SET-VALUED STATISTICS AND RANDOM SETS

Wang Peizhuang*
Liu Xihui**

*Beijing Normal University
**China Academy of Building Research

ABSTRACT

Based upon the theory of fall-shadow of random sets, this paper proposes a new model of statistics — set-valued statistics, which is a generalization from classical statistics and fuzzy statistics. As applications of set-valued statistics, the methods of fall-shadow smoother have been presented.

INTRODUCTION

In the classical statistics, each trial gives us a definite point in the phase space (the set of possibly observational values). If this confinement may be relaxed and the result of each trial is an ordinary or a fuzzy set in the phase space, then such a kind of experiment is called experiment of set-valued statistics.

Set-valued statistics are suitable for those measurement processes which possess both randomness and fuzziness. Such a kind of statistics can partly or completely contain psychological measurement, so shows very good prospects of application to various areas.

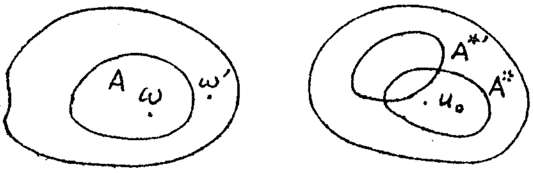
Extension like this of statistical method was starting from fuzzy stastistics. Zhang Nanlun(10), Ma Mouchao and Cao Zhiqiang(3) proposed for the purpose to establish membership function some statistical methods, which are in fact set-valued ones. E. Sanchez and author then proposed theory of fuzzy fall shadow of random sets under such background. Our work was similar to that of Goodman, but ours possesses more concrete and clear background, and a more perfect theoretical framework.

In a review, D. Dubois and H. Prade pointed out: "The above discussion suggests that fuzzy set and possibility theory can be interpreted in the framework of random sets (Goodman,Wang and Sanchez)"(1)It is just these two authors who proposed a random test model with noncrisp occurences, which is in fact a kind of set valued statistics.

TWO KINDS OF DUAL STATISTICAL MODELS

There exists some duality between the fuzzy statistical experiment model and clas-

A: fixed event
ω: basic event, movable



u : fixed element
A*: approximate of A, movable

Fig. 1

sical one. As shown in Fig. 1 the classical statistical experiment may be viewed as one where "the cycle is fixed, the point is movable", the fuzzy statistical model (generally, set-valued statistics) may be viewed as experiment where "the point is fixed and cycle is movable".

Starting from the duality mentioned here, we can transfer a kind of statistical model into another by means of interchanging "cycles" and "points". Denote

$$\dot{u} \triangleq \{A \mid U \supset A \ni u\} \quad (u \in U), \tag{2.1}$$

$$\dot{U} \triangleq \{\dot{u} \mid u \in U\} \tag{2.2}$$

The cycle A in U is a point in $\mathcal{P}(U)$, and the point u in U responds to a cycle $\dot{u}$ (an ultra-filter in Boolean algebra $\mathcal{P}(U)$). So a set-valued statistics model on U can be transformed into a classical statistics model on $\mathcal{P}(U)$.

For a given σ-field $\check{\mathcal{B}}$ containning $\dot{U}$ on $\mathcal{P}(U)$, a mapping from some probability field $(\Omega,\mathcal{F},P)$ to $(\mathcal{P}(U),\check{\mathcal{B}})$

$$\xi : \Omega \longrightarrow \mathcal{P}(U) \tag{2.3}$$

$$(\xi^{-1}(\mathcal{C}) = \{\omega \mid \xi(\omega) \in \mathcal{C}\} \in \mathcal{F} \qquad (\forall \mathcal{C} \in \check{\mathcal{B}})), \tag{2.4}$$

is called a random set on U (which is a random variable on $\mathcal{P}(U)$). $\mathcal{S}(\Omega,\mathcal{F};\ \mathcal{P}(U),\check{\mathcal{B}})$ denotes the set of all such random sets.

Each trial of probabilistic statistical experiment is an occurence of a random variable, however each trial of set-valued statisticla experiment is an occurence of a random set.

The distribution of random set ξ is $P_\xi(\mathcal{C})$, $\mathcal{C}$ is ergodic over $\check{\mathcal{B}}$. It is difficult to give it an expression. To express the constraint of P_ξ in U is easier.

Definition 2.1. Suppose $\xi \in \mathcal{S}(\Omega,\mathcal{F};\ \mathcal{P}(U),\check{\mathcal{B}})$, denote

$$\mu_\xi(u) \triangleq P(\omega \mid \xi(\omega) \in \dot{u}) = P(\omega \mid \xi(\omega) \ni u), \tag{2.5}$$

it is called the fall-shadow or covering function of ξ. The fall-shadow of random set is the most important numerical function. In fuzzy statistics it is the membership function of correspondent fuzzy subset. Many fundamental formula of random variable may be expressed in the theory of random set by the aid of fall-shadow.

SET-VALUED STATISTICS

Discussed here is only those set-valued statistics which shows an ordinary subset after each trial. For a giving $\xi \in \mathcal{S}(\Omega,\mathcal{F};\ \mathcal{P}(U),\check{\mathcal{B}})$, make n times independent observations, and get sample

$$x_1,\ x_2,\ \cdots,\ x_n \quad (x_i \in \mathcal{P}(U),\ i = 1,\cdots,n) \tag{3.1}$$

Regardless of concrete observed results, they are abstractly considered as a group of independent sets having the same distribution. For any arbitrary $u \in U$, denote

$$\bar{x}(u) \triangleq \frac{1}{n}\sum_{i=1}^{n} \chi_{x_i}(u) \tag{3.2}$$

It is called the covering frequency of ξ to u, and which is estimation function of μ_ξ.

Theorem 3.1 (The law of great numbers of fall-shadow) Suppose that $\xi_i \in (\Omega,\mathcal{F};\ \mathcal{P}(U),\check{\mathcal{B}})$ (i=1,2,$\cdots$) are independent and have the same distribution, $\mu_{\xi_i}(u) = \mu(u)$, denote

$$\bar{\xi}_n(u,\omega) = \frac{1}{n}\sum_{i=1}^{n} \chi_{\xi_i(\omega)}(u), \tag{3.3}$$

then for any $u \in U$ holds

$$\xi_n(u,\cdot) \xrightarrow{a.e.} \mu(u) \qquad (n \longrightarrow \infty) \qquad (3.4)$$

Let $\xi \in \mathcal{S}(\Omega,\mathcal{F};\mathcal{P}(U),\check{\mathcal{B}})$, $(U,\mathcal{B})$ be a measurable space, m is a positive measure on $\mathcal{B}$, denote

$$\bar{m}(\xi) = \int_U \mu_\xi(u)\, m(du) \qquad (3.5)$$

$\bar{m}(\xi)$ is an important numerical characterization. To estimate $\bar{m}(\xi)$, we can use

$$\bar{m}(x_1,\cdots,x_n) \triangleq \frac{1}{n}\sum_{i=1}^{n} m(x_i) \qquad (3.6)$$

Proposition 3.2. Suppose $\bar{m}(\xi)<\infty$, $\xi_1,\cdots,\xi_n,\cdots$ are independent random sets with the same distribution as ξ, then we have

$$\bar{m}(\xi_1,\cdots,\xi_n) = \frac{1}{n}\sum_{i=1}^{n} m(\xi_i) \xrightarrow{a.e.} \bar{m}(\xi) \qquad (n \longrightarrow \infty) \qquad (3.7)$$

TRANSFORMATION OF RANDOM TESTS INTO SET-VALUED ONES

There are unaccuracy problem with the measurement of physical guantities. The origin of errors lies in both probability and fuzziness. The observed results may be written in the form as $\theta \pm \Delta$, it is in fact a set value $[\theta-\Delta,\ \theta+\Delta]$.

Let θ, Δ be independent real random variables, $P(\Delta \geqslant 0)=1$, denote

$$\xi(\omega) \triangleq [\theta(\omega)-\Delta(\omega),\ \theta(\omega)+\Delta(\omega)] \qquad (4.1)$$

it is a random set.

Proposition 4.1. ξ has fall-shadow

$$\mu_\xi(x) = F_{\theta-\Delta}(x) - F_{\theta+\Delta}(x-0), \qquad (-\infty < x < +\infty) \qquad (4.2)$$

where $F_\tau(x) = P(\tau \leqslant x)$, $\tau = \theta-\Delta$ or $\theta+\Delta$.

Proposition 4.2. Suppose that $\xi = [\theta-\Delta,\ \theta+\Delta]$, $E\theta = a$, $E\Delta = \delta$, $(i=1,2,\cdots)$ are independent random sets with the same distribution as ξ, then for any $x \notin \{a-\delta, a+\delta\}$ holds

$$\chi_{(\frac{1}{n}\sum \xi_i(\omega))}(x) \xrightarrow{a.e.} \chi_{[a-\delta,\,a+\delta]}(x) \qquad (n \longrightarrow \infty) \qquad (4.3)$$

where $\xi_i(\omega)$ be intervals, $\frac{1}{n}\sum \xi_i(\omega)$ is understanded according to the operations among interval-numbers.

From this, for a sample $x_1,\cdots,x_n$ of ξ, we cna make

$$\bar{\bar{x}} \triangleq \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (x_i \in \mathcal{P}(R),\ i=1,\cdots,n) \qquad (4.4)$$

as an interval estemation of the expected value of θ.

We may also reversely make statistics for θ and Δ and get information concerning the fall shadow of ξ.

Proposition 4.3. Suppose θ and Δ are independent integral random variables, Δ never takes negative values. Suppose that n times of independent tests of these variables have been made and the frequency while $\theta = h$ is n_k $(\sum_{k=-\infty}^{+\infty} n_k = n)$, the frequency while $\Delta = k$ is m_k $(\sum_{k=-\infty}^{+\infty} m_k = m)$, denote

$$\mu_i = \mu_{-i} = \sum_{j \geqslant i} m_j/m \qquad (i \geqslant 0) \qquad (4.5)$$

denote also

$$\bar{x}_k \triangleq \frac{1}{n}\sum_{i=-\infty}^{+\infty} \mu_i\, n_{k+i} \qquad (k = 0, \pm 1, \cdots) \qquad (4.6)$$

then for any k, holds

$$\bar{x}_k(\omega) \xrightarrow{a.e.} \mu_\xi(k) \qquad (n \longrightarrow \infty) \qquad (4.7)$$

FALL-SHADOW SMOOTHER METHOD

Formulae (4.4) is a smoother of sequence $\{n_k\}$ by fall-shadow $\{\mu_i\}$. It is just the sliding average with period 2Δ whenever Δ is a constant.

Consider a data field $\{x_{(i_1,\cdots,i_n)}\}(i_j \in I_j = \{r|r$ is integel, $-\alpha_j \leqslant r \leqslant \alpha_j\}$ $(j=1,\cdots,n))$, which can be viewed as an occurence of a certain random field $\xi_{(i_1,\cdots,i_n)}((i_1,\cdots,i_n)\in I_1 x\cdots x I_n)$. Exame an n-dimension integral random vector, it always takes its value in $I_1 x\cdots x I_n$. Suppose the distribution of η is $\{P_{(i_1,\cdots,i_n)}\}$ $((i_1,\cdots,i_n) \in I_1 x\cdots x I_n)$, denote

$$\underline{\eta}(\omega) = \{(i_1,\cdots,i_n) \mid i_j \text{ is located between o and the } j\text{th component of } \eta(\omega)\ (j=1,\cdots,n)\}, \qquad (5.1)$$

$\underline{\eta}$ is a random set.

Proposition 5.1. $\underline{\eta}$ has fall-shadow

$$\mu_{\underline{\eta}}(i_1,\cdots,i_n) = \Sigma^* p_{(j_1,\cdots,j_n)}, \qquad (5.2)$$

Σ^* sums those terms $(j_1,\cdots,j_n)$ which are satisfied with $j_k \leqslant i_k \leqslant 0$ or $0 \leqslant i_k \leqslant j_k$ for every $k \leqslant n$.

$\mu_{\underline{\eta}}$ has following property:

$$((\forall k)\ (j_k \leqslant i_k \leqslant 0 \ \text{ or } \ 0 \leqslant i_k \leqslant j_k)) \Longrightarrow \mu(i_1,\cdots,i_n) \geqslant \mu(j_1,\cdots,j_n). \qquad (5.3)$$

The fall-shadow μ holding this property (5.3) is called centralized fall-shadow.

Definition 5.1. For a given data field $\{x_{(i_1,\cdots,i_n)}\}((i_1,\cdots,i_n) \in I_1 x\cdots x I_n)$, and a centralized fall-shadow $\mu(i_1,\cdots,i_n)$, denote

$$y_{(i_1,\cdots,i_n)} \triangleq y^{\mu}_{(i_1,\cdots,i_n)} \triangleq c\ \Sigma^* \mu(s_1,\cdots,s_n)\, x_{(t_1,\cdots,t_n)}, \qquad (5.5)$$

where Σ^* sums those terms, about which $s_k + t_k = i_k$ for any $k \leqslant n$ holds, c is an appropriately determined constant. Then call $\{y_{(i_1,\cdots,i_n)}\}$ the fall-shadow smoother of $\{x_{(i_1,\cdots,i_n)}\}$.

Let M be the set of all centalized fall-shadows.

Definition 5.2. For a given class $M_0 \subset M$, $\mu_0 \in M_0$ is called the interior correlation function of the data field,if

$$\sum_{(i_1,\cdots,i_n)} (y^{\mu_0}_{(i_1,\cdots,i_n)} - x_{(i_1,\cdots,i_n)})^2 = \min_{\mu \in M} \sum_{(i_1,\cdots,i_n)} (y^{\mu}_{(i_1,\cdots,i_n)} - x_{(i_1,\cdots,i_n)})^2. \qquad (5.6)$$

By use of the interior correlation function, we may make the interpolation and extrapolation of data field. It can be applicable to field estimation and prediction problems. The weight function regression method in mathematical statistics and the Krige method in geology ststistics(4) both may be contained in the framework suggested herein.

DEGREE ANALYSIS

The set-valued statistics is most hopefully available in those decision-making processes, which have to be relying upon psychological measurement. In practice, we should carry out lots of such estimations as evaluating the degree of some thing to fit some aim or requirement. People often use some words as "satisfying", "feasible", "compatible", "stable", "reliable", $\cdots\cdot$ and so on to express their feeling and perception to a certain object. People want to give them measurement scales and to conduct analysis about them, but effective mathematical method is lacking for this purpose. Set-valued statistics can offer an useful approach, which is the degree analysis.

1. Degree estimation for a single factor

Those evaluation methods cited in current literature are similar to that presented here. For instance, in order to evaluate the satisfactory degree of some thing, we may draw a line segment, number 1 represents "very satisfying", and 0 "very unsatisfying", 0.5 "middling", as shown in Fig. 2. Every participating one puts three points on proper locations on the line, in accordence with his feeling about the degree of satisfaction. Denote the first point from left x, and first point from right y. The interval $[x,y]$ is the result for a person, and $[x_i, y_i]$ $(i=1,2,\cdots,n)$ for all parcipating persons. Calculating $\overline{x}(u)$ according to eq.(3.2), we cna get a fuzzy satisfactory degree. Futhermore calculate

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i + y_i}{2} , \tag{6.1}$$

and

$$\overline{m} = \frac{1}{n}\sum_{i=1}^{n} (y_i - x_i) , \tag{6.2}$$

according to eq.(3.6). We may represent the point estimation of satisfactory degree by use of α. $\overline{m}$ may be called the blindness of point estimation. The smaller the value of $\overline{m}$, we are more faithful about such an estimation.

2. Degree analysis of multiple factors

Taking as a simple example, suppose that we should make an examination of several proposed plans from the view of point of "necessity" and "feasibility". As shown in Fig. 3, each trial gives us a interval $[x_1, y_1]$ on axis u(necessity) and $[x_2, y_2]$ axis v (feasibility), then as a result a rectangle $[x_1, x_2; y_1, y_2]$ too. The result of n times trials is $[x_1^{(K)}, x_2^{(K)}; y_1^{(K)}, y_2^{(K)}]$ $(k=1,\cdots,n)$. Then following result

$$\overline{x}(u,v) = \frac{1}{n}\sum_{K=1}^{n} \chi_{[x_1^{(K)}, x_2^{(K)}; y_1^{(K)}, y_2^{(K)}]}(u,v) \tag{6.3}$$

can be obtained by eq.(3.1), where

$$\chi_{[x_1^{(K)}, x_2^{(K)}; y_1^{(K)}, y_2^{(K)}]}(u,v) = \begin{cases} 1, & \text{if } x_1^{(K)} \leqslant u \leqslant y_1^{(K)},\ x_2^{(K)} \leqslant v \leqslant y_2^{(K)}; \\ 0, & \text{otherwise.} \end{cases}$$

According to the weights assigned to "necessity" and "feasibility", calculate

$$t = \int_0^1\int_0^1 (u\cdot w_1(u,v) + v\cdot w_2(u,v))\,\overline{x}(u,v)\,du\,dv \Big/ \int_0^1\int_0^1 \overline{x}(u,v)\,du\,dv \tag{6.4}$$

t is called the synthetic degree of necessity and feasibility of that plan. Here $w_1(u,v)$, $w_2(u,v)$ are varing weight functions, which satisfy

$$w_i(u,v) \geqslant 0, \qquad \sum_{i=1}^{2} w_i(u,v) \equiv 1 \tag{6.5}$$

The idea of varying weight function lies in the fact that the weights of both factors should be variable with different states of combination of these two factors.

The determination of varying weights may be carried out by investigation among experts, getting at last a "vector field" similar to that shown in Fig.5. The slope represents the ratio of w_2 and w_1. The approximate expression of slope in Fig. 4 is $w_2/w_1(u,v)=u/v$, then accordinging to eq.(6.5) we get

$$w_1(u,v) = \frac{v}{u+v} , \qquad w_2(u,v) = \frac{u}{u+v} , \tag{6.6}$$

thus

$$t = \int_0^1\int_0^1 \frac{2uv}{u+v}\,\overline{x}(u,v)\,du\,dv \Big/ \int_0^1\int_0^1 \overline{x}(u,v)\,du\,dv . \tag{6.7}$$

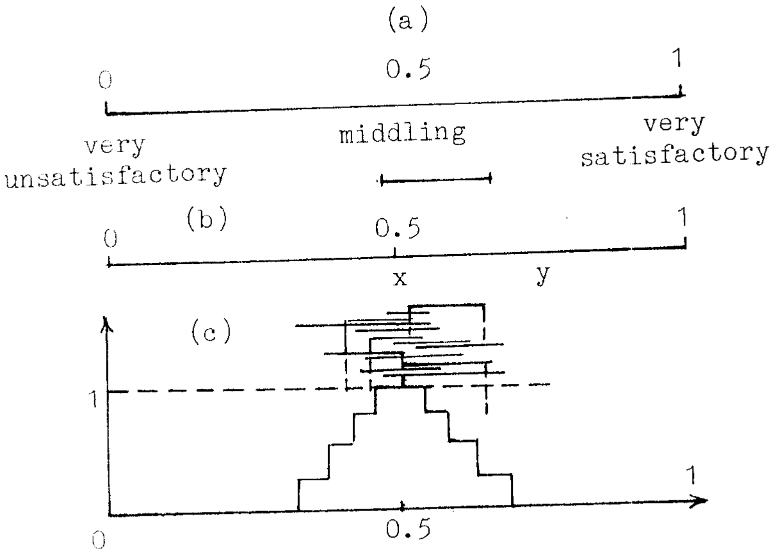


Fig. 2 Degree analysis of single factor

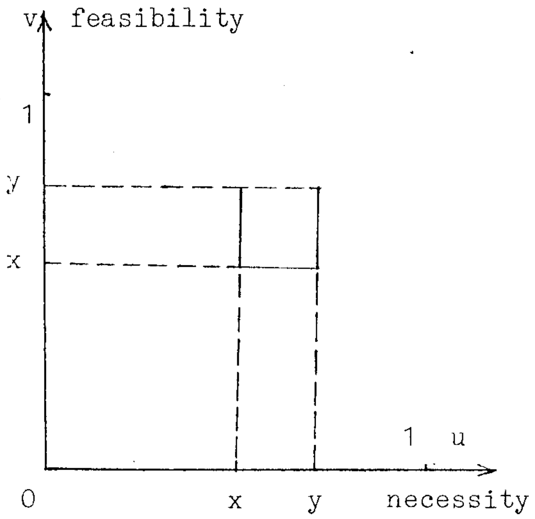


Fig. 3 Degree analysis of multiple factors

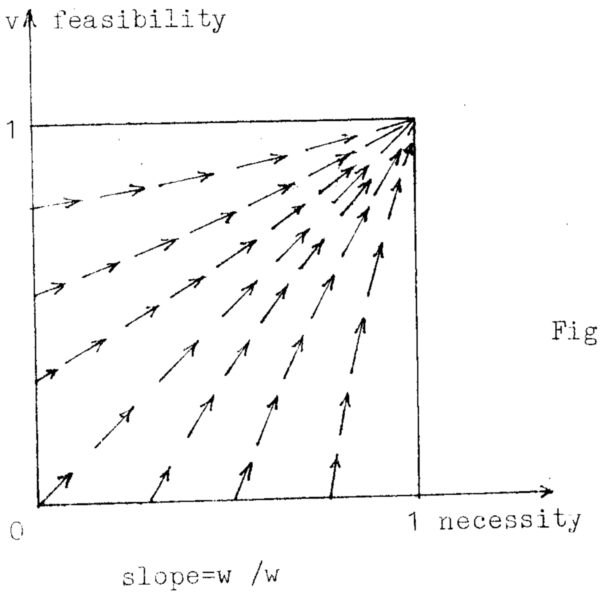


Fig. 4 Varying weights

REFERENCE

(1) D. Dubois, H. Prade, Fuzzy Sets and Statistical Data, Ensembles Flous-82, Notes, Communications, Articles ɛcrits en 1982.

(2) I. R. Goodman, Fuzzy Sets as Equivalence classes of random sets, in R. R. Yager Ed. Fuzzy set and Possibility Theory (Pergaman Press, Oxford, 1982) 327-343.

(3) Ma Mouchao, Cao Zhiqiang, The Multistage Evaluation Method in Psychological Measurement: An Application of Fuaay Sets Theory to Psychology, in Approximate Reasoning in Decision Analysis, M. M. Gupta and E. Sanchez (eds.) North-holland (1982).

(4) G. Matheron, Taite de Geostatistique Applique, Editions Technip, Prais (1962).

(5) N. T. Nguyen, On random sets and Belief Functions J. Math., Anal. & Appl. 65 (1978), 531-542.

(6) Wang Peizhuang, E. Sanchez, Treating a fuzzy subset as a fallable random subset, in Fuzzy Information and Decision Processes, M. M. Gupta, E. Sanchez (eds.) North-Holland (1982).

(7) Wang Peizhuang, From the Fuzzy Statistics to the Falling Random Subsets, in Advances on Fuzzy Set Theory and Applications, P. P. Wang (ed.) Pergamon Press (1983).

(8) Wang Peizhuang, Random Sets and Fuzzy Sets, Encyclopedia of Systems Control, Pergamon Press (1983).

(9) Wang Peizhuang, σ- Hyperfield and the Measurability of Multivalued Mappings, Kexue Tongbao, Vol.28, No.12 (1983).

(10) Zhang Nanlun, The Membership and Probability Characteristics of Random Occurences, I, II, III, J. of Wuhan Institute of Building Materials, No. 1.2.3 (1981).

(11) L. A. Zadeh, Fuzzy Sets and a Basis for Theory of Possibility, Fuzzy Sets and Systems 1 (1978) 3-28.

(12) Liu Xihui, Wang Mengmei, Wang Peizhuang, Fuzzy Intensity, Earthquake Engineering and Engineering Vibration, Vol.3, No.3, 1984.

(13) A.Kandel, On Fuzzy Statistics, in Advance in Fuzzy Set Theory and Applications, M.M.Gupta, P.K.Ragade, R.R.Yager (eds.), North-Holland (1979)