データの分析と知識発見

Introduction to Data Analysis

今回の構成

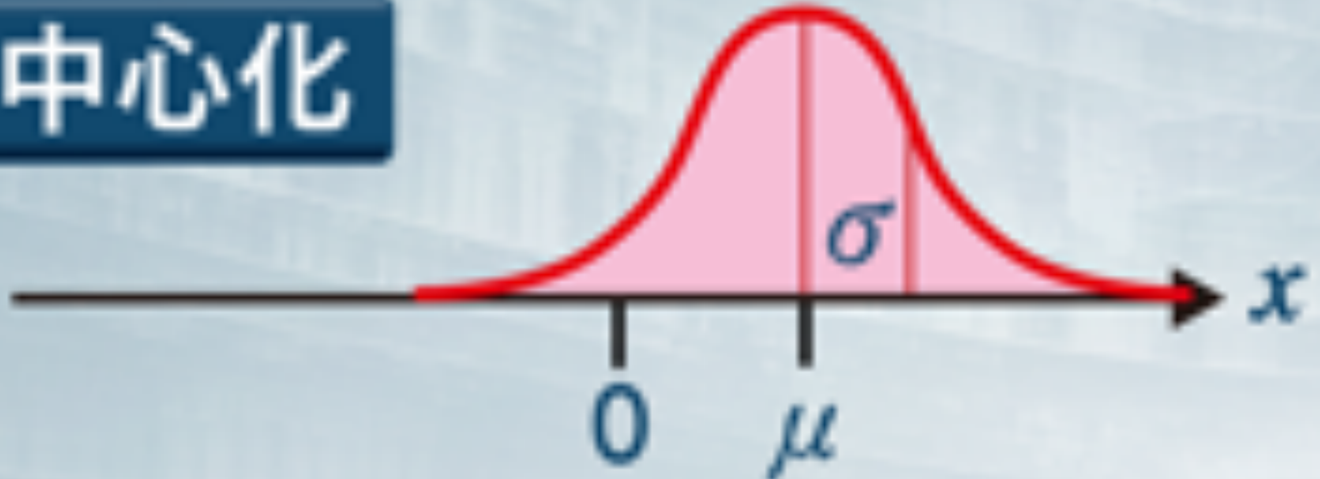
- 因子分析の考え方について学ぶ
- 因子分析の分析の手順について学ぶ
- Rを用いて因子分析を行う

主成分分析

- 変数の特徴をなるべく保持したまま、より少ない変数（主成分）で表す
- 直交回転：
  ばらつきが大きい方向を探す
- データ間の関係
- データの圧縮

中心化 と 標準化

中心化



$$x_i^{(p)\prime} = (x_i^{(p)} - \mu_i)$$

標準化

$$x_i^{(p)\prime\prime} = \frac{(x_i^{(p)} - \mu_i)}{\sigma_i}$$

因子分析とは

- 観測されたデータの背後に潜む因子を見つける
- それぞれの変数に固有の独自因子と共通して潜む共通因子を用いて表現する
- 分析を元に変数の解釈を行う

因子分析の定式化 と 用語

$$
\begin{aligned}
Y_1 &= c_{11} F_1 + c_{12} F_2 + \cdots + c_{1m} F_m + d_1 E_1 \\
Y_2 &= c_{21} F_1 + c_{22} F_2 + \cdots + c_{2m} F_m + d_2 E_2 \\
&\vdots \\
Y_n &= c_{n1} F_1 + c_{n2} F_2 + \cdots + c_{nm} F_m + d_n E_n
\end{aligned}
$$

$$Y = CF + DE$$

因子分析の定式化 と 用語

$$
\begin{aligned}
Y_1 &= c_{11} F_1 + c_{12} F_2 + \cdots + c_{1m} F_m + d_1 E_1 \\
Y_2 &= c_{21} F_1 + c_{22} F_2 + \cdots + c_{2m} F_m + d_2 E_2 \\
&\vdots \\
Y_n &= c_{n1} F_1 + c_{n2} F_2 + \cdots + c_{nm} F_m + d_n E_n
\end{aligned}
$$

共通因子

$$Y = CF + DE$$

因子分析の定式化と用語

$$
\begin{aligned}
Y_1 &= c_{11} F_1 + c_{12} F_2 + \cdots + c_{1m} F_m + d_1 E_1 \\
Y_2 &= c_{21} F_1 + c_{22} F_2 + \cdots + c_{2m} F_m + d_2 E_2 \\
&\vdots \\
Y_n &= c_{n1} F_1 + c_{n2} F_2 + \cdots + c_{nm} F_m + d_n E_n
\end{aligned}
$$

独自因子

$$Y = CF + DE$$

因子分析の定式化 と 用語

$$
\begin{aligned}
Y_1 &= c_{11} F_1 + c_{12} F_2 + \cdots + c_{1m} F_m + d_1 E_1 \\
Y_2 &= c_{21} F_1 + c_{22} F_2 + \cdots + c_{2m} F_m + d_2 E_2 \\
&\vdots \\
Y_n &= c_{n1} F_1 + c_{n2} F_2 + \cdots + c_{nm} F_m + d_n E_n
\end{aligned}
$$

因子負荷量

$$Y = CF + DE$$

共通因子と独自因子

Fのことを **共通因子** 、Eを **独自因子** という

1. 共通因子、独自因子は平均0、分散を1とする
2. 共通因子はどの独自因子とも相関がない
3. 独自因子はそれぞれ独立で相関がない
4. 共通因子間には相関がある場合とない場合の両方を考える

相関がない場合を**直交解**、ある場合を**斜行解**という

因子分析の手順

共通因子の個数を決める

- カイザーガットマン基準（相関行列の固有値が1以上の個数）
- スクリープロット（相関行列の固有値のグラフから判断）

因子負荷量を求める

- 最尤法
- 最小二乗法
- 主因子法

因子軸の回転を行う

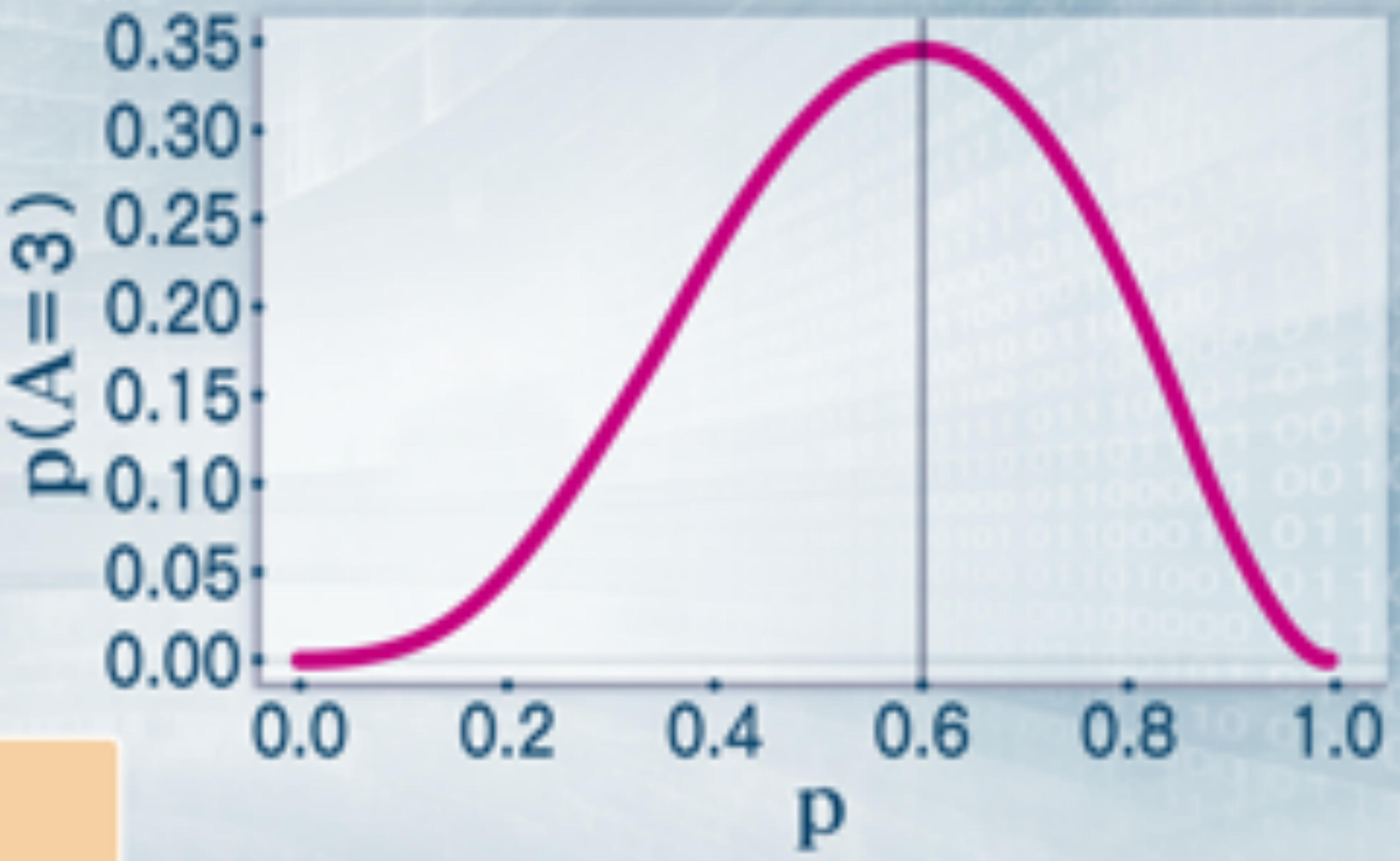
結果を解釈する

最尤推定量

AとBとが戦う
AがBに勝つ確率ｐは
毎回同じであるとする
５回戦ったらAが３回勝った
ｐをいくつと予想したら
良いか？

最尤推定量

AとBとが戦う
AがBに勝つ確率pは
毎回同じであるとする
5回戦ったらAが3回勝った
pをいくつと予想したら
良いか？



$$P(A=3)={}_5C_3\,p^3(1-p)^2$$
$$=10(p^5-2p^4+p^3)$$

因子の回転

因子負荷量は一意に決まるわけではない

回転によって、特徴が分かりやすく、より解釈しやすい軸を探す

- 直交回転
  - バリマックス法
  - オーソマックス規準を最大化
- 斜行回転
  - プロマックス法
  - バリマックス解を元にプロクラステス回転を行い特徴を際立たせる

Rによるシミュレーション

```
> w1 <- read.table("ch091.dat", header=T, row.names=1)
> w2 <- cor(w1)
> plot(eigen(w2)$values)
```

```
> w3 <- factanal(w1,2,rotation="promax")
```

```
> plot(w3$loadings)
> text(w3$loadings,rownames(w3$loadings), pos=1)
```

Rによるシミュレーション

```
> w1 <- read.table("ch091.dat", header=T, row.names=1)
> w2 <- cor(w1)
> plot(eigen(w2)$values)
```

```
> w3 <- factanal(w1,2,rotation="promax")
```

```
> plot(w3$loadings)
> text(w3$loadings,ro )
```

何もしない："none"
バリマックス法："varimax"
プロマックス法："promax"

Rによるシミュレーション

```
> w1 <- read.table("ch091.dat", header=T, row.names=1)
> w2 <- cor(w1)
> plot(eigen(w2)$values)
```

```
> w3 <- factanal(w1,2,rotation="promax")
```

```
> plot(w3$loadings)
> text(w3$loadings,rownames(w3$loadings), pos=1)
```