データ科学の意思決定写像による統一的体系化の試み

CDS 早稲田大学 データ科学センター

内容

- 意思決定写像を具体例を用いて説明

意思決定写像の標準化図

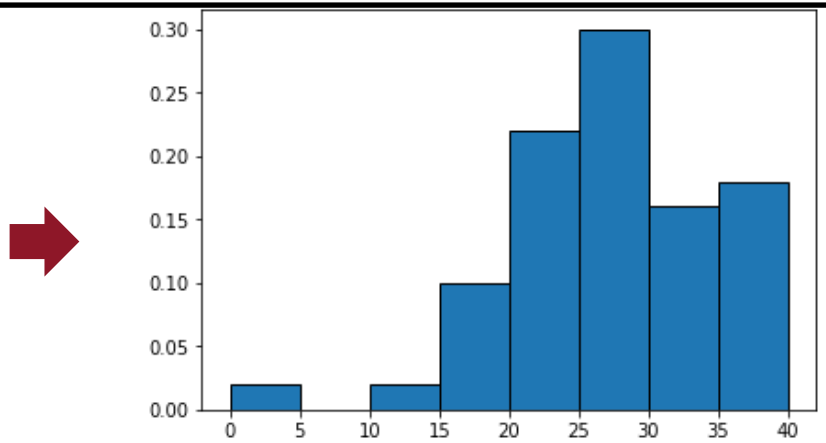
目的：
設定：
評価基準：

実際は？

意思決定写像の標準化図

問題　あるコンビニの売り上げデータ50日分がある.
このデータの特徴を視覚的に把握したい

23.5, 28.4, 37.4, 26.5, 29.5, 28.1, 31.9, 25.9, 31.5, 23.3
・
・
25.5, 19.0, 38.8, 28.1, 21.0, 21.1, 22.6, 26.8, 37.7, 30.2



単位：万円
（50日分）

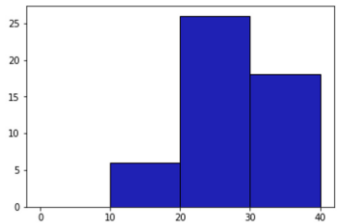
意思決定写像の標準化図

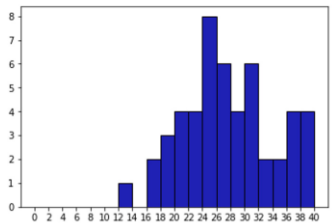
目的：数量データ $x_1, x_2, \ldots, x_{50}$ の特徴を記述
設定：視覚的に把握するためにヒストグラムを用いる
評価基準：視覚的にデータの特徴をうまく捉えているか

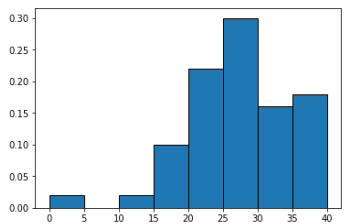
$x_1, x_2, \ldots, x_{50} \in \mathbb{R}_+^{50}$ → **意思決定写像** → 決定集合 $d \in \mathcal{B}$

視覚的
特徴記述

$\mathcal{B}$







…

意思決定写像の標準化図

問題　あるコンビニの売り上げデータ50日分がある.
このデータの特徴を数値的に把握したい

23.5, 28.4, 37.4, 26.5, 29.5, 28.1, 31.9, 25.9, 31.5, 23.3
⋮
25.5, 19.0, 38.8, 28.1, 21.0, 21.1, 22.6, 26.8, 37.7, 30.2

単位：万円
（50日分）

➡ 算術平均，中央値．．．

代表値

意思決定写像の標準化図

目的：数量データ $x_1, x_2, \ldots, x_{50}$ の特徴を記述
設定：一つの代表値 z で特徴記述
評価基準：各x_i と z との距離の合計（最小化）

$x_1, x_2, \ldots, x_{50} \in \mathbb{R}_+^{50}$ → **意思決定写像** → $z \in \mathbb{R}_+$

2乗距離 $(x_1-z)^2 + (x_2-z)^2 + \cdots + (x_{50}-z)^2$ 最小化 $z = \frac{1}{50}\sum_{i=1}^{50} x_i$

絶対値距離 $|x_1 - z| + |x_2 - z| + \cdots + |x_{50} - z|$ ↓中央値 $z = \frac{x_{(25)}+x_{(26)}}{2}$

意思決定写像の標準化図

目的：数量データ $x_1, x_2, \ldots, x_n$ の特徴を記述
設定：一つの代表値 z で特徴記述
評価基準：各x_i と z との2乗距離の合計（最小化）

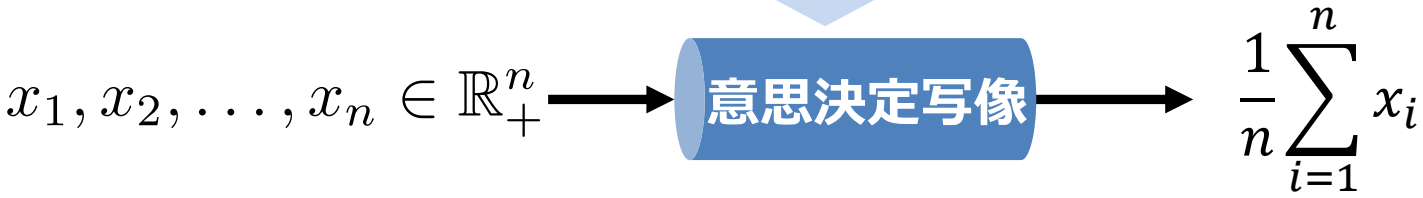
$x_1, x_2, \ldots, x_n \in \mathbb{R}_+^n$ → 意思決定写像 → $\frac{1}{n}\sum_{i=1}^{n} x_i$

目的：数量データ $x_1, x_2, \ldots, x_n$ の特徴を記述
設定：一つの代表値 z で特徴記述
評価基準：各x_i と z との絶対値距離の合計（最小化）

$x_1, x_2, \ldots, x_n \in \mathbb{R}_+^n$ → 意思決定写像 → 中央値

意思決定写像の標準化図

目的：数量データ $x_1, x_2, \ldots, x_n$ の特徴を記述
設定：一つの代表値 z で特徴記述
評価基準：各x_i と z との2乗距離の合計（最小化）



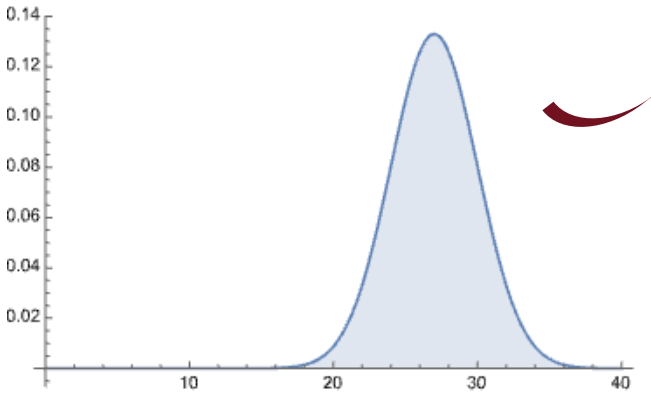
同じデータに対して異なる分析方法がありうる

意思決定写像を用いて整理
目的，設定，評価基準，写像の入出力 → 理解の一助

データの発生に関するメカニズム

問題　あるコンビニの売り上げデータ50日分がある.
このデータの背後に潜む**発生メカニズム**を知りたい　　単位：万円

23.5, 28.4, 37.4, 26.5, 29.5, 28.1, 31.9, 25.9, 31.5, 23.3
⋮
25.5, 19.0, 38.8, 28.1, 21.0, 21.1, 22.6, 26.8, 37.7, 30.2



各サンプルは独立に正規分布にしたがう

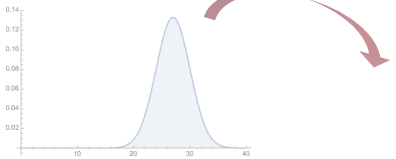
$$x_i \sim \mathcal{N}(\mu, \sigma^2)$$

確率変数　（情報理論界隈では. . X）

意思決定写像の標準化図

目的：数量データ x_i の従うメカニズムを明らかにしたい（構造推定）
設定：サンプルは独立に正規分布 $\mathcal{N}(\mu, \sigma^2)$ に従う

評価基準：尤度（最大化）



$$\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n \rightarrow \boxed{\text{意思決定写像}} \rightarrow \hat{\mu}$$

対数尤度

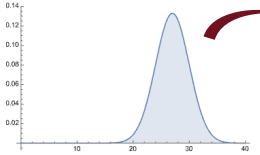
$$l(\mu, \sigma^2) = -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log 2\pi\sigma^2$$ 最大化

$$\sum_{i=1}^{n} (x_i - \mu)^2$$ 最小化

意思決定写像の標準化図

目的：数量データ x_iの従うメカニズムを明らかにしたい（構造推定）
設定：サンプルは独立に正規分布 $\mathcal{N}(\mu, \sigma^2)$に従う

評価基準：尤度（最大化）



$\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$ → **意思決定写像** → $\underset{\sim}{\hat{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{x}_i$

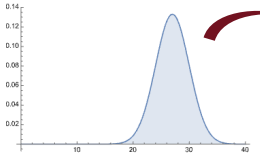
目的：数量データ $x_1, x_2, \ldots, x_n$ の特徴を記述
設定：一つの代表値 z で特徴記述

評価基準：各x_i と z との2乗距離の合計（最小化）

$x_1, x_2, \ldots, x_n$ → **意思決定写像** → $\frac{1}{n}\sum_{i=1}^{n} x_i$

意思決定写像の標準化図

目的：数量データ x_iの従うメカニズムを明らかにしたい（構造推定）
設定：サンプルは独立に正規分布 $\mathcal{N}(\mu, \sigma^2)$に従う
評価基準：尤度（最大化）



$\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$ → 意思決定写像 → $\underset{\sim}{\hat{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{x}_i$

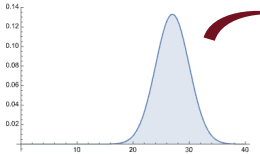
目的：数量データ $x_1, x_2, \ldots, x_n$ の特徴を記述
設定：一つの代表値 z で特徴記述
評価基準：各x_i と zとの2乗距離の合計（最

目的，設定は
異なっていても
意思決定写像の出力が
等しくなることもある

$x_1, x_2, \ldots, x_n$ → 意思決定写像 → $\frac{1}{n}\sum_{i=1}^{n} x_i$

意思決定写像の標準化図

目的：数量データ x_iの従うメカニズムを明らかにしたい（構造推定）
設定：サンプルは独立に正規分布 $\mathcal{N}(\mu, \sigma^2)$に従う
評価基準：尤度（最大化）



$\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$ → **意思決定写像** → $\underset{\sim}{\hat{\mu}} = \frac{1}{n}\sum^{n} \underset{\sim}{x}_i$

目的：数量データ $x_1, x_2, \ldots, x_n$ の特徴を記述
設定：一つの代表値 z で特徴記述
評価基準：各x_i と zとの2乗距離の合計（最

目的，設定は
異なっていても
意思決定写像の出力が
等しくなることもある

データ科学の特徴：**意思決定写像の考え方およびその図**を
用いて整理して伝えたい