\partial d} = q_1 \exp(q_1 d)E_i\{\exp(-q_1\alpha)\} - q_1 = 0$$

This gives $e^{q_1\alpha_L^*} = [E_i\{\exp(-q_1\alpha)\}]^{-1}$ and

hence we get $\alpha_L^* = -\frac{1}{q_1}\ln[E_i\{\exp(-q_1\alpha)\}]$, i=1, 2 (4)

subject to $E_i\{\exp(-q_1\alpha)\}$ exists and is finite.

It was in the year **1991**, that **Basu and Ebrahimi** revealed that Linex Loss Function was not suitable for the estimation of the scale parameter and other quantities despite its flexibility and popularity for the location parameter estimation. It was for these reasons that **Basu and Ebrahimi** defined a **Modified Linex Loss Function** as under:

$$L_5(\alpha,d) = \exp[q_2(d\alpha^{-1}-1)] - q_2(d\alpha^{-1}-1) - 1 \quad (5)$$

Here, we note that the estimation error is expressed by $[(d/\alpha)-1]$, but we clarify that such type of modification **does not change** the characteristics of **Varian"s Linex Loss Function**.

The posterior expectation of the modified Linex Loss Function $L_5(\alpha,d)$ is given by following expression:

$E_i[L_5(\alpha,d)] = \exp(-q_2)E_i[\exp(q_2 d/\alpha)] - q_2 E_i(d/\alpha) + q_2 - 1$, i= 1, 2 and the value of 'd' that minimizes $E_i[L_5(\alpha,d)]$, say $\overset{*}{\alpha}_m$, is the solution of the equation given below:

$$\frac{\partial E_i[L_5(\alpha,d)]}{\partial d} = q_2 \exp(-q_2)E_i\left\{\frac{1}{\alpha}\exp\left(-\frac{q_2 d}{\alpha}\right)\right\} - q_2 E_i\left\{\frac{1}{2}\right\} = 0$$

Here, the condition is that all the expectations must be finite. Thus, $\overset{*}{\alpha}_m$ **cannot** be given in a closed form. It is quite obvious and certain that its evaluation involves iterative procedures.

6.4 General Entropy Loss Function

Calabria and Pulcini were first to propose a suitable alternative to the Modified Linex Loss Function by introducing another type of loss function namely **General Entropy Loss Function (GEL)**, in the year **1994**. It is expressed as,

$$L_6(\alpha,d) = \left(\frac{d}{\alpha}\right)^{q_3} - q_3 \ln\left(\frac{d}{\alpha}\right) - 1 \qquad (6)$$

Its minimum value occurs at d=α. This loss is referred to the generalization of the Entropy Loss Function. It was noted by several experts such as **Dey et. al.** in the year **1987** and also by **Dey and Liu** in the year **1992**. They noted that the shape parameter q_3 was equal to 1. The more general version shown in result (4) allows different shapes of the loss function to be taken into consideration. It was found that more serious consequences were caused than a negative error when $q_3 > 0$ and a positive error (d > α). The posterior expectation of the generalized Entropy Loss Function $L_6(\alpha,d)$ is given as

$E_i[L_6(\alpha,d)] = E_i[(d/\alpha)^{q_3}] - q_3 E_i[\ln(d/\alpha)] - 1$, i= 1, 2

Also, the value of 'd' that minimizes $E_i[L_6(\alpha,d)]$, say $\overset{*}{\alpha}_E$, gives the solution of the following equation.

$$\frac{\partial E_i[L_6(\alpha,d)]}{\partial d} = q_3 E_i\left\{d^{q_3-1}/\alpha^{q_3}\right\} - q_3 E_i\left\{d^{-1}\right\} = 0$$

Finally, we get

$d^{q_3} = \{E_i[\alpha^{-q_3}]\}^{-1}$ and

$$\alpha^*_E = \{E_i[\alpha^{-q_3}]\}^{-1/q_3} \qquad (7)$$

Here also $E_i[\alpha^{-q_3}]$ must exist and must be finite. Now, the General Entropy Loss Function (GEL) for an **Integer Parameter** is expressed as

$L'_6(m,v) \propto (v/m)^{q_3} - q_3 \ln(v/m) - 1, \qquad m,v = 0,1,2,...$ (8)

Finally, we can obtain the estimate „m" by means of the nearest integer value, say $\overset{*}{m}_{GE}$ after minimizing expectation $E_m[L'_6(m,v)]$ and using posterior distribution.

7. Change Point

Here, we shall understand the phenomenon of **"Change Point"** or **"Change Detection"** or **"Change Point Detection"**. The phenomenon of change point is observed in several situations in life testing and reliability estimation problems, in practical situations. Let us now understand the concept in simple words. In practical situations, it may happen that the sequence of failure times is observed at some point of time that results into instability. We need to study that when and where this change has started occurring. This point is called a change point and the related problem is called change point inference problem. A sequence of random variables $X_1, X_2, ..., X_n$ is said to have a change point at 'm' (1≤ m ≤ n). Here, $X_i \sim F_1(x|\theta_1)$ where i=1,2,...,m and $X_i \sim F_2(x|\theta_2)$ where i=m+1,2,...,n. Also, we note that $F_1(x|\theta_1) \neq F_2(x|\theta_2)$. Here, we shall consider the situation in which F_1 and F_2 have known functional form, but the change point 'm' is unknown. Bayesian approach and ideas may play an important role in study of such change point problems. Moreover, Bayesian theory has been often proposed as a valid alternative to classical estimation procedure. In the year 1970, D.V. Hinkley and E.A. Hinkley were first to accomplish the fundamental work in obtaining asymptotic distributions for a change point in normal and binominal sequence in using random walk results. The asymptotic approximations for maximum likelihood in change point models were studied by Gombay and Horvath in a series of research papers in 1990, 1994 and 1996. Further, we note that H. Boudjellaba, B. Macgibbon and P. Sawyer have studied and proposed the estimators and confidence intervals for change point in a Poisson sequence in the year 2001.

References

[1] **Hinkley. D. V. (1970).** Inference about the change point in a sequence of random variables, Biometrika, 57,1, 1-17.
[2] **Hinkley. D. V. and Hinkley. E. A. (1970).** Inference about the change point in a sequence of binomial variables, Biometrika, 57, 3, 477-488.
[3] **Zellner, A. (1971).** *An introduction to Bayesian Inference in Econometrics.* Wiley, New York.
[4] **Smith, A. F. M. (1975).** A Bayesian approach to inference about a change point in a sequence of random variables. Biometrika, 62, 407-416.

[5] **Holbert, D., Broemeling, L. D. (1977).** Bayesian inference related to shifting sequences and two-phase regression. *Comm. Statist A6 (3): 265-275.*

[6] **Berger, J. O. (1984).** The Bayesian view point in Robustness of Bayesian Inference, (J. B. Kadane, ed.), North-Holland Publ., Amsterdam. 63-124.

[7] **Abraham, B., Wei, W. W. S. (1984).** Inference about the parameters of a time series model with changing variance, Metrika 31: 183-194.

[8] **Bell, C. B. and Smith, E. P. (1986).** Inference for Non-negative Auto-regressive Schemes. *Communication Statistics- Theory Meth.,*15(8), 2267-2293.

[9] **McLachlan, G.J. and Basford, K. E. (1988).** *Mixture models: Inferences and Applications to clustering*, Marcel Dekker Inc.

[10] **Berger, J. O. (1990).** Robust Bayesian Analysis: Sensitivity to the prior, J. Stat. Planning and Inference, 25, 303-323.

[11] **Wikipedia.**