2. Fuzzy Statistic Analysis

In order to show that fuzzy chi-square test statistic for goodness-of-fit, we must come back to see the famous Pearson's χ^2. How to get the Pearson's χ^2 is the most important thing for us to find out the fuzzy χ^2.

2.1 Chi-square Test Statistic for Goodness-of-Fit

Goodness-of-fit is an important subject in statistics. It means how well a statistic model fits a set of observations. Measures of goodness-of-fit typically summarize the difference between observed values and the expected values under the model in question. Such measures can be used in statistical hypothesis testing.

Now, we consider the famous Pearson's chi-square test. The underlying concepts are expounded in [3].

Let Y_1 be $B(n, p_1)$, where $0 < p_1 < 1$. By the central limit theorem, we can get that

$$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}} \tag{2.1}$$

has a distribution that is approximately $N(0,1)$ when n is large. We have already known that $Q_1 = Z^2$ is approximately $\chi^2(1)$. If we let $Y_2 = n - Y_1$ and $p_2 = 1 - p_1$, we see that Q_1 may be written as

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_1 - np_1)^2}{n(1-p_1)}.$$

Since

$$(Y_1 - np_1)^2 = [(n - Y_2) - n(1-p_2)]^2 = (Y_2 - np_2)^2,$$

we have

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}.$$

Let us now carefully consider each term in this last expression for Q_1. Of course, Y_1 is the number of "successes," and np_1 is the expected number of "successes"; that is, $E(Y_1) = np_1$. Similarly, Y_2 and nP_2 are, respectively, the number and the expected number of "failures." So each numerator consists of the square of a difference of the observe number and expected number.

Note that Q_1 can be written as

$$Q_1 = \sum_{i=1}^{2} \frac{(Y_i - np_i)^2}{np_i}, \tag{2.2}$$

and we have seen intuitively that it has an approximate chi-square distribution with

one degree of freedom. In a sense, Q_1 measure the "closeness" of the observed numbers to the corresponding expected numbers.

To generalize, we let an experiment have k mutually exclusive and exhaustive outcomes, say $A_1, A_2, \ldots, A_k$. Let $p_i = P(A_i)$ and thus

$$\sum_{i=1}^{k} p_i = 1.$$

The experiment is repeated n independent times, and we let Y_i represent the number of times the experiment results in A_i, $i = 1, 2, \ldots, k$. The joint p.d.f. is as following:

$$f(y_1, y_2, \ldots, y_{k-1}) = \frac{n!}{y_1! y_2! \cdots y_k!} p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}, \tag{2.3}$$

where $y_1, y_2, \ldots, y_{k-1}$ are nonnegative integers such that $y_1 + y_2 + \cdots + y_{k-1} \le n$. Note that we do not need to consider Y_k since once the other $k-1$ random variables are observed to equal $y_1, y_2, \ldots, y_{k-1}$, respectively, we know that

$$Y_k = n - y_1 - y_2 - \cdots - y_{k-1} = y_k, \text{ say.}$$

This joint distribution is a straightforward generalization of the binomial distribution.

Pearson then constructed an expression similar to Q_1 (see equation (2.2)), which involves Y_1 and $Y_2 = n - Y_1$, that we denote by Q_{k-1}, which involves $Y_1, Y_2, \ldots, Y_{k-1}$, and $Y_k = n - Y_1 - Y_2 - \cdots - Y_{k-1}$, namely,

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}. \tag{2.4}$$

He argued that Q_{k-1} has an approximate chi-square distribution with $k-1$ degrees of freedom in much the same way we argue that Q_1 is approximately $\chi^2(1)$. People can see the detail of proof in Arnold, 1990 [1].

We shall now show how we can use the fact that

$$Q_{k-1} \text{ is approximately } \chi^2(k-1)$$

to test hypotheses about probabilities of various outcomes. Let an experiment have k mutually exclusive and exhaustive outcomes, $A_1, A_2, \ldots, A_k$. We would like to test whether $p_i = P(A_i)$ is equal to a known number p_{i0}, $i = 1, 2, \ldots, \ldots k$. That is, we shall test hypothesis

$$H_0 : p_i = p_{i0}, \quad i = 1, 2, \ldots, \ldots k.$$

In order to test such a hypothesis, we shall take a sample of size n, that is repeating the experiment n independent times. We tend to favor H_0 if the observed number of times that A_i occurred, say y_i, and the number of times A_i was expected to occur if H_0 was true, namely np_{i0}, are approximately equal. That is, if

$$q_{k-1} = \sum_{i=1}^{k} \frac{(y_i - np_{i0})^2}{np_{i0}}$$

is “small,” we tend to favor H_0. Since the distribution of

$$Q_{k-1} \text{ is approximately } \chi^2(k-1),$$

we shall reject H_0 if

$$q_{k-1} \geq \chi_\alpha{}^2(k-1),$$

where α is the desired significance level of the test.

2.2 Fuzzy Set Theory and Fuzzy Numbers

Fuzzy set theory is treated of fuzziness in data, which was proposed by Zadeh in 1965 [21]. For the fuzzy set theory, the grade of membership can be taken as a value between 0 and 1. Although in the normal case of set theory, the grade of membership can be taken only as 0 or 1. We define the fuzzy set theory in the following.

Definition 2.1 Fuzzy set theory (Zimmermann, 1996 [12])

If X is a collection of objects denoted generically by x, then a fuzzy set $\tilde{A}$ in X is a set of ordered pairs:

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) \mid x \in X\}$$

$\mu_{\tilde{A}}(x)$ is called the membership function or grade of membership(also degree of compatibility or degree of truth) of x in $\tilde{A}$ that maps X to the membership space M (when M contains only the two points 0 and 1, $\tilde{A}$ is nonfuzzy and $\mu_{\tilde{A}}(x)$ is identical to the characteristic function of a nonfuzzy set). The range of the membership function is a subset of the nonnegative real numbers whose supremum is finite.Elements with a zero degree of membership are normally not listed.

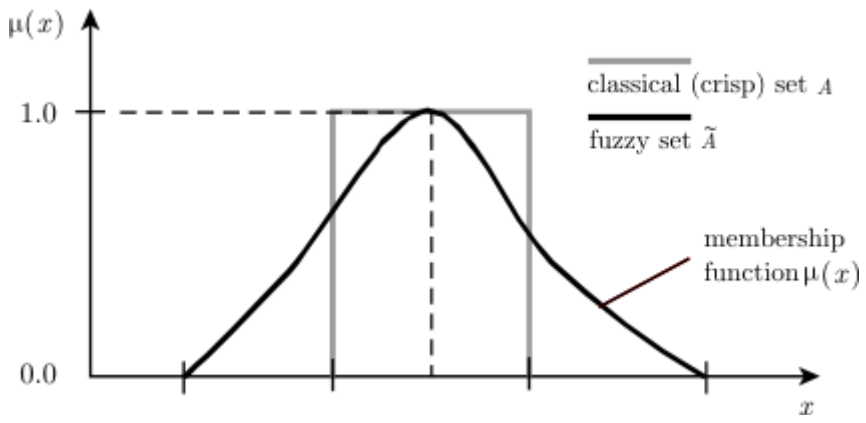

Figure 2.1. Fuzzy set and crisp set.

In considering the question related with fuzzy property, we consider that the information itself has the uncertainty and fuzzy property. The following definitions for fuzzy numbers will be used in the following text for simplicity.

Definition 2.2 Fuzzy Numbers (Nguyen and Wu 2006 [8])

Let U be an universal set,

$$A = \{A_1, A_2, \ldots, A_n\}$$

be the subset of discussion factors in U. For any term or statement X on U, it's membership corresponding to $\{A_1, A_2, \ldots, A_n\}$ is

$$\{\mu_1(X), \mu_2(X), \ldots, \mu_n(X)\}, \text{ here } \mu : U \to [0,1]$$

is a real function. If the domain of the universal set is discrete, then the fuzzy number of X can be written as following:

$$\mu_U(X) = \sum_{i=1}^{n} \mu_i(X) I_{A_i}(X) \qquad (2.3)$$

Where $I_{A_i}(x) = 1$, if $x \in A_i$; $I_{A_i}(x) = 0$, if $x \notin A_i$.

If the domain of the universal set is continuous, then the fuzzy number can be written as :

$$\mu_U(X) = \int_{A_i \subseteq A} \mu_i(X) I_{A_i}(X) \qquad (2.4)$$

Note that, in many writings, people are used to write a fuzzy number as

$$\mu_U(X) = \frac{\mu_1(X)}{A_1} + \frac{\mu_2(X)}{A_2} + \cdots + \frac{\mu_n(X)}{A_n}$$

(*where "+" stands for "or", and "$\frac{\cdot}{\cdot}$" stands for the membership $\mu_i(X)$ on A_i*) *instead of*

$$\mu_U(X) = \sum_{i=1}^{n} \mu_i(X) I_{A_i}(X).$$

In next section, we will introduce how to answer questionnaire in fuzzy numbers.

2.3 Fuzzy Sampling Surveys

In social science research, many decisions, evaluations, or purposes of evaluations are done by surveys or questionnaires to seek for people's consensus. The commonly used method is asking people to think in binary logic way from a multiple choice design. However, these processes often ignore the fuzzy concept, sometimes even ambiguous thinking behavior perceived in human logic and recognition. For instance, when people need to answer questions from the survey which lists five choices

including “Very satisfactory,” ”Satisfactory,” “Normal,” “Unsatisfactory,” “Very unsatisfactory,” traditional survey become quite exclusive. If people can use membership function to express the degree of their feelings based on their own choices, the answer presented will be closer to real human thinking.

Example 2.1

The use of fuzzy numbers in a sampling survey about which place do you want to go abroad?

Consider a fuzzy set of place as show in Table 2.1. Note that in the extreme cases when a degree is given 1 or 0, which is the five places are of a standard “Yes” and “No” in complement relationship, as in binary logic.

Table 2.1. *Response in fuzzy numbers*

Voter	Japan	America	Europe	Taiwan	China
1	0.8	0.2	0	0	0
2	0	0	0	0	1
3	0.5	0	0.5	0	0
4	0.6	0.1	0.3	0	0
5	0	0	0	1	0
6	0	0	0.8	0.2	0
7	0.5	0	0	0.2	0.3
Total	2.4	0.3	1.6	1.4	1.3

Table 2.2. *Response in traditional way*

Voter	Japan	America	Europe	Taiwan	China
1	1	0	0	0	0
2	0	0	0	0	1
3	1	0	0	0	0
4	1	0	0	0	0
5	0	0	0	1	0
6	0	0	1	0	0
7	1	0	0	0	0
Total	4	0	1	1	1

From Table 2.1, we see that there are five places to choice. If we use traditional way to answer, then we just only can choice one place (see Table 2.2). Furthermore, if we can answer with fuzzy numbers, it will be approximate our thinking. In Table2.1,

we let that A_1 represents the semantics of “Japan”, A_2 represents the semantics of “America”, A_3 represents the semantics of “Europe”, A_4 represents the semantics of “Taiwan”, and A_5 represents the semantics of “China”. Then fuzzy numbers for these seven statements can be represented as following:

$$\mu_{A_1}(X) = 0.8I_1(X) + 0I_2(X) + 0.5I_3(X) + 0.6I_4(X) + 0I_5(X) + 0I_6(X) + 0.5I_7(X)$$

$$\mu_{A_2}(X) = 0.2I_1(X) + 0I_2(X) + 0I_3(X) + 0.1I_4(X) + 0I_5(X) + 0I_6(X) + 0I_7(X)$$

$$\mu_{A_3}(X) = 0I_1(X) + 0I_2(X) + 0.5I_3(X) + 0.3I_4(X) + 0I_5(X) + 0.8I_6(X) + 0I_7(X)$$

$$\mu_{A_4}(X) = 0I_1(X) + 0I_2(X) + 0I_3(X) + 0I_4(X) + 1I_5(X) + 0.2I_6(X) + 0.2I_7(X)$$

$$\mu_{A_5}(X) = 0I_1(X) + 1I_2(X) + 0I_3(X) + 0I_4(X) + 0I_5(X) + 0I_6(X) + 0.3I_7(X)$$

From the above we can see that sampling survey with soft computing are more realistic and reasonable for the social science research.

We have introduced how to get the Pearson’s χ^2 and how to get the fuzzy sampling survey. So that people can fill the membership which is more likely their thinking. Hence, we have lots of fuzzy sample data. Now, we are interested in the hypothesis that the proportion of individuals who choose the j th category is the same or not for the different categories. However, traditional statistics reflect the result from a two-valued logic opinion. If we can find out a way to handle the fuzzy sample data, it will be good for us to work in social science survey. In next section, we begin to define some new fuzzy functions, and then to show that the fuzzy χ^2.