

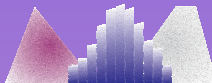
Hypothesis testing about the means of fuzzy random variables

Ana Colubi^{1,2} M. Ángeles Gil¹

¹Department of Statistics and OR

²Institute of Natural Resources and Zoning (Indurot)

Universidad de Oviedo. Spain



*Research Group on Statistical Methods
with Imprecise Random Elements*

ERCIM WG Computing & Statistics 2008

Neuchâtel, June, 19-21 2008

Outline

- **Motivation:** situations in which Fuzzy Random Variables (FRVs^{*}) are useful
 - (*) Models for random experiments where the observation of a characteristic on each outcome is imperfect
 - ✓ randomness → probability space
 - ✓ imperfect values → fuzzy sets
- **Formalization:** suitability of the considered elements to handle these experiments
 - ✓ fuzzy sets: arithmetic and distance
 - ✓ FRVs: (fuzzy valued-) expected value and (real valued-) variance
- **Statistical analysis:** methodology for hypothesis testing about the fuzzy expected value

Some experiments

- **Perceptions/judgments:** practitioners find easier to represent their perceptions about the quality/value of different items by means of fuzzy sets rather than to discard their uncertainty
- **Sociological surveys:** the usual scale from “total disagree” to “total agree” becomes richer
- **Physical measures:** the *real weight* of an item measured by a scale is not precise (measure of the machine \pm maximum error), and may be described by means of an interval
- **Fluctuations:** only the *range* of a given variable over a certain period is known/important
- **Image analysis:** gray or color scales of different objects or surfaces
- ...

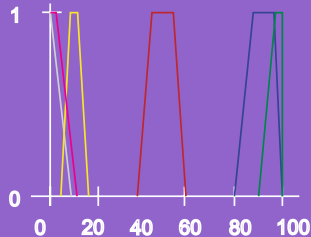
Progress of a reforestation in Asturias (Spain)-Indurot



Main characteristic: quality of the trees



Some of the fuzzy experimental data



- **Support** in percentage: 0 - absence , 100 - perfection
- Experts fix the **0-level** (set of points with positive degree of membership) as the values that they consider compatible with their opinion to a greater or lesser extent
- Experts fix the **1-level** (set of points with total degree of membership) as the values that they consider completely compatible with their opinion

Fuzzy values

$$\mathcal{F}_c(\mathbb{R}^p) = \{U : \mathbb{R}^p \rightarrow [0, 1] \mid U_\alpha \in \mathcal{K}_c(\mathbb{R}^p) \quad \forall \alpha \in [0, 1]\}$$

- $U_\alpha = \{x \in \mathbb{R}^p \mid U(x) \geq \alpha\}$ if $\alpha > 0$,
- $U_0 = \text{cl}\{x \in \mathbb{R}^p \mid U(x) > 0\}$
- $\mathcal{K}_c(\mathbb{R}^p) = \{\text{nonempty compact convex subsets of } \mathbb{R}^p\}$
 - ✓ The concept of **fuzzy set** generalizes the concept of **set**
 - ✓ We restrict ourselves to $\mathcal{K}_c(\mathbb{R}^p)$ because it is an operative and wide enough class
 - ✓ A fuzzy set is a convex $[0, 1]$ -valued **upper semicontinuous function** with compact support
 - ✓ $\mathcal{F}_c(\mathbb{R}^p)$ is a **functional space**
 - ✓ The arithmetic, distance etc. are defined as **extensions** of the analogous operations for sets

Arithmetic on $\mathcal{F}_c(\mathbb{R}^p)$: Zadeh's extension principle

- $(U + V)_\alpha = \{u + v \mid u \in U_\alpha, v \in V_\alpha\}$
 - $(\lambda \cdot U)_\alpha = \{\lambda u \mid u \in U_\alpha\}, \quad \forall \lambda \in \mathbb{R}$
- for all $U, V \in \mathcal{F}_c(\mathbb{R}^p)$, $\lambda \in \mathbb{R}$ and $\alpha \in [0, 1]$

Remarks on the Minkowski addition

- Useful when the interest is focused on the (fuzzy) set-valued data *per se*
 - ✓ A fluctuation of $[1, 2]$ added to a fluctuation of $[3, 4]$ leads to a fluctuation of $[4, 6]$
- Coherent with the product by **natural scalars**
- The structure of $(\mathcal{F}_c(\mathbb{R}^p), +, \cdot)$ is **not linear**, but *conical*
 - ✓ $[1, 2] - [1, 2] = [-1, 1] \neq \tilde{0} = \mathbb{I}_{\{0\}}$

Distances between convex compact sets

- The **Hausdorff distance** is the best known
 - ✓ Not well-suited for some statistical problems
 - ✓ the expected value coherent with the arithmetic mean in the sense of the SLLN is not a Fréchet expected value
- Some useful **distances** for statistical purposes are based on the concept of **support function**, which
 - ✓ characterizes the convex compact sets and allows us to embed $\mathcal{K}_c(\mathbb{R}^p)$ into a Hilbert space
 - ✓ can be extended level-wise in order to characterize fuzzy sets

Support function (Klement *et al.*, 1986)

- The **support function** of $V \in \mathcal{F}_c(\mathbb{R}^p)$ is

$s_V : \mathbb{S}^{p-1} \times [0, 1] \rightarrow \mathbb{R}$ defined so that

$$s_V(u, \alpha) = \sup_{v \in V_\alpha} \langle u, v \rangle \text{ for all } u \in \mathbb{S}^{p-1}, \alpha \in [0, 1]$$

- ✓ $\mathbb{S}^{p-1} =$ unit sphere in \mathbb{R}^p
- ✓ $\langle \cdot, \cdot \rangle =$ inner product on \mathbb{R}^p
- **One-to-one mapping**

$$s : \mathcal{F}_c(\mathbb{R}^p) \rightarrow \mathcal{C}^* \subset \mathcal{L}(\mathbb{S}^{p-1} \times [0, 1])$$

such that $s(U) = s_U$

- s preserves the **semilinear structure**

Family of metrics on $\mathcal{F}_c(\mathbb{R}^p)$ (Näther, 2000)

$$\begin{aligned} & [D_K(U, V)]^2 \\ &= \int_{(\mathbb{S}^{p-1})^2 \times [0,1]^2} (s_U(u, \alpha) - s_V(u, \alpha)) (s_U(v, \beta) - s_V(v, \beta)) dK(u, \alpha, v, \beta) \\ &= \langle s_U - s_V, s_U - s_V \rangle_K \end{aligned}$$

where K is a certain positive definite and symmetric kernel and $\langle \cdot, \cdot \rangle_K =$ inner product w.r.t. K in $\mathcal{L}(\mathbb{S}^{p-1} \times [0, 1])$

- $(\mathcal{F}_c(\mathbb{R}^p), D_K)$ can be embedded through an **isometry** into a cone of the Hilbert space $(\mathcal{L}(\mathbb{S}^{p-1} \times [0, 1]), \|\cdot\|_2^K)$ via the support function
- **Some elections** for K lead to **very intuitive** metrics, like Bertoluzza *et al.* (1994)'s one.

Particular case on $\mathcal{F}_c(\mathbb{R})$ (Bertoluzza *et al.*, 1994)

$$D_W^\varphi(U, V) = \left(\int_{[0,1]} \int_{[0,1]} [f_U(\alpha, \lambda) - f_V(\alpha, \lambda)]^2 dW(\lambda) d\varphi(\alpha) \right)^{\frac{1}{2}}$$

- $f_U(\alpha, \lambda) = \lambda \sup U_\alpha + (1 - \lambda) \inf U_\alpha$
- W and φ are normalized measures on $([0, 1], \mathcal{B}_{[0,1]})$ fulfilling some general properties

- **Alternative expression:**

$$D_W^\varphi(U, V) = \left(\int_{[0,1]} (\text{mid } U_\alpha - \text{mid } V_\alpha)^2 + w_0 (\text{spr } U_\alpha - \text{spr } V_\alpha)^2 \varphi(\alpha) \right)^{\frac{1}{2}}$$

- Mid and spr stand for the center and the semi-amplitude and $w_0 \in [0, 1]$
- Recently we have defined a meaningful generalization of mids and spreads to the d -dimensional case leading to very intuitive metrics in $\mathcal{F}_c(\mathbb{R}^d)$ (Trutschnig *et al.*, 2008)
- φ allows us to **weight** the importance of **each α -level**

FRVs (Puri & Ralescu, 1986)

Given a probability space (Ω, \mathcal{A}, P) , an FRV $\mathcal{X} : \Omega \rightarrow \mathcal{F}_c(\mathbb{R}^p)$ is a (D_K-) **Borel measurable** mapping

- \mathcal{X} is an FRV \Leftrightarrow the α -level mapping $\mathcal{X}_\alpha : \Omega \rightarrow \mathcal{K}_c(\mathbb{R}^p)$ is a convex compact random set $\forall \alpha \in [0, 1]$

Mean value and variance

- If $\sup_{x \in \mathcal{X}_0} \|x\| \in L^1$, the (fuzzy) **expected value** is the unique $\tilde{E}(\mathcal{X}) \in \mathcal{F}_c(\mathbb{R}^p)$ s.t. for all $\alpha \in [0, 1]$

$$\left(\tilde{E}(\mathcal{X}) \right)_\alpha = \text{Aumann's integral of } \mathcal{X}_\alpha$$

$$= \left\{ \int_\Omega X(\omega) dP(\omega) \mid X : \Omega \rightarrow \mathbb{R}^p, X \in L^1(\Omega, \mathcal{A}, P), X \in \mathcal{X}_\alpha \text{ a.s. } [P] \right\}$$

$$* \text{ If } p = 1, \left(\tilde{E}(\mathcal{X}) \right)_\alpha = [E(\inf \mathcal{X}_\alpha), E(\sup \mathcal{X}_\alpha)]$$

- If $\sup_{x \in \mathcal{X}_0} \|x\| \in L^2$, the (real-valued) **variance** is

$$\sigma_{\mathcal{X}}^2 = E([D_K(\mathcal{X}, \tilde{E}(\mathcal{X}))])^2$$

Suitability of the expected value and the variance

- The **fuzzy expected value** is **coherent** with the considered arithmetic in the sense of the SLLN even when stronger metrics are considered
- The **variance** is defined as usual in **metric spaces** when the interest is focused in quantifying the dispersion about the expected value (or, in other words, the error of predicting the values of a variable \mathcal{X} by the expected value $\tilde{E}(\mathcal{X})$)
- The tandem fuzzy expected value/variance satisfies the **Fréchet approach** when the metric D_K is considered
 - ✓ $\tilde{E}(\mathcal{X}) = \arg \min_U E(D_K^2(\mathcal{X}, U))$
 - ✓ $\sigma_{\mathcal{X}}^2 = \min_U E(D_K^2(\mathcal{X}, U))$

which is appropriate for least squares problems

Notations

- One-population problems

- ✓ **Population:** $\mathcal{X} : \Omega \rightarrow \mathcal{F}_c(\mathbb{R}^p)$ such that

$$\sup_{x \in \mathcal{X}_0} \|x\| \in L^2$$

- ✓ **Sample information:** $\mathcal{X}_1, \dots, \mathcal{X}_n$ independent and distributed as \mathcal{X}

- ✓ **Sample mean:** $\bar{\mathcal{X}} = \frac{1}{n} (\mathcal{X}_1 + \dots + \mathcal{X}_n)$

- k -populations problems

- ✓ **Populations:** $\mathcal{X}_i : \Omega_i \rightarrow \mathcal{F}_c(\mathbb{R}^p)$ for $i = 1, \dots, k$ such that

$$\sup_{x \in (\mathcal{X}_i)_0} \|x\| \in L^2$$

- ✓ **Sample information:** $\{\mathcal{X}_{ij}\}_{j=1}^{n_i}$ independent and distributed as \mathcal{X}_i for all $i = 1, \dots, k$

- ✓ **Sample means:** $\bar{\mathcal{X}}_i = \frac{1}{n_i} (\mathcal{X}_{i1} + \dots + \mathcal{X}_{in_i})$ for all

$$i = 1, \dots, k$$

Two-sided tests on expected values

- One-sample:

$$\begin{aligned} H_0 : \tilde{E}(\mathcal{X}) = V & \sim H_0 : D_K(\tilde{E}(\mathcal{X}), V) = 0 \\ H_1 : \tilde{E}(\mathcal{X}) \neq V & \sim H_1 : D_K(\tilde{E}(\mathcal{X}), V) > 0 \end{aligned}$$

- Two-samples:

$$\begin{aligned} H_0 : \tilde{E}(\mathcal{X}_1) = \tilde{E}(\mathcal{X}_2) & \sim H_0 : D_K(\tilde{E}(\mathcal{X}_1), \tilde{E}(\mathcal{X}_2)) = 0 \\ H_1 : \tilde{E}(\mathcal{X}_1) \neq \tilde{E}(\mathcal{X}_2) & \sim H_1 : D_K(\tilde{E}(\mathcal{X}_1), \tilde{E}(\mathcal{X}_2)) > 0 \end{aligned}$$

- ANOVA (k -samples):

$$\begin{aligned} H_0 : \tilde{E}(\mathcal{X}_1) = \dots = \tilde{E}(\mathcal{X}_k) \\ H_1 : \exists i, j \text{ with } \tilde{E}(\mathcal{X}_i) \neq \tilde{E}(\mathcal{X}_j) \end{aligned}$$

General procedures in $(\mathcal{F}_c(\mathbb{R}), D_W^\varphi)$

Analogue developments to the **usual** ones in hypothesis testing **for the mean of an RV** (Montenegro *et al.*, 2001, 2004, Gil *et al.*, 2006)

- **Exact** procedures under **normality** assumption
 - ✓ Normal FRVs (in Puri & Ralescu's sense) can model only a **few situations** in practice
- **Asymptotic** approaches for finitely-valued FRVs based on Taylor expansions of distances under regularity conditions
 - ✓ Conclusions from simulations: the **sample sizes** to get suitable results are **too large**
- **Bootstrap** approaches based on the asymptotic ones
 - ✓ Conclusions from simulations: **better** behaviour than the asymptotic ones **for small/moderate sample sizes**

General procedures in $(\mathcal{F}_c(\mathbb{R}^p), D_K)$

By taking advantages of known results in **Banach spaces** and the **support function**

- One-sample **asymptotic approach** (Körner, 2000)
 - ✓ it **depends on the population covariance** operator
 - ✓ it is **complex** to be applied
 - ✓ the **sample sizes** to get suitable results are **too large**
- One-sample **bootstrap** approach based on the asymptotic one (González-Rodríguez *et al.*, 2006)
 - ✓ it does **not** require further population **knowledge**
 - ✓ it is easy to be applied in practice
 - ✓ **better** behaviour for **small/moderate sample sizes**
 - ✓ it can be **generalized** to obtain **k -sample procedures**

One-sample bootstrap test algorithm (for $D = D_W^\varphi$ or $D = D_K$)

- **Step 1.** Compute the value of the statistic

$$T = \left[D(\bar{\mathcal{X}}, V) \right]^2 / \hat{S}^2$$

where $\hat{S}^2 = \sum_{i=1}^n [D(\tilde{x}_i, \bar{\mathcal{X}})]^2 / (n - 1)$

- **Step 2.** Fix the bootstrap population to be $\{\mathcal{X}_1, \dots, \mathcal{X}_n\}$
- **Step 3.** Obtain a sample of i.i.d. FRVs $(\mathcal{X}_1^*, \dots, \mathcal{X}_n^*)$ from the bootstrap population

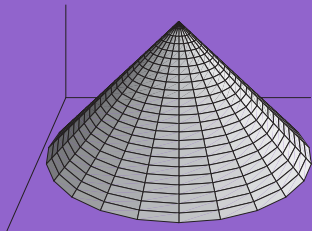
- **Step 4.** Compute the value of the bootstrap statistic

$$T^* = \left[D(\bar{\mathcal{X}}^*, \bar{\mathcal{X}}) \right]^2 / \hat{S}^{*2}$$

- **Step 5.** Steps 3 and 4 should be repeated a large number B of times to get a set of B estimators, denoted by $\{T_1^*, \dots, T_B^*\}$
- **Step 6.** Compute the bootstrap p -value as the proportion of values in $\{T_1^*, \dots, T_B^*\}$ being greater than T

Empirical behaviour. Some simulations

- Let $\mathcal{X} : \Omega \rightarrow \mathcal{F}_c(\mathbb{R}^2)$ be a FRV whose values are cones with basis having a two-dimensional random center $(\xi_{\mathcal{X}}, \eta_{\mathcal{X}})$, and a real-valued random radius, $\varrho_{\mathcal{X}}$



- Consider D_K so that

$$K(u, \alpha, v, \beta) = \begin{cases} 1 & \text{if } u \in \mathbb{S}^1, \alpha \in [0, 1], v = u, \beta = \alpha \\ 0 & \text{otherwise} \end{cases}$$

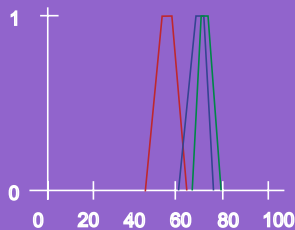
Empirical behaviour. Some simulations

- Percentage of rejections at the significance level $\alpha = .05$, for different sample sizes n .
 - ✓ Test A_1 shows the empirical results in case $\xi_{\mathcal{X}}$ and $\eta_{\mathcal{X}}$ are $\mathcal{N}(0, 1)$ and $\varrho_{\mathcal{X}}$ is $\mathcal{U}(0, 4)$.
 - ✓ Test A_2 shows the empirical results in case $\xi_{\mathcal{X}}$ is $\mathcal{N}(0, 1)$, $\eta_{\mathcal{X}}$ is $\mathcal{U}(0, 1)$ and $\varrho_{\mathcal{X}}$ is an $\exp(1)$

n	30	50	100	200
A_1	4.631	4.878	5.017	5.024
A_2	4.721	4.699	5.133	4.948

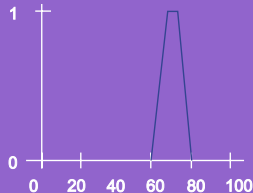
Case-study in forestry: sample data

- One of the targets was to obtain conclusions about the “mean quality” of the three main species of trees in the reforestation, namely, birch (*Betula celtiberica*), sessile oak (*Quercus petraea*) and rowan (*Sorbus aucuparia*).
- A random sample of reforested trees was picked up, and $n_1 = 133$ birches, $n_2 = 109$ sessile oaks and $n_3 = 37$ rowans were obtained
- Sample fuzzy means of the “quality” for the three species



Case-study in forestry: one-sample test

- **Objective:** to check whether or not the mean quality of the trees of each one of the species is '*moderate/high*', where this value is assumed to be described by the fuzzy set



- **Bootstrap p -values:** .24, 0 and .1 for birches, sessile oaks and rowans respectively
- **Conclusion:** at the significance level $\alpha = .05$, the 'mean quality' for the birches and rowans cannot be rejected to be '*moderate/high*', whereas this is not sustainable for the sessile oaks.

Two-sample bootstrap test algorithm

- **Step 1.** Compute the value of the statistic (without assuming equal variances)

$$T = \frac{\left[D_W^\varphi(\bar{x}_1, \bar{x}_2) \right]^2}{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}$$

- **Step 2.** Compute the bootstrap populations by adding to each one of the values of the initial sample the sample mean of the other variable, i.e.

$$\{\tilde{x}_{11} + \bar{x}_2, \dots, \tilde{x}_{1n_1} + \bar{x}_2\} \text{ and } \{\tilde{x}_{21} + \bar{x}_1, \dots, \tilde{x}_{2n_2} + \bar{x}_1\}$$

- **Step 3.** Obtain a sample of i.i.d. FRVs from each bootstrap population: $(\mathcal{X}_{11}^*, \dots, \mathcal{X}_{1n_1}^*), (\mathcal{X}_{21}^*, \dots, \mathcal{X}_{2n_2}^*)$

Two-sample bootstrap test algorithm

- **Step 4.** Compute the value of the bootstrap statistic

$$T^* = \frac{\left[D_W^\varphi(\bar{X}_1^*, \bar{X}_2^*) \right]^2}{\frac{\widehat{S}_1^{*2}}{n_1} + \frac{\widehat{S}_2^{*2}}{n_2}}$$

- **Step 5.** Steps 3 and 4 should be repeated a large number B of times to get a set of B estimators, denoted by $\{T_1^*, \dots, T_B^*\}$
- **Step 6.** Compute the bootstrap p -value as the proportion of values in $\{T_1^*, \dots, T_B^*\}$ being greater than T

Remark

- **Simulations** show the usual empirical behaviour in the Behrens-Fisher problem (bad results for unequal variances and large differences between the sample sizes)

Case-study in forestry: two-sample test

- **Objective:** to check whether or not the mean quality of the birches and the rowans on one hand, and the birches and sessile oaks on the other hand, were equal
- **Bootstrap p -values:** .26 and 0 respectively
- **Conclusion:** at the usual significance levels, we cannot reject that the 'mean quality' of birches and rowans is equal, however this hypothesis is not sustainable in connection with the birches and the sessile oaks

k -sample bootstrap test algorithm (ANOVA)

- **Step 1.** Compute the value of the statistic

$$T = \frac{\sum_{i=1}^k n_i \left[D_W^\varphi(\bar{\mathcal{X}}_i, \bar{\mathcal{X}}) \right]^2}{\sum_{i=1}^k \frac{1}{n_i} \sum_{j=1}^{n_i} \left[D_W^\varphi(\tilde{x}_{ij}, \bar{\mathcal{X}}_i) \right]^2}$$

- **Step 2.** Compute the bootstrap populations by adding to each sample the mean of the other ones; for this purpose, we will define the FRVs, \mathcal{Y}_i , taking on values $\tilde{y}_{ij} = \tilde{x}_{ij} + (\bar{\mathcal{X}}_1 + \dots + \bar{\mathcal{X}}_{i-1} + \bar{\mathcal{X}}_{i+1} + \dots + \bar{\mathcal{X}}_k)$ with the corresponding respective relative frequencies.
- **Step 3.** Obtain a sample of i.i.d. fuzzy random vectors from each bootstrap population: $(\mathcal{Y}_{i1}^*, \dots, \mathcal{Y}_{in_i}^*)$, for $i = 1 \dots k$

k -sample bootstrap test algorithm (ANOVA)

- **Step 4.** Compute the value of the bootstrap statistic

$$T^* = \frac{\sum_{i=1}^k n_i \left[D_W^\varphi(\overline{\mathcal{Y}}_i^*, \overline{\mathcal{Y}}^*) \right]^2}{\sum_{i=1}^k \frac{1}{n_i} \sum_{j=1}^{n_i} \left[D_W^\varphi(\tilde{y}_{ij}^*, \overline{\mathcal{Y}}_i^*) \right]^2}$$

- **Step 5.** Steps 3 and 4 should be repeated a large number B of times to get a set of B estimators, denoted by $\{T_1^*, \dots, T_B^*\}$
- **Step 6.** Compute the bootstrap p -value as the proportion of values in $\{T_1^*, \dots, T_B^*\}$ being greater than T

Remark

- **Simulations** show a good empirical behaviour

Case-study in forestry: k -sample test

- **Objective:** to check whether or not the mean qualities of the three species of trees were equal
- **Bootstrap p -value:** 0
- **Conclusion:** at the usual significance levels, we can conclude that the 'mean quality' is different for the three species

Concluding remarks

- These studies are part of the statistical analysis about FRVs developed by the *Research Group on Statistical Methods with Imprecise Random Elements*
- The combination of the theoretical results in **Hilbert Spaces** (through the support function) and the **resampling** techniques seems to be quite suitable in this context

Related Research

- Power analysis of the testing procedures
- 'Soften' /weaken hypotheses concerning the parameters
- Applications to clustering procedures based on the p -value, quality control ...
- Analysis of real distributions through fuzzy representation