データ解析の基礎（III）

ーその手法はどうして考えるのかー

安田晃

先週の復習（1）

	変量1	変量2	…	変量 p	y_i
1	x_{11}	x_{12}	…	x_{1p}	y_1
2	x_{21}	x_{22}	…	x_{2p}	y_2
⋮	⋮	⋮	⋱	⋮	⋮
n	x_{n1}	x_{n2}	…	x_{np}	y_n

という p 個の変量に対して1つの目的変数 $y_i\,(i=1,\cdots,n)$ というデータが n 人に対してあったとき，

$$y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + \cdots + a_p x_{pi}$$

という式を考えた．

先週の復習（2）

	変量1	変量2	…	変量 p	y_i
1	x_{11}	x_{12}	…	x_{1p}	y_1
2	x_{21}	x_{22}	…	x_{2p}	y_2
⋮	⋮	⋮	⋱	⋮	⋮
n	x_{n1}	x_{n2}	…	x_{np}	y_n

最小2乗法を使って，推定値 $\hat{y}_i\,(i=1,\cdots,n)$ と実際の $y_i\,(i=1,\cdots,n)$ との差の2乗和を最小とする係数を求める問題に帰着させた．

先週の復習（3）

- そのために，

$$f(a_0,a_1,a_2,\cdots,a_p)=\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(y_i - a_0 - a_1x_{1i} - a_2x_{2i} - \cdots - a_px_{pi}\right)^2$$

を $a_0,a_1,\cdots,a_p$ で偏微分し，0とおけば，

$$\frac{\partial f}{\partial a_0} = \sum_{i=1}^{n}\left(y_i - a_0 - a_1x_{i1} - \cdots a_px_{pi}\right)(-1) = 0$$

$$\frac{\partial f}{\partial a_1} = \sum_{i=1}^{n}\left(y_i - a_0 - a_1x_{i1} - \cdots a_px_{pi}\right)(-x_{1i}) = 0$$

$$\vdots$$

$$\frac{\partial f}{\partial a_p} = \sum_{i=1}^{n}\left(y_i - a_0 - a_1x_{i1} - \cdots a_px_{pi}\right)(-x_{pi}) = 0$$

先週の復習（4）

- $y_i(i=1,\cdots,n)$ と $\hat{y}_i(i=1,\cdots,n)$ の相関係数を重相関係数とよんで，

$$R = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(\hat{y}_i-\bar{\hat{y}}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-\bar{\hat{y}}\right)^2}}$$

と計算できた．

先週の復習（5）

実測値 y_i の変動は次のように分解される．

$$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i-\bar{\hat{y}}\right)^2 + \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$$

全変動（S_T）　回帰による変動（S_R）　回帰からの残差変動（S_e）

先週の復習 (6)

$$R^2 = \frac{S_R}{S_T} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

を考える. これによって, 全体の変動のうち回帰によって説明される部分の割合を示している. この R^2を決定係数, あるいは寄与率という.

先週の復習 (7)

例えば血圧, 年齢, 年収の相関行列を得たとき, 年齢の要因を除いたあとの血圧と年収の相関係数を, 血圧と年収の偏相関係数と呼び,

$$r_{\text{血圧年収}\cdot\text{年齢}} = \frac{r_{\text{血圧年収}} - r_{\text{血圧年齢}} r_{\text{年齢年収}}}{\sqrt{\left(1 - r_{\text{血圧年齢}}^2\right)\left(1 - r_{\text{年齢年収}}^2\right)}}$$

とする. 他の偏相関も同様に,

先週の復習 (8)

他も同様に,

$$r_{\text{年齢年収}\cdot\text{血圧}} = \frac{r_{\text{年齢年収}} - r_{\text{血圧年齢}} r_{\text{血圧年収}}}{\sqrt{\left(1 - r_{\text{血圧年齢}}^2\right)\left(1 - r_{\text{血圧年収}}^2\right)}}$$

$$r_{\text{年齢年収}\cdot\text{血圧}} = \frac{r_{\text{年齢年収}} - r_{\text{血圧年齢}} r_{\text{血圧年収}}}{\sqrt{\left(1 - r_{\text{血圧年齢}}^2\right)\left(1 - r_{\text{血圧年収}}^2\right)}}$$

今日の概要

- 平均値を比べてみる.
- その前に, 分散を考える.
- 対応のない場合, ある場合.
- n 個の平均値は一致するか？

平均値の違いを考える (1)

標本間に対応がない場合の平均値の差の検定

- M市のS大学学生食堂とI市のS大学医学部学生食堂のメニューの価格を調べて以下の表を得た. 平均的に見るとどちらの食堂が安いか. (単位は円)

	S大学	S大医学部
うどん	190	250
定食 (1)	290	350
親子丼	250	350
そば	190	メニューにない
定食 (2)	380	450

平均値の違いを考える (2)

- 平均値 $\bar{x}$ は, それぞれ

$$\bar{x}_{\text{S大}} = 260$$
$$s_{\text{S大}} = 70.99$$
$$\bar{x}_{\text{S大医}} = 350$$
$$s_{\text{S大医}} = 70.71$$

	S大学	S大医
うどん	190	250
定食 (1)	290	350
親子丼	250	350
そば	190	なし
定食 (2)	380	450

明らかにS大学の方が安いそうだ.

しかし, 今までの議論の中から, どれくらいな確率で平均値は等しいか, という数値がほしくないか.

平均値の違いを考える (3)

- 今回は「なぜ，こんな方法が導入されるのか」を避け，「この方法で行う」ということを詳述しよう．まず，

S大学，S大学医データの散らばり具合を考える．ここで新たな関数を，

$$U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

とおいて，このU^2を不偏分散と呼ぼう．

平均値の違いを考える (4)

- 2つの大学の不偏分散を計算すれば，

$$U^2_{s大} = \frac{1}{5-1}\{(190-260)^2 + \cdots + (380-260)^2\} = 6300$$

$$U^2_{s大医} = \frac{1}{4-1}\{(250-350)^2 + \cdots + (450-350)^2\} = 6667$$

この2つの不偏分散の比をとれば，

$$F_0 = \frac{U^2_{S大医}}{U^2_{S大}} = \frac{6667}{6300} = 1.058$$

となる．この F_0 は自由度4-1，5-1の F 分布に従うことが知られている．

平均値の違いを考える (5)

- F 分布って何だ．

・・・とりあえず，データの散らばりを具体的な確率で示すときに使う分布で，2つの自由度と統計量 F からなる確率密度と覚えてください．

――今ではパソコンですぐ計算できる．

平均値の違いを考える (6)

- では，先ほどの $F_0 = 1.058$ を見てみる．ここで，自由度3,4の確率密度0.05に対する F 分布からは5.41という数値が得られた．

この5.41は1.058より大きい．このような時，$S_{大}$と$S_{大医}$との分散は有意水準0.05 (あるいは危険率5%) を越えていないので，分散が等しいと見ていいことを示している．

我々が使う統計の世界ではイコールである確率が5％以下のとき，(本当は正しくないけど) 有意に差がある，といっている．

平均値の違いを考える (7)

- 以上より，$S_{大}$と$S_{大医}$では，データの分散は (とりあえず) 等しいとなった．

この状況を得たとき，平均値の差をある確率で求めるためには次のような式を用いる．

平均値の違いを考える (8)

- データ数 n_1, n_2 の標本の平均値を $\bar{x}_1, \bar{x}_2$ ，不偏分散を U_1^2, U_2^2 とすれば，検定統計量は，

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)U_1^2 + (n_2-1)U_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

この t は自由度 n_1+n_2-2 の t 分布に従う．

平均値の違いを考える (9)

- t 分布って何だ.

  ・・・とりあえず, 2つの標本間の平均値の差の検定を行うときに使う分布で, 平均値の差と分散とデータ数の関数になっている, と覚えてください

  ―――今ではパソコンですぐ計算できる.

平均値の違いを考える (10)

- 今までのデータを使って計算してみる.

  食堂データから,

$$F_0 = \frac{U^2_{S\text{大医}}}{U^2_{S\text{大}}} = \frac{6667}{6300} = 1.058$$

  ここで, 自由度(3,4)だったので, $P(F \geq F_0) = 0.05$ の F 分布表から自由度(3,4)をさがせば, 6.59で $F_0 = 1.058$ より大きい. これは, 2標本間に分散の差がないことを示している.

平均値の違いを考える (11)

- 今までのデータを使って計算してみる.

  両大学食堂データから,

$$t_0 = \frac{350 - 260}{\sqrt{\frac{(4-1)\times 6667 + (5-1)\times 6300}{4+5-2}\left(\frac{1}{4}+\frac{1}{5}\right)}} = \frac{90}{53.90} = 1.670$$

  $P(t \geq t_0) = 0.05$ を自由度7の t 分布表からみれば2.365で, t_0 より大きい. 従って, 平均値に違いはなさそうである (本当は比較する品目を例えば麺類, 丼物など統一すべきであろう).

平均値の違いを考える (12)

- もし, $F < F_0$ つまり分散に違いがあった場合, どうする.

  結論から言えば, Welchの方法を用いる. その方法は少しややこしいが, 2組の標本をとり, データ数が n_1, n_2, 平均値が $\bar{x}_1, \bar{x}_2$, 不偏分散が U_1^2, U_2^2 であるとき, 検定統計量は,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{U_1^2}{n_1} + \frac{U_2^2}{n_2}}}$$

平均値の違いを考える (13)

- 自由度 ν は,

$$\frac{1}{\nu} = \frac{p^2}{n_1 - 1} + \frac{q^2}{n_2 - 1}$$

  ここで,

$$p = \frac{U_1^2/n_1}{U_1^2/n_1 + U_2^2/n_2}$$

$$q = 1 - p$$

  Welchの方法は近似的な解法であるが, 分散が異なった場合, これを用いている.

平均値の違いを考える (14-1)

A群	100	3000	15	400	155	5825
B群	200	60	100	150	220	

$\bar{x}_A = 1582.5, \quad \bar{x}_B = 146$

$s_A = 2164, \quad s_B = 59.87$

$U_A^2 = 5619808, \quad U_B^2 = 3584$

$$F_0 = \frac{U_A^2}{U_B^2} = \frac{5619808}{3584} = 1568$$

$F_4^5(0.05) = 6.26 < F_0$

なので, 分散が異なっている. そこで, Welchの方法.

平均値の違いを考える (14-2)

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{U_1^2}{n_1} + \frac{U_2^2}{n_2}}} = \frac{1582.5 - 146}{\sqrt{\frac{5619808}{6} + \frac{3584}{5}}} = 1.484$$

$$\frac{1}{\nu} = \frac{p^2}{n_1 - 1} + \frac{q^2}{n_2 - 1} = \frac{1}{0.1997}$$

よって，

$$\nu = 5$$

$$t_5(0.05) = 2.571 > 1.484$$

A群とB群では平均値に有意な差は認められなかった．

平均値の違いを考える (14-3)

平均値にすごく差があるような2群でもこの検定は，平均値の差だけの関数ではなく，分散の関数にもなっているので，このようなこともおこる．

平均値の違いを考える (15)

標本間に対応がある場合の平均値の差の検定

同級生5名がSオールスターズのコンサート前後に血圧を測定した．コンサートは最高血圧上昇の要因になったか．

	A	B	C	D	E
見る前	112	93	107	121	104
見た後	125	94	114	136	130

平均値を計算すると，

$$\bar{x}_{\text{見る前}} = 107.4$$

$$\bar{x}_{\text{見た後}} = 121.0$$

平均値の違いを考える (16)

	1	2	3	…	n
状態A	x_{1A}	x_{2A}	x_{3A}	…	x_{nA}
状態B	x_{1B}	x_{2B}	x_{3B}	…	x_{nB}
差	$d_1 = x_{1A} - x_{1B}$	d_2	d_3	…	d_n

この検定は次のように行う．$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$，差の不偏分散 U_d^2 をとおけば，検定統計量

$$t = \frac{\bar{d}\sqrt{n}}{U_d}$$

は自由度 $n-1$ の t 分布に従う．

Sオールスターズの血圧データからは・・・，

平均値の違いを考える (17)

	A	B	C	D	E
見る前	112	93	107	121	104
見た後	125	94	114	136	130
差	13	1	7	15	26

$$\bar{d} = \frac{1}{5}(13 + \cdots + 26) = \frac{62}{5} = 12.4$$

$$U_d = \sqrt{U_d^2} = \sqrt{\frac{1}{5-1}\left\{(13-12.4)^2 + \cdots + (26-12.4)^2\right\}} = 9.370$$

$$t_0 = \frac{\bar{d}\sqrt{n}}{U_d} = \frac{12.4 \times \sqrt{5}}{9.370} = 2.964$$

平均値の違いを考える (18)

$$t_0 = \frac{\bar{d}\sqrt{n}}{U_d} = \frac{12.4 \times \sqrt{5}}{9.370} = 2.964$$

自由度は5-1だから，

$$t_4(0.05) = 2.776 < 2.964$$

Sオールスターズのコンサートでは，5名は平均的に興奮していたのだろう*)．

統計理論から本当は，「コンサートの前後の血圧はイコールである確率が5％以下であった」といい，どちらかの大小を言うわけではない．何となく差が気になるので，*)のようなことをいってしまう．

平均値の違いを考える (19)

シンプルなものだが，多重比較を考える．

次のようなデータがある．このとき，要因による変化はあるだろうか．

毎日ほぼ一定の乳量を出す牛6頭に対して，補助飼料A，B，Cを与えた．補助飼料によって1週間の合計乳量に変化があるだろうか．

⬇

平均値の違いを考える (20)

	乳牛1	乳牛2	乳牛3	乳牛4	乳牛5	乳牛6
飼料A	272	189	298	257	197	181
飼料B	300	290	256	298	285	288
飼料C	287	174	230	299	未記入	153

基礎統計量は，・・・

$$\bar{x}_A = 232.2, \quad s_A = 45.2$$
$$\bar{x}_B = 286.2, \quad s_B = 14.5$$
$$\bar{x}_c = 228.6, \quad s_c = 58.4$$

あとは，このデータから今まで行った平均値の差の検定を飼料A，B，C総あたりで行えばいい．しかし，————

平均値の違いを考える (21)

我々は今，$\bar{x}_A = \bar{x}_B = \bar{x}_C$ということを見てみたかったはずでは？ 2つの標本をt検定したとき，有意水準を0.05と考えれば，1回の検定結果を受け入れる確率は，0.95．2回続けば0.95^2=0.9025，3回だと0.95^3=0.8145.

従って，検定結果のうち少なくとも1回が有意になる確率は

1－0.8145=0.1955

この数字は最初に設定した0.05より大きくないか．

⬇

平均値の違いを考える (22)

一般化すれば，0.05の有意水準で n 群の平均値を，

${}_nC_2$ 回検定すれば，そのときの有意水準は，

$$1-(1-0.05)^n$$

となって，有意になりやすい．そこで，多重比較を行うのである．

$n=15$だと，有意水準は0.5367

平均値の違いを考える (23)

重回帰分析のところで行った

$$S_T = S_R + S_e$$

を思い出してほしい．

この概念を多重比較にも用いてみる．

⬇

先ほどの牛データを一般化すれば，——

平均値の違いを考える (24)

	データ	群平均
群1	$x_{11} \quad x_{12} \quad \cdots \quad x_{1n}$	$\bar{X}_1 = (1/n)\sum x_i$
群2	$x_{21} \quad x_{22} \quad \cdots \quad x_{2n}$	$\bar{X}_2 = (1/n)\sum x_i$
⋮	⋮	⋮
群m	$x_{m1} \quad x_{m2} \quad \cdots \quad x_{mn}$	$\bar{X}_m = (1/n)\sum x_i$

基礎統計量・・・

$$N = mn \qquad X = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} \qquad \bar{X} = \frac{1}{mn} X$$

以上の準備から，

平均値の違いを考える (25)

$$S_T = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \overline{X}\right)^2$$

$$S_B = n\sum_{i=1}^{m}\left(\overline{X}_i - \overline{X}\right)^2$$

$$S_W = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \overline{X}_i\right)^2$$

このような統計量を出したとき，

$$S_T = S_B + S_W$$

となる．本当は各自で証明してほしいが．

平均値の違いを考える (26)

$$S_T = S_B + S_W$$

全変動 ＝ 級間変動 ＋ 級内変動

不偏分散の自由度は，

$$\nu_T = mn - 1 \qquad \nu_B = m - 1 \qquad \nu_W = mn - m$$

$$\nu_T = \nu_B + \nu_W$$

である．以上の準備から，次の表を得る．

平均値の違いを考える (27)

要因	偏差平方和	自由度	不偏分散	F 値
級間	S_B	$m-1$	$U_B^2 = S_B/(m-1)$	$F = U_B^2/U_W^2$
級内	S_W	$mn-m$	$U_W^2 = S_W/(mn-m)$	
全体	S_T	$mn-1$		

これより，例えば，$F_{mn-m}^{m-1}(0.05)$ を確認して，

$$\overline{x}_1 = \overline{x}_2 = \cdots = \overline{x}_m$$

の仮説を検定する．

平均値の違いを考える (28)

では，牛データを解析してみよう．

	乳牛1	乳牛2	乳牛3	乳牛4	乳牛5	乳牛6	平均
飼料A	272	189	298	257	197	181	232
飼料B	300	290	256	298	285	288	286
飼料C	287	174	230	299	未記入	153	229

総計 $= 272 + 189 + \cdots + 153 = 4254$

全体の平均 $= \frac{1}{17} \times 4254 = 250.2$

平均値の違いを考える (29)

S_T を計算すれば

	乳牛1	乳牛2	乳牛3	乳牛4	乳牛5	乳牛6	合計
飼料A	$(272-250.2)^2 = 473.7$	3749.8	2281.5	45.8	2834.0	4793.5	14178.2
飼料B	$(300-250.2)^2 = 2476.5$	1581.2	33.2	2281.5	1208.6	1426.2	9007.2
飼料C	$(287-250.2)^2 = 1351.6$	5811.8	409.5	2378.0		9454.7	19405.6

$$S_T = 473.7 + 3749.8 + \cdots + 9454.7 = 42591.1$$

平均値の違いを考える (30)

S_B を計算すれば

	乳牛1	乳牛2	乳牛3	乳牛4	乳牛5	乳牛6	合計
飼料A	$(232.2-250)^2 = 320.5$	320.5	320.5	320.5	320.5	320.5	1922.9
飼料B	$(286.2-250)^2 = 1291.1$	1291.1	1291.1	1291.1	1291.1	1291.1	7746.4
飼料C	$(228.6-250)^2 = 468.1$	468.1	468.1	468.1		468.1	2340.4

$$S_B = 320.5 + 320.5 + \cdots + 468.1 = 6\times320.5 + 6\times1291.1 + 5\times468.1 = 12009.7$$

平均値の違いを考える (31)

S_W を計算すれば

	乳牛1	乳牛2	乳牛3	乳牛4	乳牛5	乳牛6	合計
飼料A	$(272-232.3)^2$ $=1573.4$	1877.8	4312.1	608.4	1248.4	2635.1	12255.3
飼料B	$(300-286.2)^2$ $=191.4$	14.7	910.0	140.0	1.4	3.4	1260.8
飼料C	$(287-228.6)^2$ $=3410.6$	2981.2	2.0	4956.2		5715.4	17065.2

$$S_W = 1573.4 + 1877.8 + \cdots + 5715.4 = 30581.4$$

平均値の違いを考える (32)

これらより以下の表を得る．

要因	偏差平方和	自由度	不偏分散	F 値
級間	12009.7	2	6004.9	2.7490
級内	30581.4	14	2184.4	
全体	42591.1	16		

$$F_{14}^2(0.05) = 3.74 > 2.7490$$

飼料Bの方が乳量をふやす効果がありそうに感じたが，
飼料によって乳量には変化がないようである．

平均値の違いを考える (33)

級間の差を検定する方法はないか？

・・・ある．しかも複数種．ここではシェフェ (Scheffé)
の方法を紹介しよう．他の方法については最後のス
ライドに挙げた参考書を見てください．しかも必ず．

平均値の違いを考える (34)

シェフェの方法

$$Q = \sqrt{(m-1)F_{mn-m}^{m-1}(0.05)U_B^2}$$

とおいて，i番目の群と $i+1$ 番目の群の平均値を比較するため，

$$Q\sqrt{\frac{1}{n_i}+\frac{1}{n_{i+1}}} \quad \text{と} \quad \left|\overline{x}_i - \overline{x}_{i+1}\right|$$

の大小と関係を比べる．これをすべての組み合わせで行う．

平均値の違いを考える (35)

シェフェの方法

牛データでは，平均値の差は

	飼料A	飼料B	飼料C
飼料A			
飼料B	$\lvert 232.2-286.2\rvert$ $=54$		
飼料C	$\lvert 232.2-228.6\rvert$ $=3.6$	$\lvert 286.2-228.6\rvert$ $=57.6$	

平均値の違いを考える (36)

シェフェの方法

$F_{14}^2(0.05) = 3.74$ から，

$$Q = \sqrt{(m-1)F_{mn-m}^{m-1}(0.05)U_B^2}$$
$$= \sqrt{(3-1)\times 3.74 \times 6004.9} = 211.93$$

よって，

$$Q\sqrt{\frac{1}{n_i}+\frac{1}{n_{i+1}}}$$

は次の表よって示される．

平均値の違いを考える (37)

シェフェの方法

	飼料A	飼料B	飼料C
飼料A			
飼料B	122.36		
飼料C	128.34	128.34	

平均値の差	飼料A	飼料B	飼料C
飼料A			
飼料B	54		
飼料C	3.6	57.6	

平均値の違いを考える (38)

シェフェの方法

これらの計算より，

「飼料を変えても乳量の増減は (有意水準 0.05で) なかった」.

今日話していないこと

- 平均値の差の信頼区間.
- 平均値の差の検定の数理構造.
- 多重比較におけるラテン方格，乱塊法など2元配置，3元配置．それらの交絡性.
- 多重比較における他の平均値の差の検定.
  Tuker法，Dunnet法，Limit Significant Difference 法など.

参 考 文 献

- 『やさしい統計学』. 田畑吉雄. 現代数学社.
  ——— 勉強しようかと思ったが，数理構造などを少し，あるいはほとんど忘れたことに気づいたとき，どうぞ.
- 『分散分析のはなし』. 石村貞夫. 東京図書.
  ——— 具体的なデータをたくさん使って分かりやすく書いてある.
- 『心理学のためのデータ解析テクニカルブック』.
  森敏昭，吉田寿夫編著. 北大路書房.
  ——— 具体的な数理構造を記述してある.「読んでやるぞ」と少し構えてどうぞ. 先ず上記2冊を読んでからの方がいいかも.