Special cases of lower previsions and their use in statistics

Part II: Statistics with interval data

Montpellier, July 2014

Table of contents

Where do interval data come from?

- Limited reliability of measuring instruments.
- Significant digits.
- Intermittent measurement.
- Censoring.
- Binned data.
- (Not randomly) missing data.
- Gross ignorance - Theoretical contraints.
- ...

More details in: S. Ferson *et al.*, Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty, SAND2007-0939, 2007.

Premises

- The interval specifies where the value is, and where the value is not.
- This assertion will be understood to have two mathematical components:
  - Ignorance about the distribution over the interval.
  - Full confidence.

Descriptive Statistics from interval data

- The cartesian product $[l_1, u_1] \times [l_2, u_2] \times \ldots \times [l_n, u_n]$ represents our (incomplete) knowledge about the sample $\mathbf{x} = (x_1, \ldots, x_n)$.
- What do we know about its mean, std. deviation, empirical distribution function, etc.?

Mean, median, variance...

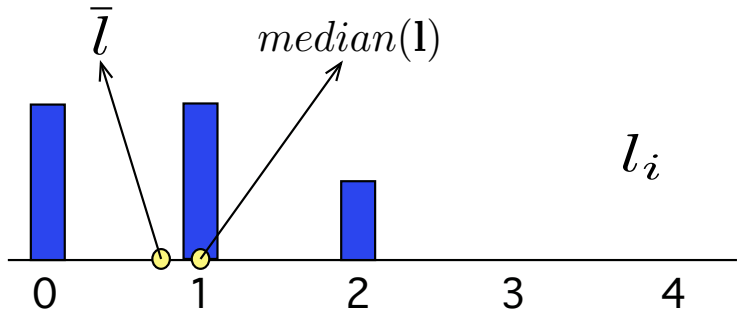
- Nomenclature: $\mathbf{l} = (l_1, \ldots, l_n)$ and $\mathbf{u} = (u_1, \ldots, u_n)$.
- We can easily determine bounds for $\overline{\mathbf{x}}$ and $\mathrm{median}(\mathbf{x})$.
  - Mean: $\overline{\mathbf{l}} \leq \overline{\mathbf{x}} \leq \overline{\mathbf{u}}$.
  - Median: $\mathrm{median}(\mathbf{l}) \leq \mathrm{median}(\mathbf{x}) \leq \mathrm{median}(\mathbf{u})$.
  - Variance: $\min\{s^2_{\mathbf{l}}, s^2_{\mathbf{u}}\} \leq s^2_{\mathbf{x}} \leq \max\{s^2_{\mathbf{l}}, s^2_{\mathbf{u}}\}$?

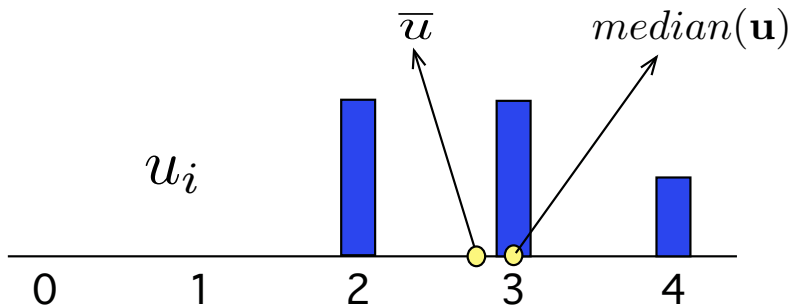
(The mean and the median are comonotonic operators, while the variance is not.)

Example

$[0,2] \times [1,3] \times [1,3] \times [2,4] \times [0,2]$ represents ill-knowledge about $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$. (Sample size $n = 5$).

- Information about the mean: $0.8 \leq \overline{\mathbf{x}} \leq 2.8$.
- Information about the median: $1 \leq \text{median}(\mathbf{x}) \leq 3$.





And what about the variance?

- The upper and lower bounds of the variance cannot be written in terms of the respective variances of **l** and **u** in general.
- We need to solve the problem:

$$\text{Calculate } \max[\overline{y^2} - (\overline{y})^2] \text{ and } \min[\overline{y^2} - (\overline{y})^2]$$

$$\text{Subject to: } l_i \leq y_i \leq u_i, \ \ i = 1, \ldots, n.$$

Information about frequencies of events and about the empirical distribution

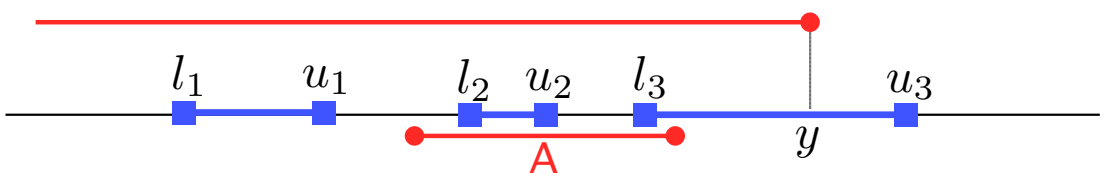
- Proportion of items in A:

$$\underline{f}_A \leq \frac{\#\{i : x_i \in A\}}{n} \leq \overline{f}_A,$$

where $\underline{f}_A = \frac{\#\{i:[l_i,u_i]\subseteq A\}}{n}$ and $\overline{f}_A = \frac{\#\{i:[l_i,u_i]\cap A\neq\emptyset\}}{n}$.

- Empirical distribution function:

$$F_{\mathbf{u}}(y) \leq F_{\mathbf{x}}(y) \leq F_{\mathbf{l}}(y), \ \forall\, y \in \mathbb{R}.$$



Exercise 1: The imprecise histogram

- Consider the following sample of size 10:

$$(2.1,\ 4.3,\ 4.2,\ 1.7,\ 3.8,\ 7.5,\ 6.9,\ 5.2,\ 6.7,\ 4.8)$$

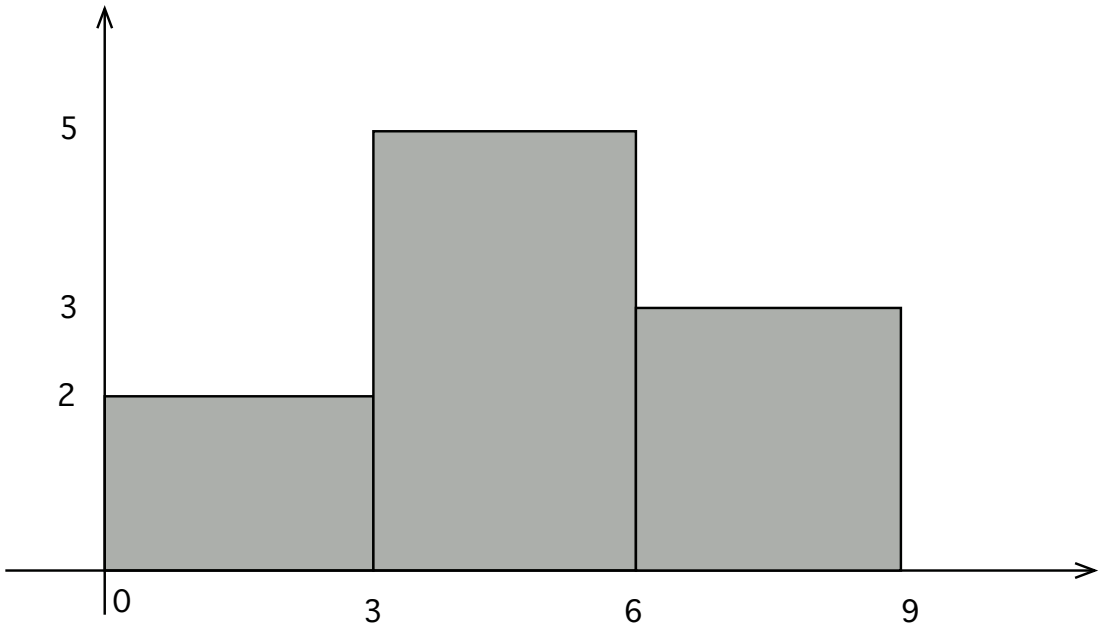
- Consider the grouping intervals $[0, 3)$, $[3, 6)$, $[6, 9]$, and draw the corresponding histogram of frequencies.
- Now suppose that someone else has imprecise information about the above data set given by means of the following cartesian product of intervals:

$$[1, 4] \times [2, 5] \times [3, 5] \times [1, 2] \times [3, 5] \times [4, 8] \times [6, 8] \times [4, 7] \times [6, 8] \times [3, 5]$$

- Consider the initial grouping intervals. For each interval, plot two lines, corresponding to its maximum and the minimum frequency. Compare the new "imprecise histogram" with the first one.

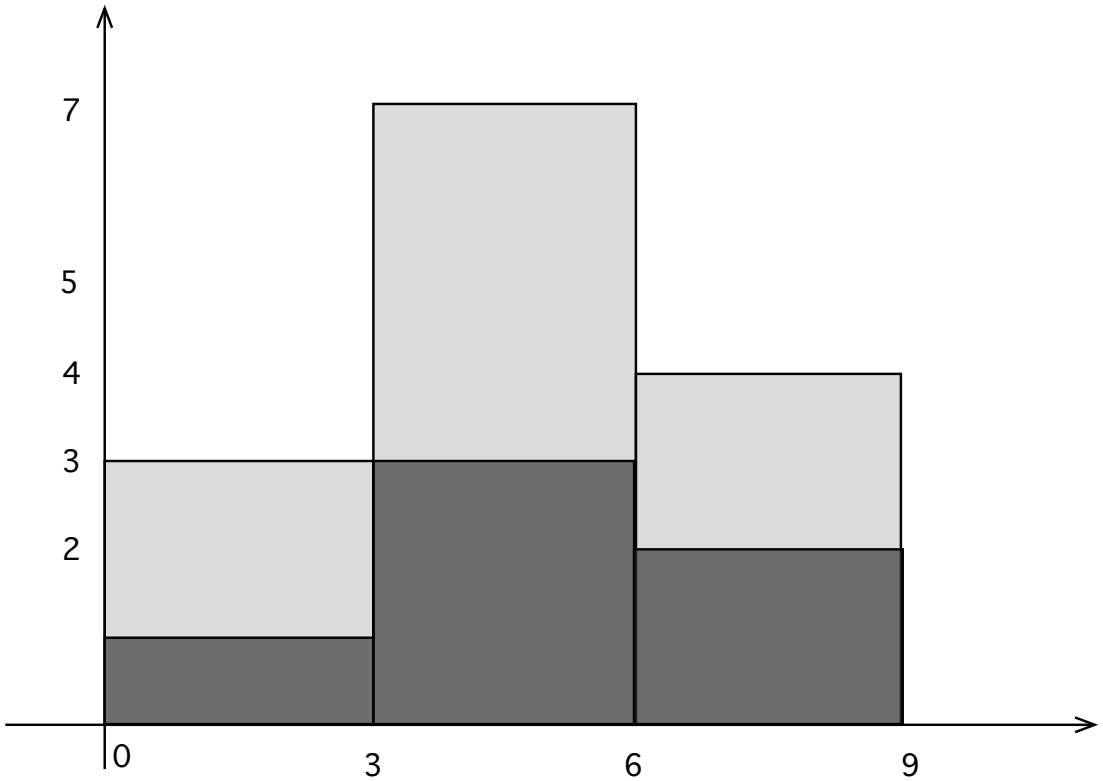
Solution to Exercise 1

The histogram associated to the original (precise) data:



Solution to Exercise 1: cont.

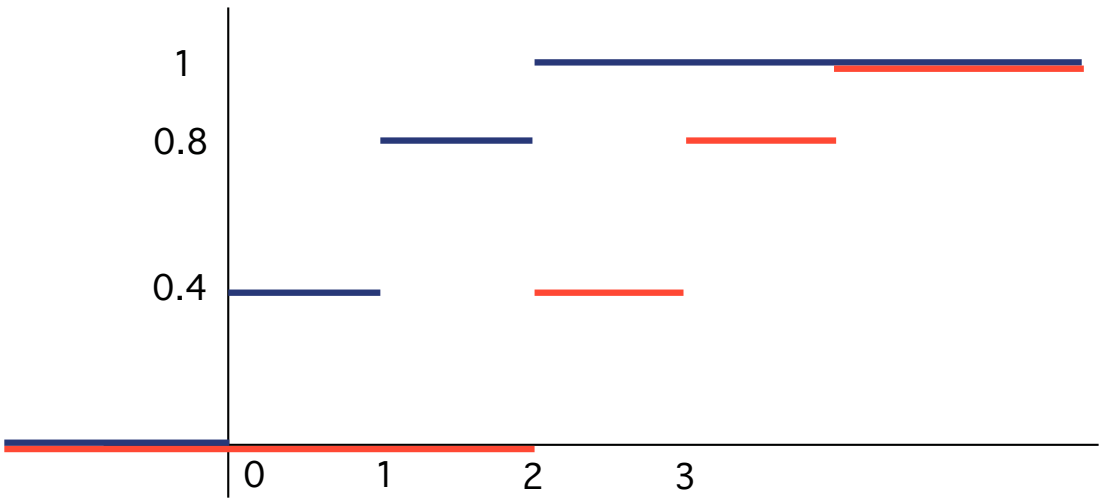
The "imprecise" histogram is the following one. It represents the collection of histograms where the respective heights are between the minimum and the maximum heights, and the sum of the three heights is equal to 10.



Example

$[0,2] \times [1,3] \times [1,3] \times [2,4] \times [0,2]$ represents ill-knowledge about $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$. (Sample size $n = 5$).

Information about empirical distribution function: p-box.



Exercise 2.- Lack of expressiveness of p-boxes

Consider the following imprecise samples of size $n = 2$:

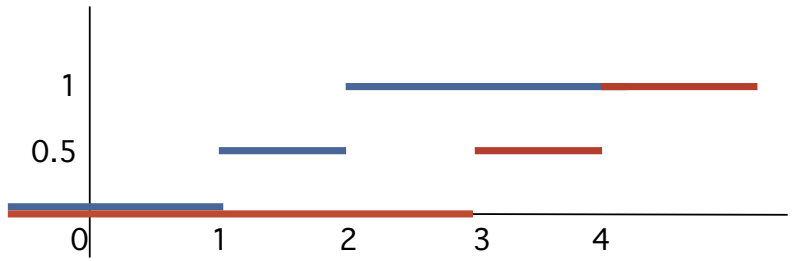
- Sample 1.- $[l_1, u_1] = [1, 4]$ $[l_2, u_2] = [2, 3]$.
- Sample 2.- $[l'_1, u'_1] = [1, 3]$ $[l'_2, u'_2] = [2, 4]$.

(a) Determine their respective empirical p-boxes. Do they coincide?

(b) Determine the upper and lower frequencies associated to the interval $[2, 3]$ in both cases. Do they coincide?

Solution to Exercise 2

(a) Both samples produce the same p-box:



(b) The respective lower and upper frequencies are:

- $\underline{f}_A = \frac{\#\{i:[l_i,u_i]\subseteq A\}}{2} = 0.5$ and $\overline{f}_A = \frac{\#\{i:[l_i,u_i]\cap A\neq\emptyset\}}{2} = 1.$
- $\underline{f}'_A = \frac{\#\{i:[l'_i,u'_i]\subseteq A\}}{2} = 0$ and $\overline{f}'_A = \frac{\#\{i:[l'_i,u'_i]\cap A\neq\emptyset\}}{2} = 1.$

(They do not coincide.)

Some curiosities about Exercise 2

- Both samples produce the same "contour function", $x \rightsquigarrow \pi(x) = P^*(\{x\})$.
- P^* is a possibility measure, $\Pi(A) = \sup \pi_{x \in A}(x)$, because the focals are nested.
- P'^* dominates Π, and therefore the set of frequency distributions compatible with $[l_1, u_1] \times [l_2, u_2]$ is more informative than the other.
- Which one is more informative, P^* or P'^*?

Some curiosities about Exercise 2

- Both samples produce the same "contour function", $x \rightsquigarrow \pi(x) = P^*(\{x\})$.
- P^* is a possibility measure, $\Pi(A) = \sup \pi_{x \in A}(x)$, because the focals are nested.
- P'^* dominates Π, and therefore the set of frequency distributions compatible with $[l_1, u_1] \times [l_2, u_2]$ is more informative than the other.
- Which one is more informative, P^* or P'^*?
- At first sight, the dataset $[l_1, u_1] \times [l_2, u_2]$ seems to be more informative than $[l'_1, u'_1] \times [l'_2, u'_2]$.

Some curiosities about Exercise 2

- Both samples produce the same "contour function", $x \rightsquigarrow \pi(x) = P^*(\{x\})$.
- P^* is a possibility measure, $\Pi(A) = \sup \pi_{x \in A}(x)$, because the focals are nested.
- P'^* dominates Π, and therefore the set of frequency distributions compatible with $[l_1, u_1] \times [l_2, u_2]$ is more informative than the other.
- Which one is more informative, P^* or P'^*?
- At first sight, the dataset $[l_1, u_1] \times [l_2, u_2]$ seems to be more informative than $[l'_1, u'_1] \times [l'_2, u'_2]$.
- But, according to the commonality functions, $[l'_1, u'_1] \times [l'_2, u'_2]$ seems to be more informative than $[l_1, u_1] \times [l_2, u_2]$. In fact, $Q(A) \geq Q'(A)$, $\forall A$, and $Q([1, 4]) > Q'([1, 4])$.

Random set: main notions

$(\Omega, \mathcal{A}, P)$ prob. space; $\Gamma : \Omega \to \wp(\mathbb{R})$ represents incomplete information about $X : \Omega \to \mathbb{R}$. Γ is said to be a *random set* if $A^* = \{\omega \in \Omega : \Gamma(\omega) \cap A \neq \emptyset\}$ is a measurable set for every $A \in \beta_{\mathbb{R}}$.

- Upper probability of A: $P^*(A) = P(\{\omega \in \Omega : \Gamma(\omega) \cap A \neq \emptyset\})$.
- Lower probability of A: $P^*(A) = P(\{\omega \in \Omega : \Gamma(\omega) \subseteq A\})$.
- Aumann expectation: $E(\Gamma) = \{E(Y) : Y \in S(\Gamma)\}$. (Aumann expectation is closely related to Choquet integral.)
- Kruse variance: $Var(\Gamma) = \{Var(Y) : Y \in S(\Gamma)\}$.

- A.P. Dempster, Upper and lower probabilities induced by multi-valued mappings, The Annals of Mathematical Statistics 38, 325-339 (1967).
- J. Aumann. Integral of set valued functions. Journal of Mathematical Analysis and Applications 12, 1-12 (1965).
- R. Kruse. On the Variance of Random Sets, Journal of Mathematical Analysis and Applications 122, 469-473 (1987).

Exercise 3.- Random sets: main notions

- Consider a set of 10 students enrolled in an international course, $\Omega = \{s_1, \ldots, s_{10}\}$.
- Consider the collection of languages: $L = \{$English, French, German, Italian, Spanish, Dutch$\}$.
- Consider the Laplace distribution over the initial set, representing the random selection of a student of the course.
- The multi-valued mapping $\Gamma : \{s_1, \ldots, s_{10}\} \to \wp(\{1, \ldots, 6\})$ reflects my knowledge about the number of those languages that each of the students can speak.

Exercise 3.- Random sets: main notions (Cont.)

$\Gamma(s_1) = \{4, 5, 6\},\ \Gamma(s_2) = \{2, 3\},\ \Gamma(s_3) = \{2\},$
$\Gamma(s_i) = \{2, \ldots, 6\},\ i = 4, \ldots, 10.$

(a) What do we know about the proportion of students that speak 3 or more different languages?

(b) Calculate the bounds of the Aumann expectation of Γ.

(c) Calculate the bounds for the actual variance of the "number of languages spoken" in the population, according to the available information.

Solution to Exercise 3

(a) $P_*([3,\infty)) = P(\{s_i \in \Omega : \Gamma(s_i) \subseteq [3,\infty), \Gamma(s_i) \neq \emptyset\}) =$
$P(\{s_1\}) = 0.1,$
$P^*([3,\infty)) = P(\{s_i \in \Omega : \Gamma(s_i) \cap [3,\infty) \neq \emptyset\}) =$
$P(\Omega \setminus \{s_3\}) = 0.9.$

(b) $\min E(\Gamma) = 0.1 \cdot 4 + 0.9 \cdot 2 = 2.2.$
$\max E(\Gamma) = 0.1 \cdot 2 + 0.1 \cdot 3 + 0.8 \cdot 6 = 5.3.$

(c) $\min Var(\Gamma) = 0.2 = Var(Y_1)$, with
$Y_1(s_1) = 4, Y_1(s_3) = 2, Y_1(s_i) = 3, i \notin \{1,3\}.$
$\max Var(\Gamma) = 4 = Var(Y_2)$, with
$Y_2(s_i) = 2, i = 2,\ldots,6, Y_2(s_i) = 6, i = 1,7,8,9,10.$

Families of probabilities associated to Γ

- P^* and P_* are ∞-order capacities. They univocally determine a pair of upper and lower previsions.
- Credal set: $M(P^*) = \{P : P \leq P^*\} = \{P : P \geq P_*\}$.
- Family of probability measures of selections:
  $\mathcal{P}(\Gamma) = \{P_Y : Y \in S(\Gamma)\}$, where
  $S(\Gamma) = \{Y : \Omega \to \mathbb{R} \;\; Y(\omega) \in \Gamma(\omega) \; \forall \, \omega \in \Omega\}$.
- $M(P^*) \supseteq \mathcal{P}(\Gamma)$.
- The lack of convexity of $\mathcal{P}(\Gamma)$ makes their difference important in some cases.

Exercise 4.- Lack of expressiveness of the credal set

- Γ represents ill-knowledge about a certain constant $c_0 = X(a)$. All we know is that $c_0 \leq k$.
  - $\Omega = \{a\}$, $\Gamma(a) = (-\infty, k]$.
  - $\mathcal{P}(\Gamma) = \{\delta_c : c \leq k\}$.
  - $M(P^*_\Gamma) = \{P : P((-\infty, k]) = 1\}$.
  - $\text{Var}(\Gamma) = \{0\}$
- Γ represents ill-knowledge about X'. All we know is that $X'(\omega) \leq k, \ \forall \omega \in [0, 1]$.
  - $\Omega' = [0, 1]$, $\Gamma'(\omega) = (-\infty, k]$
  - $\mathcal{P}(\Gamma') = \{P : P((-\infty, k]) = 1\}$.
  - $M(P^*_{\Gamma'}) = \{P : P((-\infty, k]) = 1\}$.
  - $\text{Var}(\Gamma') = [0, \infty)$.

Shafer's Evidence Theory

Consider an arbitrary finite universe U. In Evidence Theory, a mapping $m : \wp(U) \to [0,1]$ is said to be a *basic mass assigment* when it satisfies the following restrictions:

- $m(\emptyset) = 0$
- $\sum_{A \subseteq U} m(A) = 1$.

Furthermore, the *belief* and the *plausibility measure* associated to m are the respective set-functions $\mathrm{Bel} : \wp(U) \to [0,1]$ and $\mathrm{Pl} : \wp(U) \to [0,1]$ defined as follows:

- $\mathrm{Bel}(B) = \sum_{A \subseteq B} m(A), \ \forall\, B \in \wp(U)$
- $\mathrm{Pl}(B) = \sum_{A \cap B \neq \emptyset} m(A), \ \forall\, B \in \wp(U)$.

G. Shafer, A mathematical theory of evidence, Princeton University Press, 1976.

Exercise 5.- Random sets and Evidence Theory

Shafer's Evidence Theory and the theory of random sets are closely related from a formal point of view.

Consider a measurable space $(\Omega, \mathcal{A})$, a finite universe U and a random set $\Gamma : \Omega \to \wp(U)$ with non-empty images.

- Check that the lower and upper probabilities associated to Γ do respectively coincide with the belief and plausibility measures associated to some mass assignment.
- Determine such a mass assignment as a function of P_Γ, the probability measure induced by Γ on $\wp(\wp(U))$.

Solution to Exercise 5

Let us consider the set function $m : \wp(U) \to [0,1]$ defined as follows:

$$m(B) = P(\{\omega \in \Omega : \Gamma(\omega) = B\}) = P_\Gamma(\{B\}),\ \forall\, B \in \wp(U).$$

We observe that

- $m(\emptyset) = 0$ and
- $\sum_{B \in \wp(U)} m(B) = 1$,

(It is a basic mass assignment.)
Furthermore, the upper and lower probabilities induced by Γ can be defined as functions of m as follows:

$$P^*(A) = P(\{\omega \in \Omega : \Gamma(\omega) \cap A \neq \emptyset\}) = \sum_{B\,:\,B \cap A \neq \emptyset} m(B),$$

$$P_*(A) = P(\{\omega \in \Omega : \Gamma(\omega) \cap A \neq \emptyset\}) = \sum_{B\,:\,B \subseteq A} m(B).$$

Therefore, P^* and P_* do respectively satisfy the properties of plausibility and belief functions.

How do we do with interval datasets?

- How do we represent the sample information?
- How do we test hypotheses?

How do we represent the sample information?

- We take a random sample of size n. (An instance of a sequence of n i.i.d. random variables).
- Our ill-knowledge about the attribute values is represented by means of n intervals.
- Is it an instance of a sequence of n independent identically distributed random sets?

Exercise 6.- Independent random variables and dependent random sets

- We have a light sensor that displays numbers between 0 and 255.
- We take 10 measurements per second. When the brightness is higher than a threshold (255), the sensor displays the value 255 during 3/10 seconds, regardless the actual brightness value.
- Below we provide data for six measurements:
  - The actual values of brightness represent a realization of a simple random sample of size $n = 6$.
  - But what about the displayed quantities and our interval-valued information? Are them independent?

actual values	215	150	200	300	210	280
displayed quantities	215	150	200	255	255	255
set-valued information	$\{215\}$	$\{150\}$	$\{200\}$	$[255, \infty)$	$[0, \infty)$	$[0, \infty)$.

Solution to Exercise 6

The sample of the "true" values of brightness can be seen as a realization of a 6-dimensional random vector whose components are independent identically distributed random variables. Notwithstanding, our incomplete information about it does not satisfy the condition of random set independence. In fact, we have:

$$P(\Gamma_i \supseteq [255, \infty) | \Gamma_{i-1} \supseteq [255, \infty), \Gamma_{i-2} \not\supseteq [255, \infty)) = 1, \ \forall\, i \geq 3.$$

Exercise 7.- Dependent random variables and independent random sets

- X_0 and Y_0 respectively represent the temperature (in $^{\circ}$C) of an ill person taken at random in a hospital just before taking an antipyretic (X_0) and 3 hours later (Y_0).
- The random set Γ_1 represents the information about X_0 using a very crude measure (it reports always the same interval $[37, 40.5]$).
- The random set Γ_2 represents the information about Y_0 provided by a thermometer with $+/-0.5$ $^{\circ}$C of precision.

(a) Are X_0 and Y_0 stochastically independent?

(b) Are Γ_1 and Γ_2 stochastically independent?

Comments about independence

- In Exercise 6, a sequence of n i.i.d. ill-observed random variables is represented by means of non independent random sets.
- In Exercise 7, two independent random sets represent imprecise information about a pair of dependent attributes.
- Random set independence represents independence between the sources of information about the attributes, and not independence between the attributes themselves.

- I. Couso, S. Moral, P. Walley, Examples of Independence for Imprecise Probabilities First International Symposium on Imprecise Probabilities and Their Applications (ISIPTA'99).
- I. Couso, S. Moral, P. Walley, A survey of concepts of independence for imprecise probabilities, Risk, Decision and Policy, 5, 165- 187 (2000).
- I. Couso, S. Moral, Independence concepts in evidence theory, International Journal of Approximate Reasoning 51 (7), 748-758.

Frequentist Hypothesis testing: notations

- $X : \Omega \to \mathbb{R}$ random variable with distribution function F_θ.
- $\mathbf{X} = (X_1, \ldots, X_n) : \Omega^n \to \mathbb{R}^n$ simple random sample of size n.
- Null hypothesis: $H_0 : \theta \in \Theta_0$,
- Alternative hypothesis: $H_1 : \theta \in \Theta_1$.
- Test: $\varphi : \mathbb{R}^n \to \{0, 1\}$. It associates a decision to each possible sample of size n.
  - $\varphi(\mathbf{y}) = 1$ means "rejection",
  - $\varphi(\mathbf{y}) = 0$ means "no rejection" or "acceptance".

Frequentist Hypothesis testing: notations (cont.)

- Rejection region: $R = \{\mathbf{y} \in \mathbb{R}^n : \varphi(\mathbf{y}) = 1\}$.
- Size of the test: supremum of the possible values for its expectation, assuming that H_0 is true. Mathematically,

$$size(\varphi) = \sup_{\theta \in \Theta_0} E_\theta(\varphi) = \sup_{\theta \in \Theta_0} P_\theta(R).$$

- Let $(\varphi_\alpha)_{\alpha \in (0,1)}$ a sequence of tests, with nested rejection regions $(R_\alpha)_{\alpha \in (0,1)}$ and $\sup_{\theta \in \Theta_0} P_\theta(R_\alpha) = \alpha$.
- p-value of a sample $\mathbf{y}$: $p(\mathbf{y}) = \inf\{\alpha \in (0,1) : \mathbf{y} \in R_\alpha\}$.

Set-valued test functions for interval data

- Suppose that we have imprecise information about the sample realization $\mathbf{x} = (x_1, \ldots, x_n)$, expressed by means of a subset of $\mathbb{R}^n$, $\gamma \subseteq \mathbb{R}^n$.
- Consider a non-randomized test φ with rejection region R.
- Let us calculate the set-valued output of the test as follows:

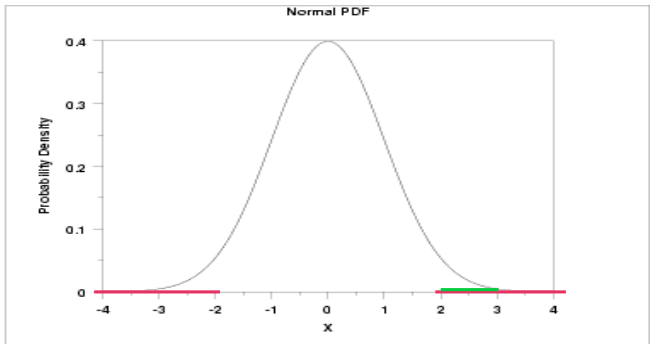
$$\varphi(\gamma) = \{\varphi(\mathbf{y}) : \mathbf{y} \in \gamma\} = \begin{cases} \{1\} & \text{if } \gamma \subseteq R \\ \{0\} & \text{if } \gamma \cap R = \emptyset \\ \{0,1\} & \text{otherwise} \end{cases}$$

- The set-valued output represents our imprecise information about $\varphi(\mathbf{x})$.

- S. Ferson *et al.*, Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty, SAND2007-0939, 2007.
- T. Denœux *et al.*, Nonparametric rank-based statistics and significance tests for fuzzy data. Fuzzy Sets and Systems, 153:1-28, 2005.
- I. Couso, L. Sánchez, Defuzzification of fuzzy p-values, In D. Dubois et al.(Eds), Soft Methods for Handling Variability and Imprecision, Advances in Soft Computing, volume 48, pages 126-132, 2008.

Set-valued test functions for interval data: example

- X is normally distributed with known variance $\sigma^2 = 1$ and unknown expectation μ.
- We consider the test $H_0 : \mu = 0$ against $H_0 : \mu \neq 0$ and take a sample of size $n = 25$.
- Under the null hypothesis, the statistic $\frac{(\overline{X}-\mu_0)}{\sigma/\sqrt{n}} = 5\overline{X}$ follows a standard normal distribution.
- Consider the 0.05-sized test function $\varphi(\mathbf{x}) = \begin{cases} 1 & \text{if } |5\,\overline{x}| > 1.96 \\ 0 & \text{otherwise.} \end{cases}$
- We obtain set-valued information about the attribute, $\gamma \subseteq \mathbb{R}^{25}$.
- Information about the sample mean: it belongs to $[0.4, 0.6]$.
- Decision: reject.

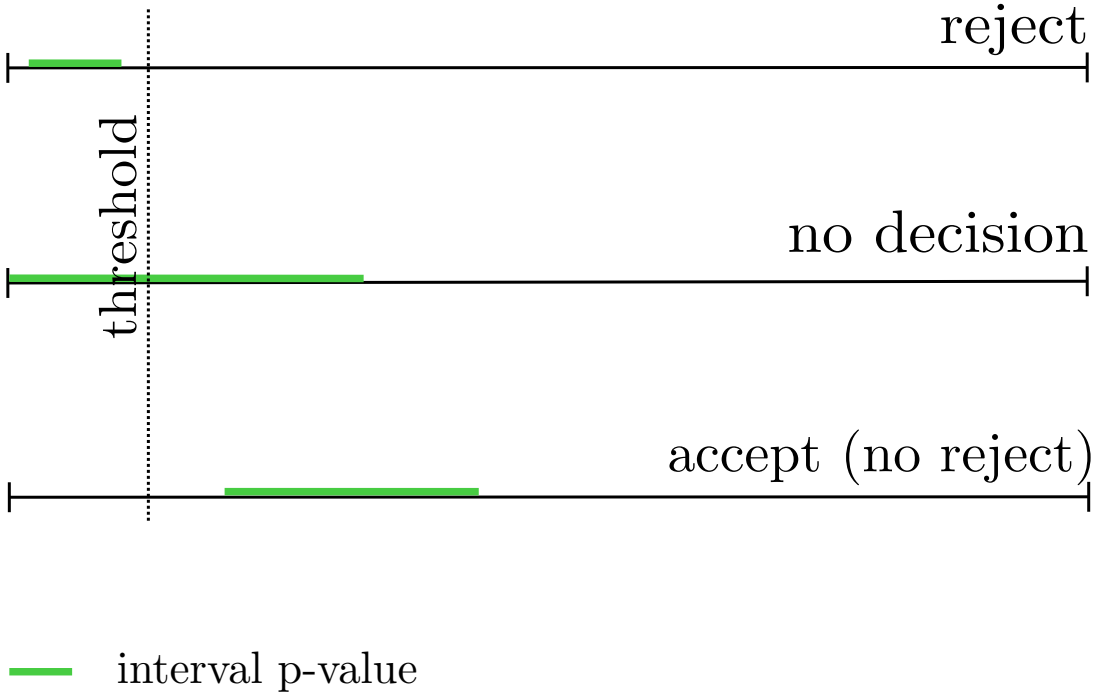


$\{5 \cdot \frac{x_1+\ldots+x_{25}}{25} : (x_1, \ldots, x_{25}) \in \gamma\}$

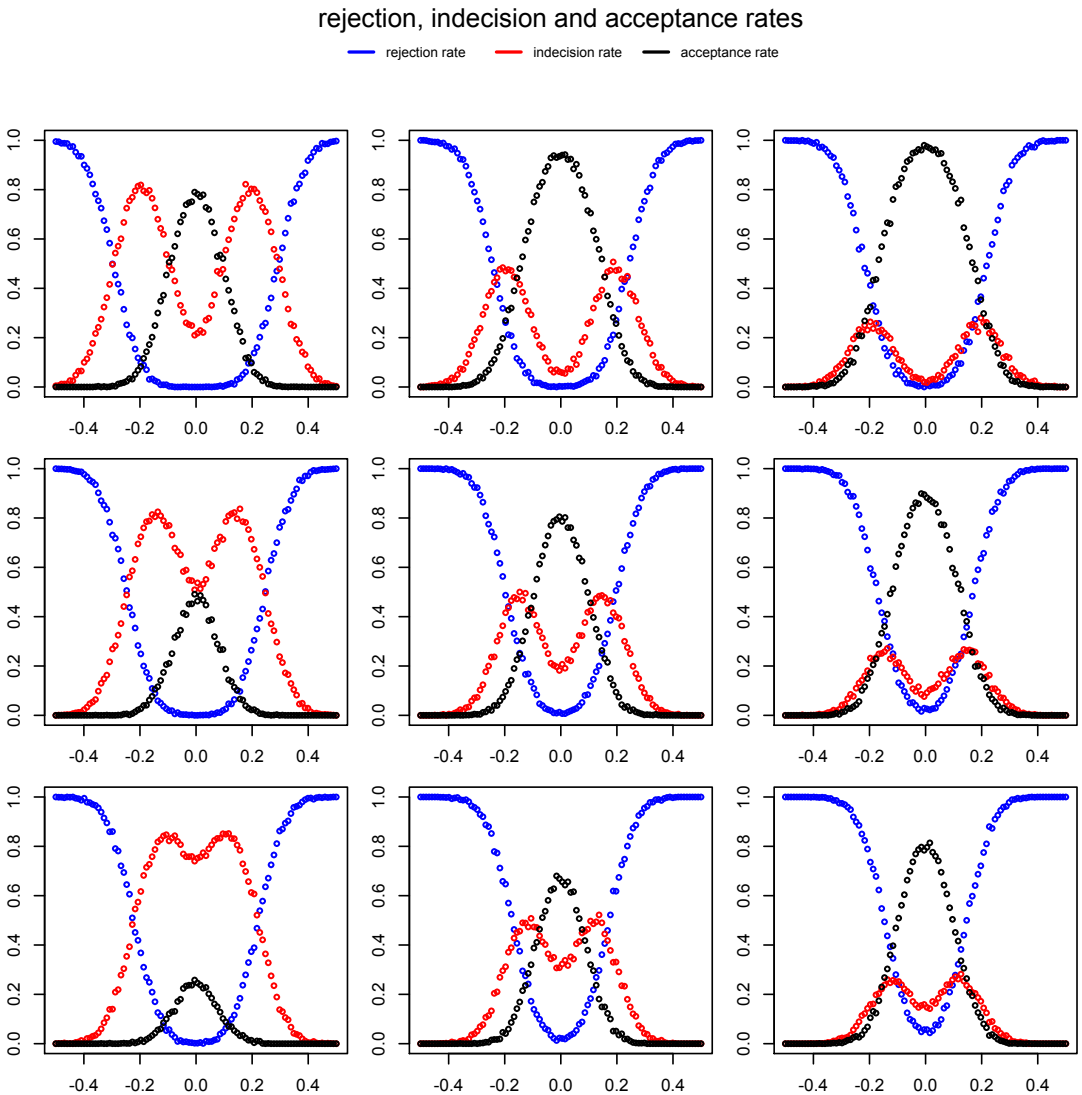
$\{5 \cdot \frac{x_1+\ldots+x_{25}}{25} : (x_1, \ldots, x_{25}) \in R\}$

Set-valued p-values and associated set-valued tests

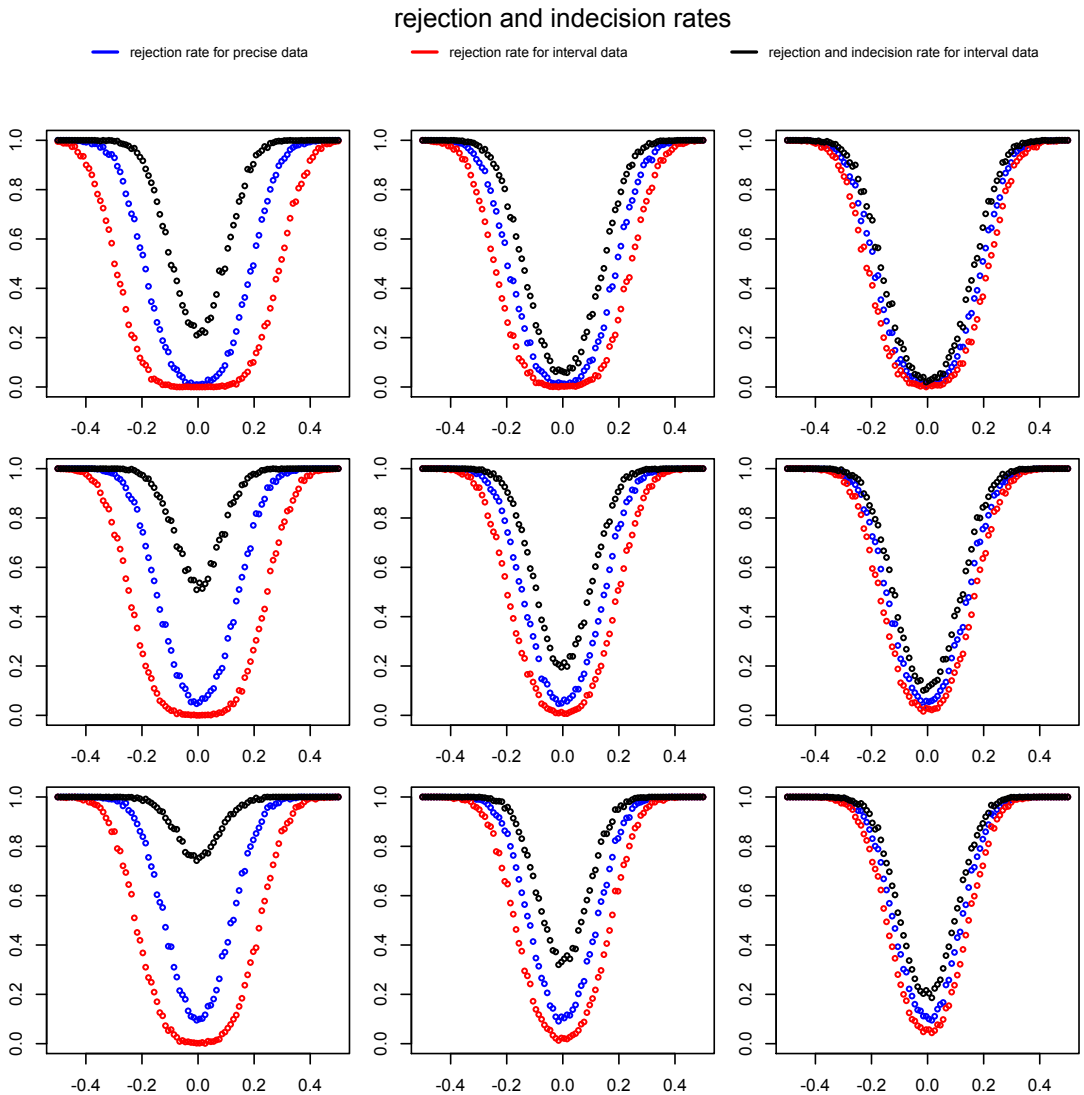
- Let $\gamma \subseteq \mathbb{R}^n$ represent our imprecise information about the sample realization $\mathbf{x} = (x_1, \ldots, x_n)$.
- Set of possible values for the p-value:
  $\mathrm{p}_{\mathrm{val}}(\gamma) = \{\mathrm{p}_{\mathrm{val}}(\mathbf{y}) : \mathbf{y} \in \gamma\}$



Example: generalized MWW test



MWW test, bounds of rejection rates



Conclusion

There exists a coherent range of set-functions combining interval and probability for the representation of uncertainty.

- Imprecise probability is the proper theoretical umbrella.
- The choice between set-functions depends on how expressive it is necessary to be in a given application.
- There exist simple practical representations of imprecise probability.
- The statistical analysis from interval data is much related to imprecise probabilities:
  - The upper and lower probabilities of the multi-valued mapping are ∞-order capacities.
  - Sometimes, the information provided by the multi-valued mapping about the distribution of the "original" random variable is not fully represented by means of the credal set associated to those capacities, but by some proper subset.