Fuzzy Data Analysis
Statistics with Fuzzy Data

Prof. Dr. Rudolf Kruse
Two Interpretations of the Research on Fuzzy Data Analysis

**FUZZY data analysis** = Fuzzy Techniques for the analysis of (crisp) data

In this course: Fuzzy Clustering

**FUZZY DATA analysis** = Analysis of Data that are described by Fuzzy Sets

In this course: Statistics with Fuzzy Data
Two Interpretations of Fuzzy Data

Epistemic view of fuzzy data

Fuzzy Sets are used to represent incomplete knowledge about an underlying object

Example: The object is Rudolf, a fuzzy set characterizes the knowledge about his (unknown) age

Memberships are subjective

Ontic view of fuzzy data

Fuzzy Sets are considered as real, complex, graded entities

Example: The object is a photo, a fuzzy set characterizes the content by grey level pixels

Memberships are objective, the object really exists
Example

datum: “R. ate 2 or 3 eggs yesterday”
This datum is imprecise, it should be modelled by the subset {2,3}.

Datum: “R. ate a low number of eggs”
This datum is subjective, it could be modelled by a fuzzy sets of the Natural Numbers (including 0) with an epistemic interpretation.

Crucial Question: What is the meaning of a membership degree and where do the numbers come from?

In real applications it is recommended to give a formal meaning together with a measurement method for the membership degrees values.
Gradual degrees are often used in Questionnaires:

**Likert Scale**

In general, how would you rate the quality of Fictionals chocolate ice cream?

- Poor
- Fair
- Good
- Very Good
- Excellent

**Slider Scale**

In general, how would you rate the quality of Fictionals chocolate ice cream?

The values of the scales can be transformed to the unit interval.
Fuzzy Membership Degrees
Fuzzy Membership Degrees

There are several different meanings of fuzzy membership degrees in real applications, often you’ll find a

- Possibilistic Interpretation
- Preference-based Interpretation
- Frequentistic Interpretation
- Similarity-based Interpretation
Fuzzy Data based on Possibilistic Scale

Fuzzy Datum: R. ate approximately two eggs.

The fuzzy data are found by a reasoning process using possibilities based on physical information (how much can R. eat?) as well as epistemic information (is it possible that R. ate x eggs?

There are close links between possibility theory to Fuzzy Set Theory. In the next chapter we will study the epistemic view in more detail.
Datum: R. ate approximately two eggs. The fuzzy data are found by a reasoning process using preferences. How many eggs does R. prefer today?

Often this interpretation is used in optimization tasks.
Datum: R. ate approximately two eggs.
The fuzzy data are found by a statistical analysis:
What did he eat in the last days?

Probabilistic uncertainty is modelled by statistical methods or by a version of subjective probability theory. Often the probabilities are finally transformed into a fuzzy scale.
Fuzzy Data based on similarity information

Datum: R. ate approximately two eggs. The fuzzy data are found by similarity based (or case based) reasoning. Is a day with a similar situation for R. is known? Often the similarities are transformed into a fuzzy scale.

Similarity analysis is used in lots of other scientific disciplines.
In real applications the data analysis you have to figure out (e.g. by interviews with expert), what exactly the meaning of the membership degrees is.

Different interpretations for the same fuzzy set can lead to completely different algorithms and evaluations.
## Descriptive Analysis of Imprecise Data

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Data (sample)</th>
<th>mean (arithmetic)</th>
<th>Deviation (standard)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>0.5, 2, 0.5</td>
<td>$x = \frac{1}{n} \sum x_i = 1.5$</td>
<td>$s = \sqrt{\frac{1}{n-1} \sum (x_i - x)^2} = 1.5$</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>$[0.1, 1, 0.2, 1, 0.1]$</td>
<td>Interval arithmetic $[\frac{2}{5}, \frac{4}{5}]$</td>
<td>$2$</td>
</tr>
<tr>
<td></td>
<td>$[0.1, 0.1, 0.1]$</td>
<td>$[0, 1]$</td>
<td>$0$</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>$[0.1, 1, 0.2, 1, 0.1]$</td>
<td>Extension principle: ${ \frac{x+y+2}{3}</td>
<td>x \in [0,1], y=2, z \in [0,1] }$</td>
</tr>
<tr>
<td></td>
<td>$[0.1, 0.1, 0.1]$</td>
<td>$[0, 1]$</td>
<td>$[0, \max { S(x,y,z)</td>
</tr>
</tbody>
</table>
Epistemic Fuzzy Data

In Fuzzy Control: similarity based fuzzy data

In Statistics: possibility based fuzzy data
Proposal 1: Similarity as Equivalence Relation?

**Definition**

Let $A$ be a set and $\approx$ be a binary relation on $A$. $\approx$ is called an equivalence relation if and only if $\forall a, b, c \in A$,

(i) $a \approx a$ (reflexivity)

(ii) $a \approx b \leftrightarrow b \approx a$ (symmetry)

(iii) $a \approx b \land b \approx c \rightarrow a \approx c$ (transitivity).

Let us try $a \approx b \iff |a - b| < \varepsilon$ where $\varepsilon$ is fixed.

$\approx$ is not transitive, $\approx$ is no equivalence relation (Poincaré paradox)

A classical equivalence relation is not able to model similarity. What about a fuzzy version?
Proposal 2: Similarity as Fuzzy Equivalence Relation?

Definition

A function $E : X^2 \rightarrow [0, 1]$ is called a fuzzy equivalence relation with respect to a t-norm $\top$ if it satisfies for all $x, y, z \in X$ the properties

1. $E(x, x) = 1$ (reflexivity)
2. $E(x, y) = E(y, x)$ (symmetry)
3. $\top(E(x, y), E(y, z)) \leq E(x, z)$ (t-transitivity).

$E(x, y)$ is the degree to which $x \approx y$ holds. The fuzzy relation $E$ is also called similarity relation, $t$-equivalence relation, indistinguishability operator, or tolerance relation.

Note that property (iii) corresponds to the fuzzy logic statement

$$\text{if } (x \approx y) \land (y \approx z) \text{ then } x \approx z.$$
Proposal 3: Similarity as Lukasiewics Equivalence Relation?

Let $\delta$ be a pseudo metric on $X$, and let $T(a, b) = \max\{a + b - 1, 0\}$ be the Łukasiewicz $t$-norm. Then $E_\delta$, defined by $E_\delta(x, y) = 1 - \min\{\delta(x, y), 1\}$ is a fuzzy equivalence relation with respect to $T$.

For the Łukasiewicz T-norm Fuzzy equivalence and distance are dual notions. This is the only T-norm with this property.

Definition

A function $E : X^2 \rightarrow [0, 1]$ is called a (Łukasiewicz) similarity relation iff

(i) $E(x, x) = 1$ (reflexivity)
(ii) $E(x, y) = E(y, x)$ (symmetry)
(iii) $\max\{E(x, y) + E(y, z) - 1, 0\} \leq E(x, z)$ (Łukasiewicz transitivity)

holds for all $x, y, z \in X$. 
Fuzzy Set describe Local Similarity

Simple Example

\[ \delta(x, y) = |x - y| \]
\[ E_\delta(x, y) = 1 - \min\{ |x - y|, 1 \} \]

Metric
Similarity relation

\[ \mu_{x_0} : X \to [0, 1] \]
\[ x \mapsto E_\delta(x, x_0) \] Fuzzy Singleton

\[ \mu_{x_0} \] describes “local” similarity of points \( x \) to \( x_0 \). The membership degree is interpreted as a similarity degree.
Given a family of fuzzy sets that describes "local" similarities.

There exists a similarity relation on $X$ with induced singletons $\mu_i$ if and only if

$$\forall i, j : \sup_{x \in X} \{\mu_i(x) + \mu_j(x) - 1\} \leq \inf_{y \in X} \{1 - |\mu_i(y) - \mu_j(y)|\}.$$  

Control Engineers often have this intuitive understanding of a fuzzy datum. Mamdani control can be seen as a similarity based interpolation.
Mamdani Control can be seen as Similarity Based Interpolation
It defines a graph that follows the pyramids

\[ \text{if } x \text{ is } \textit{large} \text{ then } y \text{ is } \textit{large} \]
How to model possibility?

Given: It was ‘cloudy’, yesterday at 18, at my home.

Fuzzy set $\mu_{\text{cloudy}} : X \rightarrow [0, 1]$, where $X = [0, 100]$. $x \in X$ clouding degrees in percent. The interpretation is as follows:

It exists a true clouding degree. This value is unknown, but there are additional information about the possibility of the different options.

The possibility of $x$ is modelled by the membership degree of $x$: 0 means impossible, 1 totally possible, degrees between 0 and 1 indicate partial possibility.
Possibility Distributions

A function $\pi : X \rightarrow [0, 1]$ is called a possibility distribution $\pi$ iff there is an $x_0 \in X$ with $\pi(x_0) = 1$.

From a mathematical point of view they are special fuzzy sets.

$\pi(u)$ is interpreted as the subjective degree of “possibility” (which is different from its probability). It quantifies a state of knowledge.

$\pi(u) = 0$: $u$ is rejected as impossible

$\pi(u) = 1$: $u$ is totally possible

Specificity of possibility distributions:

$\pi$ is at least as specific as $\pi'$ iff

for each $x$: $\pi(x) \leq \pi'(x)$ holds.
Possibility Measures

Let $\pi : X \rightarrow [0, 1]$ a possibility distribution.

Possibility degree of $A \subseteq X$: $\Pi(A) := \sup \{\pi(x) : x \in A\}$

Necessity degree of $A \subseteq X$: $N(A) := \inf \{1 - \pi(x) : x \in \overline{A}\}$.

$\Pi(A)$ evaluates to what extent $A$ is consistent with $\pi$

$N(A)$ evaluates to what extent $A$ is certainly implied.

Duality expressed by: $N(A) = 1 - \Pi(\overline{A})$ for all $A$.

It holds:

$$
\Pi(X) = 1 \\
\Pi(\emptyset) = 0, \text{ and}$$

$$\Pi(A \cup B) = \max \{\Pi(A), \Pi(B)\} \text{ for all } A \text{ and } B$$

$$\Pi(A \cap B) \leq \min \{\Pi(A), \Pi(B)\} \text{ for all } A \text{ and } B$$

Data Analysts often have this intuitive understanding of a fuzzy datum. The true original of fuzzy datum is unknown, but there is additional information about the possibility of the different options.
Standard statistical data analysis is based on random variables $X : \Omega \rightarrow R$, where $\Omega$ denotes the set of elementary event, and $R$ the set of real numbers.

The concept of a random set is a generalisation. A random set $\Gamma : \Omega \rightarrow 2^R$ is a random variable where the outcomes are subsets of $R$.

The concept of a fuzzy random set is a further generalization. A fuzzy random set $\Gamma : \Omega \rightarrow F(R)$ is a function where the outcome are fuzzy sets of $R$. The fuzzy sets are generated by a random mechanism.
Example: Mean Temperature

Let $\Omega$ denote the days in 2019, $P$ uniform probability distribution on $\Omega$

$U(\omega)$ Temperature on day $\omega$ at 18 h, we assume that only $T_{\text{min}}(\omega), T_{\text{max}}(\omega)$ i.e. the min-max temperatures per day are recorded, but the original values $U(\omega)$ are unknown. We know that $U(\omega)$ is between $T_{\text{min}}(\omega)$ and $T_{\text{max}}(\omega)$, What is the mean temperature in 2019 at 18 h? We can calculate lower and upper borders by calculating the expectations of the random variables $T_{\text{min}}$ and $T_{\text{max}}$.

The same method is used for handling random intervals

$\Gamma : \Omega \rightarrow 2^R$, $\Gamma(\omega) = [T_{\text{min}}(\omega), T_{\text{max}}(\omega)]$, is called a random set

$E(\Gamma) := [E(T_{\text{min}}), E(T_{\text{max}})]$ is a reasonable definition for the expected value of $\Gamma$.

This concept can be generalized from intervals to general sets.
(\Omega, 2^\Omega, P) \text{ Probability space, a random set is a mapping } \Gamma : \Omega \to 2^R

For an epistemic interpretation of the sets we define

\[ E(\Gamma) = \{ E(U) \mid U \text{ is random variable such that } E(U) \text{ exists and } U(\omega) \in \Gamma(\omega) \text{ for all } \omega \in \Omega \} \]

This method can be used for other quantities such as the variance.

Often subjective information about the (unknown) original data is available. In that case we can describe the data by fuzzy sets (with a possibilistic interpretation).

Using the extension principle we create a theory for descriptive statistics with fuzzy data (with an epistemic interpretation). Note that there are different models for fuzzy data with an ontic interpretation.
Expected Value of a fuzzy random variable

\[ \Gamma : \Omega \rightarrow F(\mathbb{R}) \] fuzzy random variable

The original random variable \( U^* : \Omega \rightarrow \mathbb{R} \) is unknown

Given a random variable \( U : \Omega \rightarrow \mathbb{R} \), we can evaluate the possibility, that \( U \) is the original, by

\[ \inf_{\omega \in \Omega} \{ (\Gamma(\omega))(U(\omega)) \} \]

Using the extension principle, a reasonable definition for the expected value of \( \Gamma \) is obtained:

Expected value \( E(\Gamma) : \mathbb{R} \rightarrow [0, 1] \) fuzzy set of \( X \):

\[ x \mapsto \sup_{U : E(U) = x} \left\{ \inf_{\omega \in \Omega} \{ (\Gamma(\omega))(U(\omega)) \} \right\} \]

With the same method other descriptive values can be defined.