Chapter 3. Testing on Interval Data (I)

In this chapter, we will present the character of interval data, and then define the fuzzy mean and fuzzy variance for interval data. The testing hypothesis procedure for interval data will be provided with the illustrative examples.

It is well known (see Arnold(1991)) that the $(1-\alpha)$ confidence interval for population mean μ is $(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}; \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}})$, for X_i $i = 1,...,n$ be independent, with $X_i \sim N(\mu,\sigma^2)$. We try to do the testing hypothesis with bounded closed intervals.

Also, it is well known that the optimal test for population mean is *t*-test (see Lemann(1959)), optimal test for population variance is χ^2-test and so on. That is why we try to apply these methods to define the extended confidence intervals.

3.1. Fuzzy Mean and Fuzzy Variance

3.1.1. Definitions and properties

We describe every interval by two values, named midpoint and radius. For example, an interval data $[3,5]$ will be denoted as $(4;1)$ where the first element is the center of the interval and the second is the radius of the interval, which are calculated as follows:

$$4 = (3+5)/2 \quad \text{and} \quad 1 = (5-3)/2$$

The representation is essential dealing an interval data as a single point.

Definition 3.1.1. Fuzzy Sample Mean

Let $(x_1; r_1), (x_2; r_2), \cdots, (x_n; r_n)$ *be* n *sample interval data, then the sample mean* $\overline{\boldsymbol{X}}_{\boldsymbol{I}}$ *is defined as*

$$(\frac{x_1 + \cdots + x_n}{n}; \frac{r_1 + \cdots + r_n}{n}).$$

By Moore's definition (1979), given $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2]$, $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2]$ two intervals then

$$\boldsymbol{A} + \boldsymbol{B} = [\boldsymbol{a}_1 + \boldsymbol{b}_1, \boldsymbol{a}_2 + \boldsymbol{b}_2]$$

$$\boldsymbol{A} - \boldsymbol{B} = [\boldsymbol{a}_1 - \boldsymbol{b}_2, \boldsymbol{a}_2 - \boldsymbol{b}_1]$$

$$\boldsymbol{A} \times \boldsymbol{B} = [\mathbf{min}(\boldsymbol{a}_1\boldsymbol{b}_1, \boldsymbol{a}_1\boldsymbol{b}_2, \boldsymbol{a}_2\boldsymbol{b}_1, \boldsymbol{a}_2\boldsymbol{b}_2), \mathbf{max}(\boldsymbol{a}_1\boldsymbol{b}_1, \boldsymbol{a}_1\boldsymbol{b}_2, \boldsymbol{a}_2\boldsymbol{b}_1, \boldsymbol{a}_2\boldsymbol{b}_2)]$$

$$\frac{\mathbf{1}}{\boldsymbol{A}} = [\frac{\mathbf{1}}{\boldsymbol{a}_2}, \frac{\mathbf{1}}{\boldsymbol{a}_1}]$$

for $\boldsymbol{A}$ be an interval not containing 0 and $m[a, b] = [ma, mb]$, for $m > 0$ and

$$\frac{\boldsymbol{A}}{\boldsymbol{B}} = \boldsymbol{A} \times \frac{\mathbf{1}}{\boldsymbol{B}}.$$

Moreover, Moore defines the *absolute value* of an interval $\boldsymbol{A}$ by

$$|\boldsymbol{A}| = \mathbf{max}(|\boldsymbol{a}_1|, |\boldsymbol{a}_2|)$$

and

$$\boldsymbol{A} < \mathbf{B} \text{ if and only if } \boldsymbol{a}_2 < \boldsymbol{b}_1$$

and the *distance* between $\boldsymbol{A}$ and $\mathbf{B}$ is

$$\boldsymbol{d}(\boldsymbol{A}, \boldsymbol{B}) = \mathbf{max}(|\boldsymbol{a}_1 - \boldsymbol{b}_1|, |\boldsymbol{a}_2 - \boldsymbol{b}_2|).$$

However, if $\boldsymbol{A} \cap \boldsymbol{B} \neq \phi$ then we are not able to compare $\boldsymbol{A}$ and $\mathbf{B}$. Therefore this method is not so useful for other situations.

Furthermore, for example, given *n* sample data 3, 3, …, 3 we can easily compute the sample variance is, and we have 0. On the other hand, if *n* sample data are

$$\boldsymbol{I}_1 = \boldsymbol{I}_2 = \mathbf{...} = \boldsymbol{I}_n = [\mathbf{1}, \mathbf{100}],$$

then by the above definition we have

$$\boldsymbol{d(I_i, I_j) = 0 \quad for \quad i \neq j} .$$

In this chapter we use similar but not same conception to define sample variance.

After the above discussion, we define fuzzy sample variance as follows:

Definition 3.1.2. Fuzzy Sample Variance

Let $(x_1; r_1), (x_2; r_2), \cdots, (x_n; r_n)$ *be* n *sample interval data, then sample variance* $\boldsymbol{S_I^2}$ *is defined as*

$$\left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})^2 ; \frac{1}{n-1} \sum_{i=1}^{n} (r_i - \overline{r})^2 \right)$$

and the sample standard deviation is defined as

$$\boldsymbol{S_I} = \left(\sqrt{\frac{\boldsymbol{1}}{\boldsymbol{n-1}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \overline{\boldsymbol{X}})^2} ; \sqrt{\frac{\boldsymbol{1}}{\boldsymbol{n-1}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{r_i} - \overline{\boldsymbol{r}})^2} \right)$$

Theorem 3.1.1. Let $[a_i, b_i] = (x_i; r_i)$, $i = 1, 2, \cdots, n$ be n interval samples, where $x_i = \frac{a_i + b_i}{2}$, $r_i = \frac{b_i - a_i}{2}$. Then the *sample mean* have the following property:

$$[\frac{1}{n} \sum_{i=1}^{n} a_i , \frac{1}{n} \sum_{i=1}^{n} b_i] = (\frac{1}{n} \sum_{i=1}^{n} x_i ; \frac{1}{n} \sum_{i=1}^{n} r_i)$$

Proof*:* Using simple algebra operation, we have

$$(\frac{1}{n} \sum_{i=1}^{n} x_i ; \frac{1}{n} \sum_{i=1}^{n} r_i) \quad = \quad \frac{\boldsymbol{1}}{\boldsymbol{n}} (\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_i} ; \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{r_i})$$

$$= \frac{1}{n} [\sum_{i=1}^{n} (x_i + r_i), \sum_{i=1}^{n} (x_i - r_i)]$$

$$=\frac{1}{n}[\sum_{i=1}^{n}(\frac{a_i+b_i}{2}-\frac{b_i-a_i}{2}),\sum_{i=1}^{n}(\frac{a_i+b_i}{2}+\frac{b_i-a_i}{2})]$$

$$=\frac{1}{n}[\sum_{i=1}^{n}a_i,\sum_{i=1}^{n}b_i]$$

$$=[\frac{1}{n}\sum_{i=1}^{n}a_i,\frac{1}{n}\sum_{i=1}^{n}b_i]$$

This completes the proof. ❑

3.2. Interval's Extended Confidence Interval

Usually, under normal assumption we use *t*-test for testing mean and *F*-test for testing two populations' variance. Since we do not know population variance, the $1-\alpha$ confidence interval for population mean μ is $(\overline{X}-e,\overline{X}+e)$ where $e=t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$. Further more, the $1-\alpha$ confidence interval for two populations' mean μ_1, μ_2 is

$$(\overline{X}_1-\overline{X}_2-e,\overline{X}_1-\overline{X}_2+e)$$

where

$$e=t_{K,\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

and $K=\min(n_1-1,n_2-1)$. The $1-\alpha$ confidence interval for two populations' variance is

$$(F_{n_1-1,\ n_2-1,\ 1-\alpha/2},\ F_{n_1-1,\ n_2-1,\ \alpha/2})$$

where $F=S_1^2/S_2^2$. For vague data, such as interval data, there is no $1-\alpha$

confidence interval just like classical statistics. We try to extend the concept of traditional confidence interval to interval data.

3.3. Testing Hypotheses about Mean and Variance with Interval Data

3.3.1. Extended Concept

In testing, we suppose population is normally distributed otherwise nonparametric method should be used in classical statistics. In this chapter, we extend traditional testing method to test data in interval form.

Given *n* sample data $(c_1; r_1), ..., (c_n; r_n)$, these data will firstly be tested to see what distribution they fit separately. That is we need to realize what distribution these two sets of crisp data $c_1, c_2, ..., c_n$ and $r_1, r_2, ..., r_n$ belong to. In this literature we assume they all fit normal for simplicity. At this condition, we say the population is fuzzy normally distributed.

Until now, there is no standard way to test interval data. We propose extended classical testing method to do this work. The idea is we treat intervals as points and test with the help of traditional way. This method also has what we call fuzzy extended $1-\alpha_I$ confidence interval to make decisions whether to accept or reject null hypothesis.

For testing mean of population of fuzzy interval data, we set up the following process:

1. Null Hypothesis: $\boldsymbol{H_0 : \mu_I = \mu_{I_0}}$

$$\boldsymbol{H_1 : \mu_I \neq \mu_{I_0}} \quad \text{where} \quad \boldsymbol{\mu_{I_0} = [h, k]}$$

2. Testing Statistics: $\boldsymbol{\overline{X}_I}$

3.Under the significant level $\boldsymbol{\alpha_I}$, we use extended $\boldsymbol{1-\alpha_I}$ confidence interval

$$\boldsymbol{(\overline{X}_I - e_I, \overline{X}_I + e_I)}$$

where $\boldsymbol{e_I = t_{n-1,\alpha/2} \frac{S_I}{\sqrt{n}}}$.

If $[h, k] \subset \boldsymbol{(\overline{X}_I - e_I, \overline{X}_I + e_I)}$ then we accept $\boldsymbol{\mu_I = \mu_{I_0}}$.

.

Note that $\boldsymbol{(\overline{X}_I - e_I, \overline{X}_I + e_I)}$ is an interval calculated by the above method. We also allow the interval to be closed. This wouldn't affect final result.

Remark 3.3.1: We show how the testing rules being manipulated. For example, $\boldsymbol{\overline{X}_I = [a,b]}$, $\boldsymbol{e_I = [c,d]}$ then the extended $\boldsymbol{1-\alpha_I}$ confidence interval is (m,n) where $m = a-d$, $\boldsymbol{n = b+d}$. If $[h,k] \subset (m,n)$ we accept null hypothesis $\boldsymbol{\mu_I = [h,k]}$.

Next, for testing two populations' mean, we first review classical method: Usually, we do not know about two populations' variance σ_1^2, σ_2^2 then we accept null hypothesis $\mu_1 = \mu_2$ if $|\overline{X}_1 - \overline{X}_2| < e$, where $e = t_{K,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ and $K = \min\{n_1 - 1, n_2 - 1\}$. That is, if $\overline{X}_2 - e < \overline{X}_1 < \overline{X}_2 + e$, then we accept the null hypothesis $\mu_1 = \mu_2$.

For testing two population’s mean we set up the following process:

1. Null Hypothesis: $\boldsymbol{H_0 : \mu_{I_1} = \mu_{I_2}}$

   $\boldsymbol{H_1 : \mu_{I_1} \neq \mu_{I_2}}$

2. Testing Statistics: $\boldsymbol{\overline{X}_{I_1}}$, $\boldsymbol{\overline{X}_{I_2}}$

3. Under the significant level $\boldsymbol{\alpha_I}$, we use extended $\boldsymbol{1-\alpha_I}$ confidence interval

$$\boldsymbol{(\overline{X}_{I_2} - e_I, \overline{X}_{I_2} + e_I)}$$

where $\boldsymbol{e_I = t_{K,\alpha/2}\sqrt{\frac{S_{I_1}^2}{n_1} + \frac{S_{I_2}^2}{n_2}}}$ and $\boldsymbol{\overline{X}_{I_2}}$, $\boldsymbol{S_{I_1}^2}$, $\boldsymbol{S_{I_2}^2}$ are all intervals calculated from definition 4.1.1 and 4.1.2.

If $\boldsymbol{\overline{X}_{I_1} \subset (\overline{X}_{I_2} - e_I, \overline{X}_{I_2} + e_I)}$ then we accept $\boldsymbol{\mu_{I_1} = \mu_{I_2}}$.

Similarly, let us observe $1-\alpha$ confidence interval

$$(F_{n_1-1,\,n_2-1,1-\alpha/2}, F_{n_1-1,\,n_2-1,\,\alpha/2})$$

of two populations’ variance σ_1^2, σ_2^2 where $F = S_1^2 / S_2^2$. We accept null hypothesis $\sigma_1^2 = \sigma_2^2$ if

$$S_2^2 F_{n_1-1,\,n_2-1,1-\alpha/2} < S_1^2 < S_2^2 F_{n_1-1,\,n_2-1,\,\alpha/2}$$

For testing two population’s variance we set up the following process:

1. Null Hypothesis: $\boldsymbol{H_0 : \sigma_{I_1}^2 = \sigma_{I_2}^2}$

   $\boldsymbol{H_1 : \sigma_{I_1}^2 \neq \sigma_{I_2}^2}$

2. Testing Statistics: $\boldsymbol{S_{I_1}^2}$, $\boldsymbol{S_{I_2}^2}$ calculated from definition 3.2.2.

3. Under the significant level $\boldsymbol{\alpha_I}$, we use extended $\boldsymbol{1-\alpha_I}$ confidence interval

$$(S_{I_2}^2 F_{n_1-1,n_2-1,1-\alpha/2}, S_{I_2}^2 F_{n_1-1,n_2-1,\alpha/2}).$$

If $S_{I_1}^2 \subset (S_{I_2}^2 F_{n_1-1,n_2-1,1-\alpha/2}, S_{I_2}^2 F_{n_1-1,n_2-1,\alpha/2})$ then we accept $\sigma_{I_1}^2 = \sigma_{I_2}^2$.

The above rules are all concentrate on testing whether equal relation about mean and variance holds or not. If equal relation does not hold then we compare them. Another situation we should avoid is that do not let interval length of $\overline{X}_{I_2}$ ($S_{I_2}^2$) larger than $\overline{X}_{I_1}$ ($S_{I_1}^2$).

For example, let $\overline{X}_{I_1} = [4, 6]$, $\overline{X}_{I_2} = [1, 6]$, $e_I = [1, 1]$ then $(\overline{X}_{I_2} - e_I, \overline{X}_{I_2} + e_I) = (0, 7)$. We must conclude that the two populations' mean are equal. As a matter of fact, this conclusion is not reasonable. This is one of differences between classical statistics and fuzzy statistics.

3.4. Illustration Examples

Example 3.4.1. T-type car's owner pronounces that the cars' oil consumption is ten to twelve miles per liter. Some consumer's magazine intends to investigate their car about quality. Since most consumers come from lower income family, the magazine aims at oil consumption and stability of quality. It is clear that a car will consume more oil at city than at highway. So the investigator chooses ten persons who drive T-type car randomly. Next step, he keeps six resembling drivers in driving habit and abandons the other data. He gets six data

[10, 12], [10.5, 11], [13, 14], [9, 10], [15, 16], [8.5, 9]

which show miles per liter. After computation, we get sample variance which is equal to (7.8; 0.3). Interval variation is 0.3 and median variation is 7.8, although interval lengths are roughly equal but median variation is high, we conclude that the quality is not stable. Next, for testing population's mean, we first compute sample mean and get [10.9, 12] and then take $\boldsymbol{\alpha_I = 0.05}$ to get $\boldsymbol{e_I = [0.3, 6.8]}$. Since the extended $\boldsymbol{1-\alpha_I}$ confidence interval is

$$\boldsymbol{(\overline{X}_I - e_I, \overline{X}_I + e_I) = (4.1, 18.8)}$$

and $[10,12] \subset (4.1, 18.8)$, we accept the null hypothesis $\boldsymbol{\mu_I = [10, 12]}$.

Example 3.4.2. There are two communities *X* and *Y*, we would like to compare their income level to determine a sale strategy. We randomly choose five data from community *X* and community *Y*. The data are listed as follows:

X: [3, 4], 4, [3.5, 5], [3.8, 4.2], 4.2 (ten thousands)

Y: [2, 10], 6, 2, [3, 8], [4, 7] (ten thousands).

After simple calculation, we get

$$\boldsymbol{\overline{X}_I = (4; 0.3)}, \quad \boldsymbol{\overline{Y}_I = (5.3; 1.6)}$$

We set $\boldsymbol{\alpha_I = 0.1}$ then the extended $\boldsymbol{1-\alpha_I}$ confidence interval is

$$\boldsymbol{(\overline{X}_I - e_I, \overline{X}_I + e_I) = (0.8, 7.2)}$$

Since $(5.3; 1.6) = [3.7, 6.9] \subset (0.8, 7.2)$, we conclude that the two communities' income level are equal. Next we compute *X*'s and *Y*'s sample variances, then we have

$$\boldsymbol{S_{I_X}^2 = (0.4; 0.4)}, \quad \boldsymbol{S_{I_Y}^2 = (2.9; 3.3)}$$

Taking the significance level $\boldsymbol{\alpha_I = 0.1}$ then the extended $\boldsymbol{1-\alpha_I}$ confidence interval is

$$(\boldsymbol{S_{I_X}^2 F_{4,4,0.95}, S_{I_X}^2 F_{4,4,0.05}) = (0, 5.1)}$$

Since $\boldsymbol{S_{I_Y}^2 \not\subset (0, 5.1)}$, we conclude that the two communities' variances are not equal. Clearly, $\boldsymbol{S_{I_X}^2 < S_{I_Y}^2}$, we find community *Y*'s variance is larger than *X*'s. This situation shows that people live in community *Y* has unstable income but live in community *X* is stable. Therefore, we suggest selling middle price goods in community *X* and selling low price on most kinds of goods but high price on few kinds of goods.

Example 3.4.3. A radio factory wants to purchase special type tube to manufacture high definition stereo preamplifier. Only *X* and *Y* two brands fit for this type. Since it is big money purchase, the manager compares their quality in the following way. At first, in one year he collects 10 tubes from each brand randomly per month. So he gets 120 tubes each brand. Next, he tests all 240 tubes for using life. At last, after computing these data he gets

$$\boldsymbol{\overline{X}_I = (2.54; 0.1)}, \quad \boldsymbol{\overline{Y}_I = (2.53; 0.09)}$$

and

$$\boldsymbol{S_{I_X}^2 = (0.3; 0.24)}, \quad \boldsymbol{S_{I_Y}^2 = (0.3; 0.235)}$$

where the unit is ten thousand hours. Since $\boldsymbol{\overline{X}_I}$ is almost equal to $\boldsymbol{\overline{Y}_I}$ and $\boldsymbol{S_{I_X}^2}$ is almost equal to $\boldsymbol{S_{I_Y}^2}$. We conclude that the two brands' qualities are almost on the same level. So, the manager decides to buy cheaper one.