Thailand Statistician
2015; 13(1): 1-31
http://statassoc.or.th
Contributed paper

Statistics of Fuzzy Data: A Research Direction for Applied Statistics

Hung T. Nguyen
Department of Mathematical Sciences, New Mexico State University, USA
and Faculty of Economics, Chiang Mai University, Thailand.
E-mail: hunguyen@nmsu.edu

Received: 13 June 2014
Accepted: 7 October 2014

Abstract

We emphasize problems where fuzzy data appear naturally and need to be used and analyzed properly within the context of applied statistics. To enlarge the practice of statistics to this new types of observed data, we need to place fuzzy data within the context of probability theory. Having a theory of statistics of fuzzy data, applied statisticians can embark on a promising road of important applications, especially in economics, whereas otherwise these problems are either ignored or badly handled.

Keywords: Coarse data, fuzzy logics, fuzzy sets, game theory, random set, random fuzzy sets.

1. Introduction

This paper is about theoretical statistics for applied statisticians. Since researchers in theoretical statistics (e.g., developing more general statistical procedures to enlarge the domain of applicability of statistics) need to read empirical research in order to guide their realistic generalizations, researchers in applied statistics need to read theories in order to apply their statistical tool box correctly.

The paper will elaborate on linguistic data which need to be formulated properly before subjecting to statistical analysis. Linguistic data are modeled as fuzzy sets in the sense of Zadeh [1]. While fuzzy sets and their associated logics (see e.g., Nguyen and Walker [2]) are widely used in almost all fields of science, especially, engineering and computer science (see e.g. Nguyen and Sugeno [3]), their applications to social problems, such as economics, where statistics is a dominant investigation strategy, seem lacking. Only recently that some attempts have surfaced (e.g., Lindstrom [4]). We will elaborate on why?

The paper aims at presenting several real-world problems in which only linguistic data are available, as well as establishing the foundations for making inference with fuzzy data.

The paper is organized as follows. First, we take this opportunity to remind applied statisticians that statistics is firmly based upon probability, and as such, there is a need to understand the theory of statistics before using statistical tools. This will explain why we will devote a large portion of this paper to discuss probabilistic foundations of fuzzy data, before suggesting how to use them in applied works. Next, we address applied research works where the order of investigation should be: problems, data, then tools. The point is this. Tools come last. We review various important types of coarse data (i.e., data of low quality). The point is this. After all, statistics is about data! Data dictate which appropriate statistical tools to use. Then, we give a tutorial on modeling of linguistic data by fuzzy sets, as well as their associated logics which will be used to proceed data (in fact, to suggest statistical models). The emphasis will be on how to place fuzzy data within an appropriate statistical theory, recalling that fuzzy analysis should not be an alternative to statistical analysis, but, as Lotfi Zadeh has said over and over again that fuzziness should coexist with randomness in real-world complex (physical and social) systems.

2. What is Statistics?

Statistics is a methodology to aid discovery of knowledge in experimental sciences. This is particularly useful when there is uncertainty involved. The uncertainty due to randomness is quantified by probability (objective or subjective). The following "first thing" to remember is essential for understanding why we need to consider an appropriate probability space to place fuzzy data in a statistical context in a subsequent section.

A variable is just the name of a quantity of interest, such as X standing for annual income in some population. It is variable since income can take on different values. If we can "assign" values to X, then X is called a deterministic variable. If we cannot predict values of X with certainty, then X is called a random variable (or vector). In this case, the complete information we could have about X is its "law" governing its uncertain (random) evolution. This counterpart of law of motion in physical systems is called the distribution function of X. For random variables or vectors, i.e., random "elements" whose ranges (of values) are euclidean spaces $\mathbb{R}^d$, $d \geq 1$, the law of X is simply characterized by a distribution function $F : \mathbb{R}^d \to [0, 1]$, where $F(x) = P(X \leq x)$, with $\leq$ being the usual partial order relation, defined componentwise. But then, this concept requires the operator P (standing for probability). On the other hand, if we run into other types of random elements (observed outcomes) such as curves (e.g., daily stock price fluctuations), or sets (say, in $\mathbb{R}^d$), such as areas affected by an earthquake of some given magnitude, then how to describe (in fact characterize) the distributions of such non-euclidean outcomes? We need a general framework to handle all possible types of random elements which could surface on our paths. In fact, we should describe probability spaces first to really explain what is a random variable and its law. The familiar (and "informal") formula $F(x) = P(X \leq x)$ then comes as a by-product. This is not only rigorous but also necessary to characterize laws for general random elements. Your basic *probability course for statistics* lays down the foundations: Let $(\Omega, \mathcal{A}, P)$ be a probability space, and U is an arbitrary set. A map $X : \Omega \to U$ is called a random element if there is some $\sigma-$ field $\mathcal{U}$ of subsets of U, such that X is $\mathcal{A} - \mathcal{U}-$ measurable, i.e., $X^{-1}(\mathcal{U}) \subseteq \mathcal{A}$. Then, the law of X is the probability measure on $\mathcal{U}$ defined by $P_X = PX^{-1}$. When $U = \mathbb{R}^d$, the famous Lebesgue-Stieltjes theorem established a bijection between P_X and distribution functions F which, for example, when $d = 1$, is characterized as $F : \mathbb{R} \to [0, 1]$ satisfying

i) $F(-\infty) = 0$, $F(+\infty) = 1$
ii) $x \leq y \implies F(x) \leq F(y)$
iii) $\lim_{y \searrow x} F(y) = F(x)$

Now consider the situation where U is a finite set, say, $U = \{u_1, u_2, ..., u_k\}$, and our random outcomes (of some experiment) are subsets of U. How to describe the law governing the random evolution of X? Note that this situation is in fact the very first thing you learnt from your first course in applied statistics, namely sampling from a finite population (see e.g., Hajek [5]) where a sampling design is nothing else than a *random set* X, i.e., a random element, defined on some probability space $(\Omega, \mathcal{A}, P)$ and takes values in the power set 2^U (set of all subsets of U). While in sampling designs, we only need to specify the probability converage function $\pi(u) = P(u \in X)$, here, for other statistical analyses, we need the whole "distribution" of X. On 2^U, the partial order relation to consider is set inclusion $\subseteq$ (contained in). Thus, the counterpart of distributions on euclidean spaces becomes $F(A) = P(X \subseteq A)$, for $A \in 2^U$. Clearly, if we have $P(.)$, then we get $F(.)$ by this "formula". Just like "ordinary" distribution functions, we ask "what are the characteristic properties of a set function $F : 2^U \to [0, 1]$ so that P is determined uniquely?" Here is the answer. $F(.)$ should satisfy

(a) $F(\varnothing) = 0$, $F(U) = 1$
(b) For any $k \geq 2$, and $A_j \in 2^U, j = 1, 2, ..., k$,

$$F(\cup_{j=1}^{k} A_j) \geq \sum_{\varnothing \neq I \subseteq \{1,2,...,k\}} F(\cap_{i \in I} A_i)$$

Interested reader could find a proof in, e.g., Nguyen [6]). The point here is that distributions of general random elements are defined so that they correspond bijectively with a probability measure on its range space. We will employ the same method when considering fuzzy data, i.e., for random elements taking fuzzy sets as values.

Next, whether we wish to estimate some unknown quantities of interest, or to validate models for predictions, we need to suggest best estimators, most powerful tests, and optimal predictors. These desirable properties require results from probability. For example, even in reality we only have a finite set of data, estimators should be consistent to be used. Consistency is expressed in terms of various concepts of convergence of sequences of random variables,

and often is derived based on results in probability theory such as laws of large numbers. Any of your "new" or modified estimators are only valid if they are at least consistent. Simulations on some cases provide only *indicative,* but not *conclusive,* conclusions.

Since we are going to derive models using *rules* when dealing with fuzzy data, let's say few (important) words about statistical models. Unlike physical systems, statistical phenomena do not have the luxury of having dynamical laws, and hence have to rely only on *models*. Statistical models should be justified by data, including their *validity.* You are all familiar with your BB (Bread and Butter) tool in applied statistics, namely linear regression models, e.g., Kutner et al [7]).

As we will see, most of the problems of interests are of the form: investigating the effect of covariates on a variable of interest, such as, knowing that the determinants of income are e.g. education levels and skill, we wish to have a quantitative relation between these variables, for, say, prediction purposes. Francis Galton called such analysis a regression problem: regressing response variables upon covariates (regressors) variables. If we wish to predict a response variable Y based upon a covariate X, using the mean squared error concept (MSE), then, since the conditional mean $E(Y|X)$ is an optimal predictor, the regression model could be

$$Y = E(Y|X) + \varepsilon$$

where ε is a random error such that $E(\varepsilon|X) = 0$ (so that the above statistical model is compatible). In general, $E(Y|X)$ is a non linear function of X. The situation is much simpler if $E(Y|X)$ can be approximated by a linear function in X, i.e., considering $E(Y|X) = \theta X$, leading to the statistical linear regression model

$$Y = \theta X + \varepsilon$$

This is a "plausible" simple regression model which needs to be validated (i.e., to see whether it is a good approximation to the true relationship between X and Y).

Given, say, a random sample (i.e., an i.i.d. set of observations) (X_i, Y_i), $i = 1, 2, ..., n$, drawn from (X, Y), you usually justify and validate your above linear regression model by computing your R_n^2 and decide. It is precisely here that you need to be careful. Do not behave like empirics (people who rely

solely on practical experience rather than on scientific principles)! Any empirical measures are used as estimates for corresponding population measures (parameters). But population parameters might not exist (e.g., the mean of a Cauchy population). What is the population parameter your R_n^2 is supposed to estimate? In the above model, assuming that ε is independent of X, we have

$$Var(Y) = Var(\theta X + \varepsilon) = \theta^2 Var(X) + Var(\varepsilon)$$

so that $Var(\theta X) < Var(Y)$, and hence the ratio $\frac{\theta^2 Var(X)}{Var(Y)} \in [0, 1]$ can be used as an indication of the adequacy of the linear model. Well, if $Var(Y)$,$Var(X)$ are both finite, then by the strong law of large numbers, R_n^2 is a consistent estimator of $\frac{\theta^2 Var(X)}{Var(Y)}$, otherwise (e.g., when X, Y are heavy-tailed, see, e.g., Resnick [8]), the finite quantity R_n^2 from your data is meaningless!

The point is this. In one hand, be careful when considering a model, and on the other hand, validation is needed. As an example, suppose you are interested in the effect of the covariate X on "large" values of Y, rather its mean. With an appropriate error concept, the conditional $\alpha-$ quantile $q_\alpha(Y|X)$ is the best predictor of the $\alpha-$ quantile of Y based on X, so that a linear quantile regression model could be considered:

$$Y = \theta_\alpha X + \varepsilon_\alpha$$

with $q_\alpha(\varepsilon|X) = 0$. See Koenker [9]. How do you intend to validate it?

In our subsequent exposition, we will discuss how to propose models for inference with fuzzy data. Here is a typical example in standard statistical practice.

To increase production efficiency in agricultural economics, say, statistical analyses of firm production technologies are needed. More realistic models for quantifying technical efficiency are desirable. Such research is not only beneficial to agriculture, but, by analogy, also to other economic activities such as investments (by measuring stock market efficiency).

Quantifying technical efficiency (of firms) from empirical data (e.g., cross section data) is a problem of concern in production theory. Since an input (labor, resource, capital,...) can produce various different outputs, depending on how to "manage" the input, there is such thing as the "production frontier", namely the maximum output an input can produce

$$\varphi(x) = \max\{y : x \to y\}$$

If we know the frontier $\varphi(.)$ (of a given technology) then, when observing (X_i, Y_i) from the firm i, we can take $\frac{Y_i}{\varphi(X_i)}$ as the (degree) of technical efficiency of firm i.

Things are not just simple as we think, not only because we do not know the function $\varphi(.)$, but also, the observed output Y_i could be "above" the frontier!

Let $\Psi = \{(x, y) : x \to y\}$ be the attainable production set. From a statistical viewpoint, we view Ψ at the range of the random vector (X, Y), or equivalently, the support of their joint distribution, i.e., $P((X, Y) \in \Psi) = 1$. However, an output y could be a result, not soly of an input x, but also of random "shock", resulting in a value which could exceed the frontier value $\varphi(x)$. Thus, the concept of frontier should be extended to a "stochastic frontier" $x \to \varphi(x) + V$, with V being a symmetric random noise, to cover this situation. For $\varphi(.) + V$ to be a "frontier", we would have $Y \leq \varphi(X) + V$. Thus, let $U = \varphi(X) + V - Y$ be another random variable, nonnegative ($U \geq 0$), representing technical efficiency, we arrive at the now popular *production stochastic frontier model* (SFM), see Kumbhakar and Knox Lovell [10]:

$$Y_i = \varphi(X_i) + V_i - U_i$$

The problem is the estimation of efficiency U_i for firm i, given, say, a random sample $(X_i, Y_i), i = 1, 2, ..., n$ across firms. Note that like V_i, U_i is not observable. Suppose we could estimate the production function $\varphi(.)$ by some estimation method, then we can compute the estimated residuals $\varepsilon_i = v_i - u_i$ by $y_i - \hat{\varphi}(x_i)$.

How to predict U_i based on the relevant information $\hat{\varepsilon}_i = y_i - \hat{\varphi}(x_i)$? Well, the "observale" value $\hat{v}_i - \hat{u}_i = \hat{\varepsilon}_i$is a value of the random variable $V - U$. Thus, in MSE sense, the best predictor of U is $E(U|V - U = \hat{\varepsilon}_i)$, its estimate is taken to the the estimated technical efficiency of firm i.

Given this machinery, it remains to specify the model, examine the observed data, and then proposing appropriate estimation techniques. These could be in fully parametric, semi-parametric, or nonparametric forms. In the parametric setting, *copulas* will necessarily enter the analysis, since we need to model the joint distribution of (V, U). In the semi-parametric specification (parametric form for $\varphi(.)$, and unspecified distributions of the error terms), we need to use *semi-parametric statistics***.** In the nonparametric setting, *conditional quantile-based nonparametric estimation* seems attractive.

3. What is Applied Statistics ?

Just like applied mathematics, perhaps it seems obvious that applied statistics mean ”applying statistical methods to real-world problems”? Yes, of course, but the delicate thing is this. There is a distinction between ”applications of statistics” and ”solving problems using statistics”. First of all, while probability can be considered as a branch of pure mathematics (where theoretical research might take a central stage), statistics is, by its creation and nature, an applied science. You can say that you are a mathematician without naming any fields of applications (e.g., algebraic topology), but you cannot simply say that you are a statistician, even you are only interested in theoretical research. This is so since statistics is created to provide scientific principles in experiments to discover knowledge. As such, statistical theories are guided and developed by applications, and not just investigated at an abstract level like pure mathematics. A striking example is econometrics. It is the desire to solve real-world problems using statistics that *econometricians* developed our modern statistical theories! Like all ”applied” statisticians, they have a repertoire of statistical tools, but instead of applying statistical tools blindly (e.g., using ”popular” models without justifying, using untested assumptions, ignoring unusual features which have not yet handled by standard statistics, etc...) they seriously examine the real economic problems they face (remember: problems and data first), and then see whether there are suitable statistical tools in their tool box to solve the problems. If there is no suitable statistical tools in their tool box (such as linear regression), then they pause and *develop new statistical tools*, and after that come back to apply their new theoretical work to their intended applied problems. A striking example is James Heckman’s work on sample selection bias in labor econometrics [11], for which, together with Daniel MacFadden, they were awarded the Nobel Prize in Economic Sciences in 2000. The essential of Heckman’s work is on missing data, speeling out loud once again: do not think about your familiar statistical tools before examining honestly your problems! A pleasant but important reading about the ”danger” of using wrongly statistics is Wheelan [12].

In this paper, we will address another problem on missing data, namely missing data due to unobservability of a certain kind where we will use fuzzy sets for modeling. The lesson is this. Applied statisticians should investigate honestly their problems, examine carefully their obtained data, then *only then*,

think about how to use statistics to ”solve” their problems. If there is no suitable statistical tools to handle their actual problems, then ”pause” and do some theoretical research inspired from their problems in order to come back later to solve them with their own new statistical tools. This is what we like to call *applied statistics.*

Remark. What is theoretical statistics? First, observe that the journal *Annals of Mathematical Statistics* has been changed to simply *Annals of Statistics* since there is no need to emphasize the term ”mathematical” in it. Mathematics is the language of science and hence is implicit in statistics. However, applied statisticians tend to view this journal and similar journals as theoretical statistics, a better name for the orientation. Since statistics aims at solving real-world problems, we recognize that too much assumptions on our models and data make our machninery far from reality. Therefore, there is a need to relax standard assumptions in statistics as well as examining carefully the quality of data. Research in theoretical statistics focus on issues such as these in order to make statistical theory more efficient and close to reality. Examples of such needed research are *robustness of estimators, bootstrap methods for sampling distributions,* and *maximum entropy inference.*

4. What is a Fuzzy Set?

So far we have mentioned, at several places, the term ”fuzzy data”. It is time now to elaborate on it.

Recall that after all, statistics is about data. Data should be understood as available information concerning problems under investigation. They could be numerical or else. Data are observed ”outcomes” of variables. Depending upon the type of outcomes, the corresponding of variables are classified as *quantitative or qualitative variables.* Quantitative variables refer to variables whose values are numerical, whereas qualitative variables take linguistic labels in a natural language as values, for example, ”gender” (with values male, female), ”disability status” (not disabled, partially disabled, fully disabled). A more specific name for qualitative variables is *linguistic variables.*

We are familiar with linguistic variables in statistics in general (e.g., in the form of categorical data via contingency tables), and in linear regression in particular (qualitative predictors such as ”quality of sales management”, see, e.g, Kutner et al. [7]), as well as the way we used to handle them. What we

are going to do is introducing you to a new way to look at linguistic data. Why? and what for? These questions will be answered in section 5. For now, we are going to describe this "new look" at linguistic data.

When taking natural language to impart knowledge and information, there is a great deal of imprecision, vagueness or fuzziness. Such statements as "Mary is young", "John is tall", where "young", "tall" are *fuzzy concepts*, are simple examples. We will discuss the intrinsic notion of fuzziness in natural language, as well as how to represent, manipulate and draw inferences from such imprecise information. Fuzzy sets are mathematical objects to model fuzzy concepts (in natural language) which contain valuable information for decision-making in both physical and social sciences. Fuzziness refer to phenomena that do not have sharply defined boundaries. By modeling quantitatively fuzzy concepts, the number of objects encountered in human reasoning that can be subjected to scientific investigation is increased.

We discuss first the formal concept of a fuzzy set, and then explain why fuzzy sets are models for imprecise data in real-world applications which can be subjected to statistical analysis in section 5.

Fuzzy data are linguistic data where we use the mathematical theory of fuzzy sets to model them. As we will see, this is a much more general view than defuzzifying fuzzy concepts (by putting thresholds on imprecise quantities) or using "quantitative indicators".

To motivate the concept of a fuzzy set, consider the following simple example. As we recalled in section 2, events associated with random phenomena (man-made or natural) are subsets of a universe of discourse U (e.g., $\mathbb{R}^d$). For example, the set of monthly incomes larger than $50,000$ baht is the subset $A = \{x \in \mathbb{R}^+ : x \geq 50,000\}$ of $U = \mathbb{R}^+$. How to describe the "set" of "high income"? Well, we need to generalize ordinary (crisp) sets! The question is how?

A crisp set $A \subseteq U$ can be described by its indicator function, in an equivalent way. Specifically, there is a bijection between A and the function $1_A(.) : U \to \{0, 1\}$ where

$$1_A(u) = \begin{cases} 1 & \text{if} \quad u \in A \\ 0 & \text{if} \quad u \notin A \end{cases}$$

The indicator function $1_A(.)$ has a nice meaning: it is a *membership function*, telling us about *membership* of elements of U. Indeed, if $1_A(u) = 1$, then

the element $u \in U$ is a member of A, whereas if $1_A(v) = 0$, v is not a member of A. Membership is either full or nothing, as indicated by the range $\{0, 1\}$.

In coalitional (cooperative) games in economic competition, coalitions are subsets of the set U of all "players" involved. If we examine attentively, players, when joining a coalition, might not necessarily commit their full times, resources, energy to it. Two members of a coalition A might differ by their levels of commitment and therefore should be classified differently (for, e.g., a capital risk allocation). We can indicate their "degrees of membership" in A by the portions of their committments, which are numbers in the unit interval $[0, 1]$. By doing so, we actually model the coalition as a fuzzy set.

As an other example, observations in quality control, based on a classical six-sigma control chart (under normality assumption) have different degrees of "in-control" proportional to their positions with respect to the mean. Observations near the frontier of the control chart should be assigned lower degrees than those which are nearer to the mean. If we take this "finer" viewpoint, then we actually model our observed data as *a fuzzy random (closed) set* (see details in section 6 below).

Formally, while we cannot generalize the set A directly, we can use its equivalent representation by its indicator function to do so. Simply extend the range $\{0, 1\}$ to the whole unit interval $[0, 1]$, representing the partial (gradual) degrees of membership. For convenience, we denote also by $A(.) : U \to [0, 1]$ the *membership function*, where $A(u) \in [0, 1]$ represents the degree of membership of u in the "fuzzy set" A which could be different than 1 or 0. Fuzziness is a matter of degree.

For example, let $A =$"high income". It is a fuzzy concept since the linguistic label "high" does not have a sharply defined boundary (unless you defuzzify it by a threshold like $50,000$). This fuzzy concept is *defined* by a membership function. For example, in thousand of Baht unit,

$$A(x) = \begin{cases} 0 & \text{if} \quad x < 20 \\ \frac{x-20}{55} & \text{if} \quad 20 \leq x \leq 75 \\ 1 & \text{if} \quad x > 75 \end{cases}$$

We will discuss how to assign membership functions to fuzzy concepts later. For the time being, let's put down a definition

Definition. A fuzzy subset of a set U is a function $U \to [0, 1]$

Note that ordinary (crisp) subsets are special cases of fuzzy subsets.

Several remarks are in order.

(i) For $x = 50$, its degree of compatibility with the fuzzy concept of "high income" (or its degree of membership in the fuzzy set "high income") is $A(50) = \frac{30}{55}$ which does not mean that the probability of an income of 50 is high, in other words, it is not the probability of having a high income. *Membership functions are not probability distributions.*

If we wonder whether some connections between fuzziness and randomness exist, then here is one which says that fuzziness can be viewed as a weak form of randomness. Specifically, let $f : U \to [0, 1]$ be a given membership function (of some fuzzy subset of U). For $\alpha \in\in [0, 1]$, the $\alpha-$level set of f is $L_\alpha(f) = \{u \in U : f(u) \geq \alpha\}$. Note that $f(.)$ can be recovered from its level sets as $f(u) = \int_0^1 1_{L_\alpha(f)}(x)d\alpha$. If we randomize its level sets, i.e., choosing α as random, i.e., consider $\alpha : (\Omega, \mathcal{A}, P) \to [0, 1]$, uniformly distributed, then we create a random set $S_f : (\Omega, \mathcal{A}, P) \to 2^U$, $S_f(\omega) = \{u : f(u) \geq \alpha(\omega)\}$. Then, clearly,

$$P(u \in S_f) = P(\omega : f(u) \geq \alpha(\omega)) = f(u)$$

i.e., the membership function f on U is the coverage function of the random set S_f. This correspondence could be used to manipulate membership functions somewhat according to probability calculus. Clearly, given a membership function f, there might exist many random sets whose coverage functions are equal to f, i.e., each fuzzy set determines an equivalent class of random sets. This fact is what we referred to as a weak form of randomness.

The above connection between fuzzy sets and random sets does not imply that fuzziness is subsumed by probability theory. However, it could suggest a way to obtain membership functions. Suppose we wish to obtain the membership function for "seriousness" of some illness. Suppose that the illness under consideration is manifested as subsets of the set of possible symptoms $\Omega = \{\omega_1, \omega_2, ..., \omega_k\}$. Let U be a set of humans. Let $S : \Omega \to 2^U$ be $S(\omega) = \{u \in U : u$ has the symptom $\omega\}$. A measure of seriousness of a person u could be some numerical measure of the set $\{\omega \in \Omega : u \in S(\omega)\}$. Medical experts often can provide assessments that can be described as a function $\mu : 2^\Omega \to [0, 1]$, with $\mu(A)$ being the degree of seriousness of the illness for a person having all symptoms in A. As such, a membership function can be taken as $f(u) = \mu\{\omega : u \in S(\omega)\}$.

(ii) Having defined fuzzy sets by their membership functions, we can view functions $U \to [0,1]$ as representing fuzzy sets. For example, we can view *randomized tests* as fuzzy sets. Specifically, recall that the fundamentals of the theory of testing statistical hypotheses are as follows (see e.g., Nguyen and Rogers [13]). Since our observable is $V = (X_1, X_2, ..., X_n)$, we look at the sample space $\mathcal{X} = \mathbb{R}^n$. Consider testing a simple hypothesis H_o against a simple alternative H_a. A statistical test is a rule for choosing one of H_o and H_a. It is sensible to base the test on V so that we select a set B as the critical region: if the observable $V \in B$, we reject H_o, otherwise, we do not reject H_o. By doing so, we commit two types of errors. Type I error (rejecting H_o when it is true): $\alpha = P(V \in B|H_o)$, and type II error (not rejecting H_o when H_a is true): $\beta = P(V \notin B|H_a)$. Since a decrease in α can be accomplished only be a "decrease" in B, accompagnized by an "increase" in B^c and β, thus in general, there is no way to decrease α and β simultaneously with V fixed. J. Neyman and E.S. Pearson in 1933 suggested a slight change in the problem. For fixed α, find the set B to minimize β (leading to the most powerful critical region B_*, where the power of a test is $1 - \beta$). If we denote by L_o and L_a the densities of the observable under H_o and H_a, respectively, then

$$\alpha = \int_B L_o ....., \quad \beta = \int_{B^c} L_a$$

The problem is to find B_* in $\mathcal{C} = \{B \in \mathcal{B}(\mathbb{R}^n)$ (Borel sets), $\int_B L_o \leq \alpha\}$ so that $\int_{B_*} L_o = \alpha$ and $\int_{B_*^c} L_a \leq \int_{B^c} L_a$ for all $B \in \mathcal{C}$.

The famous Neyman-Pearson lemma says this. Let $k > 0$ be a constant such that

$B_* = \{V \in \mathcal{X} : L_a(V) > kL_o(V)\}$ has $\int_{B_*} L_o = \alpha$, then $\int_{B_*} L_a \geq \int_B L_a$ for all $B \in \mathcal{C}$.

However, while the lemma shows that the given B_* is most powerful, it says nothing about whether such a k actually exists. The hypothesis of the lemma is a sufficient condition for the integral inequalities to hold. Now observe that

$$A(k) = P(\{V : L_a(V) > kL_o(V)|H_o) = P(\frac{L_a(V)}{L_o(V)} > k) = 1 - F(k)$$

where F denote the distribution function of the random variable $W = \frac{L_a(V)}{L_o(V)}$. If F is continuous then there is a k_α such that $1 - F(k_\alpha) = \alpha$ and the lemma applies to k_α. If F is not continuous (in particular if V is discrete) it can happen

that the given $1-\alpha$ "meets a jump" in F, and hence there is no solution to $A(k) = 1 - F(k)$. Suppose we have that

$$1 - F(k_o) < \alpha < 1 - F(k_o-)$$

with $P(W = k_o) = F(k_o) - F(k_o-) > 0$. Then the testing rule is: when $L_a(V) > k_o L_o(V)$, reject H_o; when $L_a(V) < k_o L_o(V)$, do not reject H_o; and when $L_a(V) = k_o L_o(V)$, reject H_o with probability $\pi_0 = \frac{\alpha - 1 + F(k_o)}{F(k_o) - F(k_o-)}$. With this rule, we do get the exact significance level prescribed. Note that many statisticians are not impressed by this randomized test procedure, and prefer the p-value approach!

In any case, the above randomized test (of size α) is described by a function $\tau : \mathcal{X} \to [0,1]$ with $E(\tau(V)|H_o) = \alpha$. Specifically,

$$\tau(V) = \begin{cases} 1 \quad \text{for} \quad L_a(V) > kL_o(V) \\ \pi_o \text{...for} \quad L_a(V) = kL_o(V) \\ 0 \quad \text{for} \quad L_a(V) < kL_o(V) \end{cases}$$

Thus, "formally" the usual randomized test, given as a function $\tau : \mathcal{X} \to [0,1]$, is a fuzzy set on $\mathcal{X}$. Note also that often in statistics we need to use linguistic terms to make decisions, such as "a rationale for rejecting a null hypothesis H_o is based on the fact that, if H_o is true, then it is "unlikely" that the data behave as we observed", where the linguistic term "unlikely" is a *fuzzy probability*, i.e., a fuzzy subset of $[0,1]$. But, as usual, we defuzzify it to $\alpha \in [0,1]$.

It is interesting to mention also that R.J. Aumann and L.S. Shapley [14], in their famous book *Values of Non-Atomic Games*, needed to generalize ordinary sets to obtain what they called "evenly spread sets" to arrive at their notion of values for games with large masses of players. As they put down in a footnote (page 142), their "ideal sets" are formally Zadeh's fuzzy sets.

We turn now to the manipulation of fuzzy sets. Note that what we have in mind is using fuzzy sets as our data in statistical analysis. As such, we need to operate them as generalized sets by extending ordinary set operations, as well as viewing them as "quantities" where extensions of arithmetic operations are required.

Before we extend logical operations of ordinary sets, we recall the reader that sets are viewed as propositions in a language. Binary logic is a logic for true-or-false propositions. Fuzzy logic is not a logic which is fuzzy! It is a logic

for fuzzy propositions, i.e. propositions (in a natural language) which could be true, false or partially true since they might contain fuzzy concepts. In other words, it is a multi-valued logic whose truth values lie in the unit interval $[0, 1]$.

Operations on ordinary sets can be expressed in terms on indicator functions. For $A, B \in 2^U$, we have

$$1_{A^c}(.) = 1 - 1_A(.);\ 1_{A\cap B}(.) = (1_A.1_B)(.) = (1_A \wedge 1_B)(.),\ 1_{A\cup B}(.) = (1_A \vee 1_B)(.)$$

where $x' = 1 - x$, $x \wedge y = \min\{x, y\}$, $x \vee y = \max\{x, y\}$.

Replacing indicator functions by membership functions for A, B fuzzy subsets of U, we can first consider the simplest fuzzy logic with fuzzy connectives for "not", "and" and "or", respectively as: negation of A is $1 - A(.)$, "A and B" is $(A \wedge B)(.)$, and "A or B" is $(A \vee B)(.)$. For example, a (finite) *fuzzy partition* of a set U is a collection $\{A_i, i = 1, 2, ..., k\}$ of fuzzy subsets of U such that $\sum_{i=1}^{k} A_i(u) = 1$ for all $u \in U$. Such a partition is usually used to coarsen the domain U (see next section).

More "sophisticated" fuzzy logics are defined in terms on general negation operators, t-norms and t-conorms (for details, see e.g., Nguyen and Walker [2]).

Extensions of arithmetic operations of numbers are generalized to sets and to fuzzy sets via the *extension principle* as follows.

Let $f : U \times V \to W$. For A, B fuzzy subsets of U, V, respectively. Then $f(A, B)$ is the fuzzy subset of W whose membership function is given by

$$f(A, B)(w) = \max_{\{(u,v):f(u,v)=w\}} (A(u) \wedge B(v))$$

For example, addition of fuzzy sets is obtained as

$$(A + B)(w) = \max_{\{(u,v):u+v=w\}} (A(u) \wedge B(v))$$

It is sometimes convenient to work with level sets of membership functions. The following result, known in the literature as *Nguyen's theorem* (Nguyen [15]; Fuller and Keresztfalvi [16]; Fuller [17]), is useful.

For $f : U_1 \times ... \times U_n \to V$, and $A^{(i)}$, $i = 1, 2, ..., n$, fuzzy subsets of U_i, and $\alpha \in [0, 1]$, we have

$$f(A_\alpha^{(1)}, ..., A_\alpha^{(n)}) = [f(A^{(1)}, ..., A^{(n)})]_\alpha$$

if and only if for each $v \in V$,

$$\max_{\{(u_1,\ldots,u_n)\in f^{-1}(v)\}} [\wedge_{i=1}^{n} A^{(i)}(u_i)]$$

is attained.

As a final note, for representation of fuzzy information on computer, see Nguyen and Kreinovich [18].

5. Where Do We Run into Fuzzy Data?

We come now to what you are waiting for! OK, fuzzy concepts in natural language are informative (for scientific investigations) and can be modeled mathematically as fuzzy sets, but where do we ”run into” fuzzy data in applied statistics?

At least as far as applying statistics to economics is concerned, econometricians borrow almost all tools and methodologies from physical sciences, especially engineering, by obvious reasons: like physical systems, economic systems are uncertain, dynamical systems. The ”borrowed” tools include Kalman filter, quantum mechanics (path integrals and Hamiltonian for options and interest rates, see e.g., Baaquie, B.E. [19]).

Since 1965, approaches to modeling and control based on fuzzy methodology are familiar in engineering, including a combined method from machine learning, namely *adaptive neural fuzzy inference systems (ANFIS)* (see. e.g., Nguyen, Prasad, Walker and Walker [20]). The situation is somewhat different in applied statistics, especially in econometrics. Specifically, why fuzzy analysis in engineering fields seems not to be in the ”mean stream” of statistics in general, and econometrics, in particular?

In our view, one of the main reasons is that the various attempts to bring in fuzzy analysis to statistics did not possess an underlying rigorous probability theory to support them. Some approaches treat fuzzy data without relations to randomness, while others tend to ”fuzzify” statistics rather than putting fuzzy data within standard theory of statistics. In either case, there is *no statistical inference involved*, so that it looks like applications without theory! A state-of-the-art of the above description is contained in a special issue of the Journal *Computational Statistics and Data Analysis*, Volume 51 (2006), devoted to ”The fuzzy approach to statistical analysis”. In it, the only contributed paper by Nguyen and Wu [21] is about ”statistics with fuzzy data”, i.e., viewing fuzzy

data as bona fide statistical data and treating them within the theory of probability. Indeed, in the section "An overview of the contributions to this issue", the guest editors wrote *"The contribution of Nguyen and Wu focuses on Fuzzy Statistics seen as "Statistics with fuzzy data". In this specific context, statistical data may be point-valued, set-valued or fuzzy-set valued observations. Random sets (viewed as elements of separable metric spaces) are proposed at the appropriate mathematical model for set-valued observations. Likewise, random fuzzy sets are suitable for analyzing random fuzzy data. Some issues related to these modelization are examined. The proposals stemming from this investigation may help in strengthening the bases of a sound methodology for the analysis of imprecise data. In particular, the notion of "coarsening" is thoroughly discussed along with the process of generation of a membership function from available information".*

Another important reason is that, not only real-world situations where fuzzy data surface seem lacking, but also, when such a situation did occur, there is no serious discussions about the advantage of considering fuzzy data, rather than just defuzzifying them.

To elaborate on this reason, let's look back at how statisticians used to handle linguistic variables, in the popular tool of linear regression (see a cook book like Kutner et al. [7]).

Linguistic variables could be intrinsically linguistic (e.g., ability in mathematics, skill of workers), i.e., variables whose values can only be described in linguistic terms, or due to *coarsening schemes* (for more on coarsening, see Nguyen [22]). The second kind of linguistic variables is very important in data analysis. An essential aspect of human intelligence, say, in making every day life decisions, is coarsening domains of numerical variables. When we can not guess with precision the temperature at some location, we coarsen a domain $[a, b]$ of the variable "temperature", i.e., transforming it into a *fuzzy partition*, such as "very cold , cold, medium, hot, very hot", in order to obtain a correct, but imprecise, useful information. Formally, a fuzzy partition of a set U is a collection $\{A_i : i = 1, 2, ...n\}$ of fuzzy subsets of U such that the sum of their membership functions is one: $\sum_{i=1}^{n} A_i(u) = 1$, for all $u \in U$. This is clearly a generalization of the ordinary concept of a (crisp) partition of a set. This explains why in statistics we consider linguistic values like "very low, low, normal, high, excessive" for the variable "unemployment rate" in regression analysis,

for example. By doing so, we actually transform a quantitative variable into a fuzzy variable (in the sense that the values of the latter are fuzzy sets). But, that transformation, in classical statistics, is only for the purpose of classification to collect (counting) data, and not viewing these linguistic values as data per se.

Here is an example of using "quantitative indicators" in regression with a qualitative predictor. In the regression of advertizing expenditures X (quantitative) on the quality of sales management Y (qualitative with two values "low, high"), the quantitative indicator of Y is

$$Y = \begin{cases} 1 & \text{if the quality of sales management is high} \\ 0 & \text{otherwise} \end{cases}$$

Well, how "high" is defined here to obtain the indicator of Y?! Of course, "high" is defuzzified by using some threshold.

What we have in mind when talking about fuzzy data is at least twofold. First, even in classical problems as above, defuzzification might entail loss of information. Is there a better way then sharp defuzzification? e.g., some smooth procedures. This is perhaps the main reason in using fuzzy modeling in the newly developed *Regression-Discontinuity Analysis*, see Klaauw [23].

Secondly, as we will illustrate below, there are important situations where fuzzy data need to be treated as data, i.e., just like a random sample of numerical observations, so that manipulation, processing of them are necessary. This is not considered in classical regression with qualitative variables. In fact, that is impossible since there is no fuzzy modeling available.

Below is just a short list of problems where either fuzzy data appear as valuable information, or fuzzy theory is helpful to assist statistical analysis/decision-making. Typical problems such as these could trigger research for interested applied statisticians.

Clearly, the first class of problems contains *regressions with linguistic variables* (regressors or responses, or both). Regressions with imprecise data are special cases.

The second class of problems can be viewed as regression with *seemingly unobservable variables (SUV).* In labor economics, the studies of effects of covariates (such as education level, skill) on a response variable (such as salaries of workers) are conducted using linear regression models, in which the unobservable covariate "skill" is often ignored. If this covariate is to be taken

into account, it will be a linguistic variable by its own nature. Clearly, taking it into account should shed more light on the response variable, since it is (also clearly!) that skill is an important determinant for workers' salaries. Usually, interviews for jobs could reveal applicants' skill. In our view, this class of problems is typical for having fuzzy data (on the seemingly unobservable variables) with which regression with fuzzy data could be developed, thus, improving conventional models in econometrics.

As a concrete example, the problem of finding out how underground economy (u.e.) affects national economy is important, and has been, as a start, investigated by Draeseke and Giles [24] and Ene and Hurduc [25]. Note that these authors only considered fuzzy methods to "estimate" u.e. size but did not pursue an SUV regression using statistics of fuzzy data.

Regression of u.e. on national economy is a problem of regression with unobservable covariate. How can we even consider such a regression, although it is an important economic problem? If such a regression is to be feasible, we need data on u.e.. But, by definition (!), u.e. is not reported (to evade taxes), and hence "unobservable". It is precise here that fuzzy data is needed. Not knowing the u.e. size, we could seek to estimate it from several obvious causal variables (e.g., taxation rate, GDP per capital, unemployment rate), resulting in fuzzy data for the estimate. A generated time series (say, annually) of fuzzy data on u.e. then could be used as inputs to the original SUV regression.

Let's elaborate a bit on how to estimate u.e. size using fuzzy methods. Since the "regression" of causal variables on the response variable "size of u.e." can not be put in the standard format of a linear regression model, we make the following observation. A standard linear regression model $Y_i = aX_i + b + \varepsilon_i$, $i = 1, 2, ..., n$, is in fact a collection of "If...Then.." rules, since that equation (in fact any equation) reads: "If X is X_i then Y is $Y_i = aX_i + b + \varepsilon_i$". Thus, when (X_i, Y_i) is linguistic (modeled by fuzzy sets (A_i, B_i)), the "rule" becomes

$R_i =$"If X is A_i then Y is B_i" or "$A_i(X) \Longrightarrow B_i(Y)$". For example " If taxation is *high* and unemployment is *low*, then the size of u.e. is *medium*". These rules are, in a sense, common sense. For reasoning with rules that have exceptions, see Bamber et al. [26].

Given $Y_i = aX_i + b + \varepsilon_i$, $i = 1, 2, ..., n$, we "combine" them to arrive at a prediction "formula" for Y when observing X, by using a statistical procedure, namely least squares. Now, the rule base $\{R_i = A_i(X) \Longrightarrow B_i(Y), i =$

$1, 2, ...n\}$ is combined as follows to form a *nonlinear model* for prediction, inspired from fuzzy control methodology in engineering (see e.g., Nguyen et al [20].

First, the rule $A_i(X) \Longrightarrow B_i(Y)$ represents a fuzzy relation between X and Y, i.e., for a value of X, the degree to which Y is compatible with the rule is , say, $A_i(X) \wedge B_i(Y)$. Combining these fuzzy sets leads to a fuzzy set representing the "qualitative" predictor for Y, given a new X: $\max_{i=1,2,...,n}(A_i(X) \wedge B_i(Y))$.

A third class of problems consists of introducing fuzzy data (in the form of membership functions of conventional data) into classical statistical setting in order to improve the performance of statistical procedures. The typical problem is *statistical quality control.*

Current research in Statistical Quality Control (SQC) addresses the more realistic statistical models in which characteristics of manufacturing products need not follow multivariate normal distributions. In other words, the research aims at deriving tolerance regions (leading to control charts) in the setting of multivariate, nonparametric models. This is carried out by recognizing that traditional tolerance regions are nothing else than level sets of probability density functions. The recent paper by Verdier [27] brings out the usefulness of using copulas in modern SQC. The multivariate SQC (see Montgomery [28] is essentially based on (parametric) normal distributions.

In the univariate case, Shewhart in 1924 first observed that, if the (single) product characteristic is modeled by a random variable X (due to its possible variations), then we can detect whether it is "out of range" (out-of-control) if the new value is far away from its mean $\mu = EX$ by 3 standard deviation $\sigma = \sqrt{Var(X)} = \sqrt{E(X-\mu)^2}$, by using Chebyshev's inequality:

$$P(|X - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$$

For example, for $k = 3$,

$$P(|X - \mu| \leq 3\sigma) \geq 0.8889$$

Remark. Using extension of Chebyshev's inequality in higher dimensions (i.e., for random vectors), similar assessments can be obtained.

If we insist that X is normal $N(\mu, \sigma^2)$, then the above lower bound is more accurate, namely

$$P(|X - \mu| \leq 3\sigma) \geq 0.997$$

so that the interval $[\mu - 3\sigma, \mu + 3\sigma]$ could be used as a "tolerance" zone for the variations of X. Specifically, since $(P(|X-\mu| > 3\sigma)$ is so small, it is unlikely that a value of X in $|X - \mu| > 3\sigma$ could come from X. Of course, false alarms could arise!

The following observation is essential for considering multivariate SQC when traditional multivariate normal distribution assumption is dropped.

Remark. It is important to remember that making too much model assumptions takes us further from realities! The task of a statistician is trying to obtain models as general as possible.

If we look at the tolerance interval $[\mu - 3\sigma, \mu + 3\sigma]$, we realize that it is precisely the set

$$\{x \in \mathbb{R} : f(x) \geq c\}$$

where

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

and $c = f(\mu + 3\sigma)$, with $\mu + 3\sigma$ being a quantile of X.

Thus, for general multivariate (joint) density function f, a tolerance region is of the form

$$\{x \in \mathbb{R}^d : f(x) \geq c\}$$

In the univariate normal distribution case, since both μ, σ are unknown, the tolerance interval $[\mu - 3\sigma, \mu + 3\sigma]$ is estimated by $[\bar{X}_n - 3S_n, \bar{X}_n + 3S_n]$, where $\bar{X}_n$ and S_n are sample mean and sample standard deviation of an i.i.d. random sample $X_1, X_2, ..., X_n$ drawn from X. Note that, in the multivariate normal case, we use T^2-Hotelling statistic.

In general, the meaning of

$$\{x \in \mathbb{R}^d : f(x) \geq c\}$$

is that

$$P(f(X) \geq c_\alpha) = \alpha$$

i.e., the probability that a new observation, say, $X_{n+1}(\omega)$, is in the level set is some predetermined α.

Now, of course, the joint density f on $\mathbb{R}^d$ (e.g., when the manufacturing product depends on d (related) characteristics) is unknown. As such, the (population) "parameter" $\{x \in \mathbb{R}^d : f(x) \geq c\}$, which is a set, needs to be estimated (by some set statistics, i.e., random sets). Such a set statistic is the

statistical tolerance region for deriving multivariate control charts. Basically, it consists of, first, estimating the joint density f, nonparametrically by (say, using Kernel method) f_n, then "plug-in" to obtain the statistical tolerance region $\{x \in \mathbb{R}^d : f_n(x) \geq c\}$.

Verdier [27] suggested using copulas to model process control in which a semiparametric model is appropriate. It consists of two steps: First, estimate nonparametrically the marginals densities, second: use some appropriate parametric family of copulas, and use Sklar's theorem to obtain multivariate models.

With the above direction of research towards more realistic control charts, we add to it another way to improve SQC by looking at observed data. As mentioned previously, data should be classified according to their degrees of "in or out of control". As such, we are in fact "fuzzify" (as opposed to defuzzify) crisp data, also in order to obtain more realistic control charts.

Another situation where fuzzifying concepts (rather than data) is useful is in financial econometrics. The following example brings out also an advantage of using fuzzy methodology in investing economic problems, showing that, in some cases, fuzzy theory is really indispensable.

The capital risk allocation (CRA) is an important problem in financial risk management (see, e.g. Denault, [29]). The most recent approach to the solution of CRA is based upon coalitional game theory, since cost functions can be expressed in terms of characteristic functions of such games. Unfortunately, the Shapley value cannot be inside the core of the game for coherent risk measures. It was suggested that extending (crisp) coaltional games to *fuzzy games* (Aubin [30,31]) could lead to a solution.

Extending a coalitional game to a fuzzy game is in the same spirit of SQC above. Here, it is conceivable that members of a coalition might not always commit their full resources when joining a coalition. As such, degrees of participation in a coalition should be taken into account, for a fair capital risk allocation. When doing so, we actually consider fuzzy coalitions, thus, enlarging the coalitional game.

6. How to Put Fuzzy Data in a Statistical Setting?

This last section is rather theoretical. But it is a must since we need to put fuzzy data in a firm footing entirely within probability theory in order to use its

associated statistical theory for making inferences in an "acceptable" manner. Applied statisticians will have an opportunity, not only to know the theory before applications, but also to learn how "theoretical" statisticians conduct their research!

The problem is how to "view" a new kind of data, namely fuzzy data, as statistical data? It is here that we need probability theory which is the "background" of statistics. It is in a situation such as this that we need to develop a new theory before considering applications. The new theory is inspired from empirical observations, and is not from an "abstract" generalization. In this sense, according to our view expressed in section 3, it is applied statistics.

Since fuzzy sets generalize crisp (ordinary) sets, let's us start out with the theory of *random sets.* Roughly speaking, a random set is a set obtained at random! Here is an interesting example of random set observations which need theoretical justifications to use in statistics.

A well-known graphical method in exploratory data analysis to test ("informally") the goodness-of-fit of a sample (e.g., that the sample came from a normal distribution) is the Quantile-Quantile plot (QQ plot). Let $(X_1, X_2, ..., X_n)$ be a random sample drawn from a population X with unknown distribution function F. To see whether the sample comes from a specific distribution F^o, we compare various quantiles of F^o with corresponding empirical quantiles, i.e., for $\alpha \in (0, 1)$, compare

$$q_\alpha(F^o) = \inf\{x \in \mathbb{R} : F^o(x) \geq \alpha\}$$

with $q_\alpha(F_n)$ where the empirical distribution function is

$$F_n(x|X_1, X_2, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} 1_{(X_i \leq x)}$$

Specifically, we plot $q_\alpha(F_n)$ versus $q_\alpha(F^o)$ for several values of α. Note that, unlike moments, a distribution function is *characterized* by its quantiles, i.e., can be recovered from its quantiles. As such, coincidence of quantiles is a good *indication* for goodness-of-fit. In the QQ plot, the indication of a good fit is detected when the QQ plots "hugs" a straight line through the origin at an angle of 45 degrees (as often said in the literature, this is a "quick and dirty" way of doing statistics!). We have just said that such an inspection reveals only an indicative conclusion. It is not a conclusive conclusion, i.e., not "rigorous". Again, this is a good place to remind applied statisticians about validity

of empirical analyses! Empirical analyses need to be validated to draw final conclusions. How to ”validate” the QQ plot? Well, if the sample did come from F^o, then the QQ plot should converge to a straight line as the sample size increases. Here the ”target” is a straight line, a subset of $\mathbb{R}^2$ and the QQ plot is a sequence of random subsets of $\mathbb{R}^2$ (see below). As such, we need to know what do we mean by ”convergence of a sequence of random sets to another set?”!. If you look at your statistical ”tool box”, you cannot find any ”tool” in it to investigate the needed problem. It is so since ”standard” data are either numbers or vectors, but not sets, and we do not have a theory of random sets in probability theory for statistical inference. We need to pause to do some theoretical research and then come back to ”solve” our needed statistical problem of validating the simplest goodness-of-fit test. It was Das and Resnick [32] who investigated this problem in a formal theory of random sets.

Remark. Historically, while random sets appeared naturally in many places, such as stochastic geometry, its formal theory was not rigorously established until 1975 (by Matheron [33]). When estimating the ”size” (area, volume) of a random set, Robbins [34] did not really consider a formal concept of a random set. This so since the size of a random set $\mu(S)$ (where μ the Lebesgue measure on $\mathbb{R}^d$) is in fact a numerical random variable, although it depends on the random set S. Without a formal concept of random sets, it is not possible to find the distribution of the nonnegative random variable $\mu(S)$ which is a function of S. The clever result of Robbins is this. As far as the expected value of $\mu(S)$ is concerned, we need much less than the distribution of $\mu(S)$. Specifically, the knowledge on the coverage function of the informal random set S is sufficient to determine $E\mu(S)$, a ”weaker” form of information. Note that if S is a *confidence interval,* which is a random set, (say, at $1-\alpha$ confidence level), then $\mu(S)$ is its length, and an optimal confidence interval is the one with smallest length (maximum precision at a given confidence level). The computation of the expected length of a random set of the form $S=[0,X]$ where $X \geq 0$ is a random variable is simple: the length of S is X, so that

$$E\mu(S) = EX = \int_0^\infty P(X > x)dx = \int_0^\infty \pi(x)dx$$

where

$$\pi(x) = P(x \in S) = P(x \in [0,X]) = P(X > x)$$

is the coverage function of the random set S. The Robbins' formula says that

the above formula is in fact general: For any random set S on $\mathbb{R}^d$, we have

$$E\mu(S) = \int_{\mathbb{R}^d} \pi(x) d\mu(x)$$

where $\pi(.) : \mathbb{R}^d \to [0,1]$ is the coverage function of S.

Now, while the $q_\alpha(F^o)$ are deterministic, the $q_\alpha(F_n)$ are random (depending upon the sample), and as such the points $(q_\alpha(F_n), q_\alpha(F_n))$, for various α, are random points in the plane, forming a random set (of points). Since the sample is finite, this random set is a *closed* subset of the plane $\mathbb{R}^2$. Note that a straight line in $\mathbb{R}^2$ is also a closed set, and hence we are facing the problem of, say, almost sure convergence of *random closed sets*.

We now outline Matheron's theory of random closed sets on euclidean spaces $\mathbb{R}^d$.

The general framework of probability is this. We always consider an abstract probability space $(\Omega, \mathcal{A}, P)$ on which all random elements are defined. To specify a type of random elements X, we specify a measurable space $(U, \mathcal{U})$ consisting of a set U which is the range of the random element X we have in mind, and $\mathcal{U}$ a suitable $\sigma-$field of subsets of U (for the domain of the probability measures governing the random evolution of X, elements of $\mathcal{U}$ are events).

For example, if $U = 2^V$ (power set of V), with V being a finite set, i.e., U is the set of all subsets of a finite set V, then just take $\mathcal{U} =$ power set of 2^V since probability measures can be defined on such $\mathcal{U}$: simply assign $Q(A) \in [0,1]$ such that $\sum_{A \subseteq V} Q(A) = 1$, and define $P(.)$ on 2^{2^V} by $P(\mathbb{A}) = \sum_{A \in \mathbb{A}} Q(A)$.

As for $U = \mathbb{R}$, an infinite, uncountable set, the situation is more delicate. The power set of $\mathbb{R}$ is too big to define probability measures on it. We seek a largest collection $\mathcal{U}$ of subsets of $\mathbb{R}$ (but strictly contained in the power set of $\mathbb{R}$) to be the domain of all probability measures. Inspired by measure theory in real analysis, it turns out that there is a canonical way of getting such $\mathcal{U}$. We equip $\mathbb{R}$ with a *topology* (i.e., declare a collection $\mathcal{O}$ of subsets as "open" sets). For $\mathbb{R}$, the canonical topology is the smallest collection of subsets containing the open intervals (a, b). Then take the smallest $\sigma-$field containing all open sets, denoted as $\mathcal{B}(\mathcal{O})$ (we also say that it is the $\sigma-$field generated by $\mathcal{O}$). The "canonical" $\sigma-$field obtained this way is refered to as the Borel $\sigma-$field associated with the topology $\mathcal{O}$. Just to be self-contained, a $\sigma-$field is a collection of subsets, suitable for defining probability measures on it. A collection $\mathcal{B}$ of subsets (events) of a set U is a $\sigma-$field if it satisfies the conditions (i) $U \in \mathcal{B}$,

(ii) If $A \in \mathcal{B}$ then its complement $A^c \in \mathcal{B}$, and (iii) For any countable collection of elements of $\mathcal{B}$, $\{A_n, n \geq 1\}$, $\cup_{n\geq 1} A_n \in \mathcal{B}$.

Now, consider $U = \mathcal{F}(\mathbb{R}^d)$, the set of closed subsets of $\mathbb{R}^d$. We will proceed to equip U with a topology τ and take $\mathcal{U} = \mathcal{B}(\tau)$.

Let $\mathcal{F}, \mathcal{G}, \mathcal{K}$ denote the classes of closed, open and compact subsets of $\mathbb{R}^d$, respectively. For $A \subseteq \mathbb{R}^d$, let

$$\mathcal{F}_A = \{F \in \mathcal{F} : F \cap A \neq \varnothing\}, \;\; \mathcal{F}^A = \{F \in \mathcal{F} : F \cap A = \varnothing\}$$

$$\mathcal{F}^K_{G_1 G_2,...,G_n} = \mathcal{F}^K \cap \mathcal{F}_{G_1} \cap \mathcal{F}_{G_2} \cap ... \cap \mathcal{F}_{G_b}$$

$$\mathbb{B} = \{\mathcal{F}^K_{G_1 G_2,...,G_n} : K \in \mathcal{K}, G_i \in \mathcal{G}, n \neq 0\}$$

Let τ be the topology generated by the base $\mathbb{B}$. This topology is called the hit-or-miss topology of $\mathcal{F}$. The associated Borel $\sigma-$field is denoted as $\mathcal{B}(\mathcal{F})$

Definition. Let $(\Omega, \mathcal{A}, P)$ be a probability space. By a *random closed set* on $\mathbb{R}^d$, we mean a map $X : \Omega \to \mathcal{F}$ which is $\mathcal{A} - \mathcal{B}(\mathcal{F})-$ measurable. The probability law of X is the probability $P_X = PX^{-1}$ on $\mathcal{B}(\mathcal{F})$, i.e., for $\mathbb{A} \in \mathcal{B}(\mathcal{F})$, $P_X(\mathbb{A}) = P(X \in \mathbb{A})$.

For an elementary exposition on the whole theory of random closed sets, the reader can consult Nguyen [6]. Here, we just indicate the counterpart of Lebesgue-Stieltjes theorem for random closed sets. Observe that, if we define $T : \mathcal{K} \to [0, 1]$ by

$$T(K) = P(\mathcal{F}_K) = P(F \in \mathcal{F} : F \cap K \neq \varnothing\}$$

then T satisfies the following axioms

(1) $T(\varnothing) = 0$

(2) T is alternating of infinite order, i.e., for any $n \geq 2$ and $K_1, K_2, ..., K_n$ in $\mathcal{K}$,

$$T(\cap_{i=1}^n K_i) \leq \sum_{\varnothing \neq I \subseteq \{1,2,...,n\}} T(\cup_{i\in I} K_i)$$

(3) If $K_n \searrow K$ in $\mathcal{K}$, then $T(K_n) \searrow T(K)$

Any function $T : \mathcal{K} \to [0, 1]$ satisfying the above three axioms is called a *capacity functional.* Capacity functionals play the role of distribution functions of

random variables. The collection of closed sets $\mathcal{F}_K = \{F \in \mathcal{F} : F \cap K \neq \varnothing\}$ plays the role of intervals $(-\infty, y]$ on the real line, in the determination of the distribution function of a real-valued random variable Y: $F_Y(y) = P(Y \leq y) = P_Y((-\infty, y])$.

Like Lebesgue-Stieltjes theorem, the following result simplifies the search for probability laws governing random evolution of random sets.

Choquet Theorem. If $T : \mathcal{K} \to [0, 1]$ is a capacity functional, then there exists a unique probability P on $\mathcal{B}(\mathcal{F})$ such that $P(\mathcal{F}_K) = T(K)$ for all $K \in \mathcal{K}$.

Here is another important situation in *set estimation* where the theory of random (closed) sets is essential. Let Y be a real-valued random variable. Suppose that its distribution function F is absolutely continuous so that its probability density function f exists. Given a random sample $(Y_1, Y_2, ..., Y_n)$ drawn from Y, of course we can estimate F (pointwise) by the empirical distribution function $F_n(y) = \frac{1}{n} \sum_{i=1}^{n} 1_{(Y_i \leq y)}$. However, we cannot derive an estimator for $f(.)$ from it since the (a.e.) derivative of $F_n(.)$ is identically zero. Nonparametric estimation of probability density functions reveal more information than that of their distribution functions. There are various methods for *nonparametric estimation of density functions* (e.g. kernel, orthogonal functions,...) which require lots of assumptions. An alternative approach was suggested by Hartigan [35] when only some simple qualitative information is available. Let Y be a random vector with values in $\mathbb{R}^d$. Since its (unknown) density f can be recovered from its level sets $A(\alpha) = \{y \in \mathbb{R}^d : f(y) \geq \alpha\}$, $\alpha \geq 0$, as

$$f(y) = \int_0^\infty 1_{A(\alpha)}(y) d\alpha$$

It suffices to estimate the sets $A(\alpha)$, by, of course, some *random set estimator* $A_n(\alpha)$, and use the plug-in estimator

$$f_n(y) = \int_0^\infty 1_{A_n(\alpha)}(y) d\alpha$$

as estimator of $f(.)$.

What is not obvious is how to suggest a "good" random set estimator $A_n(\alpha)$, i.e., a *consistent estimator* of the set $A(\alpha)$. The following idea, called the excess mass approach, is due to Hartigan. Recall the way *extremum estimators* in statistics are derived: if a population parameter optimizes a theoretical objective function, then a plausible estimator for it is the statistic optimizing the empirical counterpart of that objective function. Let μ denote the Lebesgue measure

on $\mathbb{R}^d$ and consider the signed measure $(dF - \alpha\mu)(.) = \varepsilon_\alpha(.)$ on $\mathcal{B}(\mathbb{R}^d)$. For $B \in \mathcal{B}(\mathbb{R}^d)$, writing $B = [A(\alpha) \cap B] \cup [A^c(\alpha) \cap B]$, we have $\varepsilon_\alpha(B) \leq \varepsilon_\alpha(A(\alpha))$, so that $A(\alpha)$ maximizes the objective function $B \to \varepsilon_\alpha(B)$. The empirical counterpart of $\varepsilon_\alpha(.)$ is $\varepsilon_{\alpha,n}(.) = (dF_n - \alpha\mu)(.)$, so that, a plausible estimator of $A(\alpha)$ could be the random set statistic $A_n(\alpha)$ maximizing $\varepsilon_{\alpha,n}(B)$ over all $B \in \mathcal{C} \subseteq \mathcal{B}(\mathbb{R}^d)$, where $\mathcal{C}$ is some specified class of Borel sets, such as closed convex sets, ellipsoids. How to "solve" this set-function optimization? See, Nguyen [6]. To establish the consistency of $A_n(\alpha)$, we need a formal theory of random sets.

We turn now to *random fuzzy (closed) sets.* First, a fuzzy subset of $\mathbb{R}^d$, say, is a function $A(.) : \mathbb{R}^d \to [0, 1]$. Since a crisp subset A is a closed set if and only if it indicator function $1_A(.)$ is upper semicontinuous (u.s.c.), i.e., for any $\alpha \geq 0$, its level set $\{y : 1_A(y) \geq \alpha\}$ is a closed set, we say that the fuzzy set $A(.)$ is a *fuzzy closed subset* of $\mathbb{R}^d$ if it is u.s.c., i.e., for any $\alpha \geq 0$, $\{y : A(y) \geq \alpha\} \in \mathcal{F}(\mathbb{R}^d)$, the set of closed subsets of $\mathbb{R}^d$. We denote by $\mathcal{F}^*(\mathbb{R}^d)$ the set of fuzzy closed subsets of $\mathbb{R}^d$.

How to extend Matheron's hit-or-miss topology to $\mathcal{F}^*(\mathbb{R}^d)$?

Remark. We take this opportunity to say an important thing. Fuzzy sets are "new" mathematical objects. If we are going to talk about topology, it should be (ordinary) topology of fuzzy sets, and *not* "fuzzy topology", i.e., a "new" concept of topology generalizing ordinary topology in mathematics! a kind of "topology" where ordinary neighborhoods of "points" in $\mathbb{R}^d$ become fuzzy. This should be so since the objects under considerations are fuzzy sets and not points in $\mathbb{R}^d$. This applied also to the wrong approach by trying to treat fuzzy data in a context of "fuzzy statistics"! Fuzzy statistics should be (like fuzzy logics) ordinary statistics of fuzzy data, and not "fuzzifying ordinary statistics"!

Now, again, a direct extension of Matheron's topology for (crisp) closed sets seems difficult. We need an equivalent way of looking at his topology for the purpose of extension. Note that this is a "routine" in mathematical investigation, as far as extensions of concepts, theories, are concerned. This includes the extension of Black-Scholes option pricing formula in financial econometrics, based on PDE: it is the equivalence in terms of martingales which allows extensions.

Another look at the hit-or-miss topology of closed sets is this. If we consider the set containment $\supseteq$ as a partial order relation on $\mathcal{F}$, i.e., for $A, B \in \mathcal{F}$, we

say that B is "less informative" than A if $B \supseteq A$ (B contains A, the reverse of standard partial order relation among sets), then $(\mathcal{F}, \supseteq)$ happens to be a *continuous lattice* (see, Gierz et al [36]). As such, there is a canonical topology, called Lawson topology, generated by $\supseteq$. Without going into technical details, we simply say that this Lawson topology is precisely the Matheron topology, see however, Nguyen and Tran [37]. Thus, the $\sigma-$ field, for defining random closed sets, is the Borel $\sigma-$field of the Lawson topology.

Now, on $\mathcal{F}^*(\mathbb{R}^d)$, if we consider the partial order relation among u.s.c. functions : f is "less" than g if $f(.) \geq g(.)$ (as extension of $\supseteq$ among sets), then, $(\mathcal{F}^*(\mathbb{R}^d), \geq)$ is also a continuous lattice, and as such we simply take the Borel $\sigma-$field $\mathcal{B}(L)$ of its Lawson topology to define random fuzzy (closed) sets.

A random fuzzy (closed) set is then a map $\Omega \to \mathcal{F}^*$, $\mathcal{A}-\mathcal{B}(L)-$ measurable.

For the extension of Choquet theorem to random fuzzy sets, see Nguyen, Wang and Wei [38]. See also Nguyen et al. [39].

The point is this. We have put fuzzy data in a rigorous theory of probability from which standard concepts for random elements can be meaningfully defined for statistical inference. As such, applied statisticians can be "confident" that their empirical works are supported firmly by theory.

Note: Dedicated to Lotfi Zadeh.

References

[1] Zadeh, L., Fuzzy sets, *Information and Control*, 1965; 8: 338-353.

[2] Nguyen, H. T. and Walker, E. A. *A First Course in Fuzzy Logic*, 3^{rd} Edition, Chapman and Hall/CRC, 2006.

[3] Nguyen, H. T. and Sugeno, M., *Fuzzy Systems: Modeling and Control*, Kluwer Academic, 1998.

[4] Lindstrom, T., A fuzzy design of the willingness to invest in Sweden, *J. Econ. Behav. Organization*, 1998; 36: 1-17.

[5] Hajek, K., *Sampling From a Finite Population*, Marcel-Dekker, 1981.

[6] Nguyen, H. T., *An Introduction to Random Sets*, Chapman and Hall/CRC, 2006.

[7] Kutner, M. H. , Nachtsheim, C. J. and Neter, J., *Applied Linear Regression Models*, McGraw-Hill, 2004.

[8] Resnick, S. I., *Heavy-Tail Phenomena*, Springer, 2007.

[9] Koenker, R., *Quantile Regression*, Cambridge University Press, 2005.

[10] Kumbhakar, S. C. and Knox Lovell, C.A., *Stochastic Frontier Analysis*, Cambridge University Press, 2003.

[11] Heckman, J., Sample selection bias as a specification error, *Econometrica*, 1979; 47: 153-161.

[12] Wheelan, C., *Naked Statistics*, W.W. Norton, 2013.

[13] Nguyen, H. T. and Rogers, G. S., *Fundamentals of Mathematical Statistics, Volume II: Statistical Inference*, Springer-Verlag, 1989.

[14] Aumann, R.J. and Shapley, L.S., *Values of Non-Atomic Games*, Princeton University Press, 1974.

[15] Nguyen, H. T., A note on the extension principle for fuzzy sets, *J. Math. Anal. and Appl.*, 1978; 64: 369-380.

[16] Fuller, R. and Keresztfalvi, T., On generalization of Nguyen's theorem, *Fuzzy Sets and Systems*, 1991; 41: 371-374.

[17] Fuller, R., On generalization of Nguyen's theorem: A short survey of recent developments, in *Advances in Soft Computing , Robotics and Control*, Springer, 2014: 183-190.

[18] Nguyen, H. T. and Kreinovich, V., How to fully represent expert information about imprecise properties in a computer system: random sets, fuzzy sets, and beyond: an overview, *Intern. J. General Systems* (to appear), 2014.

[19] Baaquie, B. E., *Quantum Finance,* Cambridge University Press, 2007.

[20] Nguyen, H. T. , Prasad, N. R., Walker, C. and Walker, A., *A first Course in Fuzzy and Neural Control*, Chapman and Hall/ CRC, 2003.

[21] Nguyen, H. T. and Wu, B., Random sets and fuzzy sets in coarse data analysis, *Comp. Statist. and Data Anal*, 2006; 41: 70-85.

[22] Nguyen, H. T., Survey sampling revisited and coarse data analysis, *Thailand Statistician*, 2004; 2: 1-19.

[23] Klaauw, W. van der, Regression-Discontinuity Analysis: A survey of recent developments in economics, *Journal of Compilation*, 2008: 219-245.

[24] Draeseke, R. and Giles, D. E. A., A fuzzy logic approach to modelling the New Zeland underground economy, *Math. and Comp. in Simul.*,2002; 59, 115-123.

[25] Ene, C. M. and Hurduc, N., A fuzzy model to estimate Romanian underground economy, *Internal Auditing and Risk Management*, 2010; 2(18): 1-10.

[26] Bamber, D. , Goodman, I. R. and Nguyen, H. T., Robust reasoning with rules that have exceptions, *Ann. Math. Art. Intell*, 2005; 45: 83-171.

[27] Verdier, G., Application of copulas to multivariate control charts, *J. Statist. Planning and Inference*, 2013; 143: 2151-2159.

[28] Montgomery, D.C., *Introduction to Statistical Quality Control*, 6^{th} Edition, J. Wiley, 2009.

[29] Denault, M., Coherent allocation of risk capital, *J. Risk*, 2001; 4: 1-34.

[30] Aubin, J. P.,*Optima and Equilibria*, Springer-Verlag, Aubin, 1993.

[31] Aubin, J. P., Cooperative fuzzy games, *Math. Oper. Res.*, 1981; 6(1): 1-13.

[32] Das, B. and Resnick, S. I. QQ Plots, random sets and data from a heavy tailed distribution, *Stochastic Models* , 2008; 103-132.

[33] Matheron, G., *Random Sets and Integral Geometry*, J. Wiley, 1975.

[34] Robbins, H. E., On the measure of a random set, *Ann. Math. Statist.*, 1944; 14: 70-74.

[35] Hartigan, J. A., Estimation of a convex density contour in two dimensions, *J. Amer. Statist. Assoc.*, 1975; 82: 267-270.

[36] Gierz, G. et al., *Continuous Lattices and Domains*, Cambridge University Press, 2003.

[37] Nguyen, H. T. and Tran, H., On a continuous lattice approach to modeling of coarse data in systems analysis, *J. Uncertain Systems*, 2007; 1(1): 62-73.

[38] Nguyen, H. T., Wang, Y. and Wei, G., On Choquet theorem for upper semi-continuous functions, *Intern. Journ. Approximate Reasoning*, 2007; 46: 3-16.

[39] Nguyen, H. T. , Kreinovich, V., Wu, B. and Xiang, G., *Computing Statistics Under Interval and Fuzzy Uncertainty*, Springer-Verlag, 2012.