応用統計学

阪本雄二

目 次

1 多変量データ

1.1 2 変量データ

右の表は，ある町の 10 地区における世帯数と 1 か月間のごみ排出量 (0.1 トン) である．地区番号 i の世帯数を x_i，ごみ排出量を y_i と表し，

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i,$$
$$s_{xx} := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$
$$s_x := \sqrt{s_{xx}}$$

と定義すると，$\bar{x}$, s_{xx}, s_x はそれぞれ，世帯数の**平均** (mean)，**分散** (variance)，**標準偏差 (SD)**(standard deviation) と呼ぶのであった．ここで，n は調査した地区の数を表し，この表では $\bar{x} = 52.1$, $s_{xx} = 475.3$, $s_x = 21.8$, $n = 10$ である．ごみ排出量 y_i に対しても同様に定義すると $\bar{y} = 24.9$, $s_{yy} = 105.7$, $s_y = 10.28$ となる．

表 **1.1** ごみ排出量

地区番号	世帯数	排出量
1	73	37
2	63	27
3	31	18
4	24	11
5	79	39
6	84	40
7	32	14
8	33	18
9	66	28
10	36	17
単位	人	0.1 トン
平均	52.1	24.9
分散	475.3	105.7
SD	21.8	10.28
共分散	219.4	
相関係数	0.979	

データが調査対象からの**無作為標本** (random sample) であると考える場合は，調査対象全体である**母集団** (population) の平均や分散，標準偏差を標本から見積もる必要がある．そのような場合は，

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

を用いて，母集団の分散 (**母分散** (population variance)) が見積もることが望ましい．なぜなら，s_{xx} は母分散より小さくなる傾向があるが，u_{xx} はその傾向が修正されているからである．これらを区別するため，s_{xx} を**標本分散** (sample variance)，u_{xx} を**標本不偏分散** (sample unbiased variance) と呼ぶ．

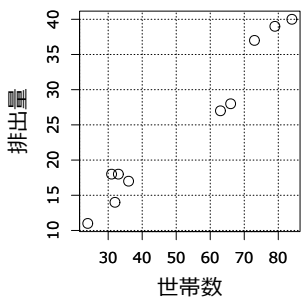
標本の平均である $\bar{x}$ と母集団の平均を区別するため，それぞれ，**標本平均** (sample mean)，**母平均** (population mean) と呼ぶ．標本平均は母平均に対して偏る傾向がないので，見積もる場合もそのまま用いられる．

さらに，世帯数と排出量の関係を視覚的にとらえるのが右の**散布図** (scatter diagram) である．散布図では世帯数が増えると排出量が増える直線的な傾向があるように見えるが，直線的関係の強さは，

$$s_{xy} := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$
$$r_{xy} := \frac{s_{xy}}{s_x\, s_y}$$

と定義される**共分散** (covariance) s_{xy} と **相関係数** (correlation coefficient) r_{xy} によって測ることがで

図 **1.1** ごみ排出量と世帯数の散布図



きる．この表では，$s_{xy} = 219.4$, $r_{xy} = 0.979$ である．ただし，共分散 s_{xy} は標準偏差 s_x と s_y が大きくなればいくらでも大きくなるのに対して，相関係数はどんなに s_x と s_y が大きくなっても

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

の範囲にとどまる．この表における r_{xy} は上限の 1 に近いため，強い直線関係があると考えてよい．

直線関係を具体的に求める時は，

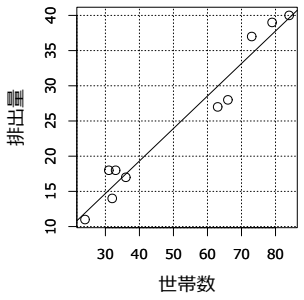
$$S(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

を最小にするように a, b の値を求めればよいが，

$$\hat{a} := \frac{s_{xy}}{s_{xx}}, \quad \hat{b} := \bar{y} - \hat{a}\bar{x}$$

が $S(a,b)$ を最小にする a, b の値であることが示せる．この $\hat{a}$, $\hat{b}$ を**回帰係数** (regression coefficient)，それらを傾きと切片に持つ直線 $y = \hat{a}x + \hat{b}$ を**回帰直線** (regression line) と呼ぶ．このように直線関係を具体的にする方法を**最小 2 乗法** (least squares method) という．

図 **1.2** ごみ排出量の回帰直線



共分散 $s_{xy} = 219.4$, 分散 $s_{xx} = 475.3$, 平均 $\bar{x} = 52.1$, $\bar{y} = 24.9$ の値を用いると，$\hat{a} = 0.462$, $\hat{b} = 0.849$ となり，回帰直線は $y = 0.462x + 0.849$ となる．それを散布図に書くと右上図のようになる．

共分散も，母集団の推測に用いる場合は，

$$u_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

が用いられる．これを**標本不偏共分散** (sample unbiased covariance) と呼び，s_{xy} を **標本共分散** (sample covariance) と呼ぶ．

一般に，2 変量データ $(x_1, y_1), \ldots, (x_n, y_n)$ は右の表のようにまとめられ，基本統計量は平均 $\bar{x}$, $\bar{y}$，分散 s_{xx}, s_{yy}，標準偏差 s_x, s_y と，変量間の直線関係の強さを表す s_{xy}, r_{xy} である．また，散布図は関係を視覚的に捉えるために有効であり，直線関係がある時は，最小 2 乗法で回帰直線 $y = \hat{a}x + \hat{b}$ を求めればよい．求めた回帰直線の統計的な妥当性を検証する (区間推定や仮説検定を行う) には，回帰モデルを立て，データの出現確率に仮定を置く必要がある．

表 **1.2** 2 変量データ

i	x_i	y_i
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n
平均	$\bar{x}$	$\bar{y}$
分散	s_{xx}	s_{yy}
SD	s_x	s_y
共分散	s_{xy}	
相関係数	r_{xy}	

演習 1.1　次の 2 変量データの平均，分散，共分散，相関係数を求め，それらから，回帰直線を求めよ．さらに，分散共分散行列の固有値と固有ベクトルを求め，固有値はすべて正であること，固有ベクトルは互いに直交していることを確かめよ．

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	4	2	1	4	-9	-8	-4	4	6	-3	-1	-8
y_i	-2	5	-9	4	-3	-7	-9	8	5	-1	-7	-8

コーシー=シュワルツ (Cauchy-Schwarz) の不等式)

(1) ベクトル $\vec{a} = (a_1, a_2)$, $\vec{b} = (b_1, b_2)$ に対して，

$$|\vec{a} \cdot \vec{b}| \leqq ||\vec{a}|| \cdot ||\vec{b}||.$$

ただし，$\vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2 (= ||\vec{a}|| \cdot ||\vec{b}|| \cos\theta)$, $||\vec{a}|| = \sqrt{a_1^2 + a_2^2}$ であり，等号成立のための必要十分条件は $\exists p, q \neq (0,0)$, $p\vec{a} + q\vec{b} = \vec{0}$.

(2) ベクトル $\boldsymbol{a} = (a_1, \ldots, a_n)$, $\boldsymbol{b} = (b_1, \ldots, b_n)$ に対して，

$$|\langle \boldsymbol{a}, \boldsymbol{b} \rangle| \leqq ||\boldsymbol{a}|| \cdot ||\boldsymbol{b}||.$$

ただし，$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = a_1 b_1 + \cdots + a_n b_n = \sum_{i=1}^n a_i b_i$,
$||\boldsymbol{a}|| = \sqrt{a_1^2 + \cdots + a_n^2} = \sqrt{\sum_{i=1}^n a_i^2}$

(3) $(x_1, y_1), \ldots, (x_n, y_n)$ に対して，

$$|s_{xy}| \leqq s_x s_y.$$

これは (2) において $a_i = x_i - \bar{x}$, $b_i = y_i - \bar{y}$ とすれば証明できる．等号成立のための必要十分条件は，

$$\exists (p, q) \neq (0,0), \exists r \in \mathbb{R}, \forall i = 1, \ldots, n, px_i + qy_i + r = 0.$$

(4) 数列 $\{a_n\}$, $\{b_n\}$ に対して，

$$\left| \sum_{n=1}^{\infty} a_n b_n \right| \leqq \sqrt{\sum_{n=1}^{\infty} a_n^2} \sqrt{\sum_{n=1}^{\infty} b_n^2}.$$

(5) 関数 f, g に対して，

$$\left| \int_{t_0}^{t_1} f(t) g(t) dt \right| \leqq \sqrt{\int_{t_0}^{t_1} f^2(t) dt} \sqrt{\int_{t_0}^{t_1} g^2(t) dt}.$$

(6) 確率変数 X, Y に対して，

$$|E[XY]| \leqq \sqrt{E[X^2]} \sqrt{E[Y^2]}$$

$$|Cov[X, Y]| \leqq \sqrt{Var[X]} \sqrt{Var[Y]}$$

統計解析ソフト R による実行例

```
> setwd("c:/usr/rtemp")                    # 作業フォルダを c:/usr/rtemp にする
> gomi<-read.csv("gomi.csv")               # データの読み込み，gomi に代入
> gomi                                     # 読み込み結果の確認
   地区番号 世帯数 排出量
1         1     73     37
2         2     63     27
3         3     31     18
4         4     24     11
5         5     79     39
6         6     84     40
7         7     32     14
8         8     33     18
9         9     66     28
10       10     36     17
> n<-nrow(gomi)                            # データ数を取得して，n に代入
> n                                        # データ数の確認
[1] 10
> mean(gomi$世帯数)                        # 世帯数の平均を求める
[1] 52.1
> var(gomi$世帯数)*(n-1)/n                 # 標本分散を求める．var は標本不偏分散
[1] 475.29
> sqrt(var(gomi$世帯数)*(n-1)/n)           # 標準偏差を求める
[1] 21.80115
> var(gomi$世帯数,gomi$排出量)*(n-1)/n     # 共分散を求める
[1] 219.41
> cor(gomi$世帯数,gomi$排出量)             # 相関係数を求める
[1] 0.9789491

> plot(排出量~世帯数,data=gomi)           # 散布図の描画
> gomi.lm<-lm(排出量~世帯数,data=gomi)     # 最小 2 乗法で回帰直線を求める
> coefficients(gomi.lm)                    # 回帰係数を表示する
> abline(gomi.lm)                          # 回帰直線を散布図に書き加える
```

1.2 多変量データ

右のデータは中学 2 年生男子の遠投，握力，身長，体重に関するデータである．出席番号 (No) が i の男子の遠投を x_{i1}, 握力を x_{i2}, 身長を x_{i3}, 体重を x_{i4} と表すことにし，それぞれの平均を $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, $\bar{x}_4$, 分散を s_{11}, s_{22}, s_{33}, s_{44}, 標準偏差を s_1, s_2, s_3, s_4 と表すことにする．すなわち，

$$\bar{x}_j := \frac{1}{n}\sum_{i=1}^{n} x_{ij},$$

$$s_{jj} := \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2,$$

$$s_j := \sqrt{s_{jj}}$$

と定義する．ここで，n は男子生徒の人数であり，右の表では $n = 15$ である．x_{ij} と書く時，最初の添え字 i は出席番号を表しており，2 番目の添え字 j は，遠投を 1 番目，握力を 2 番目，身長を 3 番目，体重を 4 番目の特性量とする時の，j 番目の特性量であることを表している．すなわち，x_{ij} は出席番号 i の男子生徒の第 j 特性量である．

表 **2.1** 体力測定

No	遠投	握力	身長	体重
1	22	28	146	34
2	36	46	169	57
3	24	39	160	48
4	22	25	156	38
5	27	34	161	47
6	29	29	168	50
7	26	38	154	54
8	23	23	153	40
9	31	42	160	62
10	24	27	152	39
11	23	35	155	46
12	27	39	154	54
13	31	38	157	57
14	25	32	162	53
15	23	25	142	32

さらに，第 j 特性量と第 j' 特性量の共分散を $s_{jj'}$，相関係数を $r_{jj'}$ と表す．すなわち，

$$s_{jj'} := \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}),$$

$$r_{jj'} := \frac{s_{jj'}}{s_j\, s_{j'}}$$

と定義する．これらをまとめて，第 (j, j') 成分が $s_{jj'}$ である行列を S，$r_{jj'}$ である行列を R と表し，それぞれ**標本分散共分散行列** (sample variance-covariance matrix)，**標本相関行列** (sample correlation matrix) と呼ぶ．すなわち，

$$S := \begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ s_{21} & s_{22} & s_{23} & s_{24} \\ s_{31} & s_{32} & s_{33} & s_{34} \\ s_{41} & s_{42} & s_{43} & s_{44} \end{bmatrix}, \quad R := \begin{bmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{bmatrix}$$

と定義する．ここで，s_{11}, s_{22}, s_{33}, s_{44} の S の対角成分は 4 つの特性値の分散を表していることに注意しよう．また，$r_{11} = \dfrac{s_{11}}{s_1\ s_1} = 1$ であり，同様に $r_{22} = r_{33} = r_{44} = 1$ であることにも注意しよう．

中 2 男子のデータに関して，平均，分散，標準偏差を求めると $\bar{x}_1 = 26.2$, $\bar{x}_2 = 33.33$, $\bar{x}_3 = 156.6$, $\bar{x} = 47.4$, $s_{11} = 15.2$, $s_{22} = 45.4$, $s_{33} = 48.8$, $s_{44} = 77.0$, $s_1 = 3.90$, $s_2 = 6.74$, $s_3 = 6.98$, $s_4 = 8.78$ となる．また，標本分散共分散行列 S と標本相関行列 R は

$$S = \begin{bmatrix} 15.23 & 19.67 & 18.68 & 26.99 \\ 19.67 & 45.42 & 24.93 & 50.07 \\ 18.68 & 24.93 & 48.77 & 42.56 \\ 26.99 & 50.07 & 42.56 & 77.04 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0.748 & 0.685 & 0.788 \\ 0.748 & 1 & 0.530 & 0.846 \\ 0.685 & 0.530 & 1 & 0.694 \\ 0.788 & 0.846 & 0.694 & 1 \end{bmatrix}$$

となる．S も R も対称行列であることに注意しよう．さらに，右図のように特性値のすべての組み合わせに対する散布図を行列上に並べたものを**散布図行列** (scatterplot matrices) と呼ぶ．対角線上にはヒストグラムが書かれていて，各散布図には回帰直線が描かれている．遠投 (throw) は体重 (weight) と強い相関があり，握力 (grip) と体重 (weight) も強い相関があることが，散布図からもわかる．また，遠投 (throw) の分布は歪んでいることもヒストグラムから読み取れる．

このように，散布図行列を描くと，多くの変数の関係を一度に捉えることができて，関係の強い変数の組を即座に見分けることができる．多変量データを解析する上で，まず行うべき有効なツールである．

各個体の特性値が多数あるデータは右の表のようにまとめられる．理論的な議論を行うときは，各データを二つの添え字を付けて x_{ij} のように表す．第 1 添え字 i が個体の違いを表し，第 2 添え字 j が特性値の違いを表すことに注意しよう．平均や分散，SD(標準偏差) の添え字はすべて特性値に対応するものである．

図 **2.1** 中 2 男子の体力の散布図行列

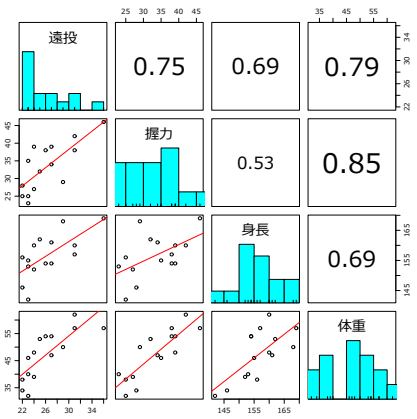


表 **2.2** 多変量データ

i	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{ip}
1	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1p}
2	x_{21}	$\cdots$	x_{2j}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	$\cdots$	x_{nj}	$\cdots$	x_{np}
平均	$\bar{x}_1$	$\cdots$	$\bar{x}_j$	$\cdots$	$\bar{x}_p$
分散	s_{11}	$\cdots$	s_{jj}	$\cdots$	s_{pp}
SD	s_1	$\cdots$	s_j	$\cdots$	s_p

- **分散共分散行列**

  表 **2.2** の多変量データに対して，

$$S = \begin{bmatrix} s_{11} & \cdots & s_{1i} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ s_{i1} & \cdots & s_{ii} & \cdots & s_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pi} & \cdots & s_{pp} \end{bmatrix}$$

  を分散共分散行列と呼ぶ．$s_{ij} = s_{ji}$ だから，分散共分散行列は**対称行列**である．

- **対称行列** (symmetric matrix)

  (1) 対称行列の固有値はすべて実数である．

  (2) 対称行列の異なる固有値に属する固有ベクトルは直交する．

  (3) A が対称行列のとき，直交行列 P と対角行列 Λ が存在して，

$$A = P\Lambda P'$$

  が成り立つ．ここで，直交行列とはすべての列ベクトルの大きさが 1 であり，互いに直交する行列のことである．また，P' は P の転置行列を表す．

- **正定値対称行列** (positive definite matrix)

  すべての固有値が正である対称行列を正定値対称行列と呼ぶ．行列 A が正定値対称行列であることと，

$$\forall \boldsymbol{a} \neq \mathbf{0}, \boldsymbol{a}' A \boldsymbol{a} > 0$$

  は同値である．

- **分散共分散行列の正定値性**

  分散共分散行列は常に正定値対称行列である．

統計解析ソフト R による実行例

```
> tairyoku<-read.csv("tairyoku.csv")
> tairyoku
   遠投 握力 身長 体重
1    22   28  146   34
2    36   46  169   57
3    24   39  160   48
4    22   25  156   38
5    27   34  161   47
6    29   29  168   50
7    26   38  154   54
8    23   23  153   40
9    31   42  160   62
10   24   27  152   39
11   23   35  155   46
12   27   39  154   54
13   31   38  157   57
14   25   32  162   53
15   23   25  142   32
> n<-nrow(tairyoku)
> n
[1] 15
> apply(tairyoku,2,mean)
     遠投       握力       身長       体重
 26.20000  33.33333 156.60000  47.40000
> var(tairyoku)*(n-1)/n
         遠投     握力     身長     体重
遠投 15.22667 19.66667 18.68000 26.98667
握力 19.66667 45.42222 24.93333 50.06667
身長 18.68000 24.93333 48.77333 42.56000
体重 26.98667 50.06667 42.56000 77.04000
> cor(tairyoku)
          遠投       握力       身長       体重
遠投 1.0000000 0.7478150 0.6854618 0.7879319
握力 0.7478150 1.0000000 0.5297304 0.8463624
身長 0.6854618 0.5297304 1.0000000 0.6943082
体重 0.7879319 0.8463624 0.6943082 1.0000000
> install.packages("psych", dep=TRUE)
> library(psych)
> pairs.panels(tairyoku,smooth=FALSE,density=FALSE,ellipses=FALSE,
  scale=TRUE,pch=1,lm=TRUE)
```

2 回帰分析

2.1 回帰モデル

第 1.2 節 (p.5) の中学 2 年生男子の体力測定データにおいて，遠投の距離が他の 3 つの変量で説明できるかどうかを考えるために，以下のようなモデルを仮定する．まず，出席番号 (No) が i の握力を x_{i1}, 身長を x_{i2}, 体重を x_{i3} と記号を付け替えることにする．さらに，遠投を y_i と表し，y_i は

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad \epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

のように定義される確率変数 Y_i の実現値であるとする．つまり，握力が x_{i1} ，身長が x_{i2}，体重が x_{i3} ならば，遠投の距離は $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$ 程度になることは決まるが，測定ごとに多少変化し，測定してみないとわからない量 (確率変数) であると考えるわけである．

一般に，表 **1.1** のようなデータに対して，

(**回帰モデル**)
$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad \epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

のように定義される確率変数 Y_i の実現値を y_i と仮定して，係数 $\beta_0, \beta_1, \ldots, \beta_p$ に関する統計的な推測を行うことを**回帰分析** (regression analysisi) という．ここで，$x_{i1}, \ldots, x_{ip}$ を**説明変数** (explanatory variable)，y_i と Y_i を**目的変数** (objective variable)，$\beta_0, \beta_1, \ldots, \beta_p$ を**回帰係数** (regression coefficient)，ε_i を**誤差** (error) と呼ぶ．説明変数は測定ごとに変化しない制御できる変数であり，確率的には一定の値をとるが，目的変数は説明変数以外の誤差的要因 ϵ_i の影響で測定ごとにランダムに変動し，制御しきれない量であると考えていることに注意しよう．回帰モデルの下では，Y_i は平均が $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ で分散が σ^2 の正規分布する確率変数であり，$Y_1, \ldots, Y_n$ は互いに独立である．目的変数と説明変数の平均的な関係 $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ を**回帰式**と呼ぶ．

表 **1.1** 回帰分析のデータ

i	y_i	x_{i1}	x_{i2}	$\cdots\cdots$	x_{ip}
1	y_1	x_{11}	x_{12}	$\cdots\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots\cdots$	x_{np}

2.2 最小 2 乗法

表 **1.1** のようなデータが与えられた時，回帰係数の推定は，以下のような最小 2 乗法と呼ばれる方法で実行できる．

回帰係数 $\beta_0, \beta_1, \ldots, \beta_p$ の候補を $b_0, b_1, \ldots, b_p$ と表すと，回帰式の候補は $y = b_0 + b_1 x_1 + \cdots + b_p x_p$ となる．説明変数が $x_{i1}, \ldots, x_{ip}$ であるときの目的変数の値は，この候補の式から $b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}$ と予想されるが，この値と実際の測定値のズレの大きさを

$$S = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}))^2$$

で測ることにする．S は回帰式の候補と実際のデータの全体的なズレである．この S を最小にする $b_0, b_1, \ldots, b_p$ が回帰式における回帰係数 $\beta_0, \beta_1, \ldots, \beta_p$ であろうと見積もる方法を**最小 2 乗法** (least squares method) と呼ぶ．

今，それぞれの変量の偏差を $\tilde{y}_i = y_i - \bar{y}$, $\tilde{x}_{ij} = x_{ij} - \bar{x}_j$ と置き，$B = -\bar{y} + b_0 + b_1\bar{x}_1 + \cdots + b_p\bar{x}_p$ と置くと，

$$S = \sum_{i=1}^{n} \left(\tilde{y}_i - (B + b_1\tilde{x}_{i1} + \cdots + b_p\tilde{x}_{ip})\right)^2$$

と表される．したがって，S を最小にする $b_0, b_1, \ldots, b_p$ は連立方程式 $\dfrac{\partial S}{\partial B} = 0$, $\dfrac{\partial S}{\partial b_1} = 0, \ldots, \dfrac{\partial S}{\partial b_p} = 0$ の解である．S を B, b_j で偏微分すると，

$$\frac{\partial S}{\partial B} = -2\sum_{i=1}^{n} \left(\tilde{y}_i - (B + b_1\tilde{x}_{i1} + \cdots + b_p\tilde{x}_{ip})\right),$$
$$\frac{\partial S}{\partial b_j} = -2\sum_{i=1}^{n} \left(\tilde{y}_i - (B + b_1\tilde{x}_{i1} + \cdots + b_p\tilde{x}_{ip})\right)\tilde{x}_{ij}$$

となるが，$s_{yj} = \dfrac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_{ij} - \bar{x}_j)$ と置き，$s_{jj'} = \dfrac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$, $\sum_{i=1}^{n}\tilde{y}_i = 0$, $\sum_{i=1}^{n}\tilde{x}_{ij} = 0$ だったことを思い出すと，

$$\frac{\partial S}{\partial B} = -2nB, \quad \frac{\partial S}{\partial B_j} = -2n(s_{yj} - s_{j1}b_1 - \cdots - s_{jp}b_p)$$

となる．したがって，S を最小にする $b_0, b_1, \ldots, b_p$ を $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ とすると，

(**正規方程式 1**)
$$\begin{array}{rcl} \bar{x}_1\hat{\beta}_1 + \cdots + \bar{x}_p\hat{\beta}_p & = & \bar{y} - \hat{\beta}_0 \\ s_{11}\hat{\beta}_1 + \cdots + s_{1p}\hat{\beta}_p & = & s_{y1} \\ \vdots & & \vdots \\ s_{p1}\hat{\beta}_1 + \cdots + s_{pp}\hat{\beta}_p & = & s_{yp} \end{array}$$

である．この連立方程式を**正規方程式** (normal equation) と呼ぶ．ここで，

$$S = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}, \quad \boldsymbol{s} = \begin{bmatrix} s_{y1} \\ \vdots \\ s_{yp} \end{bmatrix}$$

と置くと，S は $x_{i1}, \ldots, x_{ip}$ の標本分散共分散行列であり，$\boldsymbol{s}$ は $\{y_i\}_{i=1,\ldots,n}$ と $\{x_{ij}\}_{i=1,\ldots,n}$ の共分散を j 成分に持つベクトルである．これらを用いると，正規方程式は

(**正規方程式 2**)
$$S\hat{\boldsymbol{\beta}} = \boldsymbol{s}, \quad \hat{\beta}_0 = \bar{y} - \bar{x}_1\hat{\beta}_1 - \cdots - \bar{x}_p\hat{\beta}_p$$

と表すこともできる．$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ を (母) 回帰係数 $\beta_0, \beta_1, \ldots, \beta_p$ の**最小 2 乗推定量** (least square estimator) と呼ぶ．$E[s_{yj}] = \beta_1 s_{j1} + \cdots + \beta_p s_{jp}$ が成り立つので，$E[\boldsymbol{s}] = S\boldsymbol{\beta}$．よって，(正規方程式 2) の両辺の期待値をとると，$SE\left[\hat{\boldsymbol{\beta}}\right] = S\boldsymbol{\beta}$ となり，$E\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$ がわかる．つまり，最小 2 乗推定量は不偏推定量である．ここでは $\epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$ は仮定していない．

2.3 予測値・残差・寄与率

説明変数が $x_{i1}, \ldots, x_{ip}$ である時に目的変数を再度測定したら得られる値は，$\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$ と予測されるので，この $\hat{y}_i$ ををを**予測値** (prediction value) という．また，実際の測定値 y_i と予測値 $\hat{y}_i$ の差 $\hat{e}_i := y_i - \hat{y}_i$ を**残差** (residual) と呼び，

$$S_{ee} := \sum_{i=1}^{n} \hat{e}_i^2$$

を**残差平方和** (residual sum of squares) という．残差は $\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}$ であり，誤差は $\epsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}$ なので，$\hat{e}_i$ は ϵ_i の推定量であると考えられる．

最小 2 乗推定量は，(正規方程式 2) を満たすので，予測値 $\hat{y}_i$ の平均 $\bar{\hat{y}}$ と $\bar{y}$ は等しい．つまり，$\bar{\hat{y}} = \bar{y}$. したがって，残差の平均 $\bar{\hat{e}}$ は，

$$\bar{\hat{e}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \bar{y} - \bar{\hat{y}} = 0.$$

これは誤差の平均が 0 であること，つまり，$E[\epsilon_1] = \cdots = E[\epsilon_n] = 0$ であることに対応する．

一方，

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \hat{e}_i^2 + \sum_{i=1}^{n} (\hat{y}_i - E[\hat{y}_i])^2$$

であり，$E\left[\epsilon_i^2\right] = \sigma^2$, $\sum_{i=1}^{n} Var[\hat{y}_i] = (p+1)\sigma^2$ なので，$E[S_{ee}] = (n-p-1)\sigma^2$．したがって，

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} \hat{e}_i^2 = \frac{1}{n-p-1} S_{ee}$$

とおくと，$E\left[\hat{\sigma}^2\right] = \sigma^2$ となることがわかる．つまり，$\hat{\sigma}^2$ は誤差 ϵ_i の分散 σ^2 の不偏推定量である [*]. (正規方程式 l) を用いると，残差平方和 S_e は

$$S_{ee} = n\left\{s_{yy} - s_{y1}\hat{\beta}_1 - \cdots - s_{yp}\hat{\beta}_p\right\}$$

と変形できることに注意しよう．ただし，$s_{yy} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

観測値 y_i と予測値 $\hat{y}_i$ の相関係数 $R = \frac{s_{y\hat{y}}}{\sqrt{s_{yy}s_{\hat{y}\hat{y}}}}$ を y と $x_1, \ldots, x_p$ の**重相関係数** (multiple correlation coefficient) と呼び，観測値と回帰モデルによる予測値の関係の強さを測る指標として用いる．回帰モデルが観測値にとって適当なモデルであると，予測値が観測値に近い値をとるようになるので，重相関係数の 2 乗 R^2 は回帰モデルの妥当性を測る指標として用いることができる． R^2 を**寄与率** (contribution), あるいは，**決定係数** (coefficient of determination) と呼ぶが，寄与率が 1 に近いほどモデルが観測値に適合していると考えることができる．寄与率は目的変数の偏差平方和 $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ と残差平方和 S_{ee} を用いて，$R^2 = 1 - \frac{S_{ee}}{S_{yy}}$ のように計算できることが知られているが，説明変数の個数が多いと R^2 が大きくなるので，説明変数の個数が異なるモデルを比較するには，その欠点を修正した

$$R^{*2} = 1 - \frac{S_{ee}/\phi_e}{S_{yy}/\phi_T}, \quad \phi_e = n-p-1, \ \phi_T = n-1$$

を用いる方がよい．R^{*2} を**自由度調整済み寄与率** (adjusted contribution) と呼ぶ．

[*] 一般に母数 θ の推定量 $\hat{\theta}$ が $E\left[\hat{\theta}\right] = \theta$ を満たすとき，$\hat{\theta}$ は不偏推定量であるという．

2.4 回帰係数の検定

$S_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ とおくと，$S_{yy} = S_R + S_{ee}$ という関係が成り立つ．このことは，目的変数の変動 S_{yy} が回帰モデルによる予測値による変動 S_R と回帰モデルでは説明できない変動 S_{ee} に分解できることを示している．回帰モデルがデータによくあてはまっているならば，目的変数の変動の大半は回帰モデル由来の S_R であるはずである．ただし，S_R は説明変数の個数 p が多くなれば大きくなり，S_{ee} は小さくなることが知られているので，そのことを調整した

$$W := \frac{\frac{S_R}{p}}{\frac{S_e}{n-p-1}}$$

を考えることにする．W は説明変数の効果が一切ないと考えられる $\beta_1 = \cdots = \beta_p = 0$ の状況で，自由度 $(p, n-p-1)$ の F 分布することが知られているので，自由度 $(p, n-p-1)$ の F 分布の上側 $100\alpha\%$ 点 $f_{p,n-p-1}(\alpha)$ より大きい時，仮説 $H_0 : \beta_1 = \cdots = \beta_p = 0$ を棄却するという検定を考えればいいことがわかる．つまり，仮説 $\begin{cases} H_0 : \beta_1 = \cdots = \beta_p = 0 \\ H_1 : H_0 \text{ ではない} \end{cases}$ に対して，有意水準 α の仮説検定の棄却域は $R = \{W > f_{p,n-p-1}(\alpha)\}$. とすればよいことがわかる．これらの検定に用いる統計量を以下のような表にまとめたものを**分散分析表** (analysis of variance table) と呼ぶ．なお，p-値とは，与えられた F-値 W が棄却される有意水準の中で最小の値のこと，つまり，$W > f_{p,n-p-1}(\alpha)$ を満たす最小の α のことである．

表 **4.1** 分散分析表

変動要因	自由度 (DF)	平方和 (SS)	不偏分散 (MS)	F-値	p-値
回帰	p	S_R	$V_R = S_R/p$	$W = V_R/V_e$	
残差	$n-p-1$	S_e	$V_e = S_e/(n-p-1)$		
全体	$n-1$	S_y			

2.5 遠投の回帰分析

説明変数 (握力 x_1，身長 x_2, 体重 x_3) の平均と目的変数 (遠投 y) の平均は，$\bar{x}_1 = 33.34$, $\bar{x}_2 = 156.6$, $\bar{x}_3 = 47.4$, $\bar{y} = 26.2$ であった．また，説明変数のの標本分散共分散行列 S と 目的変数 (遠投 y) との共分散 $\boldsymbol{s}$，および目的変数の分散 s_{yy} は

$$S = \begin{bmatrix} 45.42 & 24.93 & 50.07 \\ 24.93 & 48.77 & 42.56 \\ 50.07 & 42.56 & 77.04 \end{bmatrix}, \quad \boldsymbol{s} = \begin{bmatrix} 19.67 \\ 18.68 \\ 26.99 \end{bmatrix}, \quad s_{yy} = 15.23$$

であった．したがって，(正規方程式 1) は

$$\begin{cases} 33.34\,\hat{\beta}_1 + 156.6\,\hat{\beta}_2 + 47.4\,\hat{\beta}_3 = & 26.2 - \hat{\beta}_0 & \cdots & ① \\ 45.42\,\hat{\beta}_1 + 24.93\,\hat{\beta}_2 + 50.07\,\hat{\beta}_3 = & 19.67 & \cdots & ② \\ 24.93\,\hat{\beta}_1 + 48.77\,\hat{\beta}_2 + 42.56\,\hat{\beta}_3 = & 18.68 & \cdots & ③ \\ 50.07\,\hat{\beta}_1 + 42.56\,\hat{\beta}_2 + 77.04\,\hat{\beta}_3 = & 26.99 & \cdots & ④ \end{cases}$$

となり，②,③,④ を解くと，$\hat{\beta}_1 = 0.201$, $\hat{\beta}_2 = 0.171$, $\hat{\beta}_3 = 0.125$ となり，① に代入すると $\hat{\beta}_0 = -13.217$ が得られる．まとめると，

$$(\text{遠投 } y\text{m}) = -13.217 + 0.201 \times (\text{握力 } x_1\text{kg}) + 0.171 \times (\text{身長 } x_2\text{cm}) + 0.125 \times (\text{体重 } x_3 \text{ kg})$$

という関係が平均的に成り立つと考えられる．また，残差平方和 S_{ee} は

$$\begin{aligned} S_{ee} =& n\{s_{yy} - s_{y1}\hat{\beta}_1 - s_{y2}\hat{\beta}_2 - s_{y3}\hat{\beta}_3\} \\ =& 15 \times \{15.23 - 19.67 \times 0.201 - 18.68 \times 0.171 - 26.99 \times 0.125\} = 70.50 \end{aligned}$$

なので，平均的関係からのズレ (誤差) の標準偏差 σ は

$$\hat{\sigma} = \sqrt{\frac{1}{n-3-1} S_{ee}} = \sqrt{\frac{1}{11} \times 70.50} = 2.53$$

と推定される．また，$S_{yy} = n s_{yy} = 15 \times 15.23 = 228.4$ なので，

$$R^2 = 1 - \frac{S_{ee}}{S_{yy}} = 1 - \frac{70.50}{228.4} = 0.69, \quad R^{*2} = 1 - \frac{S_{ee}/(n-3-1)}{S_{yy}/(n-1)} = 1 - \frac{70.50/11}{228.4/14} = 0.61$$

のように寄与率 R^2 は 0.69 であり，自由度調整済み寄与率 R^{*2} は 0.61 であることがわかる．

さらに，$S_R = S_{yy} - S_{ee} = 157.9$ なので，$W = (S_R/3)/(S_{ee}/(n-3-1)) = (157.9/3)/(70.50/11) = 8.21$ であり，F 分布表より $f_{3,15-3-1}(0.01) = 6.22$ だから，帰無仮説 $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ は有意水準 0.01% で棄却されることがわかる．つまり，遠投が握力と身長，体重で説明できると言える．

演習 2.1 15 個体について，説明変数 x_1, x_2, x_3 の平均が $\bar{x}_1 = 2, \bar{x}_2 = 3, \bar{x}_3 = -1$ で目的変数 y の平均が $\bar{y} = 6$，説明変数の標本分散共分散行列 S と説明変数と目的変数の共分散 $\boldsymbol{s}$，目的変数の分散 s_{yy} が下のように与えられている時，回帰式を最小 2 乗法により推定せよ．また，誤差分散の推定量，寄与率，自由度調整済み寄与率を求めよ．

$$S = \begin{bmatrix} 4 & 4 & 1 \\ 4 & 6 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad \boldsymbol{s} = \begin{bmatrix} 3 \\ 4 \\ 2 \end{bmatrix}, \quad s_{yy} = 5$$

検定統計量 W の値を求め，回帰係数 a_1, a_2, a_3 に関する

$$\begin{cases} H_0 : & a_1 = a_2 = a_3 = 0 \\ H_1 : & H_0 \text{ でない} \end{cases}$$

を検定せよ．ただし，棄却限界値として，自由度 $(3, 11)$ のエフ分布の上側 5% 点 3.20 を用いよ．

統計解析ソフト R による実行例

```
> tairyoku.lm<-lm(遠投~握力+身長+体重,data=tairyoku)
> summary(tairyoku.lm)

Call:
lm(formula = 遠投 ~ 握力 + 身長 + 体重, data = tairyoku)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9976 -1.3822  0.3611  1.1549  3.9291

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.2173    17.6038  -0.751    0.469
握力          0.2014     0.1842   1.093    0.298
身長          0.1710     0.1316   1.300    0.220
体重          0.1249     0.1667   0.750    0.469

Residual standard error: 2.532 on 11 degrees of freedom
Multiple R-squared:  0.6913,    Adjusted R-squared:  0.6072
F-statistic: 8.213 on 3 and 11 DF,  p-value: 0.003774

> anova(tairyoku.lm)
Analysis of Variance Table

Response: 遠投
          Df  Sum Sq Mean Sq F value    Pr(>F)
握力       1 127.727 127.727 19.9303 0.0009558 ***
身長       1  26.576  26.576  4.1469 0.0665153 .
体重       1   3.600   3.600  0.5618 0.4692729
Residuals 11  70.496   6.409
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

3 主成分分析

3.1 目的

右の表は，ある中学 2 年生の試験結果である．5 教科の得点で学力を区別するためには，合計点で総合的な学力を見たり，数学と理科の得点で理系的な能力を見たりすることが多い．さて，5 教科の得点から学力を評価する客観的な方法を考えてみよう．

表 **1.1** 中 2 学力試験

No	国語	社会	数学	理科	英語
1	29	33	55	79	84
2	71	68	72	64	97
3	74	91	79	76	100
4	52	56	58	60	85
5	77	92	96	88	98
6	60	85	66	66	88
7	81	91	73	63	95
8	61	84	72	78	92
9	70	75	81	67	96
10	53	70	73	51	92
11	69	64	96	57	97
12	87	89	90	85	100
13	83	75	96	81	98
14	76	61	67	57	86
15	87	82	78	82	97
16	77	80	78	70	94
17	38	43	45	12	96
18	67	73	78	67	95
19	83	77	80	67	100
20	47	61	56	21	95
21	70	62	88	51	96
22	81	51	63	66	92
23	51	16	36	48	84

一般に，表 **2.2**(p.6) のような，i 番目の個体に関するデータ $\boldsymbol{x}_i$ が

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

のように p 個の成分を持つベクトル (p 次元ベクトル) である時，個体差は変量ごとに異なるので，p 個の変量を合成した

$$z_i = a_1 x_{i1} + a_2 x_{i2} + \ldots + a_p x_{ip}$$

という変量を考えることにする．合成変量を決めるのは係数 a_1, a_2, ..., a_p であるが，これらの係数はある意味で個体差をよく表すものが望ましい．上の例では，何らかの意味の学力差があらわれるように係数を決定すべきである．ここでは，n 個体の合成変量 $z_1, \ldots, z_n$ の分散 s_z^2 が個体差を表す指標と考えることにする．つまり，s_z^2 を大きくするように a_1, a_2, ..., a_p を決める方法について考えることにする．

このように，分散を最大にする合成変量をつくる方法を **主成分分析** (principal component analysis) という．

3.2 2 次元の時

簡単のため，$p = 2$ のときを考える．そのとき，

$$\begin{aligned} s_z^2 &= \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \frac{1}{n}\sum_{i=1}^{n}\{a_1(x_{i1} - \bar{x}_1) + a_2(x_{i2} - \bar{x}_2)\}^2 \\ &= s_{11}a_1^2 + 2s_{12}a_1a_2 + s_{22}a_2^2 \\ &= [a_1, a_2]\begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \end{aligned}$$

となる．ここで，

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad s_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}).$$

いま，$r = \sqrt{a_1^2 + a_2^2}$, $\bar{a}_1 = a_1/r$, $\bar{a}_2 = a_2/r$ とおくと，

$$s_z^2 = r^2\,[\bar{a}_1, \bar{a}_2] \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \begin{bmatrix} \bar{a}_1 \\ \bar{a}_2 \end{bmatrix}.$$

この形から，r を大きくすると s_z^2 はいくらでも大きくすることができるので，$r = 1$ として，s_z^2 を大きくする $\bar{a}_1, \bar{a}_2$ を決めることにする．どんなデータの分散行列でも，$\lambda_1 \geqq \lambda_2 > 0$ と

$$\sum_{j=1}^{2} u_{jk}u_{jl} = \begin{cases} 1 & (k = l) \\ 0 & (\text{その他の時}) \end{cases}$$

なる u_{jk} を適当に選ぶと，

$$\begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \end{bmatrix}$$

と分解できるので，

$$s_z^2 = \lambda_1 v_1^2 + \lambda_2 v_2^2$$

と表せる．ただし，

$$v_j = u_{1j}\bar{a}_1 + u_{2j}\bar{a}_2.$$

$v_1^2 + v_2^2 = 1$ が成り立つので，したがって，$w_1^2 = 1$, $w_2^2 = 0$ のとき，s_z^2 は最大値 λ_1 をとることがわかる．このとき，$(\bar{a}_1, \bar{a}_2) = \pm(u_{11}, u_{21})$ が成り立つ．したがって，分散が大きな総合指標は

$$z_i = \pm(u_{11}x_{i1} + u_{21}x_{i2})$$

であることがわかった．

次に，z_i と相関がなく，分散が大きな総合指標

$$w_i = b_1 x_{i1} + b_2 x_{i2}$$

を考えることにしよう．$\bar{b}_1 = b_1/s$, $\bar{b}_2 = b_2/s$, $s = \sqrt{b_1^2 + b_2^2}$ とおいて，$s = 1$ とし，$(\bar{a}_1, \bar{a}_2) = \pm(u_{11}, u_{21})$ とすると，w_i の分散 s_w^2 と，z_i と w_i の共分散は

$$\begin{aligned} s_w^2 &= \lambda_1 (v_1')^2 + \lambda_2 (v_2')^2, \\ s_{zw} &= \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(w_i - \bar{w}) \\ &= a_1 b_1 s_{11} + (a_1 b_2 + a_2 b_1) s_{12} + a_2 b_2 s_{22} \\ &= \lambda_1 v_1' \end{aligned}$$

となる．ただし，

$$v_j' = u_{1j}\bar{b}_1 + u_{2j}\bar{b}_2.$$

したがって，$s_{zw} = 0$ となるのは，$(\bar{b}_1, \bar{b}_2) = \pm(u_{12}, u_{22})$ であり，その時，$s_w^2 = \lambda_2$ となる．したがって，z_i と相関がなく，分散が大きな指標は

$$w_i = \pm(u_{12}x_{i1} + u_{22}x_{i2})$$

である．

3.3 一般の次元では

2 次元の時と同様にすると，

$$s_z^2 = [a_1, \ldots, a_p] \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}$$

と表すとことができて，

$$\bar{a}_i = \frac{a_i}{r}, \quad r = \sqrt{a_1^2 + \cdots + a_p^2}$$

とおくと，

$$s_z^2 = r^2 \, [\bar{a}_1, \ldots, \bar{a}_p] \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix} \begin{bmatrix} \bar{a}_1 \\ \vdots \\ \bar{a}_p \end{bmatrix}$$

となる．したがって，$r = 1$ として，$\bar{a}_i$ の値を求めることにする．いま，$\lambda_1 \geqq \cdots \geqq \lambda_p \geqq 0$ なる λ_j と

$$\sum_{j=1}^{p} u_{jk} u_{jl} = \begin{cases} 1 & (k = l) \\ 0 & (\text{その他の時}) \end{cases}$$

を満たす u_{jk} を適当に選ぶと,,

$$\begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix} = \begin{bmatrix} u_{11} & \cdots & u_{1p} \\ \vdots & \ddots & \vdots \\ u_{p1} & \cdots & u_{pp} \end{bmatrix} \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix} \begin{bmatrix} u_{11} & \cdots & u_{p1} \\ \vdots & \ddots & \vdots \\ u_{1p} & \cdots & u_{pp} \end{bmatrix}$$

と分解できる．したがって，

$$s_z^2 = \lambda_1 v_1^2 + \cdots + \lambda_p v_p^2$$

と表せる．ただし，

$$v_j = u_{1j} \bar{a}_1 + \cdots + u_{pj} \bar{a}_p.$$

$v_1^2 + \cdots + v_p^2 = 1$ が成り立つので，$v_1^2 = 1, v_2^2 = \cdots = v_p^2 = 0$ のとき，s_z^2 が最大値 λ_1 をとることがわかる．このとき，$(\bar{a}_1, \ldots, \bar{a}_p) = \pm(u_{11}, u_{21}, \ldots, u_{p1})$ となる．したがって，分散が大きな総合指標は

$$z_i = \pm(u_{11} x_{i1} + u_{21} x_{i2} + \cdots + u_{p1} x_{ip})$$

であることがわかる．

2 次元の時と同様に考えると，z_i と相関がなく分散が大きな総合指標は

$$w_i = \pm(u_{12} x_{i1} + u_{22} x_{i2} + \cdots + u_{p2} x_{ip})$$

であり，その分散 s_w^2 は λ_2 であることがわかる．

3.4 色々な統計量

最も大きな分散を持つ総合指標

$$z = a_1x_1 + \cdots + a_px_p$$

を第 1 **主成分** (component)，それと相関のなく，2 番目に大きな分散を持つ総合指標

$$w_i = b_1x_1 + \cdots + b_px_p$$

を第 2 主成分と呼ぶ．以下同様に，第 1 主成分から第 $k-1$ 主成分までと相関がなく，k 番目に大きな分散を持つ総合指標を第 k 主成分とよぶ．

データ

$$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$$

に対する第 1 主成分の値

$$z_i = a_1x_{i1} + \cdots + a_px_{ip}$$

を第 1 **主成分得点** (component score)，または，**第 1 主成分スコア**という．**第** k **主成分得点**も同様に定義する．

分散と共分散を並べた

$$S_{\boldsymbol{x}} = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix}$$

を**分散共分散行列**と呼ぶが，その固有値は全て正の値をとり，それを大きさの順に並べて，$\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_p$ とする．λ_k の固有ベクトルを

$$\boldsymbol{u}_k = [u_{1k}, \cdots, u_{pk}]$$

とすると，第 k 主成分は

$$u_{1k}x_1 + \cdots + u_{pk}x_p$$

と表され，また，第 k 主成分得点の分散が λ_k であった．分散 s_{jj} と固有値 λ_k には

$$s_{11} + \cdots + s_{pp} = \lambda_1 + \cdots + \lambda_p$$

という関係があるが，両辺の値を T と置き，λ_k/T, $(\lambda_1 + \cdots + \lambda_k)/T$ をそれぞれ，第 k 主成分の**寄与率** (contribution)，第 k 主成分までの**累積寄与率** (cumulative contribution) と呼ぶ．累積寄与率が一定の値以上となる主成分までを意味のあるものと考えることが一般的であるが，よく用いられる基準は，0.7 ∼ 0.9 以上，また，固有値の平均値以上 (**カイザー基準** (Kaiser Criterion)) である．

さらに，主成分と元の変数の相関係数を**因子負荷量** (factor loading) と呼び，主成分の意味や性質を考える上で重要な指標である．

3.5 標準化変量に対する主成分

例えば，第 1 主成分の重み $a_1, \ldots, a_p$ の大きさは，各変数の第 1 主成分に対する重要度を表すものと考えられるが，変量の単位を取り替えると，重みも変化する．したがって，重みを比較するうえでは，変量を標準化してから，主成分を考えることがある．

第 j 変量の平均と標準偏差は

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad \sqrt{s_{jj}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$$

であった．これらを用いて，x_{ij} を標準化すると

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_{jj}}}$$

となる．この y_{ij} に対して，

$$z_i = a_1 y_{i1} + \cdots + a_p y_{ip}$$

のように表される主成分を考えると，その重み $a_1, \ldots, a_p$ の大きさに意味がある．

標準化変量 y_{ij} の平均は 0, 分散は 1 であり，y_j と $y_{j'}$ の共分散は x_j と $x_{j'}$ の相関係数になる．したがって，標準化変量 y_{ij} を用いた主成分分析では，標本相関係数のかわりに，標本相関行列

$$R = \begin{bmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{bmatrix}, \quad r_{jj} = 1$$

を用いることになる．ここで，r_{jk} は第 j 変量 (x_j) と第 k 変量 (x_k) の相関係数である．R の固有値も全て正の値をとり，それらを大きさの順にならべたものを $\lambda_1 \geqq \cdots \geqq \lambda_p$ として，λ_k に対応する固有ベクトルを

$$\boldsymbol{u}_k = [u_{1k}, \ldots, u_{pk}]$$

とすると，第 k 主成分が

$$u_{1k} y_1 + \cdots + u_{pk} y_p$$

であり，その分散が λ_k となる．$\lambda_1 + \cdots + \lambda_p = 1 + \cdots + 1 = p$ が成り立つので，第 k 主成分の寄与率は λ_k/p, 累積寄与率は $(\lambda_1 + \cdots + \lambda_k)/p$ となる．

3.6 中学 2 年生学力試験

表 **1.1**(p.15) のデータの主成分を求めてみる．分散共分散行列 S は，

$$S = \begin{bmatrix} 239.94 & 193.698 & 168.147 & 156.134 & 45.4631 \\ 193.698 & 360.662 & 214.433 & 178.062 & 59.9244 \\ 168.147 & 214.433 & 239.244 & 163.344 & 53.0151 \\ 156.134 & 178.062 & 163.344 & 326.82 & 14.2836 \\ 45.4631 & 59.9244 & 53.0151 & 14.2836 & 24.9527 \end{bmatrix}$$

であり，固有値と固有ベクトルを表にすると，

i	λ_i	寄与率	u_{1i}	u_{2i}	u_{3i}	u_{4i}	u_{5i}
1	845.03	0.709	-0.448	-0.578	-0.469	-0.484	-0.106
2	176.21	0.148	-0.101	-0.471	-0.149	0.842	-0.19
3	95.17	0.08	0.753	-0.595	0.189	-0.189	0.085
4	67.02	0.056	-0.463	-0.293	0.828	-0.048	0.11
5	8.19	0.007	-0.082	-0.071	-0.191	0.135	0.966

となる．主成分得点を求めると，次の表のようになる．

No	国語	社会	数学	理科	英語	平均	第 1 主成分	第 2 主成分
1	29	33	55	79	84	125.2	105	-23.9
2	71	68	72	64	97	166.4	146.1	14.5
3	74	91	79	76	100	187.8	170.1	17.2
4	52	56	58	60	85	139.1	120.9	5.9
5	77	92	96	88	98	201.7	185.6	10
6	60	85	66	66	88	163.2	148.2	17.1
7	81	91	73	63	95	180.2	163.6	27
8	61	84	72	78	92	173.1	157.1	8.3
9	70	75	81	67	96	174	155.3	16.3
10	53	70	73	51	92	151.6	132.8	23.8
11	69	64	96	57	97	171.3	150.8	21.9
12	87	89	90	85	100	201.7	184.3	11.6
13	83	75	96	81	98	193.6	175.1	8.5
14	76	61	67	57	86	155.2	137.4	14.8
15	87	82	78	82	97	190.5	172.9	8.5
16	77	80	78	70	94	178.4	161.1	16.1
17	38	43	45	12	96	104.6	78.9	39
18	67	73	78	67	95	169.9	151.3	14.5
19	83	77	80	67	100	182	162.2	19.2
20	47	61	56	21	95	125.2	102.8	42.2
21	70	62	88	51	96	164.1	143.3	24.7
22	81	51	63	66	92	157.9	137	3.5
23	51	16	36	48	84	105.1	81.1	-6.4
分散	240	361	239	327	25	737	844.6	176.4

演習 3.1　演習 **1.1**(p.2) の 2 変量データに対して，第 1 主成分 $z = a_1x + a_2y$ と第 2 主成分 $w = b_1x + b_2y$ を求めよ．また，第 1 主成分の寄与率を求めよ．さらに，標準化変量に対する第 1 主成分と第 2 主成分を求め，第 1 主成分の寄与率を求めよ．

演習 3.2　対称行列 $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ の固有値を λ_1, λ_2 とし，λ_i に属する固有ベクトルを $\boldsymbol{p}_i = \begin{bmatrix} p_{1i} \\ p_{2i} \end{bmatrix}$ とする．ただし，$p_{1i}^2 + p_{2i}^2 = 1$ のものだけを考えることにする．このとき，以下の問いに答えよ．ただし，a, b, c は実数である．

(1)　固有方程式を求め，その解である固有値 λ_1, λ_2 はともに実数であることを示せ．

(2) $\boldsymbol{q}_i = \begin{bmatrix} a-\lambda_i \\ b \end{bmatrix}$ とすると，$\boldsymbol{p}_i$ と $\boldsymbol{q}_i$ が垂直であることを示せ．また，$\boldsymbol{q}_1$ と $\boldsymbol{q}_2$ が垂直であることを示せ（ヒント：固有方程式の解と係数の関係).

(3) 上の結果から，$\boldsymbol{p}_1, \boldsymbol{p}_2$ が垂直であることを示せ．さらに，$P = [\boldsymbol{p}_1, \boldsymbol{p}_2] = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$ に対して，${}^tPP = I_2$(単位行列) であることを示せ.（このことから，${}^tP = P^{-1}$ がわかる)

(4)

$$A = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{bmatrix}$$

と表せることを示せ.

統計解析ソフト R による実行例

```
> seiseki<-read.csv("seiseki.csv")
> n<-nrow(seiseki)
> seiseki.ss<-var(seiseki)*(n-1)/n
          国語       社会       数学       理科      英語
国語 239.93951 193.69754 168.14745 156.13422 45.46314
社会 193.69754 360.66163 214.43289 178.06238 59.92439
数学 168.14745 214.43289 239.24386 163.34405 53.01512
理科 156.13422 178.06238 163.34405 326.82042 14.28355
英語  45.46314  59.92439  53.01512  14.28355 24.95274

> eigen(seiseki.ss)
$values
[1] 845.030828 176.209851  95.166066  67.023864   8.187538

$vectors
           [,1]       [,2]        [,3]        [,4]        [,5]
[1,] -0.4481043  0.1012605  0.75339195  0.46321069  0.08237283
[2,] -0.5777951  0.4712322 -0.59458373  0.29255806  0.07051953
[3,] -0.4687250  0.1494315  0.18924463 -0.82794928  0.19145111
[4,] -0.4842098 -0.8421800 -0.18941761  0.04770349 -0.13460715
[5,] -0.1057972  0.1899728  0.08472421 -0.10986357 -0.96616205

> seiseki.pc<-prcomp(seiseki)
> summary(seiseki.pc)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5
Standard deviation     29.7227 13.5727 9.97456 8.37081 2.92570
Proportion of Variance  0.7091  0.1479 0.07986 0.05625 0.00687
Cumulative Proportion   0.7091  0.8570 0.93688 0.99313 1.00000
> seiseki.pc$rotation
           PC1        PC2         PC3         PC4         PC5
国語 0.4481043  0.1012605  0.75339195 -0.46321069  0.08237283
社会 0.5777951  0.4712322 -0.59458373 -0.29255806  0.07051953
数学 0.4687250  0.1494315  0.18924463  0.82794928  0.19145111
理科 0.4842098 -0.8421800 -0.18941761 -0.04770349 -0.13460715
英語 0.1057972  0.1899728  0.08472421  0.10986357 -0.96616205
```

4 判別分析

4.1 目的

右の表は，ある高校の 3 年生に対して去年実施した実力試験の結果 (3 教科合計) であり，その後の大学受験において合格したものと不合格であったものを群分けしたものである．この情報を利用して，今年度の実力試験の結果より，生徒が合格するかどうかを判定したいとする．どのような判定基準を作ればいいであろうか？

一般に，複数の群 $\Pi_1, \Pi_2, \ldots, \Pi_m$ のいずれかから得られた p 次元データ

$$\boldsymbol{z} = (z_1, \ldots, z_p)$$

が，どの群に属するかを判定する統計手法を**判別分析** (discriminant analysis) と呼ぶ．各群からの距離と判別したいデータの距離を測り，もっとも近い群が，属する群で判定するのが素朴な方法であるが，どのように距離を測ればいいのかを順に説明する．

表 **2.1** 高 3 実力試験 (合格・不合格群別)

不合格群		合格群	
No	合計	No	合計
1	150	1	190
2	127	2	168
3	160	3	215
4	162	4	204
5	129	5	171
6	175	6	226
7	136	7	150
8	198	8	245
9	140	9	178
10	147	10	203
11	126	11	157
12	123	12	205
13	131	13	168
14	128	14	152
15	114	15	165
16	146	16	159
17	180	17	182
18	122	18	199
19	153	19	151
20	177	20	179
平均	146.2	平均	183.4
分散	502	分散	681
標準偏差	22.4	標準偏差	26.1

4.2 1 変量の場合

簡単のために $m = 2$，$p = 1$ の場合について考える．つまり，判別したいデータがスカラー z で，それが 2 つの群 Π_1, Π_2 のどちらから出現したのかを判定することにする．Π_i の平均と分散を μ_i, σ_i^2 として，$\mu_1 < \mu_2$ と仮定する．

このとき，データ z と群 Π_i の距離を z と平均の距離 $z - \mu_i$ で測るのが自然であるが，分散の違いを考慮して $\dfrac{z-\mu_i}{\sigma_i}$，あるいは，

$$D^2(z, \Pi_i) = \frac{(z-\mu_i)^2}{\sigma_i^2}$$

で測るのがより適当であると考えられる．したがって，

$$\begin{cases} D^2(z,\Pi_1) < D^2(z,\Pi_2) \;\Rightarrow\; z \text{ は } \Pi_1 \text{ に属する} \\ D^2(z,\Pi_1) > D^2(z,\Pi_2) \;\Rightarrow\; z \text{ は } \Pi_2 \text{ に属する} \end{cases}$$

と判定するのが妥当であろう．この判定基準を $\sigma_1 = \sigma_2$ の場合と $\sigma_1 \neq \sigma_2$ の場合に分けて整理する．

$\underline{\sigma_1 = \sigma_2 =: \sigma \text{ の場合}}$

$$\begin{aligned} D^2(z,\Pi_2) - D^2(z,\Pi_1) &= \frac{-2\mu_2 z + \mu_2^2 + 2\mu_1 z - \mu_1^2}{\sigma^2} \\ &= -\frac{2(\mu_2-\mu_1)}{\sigma^2}\left(z - \frac{\mu_1+\mu_2}{2}\right) \end{aligned}$$

であり，$\mu_1 < \mu_2$ を仮定しているので，

$$\begin{cases} D^2(z,\Pi_1) < D^2(z,\Pi_2) & \Longleftrightarrow \quad z < \frac{\mu_1+\mu_2}{2} =: \bar{\mu} \\ D^2(z,\Pi_1) > D^2(z,\Pi_2) & \Longleftrightarrow \quad z > \frac{\mu_1+\mu_2}{2} =: \bar{\mu} \end{cases}$$

となる．つまり，平均の中央の値 $\bar{\mu}$ の値より小さい時，平均が小さい Π_1 に，$\bar{\mu}$ より大きい時，平均が大きい Π_2 に属すると判定することになる：

$$\begin{cases} z < \bar{\mu} \ \Rightarrow \ z \text{ は } \Pi_1 \text{ に属する} \\ z > \bar{\mu} \ \Rightarrow \ z \text{ は } \Pi_2 \text{ に属する} \end{cases}$$

$\sigma_1 > \sigma_2$ の場合

$$D^2(z,\Pi_2) - D^2(z,\Pi_1) = \frac{(\sigma_1+\sigma_2)(\sigma_1-\sigma_2)}{\sigma_1^2\sigma_2^2}\,(z-c_1)\,(z-c_2)$$

ただし，

$$c_1 = \frac{\sigma_1\mu_2 + \sigma_2\mu_1}{\sigma_1+\sigma_2},\ c_2 = \frac{\sigma_1\mu_2 - \sigma_2\mu_1}{\sigma_1-\sigma_2}.$$

このとき，

$$c_1 - c_2 = \frac{2\sigma_1\sigma_2(\mu_1-\mu_2)}{(\sigma_1+\sigma_2)(\sigma_1-\sigma_2)}$$

なので，$\sigma_1 > \sigma_2$, $\mu_1 < \mu_2$ のときは，$c_1 < c_2$．したがって，

$$\begin{cases} c_1 < z < c_2 \ \Rightarrow \ z \text{ は } \Pi_2 \text{ に属する} \\ z < c_1, c_2 < z \ \Rightarrow \ z \text{ は } \Pi_1 \text{ に属する} \end{cases}$$

となる．

$\sigma_1 < \sigma_2$ の場合．　$\sigma_1 > \sigma_2$ の場合と同様にすると，$c_2 < c_1$ であり，

$$\begin{cases} c_2 < z < c_1 \ \Rightarrow \ z \text{ は } \Pi_1 \text{ に属する} \\ z < c_2, c_1 < z \ \Rightarrow \ z \text{ は } \Pi_2 \text{ に属する} \end{cases}$$

となる．

表 **2.1**(p.23) のデータから，μ_1, μ_2 の推定値を求めると $\hat{\mu}_1 = 146.2$, $\hat{\mu}_2 = 183.35$ であり，σ_1^2, σ_2^2 の推定値は $\hat{\sigma}_1^2 = 528.5895$, $\hat{\sigma}_2^2 = 717.3974$ である．$\sigma_1 = \sigma_2$ であるかどうかは，不偏分散の比 U_1^2/U_2^2 が自由度 (n_1-1, n_2-1) のエフ分布の上側 $100(\alpha/2)\%$ 以下，下側 $100(\alpha/2)\%$ 点以上であれば $\sigma_1 = \sigma_2$，そうでなければ，$\sigma_1 \neq \sigma_2$ と判定すればよいが，このデータでは，$U_1^2/U_2^2 = 528.6/717.4 = 0.737$, $n_1 = n_2 = 20$, 上側 2.5% 点は 2.53, 下側 2.5% 点は 0.395 なので，　有意水準 5% では $\sigma_1 = \sigma_2$ と判定される．したがって，$\bar{\mu}$ の推定値 $\hat{\mu} = 164.8$ より大きいかどうかで判定すればよいことがわかる．

4.3　2 変量の場合

$m = 2$, $p = 2$ の場合を考える．群 Π_j のから

$$\begin{bmatrix} x_{j1} \\ y_{j1} \end{bmatrix},\ \cdots,\ \begin{bmatrix} x_{jn_j} \\ y_{jn_j} \end{bmatrix}$$

のデータが得られているとする．このとき，標本平均ベクトルと標本分散共分散行列は次のように定義される．

$$\begin{bmatrix} \bar{x}_j \\ \bar{y}_j \end{bmatrix}, \quad \begin{bmatrix} s_{j,xx} & s_{j,xy} \\ s_{j,xy} & s_{j,yy} \end{bmatrix}.$$

ここで，$\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ji}$, $\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ji}$, $s_{j,xx} = \frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ji}-\bar{x}_j)^2$,
$s_{j,xy} = \frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ji}-\bar{x}_j)(y_{ji}-\bar{y}_j)$，$s_{j,yy} = \frac{1}{n_j}\sum_{i=1}^{n_j}(y_{ji}-\bar{y}_j)^2$ である．

群 Π_i の第 1 主成分得点を $z_{ji} = u_{j1}x_{ji} + u_{j2}y_{ji}$, 第 2 主成分得点を $w_{ji} = v_{j1}x_{ji} + v_{j2}y_{ji}$ とおくと，$(z_{j1}, w_{j1}), \ldots, (z_{jn_j}, w_{jn_j})$ は無相関であり，それぞれの分散 $s_{j,zz}$, $s_{j,ww}$ は標本分散共分散行列の固有値 λ_j, μ_j であった．もちろん，(u_{j1}, u_{j2}) は λ_j に対応する長さ 1 の固有ベクトルであり，(v_{j1}, v_{j2}) は μ に対応する長さ 1 の固有ベクトルである．また，

$$\begin{bmatrix} s_{j,xx} & s_{j,xy} \\ s_{j,xy} & s_{j,yy} \end{bmatrix} = \begin{bmatrix} u_{j1} & v_{j1} \\ u_{j2} & v_{j2} \end{bmatrix} \begin{bmatrix} \lambda_j & 0 \\ 0 & \mu_j \end{bmatrix} \begin{bmatrix} u_{j1} & u_{j2} \\ v_{j1} & v_{j2} \end{bmatrix}$$

が成り立つ．

$$\begin{bmatrix} u_{j1} & u_{j2} \\ v_{j1} & v_{j2} \end{bmatrix} = \begin{bmatrix} u_{j1} & v_{j1} \\ u_{j2} & v_{j2} \end{bmatrix}^{-1}$$

なので，

$$\begin{bmatrix} s_{j,xx} & s_{j,xy} \\ s_{j,xy} & s_{j,yy} \end{bmatrix}^{-1} = \begin{bmatrix} u_{j1} & v_{j1} \\ u_{j2} & v_{j2} \end{bmatrix} \begin{bmatrix} 1/\lambda_j & 0 \\ 0 & 1/\mu_j \end{bmatrix} \begin{bmatrix} u_{j1} & u_{j2} \\ v_{j1} & v_{j2} \end{bmatrix} \cdots ①$$

も成り立つことに注意しよう．

さて，データ $\begin{bmatrix} x \\ y \end{bmatrix}$ の群 Π の中心までの距離として二つの主成分を利用しよう．第 1 主成分の平均は $u_{j1}\bar{x}_j + u_{u2}\bar{y}_j$ であり，データの第 1 主成分は $u_{j1}x + u_{j2}y$ だから，第 1 主成分の分散 λ_j で標準化して，

$$\left(\frac{u_{j1}x + u_{j2}y - (u_{j1}\bar{x}_j + u_{j2}\bar{y}_j)}{\sqrt{\lambda_j}}\right)^2 = \frac{\{u_{j1}(x-\bar{x}_j) + u_{j2}(y-\bar{y}_j)\}^2}{\lambda_j}$$

が第 1 主成分の平均までの距離である．同様に，第 2 主成分の平均までの距離を考えて，それらを加えると

$$D^2(x, y, \Pi_i) := \frac{\{u_{j1}(x-\bar{x}_j) + u_{j2}(y-\bar{y}_j)\}^2}{\lambda_j} + \frac{\{v_{j1}(x-\bar{x}_j) + v_{j2}(y-\bar{y}_j)\}^2}{\mu_j}$$

となる．第 1 主成分と第 2 主成分は相関がないので，標準化した距離の 2 乗を単純に加えてよいと考えられる．$\tilde{x}_j = x - \bar{x}_j$, $\tilde{y}_j = y - \bar{y}_j$ とおいて，$D^2(x, y, \Pi_i)$ を行列を用いて表現すると，

$$\begin{aligned} D^2(x, y, \Pi_i) =& [u_{j1}\tilde{x}_j + u_{j2}\tilde{y}_j, v_{j1}\tilde{x}_j + v_{j2}\tilde{y}_j] \begin{bmatrix} \frac{1}{\lambda_j} & 0 \\ 0 & \frac{1}{\mu_j} \end{bmatrix} \begin{bmatrix} u_{j1}\tilde{x}_j + u_{j2}\tilde{y}_j \\ v_{j1}\tilde{x}_j + v_{j2}\tilde{y}_j \end{bmatrix} \\ =& [\tilde{x}_j, \tilde{y}_j] \begin{bmatrix} u_{j1} & v_{j1} \\ u_{j2} & v_{j2} \end{bmatrix} \begin{bmatrix} \frac{1}{\lambda_j} & 0 \\ 0 & \frac{1}{\mu_j} \end{bmatrix} \begin{bmatrix} u_{j1} & u_{j2} \\ v_{j1} & v_{j2} \end{bmatrix} \begin{bmatrix} \tilde{x}_j \\ \tilde{y}_j \end{bmatrix}. \end{aligned}$$

① を用いると，

$$D^2(x,y,\Pi_i)=[x-\bar{x}_j,y-\bar{y}_j]\begin{bmatrix} s_{j,xx} & s_{j,xy} \\ s_{j,xy} & s_{j,yy} \end{bmatrix}^{-1}\begin{bmatrix} x-\bar{x}_j \\ y-\bar{y}_j \end{bmatrix}.$$

このように定義される距離をマハラノビスの距離と呼ぶ．

4.4 多変量の場合

群 Π_i からの p 変量データが n_j 個

$$\boldsymbol{x}_{j1}=\begin{bmatrix} x_{j,11} \\ \vdots \\ x_{j,p1} \end{bmatrix},\ldots,\boldsymbol{x}_{jn_j}=\begin{bmatrix} x_{j,1n_j} \\ \vdots \\ x_{j,pn_j} \end{bmatrix}$$

が与えられている時，その平均 $\bar{x}_{jk}=\frac{1}{n}\sum_{s=1}^{n_j}x_{j,ks}$，と共分散 $s_{j,kk'}=\frac{1}{n}\sum_{s=1}^{n_j}(x_{j,ks}-\bar{x}_{jk})(x_{j,k's}-\bar{x}_{jk'})$ と表し，群 Π_j の平均ベクトル $\bar{\boldsymbol{x}}_j$ を第 k 成分が $\bar{x}_{jk}$ である縦ベクトル，群 Π_j の標本分散共分散行列 S_j を (k,k') 成分が $s_{j,kk'}$ である $p\times p$ 行列とする．このとき，$\boldsymbol{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$ と群 Π_j の距離を

$$D^2(\boldsymbol{x},\Pi_j)=(\boldsymbol{x}-\bar{\boldsymbol{x}}_j)'S_j^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}_j)$$

で測る．この距離をマハラノビスの距離という．

演習 4.1 $n_1=10$, $n_2=12$, $(\bar{x}_{1\cdot1},\bar{x}_{1\cdot2})=(1,2)$, $(\bar{x}_{2\cdot1},\bar{x}_{2\cdot2})=(2,1)$ で

$$S_1=\frac{1}{10}\begin{bmatrix} 4 & 1 \\ 1 & \frac{3}{4} \end{bmatrix},\quad S_2=\frac{1}{12}\begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}$$

のとき，$\boldsymbol{z}=(3,3)$ がどちらの集団からのデータであるか，マハラノビスの距離により判別せよ．

4.5 B-W 法

多変量データを，合成変量を用いて判別する方法について考えよう．簡単のために，$m=2$, $p=2$ の場合について考える．マハラノビスの距離による判別を考えたのと同様に，群 Π_j のから

$$\begin{bmatrix} x_{j1} \\ y_{j1} \end{bmatrix},\ \cdots,\ \begin{bmatrix} x_{jn_j} \\ y_{jn_j} \end{bmatrix}$$

のデータが得られているとする．このとき，群 Π_j 内部の x の平均と分散を $\bar{x}_j$, $s_{j,xx}$, y の平均んと分散を $\bar{y}_j$, $s_{j,yy}$, x と y の共分散を $s_{j,xy}$ と表すことにする．また，群で分けないで，データ全体で x, y の平均と分散，共分散を次のように定義する:

$$\bar{x}=\frac{1}{n_1+n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}x_{ji},\quad \bar{y}=\frac{1}{n_1+n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}y_{ji},\quad s_{xx}=\frac{1}{n_1+n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(x_{ji}-\bar{x})^2,$$

$$s_{yy}=\frac{1}{n_1+n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(y_{ji}-\bar{y})^2,\quad s_{xy}=\frac{1}{n_1+n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(x_{ji}-\bar{x})(y_{ji}-\bar{y}).$$

このとき，

$$\begin{aligned} s_{xx} =& \frac{1}{n_1+n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(x_{ji}-\bar{x}_j+\bar{x}_j-\bar{x})^2 \\ =& \frac{1}{n_1+n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}\left\{(x_{ji}-\bar{x}_j)^2+2(\bar{x}_j-\bar{x})(x_{ji}-\bar{x}_j)+(\bar{x}_j-\bar{x})^2\right\} \end{aligned}$$

ここで，$\sum_{i=1}^{n_j}(x_{ji}-\bar{x}_j)=0$ に注意すると

$$\begin{aligned} =& \frac{1}{n_1+n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}\left\{(x_{ji}-\bar{x}_j)^2+(\bar{x}_j-\bar{x})^2\right\} \\ =& \frac{1}{n_1+n_2}\sum_{j=1}^{2} n_j\left\{s_{j,xx}+(\bar{x}_j-\bar{x})^2\right\}. \end{aligned}$$

したがって，$W_{xx}=\sum_{j=1}^{2} n_j s_{j,xx}$, $B_{xx}=\sum_{j=1}^{2} n_j(\bar{x}_j-\bar{x})^2$ とおくと，

$$s_{xx}=\frac{1}{n_1+n_2}(W_{xx}+B_{xx})$$

と表せる．$W_{xx}=\sum_{j=1}^{2}\sum_{i=1}^{n_j}(x_{ji}-\bar{x}_j)^2$ だから，これは群内部の偏差平方和をすべての群について合計したものであり，群内部のバラツキの全体を表す．また，B_{xx} は群 Π_j のデータがすべて $\bar{x}_j$ であると考えた時の，データ全体のバラツキを表している．つまり，群内部ではばらつかず一定の値 $\bar{x}_j$ をとると考えた時のバラツキなので，群間のバラツキを表すと考えることができる．W_{xx} を x の**群内偏差平方和**，B_{xx} を x の**群間偏差平方和**と呼ぶことにする．同様に，$W_{yy}=\sum_{j=1}^{2} n_j s_{j,yy}$, $B_{yy}=\sum_{j=1}^{2} n_j(\bar{y}_j-\bar{y})^2$, $W_{xy}=\sum_{j=1}^{2} n_j s_{j,xy}$, $B_{xy}=\sum_{j=1}^{2} n_j(\bar{x}_j-\bar{x})(\bar{y}_j-\bar{y})$ と置くと，

$$s_{yy}=\frac{1}{n_1+n_2}(W_{yy}+B_{yy}),\quad s_{xy}=\frac{1}{n_1+n_2}(W_{xy}+B_{xy}).$$

W_{xy} を**群内偏差積和**，B_{xy} を**群間偏差積和**と呼ぶことにする．

さて，判別を容易にするための合成変量

$$z_{ji}=ax_{ji}+by_{ji},\quad j=1,2(=m), i=1,\ldots,n_j$$

について考えよう．x_{ji} や y_{ji} と同様に，群 Π_j 内部の平均と分散を $\bar{z}_j$, $s_{j,zz}$ と表し，群で分けないで全体にわたる平均を $\bar{z}$, 分散を s_{zz} と表すことにする．さらに，$W_{zz}=\sum_{j=1}^{2} n_j s_{j,zz}$, $B_{zz}=\sum_{j=1}^{2} n_j(\bar{z}_j-\bar{z})^2$ とおくと，

$$s_{zz}=\frac{1}{n_1+n_2}(W_{zz}+B_{zz})$$

となる．判別を容易にするためには，各群がバラツキが小さく，群間の違いがはっきりすることが望ましいので，群内偏差平方和 W_{zz} が小さく，群間偏差平方和 B_{zz} が大きいほうがよい．そこで，

$$R(a,b):=\frac{B_{zz}}{W_{zz}}$$

として，$R(a,b)$ を最大にする (a,b) を求めることにしよう．

主成分分析の時と同様に，

$$W_{zz}=[a,b]W\begin{bmatrix} a \\ b \end{bmatrix},\quad B_{zz}=[a,b]B\begin{bmatrix} a \\ b \end{bmatrix}$$

と 2 次形式で表すことにする．ここで，

$$W = \begin{bmatrix} W_{xx} & W_{xy} \\ W_{xy} & W_{yy} \end{bmatrix}, \quad B = \begin{bmatrix} B_{xx} & B_{xy} \\ B_{xy} & B_{yy} \end{bmatrix}.$$

W は対称行列で固有値が正なので，W の長さが 1 である固有ベクトルを列ベクトルとする正方行列を P_W，固有値の平方根を対角成分に持つ対角行列を $\Lambda_W^{1/2}$ とおき，$W^{1/2} = P_W \Lambda_W^{1/2} P_W'$ とすると，$W = W^{1/2} W^{1/2}$ が成り立つ．したがって，

$$\boldsymbol{\gamma} := \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = W^{1/2} \begin{bmatrix} a \\ b \end{bmatrix}$$

とおくと，

$$W_{zz} = \boldsymbol{\gamma}'\boldsymbol{\gamma}, \quad B_{zz} = \boldsymbol{\gamma}' W^{-1/2} B W^{-1/2} \boldsymbol{\gamma}.$$

と表せる．ここで，$W^{-1/2} = P_W \Lambda_W^{-1/2} P_W'$ であり，$\Lambda_W^{-1/2}$ は W の固有値の $-1/2$ 乗を対角成分とする対角行列である．こうすると，$W^{-1/2}$ は $W^{1/2}$ の逆行列であることもわかる．さらに，$\tilde{\boldsymbol{\gamma}} = \frac{1}{||\boldsymbol{\gamma}||}\boldsymbol{\gamma}$ とおくと，$||\tilde{\boldsymbol{\gamma}}|| = 1$ であり，

$$R(a, b) = \tilde{\boldsymbol{\gamma}}' W^{-1/2} B W^{-1/2} \tilde{\boldsymbol{\gamma}}.$$

この表現より，$\tilde{\boldsymbol{\gamma}}$ が $W^{-1/2} B W^{-1/2}$ の最大固有値に対応する固有ベクトルであるとき，$R(a, b)$ は最大となることがわかる．$W^{-1/2} B W^{-1/2}$ の固有値を λ，それに対応する固有ベクトルを $\boldsymbol{q}$ とすると，$W^{-1/2} B W^{-1/2} \boldsymbol{q} = \lambda \boldsymbol{q}$ が成り立つ，両辺左から $W^{-1/2}$ をかけるて，$\boldsymbol{h} = W^{-1/2}\boldsymbol{q}$ とおくと，$W^{-1} B \boldsymbol{h} = \lambda \boldsymbol{h}$．つまり，$\lambda$ が $W^{-1}B$ の固有値であるとき，λ はまた $W^{-1/2} B W^{-1/2}$ の固有値でもある．λ に対応する $W^{-1}B$ の固有ベクトルが $\boldsymbol{h}$ である時，λ に対応する $W^{-1/2} B W^{-1/2}$ の固有ベクトルは $\boldsymbol{q} = W^{1/2}\boldsymbol{h}$ であることもわかる．そこで，$W^{-1}B$ の最大固有値を λ_1，それに対応する $W^{-1}B$ の固有ベクトルを $\boldsymbol{h}_1$ とすると，$\tilde{\boldsymbol{\gamma}} = W^{1/2}\boldsymbol{h}_1$ のとき，$R(a, b)$ は最大となることがわかる．したがって，

$$\begin{bmatrix} a \\ b \end{bmatrix} = W^{-1/2} \tilde{\boldsymbol{\gamma}} = \boldsymbol{h}_1$$

のとき，$R(a, b)$ が最大値 λ_1 をとることがわかる．

以上をまとめると，群間偏差平方和 B_{zz} を大きくし，群内偏差平方和 W_{zz} を小さくする合成変量は，$W^{-1}B$ の最大固有値 λ_1 に対応する固有ベクトルを $\boldsymbol{h}_1$ を用いて，

$$z_{ji} = a_1 x_{ji} + b_1 y_{ji}, \quad \boldsymbol{h}_1 = \begin{bmatrix} a_1 \\ b_1 \end{bmatrix}$$

と表される．またその時，B_{zz}/W_{zz} は最大値 λ_1 をとる．

一般に，群 Π_j からのデータ，

$$\boldsymbol{x}_{j1} = \begin{bmatrix} x_{j11} \\ \vdots \\ x_{j1p} \end{bmatrix}, \ldots, \boldsymbol{x}_{jn_j} = \begin{bmatrix} x_{jn_j1} \\ \vdots \\ x_{jn_jp} \end{bmatrix}$$

が与えれれている時，群 Π_j 内部の平均，分散，共分散を $\bar{x}_{jk}$, $s_{j,kk}$, $s_{j,kk'}$ と表し，群で分けないでデータ全体にわたる平均，分散，共分散を $\bar{x}_k$, s_{kk}, $s_{kk'}$ とすると，

$$s_{kk'} = \frac{1}{n_1 + n_2} \sum_{j=1}^{2} n_j \left\{ s_{j,kk'} + (\bar{x}_{jk} - \bar{x}_k)(\bar{x}_{jk'} - \bar{x}_{k'}) \right\}$$

が成り立つ．(k,k') 成分を $\sum_{j=1}^{2} n_j s_{j,kk'} = \sum_{j=1}^{2}\sum_{i=1}^{n_j}(x_{jik}-\bar{x}_{jk})(x_{jik'}-\bar{x}_{jk'})$ とする $p\times p$ 行列を W と表す．また，(k,k') 成分が $\sum_{j=1}^{2} n_j(\bar{x}_{jk}-\bar{x}_k)(\bar{x}_{jk'}-\bar{x}_{k'})$ である $p\times p$ 行列を B と表す．この時，群間のバラツキを大きく，郡内のバラツキを小さくする合成変量は，$W^{-1}B$ の最大固有値 λ_1 に対応する固有ベクトル $(u_{11},\ldots,u_{p1})$　を用いて，

$$z_{ji} = u_{11}x_{ji1} + \cdots u_{p1}x_{jip}, \quad , j = 1,2, i = 1,\ldots,n_j$$

で与えれれる．主成分分析のときと同じように，1 つの合成変量で不十分な時は，第 2，第 3 の合成変量を用いることもある．

以上をまとめると，次のようになる．$W^{-1}B$ の固有値を $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_p \geqq 0$ とし，λ_t に対応する固有ベクトルを $\boldsymbol{u}_t = (u_{1t},\ldots,u_{pt})$ とする．ただし，$||\boldsymbol{u}_t|| = 1$．母集団 Π_1, Π_2 のいずれかから出現したデータ $\boldsymbol{z} = (z_1,\ldots,z_p)'$ に対して，$\zeta_t = u_{1t}z_1 + \cdots + u_{pt}z_p$ を求め，さらに，

$$D^2_{BW}(\boldsymbol{z},\Pi_j) = \sum_{t=1}^{m}(\zeta_t - \bar{\xi}_{jt})^2, \quad \bar{\xi}_{jt} = u_{1t}\bar{x}_{j1} + \cdots + u_{pt}\bar{x}_{jp}$$

を求める．$D^2_{BW}(\boldsymbol{z},\Pi_1) < D^2_{BW}(\boldsymbol{z},\Pi_2)$ なら $\boldsymbol{z}$ は Π_1 から得られたと判定し，そうでなければ Π_2 から得られたと判定する．ただし，m は

$$\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p}$$

が一定割合，たとえば，0.9 以上になるように選ぶ．

演習 4.2

(1) 任意の u_1, u_2 に対して，$2u_1^2 + 2u_1u_2 + 4u_2^2 = [u_1,u_2]\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ を満たす $b_{ij}, i,j = 1,2$ を求めよ．ただし，$b_{12} = b_{21}$ とする．

(2) 任意の u_1, u_2 に対して，$2u_1^2 + 2u_1u_2 + u_2^2 = [u_1,u_2]\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ を満たす $w_{ij}, i,j = 1,2$ を求めよ．ただし，$w_{12} = w_{21}$ とする．

(3) $V = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}^{-1}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$ の固有値と固有ベクトルを求めよ．

(4) $\dfrac{2u_1^2 + 2u_1u_2 + 4u_2^2}{2u_1^2 + 2u_1u_2 + u_2^2}$ を最大にする u_1, u_2 の値と最大値を求めよ．

(5) $\bar{x}_{1\cdot1} = 2$, $\bar{x}_{1\cdot2} = 1$, $\bar{x}_{2\cdot1} = 1$, $\bar{x}_{2\cdot2} = 2$ の時，$\boldsymbol{z} = (3,3)$ を $B - W$ 法で判別せよ．ただし，B は (1) の b_{ij} を成分とする行列，W は (2) の w_{ij} を成分とする行列.

4.6 ベイズ法による判別

$\boldsymbol{z}$ が Π_g から出現する確率を p_g とし，$\boldsymbol{z}$ が Π_g から出現したにもかかわらず $\Pi_{g'}$ から出現したと判別するときのコストを $c(g'|g)$ とする．このとき，

$\boldsymbol{z} \in D_g$ ならば，$\boldsymbol{z}$ は Π_g から出現した

と判定する．ただし，$D_1 = \{\boldsymbol{v} \mid p_1 c(2|1) f_1(\boldsymbol{v}) > p_2 c(1|2) f_2(\boldsymbol{v})\}$, $D_2 = D_1^c$, であり，f_g は Π_g の確率密度関数．この判別法では，

$$p_1 c(2|1) \int_{D_2} f_1(\boldsymbol{v}) d\boldsymbol{v} + p_2 c(1|2) \int_{D_1} f_2(\boldsymbol{v}) d\boldsymbol{v}$$

を最小にすることに注意しよう．特に，$\pi_1 c(2|1) = \pi_2 c(1|2)$ であり，f_g が正規分布であるとき，

$$D_1 = \{\boldsymbol{v} \mid MD_1(\boldsymbol{v}) - MD_2(\boldsymbol{v}) < \log|\Sigma_2| - \log|\Sigma_1|\}$$

となり，さらに $\Sigma_1 = \Sigma_2 = \Sigma$ のときは，

$$D_1 = \{\boldsymbol{v} \mid (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \Sigma^{-1} (\boldsymbol{v} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}) > 0\}$$

また，f_g か未知のときは，Π_g の教師データから f_g の推定量を構成して，それを用いる．

4.7 誤判別確率

$\boldsymbol{x}$ が母集団 Π_j から出現したにもかかわらず，$\Pi_{j'}$ から出現したと判別する確率を $e(j'|j)$ として，

$$e(1|2) + e(2|1)$$

を**誤判別確率**と呼ぶ．また，$e(j'|j)$ を表にしたものを**判別表**と呼ぶ．$e(j'|j)$ は群 Π_j, $\Pi_{j'}$ の確率分布に依存するので，推定する必要があるが，もっとも簡便な推定量は，

$$\hat{e}(j'|j) = \frac{n(j'|j)}{n_j}$$

である．ただし，$n(j'|j)$ は，Π_j 群のデータ $\boldsymbol{x}_{ji}$, $i = 1, \ldots, n_j$ を何らかの方法で判別した時，誤って $\Pi_{j'}$ に判別されるデータの個数を表すものとする．

5 クラスター分析

5.1 目的

n 個の p 変量データ $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots, n$ をある類似度 (距離) を用いて，いくつかのグループ (クラスター) に分けること．

5.2 データの距離とクラスターの距離

(1) データの距離： p 変量データ $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ と $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jp})'$ の距離を $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ で表す．

(a) **ユークリッド距離** $d(\boldsymbol{x}_i, \boldsymbol{x}_j) := \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2}$

(b) **マンハッタン距離** $d(\boldsymbol{x}_i, \boldsymbol{x}_j) := |x_{i1} - x_{j1}| + \cdots + |x_{ip} - x_{jp}|$

(2) クラスター間の距離： クラスター C_k と C_l には，それぞれ，n_k 個の p 変量データ $\boldsymbol{x}_{ki} = (x_{ki1}, \ldots, x_{kip})'$, $i = 1, \ldots, n_k$ と n_l 個の p 変量データ $\boldsymbol{x}_{lj} = (x_{lj1}, \ldots, x_{ljp})'$, $j = 1, \ldots, n_l$ が存在するとき，クラスター間の距離を D_{kl} で表す．

(a) **単連結法 (最短距離法)** $D_{kl} := \min\{d(\boldsymbol{x}_{ki}, \boldsymbol{x}_{lj})\ ;\ i = 1, \ldots, n_k, j = 1, \ldots, n_l\}$

(b) **完全連結法 (最長距離法)** $D_{kl} := \max\{d(\boldsymbol{x}_{ki}, \boldsymbol{x}_{lj})\ ;\ i = 1, \ldots, n_k, j = 1, \ldots, n_l\}$

(c) **ウォード法** $D_{kl} := \dfrac{n_k n_l}{n_k + n_l} \displaystyle\sum_{u=1}^{p} (\bar{x}_{k\cdot u} - \bar{x}_{l\cdot u})^2$,

ただし，$\bar{x}_{k\cdot u} = \dfrac{1}{n_k} \displaystyle\sum_{i=1}^{n_k} x_{kiu}$，$\bar{x}_{l\cdot u} = \dfrac{1}{n_l} \displaystyle\sum_{j=1}^{n_l} x_{lju}$

5.3 階層的なクラスター分析

n 個の p 変量データ $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots, n$, に対して，

Step 1. k 番目のクラスター C_k には，k 番目のデータ $\boldsymbol{x}_k$ が 1 つだけ含まれるように，n 個のクラスター $C_1, \ldots, C_n$ を作る．

Step 2. クラスター間の距離 D_{kl} を，すべてのクラスターの組み合わせに対して求める．

Step 3. クラスター間の距離 D_{kl} が最小となる 2 つのクラスターを 1 つに合弁する．

Step 4. クラスターの個数が 1 つならば終了する．そうでなければ，合弁して 1 個少なくなったクラスターに対して，Step 2 へ戻る．

5.4 非階層的なクラスター分析

n 個の p 変量データ $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots, n$, に対して，

Step 1. クラスター数 k を決める．

Step 2. n 個のデータから，k 個のデータ $\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_k}$ を選び，それらを順にクラスター C_1, ..., C_k の代表点 (シード点) とする．

Step 3. n 個のデータ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ それぞれを，代表点まで最も近いクラスターに割り振る．

Step 4. 各クラスターの平均が代表点と同じならば終了する．そうでなければ，平均を代表点として，Step 3 へ戻る．

演習 5.1

$$S_k = \sum_{i=1}^{n_k}\sum_{j=1}^{p}(x_{kij} - \bar{x}_{k\cdot j})^2, \quad S_{kl} = \sum_{i=1}^{n_k}\sum_{j=1}^{p}(x_{kij} - \bar{x}_j)^2 + \sum_{i=1}^{n_l}\sum_{j=1}^{p}(x_{lij} - \bar{x}_j)^2$$

$$\bar{x}_{k\cdot j} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{kij}, \quad \bar{x}_j = \frac{1}{n_k + n_l}(n_k\bar{x}_{k\cdot j} + n_l\bar{x}_{l\cdot j})$$

のとき，

$$S_{kl} = S_k + S_l + D_{kl}$$

を示せ．ただし，D_{kl} はウォード法におけるクラスター間の距離．

演習 5.2 下の表のように距離行列が与えらる点 A, B, C,D,E を最短距離法でクラスター分析し，デンドログラムを描け．また，最長距離法でも同様にクラスター分析し，デンドログラムを描け．それらの結果が違うか確認せよ．

			A	B	C	D	E
			2	6	4	12	12
			8	8	0	3	1
A	2	8	0.0	4.0	8.2	11.2	12.2
B	6	8	4.0	0.0	8.2	7.8	9.2
C	4	0	8.2	8.2	0.0	8.5	8.1
D	12	3	11.2	7.8	8.5	0.0	2.0
E	12	1	12.2	9.2	8.1	2.0	0.0

演習 5.3 飲み物に対する嗜好調査の結果，飲み物間の距離を測ったら以下の表のようになった．最短距離法でクラスター分析を行え．

		B	C	D	E	F	G	H	I	J	K
A	ワイン	7.4	43.6	17.4	21.9	30.3	48.8	34.8	13.6	26.5	18.8
B	ビール		54.0	28.9	29.3	36.0	64.4	49.2	23.3	44.4	40.9
C	緑茶			9.3	8.5	12.8	8.9	12.6	39.9	15.4	39.9
D	紅茶				1.7	6.6	11.0	7.3	12.4	4.8	15.3
E	コーヒー					7.2	12.7	10.1	16.8	8.4	20.4
F	牛乳						7.2	5.0	16.0	7.2	25.6
G	カルピス							2.1	28.7	6.1	27.8
H	オレンジジユース								16.4	2.5	17.1
I	コーラ									12.1	10.4
J	サイダー										9.4
K	ネクター										

6 因子分析

6.1 目的

データの相関構造を少ない潜在因子で説明することが，因子分析の目的である．

たとえば，各種学力 (国語，数学，英語など) の相関関係を,「記憶力」,「論理的思考力」,「表現力」,「作業能力」などの共通する潜在因子で説明しするために利用される．

6.2 モデル

n 個の p 変量データ $\boldsymbol{z}_i = (z_{1i}, \ldots, z_{pi})'$ に対して，

$$\begin{cases} z_{1i} = & a_{11}f_{1i} + \cdots + a_{1m}f_{mi} + e_{1i} \\ \vdots & \vdots \\ z_{pi} = & a_{p1}f_{1i} + \cdots + a_{pm}f_{mi} + e_{pi} \end{cases}$$

というモデルを考える．ここで，f_{ki} を**共通因子** f_k の固体 i の**因子得点**，a_{jk} を因子 f_k の変量 z_j に対する**因子負荷量**，e_{ji} を固体 i の 変量 z_j に対する**独自因子**と呼ぶ．

すべての因子得点 f_{ki} と独自因子 e_{ji} は互いに独立であると仮定し，f_{ki} は平均 0，分散 1，e_{ji} は平均 0，分散 d_j であると仮定する．このとき，z_{ji} の分散 σ_{jj}，z_{ji} と $z_{j'i}$ の共分散 $\sigma_{jj'}$ は

$$\begin{aligned} \sigma_{jj} =& a_{j1}^2 + \cdots + a_{jm}^2 + d_j^2 \\ \sigma_{jj'} =& a_{j1}a_{j'1} + \cdots + a_{jm}a_{j'm} \end{aligned}$$

と表せる．行列を用いると,

$$\Sigma = AA' + D.$$

ただし，$\Sigma = (\sigma_{jj'})$, $A = (a_{jk})$，$D = \mathrm{diag}(d_1^2, \ldots, d_p^2)$. $h_j^2 = \sigma_{jj} - d_j^2$ とおき，h_j^2 を**共通性**と呼ぶ. z_{ji} と f_{ki} の共分散が a_{jk} であることに注意しよう．

6.3 モデルの不定性

$\boldsymbol{f}_i = (f_{1i}, \ldots, f_{mi})'$, $\boldsymbol{e}_i = (e_{1i}, \ldots, e_{pi})'$ と置くと，モデルは

$$\boldsymbol{z}_i = A\boldsymbol{f}_i + \boldsymbol{e}_i$$

と表せるが，可逆な行列 T を用いて，$A^* = AT$，$\boldsymbol{f}_i^* = T^{-1}\boldsymbol{f}_i$ とおくと，

$$\boldsymbol{z}_i = A^*\boldsymbol{f}_i^* + \boldsymbol{e}_i$$

とも表せる．つまり，A，$\boldsymbol{f}_i$ の選び方は一意ではない．

6.4 標準化

n 個の p 変量データ $\boldsymbol{z}_i = (z_{1i}, \ldots, z_{pi})'$ は標準化されているものとする：つまり，素データ $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})$ に対して，

$$z_{ji} = \frac{x_{ji} - \bar{x}_j}{s_j}, \quad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ji}, \quad s_j = \sqrt{\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2}.$$

このように標準化することで，変数の大きさや散らばりの度合いが変数間で調整される．したがって，因子負荷量の大きさが変数間で比較することができるようになる．

また，このような標準化を行うことで，$\bar{z}_j = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} z_{ji} = 0$ となるので，モデルにおいて，因子得点と独自因子の平均を 0 にしたことが正当化される．さらに，

$$s_{z_j} = \frac{1}{n}\sum_{i=1}^{n}(z_{ji} - \bar{z}_j)^2 = 1$$

であることから，$\sigma_{jj} = a_{j1}^2 + \cdots + a_{jm}^2 + d_j^2 = 1$ と仮定することも多い．この仮定の下では，因子負荷量 a_{jk} は z_{ji} と f_{ki} の相関係数であり，共通性 h_j^2 は $h_j^2 = 1 - d_j^2$ と表される．以下では，このことも仮定する．

6.5 因子数 m の決め方

$\boldsymbol{z}_i$, $i = 1, \ldots, n$ の分散共分散行列 R，つまり，素データ $\boldsymbol{x}_i$, $i = 1, \ldots, n$, の相関行列 R の固有値の中で，1 より大きい固有値の個数を m とする．この方法は主成分の個数を決める時にも用いられる．主成分や共通因子は観測値を総合する指標であると考えているので，観測値そのものより多くの情報を持っているべきである．情報の多さを分散の大きさと考えると，固有ベクトルの成分で重みづけた総合指標は，(標準化した) 観測値より多くの情報をもっていると考えられる．このことから，総合指標の個数として 1 より大きい固有値の個数とするのが適当であると考えられる（※他にも，スクリーテストや AIC など色々な方法がある).

6.6 因子負荷量の推定法

標準化変量 $\boldsymbol{z}_i$, $i = 1, \ldots, n$ の標本分散共分散行列を R として，主成分を利用した因子負荷量の推定法を説明する．因子負荷量 a_{jk} の行列 $A = (a_{jk})$ と分散共分散行列 Σ，独自因子の分散行列 $D = \mathrm{diag}(d_1^2, \ldots, d_p^2)$ には $\Sigma - D = AA'$ と言う関係が成り立ち，Σ の推定量として，R が適当であることに注意しよう．

Step 1. D の初期推定量 $\hat{D}$ を (適当に) 設定する．例えば，R^{-1} の対角要素を $r^{11}, \ldots, r^{pp}$ として，$\hat{D} = \mathrm{diag}(1/r^{11}, \ldots, 1/r^{pp})$ とする．

Step 2. $R - \hat{D}$ の固有値を大きいほうから順に m 個 $\lambda_1 \geq \cdots \geq \lambda_m$ 取り出し，それに対応する固有ベクトル $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m$ を用いて，$\hat{A} = [\sqrt{\lambda_1}\boldsymbol{c}_1, \ldots, \sqrt{\lambda_m}\boldsymbol{c}_m]$ とする．ここで，$m = \mathrm{rank}(R - \hat{D})$ なら，$R - \hat{D} = \hat{A}\hat{A}'$ となるが，一般には $m < \mathrm{rank}(R - \hat{D})$ なので，$R - \hat{D} \neq \hat{A}\hat{A}'$ であることに注意しよう．

Step 3. $R-\hat{D}$ の対角要素 $h_{11},\ldots,h_{pp}$ と $\hat{A}\hat{A}'$ の対角要素 $\alpha_{11},\ldots,\alpha_{pp}$ を比較して，与えられた判定基準 $\epsilon>0$ に対して，

$$|h_{jj}-\alpha_{jj}|<\epsilon$$

となれば，$\hat{A}$ の (j,k) 成分を因子負荷量 a_{jk} として，終了する．そうでなければ，$\hat{D}=I_p-\mathrm{diag}(\alpha_{11},\ldots,\alpha_{pp})$ として，Step 2 へ戻る．

6.7 因子負荷量の回転

因子負荷量 $A=(a_{jk})$ と因子得点 $\boldsymbol{f}_i=(f_{1i},\ldots,f_{mi})'$ には不定性があるので，任意の回転行列 T に対して，$A^*=AT$, $\boldsymbol{f}_i^*=T'\boldsymbol{f}_i$ も因子負荷量と因子得点となりうる．A^* の要素 a_{jk}^* の 2 乗を共通性 $(h_j^*)^2=(a_{j1}^*)^2+\cdots+(a_{jm}^*)^2$ で割って調整した量 $(a_{jk}^*)^2/(h_j^*)^2$ の分散，つまり，

$$V=\sum_{k=1}^{m}\frac{1}{p}\left\{\sum_{j=1}^{p}\frac{(a_{jk}^*)^4}{(h_j^*)^4}-\frac{1}{p}\left(\sum_{j=1}^{p}\frac{(a_{jk}^*)^2}{(h_j^*)^2}\right)^2\right\}$$

を最大にする回転 T を選ぶと，因子負荷量 a_{jk}^* は絶対値が大きいものと小さいものに分かれるので，比較的解釈が容易である．この回転のことを **バリマックス回転** と呼ぶ．他にも提案されている回転があるが，結果的に解釈のしやすいものを選ぶことが多い．

他にも，オーソマックス法（直交回転) やプロマックス法 (斜交回転）など色々あるが省略．

6.8 因子得点の推定法

因子得点 (f_{ki}) は，データ (z_{ji}) の線形結合 $\hat{f}_{ki}=\sum_{j=1}^{p}b_{kj}z_{ji}$ の形で推定されることが多い．このとき，係数 b_{kj} は

$$Q=\sum_{k=1}^{m}\sum_{i=1}^{n}(f_{ki}-\hat{f}_{ki})^2$$

を最小にするように選ぶ．このような推定法を **回帰推定法** と呼ぶが，他にも色々な推定法が提案されている．この係数 b_{kj} は，$f_{k1},\ldots,f_{kn}$ の平均が 0，分散が 1 であること，$e_{j1},\ldots,e_{jn}$ の平均が 0 であること，さらに，$(f_{k1},\ldots,f_{kn})$, $(f_{k'1},\ldots,f_{k'n})$, $(e_{j1},\ldots,e_{jn})$ が無相関であることを仮定すると，

$$b_{kj}=\sum_{j'=1}^{p}a_{j'k}r^{jj'}$$

となることが分かる．ここで，$r^{jj'}$ は相関行列 R の逆行列の (j,j') 成分．

6.9 数値例

	-1	-2	-3	-4	-5	-6
（1）国語	1	0.43231	0.37929	0.21867	0.30187	0.14674
（2）英語	0.43231	1	0.41695	0.35116	0.26784	0.28626
（3）社会	0.37929	0.41695	1	0.25768	0.4349	0.12058
（4）数学 1	0.21867	0.35116	0.25768	1	0.6952	0.623
（5）数学 2	0.30187	0.26784	0.4349	0.6952	1	0.67572
（6）理科	0.14674	0.28626	0.12058	0.623	0.67572	1

$\lambda_1=2.919$, $\lambda_2=1.263$, $\lambda_3=0.672$, $\lambda_4=0.585$

	因子	回転前 (a)		バリマックス回転後 (b)		共通性
変量		1	2	1	2	
(1) 国語		0.4179	0.1805	0.161	0.4257	0.2072
(2) 英語		0.4778	-0.1375	0.2336	0.4389	0.2472
(3) 社会		0.7002	-0.5913	0.0622	0.9143	0.8399
(4) 数学 1		0.6986	0.3781	0.7576	0.2389	0.631
(5) 数学 2		0.8388	0.2726	0.7793	0.413	0.7779
(6) 理科		0.6266	0.5309	0.8173	0.0809	0.6745
寄与率 (%)		41.3	15	32.2	24.1	–
累積寄与率 (%)		41.3	56.3	41.3	56.3	–

7 数量化 III 類

7.1 目的

2 次元の質的変数 A, B を，その関係をよく表すように，量的変数 x, y に変換するための方法.

		A_1	$\cdots$	A_p	
		x_1	$\cdots$	x_p	
B_1	y_1	f_{11}	$\cdots$	f_{1p}	b_1
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
B_n	y_n	f_{n1}	$\cdots$	f_{np}	b_n
		a_1	$\cdots$	a_p	N

7.2 方法

相関係数が最大になるように (x_j, y_i) を決定する. ただし，一般性を失うことなく，

$$\bar{x} = \sum_{j=1}^{p} a_j x_j = 0, \quad , \bar{y} = \sum_{i=1}^{n} b_i y_i = 0, \quad s_x^2 = \sum_{j=1}^{p} x_j^2 a_j = 1, \quad s_y^2 = \sum_{i=1}^{n} y_i^2 b_i = 1$$

という制約条件をおく. このとき，x と y の相関係数 R は

$$R = \sum_{i=1}^{n} \sum_{j=1}^{p} x_j y_i f_{ij}$$

である. ここで，$u_j = \sqrt{a_j} x_j$, $v_i = \sqrt{b_i} y_i$ とおくと, 制約条件は

$$\sum_{j=1}^{p} \sqrt{a_j} u_j = 0, \quad \sum_{i=1}^{n} \sqrt{b_i} v_i = 0, \quad \sum_{j=1}^{p} u_j^2 = 1, \quad \sum_{i=1}^{n} v_i^2 = 1$$

であり，

$$R = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{f_{ij}}{\sqrt{a_j b_i}} u_j v_i$$

となる. ラグランジュの未定乗数法を使って，R の最小値を求める.

$$H = R - \frac{\lambda}{2}(\sum_{j=1}^{p} u_j^2 - 1) - \frac{\eta}{2}(\sum_{i=1}^{n} v_i^2 - 1)$$

とおき，

$$\begin{cases} \frac{\partial H}{\partial u_j} = \sum_{i=1}^{n} \frac{f_{ij}}{\sqrt{a_j b_i}} v_i - \lambda u_j = 0 \\ \frac{\partial H}{\partial v_i} = \sum_{j=1}^{p} \frac{f_{ij}}{\sqrt{a_j b_i}} u_j - \eta v_i = 0 \end{cases}$$

として，u_j, v_i を求める.

$$\sum_{j=1}^{p} u_j \frac{\partial H}{\partial u_j} = R - \lambda \sum_{j=1}^{p} u_j^2 = R - \lambda = 0, \quad \sum_{i=1}^{n} v_i \frac{\partial H}{\partial v_i} = R - \eta = 0$$

だから，$R = \lambda = \eta$. これより，

$$\sum_{i=1}^{n} \frac{f_{ij}}{\sqrt{a_j b_i}} \left(\frac{1}{\lambda} \sum_{j'=1}^{p} \frac{f_{ij'}}{\sqrt{a_{j'} b_i}} u_{j'} \right) = \lambda u_j$$

が成り立つことがわかり，

$$C = (c_{jj'}), \quad c_{jj'} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij'}}{\sqrt{a_j a_{j'}} b_i}, \quad \boldsymbol{u} = [u_1, \ldots, u_p]'$$

とおくと，

$$C\boldsymbol{u} = \lambda^2 \boldsymbol{u}.$$

つまり，λ^2 は C の固有値であり，$\boldsymbol{u}$ はそれに対応する固有ベクトルであることがわかる．問題は，R を大きくする (x_j, y_i) を求めることであったが，$R = \lambda$ なので，C の大きな固有値に対応する固有ベクトルを求めるとよいになる．ただし，$|R| \leqq 1$ なので，$0 \leqq \lambda^2 \leqq 1$ に注意しよう．また，$\boldsymbol{u} = [\sqrt{a_1}, \ldots, \sqrt{a_p}]'$ とすると $C\boldsymbol{u} = \boldsymbol{u}$ が成り立つので，最大の固有値は 1 であり，その固有ベクトルは　$\boldsymbol{u} = [\sqrt{a_1}, \ldots, \sqrt{a_p}]'$ であることがわかるが，この $\boldsymbol{u}$ は制約条件 $\sum_{j=1} \sqrt{a_j} u_j = 0$ を満たさないので，求めるものとして不適当である．

一般に，固有値の大きいものの順に $\lambda_1^2 \geqq \lambda_2^2 \geqq \cdots \geqq \lambda_p^2$ と番号付けすると，$1 = \lambda_1^2 > \lambda_2^2 \geqq \cdots \geqq \lambda_p^2 \geqq 0$ となり，$\lambda_k^2, k = 2, \ldots, p$ に対応する固有ベクトル $\boldsymbol{u}_k$ は制約条件を満たすことが知られている．したがって，$\boldsymbol{u}_k$ に対応する $\boldsymbol{v}$ を $\boldsymbol{v}_k$, それらに対応する x, y を $\boldsymbol{x}_k, \boldsymbol{y}_k$ として，適当な固有値の大きさまでの $\boldsymbol{x}_k, \boldsymbol{y}_k$ を量的変数として用いればよい．主成分分析などと同様に，

$$\frac{\lambda_2^2 + \cdots + \lambda_k^2}{\lambda_2^2 + \cdots + \lambda_k^2 + \cdots + \lambda_p^2}$$

により定義される累積寄与率によって判断することが多いが，客観的な基準はない．

7.3 例

	A 君	B さん	C さん	D 君
学食	0.8	0.4	0.6	0.5
コンビニ	0	0.2	0.1	0.3
弁当	0.1	0.2	0.1	0.1
レストラン	0	0.2	0	0.1
なし	0.1	0	0.2	0

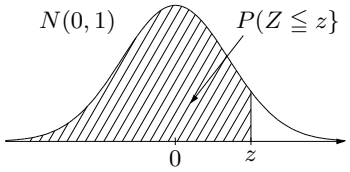
	A 君	C さん	D 君	B さん
学食	0.8	0.6	0.5	0.4
コンビニ	0	0.1	0.3	0.2
弁当	0.1	0.1	0.1	0.2
レストラン	0	0	0.1	0.2
なし	0.1	0.2	0	0

	A 君	C さん	D 君	B さん
なし	0.1	0.2	0	0
弁当	0.1	0.1	0.1	0.2
学食	0.8	0.6	0.5	0.4
コンビニ	0	0.1	0.3	0.2
レストラン	0	0	0.1	0.2

1 分布表

A.1 正規分布表.

$Z \sim N(0,1)$ の時の $P(Z \leqq z)$ の値.

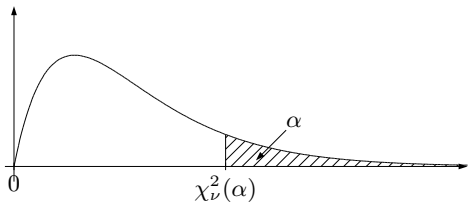


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995

A.2 正規分布上側パーセント点

α	0.005	0.01	0.025	0.05	0.1
$z(\alpha)$	2.576	2.326	1.960	1.645	1.282

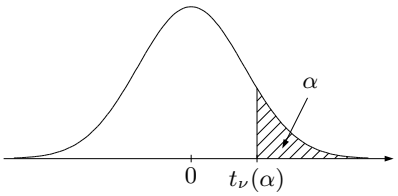
A.3 カイ 2 乗分布表．$W \sim \chi^2_\nu$ の時の，$P(W > x) = \alpha$ となる $x = \chi^2_\nu(\alpha)$ の表.



自由度 ν のカイ 2 乗分布 χ^2_ν

$\nu\backslash\alpha$	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
120	83.852	86.923	91.573	95.705	100.624	140.233	146.567	152.211	158.950	163.648

A.4 ティー分布表. $T \sim t_\nu$ のときの，$P(T > x) = \alpha$ を満たす $x = t_\nu(\alpha)$ の表.

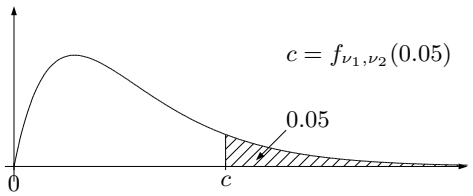


自由度 ν のティー分布 t_ν

0 で対称な分布だから，$P(|T| > x) = \alpha$ の時，$P(T > x) = P(T < -x) = \alpha/2$.

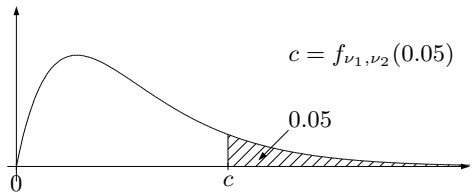
$\nu\backslash\alpha$	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576

A.5 エフ分布表 1. $W \sim F_{\nu_1,\nu_2}$ の時の，$P(W > c) = 0.05$ となる c の表



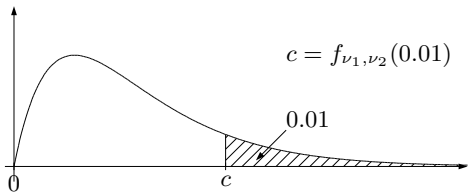
$\nu_2\backslash\nu_1$	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.13	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.10	3.01	2.95	2.90	2.85
12	4.75	3.88	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.02	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.56	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.33	3.47	3.07	2.84	2.69	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.38	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.38	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.17
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09
40	4.09	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
∞	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83

A.6 エフ分布表 2. $W \sim F_{\nu_1,\nu_2}$ の時の，$P(W > c) = 0.05$ となる c の表



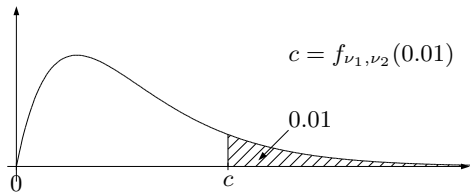
$\nu_2\backslash\nu_1$	11	12	14	16	20	24	30	40	50	∞
1	242.98	243.91	245.36	246.46	248.01	249.05	250.10	251.14	251.77	254.31
2	19.41	19.41	19.42	19.43	19.45	19.45	19.46	19.47	19.48	19.50
3	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.59	8.58	8.53
4	5.94	5.91	5.87	5.84	5.80	5.77	5.75	5.72	5.70	5.63
5	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.37
6	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.67
7	3.60	3.58	3.53	3.49	3.44	3.41	3.38	3.34	3.32	3.23
8	3.31	3.28	3.24	3.20	3.15	3.12	3.08	3.04	3.02	2.93
9	3.10	3.07	3.02	2.99	2.94	2.90	2.86	2.83	2.80	2.71
10	2.94	2.91	2.87	2.83	2.77	2.74	2.70	2.66	2.64	2.54
11	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.51	2.40
12	2.72	2.69	2.64	2.60	2.54	2.50	2.47	2.43	2.40	2.30
13	2.63	2.60	2.55	2.52	2.46	2.42	2.38	2.34	2.31	2.21
14	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.13
15	2.51	2.48	2.42	2.38	2.33	2.29	2.25	2.20	2.18	2.07
16	2.46	2.42	2.37	2.33	2.28	2.23	2.19	2.15	2.12	2.01
17	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.10	2.08	1.96
18	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.06	2.04	1.92
19	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.03	2.00	1.88
20	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.97	1.84
21	2.28	2.25	2.20	2.16	2.10	2.05	2.01	1.97	1.94	1.81
22	2.26	2.23	2.17	2.13	2.07	2.03	1.98	1.94	1.91	1.78
23	2.24	2.20	2.15	2.11	2.05	2.00	1.96	1.91	1.89	1.76
24	2.22	2.18	2.13	2.09	2.03	1.98	1.94	1.89	1.86	1.73
25	2.20	2.17	2.11	2.07	2.01	1.96	1.92	1.87	1.84	1.71
26	2.18	2.15	2.09	2.05	1.99	1.95	1.90	1.85	1.82	1.69
27	2.17	2.13	2.08	2.04	1.97	1.93	1.88	1.84	1.81	1.67
28	2.15	2.12	2.06	2.02	1.96	1.92	1.87	1.82	1.79	1.65
29	2.14	2.10	2.05	2.01	1.95	1.90	1.85	1.81	1.77	1.64
30	2.13	2.09	2.04	2.00	1.93	1.89	1.84	1.79	1.76	1.62
32	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.77	1.74	1.59
34	2.08	2.05	2.00	1.95	1.89	1.84	1.79	1.75	1.71	1.57
36	2.07	2.03	1.98	1.93	1.87	1.82	1.78	1.73	1.69	1.55
38	2.05	2.02	1.96	1.92	1.85	1.81	1.76	1.71	1.68	1.53
40	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.51
50	1.99	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.44
∞	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.39	1.35	1.00

A.7 エフ分布表 3. $W \sim F_{\nu_1,\nu_2}$ の時の，$P(W > c) = 0.01$ となる c の表



$\nu_2\backslash\nu_1$	1	2	3	4	5	6	7	8	9	10
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.74	10.93	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.90	3.81
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.67	4.18	3.86	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.86
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.79	2.70
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

A.8 エフ分布表 4. $W \sim F_{\nu_1,\nu_2}$ の時の，$P(W > c) = 0.01$ となる c の表



$\nu_2\backslash\nu_1$	11	12	14	16	20	24	30	40	50	∞
1	6083.32	6106.32	6142.67	6170.10	6208.73	6234.63	6260.65	6286.78	6302.52	6365.86
2	99.41	99.42	99.43	99.44	99.45	99.46	99.47	99.47	99.48	99.50
3	27.13	27.05	26.92	26.83	26.69	26.60	26.50	26.41	26.35	26.13
4	14.45	14.37	14.25	14.15	14.02	13.93	13.84	13.74	13.69	13.46
5	9.96	9.89	9.77	9.68	9.55	9.47	9.38	9.29	9.24	9.02
6	7.79	7.72	7.61	7.52	7.40	7.31	7.23	7.14	7.09	6.88
7	6.54	6.47	6.36	6.28	6.16	6.07	5.99	5.91	5.86	5.65
8	5.73	5.67	5.56	5.48	5.36	5.28	5.20	5.12	5.07	4.86
9	5.18	5.11	5.00	4.92	4.81	4.73	4.65	4.57	4.52	4.31
10	4.77	4.71	4.60	4.52	4.41	4.33	4.25	4.17	4.12	3.91
11	4.46	4.40	4.29	4.21	4.10	4.02	3.94	3.86	3.81	3.60
12	4.22	4.16	4.05	3.97	3.86	3.78	3.70	3.62	3.57	3.36
13	4.03	3.96	3.86	3.78	3.67	3.59	3.51	3.42	3.38	3.17
14	3.86	3.80	3.70	3.62	3.50	3.43	3.35	3.27	3.21	3.00
15	3.73	3.67	3.56	3.48	3.37	3.29	3.21	3.13	3.08	2.87
16	3.62	3.55	3.45	3.37	3.26	3.18	3.10	3.02	2.97	2.75
17	3.52	3.46	3.35	3.27	3.16	3.08	3.00	2.92	2.87	2.65
18	3.43	3.37	3.27	3.19	3.08	3.00	2.92	2.84	2.78	2.57
19	3.36	3.30	3.19	3.12	3.00	2.92	2.84	2.76	2.71	2.49
20	3.29	3.23	3.13	3.05	2.94	2.86	2.78	2.69	2.64	2.42
21	3.24	3.17	3.07	2.99	2.88	2.80	2.72	2.64	2.58	2.36
22	3.18	3.12	3.02	2.94	2.83	2.75	2.67	2.58	2.53	2.31
23	3.14	3.07	2.97	2.89	2.78	2.70	2.62	2.54	2.48	2.26
24	3.09	3.03	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.21
25	3.06	2.99	2.89	2.81	2.70	2.62	2.54	2.45	2.40	2.17
26	3.02	2.96	2.86	2.78	2.66	2.59	2.50	2.42	2.36	2.13
27	2.99	2.93	2.82	2.75	2.63	2.55	2.47	2.38	2.33	2.10
28	2.96	2.90	2.79	2.72	2.60	2.52	2.44	2.35	2.30	2.06
29	2.93	2.87	2.77	2.69	2.57	2.50	2.41	2.33	2.27	2.03
30	2.91	2.84	2.74	2.66	2.55	2.47	2.39	2.30	2.25	2.01
32	2.86	2.80	2.70	2.62	2.50	2.42	2.34	2.25	2.20	1.96
34	2.82	2.76	2.66	2.58	2.46	2.38	2.30	2.21	2.16	1.91
36	2.79	2.72	2.62	2.54	2.43	2.35	2.26	2.17	2.12	1.87
38	2.75	2.69	2.59	2.51	2.40	2.32	2.23	2.14	2.09	1.84
40	2.73	2.67	2.56	2.48	2.37	2.29	2.20	2.11	2.06	1.80
50	2.63	2.56	2.46	2.38	2.27	2.18	2.10	2.01	1.95	1.68
∞	2.25	2.19	2.08	2.00	1.88	1.79	1.70	1.59	1.52	1.00

索 引