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（10）

～統計的推論：分散分析～

(C) Sadami SUZUKI

複数の集団における集団間の差異

- 前回までに学んだ店舗Xと店舗Yにおける満足度の比較に対し，店舗Zを新たに加え，3つの店舗間で顧客満足度に差があるかどうかを考えたい

番号	店舗X	店舗Y	店舗Z
1	93	84	87
2	75	97	79
3	72	89	83
4	79	95	98
5	80	85	86
6	82	81	98
7	72	77	85
8	78	82	97
9	89	75	88

帰無仮説

$$H_0 : \mu_X = \mu_Y = \mu_Z$$

- 3つの店舗間で平均満足度に差はない
- 3つの店舗間のどの組合せにおいても平均満足度に差はない

対立仮説

このときの対立仮説は？

(C) Sadami SUZUKI

複数の集団における集団間の差異

- 店舗X－店舗Y，店舗X－店舗Z，店舗Y－店舗Zと**t検定を3回繰り返す**
- 「コインを1回投げて表が出る確率は？」・・・
- 「コインを2回投げて少なくとも1回は表が出る確率は？」・・・

(C) Sadami SUZUKI

分散分析（*ANOVA; Analysis of Variance*）

- 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・・・店舗
- 水準数：因子の状態（値）の数

【一元配置データ】

$$A_1 : X_{11}, X_{12}, \cdots, X_{1r} \sim N(\mu_1, \sigma^2)$$
$$A_2 : X_{21}, X_{22}, \cdots, X_{2r} \sim N(\mu_2, \sigma^2)$$
$$\vdots$$
$$A_a : X_{a1}, X_{a2}, \cdots, X_{ar} \sim N(\mu_a, \sigma^2)$$

因子　水準（水準数：a）

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

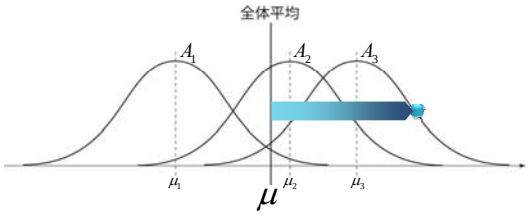
繰り返し数：n

水準 A_i における j 番目の観測値

(C) Sadami SUZUKI

一元配置分散分析の考え方

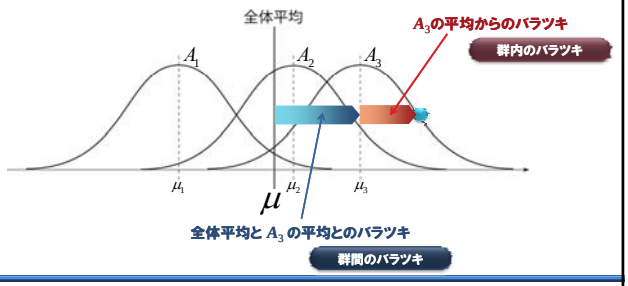
- まずは各サンプルの**全体平均からのズレ（バラツキ）**を考えよう



(C) Sadami SUZUKI

一元配置分散分析の考え方

- 全体平均からのバラツキを**分解**して考えると・・・



(C) Sadami SUZUKI

講義にて配布した資料の抜粋・縮小版です

一元配置分散分析の考え方

バラツキの分解

・（全体の平均からのバラツキ）=（群間のバラツキ）+（群内のバラツキ）

群間のバラツキ…因子の水準間（店舗間）の違い

群内のバラツキ…同一水準の群（各店舗）の中で
個々のデータがどれほどバラついているか
- 残差，誤差，個人差

(C) Sadami SUZUKI

群間および群内のバラツキの比較

分散分析の判断枠組み

・群内のバラツキと比較して，群間のバラツキが大きければ・・・
- 因子の水準間（標本集団の間）の違いが大きいということを示し，
  「母集団の平均に差がない」との帰無仮説を棄却することが可能となる
- 群内のバラツキ（誤差）に比べ，群間のバラツキが小さければ，標本集団の間の違いが大きいとはいえず，帰無仮説は棄却できない

(C) Sadami SUZUKI

一元配置データにおけるバラツキの分解

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

・全体平均からのバラツキ・・・

総平方和

$$S_T = \sum_{i=1}^{a}\sum_{j=1}^{r}\left(x_{ij} - \bar{\bar{x}}\right)^2$$

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{r} x_{ij}}{ar} \quad \cdots\text{全体平均}$$

・水準 A_i における平均・・・ $\bar{x}_i = \dfrac{\sum_{j=1}^{r} x_{ij}}{r} \qquad (i = 1, 2, \cdots, a)$

$$S_T = \sum_{i=1}^{a}\sum_{j=1}^{r}\left(x_{ij} - \bar{\bar{x}}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{r}\left\{\left(x_{ij} - \bar{x}_i\right) + \left(\bar{x}_i - \bar{\bar{x}}\right)\right\}^2$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{r}\left(x_{ij} - \bar{x}_i\right)^2 + r\sum_{i=1}^{a}\left(\bar{x}_i - \bar{\bar{x}}\right)^2 + 2\sum_{i=1}^{a}\sum_{j=1}^{r}\left(x_{ij} - \bar{x}_i\right)\left(\bar{x}_i - \bar{\bar{x}}\right)$$

$$\sum_{i=1}^{a}\sum_{j=1}^{r}\left(x_{ij} - \bar{x}_i\right)\left(\bar{x}_i - \bar{\bar{x}}\right) = \sum_{i=1}^{a}\left(\bar{x}_i - \bar{\bar{x}}\right) \times \sum_{j=1}^{r}\left(x_{ij} - \bar{x}_i\right) = 0$$

(C) Sadami SUZUKI

一元配置分散分析（*One-way ANOVA*）

Step3：検定に用いる統計量の選択とその分布に関する検討

・考察の対象：群内のバラツキと比較した群間のバラツキの大きさ・・・ $\dfrac{S_A}{S_e}$
- 総平方和の自由度：総サンプル数－1・・・ $\phi_T = ar - 1$
- 群間のバラツキ（群間平方和）の自由度：群の数（水準数）－1・・・ $\phi_A = a - 1$
- 群内のバラツキ（群内平方和）の自由度：水準数×（各水準におけるサンプル数－1）・・・ $\phi_e = \phi_T - \phi_A = a(r-1)$

・検定量：群間平方和および群内平方和をそれぞれの自由度で割った値（平均平方）の比

$$F_0 = \frac{S_A / \phi_A}{S_e / \phi_e}$$

自由度 a-1, $a(r$-1) の F 分布に従う

(C) Sadami SUZUKI

平方和計算の実際

・実際に各平方和を計算するときは，次のようにすると便利である

1. 修正項の計算

$$CT = \frac{\left\{\sum\left(\text{個々の値}\right)\right\}^2}{\left(\text{全データ数}\right)} = \frac{\left\{\sum_{i=1}^{a}\sum_{j=1}^{r} x_{ij}\right\}^2}{ar}$$

2. 総平方和の計算

$$S_T = \sum\left(\text{個々の値}\right)^2 - CT = \sum_{i=1}^{a}\sum_{j=1}^{r} x_{ij}^2 - CT$$

3. 群間平方和の計算

$$S_A = \frac{\sum\left(A_i\text{での値の和}\right)^2}{\left(A_i\text{でのデータ数}\right)} - CT = \frac{\sum_{i=1}^{a}\left(\sum_{j=1}^{r} x_{ij}\right)^2}{r} - CT$$

4. 群内平方和の計算

$$S_e = S_T - S_A$$

(C) Sadami SUZUKI

一元配置分散分析（*One-way ANOVA*）

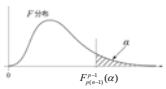
Step4：検定量を用いて仮説が正しいか否かを判定する

・分散分析表の活用

要因	平方和	自由度	平均平方（不偏分散）	F値
群間（因子A）	S_A	$\phi_A = a - 1$	S_A / ϕ_A	$F_0 = \dfrac{S_A / \phi_A}{S_e / \phi_e}$
群内（誤差）	S_e	$\phi_e = a(r-1)$	S_e / ϕ_e	
計	S_T	$\phi_T = ar - 1$		

$F_0 > F_{a(r-1)}^{a-1}(\alpha)$ となる確率がα

有意水準αで仮説は棄却される



(C) Sadami SUZUKI