Chapter 2. Literature Review

A statistical test is generally conducted by means of a hypothesis testing from which the probability distribution is determined by the assumption that the null hypothesis H_0 is true. Under the significant level α, a critical region is computed such that if the observed statistics falls in the critical region, we reject the null hypothesis. The statistical test provides information from which we can decide the significance of the increase (or decrease) in any experimental results.

However some vague information, formulated by terms from natural language, is not easy to describe in statistical terms. To handle this information and knowledge, it is natural to use intelligent computing techniques. Fortunately, in many expositions of intelligent computing methods, fuzzy techniques are described as an alternative to a more traditional statistical approach. Such a description makes fuzzy techniques difficult to be understood and accepted for researchers who are accustomed to statistical methods.

Although the statistical methods used to testing fuzzy sample mean are based on the traditional decision theory, which is to extend Neyman-Peason's lemma about most powerful test, they are not austerely investigated for two reasons: (i) soft-computing for critical region with fuzzy number is still not identified; (ii) distribution for fuzzy population are vague, incomplete or unknown.

In this paper we first use new conception on testing hypotheses of fuzzy mean and fuzzy variance with interval data. About testing, the method and theory are complete in classical statistics for those data and parameters in point form. It is not easy to handle interval data. For example, $f(x)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$ is a density function for normal distribution data *x*, but on the other hand, provided data are intervals, how do you know if the data fit normal distribution and what $\mu=[a,b]$, $\sigma^2=[c,d]$, in $f(x)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$ means? Another question is how to apply traditional testing hypothesis method such as *t*-test and *F*-test in testing interval data. In this paper, we propose extended conception to test mean and variance, which are intervals.

It is known that there are two directions in testing. One is that we assume the data are normally distributed. The other one is nonparametric method. The sample data is single value, which is called point. Test statistics are precise value and the theory is complete. But in reality, usually when we do the survey work, the vague answers may appear. For example, when you ask a job seeker how much he demands per month. The answer always appears an interval form such as two thousand dollars to three thousand dollars will be O.K. For traditional statistics, it is a failure data. If all data or most data are vague, then the traditional testing method fails. Another situation is that the Minister of Finance announces that the salary of fresh university graduate is two thousand dollars per month in average. It seems that this value cannot properly show real situation. In different way, if he says that the demanding salary of fresh university graduates is about two thousand dollars to two and half thousand dollars, it would be more acceptable for such data form. The next problem is to examine whether to accept or reject this interval data [2000, 2500]. This problem becomes more and more eager to be solved. In this paper we try to apply traditional method by extended conception to do the testing work. The idea is we treat each interval as point and test by confidence interval. When using *t*-test or *F*-test, we need to assume sample data are normal. So, we first assume these interval data are fuzzy normally distributed.

Secondly, we will discuss how to test interval data in another point of view. How do we make a statistical decision process when the data illustrates the interval type? In this paper we proposed an innovative distribution, i.e. fuzzy normal for the interval sample. The sample mean and sample variance of interval data on population with fuzzy normal distribution are discussed in detail. A testing hypothesis processing and decision rule for the parameters are suggested by extending the conventional *t*-test and *F*-test procedures. The empirical studies demonstrate that our proposed decision rules are quite efficient and realistic.

Until now there are seldom papers that have mentioned about how to test with interval data. Almost all papers aim at testing fuzzy numbers. This is our innovation.

At last we introduce the concept of fuzzy statistics via discrete fuzzy sample and continuous fuzzy sample. The definitions of fuzzy mean for two kinds of fuzzy data are proposed. Using these definitions we are able to set up the fuzzy testing hypothesis such as fuzzy equals to and the fuzzy belongs to. These testing processes are very useful tools for decision make in a fuzzy system.

Since our main objective is to promote the understanding of these two classes of techniques - statistics and fuzzy - to researchers who may only know well one of theses techniques, we go into some detailed explanation of the basic techniques. Any reader who is well familiar with one or both of these techniques is advised to at least browse through or exposition of their basics.

In reality, more complex situations are possible, in which an expert is not 100% sure whether a given estimate x_i is possible; in this case, we can no longer use polling to get numerical characteristics of the expert knowledge. Fuzzy set methodology can handle such more complex situation as well. The above polling method of eliciting the values of the membership functions is only one of the many known elicitation techniques (Wu and Yang, 1998; Dubois and Prade, 1991; Nguyen and Wu, 2000; Liang and Wang, 1991; Stojakovic, 1994).

In this content, we present a class of fuzzy statistical decision process in which testing hypothesis can be naturally reformulated in terms of interval-valued statistics. To describe these situations we will start with a brief motivation of traditional statistical techniques, and then give a brief motivation of the corresponding fuzzy methods, and then describes the relation between these two classes of techniques. We provide the definitions of fuzzy mean, fuzzy distance as well as investigation of their related properties. We also give some empirical examples to illustrate the techniques and to analyze fuzzy data. Empirical studies show that fuzzy hypothesis testing with soft computing for interval data is more realistic and reasonable in the social science research.

Our result is in good agreement with a general one from Nguyen and Wu (2000), according to which an arbitrary fuzzy set can be interpreted in statistical terms: namely, as a random set. For the latest developments in this area (Goutsias, Mahler, and Nguyen, 1997), we hope that this reformation will make the corresponding fuzzy techniques more acceptable to researchers whose only experience is in using traditional statistical methods. Also, in Chapter 5, we refer Nguyen and Wu(2006) to do the testing work in symbols and fuzzy number's property.

Fuzzy testing hypotheses was discussed by many researchers, including Delgado et al. (1985), Saade and Schwarzlander (1990), Saade (1994), Watanabe and Imaizumi (1993), Kruse (1982), as well as Kruse and Meyer (1987, 1988) who considered the problem of testing vague hypotheses in the presence of vague data. Grzegorzewski (2000) proposed the grade of acceptability of null and alternative hypothesis. Also, in Grzegorzewski (2001) a definition of a fuzzy test for testing hypothesis with vague data was proposed and then two methods for fuzzy test, defuzzification, based on the maximum value and randomize operators was also proposed.

Recently, in testing fuzzy data, Körner (2000) and Montenegro et al. (2004) have presented asymptotic one-sample procedures. Montenegro et al. (2000, 2001) have developed asymptotic methods for the two-sample case. Montenegro et al. (2004a) have discussed the asymptotic ANOVA test. Montenegro et al. (2004b) say their bootstrap techniques are more valuable at testing mean. However, for accurate data, it seems not so valuable.

Gil et al. (2006) extended the bootstrap study from the generalized metric by Körner and Näther (2002). Körner (2000) followed the method proposed by Kruse and Meyer (1987) for those sets of normal compact convex fuzzy subsets of R^n denoted by F_c.

Moreover, Körner showed that each fuzzy set $A \in F_c$ corresponds uniquely to its support function

$$s_A(\alpha, u) = \sup\{\langle u, a \rangle : a \in A^{\alpha}\}, \quad u \in S^{n-1}, \quad \alpha \in [0,1]$$

where S^{n-1} is the $(n-1)$-dimensional unit sphere of R^n and $\langle a,b\rangle$ is the inner product of the Euclidean space R^n. He defined a metric on F_c by the L_2-metric of Lebesgue integrable function

$$d_2(A,B)=\| s_A - s_B \|_2 = [n\int_0^1\int_{S^{n-1}} |s_A(\alpha,u)-s_B(\alpha,u)|^2 \ \mu(du)d\alpha]^{1/2} \qquad (2.1)$$

for all $A,B \in F_c$ and the norm

$$\| A \|_2 = \| s_A \|_2 = [n\int_0^1\int_{S^{n-1}} |s_A(\alpha,u)|^2 \ \mu(du)d\alpha]^{1/2}.$$

On the other hand, Diamond and Kloeden (1994), have shown that the space (F_c, d_2) is a complete separable space.

The approach of Fréchet (1948), handles with an expectation and a variance in metric space. The expectation of a random element X in M is defined by the set of all elements $A \in M$ with

$$d^2(X,A) = \inf_{B\in M} E[d^2(X,B)] \qquad (2.2)$$

If there is an $A \in M$ with $E[d^2(X,A)] < \infty$. He called the infimum of (2.2) the variance of the random element X.

Körner defined the expectation and the variance of a fuzzy random variable X in the metric space of normal compact convex fuzzy sets equipped with the metric (2.1), i.e.

$$E_F X = \{A \in F_c : Ed_2^2(X,A) = \inf_{B\in F_c} Ed_2^2(X,B)\}$$

and

$$Var(X) = \inf_{B\in F_c} Ed^2(X,B)$$

We note that in Gil et al. (2006), the method can not work on so-called discrete fuzzy numbers and non-normal fuzzy numbers.

For the purpose of fluent process, we do the testing work with three parts independently.