6

Fuzzy Statistics

The uncertainty characterizing decisions and indeed the process of decision-making on the basis of statistical reasoning can be traced, in many cases, to the lack of imprecise information coming from vagueness in the data considered. In this sense, fuzzy set theory comes to the forefront and, as it is plausibly expected, it plays a prominent role. Before inserting fuzziness, we will present a brief review of *random variables* (also known as *stochastic variables*) and their properties as well as the classical statistical notions of *point estimation*, *interval estimation*, *hypothesis testing*, and *regression*. Then, their fuzzy analogues will be introduced.

6.1 Random Variables

A mapping from the set Ω of possible outcomes (sample points) of an experiment to a subset of real numbers $\mathbb{R}$ is a random variable. A rigorous definition of a random variable is the following:

Definition 6.1.1 A (real) measurable function X from a sample space Ω of a probability model to the set of real numbers, $X : \Omega \to \mathbb{R}$, is called a (real) *random variable*.

Some remarks are in order at this point:

(i) a random variable is not a variable in the usual sense, but a function with the domain Ω and the range $\mathbb{R}$;
(ii) the random variable X may be undefined or infinite for a subset of Ω with zero probability;
(iii) the mapping $X(\omega)$ (i.e. the sample value for a sample point ω) must be such that

$$\{\omega \in \Omega \mid X(\omega) = x\}$$

A Modern Introduction to Fuzzy Mathematics, First Edition.
Apostolos Syropoulos and Theophanes Grammenos.

for $X = x$, where ω is an *event* for a fixed sample value x, for all $x \in \mathbb{R}$. In fact, one can similarly define the events

$$\{\omega \mid X(\omega) \leq x\}$$

for $X \leq x$, or

$$\{\omega \mid X(\omega) > x\}$$

for $X > x$, or even

$$\{\omega \mid x_1 < X(\omega) \leq x_2\}$$

for $x_1 < X \leq x_2$ (see, e.g. [133, 233]).

It is possible to assign probabilities corresponding to the aforesaid events, for example,

$$\Pr(X = x) = \Pr\{\omega \mid X(\omega) = x\},$$

and so on. Now, let us define the *distribution function of a random variable*:

Definition 6.1.2 The *cumulative distribution function* or simply distribution function of a random variable X is defined as

$$D_X(x) = \Pr(X \leq x) = \Pr\{\omega \mid X(\omega) \leq x\}.$$

The cumulative distribution function has to satisfy the following properties:

(i) $D_X(x) \in [0, 1]$,
(ii) $D_X(x_1) \leq D_X(x_2)$, for $x_1 \leq x_2$,
(iii) $\lim_{x \to \infty} D_X(x) = 1$,
(iv) $\lim_{x \to -\infty} D_X(x) = 0$,
(v) $\lim_{\varepsilon^+ \to 0} D_X(x + \varepsilon) = D_X(x)$,

with the second property showing that $D_X(x)$ is a nondecreasing function and the fifth property pointing out its continuity on the right.

When the range of X is finite or countably infinite, then the random variable is *discrete* and a *discrete probability distribution* (known as *probability mass function*) $p_X(x)$ can be defined assigning a certain probability to each value in the range of X, that is, $\Pr(X = x_i) = p_X(x)$ for each sample value x_i. The probability mass function has to satisfy the following properties:

(i) $p_X(x_i) \in [0, 1]$, $i \in \mathbb{N}$,
(ii) $p_X(x) = 0$ for $x \neq x_i$,
(iii) $\sum_i p_X(x_i) = 1$.

Then, the cumulative distribution function $D_X(x)$ of a discrete random variable is given by

$$D_X(x) = \Pr(X \leq x) = \sum_{x_i \leq x} p_X(x_i).$$

In the case of an uncountably infinite range, X is a *continuous* random variable and, if it has a first derivative that is piecewise continuous and exists everywhere except possibly at a finite number of points, then a *probability density function* can be defined:

$$f_X(x) = \frac{d}{dx} D_X(x),$$

which can be integrated in order to find the probability. The probability density function has to satisfy the following properties:

(i) $f_X(x) \geq 0$,
(ii) $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$,
(iii) $\Pr(a < X \leq b) = \int_a^b f_X(x)\, dx$.

Then, the cumulative distribution function $D_X(x)$ of a continuous random variable X is given by

$$D_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt.$$

Now, one can determine the *expectation value* (or *mean*) μ_X of a random variable X:

$$\mu_X = \sum_i x_i p_X(x_i), \quad i \in \mathbb{N}, X \text{ discrete},$$
$$\mu_X = \int_{-\infty}^{\infty} x f_X(x) \quad dx, X \text{ continuous}.$$

Based upon the above relations, the *n th moment $E(X^n)$ of a random variable X* can be introduced:

$$E(X^n) = \sum_i x_i^n p_X(x_i), \quad i \in \mathbb{N}, X \text{ discrete},$$
$$E(X^n) = \int_{-\infty}^{\infty} x^n f_X(x)\, dx, \quad X \text{ continuous}.$$

Clearly, the first ($n = 1$) moment of X is its expectation value μ_X.

Finally, the concepts of *variance* and *standard deviation* of a random variable X can be defined:

$$\sigma_X^2 = E[(X - \mu_X)^2] \geq 0,$$

which for a discrete random variable becomes

$$\sigma_X^2 = \sum_i (x_i - \mu_X)^2 p_X(x_i),$$

while for a continuous random variable, one obtains

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\, dx.$$

In fact, by expanding the expression for σ_X^2, one gets the useful formula for the variance of a random variable X:

$$\sigma_X^2 = E[X^2] - (E[X])^2 = E[X^2] - \mu_X^2.$$

Finally, the positive square root of σ_X^2 yields the *standard deviation*, σ_X, of a random variable X.

There are many important distributions for random variables, most notably the binomial distribution, the Poisson distribution, and the normal (Gaussian) distribution (see, e.g. [233]). At this point, one last remark concerning the so-called conditional distributions is deemed necessary. Following the definition of the *conditional probability* of an event A given event B:

Definition 6.1.3

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \quad \Pr(B) > 0,$$

we can find the cumulative distribution function of a random variable X given the event B:

$$D_X(x|B) = \Pr(X \leq x|B) = \frac{\Pr((X \leq x) \cap B)}{\Pr(B)}.$$

Again, one can distinguish between a discrete and a continuous random variable. Thus, for a discrete random variable, the above equation yields for the *conditional probability mass function*

$$p_X(x_i|B) = \Pr(X = x_i|B) = \frac{\Pr((X = x_i) \cap B)}{\Pr(B)},$$

while in the case of a continuous random variable, we have for the *conditional probability density function*

$$f_X(x|B) = \frac{d}{dx} D_X(x|B).$$

6.2 Fuzzy Random Variables

In Section 6.1, we have reviewed classical random variables and their basic properties. In order to handle fuzzy data or observations, one can introduce fuzzy-valued random variables to grasp vagueness (see, e.g. [78] for a short review), in other

words, randomness and vagueness are now allowed to appear simultaneously. In the literature, one can find different approaches to the notion of a fuzzy random variable, most notably the (mathematically equivalent) definitions introduced by Erich Peter Klement et al. [176] and Madan L. Puri, and Dan A. Ralescu [245], Huibert Kwakernaak [186, 187], and Rudolf Kruse and Klaus Dieter Meyer [183]. In what follows, we will adopt the Kruse–Meyer approach in which a fuzzy random variable is studied as a fuzzy perception/observation of a classical real-valued random variable. This approach is actually a combination of the other authors' considerations. However, we stress that we are not going to introduce the notions of expected value, variance, and distribution function for fuzzy random variables. The reader is referred to [183] for a detailed presentation.

First, let us see what is meant by perception/observation in this context. Assume a measurable space (Ω, A) with Ω denoting the set of all possible outcomes of a random experiment, A a σ-algebra of subsets of Ω[1] and $(\mathbb{R}, B)$ the Borel-measurable space (see footnote in page 241). Further, let the probability space (Ω, A, P) with P a set function defining a probability measure on the space (Ω, A). Suppose that the results of the random experiment are described by $u : \Omega \to \mathbb{R}$ which assigns a random value u to each random choice. Then, u is a random variable.

The perception/observation of the aforementioned random variable means that for each $\omega \in \Omega$, we can investigate whether $u(\omega) \in V_i$ for some i, where $V_i \subseteq \mathbb{R}$. Now, if we examine another mapping, say $X : \Omega \to B(\mathbb{R})$ with $X(\Omega) = V_i \iff V(\omega) \in V_i$, then we associate with each $\omega \in \Omega$ not a real number $u(\omega)$ as in the case of ordinary random variables, but a set V called *random set* (see, e.g. [210] for a detailed study of random sets).[2] The random variable u of which X is a perception is called an *original* of X. In general, given a random set X, the corresponding true original u is not known, we only have a possible set of originals. If no further information is available, then each random variable u with $u(\omega) \in X(\omega)$, for all $\omega \in \Omega$, is a possible candidate for being the original.

Definition 6.2.1 Suppose we have a random sample with fuzzy outcomes (*fuzzy random sample*) defined by the mapping $X : \Omega \to F(\mathbb{R})$ or $(X_1, \ldots, X_n) : \Omega \to (F(\mathbb{R}))^n$. Then, this mapping is called a *fuzzy random variable* if and only if

---

1 Assume that X is a set. Then, a σ-algebra F is a nonempty collection of subsets of X such that the following hold:

(i) X is in F;
(ii) if A is in F, then so is the complement of A; and
(iii) if A_n is a sequence of elements of F, then the union of the elements A_n is in F.

2 A random set is a Borel-measurable function from Ω to the set of all nonempty and compact subsets of $\mathbb{R}$.

there exists a system $X_\alpha(\omega)$, for all $\omega \in \Omega$, and for all $\alpha \in [0,1]$ of subsets of $\mathbb{R}$, such that

$$X_\alpha^{\mathrm{l}} : \Omega \to \mathbb{R}, \quad X_\alpha^{\mathrm{u}} : \Omega \to \mathbb{R}, \quad \forall \alpha \in (0,1]$$

are real-valued random variables and $X_\alpha(\omega) = [X_\alpha^{\mathrm{l}}(\omega), X_\alpha^{\mathrm{u}}(\omega)]$, where l, u stand for "lower" and "upper," respectively.

In other words, a fuzzy random variable is a (fuzzy) perception of an unknown random variable $u : \Omega \to \mathbb{R}$, with u being a possible original of X.

Definition 6.2.2 The set of all possible originals is given by

$$\chi = \{u : \Omega \to \mathbb{R}\},$$

where u is a Borel-measurable mapping.

Now, an interesting theorem (see Ref. [183] for a proof) describing the behavior of fuzzy random variables is the following:

Theorem 6.2.1 *(i) If $n \in \mathbb{N}$, $(X_1, \ldots, X_n)$ is a fuzzy random vector, and $(x_1, \ldots, x_n) \in \mathbb{R}^n$, then the linear combination $\sum_{i=1}^n x_i X_i$ is a fuzzy random variable, (ii) if $k \in \mathbb{N}$ is an odd number and X is a fuzzy random variable, then X^k is a fuzzy random variable, (iii) if $k \in \mathbb{N}$ is an even number and $X : \Omega \to F(\mathbb{R})$ is such that $|X|$ is a fuzzy random variable, then X^k is a fuzzy random variable.*

Now, we come to the definition of the moment of a fuzzy random variable. Let the fuzzy random variable $X : \Omega \to F(\mathbb{R})$, and $k \in \mathbb{N}$. Then, the *k -th moment* of X with respect to the original χ is defined as

Definition 6.2.3

$$(EX)^k(q) = \bigvee \{\mu_X(u) \mid u \in \chi \text{ and } E|u^k| < \infty \text{ and } Eu^k = q\},$$

for all $q \in \mathbb{R}$ and where $\mu_X(u)$ is the so-called *acceptability degree*, that is, the degree for a number $q \in \mathbb{R}$ to be the *expected value* of X is the maximal value of $\mu_X(u)$ such that u is an original of X. Further, one can see by use of the extension principle, that EX^k is the image of the fuzzy subset $\mu_X : \chi \to [0,1]$ if one considers the mapping $E^k : \chi \to \mathbb{R}, u \to Eu^k$ [183].

In practice, to compute the expected value for a series of results given by random experimental results that are in the class of fuzzy sets $K_n(\mathbb{R})$, $n \in \mathbb{N}$, one proceeds as follows:

Given a finite probability space $\Omega = \{\omega_i,\ i = 1, \dots, k\}$, the corresponding probabilities p_i, $i = 1, \dots, k$ for the ω_is, respectively, and the fuzzy random variable $X : \Omega \to K_n(\mathbb{R})$, then the expected value is calculated as

$$EX = \sum_{i=1}^{k} p_i X(\omega_i),$$

and the fuzzy number $E(X) \in F(\mathbb{R})$ is the *fuzzy expectation value* of X if $E(X_\alpha) = [EX_\alpha^{\mathrm{l}}, EX_\alpha^{\mathrm{u}}]$, for all $\alpha \in (0, 1]$.

Next, we consider the concept of *variance of a fuzzy random variable* with respect to χ, whereby we will use the notion of the moment of a fuzzy random variable.

Definition 6.2.4

$$\operatorname{var} X = \bigvee \{\mu_X(u) \mid u \in \chi \text{ and } E|u - Eu|^2 < \infty \text{ and } E(u - Eu)^2 = q\},$$

for all $q \in \mathbb{R}$.

Finally, we come to the definition of the *distribution function of a fuzzy random variable*. First, we need the notion of a *normal set representation* of a fuzzy set.

Definition 6.2.5 We say that a fuzzy set $S \in F(\mathbb{R})$ belongs to the class $Q(\mathbb{R})$ of all fuzzy sets with a *normal set representation* if and only if there exists a set representation $\{A_\alpha \mid \alpha \in (0, 1)\}$ of S such that the following hold:

(i) $\bigwedge A_\alpha > -\infty \Rightarrow \bigwedge A_\alpha \in A_\alpha$;
(ii) $\bigvee A_\alpha < \infty \Rightarrow \bigvee A_\alpha \in A_\alpha$;
(iii) $\bigvee A_\alpha = -\infty$ or $\bigvee A_\alpha = \infty \Rightarrow A_\alpha$ is convex for all $\alpha \in (0, 1)$.

With the help of this definition, we come to the notion of the distribution function:

Definition 6.2.6 The (one-dimensional) distribution function $(F_X(x))(p)$ of a fuzzy random variable $X : \Omega \to Q(\mathbb{R})$ is a mapping $F_X : \mathbb{R} \to F(\mathbb{R})$ such that

(i) $(F_X(x))(p) = \bigvee \{\mu_X(V) \mid V \in \overline{\chi} \text{ and } (P \otimes P')[V \le x] = p\}$, for all $p \in [0, 1]$;
(ii) $(F_X(x))(p) = 0$, for all $p \in \mathbb{R} \setminus [0, 1]$,

where μ_X is the fuzzy set of all possible originals of the fuzzy random vector X, $\overline{\chi}$ is the class of all $A \otimes A' \to$ Borel-measurable random vectors with $A \otimes A'$ a product σ-algebra, and $P \otimes P'$ is a product probability measure. The latter is associated with the product space $(\Omega \otimes \Omega', A \otimes A', P \otimes P')$ that is the probability space from which (Ω, A, P) is the perception. For the notion of a multidimensional distribution function, the reader is referred to [183].

Finally, two fuzzy random variables X and Y are *identically distributed* if X_α^1, Y_α^1 and X_α^2, Y_α^2 are identically distributed for all $\alpha \in (0, 1]$, while X and Y are *independent* if each variable from the set

$$\{X_\alpha^1, X_\alpha^2 \mid \alpha \in (0, 1]\}$$

is independent from each variable from the set

$$\{Y_\alpha^1, Y_\alpha^2 \mid \alpha \in (0, 1]\}$$

(see Ref. [183]). Furthermore, $(X_1, \ldots, X_n)$ is a *normal* (or *Gaussian*) *fuzzy random sample* of size n if all the X_i, $i = 1, \ldots, n$, are *independent and identically distributed* (iid) normal fuzzy random variables, whereby X is a normal fuzzy random variable when $X = E(X) \oplus R$ with $\oplus$ denoting the extended operation of addition, and R is a normal (not fuzzy) random variable with zero mean and variance σ^2, so that $R \sim N(0, \sigma^2)$ [122].

6.3 Point Estimation

Classical statistical analysis is based on random variables, point estimations, statistical hypotheses, and so on. The first major question encountered in statistical inference concerns the *point estimation for one of more unknown parameters*. Just to give an idea, a point estimator estimates a parameter by giving a specific numerical value. Thus, for example, the best point estimate of the population mean μ is the calculated sample mean $\overline{x}$. So the general question becomes how can we choose an estimator on a sample of a fixed size taken on a random variable with a probability density function containing one or more unknown parameters, in order to have a best estimate of that parameter?

Let us formulate the classical problem as it is encountered in statistical inference theory and, in fact, best considered as a problem of decision theory: Let a random variable X with a probability density function $f_X(x; \xi_1, \xi_2, \ldots, \xi_k)$, where ξ_i, $i = 1, \ldots, k$ are unknown parameters. Then, given the value $(x_1, \ldots, x_n)$ of a random sample $(X_1, \ldots, X_n)$ of size n from the population $f(x; \theta_1, \ldots, \theta_k)$, we ask, based on this value, for the estimation ("best guess") of the parameters θ_i. If the estimation of the parameters is given as a single value, then we speak of a *point estimation*, otherwise, we refer to an *interval estimation*. In this sense, an interval estimation of a parameter gives an estimation of the parameter as an interval or a range of values.

In the present section, we shall examine the point estimation. Let $\hat{\theta}_i$ be a point estimation of the parameter θ_i, $i = 1, \ldots, n$. In fact, this estimation is but a *decision* d_i, so that $\hat{\theta}_i = d_i$. This decision depends on the parameter θ_i, and it is a function of the value $(x_1, \ldots, x_n)$ of the random sample $(X_1, \ldots, X_n)$, so $\hat{\theta}_i = d_i(x_1, \ldots, x_n)$.

Further, $\hat{\theta}_i = d_i(x_1, \dots, x_n)$ is a value of the function $\hat{\Theta}_i = d_i(X_1, \dots, X_n)$. The latter is called the *estimator function* (or *decision function*) or simply *estimator* of the parameter θ_i and, indeed, the process of finding the point estimation of the parameters $\theta_1, \dots, \theta_k$ amounts to finding the estimators $\hat{\theta}_1, \dots, \hat{\theta}_k$ of these parameters. In fact, the finding of the estimator $\hat{\theta}_i$ depends on the properties of this function and, usually, the determination of these properties leads to the way of finding the estimator. In this sense, we can seek for various kinds of point estimators, such as the *sufficient*, the *unbiased*, the *consistent*, the *efficient*, or the *maximum likelihood estimator*. For more details see Ref. [161] or [191] where a more advanced approach to point estimators is provided.

In what follows, we choose to briefly present three, very commonly used, classical point estimators, the unbiased estimator, the consistent estimator, and the maximum likelihood estimator.

6.3.1 The Unbiased Estimator

Suppose we have a random sample $(X_1, \dots, X_n)$ from a population $f(x; \theta)$ and the point estimator $\hat{\theta} = d(x_1, \dots, x_n)$. Then, the following theorem holds

Theorem 6.3.1

$$E[(\hat{\theta} - \theta)^2] = \text{var}(\hat{\theta}) + [\theta - E(\hat{\theta})]^2$$

Proof:

$$\begin{aligned} E[(\hat{\theta} - \theta)^2] &= E\{\{[\hat{\theta} - E(\hat{\theta})] - [\theta - E(\hat{\theta})]\}^2\} \\ &= E\{[\hat{\theta} - E(\hat{\theta})]^2 + [\theta - E(\hat{\theta})]^2 - 2[\hat{\theta} - E(\hat{\theta})][\theta - E(\hat{\theta})]\} \\ &= E\{[\hat{\theta} - E(\hat{\theta})]^2 + [\theta - E(\hat{\theta})]^2 - 2[E(\hat{\theta}) - E(\hat{\theta})][\theta - E(\hat{\theta})]\} \\ &= \text{var}(\hat{\theta}) + [\theta - E(\hat{\theta})]^2 \end{aligned}$$

The term $\theta - E(\hat{\theta})$ is called the *bias* of the estimator $\hat{\theta}$. □

Consequently, if it is possible to find an estimator $\hat{\theta}$ of the parameter θ that has very small bias $\theta - E(\hat{\theta})$ and $\text{var}(\hat{\theta})$, then the mean squared error $E[(\hat{\theta} - \theta)^2]$ will be correspondingly small. Naturally, one desires a zero bias, $\theta - E(\hat{\theta}) = 0$, or $E(\hat{\theta}) = \theta$. So we come to the following definition:

Definition 6.3.1 Suppose we have a random sample $(X_1, \dots, X_n)$ from a population $f(x; \theta)$. Then, the point estimator $\hat{\theta} = d(x_1, \dots, x_n)$ of the parameter θ is called *unbiased* when its expected value equals θ, that is, when $E(\hat{\theta}) = \theta$. Otherwise, $\hat{\theta}$ is a *biased* estimator.

6.3.2 The Consistent Estimator

Again, suppose that we have a random sample $(X_1, \dots, X_n)$ from a population $f(x;\theta)$ and let an estimator of θ be $\hat{\theta}_n = d_n(x_1, \dots, x_n)$, whereby the index n denotes the quantity of the sample. Intuitively speaking, a good estimator is one for which the so-called *risk function* $R(d_n;\theta)$ will decrease with increasing n. Let us briefly introduce the notion of the aforementioned risk function.

We know from decision theory that the estimators $\hat{\theta}_1, \dots, \hat{\theta}_k$ (i.e. the decision rules) of the parameters $\theta_1, \dots, \theta_k$ for the values $(x_1, \dots, x_n)$ of the random sample $(X_1, \dots, X_n)$ establish a mapping of the sample space to the decision space. The wrong choice of the estimators produces a loss or cost, expressing the difference between estimated and true values. This loss is quantified by the so-called *loss function* $L(\hat{\theta}_1, \dots, \hat{\theta}_k; \theta_1, \dots, \theta_k)$ and its expected value is the aforementioned risk function:

$$R(d_1, \dots, d_k; \theta_1, \dots, \theta_k) = E[L(\hat{\theta}_1, \dots, \hat{\theta}_k; \theta_1, \dots, \theta_k)]$$

Hence, good estimators are those which minimize the risk function.

So, suppose we have a sequence of estimators $\{\hat{\theta}_n\} = \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \dots$ with $\hat{\theta}_1 = d_1(x_1), \hat{\theta}_2 = d_2(x_1, x_2), \hat{\theta}_3 = d_3(x_1, x_2, x_3), \dots$ for the parameter θ, generated by $\hat{\theta}_n = d_n(x_1, x_2, \dots, x_n)$ for $n = 1, 2, 3, \dots$. Then, we demand

$$\lim R(d_n;\theta) \xrightarrow[n\to\infty]{} 0$$

and if the risk function is the mean squared error $E[(\hat{\theta}_n - \theta)^2]$, then we have the following theorem:

Theorem 6.3.2

$$\lim E[(\hat{\theta}_n - \theta)^2] \xrightarrow[n\to\infty]{} 0.$$

Proof: From Theorem 6.3.1 we have $E[(\hat{\theta}_n - \theta)^2] = \operatorname{var}(\hat{\theta}_n) + [\theta - E(\hat{\theta}_n)]^2$. But $E(\hat{\theta}_n) = \theta$ and $\lim \operatorname{var}(\hat{\theta}_n) \xrightarrow[n\to\infty]{} 0$. Therefore, $\lim E[(\hat{\theta}_n - \theta)^2] \xrightarrow[n\to\infty]{} 0$. □

Now, having said all that, we finally come to the following definition:

Definition 6.3.2 Suppose that we have a random sample $(X_1, \dots, X_n)$ from a population $f(x;\theta)$. The estimator $\hat{\theta}_n = d_n(x_1, \dots, x_n)$ of the parameter θ is called *consistent* if it is unbiased and if it holds that $\lim \operatorname{var}(\hat{\theta}_n) \xrightarrow[n\to\infty]{} 0$ and $E(\hat{\theta}_n) = \theta$.

As an example, one can readily show that for a random sample $(X_1, \dots, X_n)$ from a population with the normal distribution $N(\mu, \sigma^2)$, $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is a consistent estimator of the population variance σ^2.

6.3.3 The Maximum Likelihood Estimator

First, we shall present the definition of the *likelihood function*:

Definition 6.3.3 A likelihood function of n random variables $(X_1, \dots, X_n)$ is the joint distribution

$$L(\theta_1, \dots, \theta_k) = g(X_1, \dots, X_n; \theta_1, \dots, \theta_k), \quad k \geq 1$$

Apparently, when $X_1, \dots, X_n$ is a random sample from the population $f(x; \theta_1, \dots, \theta_k)$, then the likelihood function of this sample is

$$L(\theta_1, \dots, \theta_k) = g(X_1, \dots, X_n; \theta_1, \dots, \theta_k) = \prod_{i=1}^{n} f(x_i; \theta_1, \dots, \theta_k).$$

Now, we can define the maximum likelihood estimator as follows:

Definition 6.3.4 Let $X_1, \dots, X_n$ be a random sample from the population $f(x; \theta_1, \dots, \theta_k)$. The estimators $\hat{\Theta}_i = d_i(X_1, \dots, X_n)$ of the parameters θ_i, $i = 1, 2, \dots, k$ are the *maximum likelihood estimators* if their values $\hat{\theta}_i = d_i(x_1, \dots, x_n)$ maximize the likelihood function $[L(\theta_1, \dots, \theta_k)$ of the sample:

$$L(\hat{\theta}_1, \dots, \hat{\theta}_k) = \max L(\theta_1, \dots, \theta_k).$$

When the likelihood function contains only one parameter θ and is differentiable w.r.t. θ, then the maximum likelihood estimator $\hat{\theta}$ of θ is the solution of the equation

$$\frac{dL(\theta)}{d\theta} = 0.$$

In fact, often the equation

$$\frac{d \ln L(\theta)}{d\theta} = 0$$

is used in applications, since $L(\theta)$ and $\ln L(\theta)$ are maximized for the same values of θ.

In the case where the likelihood function L depends on more than one parameters $\hat{\theta}_1, \dots, \hat{\theta}_k$, then the maximum likelihood estimators $\hat{\theta}_1, \dots, \hat{\theta}_k$ are found as the solution of the system of equations

$$\frac{\partial L(\theta_1, \dots, \theta_k)}{\partial \theta_i} = 0, \quad i = 1, 2, \dots, k$$

or

$$\frac{\partial \ln L(\theta_1, \dots, \theta_k)}{\partial \theta_i} = 0, \quad i = 1, 2, \dots, k.$$

6.4 Fuzzy Point Estimation

We start with the problem of fuzzy point estimation that generalizes what has been presented in Section 6.3. First of all, we should point out that often fuzzy point estimation is considered as more basic than fuzzy interval estimation, since the latter can be determined by the point estimations of the lower and upper boundaries of the interval. In fact, if the point estimation is characterized by very low confidence, it can be relaxed to an interval estimation (for an application see, e.g. [7]).

Now, let us adopt the more formal approach to fuzzy point estimation and suppose we have a fuzzy random sample, that is, a fuzzy random vector, $(X_1, \dots, X_n) : \Omega \to (F(\mathbb{R}))^n$. A fuzzy parameter for our fuzzy point estimation is a perception of the unknown parameter we seek.

Definition 6.4.1 The *fuzzy parameter* $\theta_C(X_1, \dots, X_n)$ of a fuzzy random sample $(X_1, \dots, X_n) : \Omega \to (F(\mathbb{R}))^n$ with respect to a mapping assigning to each probability distribution function its parameter, is a fuzzy set

$$\theta_C[X_1, \dots, X_n](m) = \bigvee \{\mu_{(X_1,\dots,X_n)}(V_1, \dots, V_n) \mid (V_1, \dots, V_n) \in \chi_C^n \text{ and } \theta_C(D_{V_1}) = m\},$$

where $m \in \mathbb{R}$, $n \in \mathbb{N}$, $\theta_C : C \to \mathbb{R}$, C is the class of probability distribution functions, χ_C^n is the set of all possible originals, $\mu_{(X_1,\dots,X_n)}(V_1, \dots, V_n)$ is the acceptability that the random vector $(V_1, \dots, V_n)$ on $\Omega \times \Omega'$ is the original of $(X_1, \dots, X_n)$, and D_V is the probability distribution function of V from the class C.

So we come to the following definition of the fuzzy point estimation:

Definition 6.4.2 Suppose $(X_1, \dots, X_n)$ is a fuzzy random sample from the class of probability distribution functions C and $\theta_C : C \to \mathbb{R}$. Then, a *fuzzy point estimator* of the parameter θ_C is

$$G_n : \Omega \to E(\mathbb{R}), \quad \omega \to G_n[(X_1)_\omega, \dots, (X_n)_\omega],$$

where $E(\mathbb{R})$ is the class of all fuzzy subsets of $\mathbb{R}$.

As we have done with classical point estimation, we can distinguish between different kinds of fuzzy point estimators.

Definition 6.4.3 Suppose $G_n : \Omega \to F(\mathbb{R})$ is a fuzzy random vector and $\mu \in F(\mathbb{R})$. Then, G_n is an *unbiased fuzzy point estimator* if $E(G_n) = \mu$ holds.

Similarly, when we desire to obtain a parameter estimation by using a sequence of fuzzy random variables, then we have the following:

Definition 6.4.4 Suppose C is a class of distribution functions, $\theta_C : C \to \mathbb{R}$ and $G_n : [F(\mathbb{R})]^n \to F(\mathbb{R})$. Let S be the class of sequences of fuzzy random variables (X_k) such that the following holds for the *convex hulls* for every $(X_k) \in S$, $k \in \mathbb{N}$:

$$\operatorname{conv}(\theta_C[X_i]) = \operatorname{conv}(\theta_C[X_1, \dots, X_n]).$$

Then, (G_n) is a *consistent estimator* of θ_C with respect to S if for all sequences $(X_k) \in S$ one has the convergence

$$(G_n[X_1, \dots, X_n])_{n\in\mathbb{N}} \to \operatorname{conv}(\theta_C[X_1]).$$

Concerning the maximum likelihood estimator presented in Section 6.3.3 due to its simplicity and the consequent usefulness in statistical inference, its implementation in the realm of fuzzy data is fairly complex in most practical situations. Indeed, one "classical" approach is to consider a maximum likelihood estimation of a desired parameter as the crisp value that maximizes the probability of observing the fuzzy data (see, e.g. [59]). We shall not go into details, however, we point out that there have been more efficient ways to handle the aforementioned difficulty and the reader is referred, for instance, to the so-called "expectation–maximization algorithm" (see Ref. [95] and references therein).

A final, worth noticing, comment is deemed proper at this point. It concerns the interesting and different approach to the problem of fuzzy point estimation presented in [3] by introducing the methods of the *fuzzy uniformly minimum variance unbiased estimation* and the *fuzzy Bayesian estimation*, both based on the notions of the Yao-Wu signed distance and the L_2-metric. When the fuzzy random variables become crisp random variables, the aforementioned methods are reduced to the classical uniformly minimum variance unbiased estimation and the Bayesian estimation.

6.5 Interval Estimation

In order to determine an interval of plausible values for an unknown sample population parameter, the notion of *interval estimation* is used in classical statistics. In other words, the unknown parameters are estimated (notably, but not only, these parameters are the mean and the variance of the population) as an interval or as an entire range of numerical values within which the aforesaid parameter is estimated to lie. One of the most prevalent forms of interval estimation is the frequentist approach known as *confidence interval* (see, e.g. [215]) that was introduced by the Polish mathematician Jerzy Neyman [227] in 1937.

Let a random sample $(X_1, \dots, X_n)$ of size n with the value $(x_1, \dots, x_n)$. This value consists in measured data characterized, in general, by uncertainties expressed as errors that have been neglected in the previously presented process of finding a

point estimator $\hat{\theta}$. So, in order to increase the reliability of the estimate, one ought to take into account these errors and give the point estimator $\hat{\theta}$ in some interval, say, $(\hat{\theta}-\varepsilon, \hat{\theta}+\varepsilon)$, $\varepsilon > 0$. Let us see how this can be realized in our case. Suppose that our random sample has a continuous distribution $f(x)$ with an unknown mean μ and a known variance σ^2. Now, it is a well-known fact from classical statistics that if X is a random variable with the distribution $f(x)$, then for the random variable defined as

$$Y = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we can always determine the distribution $g(y)$, where $Y = y(X)$. For large samples, $g(y)$ is the standard normal distribution (i.e. the Gaussian distribution with $\mu = 0$ and $\sigma^2 = 1$), as it is inferred by the Central Limit Theorem (see Ref. [161]).

Now, for two arbitrary values y_1 and y_2 of Y, we have

$$\Pr(y_1 < Y < y_2) = \int_{y_1}^{y_2} g(y)\, dy = 1 - \delta$$

or

$$\Pr\left(\overline{X} - \frac{\sigma}{\sqrt{n}} y_2 < \mu < \overline{X} - \frac{\sigma}{\sqrt{n}} y_1\right) = 1 - \delta.$$

Here, δ denotes the *confidence level* (also known as *significance level*).[3] The interval defined by the inequality

$$\overline{X} - \frac{\sigma}{\sqrt{n}} y_2 < \mu < \overline{X} - \frac{\sigma}{\sqrt{n}} y_1$$

is called *confidence interval* of the parameter μ with the probability $100(1-\delta)\%$. The length of this interval is the *confidence length* or *width*:

$$\overline{x} - \frac{\sigma}{\sqrt{n}} y_1 - \left(\overline{x} - \frac{\sigma}{\sqrt{n}} y_2\right) = (y_2 - y_1)\frac{\sigma}{\sqrt{n}}.$$

The confidence length depends mainly on the size of the sample but also on its variability as well as on the confidence level. One always desires the confidence length to be the least possible. Obviously, this can be achieved by minimizing the difference $y_2 - y_1$. In fact, for a constant $1 - \delta$, it can be rather easily shown that the difference $y_2 - y_1$ gets its minimum value when $g(y_2) = g(y_1)$.

---

3 In the literature, it is often the case that the confidence level is designated by $1 - \delta$ when δ is the significance level. Furthermore, usually the desired confidence level is chosen prior to examining the measurements. Very often, the 95% confidence level is applied. However, attention should be paid to the following common misunderstanding: a 95% confidence level does ***not*** mean that 95% of the sample data lie within the confidence interval.

6.6 Interval Estimation for Fuzzy Data

The study of the notion of the fuzzy interval or confidence estimation has started mainly in the 1980s by Roger A. McCain [212] and begun to escalate after 2000 with a plethora of results (see, e.g. [169] for more information on the historical development and on various approaches to the problem). Here, we have chosen to present a rather practical method presented by Norberto Corral and Maria Ángeles Gil [79] for the construction of an interval estimation of an unknown parameter θ for a given sample fuzzy information. One has to stress that, although the method yields rather crude confidence intervals (the probability of the parameter being within the estimated interval may be larger than the confidence level), it has the great advantage of being always applicable, that is, for any membership function and any class of distribution functions.

Suppose that we have a random experiment X with probability space (Ω, A, P), where Ω denotes the set of all possible outcomes of a random experiment, A being a σ-algebra of subsets of Ω, and P defining a probability measure. Let θ be the parameter whose value lies in an interval of the experiment. Further, suppose that the set of all fuzzy observations defines our fuzzy information. Let us first recall some basic definitions:

Definition 6.6.1 A *fuzzy information* for a random experiment is a fuzzy event $\tilde{x}$ on Ω characterized by a Borel-measurable function $\mu_{\tilde{x}}$ that represents the grade of membership of x to $\tilde{x}$.

Definition 6.6.2 A fuzzy partition (orthogonal system) with fuzzy events on X, is called *fuzzy information system* S if $\sum_{\tilde{x}\in S} \mu_{\tilde{x}}(x) = 1$, $\forall x \in X$.

It is assumed that the increased sampling from X will not yield a precise observation but a *sample fuzzy information*:

Definition 6.6.3 A sample fuzzy information of size n from an experiment X is the algebraic product of n elements belonging to a sample information system S of the experiment. Indeed, the fuzzy information system on a random sample of size n of X is called a *fuzzy random sample* of size n from S.

Now, based on Definition 6.6.3, we can proceed to a formal definition of the confidence interval:

Definition 6.6.4 For a given fuzzy random sample of size n from S, denoted by $S^{(n)}$, and an interval $[\theta_{\mathrm{l}}(S^n), \theta_{\mathrm{u}}(S^n)]$, where the subscripts l and u stand for "lower" and "upper," respectively, we have

$$P_\theta\{\theta_{\mathrm{l}}(S^n) \le \theta \le \theta_{\mathrm{u}}(S^n)\} = \sum_{\substack{(\tilde{x}_1,\dots,\tilde{x}_n) \in S^{(n)} \\ \theta_{\mathrm{l}}(\tilde{x}_1,\dots,\tilde{x}_n) \le \theta \le \theta_{\mathrm{u}}(\tilde{x}_1,\dots,\tilde{x}_n)}} \int_X \mu_{(\tilde{x}_1,\dots,\tilde{x}_n)}(x_1,\dots,x_n) dP_\theta(x_1,\dots,x_n) \ge \delta,$$

as a δ-*confidence interval* for the parameter θ with $\delta \in [0,1]$ being the *confidence level*, and $\theta_{\mathrm{l}}(S^n)$, $\theta_{\mathrm{u}}(S^n)$ the *lower* and *upper confidence limits* of θ, respectively. A constant interval $[\theta_{\mathrm{l}}(\tilde{x}_1,\dots,\tilde{x}_n), \theta_{\mathrm{u}}(\tilde{x}_1,\dots,\tilde{x}_n)]$ is often also called a δ-confidence interval for the parameter θ.

Suppose that we have a sample fuzzy information $(\tilde{x}_1,\dots,\tilde{x}_n)$ with

$$s_{(\tilde{x}_1,\dots,\tilde{x}_n)} = \{(x_1,\dots,x_n) \in X^n \mid \mu_{(\tilde{x}_1,\dots,\tilde{x}_n)}(x_1,\dots,x_n) > 0\}$$

its support and n the size of the sample. Then, the following theorem provides a way to determine a δ-confidence interval for the parameter θ:

Theorem 6.6.1 *If* $\theta_{\mathrm{l}}(\tilde{x}_1,\dots,\tilde{x}_n) = \bigwedge\{\Theta_{\mathrm{l}}(s_{(\tilde{x}_1,\dots,\tilde{x}_n)})\}$ *and* $\theta_u(\tilde{x}_1,\dots,\tilde{x}_n) = \bigvee\{\Theta_u(s_{(\tilde{x}_1,\dots,\tilde{x}_n)})\}$, *with the* δ-*confidence interval* $[\Theta_{\mathrm{l}}(X^{(n)}), \Theta_u(X^{(n)})]$ *for* θ *such that*

$$P_\theta\{\Theta_l(X^{(n)}) \le \theta \le \Theta_u(X^{(n)})\} = \sum_{\substack{(x_1,\dots,x_n) \in X^n \\ \Theta_{\mathrm{l}}(x_1,\dots,x_n) \le \theta \le \Theta_{\mathrm{u}}(x_1,\dots,x_n)}} \int_X dP_\theta(x_1,\dots,x_n) \ge \delta,$$

then the interval $[\theta_{\mathrm{l}}(\tilde{x}_1,\dots,\tilde{x}_n), \theta_{\mathrm{u}}(\tilde{x}_1,\dots,\tilde{x}_n)]$ *determines a* δ-*confidence interval for the parameter* θ.

(For a proof of this theorem, see Ref. [79]).

6.7 Hypothesis Testing

Very often, samples of measurement data can be interpreted by a priori assuming a structure (i.e. a specific distribution) of the measurement results and then apply certain statistical tests to determine the probability of the initial assumption (hypothesis) being true or not. In other words, a statistical hypothesis is the assumption about a population parameter, and it is an important part of empirical evidence-based research (see, e.g. [192] for a general overview).

The procedure starts with the examination of a random sample of the population considered. First, it is assumed either that the sample data are the result of pure chance (*null hypothesis*, denoted as H_0) or they are sufficiently affected by some nonrandom influence (*alternative hypothesis*, denoted as H_1). Then, an appropriate statistical test is chosen in order to assess the truth of the null hypothesis. In the next step, one determines the probability that the given data would occur when H_0 is assumed. This probability is called the *p-value* (or *p-level*), and it is used to interpret the result obtained. The smaller the *p-value*, the stronger is the evidence against H_0. In other words, the p-value is a measure of how likely the data would be observed if H_0 were true. Finally, one compares the calculated p-value with the selected *significance* or *confidence level* δ (see Section 6.3). If $p \leq \delta$, then the observed influence is statistically significant, H_0 must be rejected, and H_1 holds. If $p > \delta$, then H_0 is true.

Now, concerning the magnitude of the p-value, one of two types of error may appear. When the p-value is small, there is a possibility that H_0 is true, but an unlikely event has been measured (*type I error* or *false positive*) and we have incorrectly rejected H_0, while if the p-value is large, there is a possibility that H_0 is false, but an unlikely event has been measured (*type II error* or *false negative*), and we have incorrectly accepted H_0. The probability of making a type I error, that is, the probability of rejecting H_0, given that it is true, is δ (i.e. the significance level), while the probability of making a type II error, that is, the probability of accepting H_0, given that H_1 is true, is denoted by β. A usual way out of this apparent impasse is provided by the demand for independent verification of the data.

In practice, in order to find when the null hypothesis can be rejected or not, the concept of the test function can be used. To this purpose, suppose that we have the null hypothesis $H(\theta_0) : \theta = \theta_0$, and let $A(\theta_0)$ be the acceptance set of $H(\theta_0)$ on a δ significance level. Then, with the set $G(x) = \{\theta \mid x \in A(\theta)\}$ we have

$$\theta \in G(x) \leftrightarrow x \in A(\theta),$$

so that

$$P_\theta\{\theta \in G(x)\} \geq 1 - \delta.$$

In other words, any set of acceptance on a δ significance level yields a *confidence set* $G(x)$ on the confidence level $1 - \delta$. The confidence set lets us conclude, for each θ_0, whether the null hypothesis $H(\theta_0)$ should be accepted or rejected on the δ significance level for the measured x. Indeed, we can define a *test function* $\Phi(x; \theta_0)$:

Definition 6.7.1

$$\Phi(x; \theta_0) = \begin{cases} 0, & \theta_0 \in G(x), \\ 1, & \theta_0 \notin G(x), \end{cases}$$

where $\Phi(x;\theta_0)=0$ signifies the acceptance, while $\Phi(x;\theta_0)=1$ signifies the rejection of the null hypothesis. Equivalently, by introducing the so-called *indicator function* $I_{G(x)}(\theta)$ of the set $G(x)$, one has

$$\Phi(x;\theta_0)=\begin{cases}0, & I_{G(x)}(\theta_0)=1,\\ 1, & I_{G(x)}(\theta_0)=0.\end{cases}$$

The following example borrowed from [62] illustrates very clearly the use of the test function.

Example 6.7.1 Let $X_1,\dots,X_n$ be iid from the normal distribution $N(\theta,1)$ with the unknown mean θ. Then, one can deduce a confidence interval for θ at the level $1-\delta$ of the form $G(X)=\left[\overline{X}-\frac{1}{\sqrt{n}}z_{1-\frac{\delta}{2}},\overline{X}+\frac{1}{\sqrt{n}}z_{1-\frac{\delta}{2}}\right]$, with z_δ the δ-quantile of the standard normal distribution. Suppose that in a random sample of size $n=25$ one observes $\overline{x}=0.75$ and the null hypothesis $H_0:\theta=0.5$ is to be tested against the alternative hypothesis $H_1:\theta\neq 0.5$ at the significance level $\delta=0.05$. Here $G(x)=[0.358,1.142]$, consequently the test function reads

$$\Phi(x;0.5)=1\begin{cases}0, & I_{G(x)}(0.5)=1,\\ 1, & I_{G(x)}(0.5)=0.\end{cases}$$

Hence, on the basis of the observed value $\overline{x}=0.75$, the null hypothesis H_0 is accepted at the significance level $\delta=0.05$.

6.8 Fuzzy Hypothesis Testing

Starting in the 1980s with the work of María Rosa Casals et al. [59], a rather large amount of work has been published in the field of fuzzy statistics (see Ref. [282] for a review and references therein). More generally, statistical hypothesis testing in a fuzzy environment was taken up by Przemysław Grzegorzewski and Olgierd Hryniewicz [151]. The problem of fuzzy hypothesis testing has been linked to the notion of the *p*-value described in Section 6.7 by Glen Meeden and Siamak Noorbaloochi [214], who instead of determining a null hypothesis and an alternative hypothesis have given a reformulation of the problem "as the problem of estimating the membership function of the set of good or useful or interesting parameter points." Here, we shall present a different approach introduced by Jalal Chachi et al. [62], who alternatively, in the form of six steps, have given a constructive method to connect fuzzy hypothesis testing with confidence intervals.

Now, let us assume that both the hypothesis parameter θ and the confidence interval are fuzzy [61], that is, $\tilde{\theta}=\tilde{\theta}_0$ and $I=I(\tilde{\theta}_0)$ are, respectively, the fuzzy

parameter value (according to what we have said in Section 6.4, $\tilde{\theta}$ is considered as a fuzzy perception of θ) and the degree of hypothesis acceptance that depends on $\tilde{\theta}_0$, then we have for the set of the parameter values for which the tested hypothesis is accepted

$$\{\tilde{\theta}_0 \in \tilde{I}(X) \mid I(\tilde{\theta}_0) > 0\},$$

while

$$\{\tilde{\theta}_0 \in \tilde{I}(X) \mid 1 - I(\tilde{\theta}_0) > 0\},$$

for the set of the parameter values for which the tested hypothesis is rejected, with $\tilde{I}(X)$ the fuzzy confidence interval. The hypothesis to be tested is the null hypothesis $H_0 : \tilde{\theta} = \tilde{\theta}_0$ against the alternative $H_1 : \tilde{\theta} \neq \tilde{\theta}_0$ for observational data with unknown fuzzy mean $\tilde{\theta}$ but known variance σ^2, so that $(X_1, \ldots, X_n) \overset{\text{iid}}{\sim} N(\tilde{\theta}, \sigma^2)$. To that purpose, we must find the degrees of acceptability for H_0 and H_1 and the "Chachi–Taheri–Viertl algorithm" is codified as follows (see Ref. [62] and the nice numerical examples therein):

(i) Convert the hypothesis to be tested to a set of crisp problems on (for $\alpha \in [0,1]$) the fuzzy parameter. Then, for each α-level for the samples $X_\alpha^{\text{l}} = (X_{1\alpha}^{\text{l}}, \ldots, X_{n\alpha}^{\text{l}})$ and $X_\alpha^{\text{u}} = (X_{1\alpha}^{\text{u}}, \ldots, X_{n\alpha}^{\text{u}})$, one must solve at the confidence level δ the classical hypothesis testing problems

$$H_0 : \theta_\alpha^{\text{l}} = \theta_{0\alpha}^{\text{l}} \text{vs. } H_1 : \theta_\alpha^{\text{l}} \neq \theta_{0\alpha}^{\text{l}}, \tag{6.1}$$

$$H_0 : \theta_\alpha^{\text{u}} = \theta_{0\alpha}^{\text{u}} \text{vs. } H_1 : \theta_\alpha^{\text{u}} \neq \theta_{0\alpha}^{\text{u}}, \tag{6.2}$$

where the fuzzy parameters are $\tilde{\theta}_\alpha = [\theta_\alpha^{\text{l}}, \theta_\alpha^{\text{u}}]$, $\tilde{\theta}_{0\alpha} = [\theta_{0\alpha}^{\text{l}}, \theta_{0\alpha}^{\text{u}}]$.

(ii) Determine the $1-\delta$ confidence intervals $[L_1(X_\alpha^{\text{l}}), L_2(X_\alpha^{\text{l}})]$ and $[U_1(X_\alpha^{\text{u}}), U_2(X_\alpha^{\text{u}})]$ for the crisp parameters θ_α^{l} and θ_α^{u}, respectively, for each $\alpha \in (0,1]$.

(iii) Test the hypotheses (6.1) and (6.2) through the examination of the $1-\delta$ confidence intervals $[L_1(X_\alpha^{\text{l}}), L_2(X_\alpha^{\text{l}})]$ and $[U_1(X_\alpha^{\text{u}}), U_2(X_\alpha^{\text{u}})]$ to see whether they contain $\theta_{0\alpha}^{\text{l}}$ and $\theta_{0\alpha}^{\text{u}}$, respectively. Here, the corresponding test functions are

$$\Phi(X_\alpha^{\text{l}}; \theta_{0\alpha}^{\text{l}}) = \begin{cases} 0, & \theta_{0\alpha}^{\text{l}} \in [L_1(X_\alpha^{\text{l}}), L_2(X_\alpha^{\text{l}})], \\ 1, & \theta_{0\alpha}^{\text{l}} \notin [L_1(X_\alpha^{\text{l}}), L_2(X_\alpha^{\text{l}})], \end{cases}$$

and

$$\Phi(X_\alpha^{\text{u}}; \theta_{0\alpha}^{\text{u}}) = \begin{cases} 0, & \theta_{0\alpha}^{\text{u}} \in [U_1(X_\alpha^{\text{u}}), U_2(X_\alpha^{\text{u}})], \\ 1, & \theta_{0\alpha}^{\text{u}} \notin [U_1(X_\alpha^{\text{u}}), U_2(X_\alpha^{\text{u}})]. \end{cases}$$

(iv) Gather the results in the third case to get a fuzzy confidence interval in order to proceed to the construction of a fuzzy test on the basis of the membership degree of each fuzzy parameter $\tilde{\theta}$ in the fuzzy confidence interval. To this

purpose, it is necessary to group the α-values for which the null hypotheses (6.1) and (6.2) are accepted or rejected. This grouping can be performed by making a graph of α vs. the $1-\delta$ confidence intervals so that every confidence interval has α as its height, and then determine confidence bounds from their intersection (obviously, $\alpha = 1$ is the maximal height). Then, the membership function of the fuzzy parameter $\tilde{\theta}$ is compared with the confidence bound obtained. From this comparison, the α-values for which the null hypotheses are accepted or rejected are found.

(v) Construct the fuzzy set $\tilde{C} = \{(\tilde{\theta}, C(\tilde{\theta})) : \tilde{\theta} \in F(\Theta)\}$ by applying the method given in [61]. This set is a fuzzy confidence interval for the fuzzy parameter $\tilde{\theta}$.

(vi) Construct the fuzzy test function

$$\tilde{\Phi}(X; \tilde{\theta}_0)(r) = \begin{cases} C(\tilde{\theta}_0), & r = 0, \\ 1 - C(\tilde{\theta}_0), & r = 1, \end{cases}$$

for the fuzzy random sample $X = (X_1, \dots, X_n)$. Evidently, the above fuzzy test function $\tilde{\Phi}(X; \tilde{\theta}_0) : (F(\mathbb{R}))^n \to F\{0, 1\}$ is described by a fuzzy set leading to the acceptance (with degree of acceptance $C(\tilde{\theta}_0)$) of the tested null hypothesis, or to its rejection with degree $1 - C(\tilde{\theta}_0)$.

At this point, we must stress that if $C(\tilde{\theta}_0)$ is between zero and one so that it is not absolutely clear what to do, then necessarily one has to make a subjective "fuzzy" decision on the acceptance or rejection of the null hypothesis. In such a case, as it is expected, the null hypothesis is accepted, the more the values of $C(\tilde{\theta}_0)$ tend to one, and rejected, the more they tend to zero. Naturally, the most difficult decision to be made is when the value of $C(\tilde{\theta}_0)$ is $\frac{1}{2}$.

6.9 Statistical Regression

Regression analysis is a statistical methodology for the estimation of the conditionally expected value of a dependent random variable y (the *response variable*) given one or more independent nonrandom variables x (the *predictor variables*) and one or more involved unknown parameters β to be estimated from the data, in other words, we have $E(y|x) = f(x, \beta)$. According to the linearity of the parameters β in the function $f(x, \beta)$, one can build linear or nonlinear regression models to fit the measured data. If one has n pairs of data points (x_i, y_i), $i = 1, \dots, n$ and $n < m$, where m is the length of the vector of the unknown parameters β, the regression model is underdetermined and β cannot be specified. If $n = m$ and the function $f(x, \beta)$ is assumed linear, then the *regression equation* $y = f(x, \beta)$ can be exactly solved, i.e. to find β one can solve a $n \times n$ quadratic system which has one unique

solution provided the xs are linearly independent. Finally, if $n > m$, which is the main problem in regression analysis, one can estimate values for the βs that best fit the data. The method of least squares belongs to the latter case. Further, the performance of regression analysis depends on the chosen process for collecting and measuring the data and the assumptions made in this process (for a more detailed account of regression see, e.g. [106]).

As an example, let us consider the most simple *linear regression* model. It involves two-dimensional data points (given, say, in Cartesian coordinates) and contains one independent and one dependent variable. So let us assume that the model function is linear in β and of first order in x with the regression equation

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{6.3}$$

where the unknowns are β_0, β_1, ε and ε is a random error term denoting the deviation from the true line. Usually, the distribution of the random errors is modeled as a normal distribution with zero mean. In order to estimate the parameters, we shall apply the *method of least squares* and $\hat{\beta}_0$, $\hat{\beta}_1$ will denote the estimated by the given data values of β_0, β_1. Then, the predicted value of y, denoted by $\hat{y}$, is obtained as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{6.4}$$

that is, a straight-line fit. Now, based on the n pairs of data points, we can write (6.3) as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n \tag{6.5}$$

and we have

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$ generate the least possible value of S. For their calculation, we start with the derivatives

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \textstyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i), \\ \frac{\partial S}{\partial \beta_1} &= -2 \textstyle\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i), \end{aligned}$$

from which it follows that

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0, \\ \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0, \end{aligned}$$

respectively. From these equations, we get

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0,$$
$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0,$$

respectively. Hence,

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i,$$
$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2.$$

From these equations, one readily obtains the estimated parameter $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}.$$

By denoting the mean as $\overline{x}$ and $\overline{y}$, the latter can be written in the more usual form

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

while the other estimated parameter is found to be

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

Therefore, (6.4) becomes now

$$\hat{y} = \overline{y} + \hat{\beta}_1 (x - \overline{x}).$$

Evidently, for $x = \overline{x}$ the last equation yields $\hat{y} = \overline{y}$, so $(\overline{x}, \overline{y})$ is a point of the fitted line.

The estimates of the errors ε_i are given by $\hat{\varepsilon}_i = y_i - \hat{y}_i$, thus, we have

$$\hat{\varepsilon}_i = (y_i - \overline{y}) - \hat{\beta}_1 (x_i - \overline{x})$$

from which it follows that

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} (y_i - \overline{y}) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \overline{x}) = 0.$$

However, one should point out that, in practical applications, the rounding of the measured data always results in a nonzero error.

6.10 Fuzzy Regression

Simply put, statistical regression as presented in Section 6.9 is based on crisp random errors, while fuzzy regression is based on fuzzy errors. The difference between these two kinds of uncertainty has led to the necessity of extending statistical regression for the case of a fuzzy environment. This extension started in 1982 with the pioneering work of Hideo Tanaka et al. [283], who applied the methodology of linear programming to develop a *fuzzy linear regression* model. In fact, Tanaka's approach constitutes one of two main approaches in fuzzy regression, namely the so-called *possibilistic regression analysis* that relies on the notion of possibility, and the approach known as *fuzzy least squares method* that aims at the minimalization of the errors between the observed and the estimated data.

Returning to Tanaka's approach, it must be stressed that it can be applied only to linear functions. However, it is very simple in its computational implementation, while the fuzzy least squares method has an advantage compared to Tanaka's approach on that it keeps the degree of fuzziness between observed and estimated results to a minimum (see, e.g. [103] and references therein for some critiques on Tanaka's model). This possibilistic regression model is based on the idea of fuzziness minimization through a minimization of the total support of the regression fuzzy coefficients [258], subject to the inclusion of all the observed data.

The basic form of the model is the general linear function

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 x_1 + \tilde{A}_2 x_2 + \cdots + \tilde{A}_n x_n,$$

with $\tilde{Y}$ the fuzzy response, $\tilde{A}_i$, $i = 1, \ldots, n$ the fuzzy coefficients (or parameters), and $x_1, \ldots, x_n$ the components of a nonfuzzy input vector. The $\tilde{A}_i$s are assumed to be triangular fuzzy numbers and the fuzzy coefficients are characterized by a membership function $\mu_A(\alpha)$. As an example, suppose that we have a fuzzy relationship between the variables (i.e. fuzzy-dependent variables), while the *observed data are crisp*. Further, the triangular fuzzy numbers are assumed symmetric. Then, the membership function of the jth coefficient can be defined as

$$\mu_{A_j}(a) = \max\left\{1 - \frac{a - a_j}{c_j}, 0\right\},$$

where a is the mean value of the fuzzy number $\tilde{A}_j$, a_j is the mode (center value of A_j), and c_j is the spread (width around the center value). The choice of symmetric triangular fuzzy numbers assures that the structure of the model depends only on the data involved in the determination of the upper and lower bounds, while other data points do not play any role. So we have [258]

$$\tilde{A}_j = \{a_j, c_j\}_L = \{\tilde{A}_j \mid a_j - c_j \le \tilde{A}_j \le a_j + c_j\}_L, \quad j = 0, 1, \ldots, n$$

Thus, we obtain

$$\tilde{Y}_i = \tilde{A}_0 + \sum_{j=1}^{n} \tilde{A}_j x_{ij} = (a_0, c_0)_L + \sum_{j=1}^{n} (a_j, c_j)_L x_{ij}.$$

Now, if the support is just enough to contain all the sample's data points, then the confidence in an out-of-sample projection is limited unless one extends the support. To this purpose, one chooses a value $h < 1$ (called the *h-certain factor*) of $\mu_{A_j}(a)$ with the h-line cutting the two sides of the triangle graph of $\mu_{A_j}(a)$ at points with coordinates $(a_j - (1-h)c_j^L, a_j + (1-h)c_j^R)$ on the a-axis and L, R denoting "left" and "right" to a_j, respectively. The interval $[a_j - (1-h)c_j^L, a_j + (1-h)c_j^R]$ on the a-axis is the feasible data interval. This h-certain factor extends, by controlling the size of this data interval, the support of the membership function. In fact, the increase of h leads to the increase of c_j^L and c_j^R.

When the observed data are fuzzy, the application of the aforementioned h-certain factor is also possible. Assuming that the observed fuzzy result can be described by a symmetric triangular fuzzy number $\tilde{Y}_i = (y_i, e_i)$, where y_i is the mode (center value) and e_i is the spread, then the actual data points belong to the interval $[y_i - (1-h)e_i, y_i + (1-h)e_i]$. So the optimization of the model requires the minimization of the spread (width around the center value),

$$\min\left(c_0 + \sum_{j=1}^{n} c_j |x_{ij}|\right), \quad c_j \geq 0.$$

Then, the observed fuzzy data that are adjusted for the h-certain factor, are contained in the estimated fuzzy result [258]:

$$a_0 + \sum_{j=1}^{n} a_j x_{ij} + (1-h)\left(c_0 + \sum_{j=1}^{n} c_j |x_{ij}|\right) > y_i + (1-h)e_i,$$

$$a_0 + \sum_{j=1}^{n} a_j x_{ij} - (1-h)\left(c_0 + \sum_{j=1}^{n} c_j |x_{ij}|\right) < y_i - (1-h)e_i,$$

$$c_j \geq 0, \quad i = 0, 1, \ldots, m, \quad j = 0, 1, \ldots, n.$$

Thus, the increase of the h-certain factor extends the confidence interval and, consequently, the probability for out-of-sample values to be covered by the model.

Let us now examine the *fuzzy least-squares regression*. Remembering (6.5) from Section 6.9 one has, in a fuzzy environment,

$$\tilde{Y}_i = \beta_0 + \beta_1 \tilde{X}_i + \tilde{\varepsilon}_i, \quad i = 1, \ldots, m$$
$$\iff \tilde{\varepsilon}_i = \tilde{Y}_i - \beta_0 - \beta_1 \tilde{X}_i$$

and one has to optimize

$$\min \sum_{i=1}^{n} (\tilde{Y}_i - \beta_0 - \beta_1 \tilde{X}_i)^2.$$

The most commonly used approach for this is the *method of distance measures* that was introduced by Phil M. Diamond [98]. Namely, by defining a measure of the distance d between two triangular fuzzy numbers $\mathrm{tfn}(r_1, s_1, t_1)$, $\mathrm{tfn}(r_2, s_2, t_2)$:

$$\begin{aligned} &d(\mathrm{tfn}(r_1, s_1, t_1), \mathrm{tfn}(r_2, s_2, t_2))^2 \\ &\quad = (r_1 - r_2)^2 + [(r_1 - s_1) - (r_2 - s_2)]^2 + [(r_1 + t_1) - (r_2 + t_2)]^2, \end{aligned}$$

with the model written as

$$\tilde{Y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{X}_i + \tilde{\varepsilon}_i, \quad i = 1, \ldots, m \tag{6.6}$$

one has to optimize

$$\min_{A,B} \sum_{i=1}^{m} d(\tilde{A} + \tilde{B}x_i, \tilde{Y}_i)^2.$$

Assuming $\tilde{B} > 0$, it follows for the distance from (6.6)

$$\begin{aligned} d(\tilde{A} + \tilde{B}x_i, \tilde{Y}_i)^2 = (a + bx_i - y_i)^2 + (a + bx_i - c_A^L - c_B^L x_i - y_i + c_{Y_i}^L)^2 \\ + (a + bx_i + c_A^R + c_B^R x_i - y_i + c_{Y_i}^R)^2, \end{aligned}$$

while a similar expression can be derived for $\tilde{B} < 0$.

A 6×6 system of equations yields the parameters of $\tilde{A}$, $\tilde{B}$ (see, e.g. [86] and references therein for an implementation of the above algorithm).

Exercises

The following exercises are inspired by examples presented in [41].

6.1 Let the fuzzy random vector $X_1, \ldots, X_n$ be independent and identically distributed (iid) from the normal distribution $N(\theta, 1)$, where θ is the unknown mean. Suppose there is a random sample of size $n = 50$, and by performing an experiment, we observe that $\overline{x} = 0.65$. Test the null hypothesis $H_0 : \theta = 0.5$ against the alternative hypothesis $H_1 : \theta \neq 0.5$ at the significance level $\delta = 0.06$.

6.2 Suppose we have a random variable X with the Poisson probability mass function. The probability for $X = x$ is given by $R(x) = \frac{a^x e^{-a}}{x!}$, $x \in \mathbb{N}_0$, $a \in \mathbb{R}^+$. The fuzzy Poisson probability mass function $\overline{P}(x)$ is obtained when a is replaced by the positive fuzzy number $\overline{a}$. Let $x = 10$ and $\overline{a} = \mathrm{tfn}(5, 7, 9)$. Find the α-cut of the fuzzy probability $\overline{P}(8)[\alpha]$, $\alpha \in [0, 1]$.

6.3 Let the random variable X with the normal probability density function $N(\mu, 500)$ and a random sample $X_1, \ldots, X_n$ from $N(\mu, 500)$ with sample size

$n = 70$ and a mean equal to 50. Find the fuzzy estimator $\overline{\mu}$ as a triangular fuzzy number.

6.4 Let the random variable X with the normal probability density function $N(\mu, \sigma^2)$ and a random sample $X_1, \ldots, X_n$ from $N(\mu, \sigma^2)$ with sample size $n = 10$. Find an unbiased fuzzy estimator $\overline{\sigma}^2$ for the variance.

6.5 An experiment is performed and the following ten crisp data pairs (x, y) are measured:

i	x	y
1	20	27
2	24	44
3	22	38
4	18	30
5	8	21
6	4	26
7	32	38
8	14	30
9	30	40
10	11	19

Conclude whether there is any additional information missing in order to find the fuzzy coefficients $\tilde{A}_i$, $i = 1, \ldots, 10$, of the fuzzy linear regression model for this data set. Then, by assuming that the necessary missing information is known, determine the fuzzy coefficients $\tilde{A}_i$.