Available online at http://scik.org
J. Math. Comput. Sci. 2 (2012), No. 3, 624-636
ISSN: 1927-5307

THEORY OF IMPRECISE SETS: IMPRECISE RANDOMNESS AND BAYESIAN STATISTICS

HEMANTA K. BARUAH*

Department of Statistics, Gauhati University, Guwahati-781014, India

Abstract. If the realizations of a random variable are imprecise in the sense that two independent laws of randomness can define the presence level of values of the variable in a given interval, we would have to deal with the matters using the idea of imprecise randomness. In Bayesian methodology, inferences are made with reference to a probability law which is followed by a parameter that describes a probability law which is followed by a random variable. In this article, the mathematics of imprecise randomness has been introduced, and an attempt has been made to link the concept of imprecise randomness with that of Bayesian statistical inference.

Keywords: Imprecise uncertainty, Bayesian statistical inference .

2000 AMS Subject Classification: 00A30, 62A86, 62C10.

1. Introduction

There are two distinct paradigms of statistical inferences: Bayesian and non-Bayesian. In the Bayesian inferential procedure, it is assumed that a parameter describing a probability law of a random variable follows a probability law on its own, and the procedure is aimed at finding not just one single estimate of the parameter, but at making inferences

*Corresponding author

E-mail addresses: hemanta_bh@yahoo.com, hkb.gauhati@gmail.com (H. K. Baruah)

Received February 10, 2012

with reference to the probability law followed by the parameter (see for example [1, 2]). In the non-Bayesian or the conventional approach, one aims at finding one single estimate of the parameter. The Bayesian approach starts with a subjective *judgement* regarding the probability law followed by the parameter. This subjectivity however is something that has not been really accepted by the non-Bayesians since the beginning.

Impreciseness as a concept of uncertainty has been defined by the present author as that kind of uncertainty in which the elements in an interval are partially present [3]. It is different from the concept of fuzziness in the sense that unlike the fuzzy sets, the imprecise sets do conform to the classical measure theoretic as well as field theoretic formalisms. Such matters have been discussed in detail by the present author while defining fuzziness logically from his standpoint [4, 5, 6, 7].

When the observations of a random variable following a probability law are imprecise, we can call that variable an *imprecise random variable*. The randomness concerned would then be called *imprecise randomness*. In this article, we are going to introduce the *mathematics of imprecise randomness*, and thereafter we are going to discuss its link with the Bayesian statistical inferential methodology. We have in fact mentioned about this link between the two concepts in an earlier work [7, page 18]. In what follows, we would discuss certain very basic things in short about non-Bayesian and Bayesian matters. That would be followed by a discussion on how to make probabilistic conclusions from imprecise data. Finally, we shall make an attempt to link imprecise randomness and Bayesian principles.

2. Statistical Inference : Conventional and Bayesian

In conventional statistical analysis, it is presumed that observations $x_i, i = 1(1)n$, of a random variable X lie around an unknown parameter μ such that

$$x_i = \mu + \epsilon_i$$

where ϵ_i is an error following a probability law of errors. We are using the term random variable here in the statistical sense and not in the broader measure theoretic sense. The parameter here is presumed to be fixed, so that the probability law followed by the error variable can be translated to the random variable X. For estimation of the parameter, the three Gauss-Markov conditions for estimation of a linear parameter

(1) mathematical expectation of ϵ_i, $E(\epsilon_i) = 0$,

(2) variance of ϵ_i, $V(\epsilon_i) =$ a constant independent of i, and

(3) covariance of ϵ_i and ϵ_j , $E(\epsilon_i \epsilon_j) = 0$ for $i \neq j$,

are presumed to be followed.

For example, in a laboratory experiment if we need to measure the length μ of a rod, it would be presumed that no measuring device is ever perfect, and hence a number of observations x_i would be taken. The value of the sample mean computed from the observations thereafter would be taken as an estimate of μ. This is the essence of conventional statistical estimation of a linear parameter.

The Bayesian estimation principles are based on the Bayes' Theorem on conditional probability. In simple terms, it can be expressed as what follows. For two probabilistic events A and B, we trivially have the following relationship linking the conditional probability of happening of the event A given that the event B has already happened and the conditional probability of the event B given that the event A has already happened:

$$Prob(A|B)Prob(B) = Prob(B|A)Prob(A)$$

and therefore

$$Prob(A|B) = Prob(B|A)Prob(A)/Prob(B).$$

This innocuous looking identity has led to the Bayes' theorem on conditional probability of an event that had already happened, and based on this philosophy of finding the probability of a past event there ultimately came into existence a totally different paradigm of statistical inferential matters, that came to be known as Bayesian inference. For those who are not conversant with the Bayes' theorem, we would like to cite a classroom example of use of this identity first.

Assume that there are three identical boxes each having green and white balls in possibly different configurations. We can select a box at random, and draw a ball from it.

The probability of drawing a green ball, say, can thus be computed. Now assume that a ball has been drawn, and that the ball happens to be green. Bayes' theorem shows the way to find the probability that the second box, say, was chosen given that the ball drawn happens to be a green one. In the identity shown above, A can be the event of selecting the second box and B can be the event of drawing a green ball. The probability of happening of B given that A has happened, divided by the probability that B happens with reference to selecting every box, gives us the probability that A *had happened in the past* given that B *will happen in the future.*

As we have seen, this theorem gives the probability of an event that has already happened in the past. Evaluating the probability of a past event may not quite seem to be logical, but this identity is algebraically true anyway. Be that whatever, in the Bayesian inferential principles, it is presumed that the parameter to be estimated is not fixed; it lies in an interval, following some probability law on its own. In other words, here the parameter is taken to be a probabilistic variable on its own right, following what came to be known as a *prior* probability law giving an indirect reference to the past which is why there came up the nomenclature *Bayesian Inference.*

In our earlier example, the length of the rod must necessarily be fixed. So in that kind of a case, the Bayesian principles are not needed. Assume that we have some observations on some stock market related situation. In such a case, presumption of existence of a fixed value of a parameter may not actually be quite logical. Indeed in such a case, it can safely be presumed that the parameter lies in an interval, following some probability law. For example, the probability law followed by the parameter can be assumed to be uniform. More complicated prior probability laws are also in use making the question of subjectivity more prominent. We are not going to make any comment on acceptability or otherwise of such complicated prior probability laws in the Bayesian inferential matters. This is a debatable question, and we are not interested to enter into that at present.

3. Imprecise Numbers: Definitions and Notations

We are now going to define certain terms with reference to impreciseness [3].

Definition 3.1. An *imprecise number* $[\alpha, \beta, \gamma]$ is an interval around the real number β with the elements in the interval being *partially present.*

Definition 3.2. Partial presence of an element in an imprecise real number $[\alpha, \beta, \gamma]$ is described by the *presence level indicator function* $p(x)$ which is counted from the *reference function* $r(x)$ such that the presence level for any x, $\alpha \leq x \leq \gamma$, is $(p(x) - r(x))$, where $0 \leq r(x) \leq p(x) \leq 1$.

Definition 3.3. A *normal* imprecise number $N = [\alpha, \beta, \gamma]$ is associated with a presence level indicator function $\mu_N(x)$, where

$$\mu_N(x) = \begin{cases} \Psi_1(x), & \text{if } \alpha \leq x \leq \beta \\ \Psi_2(x), & \text{if } \beta \leq x \leq \gamma \\ 0, & \text{otherwise,} \end{cases}$$

with a *constant reference function* 0 in the entire real line. Here $\Psi_1(x)$ is continuous and non-decreasing in the interval $[\alpha, \beta]$, and $\Psi_2(x)$ is continuous and non-increasing in the interval $[\beta, \gamma]$, with

$$\Psi_1(\alpha) = \Psi_2(\gamma) = 0,$$

$$\Psi_1(\beta) = \Psi_2(\beta) = 1.$$

Here, the imprecise number would be characterized by $\{x, \mu_N(x), 0 : x \epsilon R\}$, R being the real line.

Definition 3.4. For a normal imprecise number $N = [\alpha, \beta, \gamma]$ with presence level indicator function

$$\mu_N(x) = \begin{cases} \Psi_1(x), & \text{if } \alpha \leq x \leq \beta \\ \Psi_2(x), & \text{if } \beta \leq x \leq \gamma \\ 0, & \text{otherwise,} \end{cases}$$

such that

$$\Psi_1(\alpha) = \Psi_2(\gamma) = 0,$$

$$\Psi_1(\beta) = \Psi_2(\beta) = 1,$$

with constant reference function equal to 0, $\Psi_1(x)$ is the *distribution function* of a *random variable* defined in the interval $[\alpha, \beta]$, and $\Psi_2(x)$ is the *complementary distribution function* of another random variable defined in the interval $[\beta, \gamma]$.

Definition 3.5. For a normal imprecise number $N = \{x, \mu_N(x), 0 : x \epsilon R\}$ as defined above, the complement $N^C = \{x, 1, \mu_N(x) : x \epsilon R\}$ will have constant presence level indicator function equal to 1, the reference function being $\mu_N(x)$ for $-\infty < x < \infty$.

We are using the term *random variable* here in the broader measure theoretic sense which does not require that the notion of probability needs to appear in defining randomness. What we mean is that a random variable need not be probabilistic, while a probabilistic variable is necessarily random.

Definition - 3.1 here is indeed the definition of a fuzzy number. Definition - 3.2 is based on a function of reference which is not there in the original definition of fuzzy sets. Definition - 3.3 again is indeed the definition of a normal fuzzy number with constant reference function equal to zero. Definition - 3. 4 expresses how the membership function of a normal fuzzy number should really have been defined based on two independent laws of randomness, with randomness defined in the measure theoretic sense. Here is where we have deviated from the Zadehian theory of fuzziness. The Zadehian theory assumes that one single law of *probability* can be formed from any given law of normal fuzziness and vice versa. This in our eyes is absolutely illogical as a normal law of fuzziness can be derived from two independent laws of *randomness* already [5, 6, 7]. Finally, Definition - 3.5 is on how to express the complement of a normal fuzzy set correctly [4]. The Zadehian definition of the complement of a fuzzy set is incorrect as it leads to a highly illogical conclusion that any statement and its complement do not complement each other.

4. Imprecise Randomness

We would like to start our discussions now with an example in which impreciseness is inherent. Consider the case of variability of temperature in a particular place [6]. It is obvious that both the minimum temperature and the maximum temperature in any place at any time are two probabilistic variables, and therefore they are random variables in the measure theoretic sense too. The two laws of randomness in this case would lead to an imprecise number [3]. In the case of variables related to rise and fall of price in stock exchanges, there always are two random variables in action, one for the minimum price, and the other for the maximum prices, giving rise to imprecise numbers. Now, if a random variable takes imprecise values, we shall have to deal with the situation using the mathematics of imprecise randomness.

Assume that X is a random variable following the normal probability law with location parameter μ and variance unity. Now if the location parameter is not fixed but imprecise instead, defined as

$$M = [\mu - \delta, \mu, \mu + \delta], \delta > 0$$

we would actually have to define an uncountably infinite number of normal probability density functions with the value of the location parameter ranging from $(\mu - \delta)$ to $(\mu + \delta)$ with maximum value of presence level assigned at the value μ.

Assume further that from this population, a sample of n imprecise observations around $x_1, x_2, \ldots, x_n$ has been drawn. We can then proceed to infer about the population, based on the sample data. We now proceed towards statistical analysis with reference to imprecise randomness.

The imprecise data are in terms of imprecise numbers around x_i, $i = 1, 2, \ldots, n$ defined as, say,

$$X_i = [x_i - \delta, x_i, x_i + \delta], \delta > 0.$$

The analysis can now proceed towards making an imprecise statistical conclusion. Without loss of generality, and for computational simplicity, such imprecise numbers can be taken as triangular imprecise numbers.

In this case, it is presumed that observations x_i, $i = 1(1)n$, of a random variable X lie around an unknown parameter μ such that

$$[x_i - \delta, x_i, x_i + \delta] = [\mu - \delta, \mu, \mu + \delta] + \epsilon_i,$$

where ϵ_i is an error following the normal probability law of errors.

Let the *normal* imprecise number

$$M = [\mu - \delta, \mu, \mu + \delta], \delta > 0$$

be associated with the presence level indicator function $\mu_M(x)$, where

$$\mu_M(x) = \begin{cases} \Psi_1(x), & \text{if } \mu - \delta \leq x \leq \mu \\ \Psi_2(x), & \text{if } \mu \leq x \leq \mu + \delta \\ 0, & \text{otherwise,} \end{cases}$$

with *constant reference function* 0 on the real line. In other words, we shall have for

$$X_i = [x_i - \delta, x_i, x_i + \delta]$$

the indicator function as

$$\mu_X(x) = \begin{cases} \Psi_1(x), & \text{if } x_i - \delta \leq x \leq x_i \\ \Psi_2(x), & \text{if } x_i \leq x \leq x_i + \delta \\ 0, & \text{otherwise,} \end{cases}$$

with the *constant reference function* 0 on the real line.

It can bee seen that from the distribution function $\Psi_1(x)$, for $x_i - \delta \leq x \leq x_i$, we shall get the density function

$$\frac{d\Psi_1(x)}{dx} = \varphi_1(x)$$

Similarly, from the complementary distribution function $\Psi_2(x)$, for $x_i \leq x \leq x_i + \delta$, we shall get the density function

$$\frac{d(1 - \Psi_2(x))}{dx} = \varphi_2(x).$$

Accordingly, a triangular imprecise number

$$X_i = [x_i - \delta, x_i, x_i + \delta]$$

with the presence level indicator function

$$\mu_X(x) = \begin{cases} \dfrac{(x - x_i + \delta)}{\delta}, & \text{if } x_i - \delta \leq x \leq x_i \\ \dfrac{(x_i + \delta - x)}{\delta}, & \text{if } x_i \leq x \leq x_i + \delta \\ 0, & \text{otherwise,} \end{cases}$$

would be defined by two laws of randomness with distribution functions

$$F_1(x) = \frac{(x - x_i + \delta)}{\delta}, \text{if } x_i - \delta \leq x \leq x_i,$$

and

$$F_2(x) = 1 - \frac{(x_i + \delta - x)}{\delta}, \text{if } x_i \leq x \leq x_i + \delta,$$

so that their densities

$$\frac{dF_1(x)}{dx} = \frac{1}{\delta}, \text{if } x_i - \delta \leq x \leq x_i,$$

and

$$\frac{dF_2(x)}{dx} = \frac{1}{\delta}, \text{if } x_i \leq x \leq x_i + \delta$$

are uniform.

That is how the question of imprecise randomness should come up. There should be a variable following some law of randomness. In an interval around every realization of the random variable, there should be impreciseness. The conclusions arrived at from analysis of such data will also be in terms of impreciseness.

As soon as we surmise that the data are imprecise, or in other words that the data are of the interval type with an appropriately defined presence level indicator function, we presume that there is one law of randomness in $[\mu - \delta, \mu]$ while there is another law of randomness in $[\mu, \mu + \delta]$. The readers would note here that for probabilistic conclusions based on imprecise random data, we however would need to define two probability laws in the statistical sense in the two intervals $[\mu - \delta, \mu]$ and $[\mu, \mu + \delta]$.

We now cite a numerical example [7]. Suppose we start with hypothetical imprecise data of the type $[x - 0.5, x, x + 0.5]$ with an assumption that the data are triangular. The random variable X of which this x is a realization in the sample has been assumed to be normally distributed. In other words, we have started with an assumption that the two probability laws, one on $[x-0.5, x]$ and the other on $[x, x+0.5]$, are uniform, for a normally distributed realization x with mean μ and error variance σ^2, say. Assume that we finally arrived at an imprecise value of the F - statistic for 3 and 6 degrees of freedom, say, with the following presence level indicator function:

$$\mu_F(x) = \begin{cases} \dfrac{174.303x}{(157.733x + 17.992)}, & \text{if } 0 \leq x \leq 0.482 \\ \dfrac{86.858}{(78.866 + 16.570x)}, & \text{if } 0.482 \leq x < \infty. \end{cases}$$

Accordingly, the following two distribution functions would decide the imprecise number concerned:

$$\Psi_1(x) = \frac{174.303x}{(157.733x + 17.992)}, \text{for } 0 \leq x \leq 0.482$$

and

$$(1 - \Psi_2(x)) = 1 - \frac{86.858}{(78.866 + 16.570x)}, \text{for } 0.482 \leq x < \infty.$$

In other words, the imprecise number $F = [0, 0.482, \infty)$ with $\Psi_1(F)$ and $\Psi_2(F)$ defined in the intervals $0 \leq F \leq 0.482$ and $0.482 \leq F < \infty$ respectively, would be defined by the two densities $d(\Psi_1(F))/dF$ and $d(1 - \Psi_2(F))/dF$ in the respective ranges.

In the non-imprecise situation, we would have concluded that there is no reason to reject the null hypothesis at 5% probability level of significance as the computed value of $F(= 0.482)$ is smaller than the theoretical value of $F(= 4.7571)$ for 3 and 6 degrees of freedom. We now proceed to look into the matters of making an imprecise conclusion statistically. The theoretical non-imprecise value of $F(= 4.7571)$ to the right of which the area under the probability density function of F is 0.05, is on that part of the interval on which

$$\mu_F(x) = \frac{86.858}{(78.866 + 16.570x)}, \text{if } 0.482 \leq x < \infty,$$

has been defined. We note that the presence level for $F = 4.7571$ is $\Psi_2(4.7571)$. Therefore, in this case, the probability density function concerned would be given by

$$\varphi_2(F) = \frac{1439.23706}{(78.866 + 16.570F)^2}, \text{ for } 0.482 \leq F < \infty.$$

Hence

$$Prob[F \geq 4.7571] = \text{the area under } \varphi_2(F)$$

from $F = 4.7571$ to infinity, which is the area of the right tail beyond 4.5751. The area of the left tail from 0.482 to 4.5751 is $(1 - \Psi_2(4.7571))$. Thus the area of the right tail is $\Psi_2(4.7571)$ again, which is nothing but the presence level of F at 4.7571.

$$\Psi_2(4.7571) = 0.5508$$

is therefore the probability that the imprecise null hypothesis would have to be rejected at 5% probability level of significance. In other words, when a non-rejectable hypothesis is made imprecise, there will still be a probability that the imprecise hypothesis would actually be found rejectable.

In the same way, if a rejectable hypothesis is made imprecise, there would still be a probability that the imprecise hypothesis would be found non-rejectable, the probability of non-rejection in our example being decided by

$$\mu_F(x) = \frac{174.303x}{(157.733x + 17.992)}, \text{ if } 0 \leq x \leq 0.482,$$

this time.

5. Imprecise Randomness versus Bayesian Inference

We now proceed to discuss our standpoint to link imprecise randomness and Bayesian statistics. At this point, we would like to mention two important things regarding *fuzzy* randomness. We would like to remind the readers that the definition of fuzziness does not explain in what way the membership function of a fuzzy number can be constructed. Indeed, in the Dubois - Prade nomenclature, the fuzzy membership function of a normal fuzzy number can be explained in terms of a left reference function, which is nothing but our $\Psi_1(x)$, and in terms of a right reference function, which is nothing but our $\Psi_2(x)$. However, in the literature on fuzziness, other than the present author's works, there is no reference to randomness separately with reference to $\Psi_1(x)$ and $\Psi_2(x)$. Based on an assumption that there can be framed one law of probability from a law of fuzziness, the works in the mathematics of fuzziness are continuing since the beginning. This was one reason why we have proposed a different name for this type of uncertainty [3]. For two reasons, fuzzy randomness and imprecise randomness are entirely different things. First, in terms of imprecise randomness, we can put forward probabilistic interpretations of the conclusions; this is simply not possible in the case of fuzzy randomness because in the Zadehian definition of fuzziness there is no mention that the membership function of a normal fuzzy number can be explained with the help of two independent laws of randomness. Secondly, for testing the rejectability of a fuzzy hypothesis, the workers have always referred to the alternative hypothesis as the complement of the fuzzy set defined in the currently available manner. Such a definition of complementation is not correct [4, 7]. Therefore in every case of fuzzy statistical hypothesis testing available in the literature, the alternative hypotheses must necessarily be redefined. Hence, fuzzy randomness with fuzziness defined in the Zadehian manner and imprecise randomness are two totally different things. However, fuzzy randomness from our standpoint [7] and imprecise randomness are indeed the same.

In the case of the classical Bayesian statistical mathematics, a prior probability law is assumed with reference to a parameter associated with the law of probability of the variable concerned. Assume for example, that a variable follows the normal probability law of errors with location parameter μ. Now if this μ is presumed to follow the uniform probability law for example, in the interval $[a, b]$, the Bayesian inferential procedures are there to analyse the data concerned. It is another matter that the decision of what type of probability law would be enforced on the parameter μ is of course based on subjective judgement.

In imprecise randomness, we have two laws of *randomness* defining impreciseness of a parameter associated with a random variable following some probability law of errors. We have shown that imprecise numbers can actually be constructed [6], and therefore one need not really *presume* impreciseness, which in other words means that the question of subjectivity need not really arise here.

We have mentioned earlier that not in all situations the Bayesian principles can be applied. In our earlier example of measuring the length of a rod, the question of applying Bayesian estimation does not arise. In the same way, in this case, the data can never be imprecise, and therefore the question of applying the mathematics of imprecise randomness too does not arise in this case. Indeed when the parameter is probabilistic, we can go for Bayesian inferential matters, and when the parameter is imprecise, we can go for the analysis using imprecise randomness.

Finally, it can be seen that just the distribution function $\Psi_1(x)$ defined in some interval $[a, b]$ can define an imprecise number of the type $[a, b, b]$ already. Therefore imprecise randomness with impreciseness defined by numbers of the type $[a, b, b]$ can be seen to be similar to using the Bayesian methodology in analysing data. Use of the arithmetic of impreciseness in this case would anyway be easier than using a prior probability law to draw statistical inferences.

References

[1] D. V. Lindley, the Present Position in Bayesian Statistics, Statist. Sci. 5, 1990, 44 - 89.

[2] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, Bayesian Data Analysis, Chapman & Hall, Florida, 1995.

[3] Hemanta K. Baruah, An Introduction to the Theory of Imprecise Sets: the Mathematics of Partial Presence, Journal of Mathematical and Computational Sciences, Vol. 2, No. 2, 2012, 110-124.

[4] Hemanta K. Baruah, Towards Forming a Field of Fuzzy Sets, International Journal of Energy Information and Communications, Vol. 2, Issue 1, 2010, 16 - 20.

[5] Hemanta K. Baruah, In Search of the Root of Fuzziness: The Measure Theoretic Meaning of Partial Presence, Annals of Fuzzy Mathematics and Informatics, Vol. 2, No. 1, 2011, 57 - 68.

[6] Hemanta K. Baruah, Construction of the Membership Function of a Fuzzy Number, ICIC Express Letters, Vol. 5, Issue 2, 2011, 545-549.

[7] Hemanta K. Baruah, The Theory of Fuzzy Sets: Beliefs and Realities, International Journal of Energy Information and Communications, Vol. 2, Issue 2, 2011, 1 - 22.