講義にて配布した資料の抜粋・縮小版です

平成22年度
統計工学　講義資料（12）

～変数間の関係性～

(C) Sadami SUZUKI

繰り返しのない二元配置分散分析

（繰り返しのない）
【二元配置データ】

- 特性値：収集するデータ（観測値）
- 因子：特性値に影響を及ぼす要因
  - 要因1：因子A（水準数 a）
  - 要因2：因子B（水準数 b）

x_{ij} ・・・水準 A_iB_j における観測値

		B			
		B_1	B_2	$\cdots$	B_b
A	A_1	x_{11}	x_{12}	$\cdots$	x_{1b}
	A_2	x_{21}	x_{22}	$\cdots$	x_{2b}
	$\vdots$	$\vdots$	$\vdots$	x_{ij}	$\vdots$
	A_a	x_{a1}	x_{a2}	$\cdots$	x_{ab}

(C) Sadami SUZUKI

構造模型によるデータ構造の理解

繰り返しのない二元配置データ

		B			
		B_1	B_2	$\cdots$	B_b
A	A_1	x_{11}	x_{12}	$\cdots$	x_{1b}
	A_2	x_{21}	x_{22}	$\cdots$	x_{2b}
	$\vdots$	$\vdots$	$\vdots$	x_{ij}	$\vdots$
	A_a	x_{a1}	x_{a2}	$\cdots$	x_{ab}

- 二元配置データが無作為に抽出・収集された場合
  - 全部で ab 回のデータ収集の無作為化

データ構造

$$X_{ij} = \mu_{ij} + \varepsilon_{ij} \qquad (i=1,\cdots,a,\ j=1,\cdots,b)$$

構造模型

$$X_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$$

- 交互作用項と残差項は**添字が同一**であることから厳密には**区別できない！**
- 交互作用が存在すると主効果の検定すらままならない

(C) Sadami SUZUKI

構造模型によるデータ構造の理解

一元配置データ

A_1	x_{11}	x_{12}	$\cdots$	x_{1r}
A_2	x_{21}	x_{22}	$\cdots$	x_{2r}
$\vdots$	$\vdots$	$\vdots$	x_{ij}	$\vdots$
A_a	x_{a1}	x_{a2}	$\cdots$	x_{ar}

- 一元配置データが無作為に抽出・収集された場合
  - 全部で ar 回のデータ収集の無作為化
  - 1つのデータを取るたびに因子Aの水準を設定し直す

データ構造

$$X_{ij} = \mu_i + \varepsilon_{ij} \qquad (i=1,\cdots,a,\ j=1,\cdots,r)$$

構造模型

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

【因子Aの効果に関する解析】

因子Aの効果・・・群内（残差）平方和と比較した
群間（因子A）平方和の大きさ

※ 因子Aの効果と残差の添字が異なることに注意！

(C) Sadami SUZUKI

構造模型によるデータ構造の理解

繰り返しのある二元配置データ

	B_1	$\cdots$	B_b
A_1	x_{111}	$\cdots$	x_{1b1}
	$\vdots$	$\vdots$	$\vdots$
	x_{11r}	$\cdots$	x_{1br}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_a	x_{a11}	$\cdots$	x_{ab1}
	$\vdots$	$\vdots$	$\vdots$
	x_{a1r}	$\cdots$	x_{abr}

A_1B_1	x_{111}	$\cdots$	x_{11r}
A_1B_2	x_{121}	$\cdots$	x_{12r}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_aB_b	x_{ab1}	$\cdots$	x_{abr}

- 二元配置データが無作為に抽出・収集された場合
  - 全部で abr 回のデータ収集の無作為化

データ構造

$$X_{ijk} = \mu_{ij} + \varepsilon_{ijk} \qquad (i=1,\cdots,a,\ j=1,\cdots,b,\ k=1,\cdots,r)$$

構造模型

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

- **残差平方和**と比較した**それぞれの主効果**および**交互作用平方和**の大きさを検討
- **それぞれの要因効果，交互作用項と残差項の添字が異なることから，バラツキが区別（分離）可能！**

(C) Sadami SUZUKI

構造模型によるデータ構造の理解

繰り返しのない二元配置データ

		B			
		B_1	B_2	$\cdots$	B_b
A	A_1	x_{11}	x_{12}	$\cdots$	x_{1b}
	A_2	x_{21}	x_{22}	$\cdots$	x_{2b}
	$\vdots$	$\vdots$	$\vdots$	x_{ij}	$\vdots$
	A_a	x_{a1}	x_{a2}	$\cdots$	x_{ab}

- 二元配置データが無作為に抽出・収集された場合
  - 全部で ab 回のデータ収集の無作為化

データ構造

$$X_{ij} = \mu_{ij} + \varepsilon_{ij} \qquad (i=1,\cdots,a,\ j=1,\cdots,b)$$

構造模型

$$X_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$$

- 交互作用項と残差項は**添字が同一**であることから厳密には**区別できない！**
- 交互作用が存在すると主効果の検定すらままならない

交互作用の影響と水準 A_iB_j 内のバラツキの影響を区別するためには，
水準 A_iB_j におけるデータの繰り返し測定が必要！

(C) Sadami SUZUKI

講義にて配布した資料の抜粋・縮小版です

繰り返しのない二元配置分散分析

- 店舗と商品カテゴリーの間には交互作用はないと見なし，**それぞれの要因効果（主効果）に関する検定**を行う

平方和分解

$$S_T = S_A + S_B + S_e$$

要因	平方和	自由度	平均平方（不偏分散）	F比（F値）
A	S_A	$\phi_A = a-1$	$V_A = S_A / \phi_A$	$F_A = V_A / V_e$
B	S_B	$\phi_B = b-1$	$V_B = S_B / \phi_B$	$F_B = V_B / V_e$
e	S_e	$\phi_e = (a-1)(b-1)$	$V_e = S_e / \phi_e$	
T	S_T	$\phi_T = ab-1$		

(C) Sadami SUZUKI

基本統計量の整理

1変数に関する情報

- 平均値

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- （不偏）分散
  - 平均値からのバラツキ具合を表す指標

$$\hat{\sigma}_X^2 = V_X = \frac{(X_1-\bar{X})^2 + (X_2-\bar{X})^2 + \cdots (X_n-\bar{X})^2}{n-1} = \frac{1}{n-1}\sum(X_i-\bar{X})^2 = \frac{1}{n-1}\cdot S_X$$

- 標準偏差
  - 分散の平方根
  - 当該データと同一の単位
  - 「バラツキの平均」に近いイメージの値

2変数に関する情報

- 共分散
  - 2変数間の親密度を表す指標

$$\hat{\sigma}_{XY} = \frac{(X_1-\bar{X})(Y_1-\bar{Y}) + \cdots + (X_i-\bar{X})(Y_i-\bar{Y})}{n-1} = \frac{1}{n-1}\sum(X_i-\bar{X})(Y_i-\bar{Y}) = \frac{1}{n-1}\cdot S_{XY}$$

- 相関係数

$$r_{xy} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} = \frac{S_{XY}}{\sqrt{S_X S_Y}}$$

相関係数はデータの一次変換を保持する

(C) Sadami SUZUKI

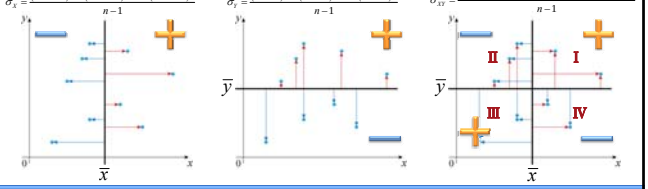
相関係数・共分散の視覚的理解

相関係数

$$r_{xy} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} = \frac{S_{XY}}{\sqrt{S_X S_Y}}$$

- 偏差積の符号より・・・
  - 領域Ⅰ，Ⅲにデータが多く分布していれば，**正の相関**
  - 領域Ⅱ，Ⅳにデータが多く分布していれば，**負の相関**
  - 全ての領域に均等に分布している場合は，**無相関**
- **両変数の標準偏差で割ることにより，基準化したものが相関係数**

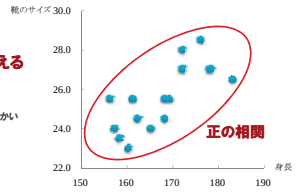
$$\hat{\sigma}_X^2 = \frac{(X_1-\bar{X})^2 + (X_2-\bar{X})^2 + \cdots (X_n-\bar{X})^2}{n-1} \qquad \hat{\sigma}_Y^2 = \frac{(Y_1-\bar{Y})^2 + (Y_2-\bar{Y})^2 + \cdots (Y_n-\bar{Y})^2}{n-1} \qquad \hat{\sigma}_{XY} = \frac{(X_1-\bar{X})(Y_1-\bar{Y}) + \cdots + (X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

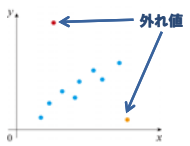


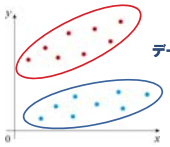
(C) Sadami SUZUKI

2変数の関係の視覚化：散布図

- **散布図**を描くことで変数間の関係性を**視覚的に捉える**
  - **外れ値（異常値）の発見**
    - 外れ値となる原因の考察
    - 他のサンプルにはない新たな特徴を変数化することで，より細かい分析が可能となることもある
  - **層別要因の発見**
    - データ群を分けて分析する必要性の発見
    - より適切な関係性の把握







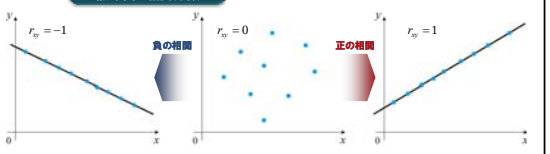
(C) Sadami SUZUKI

相関のパターン

相関係数

$$-1 \le r_{xy} \le 1$$

散布図と相関係数



負の傾きをもつ直線にすべてのデータがのるパターン

正の傾きをもつ直線にすべてのデータがのるパターン

(C) Sadami SUZUKI

無相関検定

- どの程度の相関係数であれば，**統計的に有意な**相関関係と言えるのか？
  - 同じ相関係数でもサンプル数によって**統計的な意味合いが変わる**

帰無仮説

$$H_0 : \rho = 0$$

- 確率変数XとYの間の母相関係数 $\rho = 0$ の下では・・・

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

自由度 n-2 の t 分布に従う

- 有意水準5%，1%での検定で帰無仮説が棄却される相関係数とサンプル数の関係は右表の通り

サンプル数	有意水準5%	有意水準1%
3	0.997	1.000
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708
13	0.553	0.684
14	0.532	0.661
15	0.514	0.641
16	0.497	0.623
17	0.482	0.606
18	0.468	0.590
19	0.456	0.575
20	0.444	0.561
30	0.361	0.463
40	0.312	0.403
50	0.279	0.361
60	0.254	0.330
70	0.235	0.306
80	0.220	0.286
90	0.207	0.270
100	0.197	0.256
200	0.139	0.182
300	0.113	0.149
400	0.098	0.129

(C) Sadami SUZUKI