実用統計

多変量解析編

福井正康

１章　実験計画法

1.1　１元比較実験計画法

多群間の平均や中間値に差があるかどうか検討する手法であり、いくつの変数を同時に比較するかによって１元比較、２元比較などと分かれている。ここでは理解し易い１元比較についてのみ解説する。

多群間の等分散の検定には Bartlett 検定を利用する。

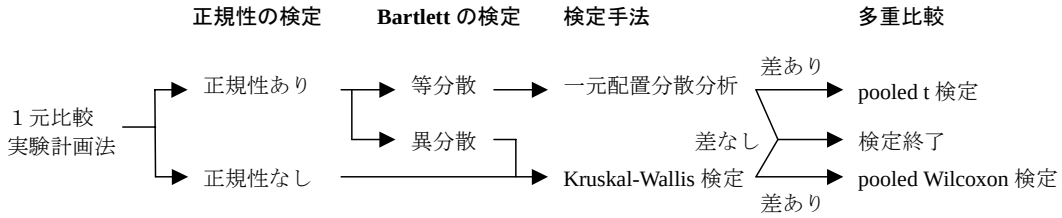


図 1.1　１元比較実験計画法の構造

1.2　１元配置分散分析

例

３つの条件である商品の売上を調査したところ、以下の結果を得た。（Samples¥分散分析 ex.txt）これらの分布が正規分布で条件間で等分散であることを仮定して、条件間に差があるといえるか、有意水準 5%で判定せよ。

条件１　115，110，108，114，120，116，108，112，115，122

条件２　121，118，124，117，119，130，121，115，118，119

条件３　116，112，120，111，112，108，114，119，104，113

解答

１元配置分散分析結果

分類名　条件１
データ数　10
平均値　114.0000
不偏分散　22.0000
：
水準間
平方和　309.8000
自由度　2
不偏分散　154.9000

水準内
平方和　566.5000
自由度　27
不偏分散　20.9815

Ｆ統計値　7.38270
片側確率Ｐ　0.00277
有意水準α　0.05
Ｐ＜αより、群別の平均間に差があるといえる。

表 1.1　1 元配置分散分析結果

	平方和	自由度	不偏分散	F 値
全変動	8.7630E+02	29		7.3827
水準間	3.0980E+02	2	1.5490E+02	P 値
水準内	5.6650E+02	27	2.0981E+01	0.0028

理論

水準間に差があるかどうか、有意水準αで検定する。

水準 1	水準 2	…	水準 k
x_{11}	x_{21}	…	x_{k1}
x_{12}	x_{22}	…	x_{k2}
:	:	:	:
x_{1n_1}	x_{2n_2}	…	x_{kn_k}

$x_{i\lambda}$は水準 i に固有な値μ_iと誤差$\varepsilon_{i\lambda}$とからなると仮定する。

$$x_{i\lambda} = \mu_i + \varepsilon_{i\lambda} \qquad \varepsilon_{i\lambda} \sim N(0,\sigma^2) \text{分布}$$

全変動は以下のように分解される。

$$S = \sum_{i=1}^{k}\sum_{\lambda=1}^{n_i}(x_{i\lambda} - \bar{x})^2 = \sum_{i=1}^{k}\sum_{\lambda=1}^{n_i}(x_{i\lambda} - \bar{x}_i)^2 + \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 = S_E + S_P$$

全変動　　水準内変動　　水準間変動

$S/\sigma^2 \sim \chi^2_{N-1}$分布，　$S_E/\sigma^2 \sim \chi^2_{N-k}$分布，　$S_P/\sigma^2 \sim \chi^2_{k-1}$分布，ここに$N = \sum_{i=1}^{k} n_i$

帰無仮説 H_0：$\mu_1 = \mu_2 = \cdots = \mu_k$　（水準間に差がない）

対立仮説 H_1：H_0でない

帰無仮説のもとで

$$F = \frac{S_P/(k-1)}{S_E/(N-k)} \sim F_{k-1,N-k} \text{分布}$$

$F = F_{k-1,N-k}(p)$として、 $p < \alpha$ ならば、水準間に差があると判定する。

1.3 Kruskal-Wallis 検定

例

３つの条件である商品の売上を調査したところ、以下の結果を得た（1.2 節参照)。分布が正規分布に従わないとして、これらの条件間に差があるかどうか有意水準 5%で判定せよ。

解答

Kruskal-Wallis 検定結果

分類名	条件 1	データ数	10
データ数	10	順位和	110.5
順位和	127.5	χ2 検定値	10.2200
分類名	条件 2	自由度	2
データ数	10	片側確率 P	0.00604
順位和	227	有意水準 α	0.05
分類名	条件 3		

P＜α より、群間の位置母数に差があるといえる。

理論

k 種類の水準の中間値に差があるかどうか、有意水準 $\alpha \times 100\%$ で判定する。

全データの小さい順に順位を付ける。

水準 1	水準 2	…	水準 k
r_{11}	R_{21}	…	r_{k1}
r_{12}	R_{22}	…	r_{k2}
:	:	…	:
r_{1n_1}	r_{2n_2}	…	r_{kn_k}
w_1	w_2	…	w_k

水準毎のデータ数 n_i， $N = \sum_{i=1}^{k} n_i$， 水準毎の合計 w_i

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\frac{w_i}{n_i} - \frac{N+1}{2} \right)^2 \sim \chi^2_{k-1} \text{分布}$$

$\chi^2 = \chi^2_{k-1}(p)$ として、 $p < \alpha$ ならば、水準間に差があると判定する。

1.4 等分散性の検定（Bartlett 検定）

例

３つの条件である商品の売上を調査したところ、1.2 節の結果を得た。各条件の分散間に差があるといえるか。分布が正規分布するものとして、有意水準 5%で判定せよ。

解答

Bartlett 検定結果

分類名	条件 1	分類名	条件 3
データ数	10	データ数	10
分散	22.0000	分散	22.9889
分類名	条件 2	χ2 値	0.15366
データ数	10	自由度	2
分散	17.9556	片側確率 P	0.92605
		有意水準 α	0.05

P≧α より、群間の分散に差があるといえない。

理論

帰無仮説 H_0： $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$

対立仮説 H_1：H_0 でない

$$V_E = \frac{S_E}{n-k} = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{\lambda=1}^{n_i}(x_{i\lambda} - \bar{x}_i)^2 , \qquad S_i^2 = \frac{1}{n_i - 1}\sum_{\lambda=1}^{n_i}(x_{i\lambda} - \bar{x}_i)^2$$

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_{j=1}^{k}\frac{1}{n_j - 1} - \frac{1}{n-k}\right] \quad \text{とすると、}$$

$$\chi^2 = \frac{1}{C}\left[(N-k)\log V_E - \sum_{i=1}^{k}(r_i - 1)\log S_i^2\right] \sim \chi_{k-1}^2 \text{分布}$$

$\chi^2 = \chi_{k-1}^2(p)$ として、 $p < \alpha$ ならば、水準間に差があると判定する。

問題１

Samples¥分散分析 1.txt は３つの工場群の不良品率を与えたものである。各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］

検定名［　　］　検定確率［　　］

判定　工場群間の不良品率に差があると［いえる・いえない］

問題２

Samples¥分散分析 2.txt は４つの群のデータであるが、各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］

検定名［　　］　検定確率［　　］

判定　群間に差があると［いえる・いえない］

1.5 多重比較

n 種の水準間の比較

$$ {}_nC_2 = \frac{n(n-1)}{2} \qquad n=5 \rightarrow {}_5C_2 = 10, \quad n=10 \rightarrow {}_{10}C_2 = 45 $$

有意水準5%として、45回も比較したら偶然だけで有意な結果が出る場合もある。

1.5.1 正規性・等分散性のある場合の多重比較

例

３つの条件である商品の売上を調査したところ、1.2節の結果を得た。これらの分布が等分散の正規分布であるとして、分散分析によって条件間に差があると判定された。ではどの条件間に差が見られるのだろうか、有意水準5%で判定せよ。

解答

検定手順　Fisher の LSD 法に従う。

	条件 1	条件 2	条件 3
データ数	10	10	10
平均	114.0000	120.2000	112.9000
不偏分散	2.2000E+01	1.7956E+01	2.2989E+01
Pooled 不偏分散	2.0981E+01		
自由度	27		
確率(両側)			
条件 1	1.00000	0.00538	0.59568
条件 2	0.00538	1.00000	0.00139
条件 3	0.59568	0.00139	1.00000

以上より、水準１と２、水準２と３の間に差があるといえる。

注）Fisher の LSD 法（least significant difference procedure）

1元配置分散分析または Kruskal-Wallis 検定を行ない差がない場合は､終了する。差がある場合のみ、pooled 推定値を用いた t 検定または、結合順位を用いた Wilcoxon の順位和検定を行なう。

注）質的指標（母比率の検定）について

χ^2 検定＋pooled 推定値を用いた比率の検定　　（省略）

理論（pooled 推定値を用いた t 検定）

k 種類の水準を考え、各水準の平均の間に差があるか有意水準 $\alpha \times 100\%$ で判定する。水準 i のデータ数を n_i、平均を $\bar{x}_i$、不偏分散を u_i^2 として、水準 i, j について考える。

$$ N = n_1 + n_2 + \cdots + n_k $$

$$u^2 = \frac{(n_1-1)u_1^2 + (n_2-1)u_2^2 + \cdots + (n_k-1)u_k^2}{N-k}$$ pooled 不偏分散

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{u\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}} \sim t_{N-k} \text{ 分布}$$ （t 検定統計量の不偏分散についての拡張）

$t_{ij} = t_{N-k}(p/2)$ として $p<\alpha$ ならば、水準間に差があると判定する。

1.5.2 正規性のない場合の多重比較

例

３つの条件である商品の売上を調査したところ、1.2 節の結果を得た。これらの分布は正規分布でないとして、Kruskal-Wallis 検定によって条件間に差があると判定された。ではどの条件間に差が見られるのだろうか、有意水準 5%で判定せよ。

解答

検定手順　Fisher の LSD 法に従う。

表　pooled Wilcoxon 結果

	条件 1	条件 2	条件 3
データ数	10	10	10
順位和	127.500	227.000	110.500
確率(両側)			
条件 1	1.00000	0.01235	0.68445
条件 2	0.01235	1.00000	0.00335
条件 3	0.68445	0.00335	1.00000

条件 1 と条件 2、条件 2 と条件 3 の間に差が見られる。

理論（結合順位による Wilcoxon の順位和検定）

k 種類の水準のどの中間値に差があるか、有意水準 $\alpha\times100\%$ で判定する。

全データの小さい順に順位を付ける。

水準 1	水準 2	…	水準 k
r_{11}	r_{21}	…	r_{k1}
r_{12}	r_{22}	…	r_{k2}
:	:	…	:
r_{1n_1}	r_{2n_2}	…	r_{kn_k}
w_1	w_2	…	w_k

水準毎のデータ数　n_i，$N = \sum_{i=1}^{k} n_i$，水準毎の合計　w_i　データ数は十分多いとする。

$$Z_{ij} = \frac{\left| \frac{w_i}{n_i} - \frac{w_j}{n_j} \right| - \frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j} \right)}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \sim N(0,1) \text{ 分布}$$

$Z_{ij} = Z(p/2)$ として、 $p < \alpha$ ならば、水準間に差があると判定する。

問題１

Samples¥分散分析 1.txt は３つの工場群の不良品率を与えたものである。各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］

検定名［　　］　検定確率［　　］

判定　工場群間の不良品率に差があると［いえる・いえない］

差があるとするとどの条件間に差があるか。差がある条件同士を工場２＜工場３（実際の結果とは関係ない）のように不等号で表せ。

検定名［　　］　結果［　　］

問題２

Samples¥分散分析 2.txt は４つの群のデータであるが、各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］

検定名［　　］　検定確率［　　］

判定　群間に差があると［いえる・いえない］

差があるとするとどの群間に差があるか。差がある群同士を群２＜群３等のように不等号で表せ。

検定名［　　］　結果［　　］

問題３

Samples¥分散分析 3.txt は３群のデータであるが、各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］

検定名［　　］　検定確率［　　］

判定　群間に差があると［いえる・いえない］

差があるとするとどの群間に差があるか。差がある群同士を群２＜群３等のように不等号で表せ。

検定名［　　］　結果［　　］

1.6 その他の関連する分析

1.6.1 対応がある場合の１元配置問題

例

３つの条件である商品の売上を調査したところ、1.2 節の結果を得た。各データに対応があるとして差があるか検定せよ。（再掲）

条件１　115，110，108，114，120，116，108，112，115，122
条件２　121，118，124，117，119，130，121，115，118，119
条件３　116，112，120，111，112，108，114，119，104，113

理論

正規性の有無により、（繰り返しのない）２元配置分散分析か Friedman 検定を利用する。これは repeated measured　1 元配置分散分析、repeated measured Kruskal-Wallis 検定とも呼ばれている。

解答

（繰り返しのない）２次元配置分散分析を用いる。

水準（列）間　（この部分を見る）
平方和　309.8000
自由度　2
不偏分散　154.9000
Ｆ統計値　6.22088
自由度　2,18
片側確率Ｐ　0.00884
有意水準α　0.05
Ｐ＜α より、水準間の平均に差があるといえる。

1.6.2 トレンド（傾向性）の検定

例　量的データ

成績順に並べた４つの群で，ある指標の点数を観察したら、以下の結果が得られた（対応はないものとする。Samples¥トレンド 1.txt)。これらのデータの母平均（中央値）μ_i に増加または減少の傾向（$\mu_1 \geq \mu_2 \geq \mu_3 \geq \mu_4$ またはその逆）があるといえるか。有意水準 5%で判定せよ。

群１　8.06,　8.27,　8.45,　8.51,　8.14
群２　7.97,　7.66,　8.05,　8.30,　8.03
群３　7.66,　7.71,　7.88,　8.05,　7.80
群４　8.00,　7.89,　7.79,　7.91,　7.40

解答

通常 Jonckheere 検定を用いる。

Ｊ統計値　　48.000（140.000）
ｚ統計値　　2.95554
両側確率Ｐ　0.00312
有意水準α　0.05
Ｐ＜αより、トレンドがあるといえる。

例　質的データ

成績順に並べた 4 つの群（各 10 名ずつ）で、ある事柄に対する興味を調べたら、以下の結果を得た。4 つの群の興味の比率に傾向があるといえるか。有意水準 5%で判定せよ。（Samples¥トレンド 2.txt のデータの 3 つのデータ形式で検定せよ。）

	群１	群２	群３	群４
興味あり	3	4	7	8
興味なし	7	6	3	2

解答

Mantel-extension 法を用いる。

２章 重回帰分析

例

以下のデータ（Samples¥重回帰分析 1.txt）をもとに体重を身長と胸囲の１次関数で予測する。

体重	身長	胸囲	体重	身長	胸囲
61.0	167.0	84.0	49.5	164.7	78.0
55.5	167.5	87.0	61.0	171.0	90.0
57.0	168.4	86.0	59.5	162.6	88.0
57.0	172.0	85.0	58.4	164.8	87.0
50.0	155.3	82.0	53.5	163.3	82.0
50.0	151.4	87.0	54.0	167.6	84.0
66.5	163.0	92.0	60.0	169.2	86.0
65.0	174.0	94.0	58.8	168.0	83.0
60.5	168.0	88.0	54.0	167.4	85.2
49.5	160.4	84.9	56.0	172.0	82.0

解説

体重 ＝ b_1 身長＋b_2 胸囲＋b_0　　の形で体重を予測する。

目的変数：体重　　説明変数：身長，胸囲

係数の値は？　　→　偏回帰係数

説明変数の重要性は？　　→　標準化偏回帰係数

どの程度予測できるか？　　→　重相関係数，寄与率（決定係数）

このモデルは有効か？　　→　Ｆ検定値と確率（要残差正規性）

それぞれの係数は有効か？　　→　ｔ検定値と確率（要残差正規性）

他の変数の影響を除いた目的変数と各説明変数の相関は？　　→　偏相関係数

どの程度予測できているのか図的に見たい　　→　散布図

どの程度予測できているのかデータ毎に見たい　　→　予測値と残差

解答

重回帰分析結果

目的変数	体重
説明変数	身長, 胸囲
データ数	20
回帰式	体重 = 0.3861*身長+0.8575*胸囲-80.7427
寄与率	0.70652
重相関係数	0.84055
自由度調整済み	0.81975
Ｆ検定値	20.46324
自由度	2 , 17
確率値	0.00003

表　重回帰分析結果

	偏回帰係数	標準化係数	t 検定値	自由度	確率値	偏相関係数
身長	0.3861	0.4333	3.23345	17	0.00488	0.61710
胸囲	0.8575	0.6401	4.77676	17	0.00018	0.75700
切片	-80.7427	0.0000	-3.57609	17	0.00233	

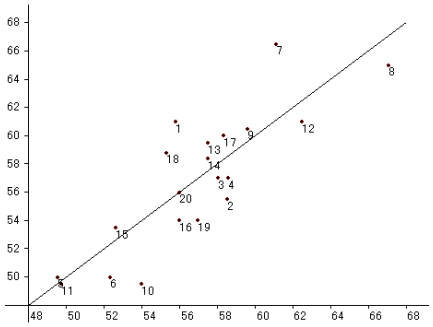


図　実測／予測値の散布図

表　予測値と残差

	実測値	予測値	残差
1	61.0	55.762	5.238
2	55.5	58.528	-3.028
3	57.0	58.018	-1.018
4	57.0	58.550	-1.550
5	50.0	49.530	0.470
6	50.0	52.312	-2.312
7	66.5	61.078	5.422
8	65.0	67.040	-2.040
9	60.5	59.579	0.921

まとめ

目的変数を体重に、説明変数を身長と胸囲にして、重回帰分析を行ったところ、以下の回帰式を得た。

体重 = 0.3861*身長+0.8575*胸囲-80.7427

予測体重と実測体重の相関である重相関係数は 0.84055 で、回帰式の寄与率は 0.70652 となった。これから体重変動の約 71%が説明できることが分かる。　各変数の予測における重要性を示す標準化偏回帰係数は、身長が 0.4333、胸囲が 0.6401 と胸囲が少し上回っている。

回帰式の妥当性の検定を行ったところ p=0.00003 となり、妥当性が有意に示された。また、各偏回帰係数が 0 と異なることを示す検定では、身長が p=0.00488、胸囲が p=0.00018、切片は p=0.00233 となり、各係数とも有意に 0 と異なっている。

以上のことからこの回帰式は予測モデルとして、かなり良いモデルになっている。

理論

標本番号	目的変数	説明変数 1	…	説明変数 p
1	y_1	x_{11}	…	x_{k1}
2	y_2	x_{12}	…	x_{k2}
:		:	…	:
n	y_n	x_{1n}	…	x_{kn}

目的

目的変数を最もよく説明する説明変数の線形モデルを与える。

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

偏回帰係数

目的変数のゆらぎ D を最も良く説明する偏回帰係数 b_0, b_i を求める。

$$Y_\lambda = b_0 + b_1 x_{1\lambda} + b_2 x_{2\lambda} + \cdots + b_k x_{k\lambda}$$

$$D = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 \qquad \text{最小化}$$

標準化偏回帰係数

$$y_\lambda * = \frac{y_\lambda - \bar{y}}{u_y}, \quad x_{i\lambda} * = \frac{x_{i\lambda} - \bar{x}_i}{u_i} \text{ として、 } y* \text{を説明する回帰式を求める。}$$

$$Y_\lambda * = b_1 * x_{1\lambda} * + b_2 * x_{2\lambda} * + \cdots + b_k * x_{k\lambda} * \qquad b_i * = b_i \frac{u_i}{u_y}$$

寄与率と重相関係数

$$SV = \sum_{\lambda=1}^{n} (y_\lambda - \bar{y})^2 = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 + \sum_{\lambda=1}^{n} (Y_\lambda - \bar{Y})^2 = EV + RV$$

全変動 SV，回帰変動 RV，残差変動 EV

寄与率 $\qquad R^2 = RV/SV$

重相関係数 $\qquad R = \sqrt{RV/SV}$ 観測値と予測値の相関係数でもある。

自由度調整済み重相関係数 $\qquad \overline{R} = \sqrt{1 - \dfrac{EV/(n-k-1)}{SV/(n-1)}}$

回帰式の有効性の検定

$$F = \frac{RV/k}{EV/(n-k-1)} \sim F_{p,n-p-1} \text{分布}$$

偏回帰係数の検定

$b_i = 0$ の検定　　自由度 $n-k-1$ の t 検定

$b_0 = 0$ の検定　　自由度 $n-k-1$ の t 検定

偏相関係数　$r_{iy\cdot 12\cdots i-1i+1\cdots k}$

X_i ：他の説明変数で作った x_i の予測回帰式

Y_i ：他の説明変数で作った y の予測回帰式

$x_i' = x_i - X_i$，　$y' = y - Y_i$ とした場合の、

x_i' と y' の相関係数（他の変数の影響を除いた相関係数）

残差

$z_\lambda = y_\lambda - Y_\lambda$

問題

Samples¥重回帰分析 2.txt はある大学の学生について調べた、卒業試験の成績、入試点数、内申点数、ある５日間の勉強時間、授業への出席率のデータである。卒業試験の成績を他の変数で予測する重回帰分析を行い、結果をまとめにならって記述せよ。

重回帰分析続き

解説

データ Samples¥重回帰分析 1.txt を用いて、体重を身長と胸囲の 1 次関数で予測する。

体重 ＝ b_1 身長＋b_2 胸囲＋b_0　　の形で体重を予測する。

　　目的変数：体重　　説明変数：身長，胸囲

係数の値は？　　→　偏回帰係数

説明変数の重要性は？　　→　標準化偏回帰係数

どの程度予測できるか？　　→　重相関係数，寄与率（決定係数）

このモデルは有効か？　　→　F 検定値と確率（要残差正規性）

それぞれの係数は有効か？　　→　t 検定値と確率（要残差正規性）

どの程度予測できているのか図的に見たい　　→　散布図

どの程度予測できているのかデータ毎に見たい　　→　予測値と残差

問題 1

Samples¥重回帰分析 2.txt について、重回帰分析を行い、以下の問いに答えよ。

1）回帰式を求めよ。

卒業試験＝　［　　］入試点数　＋［　　］内申点数

＋［　　］勉強時間　＋［　　］出席率

＋［　　］

2）この回帰式の寄与率を求めよ。［　　］

3）この場合残差の分布は正規分布といえるか。［正規分布・正規分布でない］

4）回帰式の係数の t 検定（偏回帰係数が 0 と異なるかどうかの検定）の確率値が 0.05 を超えるものの中で最大となる変数（最も不要な変数）を順次削除していくと、最終的に残るものは何か。各段階の検定確率値を記入せよ。但し、削除した変数のところは以後空欄にし、すべての確率が 0.05 未満になった場合は確定とする。

	4 変数	3 変数	2 変数	1 変数
入試点数				
内申点数				
勉強時間				
出席率				

5）最終的な回帰式はどのようになるか。不要な変数の係数欄は空欄のままでよい。

卒業試験＝　［　　］入試点数　＋［　　］内申点数

＋［　　］勉強時間　＋［　　］出席率

＋［　　］

6）上の回帰式の寄与率を求めよ。［　　］

7）上の回帰式の寄与率はすべての変数を使った場合に比べ大きく下がっているか。
［大きく下がっている・あまり下がっていない］

8）この式を新しい予測モデルとして採用するか。
［採用する・採用しない］

9）新しい予測モデルで、データ中の最初（1 番）の学生について卒業試験の実測値，その予測値，残差（実測値と予測値の差）はいくらか。
実測値［　　］　予測値［　　］　残差［　　］

10）上と同様のモデルで、質問項目の値が入試点数 70、内申点数 3.5、勉強時間 5、出席率 70%の学生の卒業試験はいくらに予測されるか。
［　　］

問題２

Samples¥重回帰分析 3.txt について、重回帰分析を行い、以下の問いに答えよ。

1）売上を従業員と資産で推測する回帰式を求めよ。
売上＝　［　　］従業員　＋［　　］資産
　　＋［　　］

2）上の回帰式の寄与率を求めよ。［　　］

3）log 売上を log 従業員と log 資産で推測する回帰式を求めよ。但し、この対数は底が 10 の常用対数である。
log 売上＝　［　　］log 従業員　＋［　　］log 資産
　　＋［　　］

4）上の回帰式の寄与率を求めよ。［　　］

5）$z = cx^a y^b$ の常用対数をとると以下のようになる。

$$\log_{10} z = a\log_{10} x + b\log_{10} y + \log_{10} c$$

ここに、$d = \log_{10} c$ とすると、$c = 10^d$（Excel で計算可能）

これを用いて３）の回帰式を以下の形に書き換えよ。
売上＝［　　］×従業員［　　］×資産［　　］

6）１）の回帰式と３）の回帰式はどちらがより優れていると思われるか。
どちらも良いモデルであるが、どちらかといえば［１・３］が優れている。

３章　判別分析

例

入学試験の合否と勉強時間・模擬試験の平均点のデータを求めたところ以下のような結果を得た（Samples¥判別分析１.txt)。合否を判定するための勉強時間と平均点の１次関数を求めよ。またこの関数によってこのデータを判別し、誤判別の確率を求めよ。

合否	勉強時間	平均点	合否	勉強時間	平均点
1	5.6	70.2	2	3.8	67.4
1	5.9	74.2	2	3.8	61.3
1	4.1	72.7	2	1.7	60.6
1	5.1	84.9	2	2.7	77.2
1	5.0	93.0	2	4.3	65.9
1	3.2	80.5	2	3.3	74.4
1	4.3	62.7	2	3.5	72.1
1	4.8	85.4	2	2.1	69.7
1	3.3	84.3	2	4.3	68.7
1	5.3	64.8	2	2.0	70.5
1	5.3	60.7	2	3.6	45.9
1	5.4	74.4	2	2.8	54.6
1	3.6	85.5	2	2.5	64.4
2	3.8	47.9	2	5.2	50.7
2	3.9	70.8	2	2.2	65.7

解説

判別分析の目的

2 群（多群）を判別する最適な 1 次式を求める。

判別得点＝b_1 勉強時間＋b_2 平均点＋b_0

判別関数

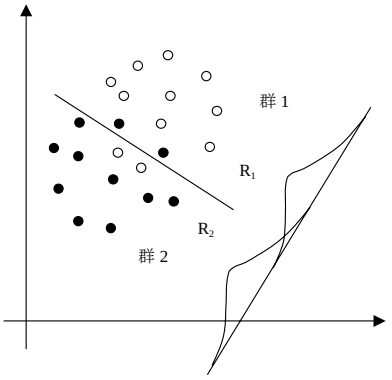


図　判別の概念図

判別分析が有効に利用できる条件は？

→　正規性・等共分散性（等共分散の検定）

判別関数の係数は？　　→　判別関数の欄

判別関数で群を分けるのは？

→　判別の分点 0（多群の場合値が最大の群）

各係数の有効性は？（要正規性・等共分散性）

→　確率の欄（係数が 0 と異なるかの検定）

誤判別の程度は？　　→　誤判別確率（実測と理論）（理論値は要正規性・等共分散性）

マハラノビス距離とは　　→　どの程度 2 群が離れているかを表わす指標

マハラノビス距離	1	4	9	16	25
誤判別確率	0.309	0.159	0.067	0.023	0.006

データ毎の判別関数の値と判別状況　　→　判別得点

事象の生起確率とは？→ 合格・不合格の現れる確率が大きく異なっている場合の措置

　　各群同じかデータ数からが実用的

誤判別損失とは？→ 間違った判断をした場合の致命傷の程度

　　大きな差がない限り、各群１とするのが実用的

解答

正規性の検定結果　→　正規分布とみなす。

等共分散性の検定結果

自由度　　3
χ2 統計値　 0.19093
片側確率Ｐ　0.97904
有意水準α　0.05
Ｐ≧αより、共分散間に差があるといえない。

判別分析結果

	勉強時間	平均点	定数項	判別の分点
判別関数	2.2461	0.2007	-23.0187	0.0000
F 検定値	19.8822	15.0274		
自由度	1,27	1,27		
確率	0.00013	0.00061		
マハラノビスの距離	5.6823			
誤判別確率	1 群を 2 群と	2 群を 1 群と		
実測から	0.07692	0.05882		
理論から	0.11665	0.11665		

判別得点結果

	所属群	判別得点	判定
1	1	3.6512	1 群
2	1	5.1279	1 群
3	1	0.7838	1 群
：	：	：	：
14	2	-4.8682	2 群
15	2	-0.0469	2 群
16	2	-0.9540	2 群
：	：	：	：

まとめ

正規性の検定から、２群とも正規性があるとみなされ、等共分散の検定でも共分散に差があるとは言えなかった。以上から判別分析が適用可能であると判断した。

２群の生起確率を同じとし、誤判別損失を等しいとすると、判別分析によって、以下の判別関数が得られた。

y=2.2461*勉強時間+0.2007*平均点-23.0187

データはこの判別関数の値をもとに、判別の分点を 0 として、2 群に分けられる。

係数の有効性の検定では、勉強時間が p=0.00013、平均点が p=0.00061 のように、両方とも有意に 0 でないことが示された。このことから２つの変数とも有効であると思われる。

マハラノビス距離 5.6823 から、理論的な誤判別確率として p=0.117 が予想される。また、実際に判定を行うと、1 群を 2 群と間違える割合が 7.7%、その逆が 5.9%となる。これらの数値から、判別はかなりうまく行われたものと思われる。

理論

群 1			群 2		
変数 1	…	変数 k	変数 1	…	変数 k
x^1_{11}	…	X^1_{k1}	x^2_{11}	…	x^2_{k1}
x^1_{12}	…	X^1_{k2}	x^2_{12}	…	x^2_{k2}
	…	:	:	…	:
$x^1_{1n_1}$	…	$x^1_{kn_1}$	$x^2_{1n_2}$	…	$x^2_{kn_2}$

判別分析の実行可能条件

分布が多変量正規分布

2 群の共分散が等しい

判別式

$$z={}^t\mathbf{x}\mathbf{S}^{-1}(\mathbf{m}^{(1)}-\mathbf{m}^{(2)})-\frac{1}{2}{}^t(\mathbf{m}^{(1)}+\mathbf{m}^{(2)})\mathbf{S}^{-1}(\mathbf{m}^{(1)}-\mathbf{m}^{(2)})$$

$$=b_0+b_1x_1+b_2x_2+\cdots+b_kx_k$$

$$\mathbf{m}^{(a)}=\frac{1}{n_a}\sum_{\lambda=1}^{n_a}\mathbf{x}^a_\lambda \quad \text{：群 } a \text{ の各変数の平均}$$

$$\mathbf{S}=\frac{1}{n_1+n_2-2}\sum_{a=1}^{2}\sum_{\lambda=1}^{n_a}(\mathbf{x}^a_\lambda-\mathbf{m}^{(a)})^t(\mathbf{x}^a_\lambda-\mathbf{m}^{(a)}) \quad \text{：共分散行列}$$

判別方法

群 j を群 i と間違える損失　C_{ij}

群 i の要素が出現する確率　P_i

1 群に属する：$z - \log_e h \geq 0$

2 群に属する：$z - \log_e h < 0$　　$h = C_{12}P_2 / C_{21}P_1$

z の確率分布

x が群 1 に属する場合　　$N(D^2/2, D^2)$

x が群 2 に属する場合　　$N(-D^2/2, D^2)$

$D^2 = {}^t(\mathbf{m}^{(1)} - \mathbf{m}^{(2)})\mathbf{S}^{-1}(\mathbf{m}^{(1)} - \mathbf{m}^{(2)})$ ：マハラノビスの距離

誤判別の理論確率

群 1 を群 2 と誤判別　　$P_{21} = Z\left(\dfrac{\log_e h - D^2/2}{D}\right)$　網掛け部分

群 2 を群 1 と誤判別　　$P_{12} = 1 - Z\left(\dfrac{\log_e h + D^2/2}{D}\right)$

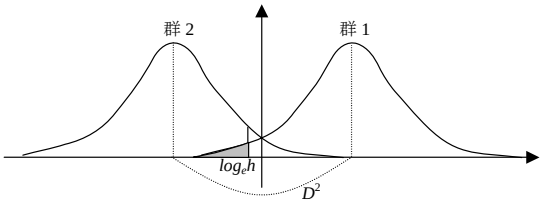


図　誤判別確率

問題

Samples¥判別分析 2.txt は、適性の有無の判定（有：1，無：2）と適性検査の結果と SPI の結果を与えたデータである。判定を適性検査と SPI で予測する判別分析を行い、結果を上のまとめにならって記述せよ。

判別分析　続き

解説（３群以上の判別）

Samples¥判別分析 1.txt を用いて試験の合否を勉強時間と平均点で予測する判別分析

判別分析の目的　　２群（多群）を判別する最適な１次式を求める。

２群の場合　　判別得点＝b_1勉強時間＋b_2平均点＋b_0　　判別関数

判別の分点 0 より大きいか小さいかで１群と２群を分ける

２群以上の場合　判別得点＝b_1勉強時間＋b_2平均点＋b_0－判別の分点

判別得点が最大となる群に属すると判定する。

判別分析が有効に利用できる条件は？　→　正規性，等共分散性（等共分散の検定）

判別関数の係数は？　　→　判別関数の欄

判別関数で群を分けるのは？　→　判別の分点 0（多群の場合は判別得点が最大の群）

各係数の有効性は？　→　確率の欄（係数が 0 と異なるかの検定）

誤判別の程度は？　　→　誤判別確率（実測と理論、３群以上は実測のみ）

データ毎の判別関数の値と判別状況　　→　判別得点

事象の生起確率とは？　→　合格・不合格の現れる確率が大きく異なっている場合の措置，各群同じかデータ数からが実用的

誤判別損失とは？　→　間違った判断をした場合の致命傷の程度

大きな差がない限り、各群１とするが実用的

問題１

Samples¥判別分析 2.txt は、適性の有無の判定（有：1，無：2）と適性検査の結果とＳＰＩの結果を与えたデータである。判定を適性検査とＳＰＩで予測する判別分析を行い、以下の問いに答えよ。但し、事象の生起確率はデータ数から、誤判別損失は２群とも１とすること。

１）このデータに判別分析は利用可能か？

正規性の検定　　正規性があると［みなす・いえない］

等共分散性　　検定確率［　　］，等共分散と［みなす・いえない］

判別分析は効率よく利用可能か。［利用可能・要注意］

２）判別関数を求めよ。

判別得点＝［　　］適性検査＋［　　］ＳＰＩ＋［　　］

３）どちらの変数が判定に影響があると思われるか。［適性検査・ＳＰＩ］

４）実測値から求めた誤判別の確率は？
適性有りを無しと［　　］　適性無しを有りと［　　］
５）厳選して新入社員を取ろうとする場合、上の誤判別でどちらの場合の損失が大きいと思われるか。［適性有りを無し・適正無しを有り］と誤判別する場合
６）上の方針に従って、大きな誤判別損失の値を 2、小さな誤判別損失の値を 1 としたとき、実測値から見た誤判別の確率はどうなるか。
適性有りを無しと［　　］　適性無しを有りと［　　］
７）上の方針で見ると、結果は改善されたか。［改善された・改善されていない］
８）誤判別損失を元に戻して、先頭（1 番）の人の判別得点はいくらか。［　　］
９）適性検査 50 点，ＳＰＩ 55 点の人の判別得点はいくらか、またその人の適性の有無を判定せよ。　判別得点［　　］　適性［有り・無し］

問題２

Samples¥判別分析 3.txt はあやめの種類をがくの長さと幅、花弁の長さと幅で３群に分類したデータである。あやめの群を他の変数の１次式で判別する３群以上の判別分析を行い、以下の問題に答えよ。

１）３つの判別得点の式を求めよ。
判別得点１＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］
判別得点２＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］
判別得点３＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］
２）実測値から求めた誤判別確率はいくらか。
群１を他と［　　］群２を他と［　　］群３を他と［　　］
３）先頭のデータの３つの判別得点を求めよ。
判別得点１［　　］判別得点２［　　］判別得点３［　　］
４）がくの長さ 4.9、がくの幅 3.4、花弁の長さ 1.2、花弁の幅 0.3 のデータはどれに判定されるか。またそのときの最大の判別得点はいくつか。
判定［群１・群２・群３］　最大判別得点［　　］
５）もう１度 Samples¥判別分析 2.txt のデータを用いて、２群の判別関数と３群以上の判別関数の関係を考えよ。

４章　主成分分析

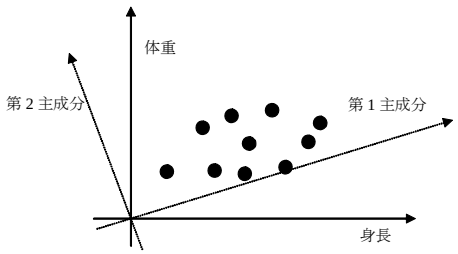
例

以下の健康診断のデータ（Samples¥主成分分析 1.txt）から、変数の１次関数として体格を表す特徴的な指標を作り、その意味を考察せよ。

身長	体重	胸囲	座高	身長	体重	胸囲	座高
148	41	72	78	139	34	71	76
160	49	77	86	149	36	67	79
159	45	80	86	142	31	66	76
153	43	76	83	150	43	77	79
151	42	77	80	139	31	68	74
140	29	64	74	161	47	78	84
158	49	78	83	140	33	67	77
137	31	66	73	152	35	73	79
149	47	82	79	145	35	70	77
160	47	74	87	156	44	78	85
151	42	73	82	147	38	73	78
157	39	68	80	147	30	65	75
157	48	80	88	151	36	74	80
144	36	68	76	141	30	67	76
139	32	68	73	148	38	70	78

解説

Samples¥主成分分析 1.txt のデータから、変数の１次関数として体格を表す特徴的な指標を作る。



主成分分析の目的

複数の変数を１次関数として組み合わせて、いくつかの特徴的な量を作り出す。

各主成分の係数値は？　→　固有ベクトルの値（全体的に符号を変えてもよい）

各主成分のばらつき（分散）は？　→　各主成分の固有値

各主成分の重要性（分散の割合）は？　→　各主成分の寄与率

各主成分と各変数の関係は？　→　因子負荷量（各主成分と各変数の相関係数）

何番目の主成分まで意味があるか？　→　等固有値の検定（要正規性）

主成分が意味がある→他の主成分と値が異なる

データごとの主成分の値は？　→　主成分得点

共分散行列からと相関行列からどちらを使う　→　実用的には相関行列が一般的

解答

分散共分散行列を元にした場合

	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分
固有値	124.8142	10.8487	2.4955	1.8634
寄与率	0.8914	0.0775	0.0178	0.0133
累積寄与率	0.8914	0.9689	0.9867	1.0000
固有ベクトル				
身長	0.6240	-0.6456	0.2236	-0.3792
体重	0.5592	0.3456	-0.7458	-0.1080
胸囲	0.4083	0.6605	0.6245	-0.0841
座高	0.3622	-0.1660	0.0621	0.9151
因子負荷量				
身長	0.9530	-0.2907	0.0483	-0.0708
体重	0.9670	0.1762	-0.1824	-0.0228
胸囲	0.8857	0.4224	0.1915	-0.0223
座高	0.9474	-0.1280	0.0230	0.2925

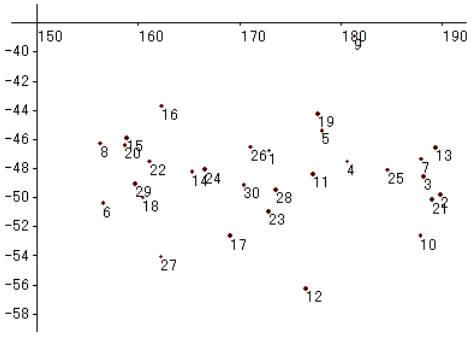
等固有値の検定

利用主成分	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分
χ 2 値	71.3232	12.3594	0.2769	
自由度	9	5	2	
等固有値確率	0.00000	0.03018	0.87071	
利用可能性	可	可	不可	不可

主成分得点

	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分
1	172.9312	-46.7674	52.3252	4.7627
2	189.8320	-49.7748	52.6616	6.2477
3	188.1962	-48.5304	57.2945	6.8066
4	180.6139	-47.4922	54.7601	6.8893
5	178.1285	-45.3882	55.4968	4.9264
:	:	:	:	:

主成分得点の散布図（y 軸：第 2 主成分／x 軸：第 1 主成分）



まとめ

変数に身長、体重、胸囲、座高の４つをとって主成分分析を行なった。各変数の値に大きな差がないことから、ここでは共分散行列を基にした方法を用いている。変数は正規分布するものとみなされ、等固有値の検定も利用可能である。

第１主成分は１次式の係数の値（固有ベクトルの値）がすべて正であることから身体の大きさを表わす変数であると考える。また、第 2 主成分は身長・座高と体重・胸囲で符号が違うことから、肥満の程度を表わす変数であると考える。

これらの主成分の寄与率をみると、第１主成分が 0.8914、第２主成分が 0.0775 であり、他はすべて 0.02 以下になっている。また等固有値の検定より、第１主成分と第２主成分が利用可能であることが分かる。第３主成分以降については意味付けが困難であり、利用しない。最後に結果を式で表わしておく。

身体の大きさを表わす主成分

第 1 主成分＝0.6240 身長＋0.5592 体重＋0.4083 胸囲＋0.3622 座高

肥満の程度を表わす主成分

第２主成分＝－0.6456 身長＋0.3456 体重＋0.6605 胸囲－0.1660 座高

理論

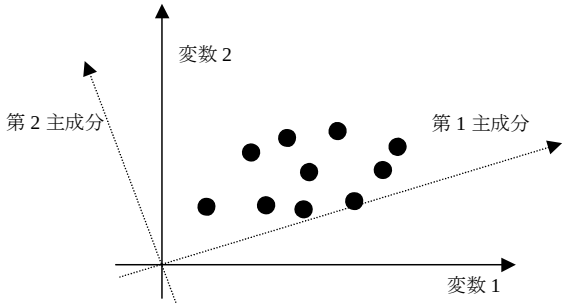
標本番号	変数 1	変数 2	…	変数 p
1	x_{11}	x_{21}	…	x_{p1}
2	x_{12}	x_{22}	…	x_{p2}
:	:	:	…	:
n	x_{1n}	x_{2n}	…	x_{pn}

これらの変数を組み合わせて x_i の変化に最も敏感な特徴的な量 y を作り出す。

$$y = a_1x_1 + a_2x_2 + \cdots + a_kx_k \qquad (\text{但し、} \sum_{i=1}^{k} a_i^2 = 1)$$

x_i の変化に最も敏感とは、

$\sigma^2 = \dfrac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda - \bar{y})^2$ を、$\sum_{i=1}^{p} a_i^2 = 1$ の条件下で最大化する。



固有方程式 $\mathbf{Su} = \lambda\mathbf{u}$ により回転角度を求める。

$\mathbf{S}$ ：共分散行列（標準化したデータから始める場合、相関行列となる。）

λ ：固有値，　${}^t\mathbf{u} = (a_1, a_2, \cdots, a_p)$ ：固有ベクトル

各固有値に応じて、固有ベクトルとして係数 a_i が決まる。

寄与率

第 i 主成分の固有値 λ_i ＝ 第 i 主成分の分散

$$c_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_k}$$　　（全変動に対する第 i 主成分の変動の割合）

因子負荷量　　第 i 主成分と各変数の相関係数

等固有値の検定

第 1 主成分が何らかの意味を持つとは､他の主成分に比べて分散に区別がつく（分散＝固有値が大きい）ことである。即ち、すべての主成分の固有値が同じだと、主成分分析は意味がない。第 i 主成分の等固有値確率とは、第 i 成分以下の固有値がすべて等しくなる確率である。

主成分得点　　データごとの主成分の値

$y = a_1x_1 + a_2x_2 + \cdots + a_kx_k$ の中に各データの値を代入したもの

問題

Samples¥主成分分析 2.txt のデータから主成分分析を行い、結果をまとめにならって記述せよ。

主成分分析 続き

解説

主成分分析の目的

　複数の変数を１次関数として組み合わせて、いくつかの特徴的な量を作り出す。

各主成分の係数値は？　→　固有ベクトルの値（全体的に符号を変えてもよい）

各主成分のばらつき（分散）は？　→　各主成分の固有値

各主成分の重要性（分散の割合）は？　→　各主成分の寄与率

各主成分と各変数の関係は？　→　因子負荷量（各主成分と各変数の相関係数）

何番目の主成分まで意味があるか？　→　等固有値の検定（要正規性）

データごとの主成分の値は？　→　主成分得点

共分散行列からと相関行列からどちらを使う　→　実用的には相関行列が一般的

問題１

　Samples¥主成分分析 2.txt は生徒の教科別の成績データである。共分散行列をもとにするモデルを用いて以下の問いに答えよ。

１）各主成分の固有値（分散の値)、寄与率、累積寄与率を求めよ。

	第１主成分	第２主成分	第３主成分	第４主成分	第５主成分
固有値					
寄与率					
累積寄与率					

２）各変数の正規性の検定　正規分布と［みなす・いえない］

　　これより等固有値の検定は［利用可能・利用不可能］

３）等固有値の検定が利用できる場合、有意に固有値が異なるといえる主成分の数は

　　［　　］個

４）第１主成分と第２主成分の関数はどのように表されるか。

　　第１主成分＝［　　］英語＋［　　］数学

　　＋［　　］国語＋［　　］理科＋［　　］社会

　　第２主成分＝［　　］英語＋［　　］数学

　　＋［　　］国語＋［　　］理科＋［　　］社会

５）これら２つの主成分で説明できるのは全体の変動の何％か。［　　］％

６）これら２つの主成分はどのように意味づけられるか。

　　第１主成分　意味［　　］

　　第２主成分　意味［　　］

7）先頭（1 番）の生徒の２つの主成分得点を求めよ。

第１主成分得点［　　］　第２主成分得点［　　］

8）先頭の生徒について軸の平行移動をした主成分得点を求めよ。

第１主成分得点［　　］　第２主成分得点［　　］

9）２つの主成分の意味を考えて、この生徒にはどんな特徴があるか。

［　　］

10）主成分得点で軸の平行移動を行わない場合と行った場合の違いは。

行った主成分得点＝行わない主成分得点－主成分得点の［　　］

問題２

Samples¥主成分分析 3.txt はある５段階の授業評価についてのデータである。相関行列をもとにするモデルを用いて以下の問いに答えよ。

1）各主成分の寄与率と累積寄与率を求めよ。

	第１主成分	第２主成分	第３主成分	第４主成分
寄与率				
累積寄与率				

2）等固有値の検定が利用可能かどうか判定し、利用できる場合、有意に固有値が異なる主成分の数を求めよ。

［利用可能・利用不可能］　　異なる主成分［　　］個

3）第１主成分と第２主成分の関数はどのように表されるか。

第１主成分＝［　　］分り易さ＋［　　］有益さ

＋［　　］私語注意＋［　　］受講態度

第２主成分＝［　　］分り易さ＋［　　］有益さ

＋［　　］私語注意＋［　　］受講態度

4）これら２つの主成分で説明できるのは全体の変動の何%か。［　　］%

5）これら２つの主成分はどのように意味づけられるか。

第１主成分　意味［　　］

第２主成分　意味［　　］

6）先頭から 4 番目の授業の２つの主成分得点を求めよ。

第１主成分得点［　　］　第２主成分得点［　　］

7）２つの主成分の意味を考えて、この授業にはどんな特徴があるか。

［　　］

8）軸の平行移動をしない場合とした場合で主成分得点に差があるか。

［ある・ない］

５章　因子分析

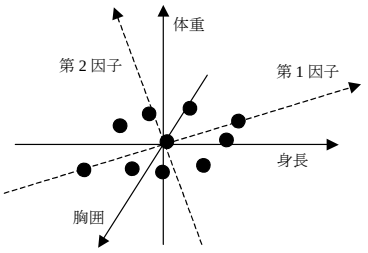
例

以下の健康診断のデータ（Samples¥因子分析 1.txt）から、変数の背後にある体格を表す共通因子を求め、その意味を考察せよ。

身長	体重	胸囲	座高	身長	体重	胸囲	座高
148	41	72	78	139	34	71	76
160	49	77	86	149	36	67	79
159	45	80	86	142	31	66	76
153	43	76	83	150	43	77	79
151	42	77	80	139	31	68	74
140	29	64	74	161	47	78	84
158	49	78	83	140	33	67	77
137	31	66	73	152	35	73	79
149	47	82	79	145	35	70	77
160	47	74	87	156	44	78	85
151	42	73	82	147	38	73	78
157	39	68	80	147	30	65	75
157	48	80	88	151	36	74	80
144	36	68	76	141	30	67	76
139	32	68	73	148	38	70	78

解説

Samples¥因子分析 1.txt のデータから、体格を表す共通因子を求める。（標準化された）各変数 z_i は因子 f_α を用いて以下の形で表されるものとする。



$$z_i \cong \sum_{\alpha=1}^{q} a_{i\alpha} f_\alpha$$

因子分析の目的

各変数の背後にある共通因子を求め、それらの１次関数として各変数が表されるように係数を求める。

各因子の係数値は？　→　因子負荷量の値（全体的に符号を変えて見てもよい）

各因子と各変数の相関係数は？　→　因子負荷量の値（因子間は無相関とした場合）

各因子の重要性は？　→　各因子の寄与率

何番目の因子まで考えるか？　→　累積寄与率が 90%程度まで（寄与率も見る）

相関行列の固有値で１より大きい固有値の数

データごとの因子の値は？　→　因子得点

解答

因子負荷量

累積寄与率から因子数を 2 として計算する。

	因子 1	因子 2	共通性
身長	0.8860	0.4316	0.9713
体重	0.6021	0.7638	0.9459
胸囲	0.4132	0.8480	0.8898
座高	0.7725	0.5467	0.8957
寄与率	0.4788	0.4469	
累積寄与率	0.4788	0.9257	

因子 1 は身長と座高を代表する因子、因子 2 は体重と胸囲を代表する因子

因子得点

	因子 1	因子 2
1	-0.1107	0.5525
2	1.5046	-0.8553
3	1.5186	-0.0360
4	0.7995	0.0424
5	0.4581	1.3368
:	:	:

まとめ

変数に身長、体重、胸囲、座高の４つをとって因子分析を行った。累積寄与率が 0.9 になるように因子数を 2 とし、主因子法を用いて計算を実行した。

バリマックス回転の実施後、各変数は２つの因子と因子負荷量を用いて以下のように予測される。

身長＝0.8860×因子 1 +0.4316×因子 2

体重＝0.6021×因子 1 +0.7638×因子 2

胸囲＝0.4132×因子 1 +0.8480×因子 2

座高＝0.7725×因子 1 +0.5467×因子 2

第 1 因子は身長と座標の因子負荷量が大きいため、体の縦方向の大きさを代表する因子、第 2 因子は体重と胸囲の因子負荷量が大きいため、体の横方向の大きさを代表する因子と考えられる。

各変数を予測するこれらの因子の寄与率は第 1 因子が 0.4788、第 2 因子が 0.4469 で、２つの因子の累積寄与率は 0.9257 である。

理論

標本番号	変数 1	変数 2	…	変数 p
1	x_{11}	x_{21}	…	x_{p1}
2	x_{12}	x_{22}	…	x_{p2}
:	:	:	…	:
n	x_{1n}	x_{2n}	…	x_{pn}

変数に内在する共通因子 $f_{\alpha\lambda}$（$\alpha = 1,2,\cdots,q \le p$）を仮定して、変数の線形予測モデルを作る。但し変数 $x_{i\lambda}$ は標準化されているとする。

$$x_{i\lambda} = \sum_{\alpha=1}^{q} a_{i\alpha} f_{\alpha\lambda} + e_{i\lambda} \qquad e_{i\lambda}：\text{誤差},\ a_{i\alpha}：\text{因子負荷量}$$

$$\sum_{\lambda=1}^{n} f_{\alpha\lambda} e_{i\lambda} = 0,\ \sum_{\lambda=1}^{n} e_{i\lambda} e_{j\lambda} = 0\ (i \ne j),\ \sum_{\lambda=1}^{n} f_{\alpha\lambda} = 0,\ \sum_{\lambda=1}^{n} f_{\alpha\lambda} f_{\beta\lambda} = \delta_{\alpha\beta}$$

因子負荷量と対角成分を共通性 h_i で置き換えた相関行列 $\mathbf{R}'$ の関係

$$\mathbf{A}^t\mathbf{A} = \mathbf{R}' \quad \text{但し、} h_i = \sum_{\alpha=1}^{q} a_{i\alpha}^2 = 1 - \sum_{\lambda=1}^{n} e_{i\lambda}^2 / n$$

因子負荷量の推定（主因子法）

$\mathbf{R}'\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$ の規格化された固有ベクトル $u_{i\alpha}$ を用いて　$a_{i\alpha} = \sqrt{\lambda_\alpha} u_{i\alpha}$

共通因子の推定（回帰推定法）

$$f_{\alpha\lambda} = \sum_{i=1}^{p} b_{i\alpha} x_{i\lambda},\qquad b_{i\alpha} = \sum_{j=1}^{p} a_{j\alpha} r^{ij}$$

（r^{ij} は相関行列の逆行列成分）

寄与率

$$P_\alpha = \sum_{i=1}^{p} a_{i\alpha}^2 \Big/ p = \lambda_\alpha / p$$

問題１

Samples¥因子分析 2.txt のデータから設定をデフォルトとして因子分析を行い、結果をまとめに習って記述せよ。

問題２

Samples¥因子分析 3.txt は北海道各地の２月の気温データである。設定はデフォルトとして以下の問いに答えよ。注）江差町（えさし：南部），寿都町（すっつ：南部），小樽市（おたる：中部），留萌市（るもい：北部），天塩町（てしお：北部）

１）各都市間の相関行列の固有値を大きい順に４つ求めよ。

1	2	3	4

2）因子数を４として、因子分析を行い、累積寄与率を求めよ。

因子１	因子２	因子３	因子４

3）これらのデータから因子数はいくつと決めるのが妥当か。[　　]個

以後因子数を２つと決めて各質問に答えよ。

4）各因子の寄与率と累積寄与率を求めよ。

	第１因子	第２因子
寄与率		
累積寄与率		

5）各因子の因子負荷量を求めよ。

	江差	寿都	小樽	留萌	天塩
第１因子					
第２因子					

6）上の因子負荷量の値から各因子の意味を解釈せよ。

第１因子：[　　]の気温を代表する因子

第２因子：[　　]の気温を代表する因子

7）最初の３日間の因子の値（因子得点）を推定せよ。

	第１因子	第２因子
1		
2		
3		

8）この３日間、北海道はどのような気候だったか。

[　　]

9）江差町における最初の３日間の標準化された実測値とその予測値を求めよ。

	実測値	予測値
1		
2		
3		

10）各地の標準化された実測値とその予測値の間の相関係数を求めよ。

江差	寿都	小樽	留萌	天塩

11）このモデルは良いモデルと思うか。

［良いと思う・あまり良いと思わない］

６章　クラスター分析

例

表　各人の好みを 1～9 の点数で表わした表（Samples¥クラスター分析 1.txt）

	日本酒	焼酎	ビール	ウィスキー	ワイン
増川	1	2	9	6	5
西山	3	1	7	5	4
三好	5	3	4	2	2
芝田	3	6	2	8	3
尾崎	4	6	9	3	4
藤田	7	2	5	4	5
細川	7	5	4	3	2

注）実在の人名とは関係ありません。

クラスター分析の目的

１）回答の類似度で個人を分類する。　→　個体（レコード）の分類

２）回答の類似度で変数を分類する。　→　変数の分類

クラスター分析は分類をどのように表示するか　→　デンドログラム（解答参照）

デンドログラムの縦軸は　→　要素またはクラスター間の距離（類似の程度を示す量）

要素間の距離とは

個体間について

量的データ：ユークリッド距離、標準化ユークリッド距離、マハラノビス距離等

質的 0/1 データ：類似比、一致係数、φ係数等を使ったもの

変数間について

量的データ：1-相関係数、1-相関係数|、1-順位相関係数、1-|順位相関係数|

質的データ：平均平方根一致係数、一致係数、クラメールのＶ等を使ったもの

要素間の距離を知るには　→　距離行列

クラスターを作る方法

最短距離法（棒状の分布に最適），最長距離法（クラスターを分離する能力が高い）

他に、群平均法、重心法、メジアン法、ウォード法

クラスター構成過程を表示するには　→　クラスター構成と距離

解答

要素間の距離測定法にはユークリッド距離、クラスター構成法には最長距離法を用いて、個体間のクラスター分析を行った。

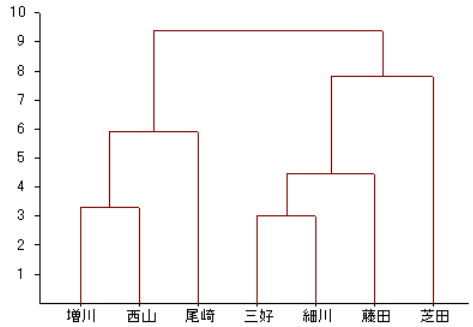
距離行列

	増川	西山	三好	芝田	尾崎	藤田	細川
増川	0.0000	3.3166	8.1854	8.7750	5.9161	7.4833	9.3808
西山	3.3166	0.0000	5.4772	7.7460	5.8310	4.7958	7.0000
三好	8.1854	5.4772	0.0000	7.3485	6.3246	4.3589	3.0000
芝田	8.7750	7.7460	7.3485	0.0000	8.7178	7.8102	6.8557
尾崎	5.9161	5.8310	6.3246	8.7178	0.0000	6.5574	6.2450
藤田	7.4833	4.7958	4.3589	7.8102	6.5574	0.0000	4.4721
細川	9.3808	7.0000	3.0000	6.8557	6.2450	4.4721	0.0000

クラスター構成法と距離

	クラスター名	クラスター名	距離
1	三好	細川	3.0000
2	増川	西山	3.3166
3	三好	藤田	4.4721
4	増川	尾崎	5.9161
5	三好	芝田	7.8102
6	増川	三好	9.3808

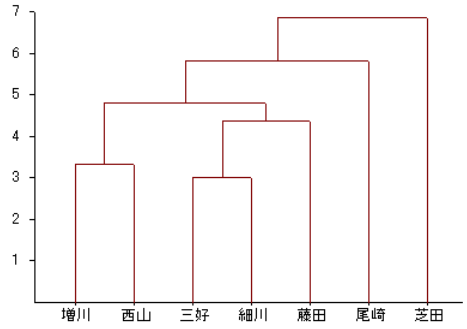
デンドログラム（最長距離法）



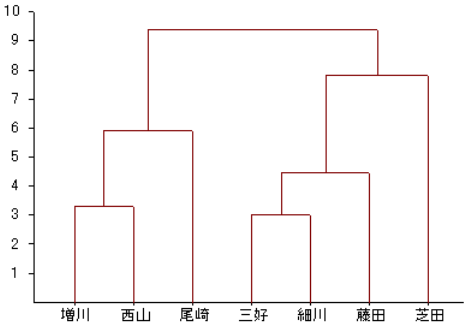
データの内容によってはさらに変数間のクラスター分析が行える。（まとめ参照）

まとめ

酒類の好みを与えたクラスター分析 1.txt のデータから、クラスター分析を用いて各人の好みの分類を行った。距離測定法はユークリッド距離、クラスター構成法は最短距離法と最長距離法を用いて以下のようなデンドログラムを得た。



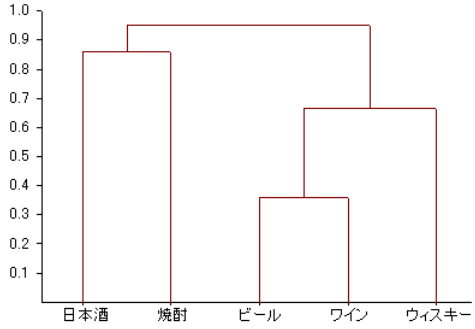
最短距離法



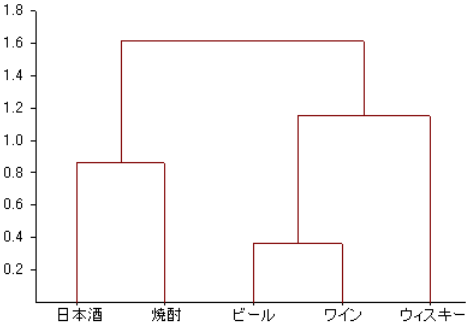
最長距離法

これらによると、クラスターは大きく増川・西山さんと三好・細川・藤田さんに分かれる。尾崎さんと芝田さんは比較的独立な存在であると思われる。

酒の種類については、距離測定法は相関係数を利用した方法で、クラスター構成法は最短距離法と最長距離法を用いて分析を行い、以下のデンドログラムを得た。



最短距離法



最長距離法

これらによると２つのクラスター構成法とも日本酒・焼酎、ビール・ワイン・ウィスキーに分かれている。

理論

距離$d_{\mu\nu}$の主な定義

量的データ

ユークリッド距離　$d_{\mu\nu}^2 = \sum_{i=1}^{p}(x_{i\mu} - x_{i\nu})^2$

標準化ユークリッド距離　$d_{\mu\nu}^2 = \sum_{i=1}^{p}\frac{1}{s_i^2}(x_{i\mu} - x_{i\nu})^2$　　等

質的 0/1 データ

類似比　　$d_{\mu\nu} = 1 - a/(a+b+c)$

一致係数　　$d_{\mu\nu} = 1 - (a+d)/(a+b+c+d)$

ファイ係数　　$d_{\mu\nu} = 1 - (ad-bc)/\sqrt{(a+b)(c+d)(a+c)(b+d)}$　等

ここに、a,b,c,d は以下のように与えられる。

$$a = \sum_{i=1}^{p} x_{i\mu}x_{i\nu}\,,\quad b = \sum_{i=1}^{p} x_{i\mu}(1-x_{i\nu})\,,\quad c = \sum_{i=1}^{p}(1-x_{i\mu})x_{i\nu}\,,\quad d = \sum_{i=1}^{p}(1-x_{i\mu})(1-x_{i\nu})$$

変数間の距離

相関　　$d_{ij} = 1 - r_{ij}$　　（1−相関係数）

順位相関　　$d_{ij} = 1 - \tilde{r}_{ij}$　　（1−順位相関係数）　等

クラスター構成法

クラスター f とクラスター g を結合してクラスター h を作り、他のクラスター l との距離を求める場合

最短距離法　　$D_{hl} = \frac{1}{2}D_{fl} + \frac{1}{2}D_{gl} - \frac{1}{2}\left|D_{fl} - D_{gl}\right|$

最長距離法　　$D_{hl} = \frac{1}{2}D_{fl} + \frac{1}{2}D_{gl} + \frac{1}{2}\left|D_{fl} - D_{gl}\right|$

重心法　　$D_{hl}^2 = \frac{n_f}{n_h}D_{fl}^2 + \frac{n_g}{n_h}D_{gl}^2 - \frac{n_f n_g}{n_h^2}D_{fg}^2$　等

問題

クラスター分析 2.txt は５教科の試験成績のデータである。これを用いてクラスター分析を行ない、結果をまとめに習って記述せよ。

クラスター分析　続き

クラスター分析の目的

　1）個体（レコード）の分類

　2）変数の分類

クラスター分析は分類をどのように表示するか　→　デンドログラム

デンドログラムの縦軸は　→　要素またはクラスター間の距離（類似の程度を示す量）

要素間の距離とは

　個体間について

　　量的データ：ユークリッド距離、標準化ユークリッド距離、マハラノビス距離等

　　質的 0/1 データ：類似比、一致係数、φ係数等を使ったもの

　変数間について

　　量的データ：相関係数、順位相関係数等を使ったもの

　　質的データ：平均平方根一致係数、一致係数、クラメールのＶ等を使ったもの

要素間の距離を知るには　→　距離行列

クラスターを作る方法

　最短距離法（棒状の分布に最適），最長距離法（クラスターを分離する能力が高い）

　他に、群平均法、重心法、メジアン法、ウォード法

クラスター構成過程を表示するには　→　クラスター構成と距離

問題 1

Samples¥クラスター分析 4.txt はある野球チームの今年度の成績である。これについてクラスター分析を行い以下の問いに答えよ。

1）ユークリッド距離及び標準化ユークリッド距離を用いた場合、山下と田中の距離はいくらか。ユークリッド距離［　　］　標準化ユークリッド距離［　　］

2）各変数の標準偏差はいくらか。

打率	安打	本塁打	打点	盗塁

3）上の結果から、距離測定法はどちらを利用すべきか。

　［ユークリッド距離・標準化ユークリッド距離］　以後はこの距離を用いる。

4）クラスター構成法を最長距離法とする場合、最初にクラスターを構成するのはどの要素とどの要素でそれらの距離はいくらか。

　［　　］と［　　］で距離［　　］

5）上の設定で、最初にクラスターとクラスター、またはクラスターと要素の結合になるのはどのようなクラスター（要素）か。それらに含まれる要素を示せ。またその際の距離はいくらか。

クラスター［　　］とクラスター（要素）［　　］で

距離［　　］

6）最長距離法の場合、４分類か５分類が適当と思われるが、４分類の場合、各クラスターにはどのような要素が含まれるか。

［　　］［　　］［　　］［　　］

7）最長距離法と最短距離法とでどちらの分類が理解しやすいと思われるか。

［最長距離法・最短距離法］

8）１－相関係数の距離測定法で最長距離法を用いて変数を３分類すると各クラスターに含まれる要素はどのようになるか。

［　　］［　　］［　　］

9）上の距離測定法で最長距離法と最短距離法とでクラスター構成に違いがあるか。

［ある・ない］

問題２

Samples¥クラスター分析 3.txt のデータを用いてクラスター分析を行い、以下の文を完成させよ。

個体の分類を行う場合、各変数の不偏分散（標準偏差）は［ほぼ等しい・異なる］と思われるので、距離測定法はユークリッド距離を利用する。またクラスター構成法については分類に大きな差が［見られるが・見られず］、最終的に３分類が妥当なように思われる。そのときの各クラスターに含まれる要素は以下のように与えられる。

［　　］［　　］［　　］

変数の分類では、１－相関係数を用いた距離測定法を使い、最短距離法と最長距離法で分類を行ったところ、最終的な２分類で、各クラスターに含まれる要素は、最短距離法で［　　］と［　　］、最長距離法で［　　］と［　　］となった。これらの分類はどちらとも納得できるものである。

７章　正準相関分析

例　正準相関分析 1.txt のデータを用いて、複数の変数間で相関の高い特徴的な量を求める。

身長	座高	体重	胸囲
148	78	41	72
160	86	49	77
159	86	45	80
153	83	43	76
:	:	:	:
148	78	38	70

正準相関分析の目的　→　複数の変数からなる２つの群の中で特徴的な量を見出し、それらの最大の相関を求める。

どのようにして相関を考えるのか。

$y = a_1$身長$+ a_2$座高
$z = b_1$体重$+ b_2$胸囲　　正準変数の組 y と z が最大の相関を持つよう係数を選ぶ。

y と z の最大の相関とは　→　正準相関係数（変数の組によって複数ある）

係数はどのように表示されるか。　→　正準相関分析で１群係数と２群係数

正準変数 y と z の各データの値を見るには　→　正準変量値

各変数と同じ群の正準変数との関係は　→　正準負荷量（相関係数）、解釈に利用

各変数と違う群の正準変数との関係は　→　交差負荷量（相関係数）、解釈に利用

複数の正準変数の組が得られるが、他の正準変数の組同士の関係は　→　相関係数 0

解答

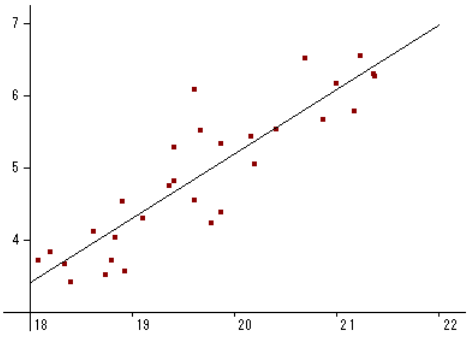
各変数の相関係数

	身長	座高	体重	胸囲
身長	1.00000	0.92046	0.86316	0.73211
座高	0.92046	1.00000	0.88273	0.78288
体重	0.86316	0.88273	1.00000	0.89651
胸囲	0.73211	0.78288	0.89651	1.00000

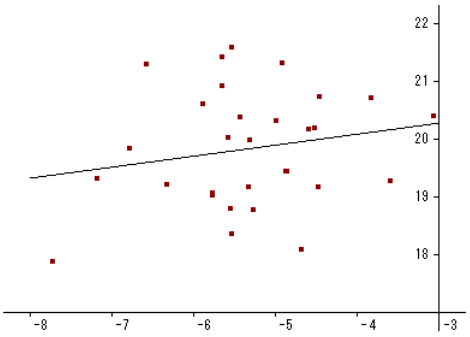
正準相関分析の結果

	正準変量 1	正準変量 2
正準相関係数	0.8935	0.1907
１群係数		
身長	0.0549	−0.3454
座高	0.1447	0.5814
２群係数		
体重	0.1720	−0.3041
胸囲	−0.0244	0.4375

各正準変数の組内での正準変数１（横軸）と正準変数２（縦軸）との関係



正準変量１



正準変量２

各変数と同じ群の正準変数との関係

	正準負荷量１	正準負荷量２
１群		
寄与率	0.9585	0.0415
身長	0.9704	-0.2415
座高	0.9876	0.1570
２群		
寄与率	0.8773	0.1227
体重	0.9985	0.0556
胸囲	0.8705	0.4922

各変数と違う群の正準変数との関係

	交差負荷量１	交差負荷量２
１群		
身長	0.8671	-0.0460
座高	0.8824	0.0299
２群		
体重	0.8921	0.0106
胸囲	0.7778	0.0938

まとめ

体形に関する変数、身長、体重、胸囲、座高について、体の縦方向の大きさを表わす身長・座高と体の横方向の大きさを表わす体重・胸囲に分け、それらでどのような特徴的な量が作れ、またそれらの間に最大どれだけの相関があるか正準相関分析を用いて調べる。

これらの変数の１次関数が最大の相関を持つようにするには、２つの正準変数を以下のようにおけばよい。その際の正準相関係数の値は、0.8935 である。

y = 0.0549 身長 + 0.1447 座高

z = 0.1720 体重 − 0.0244 胸囲

正準負荷量の値から、1 群を身長や座高を特徴付ける量、2 群を体重や胸囲を特徴付ける量といえる。

問題

正準相関分析 2.txt について、文系科目（英語・国語・社会）と理系科目（数学・理科）に分け、正準相関分析を実行し、以下の問いに答えよ。但し、共分散行列を用いたモデルにし、第 1 正準変量について考えること。

1）文系科目と理系科目の正準相関係数はいくらか。［　　］

2）文系科目と理系科目の正準変数はそれぞれどのように表されるか。

文系正準変数＝［　　］英語＋［　　］国語＋［　　］社会

理系正準変数＝［　　］数学＋［　　］理科

3）各変数の正準負荷量の値はいくらか。

英語	国語	社会	数学	理科

4）各変数の交差負荷量の値はいくらか。

英語	国語	社会	数学	理科

5）各正準変数と最も相関のある同じ組の科目は何か。

文系正準変数では［英語・国語・社会］、理系正準変数では［数学・理科］

6）各正準変数と最も相関のある違う組の科目は何か。

文系正準変数へは［数学・理科］、理系正準変数へは［英語・国語・社会］

7）第 1 正準変量の寄与率はいくつか。第 1 正準変量だけ考えた判断は正しいか。

寄与率［　　］　第 1 正準変量で［十分・不十分］

8）先頭の人の各正準変数の値を求めよ。

文系正準変数［　　］　理系正準変数［　　］

9）英語 60、国語 72、社会 66、数学 58、理科 55 の人の各正準変数の値を求めよ。

文系正準変数［　　］　理系正準変数［　　］

以後は相関行列を用いたモデルに変更して解答せよ。

10）文系科目と理系科目の正準変数はそれぞれどのように表されるか。但し、変数はすべて標準化されたものを用いること。

文系正準変数＝［　　］英語＋［　　］国語＋［　　］社会

理系正準変数＝［　　］数学＋［　　］理科

11）各変数の正準負荷量の値はいくらか。

英語	国語	社会	数学	理科

12）各変数の交差負荷量の値はいくらか。

英語	国語	社会	数学	理科

13）各正準変数と最も相関のある同じ組の科目は何か。

文系正準変数では［英語・国語・社会］、理系正準変数では［数学・理科］

14）各正準変数と最も相関のある違う組の科目は何か。

文系正準変数へは［数学・理科］、理系正準変数へは［英語・国語・社会］

15）各科目の平均と標準偏差（不偏分散からのもの）を求め、以下の式によって、９）の人の標準化値を求めよ。標準化変数＝（値－平均）／標準偏差

科目	英語	国語	社会	数学	理科
標準化値					

16）上の標準化値を利用して９）の人の正準変数の値を求めよ。

文系正準変数［　　］　理系正準変数［　　］

８章　数量化Ⅰ・Ⅱ・Ⅲ類

8.1 数量化Ⅰ類

例

以下の地域（1：都市部、2：山村部)、気候（1：温暖、2：寒冷、3：平均的)、ある商品の販売率のデータ（数量化Ⅰ類 1.txt）から販売率（目的変数）を予測する式を作り、それがどの程度有効か検討する。

販売率	地域	気候
3.0	1	2
1.8	2	1
1.5	2	2
:	:	:
2.3	1	3

上のようなアイテムのデータから、それぞれのアイテムが複数のカテゴリに分かれる以下の形のデータを作る。

販売率	地域-1	地域-2	気候-1	気候-2	気候-3
3.0	1	0	0	1	0
1.8	0	1	1	0	0
1.5	0	1	0	1	0
:	:	:	:	:	:
2.3	1	0	0	0	1

このデータをもとに以下の式で目的変数を予測する。

$$Y = a_{11}x_{11} + a_{12}x_{12} + a_{21}x_{21} + a_{22}x_{22} + a_{23}x_{23} + a_{00}$$

（基準化）カテゴリウェイト　→　上式の係数 a_{ij}

カテゴリウェイト、重回帰カテギリウェイト、基準化カテゴリウェイトの違いは

→　予測値を計算する上では同じ（予測値への影響の見易さが異なる）

予測値と実測値との相関係数　→　重相関係数

予測値は実測値をどれだけ説明しているか　→　寄与率

各アイテムの重要性は　→　相関／偏相関ボタンのウェイト範囲

予測値と実測値の散布図　→　散布図ボタン

解答（カテゴリウェイトのみ）

	地域:1	地域:2	気候:1	気候:2	気候:3	定数項
カテゴリウェイト	3.5167	1.8917	0.0000	-0.3750	-1.4667	0.0000
基準化　ウェイト	0.4875	-1.1375	0.6992	0.3242	-0.7675	2.3300
重相関係数	0.9679					
寄与率	0.9367					

8.2 数量化Ⅱ類

例

顧客が車を購入する際、３種類の特性について検討し、a かｂの車種を購入した（数量化Ⅱ類 1.txt)。顧客がどのような選択を行うかでどちらの車を購入するか判別する式を作る。

群	価格	外観	性能
a	1	1	2
a	2	1	1
a	1	2	1
a	2	2	1
b	1	1	3
:	:	:	:
b	2	1	3

上のような２分されたアイテムのデータから、以下の形のデータを作る。

群	価格:1	価格:2	外観:1	外観:2	性能:1	性能:2	性能:3
a	1	0	1	0	0	1	0
a	0	1	1	0	1	0	0
a	1	0	0	1	1	0	0
a	0	1	0	1	1	0	0
b	1	0	1	0	0	0	1
:	:	:	:	:	:	:	:
b	0	1	1	0	0	0	1

このデータからどちらの群に属するかを予測する判別関数を作る。

$$y = a_{11}x_{11} + a_{12}x_{12} + a_{21}x_{21} + a_{22}x_{22} + a_{31}x_{31} + a_{32}x_{32} + a_{33}x_{33} + a_{00}$$

（基準化）カテゴリウェイト　→　上式の係数 a_{ij}

カテゴリウェイトと基準化カテゴリウェイトの違いは

→　判別関数値を計算する上では同じ

判別方法は　→　判別得点と群別得点平均をみて、どちらの値に近いかで判定する。

各アイテムの重要性は　→　相関／偏相関ボタンのウェイト範囲

解答（カテゴリウェイトのみ）

カテゴリウェイト

	価格:1	価格:2	外観:1	外観:2	性能:1	性能:2	性能:3	切片
CW	0.0000	0.2158	0.0000	0.0896	0.0000	0.5085	0.7330	0.0000
基準化 CW	-0.1511	0.0647	-0.0358	0.0537	-0.4458	0.0627	0.2873	0.6326

判別得点

判別得点の値を見てどの群に含まれるか判断する

8.3 数量化Ⅲ類

例

各人が以下の食品について、それぞれの好み（1：好物、0：それほどでも）を与えた（Samples¥数量化Ⅲ類 1.txt)。これから好みの特徴を表す式を求め、人と食品を分類する。

ご飯	パン	うどん	そば	ラーメン	スパ
1	0	1	1	1	0
1	0	1	0	0	0
0	1	0	0	1	1
:	:	:	:	:	:
0	1	0	1	1	0

上のような項目に反応したかどうかを表わすデータ $x_{i\lambda}$ （i：カテゴリ，λ：個体）から、以下のようなカテゴリと個体とで特徴的な量を求める。

$$w_\lambda^\alpha = \sum_{i=1}^{p} u_i^\alpha x_{i\lambda} \quad （u_i^a：カテゴリウェイト，\ w_\lambda^a：個体得点）$$

カテゴリウェイトで個体得点を求め、個体得点で個体の分類を行なう。

$$y_i^\alpha = \sum_{\lambda=1}^{n} v_\lambda^a x_{i\lambda} \quad （v_\lambda^\alpha：個体ウェイト，\ y_i^\alpha：カテゴリ得点）$$

個体ウェイトでカテゴリ得点を求め、カテゴリ得点でカテゴリの分類を行なう。

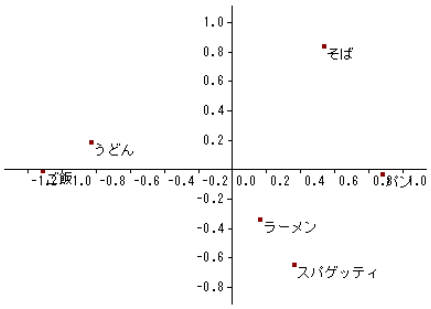
解答

個体ウェイト（２次元まで）

	第 1 次元	第 2 次元
1	-0.09367	0.14993
2	-0.30648	0.07237
3	0.15140	-0.25263
:	:	:
15	0.17225	0.14452

カテゴリ得点

	第 1 次元	第 2 次元
ご飯	-1.11223	-0.00929
パン	0.87853	-0.03002
うどん	-0.82917	0.18765
そば	0.53730	0.83608
ラーメン	0.16362	-0.33692
スパゲッティ	0.36195	-0.64749



問題

数量化Ｉ類 2.txt は店舗の売り上げを立地、人通り、競合の３段階分類データで予測しようとするものである。

１）カテゴリウェイト（定数項を０）を用いた予測式を表せ。

予測売り上げ＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］競合３

２）重回帰カテゴリウェイト（各先頭アイテムを基準）を用いた予測式を表せ。

予測売り上げ＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］競合３＋［　　］

３）基準化カテゴリウェイトを用いた（目的変数の平均値を基準）予測式を表せ。

予測売り上げ＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］競合３＋［　　］

４）予測式は実測値の変動を何％予測できるか。［　　］％

５）立地：２，人通り：２，競合：２の店舗の売り上げを予測せよ。［　　］

６）予測値に最も大きな影響を与えるアイテムは何か。［立地・人通り・競合］

７）成否は売り上げ 4000（万円）境に失敗と成功の２つに分けたものであるが、数量化Ⅱ類を用いた判別関数を表せ。

判別関数＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］競合３

８）判別関数の値がいくらのところで分けたらうまく判別できると思うか。

［ -0.4 ・ -0.2 ・ 0 ］

９）同じ分析を 0/1 データを用いた重回帰分析で行った。但し、各アイテムの第１カテゴリは係数が 0 として、変数から外した。そのときの重回帰式を示せ。

予測売り上げ＝［　　］立地２＋［　　］立地３
＋［　　］人通り２＋［　　］人通り３
＋［　　］競合２＋［　　］競合３＋［　　］

10）このことから上の重回帰分析と数量化Ｉ類は［同じ・異なる］ものと考えられる。

９章　時系列分析

時系列分析はあるデータの時間的変化を分析し、モデルを作成して今後の予測を行うことを目的とする。

9.1 変動の分解モデル

時系列データを傾向変動、季節変動、循環変動、残差に分解し、データの性質を調べると同時に予測も行う手法で、データに周期性がある場合に有効

例

Samples¥時系列分析 1.txt の売上 1 データを、傾向変動、季節変動、残差に分解し、来月の売上を予測せよ。

傾向変動　全体的な変動の傾向を表す変動
季節変動　一定の周期を持つ変動
循環変動　一定の周期ではない変動（ここでは長期の周期変動を考えている）
残差　　これらの変動を差し引いた残りの変動
時系列データを見る　→　「元データ」ラジオボックスを選択し、描画ボタン
傾向変動を分解する　→　近似モデルで見てよく適合するモデルを求める。
　変動の分解の表示で、元データ、傾向変動、残差をチェックし、実行
周期性を見る　→　コレログラム（自己相関のグラフ）とピリオドグラムで調べる。
季節変動を分解する　→　季節変動分解の周期を入力し、表示に季節変動を加え実行
循環変動を見る　→　再度周期性を調べ、循環変動分解の周期を入力し、
　　表示に循環変動を加え実行
どの程度予測があっているかの目安　→　残差 2 乗平均の値、R^2 値

手順

１）ここでは傾向変動の近似モデルは回帰モデルの 1 次式を選択する。
２）傾向変動と残差をグラフで確認する。
３）傾向変動を除いた残差から周期性をコレログラムやピリオドグラムで確認する。
４）傾向変動、季節変動、残差をグラフで確認する。
５）再度残差の周期性をコレログラムやピリオドグラムで確認する。
６）残差 2 乗平均の値を見る。（予測の良さを表す）
７）データの実測値と予測値の R^2 の値を求める。（予測の良さを表す）
８）モデルの予測値を確認する。

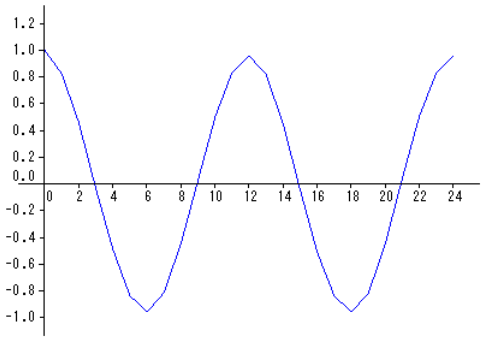


図 1 コレログラム

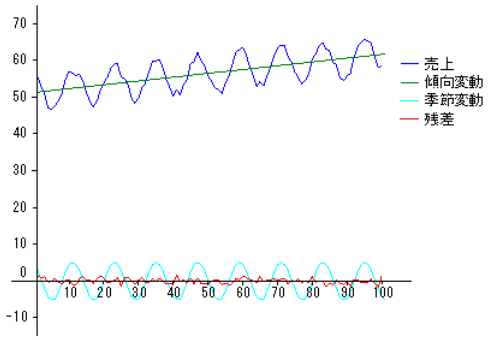


図 2 変動の分解

変動の分割

	売上	予測値	傾向変動	季節変動	循環変動	残差
89	54.5	55.43	60.45	-5.02	0.00	-0.93
90	55.6	55.66	60.55	-4.89	0.00	-0.06
91	56.5	57.89	60.65	-2.76	0.00	-1.39
92	60.6	60.33	60.76	-0.43	0.00	0.27
93	63.5	63.21	60.86	2.35	0.00	0.29
94	64.8	65.36	60.96	4.40	0.00	-0.56
95	65.7	66.16	61.07	5.10	0.00	-0.46
96	65.2	65.68	61.17	4.51	0.00	-0.48
97	64.7	64.25	61.27	2.97	0.00	0.45
98	60.9	61.72	61.38	0.35	0.00	-0.82
99	58	59.65	61.48	-1.83	0.00	-1.65
100	58.6	57.19	61.58	-4.39	0.00	1.41
@1		56.66	61.68	-5.02	0.00	

図 3 予測値の表示

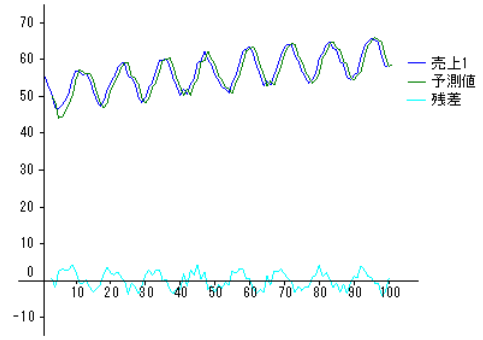
9.2 予測モデル

例

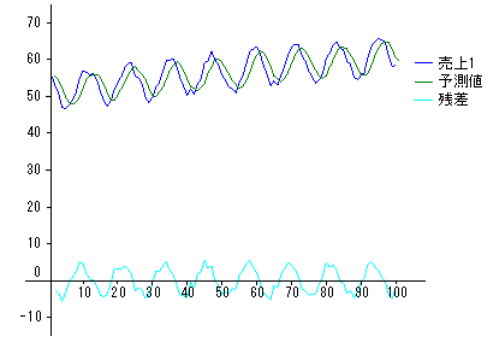
Samples¥時系列分析 1.txt の売上 1 データを予測モデルで分析せよ。

方法

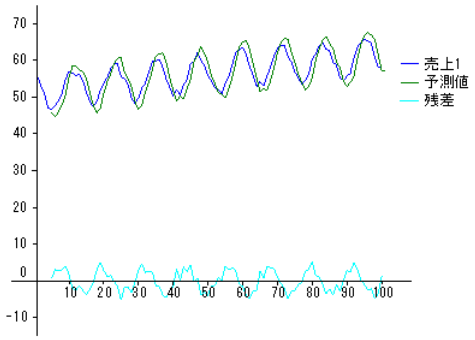
パラメータを調整しながら各分析を実行し、残差 2 乗平均などを用いて精度を確認する。



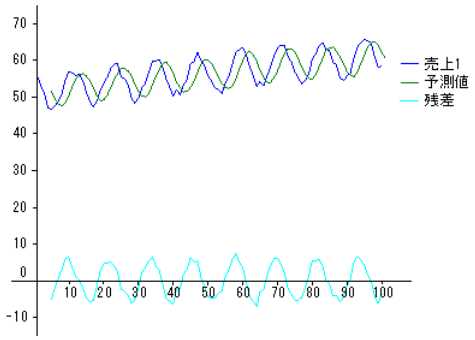
差の平均法（z^2=5.05）



指数平滑法（z^2=10.23）

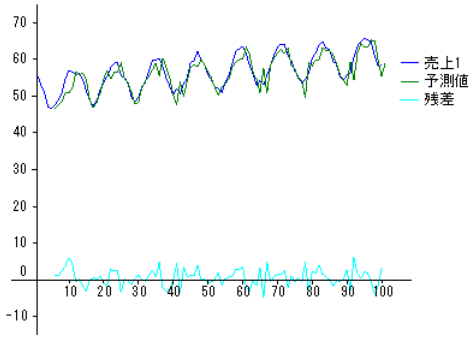


ブラウン法（z^2=7.73）

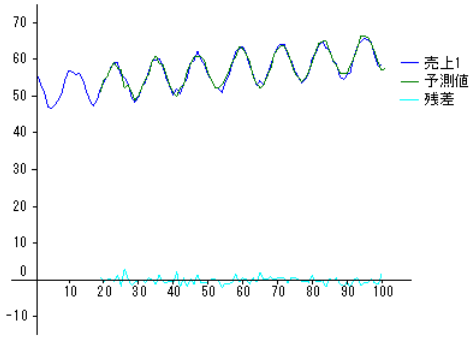


移動平均法（z^2=17.65）

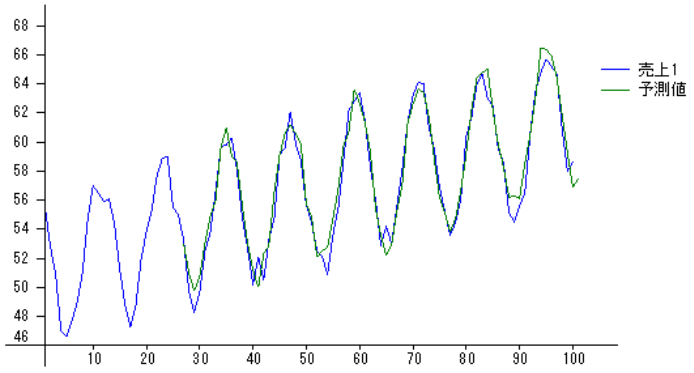
上の古典的方法は上下変動のあるデータに対してあまり役に立たない。



最近隣法（z^2=5.67）



ARIMA(6,0,0)（z^2=2.41）



傾向変動を除去した ARIMA(6,0,2)（z^2=0.97）

問題 1

Samples¥時系列分析 1.txt の売上 2 について、以下の問いに答えよ。

変動の分解モデル

1）このデータの傾向変動を 1 次式で推定するとどのような式になるか。

売上＝［　　］×時間＋［　　］

2）上の傾向変動を除いた場合の残差 2 乗平均の値はいくらか。　［　　］

3）傾向変動を除いた残差から、コレログラムを用いて季節変動の周期を求めるといくらか。　［　　］

4）上の季節変動を除いた場合の残差 2 乗平均の値はいくらか。　［　　］

5）データを上の傾向変動と季節変動で予測するモデルの R^2 の値はいくらか。

［　　］

6）このモデルでの 1 期先の予測値はいくらか。　［　　］

7）このモデルでの 5 期先の予測値はいくらか。　［　　］

予測モデル

8）以下のモデルで、最適な実測・予測の R^2 値、残差 2 乗平均と 1 期先の予測値を求めよ。

	実測・予測 R^2	残差 2 乗平均	次期予測値
差の平均法			
指数平滑法			
ブラウン法			

9）最近隣法で傾向変動の分解を行わない場合と行う場合の残差 2 乗平均を求めよ。

行わない場合［　　］，行う場合［　　］

10）上の方法で傾向変動の分解を行った場合の R^2 値を求めよ。

［　　］

11）傾向変動の分解を行った ARIMA モデルで、残差の 2 乗平均を最小にする最適なパラメータを求めよ。但し $p \le 2, d = 0, q \le 2$ とする。

（p，d，q）＝（　，0，　）

12）上の方法の場合の残差の 2 乗平均を求めよ。　［　　］

問題２

Samples¥時系列分析 1.txt の売上 4 について、以下の問いに答えよ。

変動の分解モデル

１）このデータの傾向変動は、１次近似、対数近似、べき乗近似、指数近似、多項式近似のうちどれが最適か？そのときの残差 2 乗平均の値はいくらか。
近似は［　　］，　残差 2 乗平均［　　］

２）傾向変動を除いたデータから、コレログラムなどを用いて季節変動の周期を求めるといくらか。　［　　］

３）傾向変動と季節変動を除いた残差 2 乗平均はいくらか。　［　　］

４）傾向変動と季節変動を除いたデータから、コレログラムなどを用いて循環変動（長周期）の周期を求めるといくらか。　［　　］

５）上の変動をすべて除いた残差 2 乗平均はいくらか。　［　　］

６）データを上の傾向変動、季節変動、循環変動で予測するモデルの R^2 の値はいくらか。　［　　］

７）このモデルでの 1 期先の予測値はいくらか。　［　　］

８）このモデルでの 10 期先の予測値はいくらか。　［　　］

予測モデル

９）以下のモデルで、最適な実測・予測の R^2 値、残差 2 乗平均と 1 期先の予測値を求めよ。但し、パラメータは最適なものを用いること。

	実測・予測 R^2	残差 2 乗平均	次期予測値
差の平均法			
指数平滑法			
ブラウン法			

11）最近隣法で傾向変動の分解を行わない場合と行う場合の残差 2 乗平均を求めよ。
行わない場合［　　］，行う場合［　　］

12）上の方法で傾向変動の分解を行った場合の R^2 値を求めよ。
［　　］

13）傾向変動の分解を行った ARIMA モデルで、残差の 2 乗平均を最小にする最適なパラメータを求めよ。但し $p \le 6, d = 0, q \le 3$ とする。
（p，d，q）＝（　，0，　）

14）上の方法の場合の残差の 2 乗平均と R^2 値を求めよ。
残差 2 乗平均［　　］，　R^2 値［　　］