

統計検定

Japan Statistical Society Certificate

準1級

2021年6月20日

【注意事項】

- 1 試験開始の合図があるまで、この問題冊子および部分記述問題・論述問題解答用紙の中を見てはいけません。
- 2 この問題冊子は、40ページあります。3~25ページは選択問題及び部分記述問題、26~32ページは論述問題です。
- 3 論述問題は、3問から1問のみを選択して、解答しなさい。
- 4 試験時間は120分です。
- 5 試験中に問題冊子の印刷不鮮明、ページの落丁・乱丁およびマークシートの汚れ等に気付いた場合は、手を挙げて監督者に知らせなさい。
- 6 解答用紙・マークシートのA面には次の項目があるので、それぞれの指示に従い記入あるいは確認しなさい。項目の内容に誤りがある場合は、手を挙げて監督者に知らせなさい。
 - ① マークシートの氏名
氏名を記入しなさい。
 - ② マークシートの検定種別と受験番号
受験する検定種別と受験番号を確認しなさい。
 - ③ マークシートのWeb合格発表
Web合格発表について、希望の有無をマークしなさい。
 - ④ 部分記述問題・論述問題解答用紙の表紙上部に受験番号を記入しなさい。
 - ⑤ 論述問題の解答面には、最終的な解答だけでなくそれに至る過程も記しなさい。最終的な解答が正しくないときにも点が与えられることがあります。
- 7 選択問題の解答は、マークシートのB面の解答にマークしなさい。
(この冊子裏面の記入例参照)
- 8 部分記述問題の解答は、部分記述問題解答面に解答しなさい。記述2のように表示のある問は部分記述問題です。
- 9 選択問題（問3～問11）の解答番号は28まで、部分記述問題（問1, 2, 12）の解答番号は7まであります。
- 10 33ページ以降に付表を掲載しています。必要に応じて利用しなさい。
- 11 問題冊子の余白等は適宜利用してよいが、どのページも切り離してはいけません。
- 12 試験終了後、問題冊子は持ち帰りなさい。

(冊子裏面につづく)

選択問題及び部分記述問題

問1 A, B, C の3つの事象について

$$P(A) = 0.45, \quad P(A \cup B) = 0.65, \quad P(A|B) = 0.5, \quad P(C) = 0.45,$$
$$P(A \cap C) = 0.2, \quad P(B \cap C) = 0.1, \quad P(A \cap B \cap C) = 0.05$$

のように確率が与えられているとする。

[1] 最初の3つの式を用い、確率 $P(B)$ を求めよ。 記述1

[2] 確率 $P(A \cup B \cup C)$ を求めよ。 記述2

問2 ある機械が n 台あり、 i 番目の機械が故障するまでの時間 X_i はそれぞれ独立に平均 λ の指数分布に従うとする ($i = 1, 2, \dots, n$)。ここで、平均 λ の指数分布の確率密度関数は

$$f(x) = \frac{1}{\lambda} e^{-x/\lambda} \quad (x \geq 0)$$

である。

[1] X_i の分散 $\theta = V(X_i)$ を λ の関数として表せ。 記述3

[2] [1] で求めた分散の最尤推定量 $\hat{\theta}$ を求めよ。 記述4

[3] [2] で求めた最尤推定量の漸近分散 $\lim_{n \rightarrow \infty} V(\sqrt{n}(\hat{\theta} - \theta))$ を λ の関数として表せ。 記述5

注：記述6, 7は問12にあります。

問3 $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ を次の3変量正規分布に従う確率ベクトルとする。

$$N\left(\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 & 0 & 1 \\ 0 & 3 & 2 \\ 1 & 2 & 4 \end{pmatrix}\right)$$

[1] $\begin{pmatrix} X+Y \\ Y-Z \end{pmatrix}$ が従う2変量正規分布として、次の①～⑤のうちから適切なものを一つ選べ。 1

① $N\left(\begin{pmatrix} 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 5 & 0 \\ 0 & 3 \end{pmatrix}\right)$

② $N\left(\begin{pmatrix} 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}\right)$

③ $N\left(\begin{pmatrix} 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 5 & 4 \\ 4 & 3 \end{pmatrix}\right)$

④ $N\left(\begin{pmatrix} 3 \\ -1 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}\right)$

⑤ $N\left(\begin{pmatrix} 3 \\ -1 \end{pmatrix}, \begin{pmatrix} 5 & 0 \\ 0 & 3 \end{pmatrix}\right)$

[2] $X = x$ と $Y = y$ を与えたときの Z の条件付き分布として、次の①～⑤のうちから適切なものを一つ選べ。 2

① $N\left(3, \frac{15}{4}\right)$

② $N\left(x + 2y - 2, \frac{15}{4}\right)$

③ $N\left(x + 2y - 2, \frac{13}{6}\right)$

④ $N\left(\frac{1}{2}x + \frac{2}{3}y + \frac{7}{6}, \frac{15}{4}\right)$

⑤ $N\left(\frac{1}{2}x + \frac{2}{3}y + \frac{7}{6}, \frac{13}{6}\right)$

問4 n 次元確率変数 $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^\top$ について,

$$E(\mathbf{Y}) = b\mathbf{x}, \quad V(\mathbf{Y}) = \sigma^2 I_n$$

という線形模型を仮定する。ここで, \mathbf{a}^\top はベクトル \mathbf{a} の転置を表しており, n は正の整数, $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ は要素が 0 ではなく確率的に変動しない所与の n 次元列ベクトル, b は未知のスカラー母数, $\sigma^2 (> 0)$ は未知のスカラー母数, I_n は大きさ $n \times n$ の単位行列, $E(\mathbf{Y})$ と $V(\mathbf{Y})$ はそれぞれ \mathbf{Y} の期待値と分散共分散行列である。また, 以下では $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{\mathbf{x}} = (\bar{x}, \bar{x}, \dots, \bar{x})^\top$ とする。

[1] $n = 6$, $\mathbf{x} = (1.1, 1.2, 1.9, 2.7, 2.8, 3.0)^\top$ とする。そして \mathbf{Y} の実現値 $\mathbf{y} = (0.1, 2.7, 3.3, 8.0, 6.2, 7.1)^\top$ が得られたとする。 $\mathbf{x}^\top \mathbf{x} = 30.39$, $\mathbf{x}^\top \mathbf{y} = 69.88$, $\mathbf{y}^\top \mathbf{y} = 171.04$ であることは用いてよい。

[1-1] b の最小二乗法による推定量の実現値, すなわち $(\mathbf{y} - b\mathbf{x})^\top (\mathbf{y} - b\mathbf{x})$ を最小にする b として, 次の ① ~ ⑤ のうちから最も適切なものを一つ選べ。 3

- ① 1.0 ② 1.9 ③ 2.3 ④ 3.6 ⑤ 4.1

[1-2] b の最小二乗法による推定量の実現値を \hat{b} で表す。残差平方和 $(\mathbf{y} - \hat{b}\mathbf{x})^\top (\mathbf{y} - \hat{b}\mathbf{x})$ にある数をかけて得られる σ^2 の不偏推定量の実現値として, 次の ① ~ ⑤ のうちから最も適切なものを一つ選べ。 4

- ① 1.3 ② 1.8 ③ 2.1 ④ 4.7 ⑤ 6.8

[2] 母数 b の最小二乗法による推定量に関して、次の (A) と (B) で述べられている事項の正誤について、下の ① ~ ④ のうちから最も適切なものを一つ選べ。 5

- (A) 最小二乗法による推定量は不偏性をもつが、他にも不偏性をもつ推定量が存在する。例えば、 \mathbf{c} を n 次元列ベクトルとして $\mathbf{c}^\top \mathbf{Y}$ という形の推定量を考えると、 $\mathbf{c}^\top \mathbf{x} = 1$ のとき、かつそのときに限って、推定量 $\mathbf{c}^\top \mathbf{Y}$ が不偏性をもつ。
- (B) 最小二乗法による推定量の分散より小さい分散をもつ不偏推定量が存在することはない。

① (A) と (B) の両方が正しい
③ (B) は正しいが (A) は誤り

② (A) は正しいが (B) は誤り
④ (A) と (B) の両方が誤り

[3] 次の文中に含まれる空欄「ア・イ」に当てはまる記述の組合せとして、下の ① ~ ⑤ のうちから最も適切なものを一つ選べ。ただし、平均ベクトル μ 、分散共分散行列 $\sigma^2 I_n$ の n 次元正規分布の確率密度関数を $f_n(\mathbf{t}; \mu, \sigma^2)$ (\mathbf{t} は n 次元の列ベクトル) で表し、最小二乗法による b の推定量を \hat{b} で表す。 6

$n = 6$ でさらに \mathbf{Y} の分布が 6 次元正規分布であることを仮定し、「帰無仮説： $b = 0$ 、対立仮説： $b \neq 0$ 」の尤度比検定を有意水準 5% で考える。「尤度比検定統計量 $\frac{\sup_{\sigma^2 > 0} f_6(\mathbf{Y}; \mathbf{0}, \sigma^2)}{\sup_{b, \sigma^2 > 0} f_6(\mathbf{Y}; b\mathbf{x}, \sigma^2)}$ の実現値が定数 c 未満となるとき帰無仮説を棄却する」ことを変形して得られる検定方式として「検定統計量を [ア] として棄却域を [イ, ∞) とする」を導ける。

- ① ア : $\frac{5\mathbf{x}^\top \mathbf{x}(\hat{b})^2}{(\mathbf{Y} - \hat{b}\mathbf{x})^\top (\mathbf{Y} - \hat{b}\mathbf{x})}$, イ : 6.61 ② ア : $\frac{5\mathbf{x}^\top \mathbf{x}(\hat{b})^2}{(\mathbf{Y} - \hat{b}\mathbf{x})^\top (\mathbf{Y} - \hat{b}\mathbf{x})}$, イ : 2.02
 ③ ア : $\frac{5(\mathbf{x} - \bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}})(\hat{b})^2}{(\mathbf{Y} - \hat{b}\mathbf{x})^\top (\mathbf{Y} - \hat{b}\mathbf{x})}$, イ : 2.02 ④ ア : $\frac{6(\mathbf{x} - \bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}})(\hat{b})^2}{(\mathbf{Y} - \hat{b}\mathbf{x})^\top (\mathbf{Y} - \hat{b}\mathbf{x})}$, イ : 6.61
 ⑤ ア : $\frac{6(\mathbf{x} - \bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}})(\hat{b})^2}{(\mathbf{Y} - \hat{b}\mathbf{x})^\top (\mathbf{Y} - \hat{b}\mathbf{x})}$, イ : 1.94

問5 ある素粒子と陽子をいくつかのエネルギーで衝突させることによって生成される断面積（素粒子物理学での基本的な物理量）を測定する実験を行う。衝突時のエネルギーの逆数 (x) と断面積 (y) には強い線形関係があると言われている。実データ（資料：統計分析ソフト R 内のパッケージ faraway にある strongx）から得られるエネルギーの逆数値と断面積をプロットした図1からも線形関係が見てとれる。線形関係を調べるために、次のモデルに基づく単回帰分析を行う。

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, 2, \dots, 10)$$

ここで、誤差 ε_i はそれぞれ独立に平均 0, 分散 σ_i^2 をもつ確率変数とする。また、この分散 σ_i^2 の値は既知とし、いくつかの σ_i^2 は異なる値をとるとする。

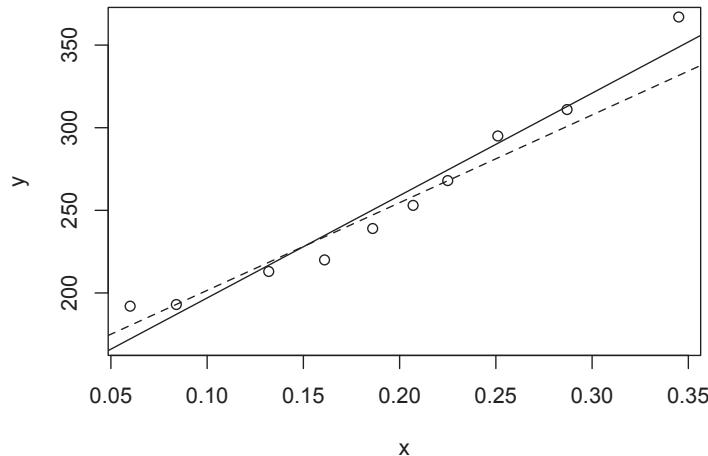


図1: 観測データの散布図と回帰直線（実線は最小二乗法による線形回帰直線、破線は一般化最小二乗法による線形回帰直線）

[1] 図1の2つの回帰直線は、最小二乗法で推定したものと、一般化最小二乗法で重み行列を 10×10 対角行列

$$\begin{pmatrix} 1/\sigma_1^2 & 0 & \cdots & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1/\sigma_9^2 & 0 \\ 0 & \cdots & \cdots & 0 & 1/\sigma_{10}^2 \end{pmatrix}$$

として推定したものである。最小二乗推定と比較し、一般化最小二乗推定を用いる利点として、次の①～⑤のうちから最も適切なものを一つ選べ。 7

- ① 推定量のバイアスが小さくなる。
- ② 推定量の分散が小さくなる。
- ③ 推定が容易になる。
- ④ 残差二乗和の値が小さくなる。
- ⑤ 回帰係数の推定値を 0 に近づけることができる。

[2] 回帰モデルの当てはまりの尺度として、最小二乗法に基づく決定係数

$$R^2 = \frac{\boxed{\text{ア}}}{\boxed{\text{イ}}}$$

がある。断面積の平均を \bar{y} , 推定された回帰モデルによる予測値を \hat{y}_i としたとき,
「**ア**・**イ**」に当てはまる式の組合せとして、次の①～⑤のうちから適切なもの
を一つ選べ。 8

- ① ア : $\sum_{i=1}^{10} (y_i - \hat{y}_i)^2$, イ : $\sum_{i=1}^{10} (y_i - \bar{y})^2$
- ② ア : $\sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2$, イ : $\sum_{i=1}^{10} (y_i - \bar{y})^2$
- ③ ア : $\sum_{i=1}^{10} (y_i - \bar{y})^2$, イ : $\sum_{i=1}^{10} (y_i - \hat{y}_i)^2$
- ④ ア : $\sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2$, イ : $\sum_{i=1}^{10} (y_i - \hat{y}_i)^2$
- ⑤ ア : $\sum_{i=1}^{10} (y_i - \bar{y})^2$, イ : $\sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2$

[3] 最小二乗推定だけでなく一般化最小二乗推定でも、[2] で与えた定義に基づき、
決定係数を求めるとする。その定義において、最小二乗推定による予測値を用いて
求めた決定係数を R_1^2 , 一般化最小二乗推定による予測値を用いて求めた決定
係数を R_2^2 とする。図 1 で与えられている実データから得られる R_1^2 と R_2^2 の組合せ
として、次の①～⑤のうちから最も適切なものを一つ選べ。 9

- | | |
|-------------------------------------|-------------------------------------|
| ① $R_1^2 = 0.355$, $R_2^2 = 0.440$ | ② $R_1^2 = 0.440$, $R_2^2 = 0.355$ |
| ③ $R_1^2 = 0.440$, $R_2^2 = 0.440$ | ④ $R_1^2 = 0.955$, $R_2^2 = 0.705$ |
| ⑤ $R_1^2 = 0.705$, $R_2^2 = 0.955$ | |

問6 2つのグループからのデータを判別する代表的な方法に、フィッシャーの線形判別がある。グループ1, グループ2の2つのグループから2次元データを収集したものとする。それぞれの標本サイズを n_1, n_2 とし, データを $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n_1}\}$, $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{n_2}\}$ とおく。また, それぞれのグループの平均ベクトルを $\bar{\mathbf{x}} = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_i$, $\bar{\mathbf{y}} = \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{y}_i$ とおく。さらに, データ全体を $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n\}$, 平均ベクトルを $\bar{\mathbf{z}} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i$ とおく。ただし, $n = n_1 + n_2$ である。

[1] 各グループの分散共分散行列 S_1, S_2 とデータ全体の分散共分散行列 S をそれぞれ

$$\begin{aligned} S_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top \\ S_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top \\ S &= \frac{1}{n} \sum_{i=1}^n (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^\top \end{aligned}$$

とおき, さらに

$$\begin{aligned} S_W &= \frac{n_1}{n} S_1 + \frac{n_2}{n} S_2 \\ S_B &= \frac{n_1}{n} (\bar{\mathbf{x}} - \bar{\mathbf{z}})(\bar{\mathbf{x}} - \bar{\mathbf{z}})^\top + \frac{n_2}{n} (\bar{\mathbf{y}} - \bar{\mathbf{z}})(\bar{\mathbf{y}} - \bar{\mathbf{z}})^\top \end{aligned}$$

と定義する。ここで[†] \top は転置を表すとする。3つの行列 S, S_W, S_B の関係について, 次の①～④のうちから最も適切なものを一つ選べ。ただし, $P > Q$ は行列 $P - Q$ の固有値がすべて正であることを意味する。 10

- ① つねに $S > S_W + S_B$ が成り立つ。
- ② つねに $S = S_W + S_B$ が成り立つ。
- ③ つねに $S < S_W + S_B$ が成り立つ。
- ④ 上記に正しいものは一つもない。

[2] フィッシャーの線形判別は、行列の固有値・固有ベクトルを計算して与えられる。具体的には、対応する固有ベクトルを v 、新しいデータを z_0 とおくとき、 $v^\top z_0$ により線形判別が行われる。 S_W, S_B が

$$S_W = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}, \quad S_B = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}$$

と与えられたとき、固有値を計算する行列と、線形判別に用いる固有ベクトル v の組合せとして、次の①～④のうちから最も適切なものを一つ選べ。

11

- ① 行列 : $\begin{pmatrix} 1 & 1/2 \\ 0 & 0 \end{pmatrix}$, 固有ベクトル : $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$
- ② 行列 : $\begin{pmatrix} 1 & 1/2 \\ 0 & 0 \end{pmatrix}$, 固有ベクトル : $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$
- ③ 行列 : $\begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix}$, 固有ベクトル : $\begin{pmatrix} -1 \\ 0 \end{pmatrix}$
- ④ 行列 : $\begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix}$, 固有ベクトル : $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$

問7 モデル選択に関する次の各間に答えよ。

[1] 2017年1月1日から2017年12月31日までの、ある地域の気象データは毎日毎に記録されたもの(サンプルサイズは365)であり、平均気温(°C)、平均湿度(%)、平均風速(m/s)、日照時間(h)が確認できる。このデータについて、平均気温を y_t 、平均湿度を x_{t1} 、平均風速を x_{t2} 、日照時間を x_{t3} と表す($t = 1, 2, \dots, 365$)。平均気温を平均湿度、平均風速、日照時間で予測するための候補モデルとして、以下の線形回帰モデルを考える:

$$\text{モデル 1: } y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \varepsilon_t$$

$$\text{モデル 2: } y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \varepsilon_t$$

$$\text{モデル 3: } y_t = \beta_0 + \beta_1 x_{t1} + \beta_3 x_{t3} + \varepsilon_t$$

$$\text{モデル 4: } y_t = \beta_0 + \beta_2 x_{t2} + \beta_3 x_{t3} + \varepsilon_t$$

$$\text{モデル 5: } y_t = \beta_0 + \beta_1 x_{t1} + \varepsilon_t$$

$$\text{モデル 6: } y_t = \beta_0 + \beta_2 x_{t2} + \varepsilon_t$$

$$\text{モデル 7: } y_t = \beta_0 + \beta_3 x_{t3} + \varepsilon_t$$

ただし、 ε_t は独立に同一の正規分布 $N(0, \sigma^2)$ に従うものとする。

[1-1] t 日目のデータ y_t , x_{t1} , x_{t2} , x_{t3} に対して、モデル1の確率密度関数として、次の①～⑤のうちから適切なものを一つ選べ。ただし、 $\phi(\cdot)$ を標準正規分布の密度関数とする。 12

$$\textcircled{1} \quad \frac{1}{\sigma^2} \phi \left(\frac{y_t + \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3}}{\sigma^2} \right)$$

$$\textcircled{2} \quad \frac{1}{\sigma^2} \phi \left(\frac{y_t - (\beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3})}{\sigma^2} \right)$$

$$\textcircled{3} \quad \frac{1}{\sigma} \phi \left(\frac{y_t + \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3}}{\sigma} \right)$$

$$\textcircled{4} \quad \frac{1}{\sigma} \phi \left(\frac{y_t - (\beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3})}{\sigma} \right)$$

$$\textcircled{5} \quad \frac{1}{\sigma} \phi \left(\frac{y_t - (\beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3})}{\sigma^2} \right)$$

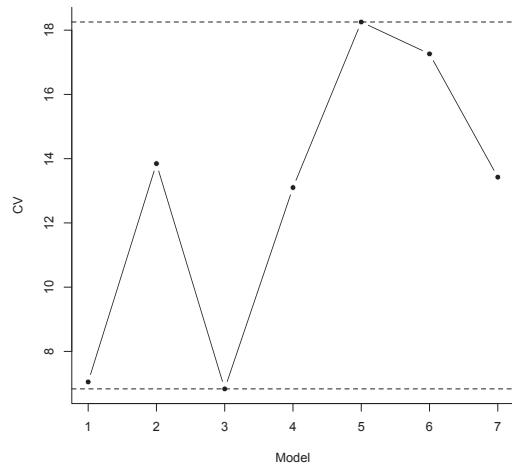
[1-2] モデル1からモデル7の各モデルにおいて最大対数尤度を計算し、以下の結果を得た。

候補モデル	最大対数尤度	パラメータ数
モデル1	-842.9193	5
モデル2	-960.4235	4
モデル3	-845.3840	4
モデル4	-951.6040	4
モデル5	-1016.1566	3
モデル6	-1001.0149	3
モデル7	-962.0377	3

このとき、AICの意味で最適なモデルとして、次の①～⑤のうちから最も適切なものを一つ選べ。 13

- ① モデル1 ② モデル3 ③ モデル4 ④ モデル5 ⑤ モデル6

[1-3] 10分割交差検証(10-fold CV)により、予測誤差の推定を行ったところ以下の図のような結果が得られた。



ここで、縦軸は10分割交差検証を行った結果、横軸は交差検証の対象となつたモデルを表している。このとき、交差検証の意味で最適なモデルとして、次の①～⑤のうちから最も適切なものを一つ選べ。 14

- ① モデル1 ② モデル3 ③ モデル4 ④ モデル5 ⑤ モデル6

[2] 情報量規準 (AIC, BIC) および交差検証の特徴を検証するために、モデル選択に関する数値実験を次の手順で行う。

ステップ1: データを生成するモデルを一つ定める (以下真のモデルと呼ぶ)。

ステップ2: データを表すモデルの候補を真のモデルも含め M 個想定する。

ステップ3: 真のモデルに従いサンプルサイズ N のデータを生成する。

ステップ4: AIC, BIC および 10 分割交差検証を用いて、ステップ2で想定した候補モデルの中からそれぞれ最適なモデルを選択する。

ステップ5: ステップ3, ステップ4を 1000 回繰り返す。

この数値実験をサンプルサイズ $N = 200, 300, 500, 1000$ の 4 通り、候補モデル数 $M = 8$ で行う。次の表は、真のモデルの選択確率 (ステップ4で真のモデルが選択された回数/1000) とステップ4を 1 回行う際にかかった時間の平均 (秒) を観測した結果である。

サンプルサイズ	選択確率			計算時間(秒)		
	手法(A)	手法(B)	手法(C)	手法(A)	手法(B)	手法(C)
200	0.456	0.679	0.663	0.183	0.184	1.635
300	0.681	0.801	0.766	0.276	0.275	2.459
500	0.912	0.847	0.821	0.459	0.461	4.115
1000	0.989	0.843	0.808	0.934	0.938	8.383

手法(A), (B), (C) は AIC, BIC, 10 分割交差検証のいずれかを表している。この手法と AIC, BIC, 10 分割交差検証の対応について、次の ① ~ ⑤ のうちから最も適切なものを一つ選べ。 15

- | | | |
|-----------------|---------------|---------------|
| ① (A) AIC | (B) BIC | (C) 10 分割交差検証 |
| ② (A) AIC | (B) 10 分割交差検証 | (C) BIC |
| ③ (A) BIC | (B) AIC | (C) 10 分割交差検証 |
| ④ (A) BIC | (B) 10 分割交差検証 | (C) AIC |
| ⑤ (A) 10 分割交差検証 | (B) AIC | (C) BIC |

問8 ある製薬会社が高血圧治療のための降圧薬Aを開発した。降圧薬Aとプラセボの効果を比較する臨床試験を計画する。各群の血圧の減少量（血圧が減少したら正の値になる）は、ともに分散 σ^2 の正規分布に従うとする。そして、降圧薬A群の平均減少量を μ_A 、プラセボ群の平均減少量を μ_P 、 $\mu_A - \mu_P$ で見込む値を $\delta (> 0)$ とし、有意水準5%，検出力80%の片側検定を行う場合の必要症例数の設計を考える。

[1] この試験の仮説検定における帰無仮説 H_0 と対立仮説 H_1 の組合せとして、次の①～⑤のうちから最も適切なものを一つ選べ。 16

- | | | |
|---|--------------------------------|-------------------------------|
| ① | $H_0: \mu_A = \mu_P,$ | $H_1: \mu_A \neq \mu_P$ |
| ② | $H_0: \mu_A = \mu_P,$ | $H_1: \mu_A < \mu_P$ |
| ③ | $H_0: \mu_A \neq \mu_P,$ | $H_1: \mu_A > \mu_P$ |
| ④ | $H_0: \mu_A = \mu_P,$ | $H_1: \mu_A > \mu_P$ |
| ⑤ | $H_0: \mu_A - \mu_P = \delta,$ | $H_1: \mu_A - \mu_P < \delta$ |

[2] 次の①～④のうちから最も適切なものを一つ選べ。 17

- ① 検出力を70%に減少させる場合、必要症例数は増加する。
- ② 有意水準を1%とする場合、必要症例数は減少する。
- ③ σ^2 が大きい値であればあるほど、必要症例数は減少する。
- ④ δ が小さい値であればあるほど、必要症例数は増加する。

[3] $\delta = 3.1$ と設定し、また $\sigma = 4.2$ を既知としたときの必要症例数として、次の①～⑤のうちから最も適切なものを一つ選べ。ただし、両群の症例数は同じであるとし、必要症例数は両群合わせた数とする。 18

- ① 22
- ② 24
- ③ 46
- ④ 58
- ⑤ 66

問9 Web調査によって10個の項目に関してデータ収集を行った。すべての項目は、1から5までの順序のある回答カテゴリのうちから1つを選ぶ形式であった。また、過去の研究から、10個の項目のうち、1番目から5番目の項目はある因子を、6番目から10番目の項目は別の因子をそれぞれ測定すること、および因子間には中程度の負の相関があることがわかっている。

なお、いくつかの項目は逆転項目になっている。逆転項目とは、他の項目とは逆のことを尋ねている項目である。たとえば、「友達は少ない方である」という項目により外向性を測定した場合、この項目は逆転項目である。逆転項目の場合、値を変換（1を5, 2を4, 4を2, 5を1）してから分析することもあるが、ここではそれは行っていない。

収集されたデータを調べたところ、一部の回答者はWeb調査に真面目に取り組んでおらず、同じ回答カテゴリばかりに回答する傾向があった。そこで、「データを同じ回答カテゴリにばかり回答した回答者（これをA群とする）」と「そうでない回答者（これをB群とする）」に分類してデータ分析を行った。ただし、A群の回答者は、10個の項目に対して必ずしも同じカテゴリに回答するわけではなく、たとえば3,3,3,4,3,3,3,3,3,3というように、いくつかの項目には他とは別のカテゴリに回答することもあるとする。

- [1] A群のデータとB群のデータに対して探索的因子分析を実行した。図1は固有値のスクリープロット、表1は因子パターン行列（2因子でプロマックス回転の場合）、表2は因子間相関である。図1、表1、表2で、数字の1,2はそれぞれA群とB群いずれかのデータの分析結果を表している。

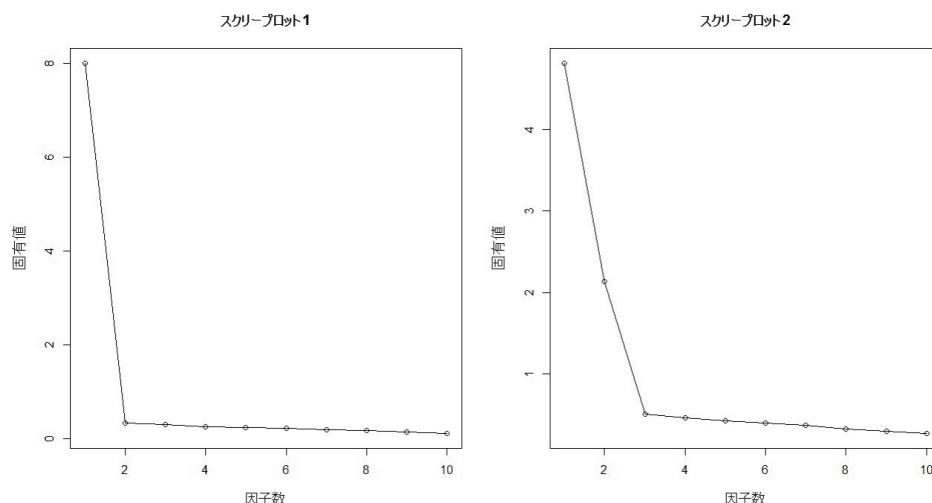


図1: 固有値のスクリープロット

表1: 因子パターン行列（プロマックス回転）

項目番号	因子パターン 1		因子パターン 2	
	因子 1	因子 2	因子 1	因子 2
1	0.02	0.76	0.31	0.61
2	0.18	-0.71	0.57	0.32
3	0.13	-0.73	0.12	0.78
4	0.23	0.88	0.73	0.20
5	-0.02	0.72	0.50	0.40
6	-0.79	0.08	0.71	0.26
7	0.78	0.00	0.31	0.62
8	0.80	0.04	0.79	0.14
9	0.81	0.08	0.36	0.50
10	0.78	-0.04	0.70	0.22

表2: 因子間相関

因子間相関 1	因子間相関 2
0.86	-0.43

B群のデータの分析結果を表しているものとして、次の①～⑤のうちから最も適切なものを一つ選べ。 19

- ① スクリープロット 1, 因子パターン 1, 因子間相関 1
- ② スクリープロット 1, 因子パターン 2, 因子間相関 1
- ③ スクリープロット 2, 因子パターン 1, 因子間相関 1
- ④ スクリープロット 2, 因子パターン 1, 因子間相関 2
- ⑤ スクリープロット 2, 因子パターン 2, 因子間相関 2

[2] 逆転項目の項目番号の組合せとして、次の①～⑤のうちから最も適切なものを一つ選べ。 20

- ① 6のみ
- ② 5と6
- ③ 2と3と6
- ④ 2と3と10
- ⑤ 2と3と5と6と10

[3] A群とB群のデータを併合した状態でも相関行列を求めたとする。なお、A群とB群でサンプルサイズは等しく、各項目の平均と標準偏差は2つのデータ間で違いはないとする。このとき、項目間の相関について、次の①～⑤のうちから最も適切なものを一つ選べ。

21

- ① A群のデータにおいて、相関が負になる項目のペアがある。
- ② B群のデータにおいて、項目1と4の相関は負になる。
- ③ B群のデータにおいて、項目1と6の相関は負になる。
- ④ B群のデータと併合データで相関の正負が異なる項目のペアはない。
- ⑤ 併合データにおいて、項目1と2の相関よりも項目1と4の相関の方が大きい。

問10 定常過程 $\{y_t\}$ に対し、時差 h の自己共分散関数 $\gamma(h)$ と自己相関関数 $\rho(h)$ は

$$\gamma(h) = \text{cov}(y_t, y_{t+h}), \quad \rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

と定義される (t と h は整数とする)。ここで $\text{cov}(y_t, y_{t+h})$ は y_t と y_{t+h} の共分散である。 $\rho(h)$ の図はコレログラムと呼ばれる。また、 $\{\varepsilon_t\}$ を独立に $N(0, \sigma^2)$ に従う誤差項とするとき、

$$y_t = \varepsilon_t + b_1\varepsilon_{t-1} + \cdots + b_q\varepsilon_{t-q}$$

で表される時系列モデルを次数 q の MA(q) モデル、

$$y_t = a_1y_{t-1} + \cdots + a_py_{t-p} + \varepsilon_t$$

で表される時系列モデルを次数 p の AR(p) モデルという。

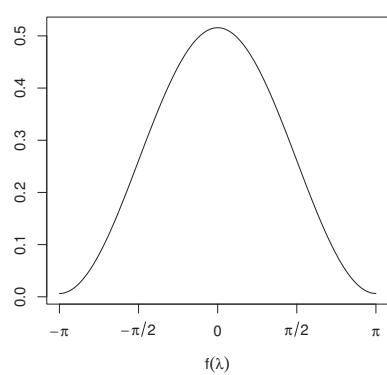
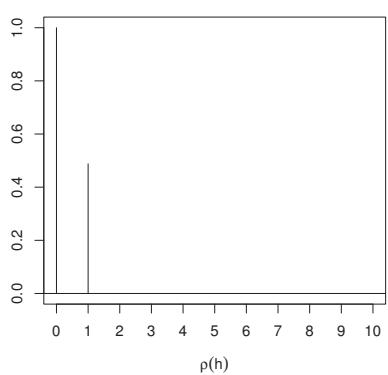
[1] $\{\varepsilon_t\}$ が独立に $N(0, 1)$ に従うときの MA(1) モデル

$$y_t = \varepsilon_t + 0.8\varepsilon_{t-1}$$

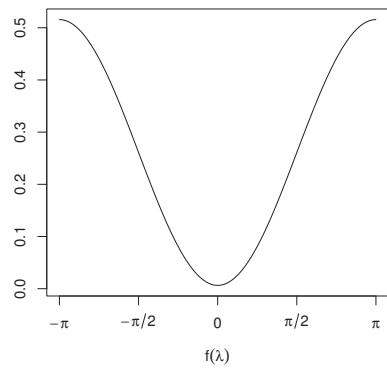
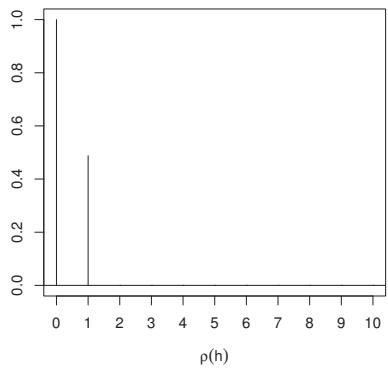
のコレログラム $\rho(h)$ とスペクトル密度関数 $f(\lambda)$ の組合せとして、次の①～④のうちから最も適切なものを一つ選べ。

22

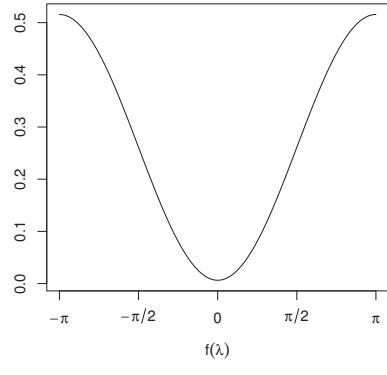
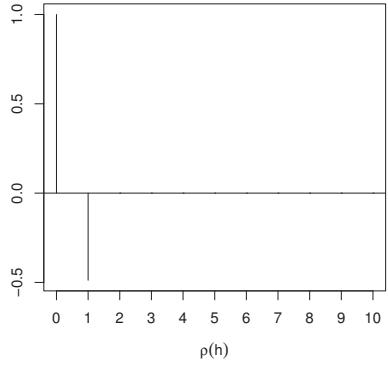
(1)



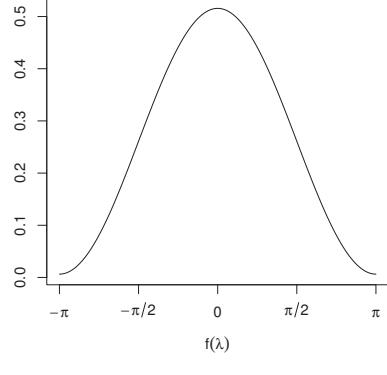
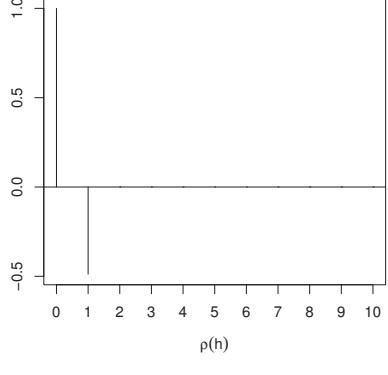
(2)



(3)



(4)

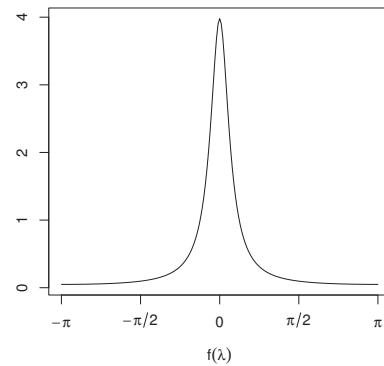
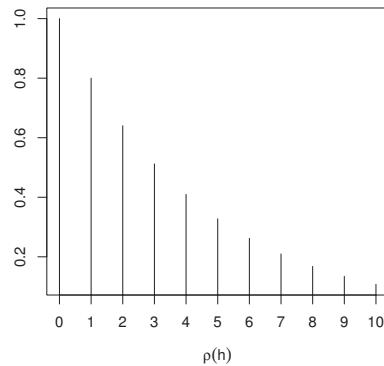


[2] $\{\varepsilon_t\}$ が独立に $N(0, 1)$ に従うときの AR(1) モデル

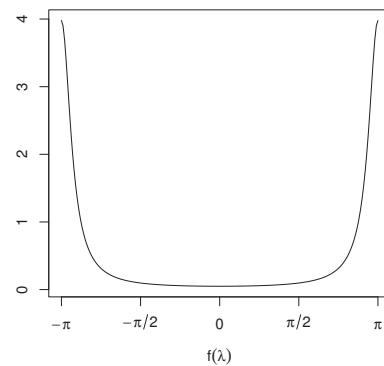
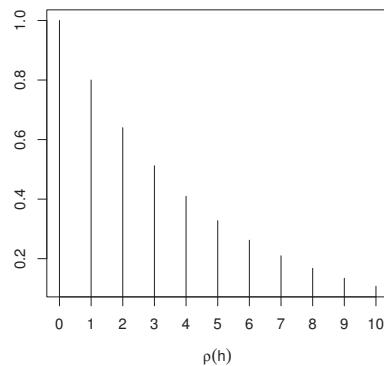
$$y_t = -0.8y_{t-1} + \varepsilon_t$$

のコレログラム $\rho(h)$ とスペクトル密度関数 $f(\lambda)$ の組合せとして、次の ① ~ ④ のうちから最も適切なものを一つ選べ。 23

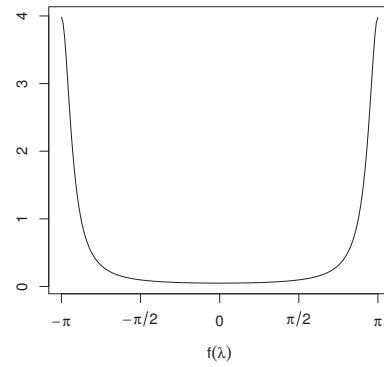
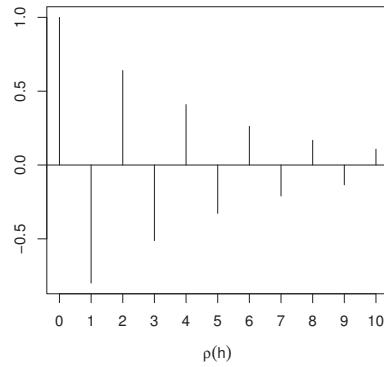
①



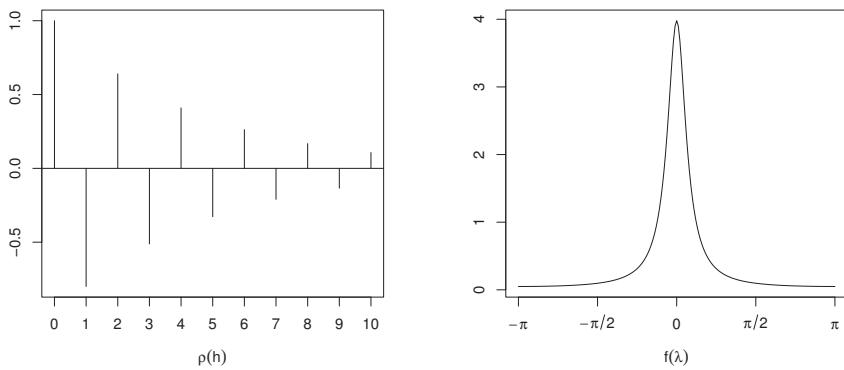
②



③



④



[3] $\{y_t\}$ の標本平均を $\bar{y}_n = \frac{1}{n} \sum_{t=1}^n y_t$ と定義する。 $\{\varepsilon_t\}$ が独立に $N(0, \sigma^2)$ に従うときの MA(1) モデル

$$y_t = \varepsilon_t + 0.8\varepsilon_{t-1}$$

における

$$\lim_{n \rightarrow \infty} \frac{nV(\bar{y}_n)}{V(y_t)}$$

の値として、次の ① ~ ④ のうちから最も適切なものを一つ選べ。ここで $V(\bar{y}_n)$ と $V(y_t)$ はそれぞれ \bar{y}_n と y_t の分散である。 **24**

① 0.02

② 0.51

③ 1.98

④ 41

[4] $\{\varepsilon_t\}$ が独立に $N(0, \sigma^2)$ に従うときの AR(2) モデル

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t$$

において、 $\rho(1) = 0.5$, $\rho(2) = -0.25$ となるとき、AR 係数 a_1, a_2 の値の組合せとして、次の ① ~ ④ のうちから最も適切なものを一つ選べ。 **25**

① $a_1 = 1.2, a_2 = -1.5$

② $a_1 = -1.5, a_2 = 1.2$

③ $a_1 = -0.67, a_2 = 0.83$

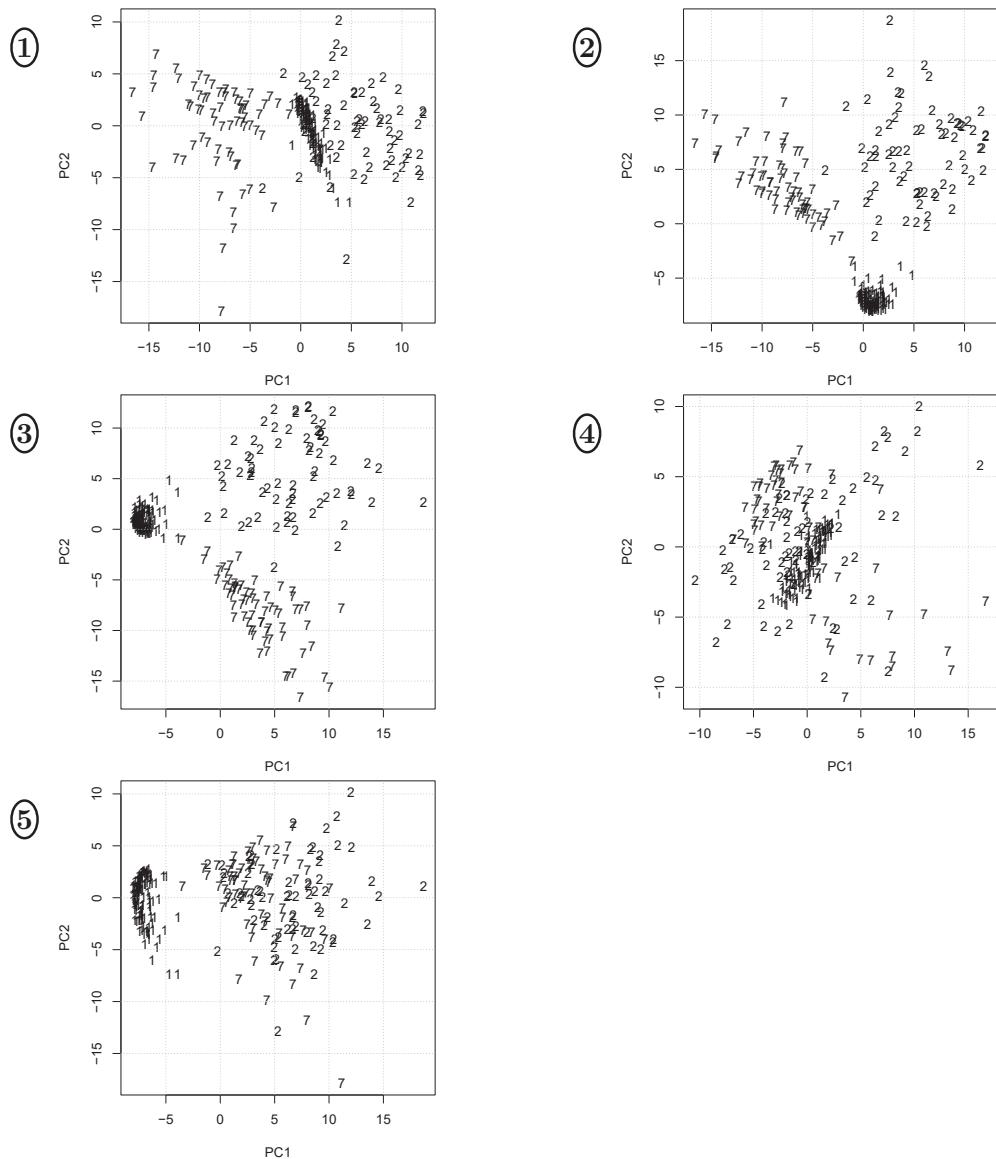
④ $a_1 = 0.83, a_2 = -0.67$

問11 16×16 ピクセルからなる0～9の手書き数字画像（資料：Le Cun et al. 1990に基づくデータ <https://web.stanford.edu/~hastie/ElemStatLearn/>）のうち、1, 2, 7の書かれた画像2381枚のデータに対して、相関行列に基づく主成分分析を行った。そして、1, 2, 7の書かれた画像をランダムに3個ずつ選び、その第1主成分得点(PC1)と第2主成分得点(PC2)を記したのが次の表である。

文字	1	1	1	2	2	2	7	7	7
PC1	-6.57	-7.86	-7.41	6.31	6.72	-0.24	2.94	1.24	8.04
PC2	0.12	1.03	0.48	9.84	1.22	6.28	-8.52	-7.54	-9.53

- [1] 横軸をPC1、縦軸をPC2としたとき、ランダムに選んだ250個の画像の主成分得点のプロットとして、次の①～⑤のうちから最も適切なものを一つ選べ。

26



[2] 主成分分析は、高次元空間に散らばるデータを特徴付ける低次元空間を推定する方法として解釈できる。低次元空間を推定するための方法として自己符号化器を考えよう。

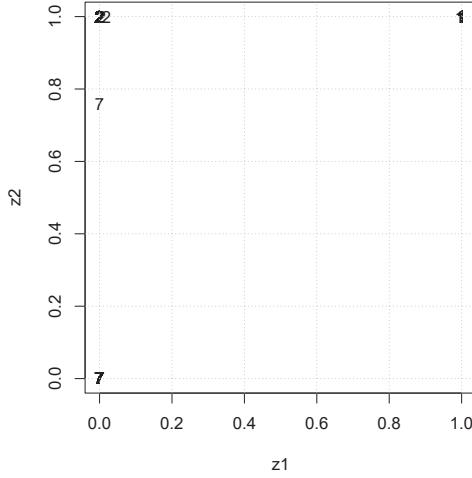
次の文章は自己符号化器についての記述である。空欄「ア」～「エ」に当てはまる単語の組合せとして、下の①～⑤のうちから最も適切なものを一つ選べ。

27

自己符号化器は、 p 次元の入力ベクトル \mathbf{x} を、 $p > q$ であるような q 次元空間へ変換する [ア] $z = f_{\theta}(\mathbf{x})$ と、 q 次元空間を p 次元空間へ変換する [イ] $\mathbf{x}' = g_{\phi}(z)$ からなる [ウ] である。ここで、 f_{θ} および g_{ϕ} は [エ] と呼ばれる非線形関数を用いて定義される関数であり、 \mathbf{x}' と \mathbf{x} が近くなるよう定められる。 θ および ϕ はそれぞれ [ア] と [イ] のパラメータベクトルである。

- ① ア：符号化器、イ：復号化器、ウ：ニューラルネットワーク、
エ：活性化関数
- ② ア：符号化器、イ：復号化器、ウ：ニューラルネットワーク、
エ：指示関数
- ③ ア：復号化器、イ：符号化器、ウ：決定木、
エ：活性化関数
- ④ ア：符号化器、イ：復号化器、ウ：決定木、
エ：指示関数
- ⑤ ア：復号化器、イ：符号化器、ウ：ニューラルネットワーク、
エ：活性化関数

[3] $q = 2$ として自己符号化による低次元空間を推定したところ、低次元空間での表のデータの散布図は次の図のようになった。



自己符号化器では、通常は誤差逆伝播法を用いてパラメータを推定する。誤差逆伝播法では、通常はデータを無作為に並べ替え、(並べ替えたデータに対して)順番にパラメータを更新する。このようなパラメータの更新規則を確率的勾配降下法とよぶ。

自己符号化器で最小化すべき目的関数を

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) = \frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|^2 = \frac{1}{2} \|\mathbf{x} - g_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(\mathbf{x}))\|^2$$

とする。 t ステップでのパラメータベクトルを $\boldsymbol{\theta}_t, \boldsymbol{\phi}_t$ とし、 $\alpha (> 0)$ を学習率係数とする。このとき、入力 \mathbf{x}_t に対する確率的勾配降下法でのパラメータの更新規則は

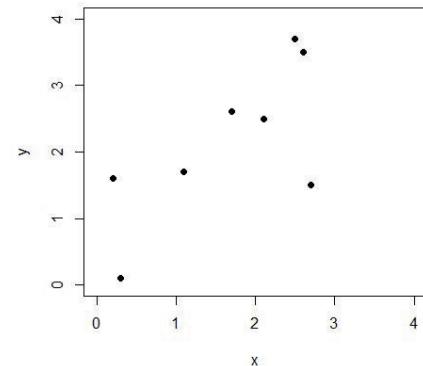
$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t - \alpha F(\mathbf{x}_t), \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha G(\mathbf{x}_t)$$

で与えられる。 $F(\mathbf{x}_t)$ および $G(\mathbf{x}_t)$ の組合せとして、次の①～④のうちから最も適切なものを一つ選べ。ただし、 $\partial_{\boldsymbol{\theta}}$ および $\partial_{\boldsymbol{\phi}}$ は、それぞれパラメータベクトル $\boldsymbol{\theta}$ および $\boldsymbol{\phi}$ でのベクトル偏微分を表す。 28

- | | | |
|---|--|---|
| ① | $F(\mathbf{x}_t) = \partial_{\boldsymbol{\phi}} L(\boldsymbol{\theta}_t, \boldsymbol{\phi}; \mathbf{x}_t) _{\boldsymbol{\phi}=\boldsymbol{\phi}_t},$ | $G(\mathbf{x}_t) = \partial_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\phi}_t; \mathbf{x}_t) _{\boldsymbol{\theta}=\boldsymbol{\theta}_t}$ |
| ② | $F(\mathbf{x}_t) = \partial_{\boldsymbol{\phi}} L(\boldsymbol{\theta}_t, \boldsymbol{\phi}; \mathbf{x}_t) _{\boldsymbol{\phi}=\boldsymbol{\phi}_t},$ | $G(\mathbf{x}_t) = \partial_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\phi}_t; \mathbf{x}_t) _{\boldsymbol{\theta}=\boldsymbol{\theta}_{t+1}}$ |
| ③ | $F(\mathbf{x}_t) = \partial_{\boldsymbol{\phi}} L(\boldsymbol{\theta}_t, \boldsymbol{\phi}; \mathbf{x}_t) _{\boldsymbol{\phi}=\boldsymbol{\phi}_{t+1}},$ | $G(\mathbf{x}_t) = \partial_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\phi}_t; \mathbf{x}_t) _{\boldsymbol{\theta}=\boldsymbol{\theta}_t}$ |
| ④ | $F(\mathbf{x}_t) = \partial_{\boldsymbol{\phi}} L(\boldsymbol{\theta}_t, \boldsymbol{\phi}; \mathbf{x}_t) _{\boldsymbol{\phi}=\boldsymbol{\phi}_{t+1}},$ | $G(\mathbf{x}_t) = \partial_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\phi}_t; \mathbf{x}_t) _{\boldsymbol{\theta}=\boldsymbol{\theta}_{t+1}}$ |

問12 次の表にある2つの量的変数 x, y についてサンプルサイズ8のデータが得られたとする。表の右側の図は、このデータの散布図を示している。

サンプル番号	変数 x	変数 y
1	0.2	1.6
2	0.3	0.1
3	1.1	1.7
4	1.7	2.6
5	2.1	2.5
6	2.5	3.7
7	2.6	3.5
8	2.7	1.5



2変量正規分布に従う母集団からサイズ8の無作為標本を抽出し、「帰無仮説：母相関係数 = 0, 対立仮説：母相関係数 ≠ 0」を有意水準5%で検定する問題を考える。表のデータをこの無作為標本の実現値とみなしたとき、次の文中の空欄を埋めよ。ただし、小数点以下3位を四捨五入せよ。変数 x の偏差平方和が 7.16, 変数 y の偏差平方和が 9.68, 変数 x と変数 y の偏差積和が 5.91 であることは用いてよい。

ピアソンの相関係数を R としたとき、検定統計量を $T = \sqrt{\frac{6R^2}{1-R^2}}$ とすると、検定方式を「 $T \geq$ **記述6** のとき帰無仮説を棄却」とできる。表のデータから計算した T の実現値は **記述7** であるから、帰無仮説を棄却する。

論述問題

(3 問中 1 問選択)

問1 $n = 0, 1, 2, \dots$ に対して, X_n と Y_n は整数値をとる確率変数（あるいは確率ベクトル）とする。

X_0, X_1, X_2, \dots は次の性質をもつものとする：任意の $n (\geq 1), i, j, i_0, i_1, \dots, i_{n-2}$ に對して

$$P(X_n = j | X_{n-1} = i, X_{n-2} = i_{n-2}, \dots, X_1 = i_1, X_0 = i_0) = P(X_n = j | X_{n-1} = i)$$

が成り立つ。このような性質をもつ $\{X_n\}$ はマルコフ性をもつという。

一方, Y_0, Y_1, Y_2, \dots は次の性質をもつものとする：任意の $n (\geq 1)$ に對して期待値は有界であり,

$$E(Y_n | Y_{n-1}, Y_{n-2}, \dots, Y_1, Y_0) = Y_{n-1}$$

が成り立つ。このような性質をもつ $\{Y_n\}$ をマルチングールという。

いま, X_n の値が i であったとき, X_{n+1} の値が j となる（条件つき）確率は n には依存しないと仮定し, この値を $p_{i,j}$ で表す, つまり

$$p_{i,j} = P(X_{n+1} = j | X_n = i)$$

とする。そして, $P(X_0 = 0) = 1$ を満たし, $\{p_{i,j}\}$ が

$$\begin{aligned} p_{i,i+1} &= p_{i,i-1} = \frac{1}{2}, & p_{i,i} &= 0, & i &= 0, \pm 1, \pm 2, \dots \\ p_{i,j} &= 0, & |i - j| &\geq 2 \end{aligned}$$

で与えられるとする。さらに, $S_n = \sum_{k=0}^n X_k$ とする。

[1] 確率過程 $\{S_n\}_{n=0,1,2,\dots}$ がマルコフ性をもつかどうか, マルチングールかどうかを調べよ。

[2] 確率ベクトル過程 $\{(X_n, S_n)\}_{n=0,1,2,\dots}$ がマルコフ性をもつかどうか, マルチングールかどうかを調べよ。

[3] $T_0 = X_0, T_1 = X_1, T_n = T_{n-1} + (X_n - X_{n-1})T_{n-2} (n = 2, 3, 4, \dots)$ としたとき, 確率過程 $\{T_n\}_{n=0,1,2,\dots}$ がマルコフ性をもつかどうか, マルチングールかどうかを調べよ。

問2 ベイズ法に関する次の各間に答えよ。

[1] 確率変数 X は、ある事象の生起確率 θ が未知の二項分布 $\text{Bin}(n, \theta)$ に従うものとする。また、 θ の事前分布としてベータ分布 $\text{Be}(\alpha_0, \beta_0)$, $\alpha_0 > 0$, $\beta_0 > 0$ を仮定すると、事後分布もベータ分布となり、これを $\text{Be}(\alpha_1, \beta_1)$ と表す。ここで、 θ がベータ分布 $\text{Be}(\alpha, \beta)$ に従うとき、その確率密度関数はベータ関数 $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$ を用いて

$$f(\theta|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \quad (0 \leq \theta \leq 1, \alpha > 0, \beta > 0)$$

で与えられる。

[1-1] 上の二項分布とベータ分布のように、標本の確率分布のパラメータに対して事前分布と事後分布が同一の分布族となるような性質を持つ事前分布を共役事前分布と呼ぶ。標本の確率分布と共役事前分布の組合せとして、次の(A)～(C)のうち正しいものののみをすべて挙げよ。

	標本の確率分布	共役事前分布
(A)	ポアソン分布	ガンマ分布
(B)	正規分布 (平均未知, 分散既知)	正規分布
(C)	正規分布 (平均既知, 分散未知)	ベータ分布

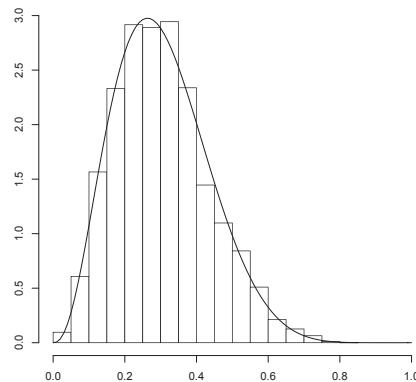
[1-2] 観測データ $X = x_0$ が得られたとき、事後分布のベータ分布のパラメータ α_1, β_1 を $x_0, n, \alpha_0, \beta_0$ を用いて表せ。

[1-3] $\alpha_0 > 1, \beta_0 > 1$ とする。観測データ $X = x_0$ が得られたとき、 θ の事後密度関数の値を最大とする θ を $x_0, n, \alpha_0, \beta_0$ を用いて表せ。

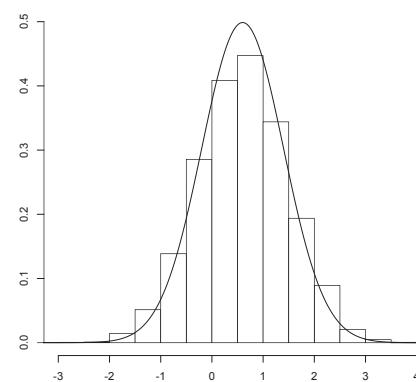
[2] 確率変数 X_1, X_2, X_3, X_4 が未知の平均 μ をもつ正規分布 $N(\mu, 4)$ に独立に従っているとする。また、 μ の事前分布として正規分布 $N(0, 1)$ を仮定する。

[2-1] 1回の観測を行い、 $X_1 = 3.0$ が得られた。このとき、醉歩連鎖によるメトロポリス・ヘイスティングス法を用いて μ の事後分布 $f(\mu|X_1 = 3.0)$ に従う乱数を 11000 個 ($\mu^{(1)}, \mu^{(2)}, \dots, \mu^{(11000)}$ と表す) 発生させ、 $\mu^{(1)}$ から $\mu^{(1000)}$ までの 1000 個を除いた残りの 10000 個の乱数から μ の事後分布を考える。下図の (A) ~ (D) のいずれかは発生させた 10000 個の乱数 $\mu^{(1001)}, \mu^{(1002)}, \dots, \mu^{(11000)}$ のヒストグラムを描いたものである。(A)~(D) のうち発生させた乱数のヒストグラムとして正しいものはどれか。理由も含めて答えよ。

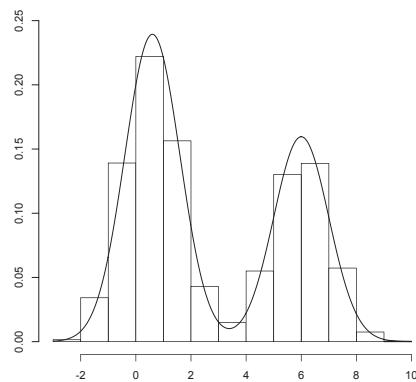
(A)



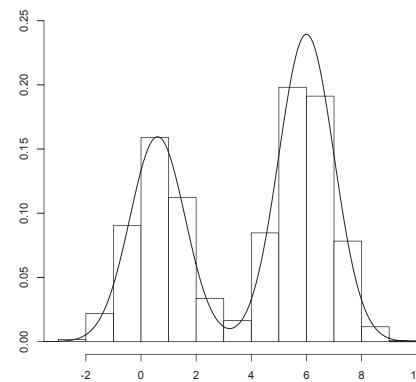
(B)



(C)



(D)



[2-2] あらためて観測を行い、 $X_2 = 2.3, X_3 = 4.2, X_4 = 1.5$ を得た。このとき、事後分布 $f(\mu|X_1 = 3.0)$ を μ の新たな事前分布とし、事後分布 $f(\mu|X_2 = 2.3, X_3 = 4.2, X_4 = 1.5)$ を求めよ。

問3 被験者 n 人に対し血圧を測定し、血圧が 130 (mmHg) 以上だった被験者を高血圧群と呼び、高血圧群の全被験者のみについて後日再測定した。1回目の血圧及び2回目の血圧を表す確率変数をそれぞれ Y_{1i}, Y_{2i} ($i = 1, 2, \dots, n$) とする。また、高血圧群かどうかを表す指示変数を H_i ($i = 1, 2, \dots, n$) とし、高血圧であれば 1、そうでなければ 0 を取る 2 値変数とする。 (Y_{1i}, Y_{2i}, H_i) は独立同一分布に従うとし、この実現値を (y_{1i}, y_{2i}, h_i) と表すことにする。 $h_i = 1$ である被験者は m 人であったとし、始めの m 人が高血圧群、残りの $n - m$ 人が非高血圧群となるようにデータを並び替えておく。

図 1 は高血圧群の観測データ y_{1i}, y_{2i} ($i = 1, 2, \dots, m$) の散布図である。直線はそれに対する単回帰直線である。

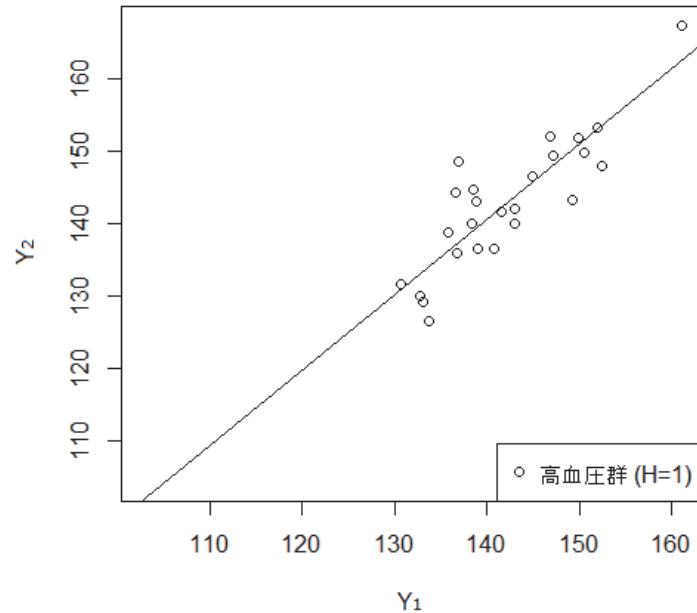


図 1: 高血圧群の観測データの散布図と回帰直線

図 2 は高血圧群に加え、非高血圧群のデータも含めた y_{1i}, y_{2i} ($i = 1, 2, \dots, n$) の散布図である。非高血圧群の2回目のデータは本当は観測されていない。破線は（その観測されていないデータも含めた）すべてのデータを用いて得られる単回帰直線である。

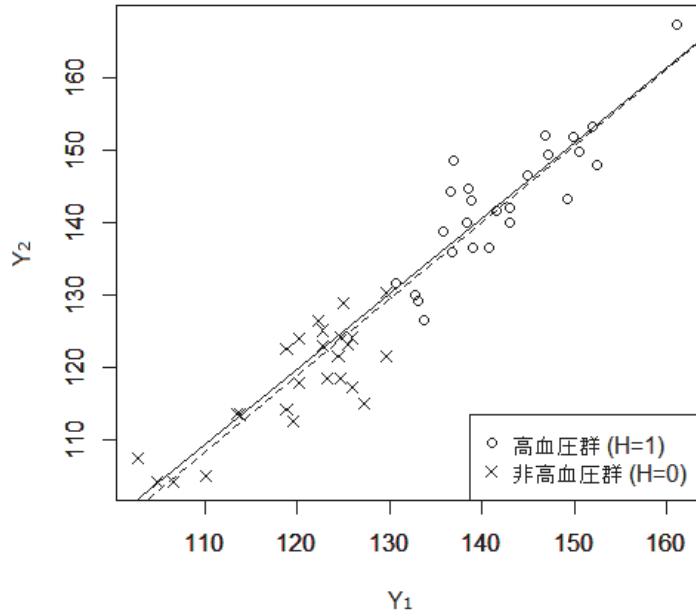


図 2: 高血圧群と非高血圧群を合わせたデータの散布図と回帰直線

- [1] 図 2 で、高血圧群のみのデータを用いた单回帰直線と、すべてのデータを用いた单回帰直線は、似たような直線となっている。実際、高血圧群のデータにおいて、関係式

$$E(Y_{2i} \mid Y_{1i} = y_1) = E(Y_{2i} \mid Y_{1i} = y_1, H_i = 1)$$

が成り立つので、单回帰分析が妥当であるならば両直線は似たものとなる。この関係式が成り立つことを示せ。

- [2] 1回目と2回目の血圧の関係に興味があるとし、次の单回帰分析を考える。

$$E(Y_{2i} \mid Y_{1i} = y_1; \boldsymbol{\beta}) = \beta_0 + \beta_1 y_1$$

ここで、 $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top \in \mathbb{R}^2$ 、 $^\top$ は転置を表すとする。いま非高血圧群の2回目のデータは観測されないため、存在している高血圧群のみのデータを用いて最小二乗法で回帰係数を推定した。このとき、この最小二乗推定値 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^\top$ は $\boldsymbol{\beta}$ の不偏推定値であることを示せ。

- [3] 2回目の血圧に関する期待値 $\mu_2 = E(Y_{2i})$ を、観測データだけを用いて推定したい。そのため、観測されていない非高血圧群のデータ y_{2i} を [2] で推定した回帰直線による予測値

$$\hat{y}_{2i} = \hat{\beta}_0 + \hat{\beta}_1 y_{1i} \quad (i = m+1, \dots, n)$$

で代用し、図3のような擬似的な完全データを作成した。いま、 μ_2 を2回目の測定値に対する擬似的な完全データから計算される標本平均 $\hat{\mu}_2$ で推定することを考える。このとき、 $\hat{\mu}_2$ は μ_2 の不偏推定値であることを示せ。

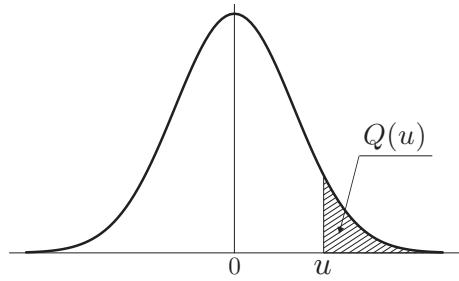
Y_1	Y_2	H
y_{11}	y_{21}	1
\vdots	\vdots	\vdots
y_{1m}	y_{2m}	1
欠測値を予測値で置き換え		
$y_{1(m+1)}$	$\hat{y}_{2(m+1)}$	0
\vdots	\vdots	\vdots
y_{1n}	\hat{y}_{2n}	0

図3: 2回目の測定の欠測値を单回帰による予測値で埋めた擬似的な完全データ

- [4] 推定値 $\hat{\beta}$ の推定精度を上げるために、[3] で得られた擬似的な完全データを用いて最小二乗推定を行うことを考える。このとき、擬似的な完全データから計算される最小二乗推定値 $\hat{\beta}_{\text{imp}}$ は [2] で計算された最小二乗推定値 $\hat{\beta}$ より良い推定値となるかどうか、両推定値を比較することで調べよ。

付 表

付表 1. 標準正規分布の上側確率

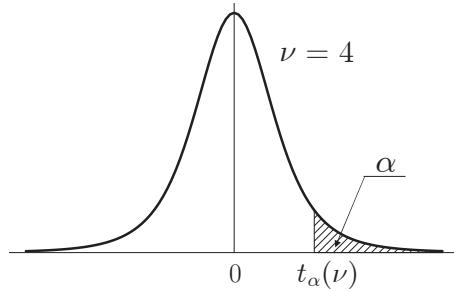


u	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

$u = 0.00 \sim 3.99$ に対する、正規分布の上側確率 $Q(u)$ を与える。

例 : $u = 1.96$ に対しては、左の見出し 1.9 と上の見出し .06 との交差点で、 $Q(u) = 0.0250$ と読む。表にない u に対しては適宜補間すること。

付表 2. t 分布のパーセント点

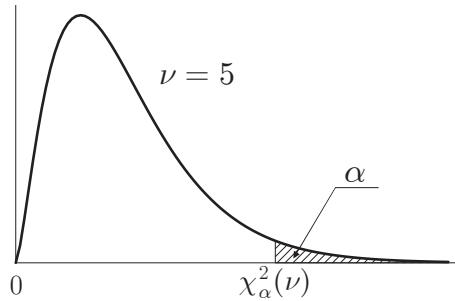


ν	α				
	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
240	1.285	1.651	1.970	2.342	2.596
∞	1.282	1.645	1.960	2.326	2.576

自由度 ν の t 分布の上側確率 α に対する t の値を $t_\alpha(\nu)$ で表す。

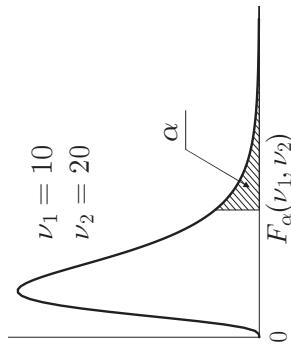
例：自由度 $\nu = 20$ の上側 5% 点 ($\alpha = 0.05$) は、 $t_{0.05}(20) = 1.725$ である。
表にない自由度に対しては適宜補間すること。

付表3. カイ二乗分布のパーセント点



ν	α							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
35	18.51	20.57	22.47	24.80	46.06	49.80	53.20	57.34
40	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69
50	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
60	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
70	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
80	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
90	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
100	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81
120	86.92	91.57	95.70	100.62	140.23	146.57	152.21	158.95
140	104.03	109.14	113.66	119.03	161.83	168.61	174.65	181.84
160	121.35	126.87	131.76	137.55	183.31	190.52	196.92	204.53
180	138.82	144.74	149.97	156.15	204.70	212.30	219.04	227.06
200	156.43	162.73	168.28	174.84	226.02	233.99	241.06	249.45
240	191.99	198.98	205.14	212.39	268.47	277.14	284.80	293.89

自由度 ν のカイ二乗分布の上側確率 α に対する χ^2 の値を $\chi^2_\alpha(\nu)$ で表す。
例：自由度 $\nu = 20$ の上側 5% 点 ($\alpha = 0.05$) は、 $\chi^2_{0.05}(20) = 31.41$ である。
表にない自由度に対しては適宜補間すること。

付表4. F 分布のパーセント点 $\alpha = 0.05$

$\nu_2 \setminus \nu_1$	1	2	3	4	5	6	7	8	9	10	15	20	40	60	120	∞
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.619	4.558	4.464	4.431	4.398	4.365
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.845	2.774	2.661	2.621	2.580	2.538
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.403	2.328	2.204	2.160	2.114	2.066
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.203	2.124	1.994	1.946	1.896	1.843
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.089	2.007	1.872	1.822	1.768	1.711
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.015	1.932	1.792	1.740	1.683	1.622
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	1.924	1.839	1.693	1.637	1.577	1.509
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993	1.836	1.748	1.594	1.534	1.467	1.389
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910	1.750	1.659	1.495	1.429	1.352	1.254

 $\alpha = 0.025$

$\nu_2 \setminus \nu_1$	1	2	3	4	5	6	7	8	9	10	15	20	40	60	120	∞
5	10.007	8.434	7.764	7.388	7.146	6.978	6.853	6.757	6.681	6.619	6.428	6.329	6.175	6.123	6.069	6.015
10	6.937	5.456	4.826	4.468	4.236	4.072	3.950	3.855	3.779	3.717	3.522	3.419	3.255	3.198	3.140	3.080
15	6.200	4.765	4.153	3.804	3.576	3.415	3.293	3.199	3.123	3.060	2.862	2.756	2.585	2.524	2.461	2.395
20	5.871	4.461	3.859	3.515	3.289	3.128	3.007	2.913	2.837	2.774	2.573	2.464	2.287	2.223	2.156	2.085
25	5.686	4.291	3.694	3.353	3.129	2.969	2.848	2.753	2.677	2.613	2.411	2.300	2.118	2.052	1.981	1.906
30	5.568	4.182	3.589	3.250	3.026	2.867	2.746	2.651	2.575	2.511	2.307	2.195	2.009	1.940	1.866	1.787
40	5.424	4.051	3.463	3.126	2.904	2.744	2.624	2.529	2.452	2.388	2.182	2.068	1.875	1.803	1.724	1.637
60	5.286	3.925	3.343	3.008	2.786	2.627	2.507	2.412	2.334	2.270	2.061	1.944	1.744	1.667	1.581	1.482
120	5.152	3.805	3.227	2.894	2.674	2.515	2.395	2.299	2.222	2.157	1.945	1.825	1.614	1.530	1.433	1.310

自由度 (ν_1, ν_2) の F 分布の上側確率 α に対する F の値を $F_\alpha(\nu_1, \nu_2)$ で表す。例：自由度 $\nu_1 = 5, \nu_2 = 20$ の上側 5% 点 ($\alpha = 0.05$) は、 $F_{0.05}(5, 20) = 2.711$ である。

表にない自由度に対しては適宜補間すること。

付表 5. 指数関数と常用対数

指数関数				常用対数			
x	e^x	x	e^x	x	$\log_{10} x$	x	$\log_{10} x$
0.01	1.0101	0.51	1.6653	0.1	-1.0000	5.1	0.7076
0.02	1.0202	0.52	1.6820	0.2	-0.6990	5.2	0.7160
0.03	1.0305	0.53	1.6989	0.3	-0.5229	5.3	0.7243
0.04	1.0408	0.54	1.7160	0.4	-0.3979	5.4	0.7324
0.05	1.0513	0.55	1.7333	0.5	-0.3010	5.5	0.7404
0.06	1.0618	0.56	1.7507	0.6	-0.2218	5.6	0.7482
0.07	1.0725	0.57	1.7683	0.7	-0.1549	5.7	0.7559
0.08	1.0833	0.58	1.7860	0.8	-0.0969	5.8	0.7634
0.09	1.0942	0.59	1.8040	0.9	-0.0458	5.9	0.7709
0.10	1.1052	0.60	1.8221	1.0	0.0000	6.0	0.7782
0.11	1.1163	0.61	1.8404	1.1	0.0414	6.1	0.7853
0.12	1.1275	0.62	1.8589	1.2	0.0792	6.2	0.7924
0.13	1.1388	0.63	1.8776	1.3	0.1139	6.3	0.7993
0.14	1.1503	0.64	1.8965	1.4	0.1461	6.4	0.8062
0.15	1.1618	0.65	1.9155	1.5	0.1761	6.5	0.8129
0.16	1.1735	0.66	1.9348	1.6	0.2041	6.6	0.8195
0.17	1.1853	0.67	1.9542	1.7	0.2304	6.7	0.8261
0.18	1.1972	0.68	1.9739	1.8	0.2553	6.8	0.8325
0.19	1.2092	0.69	1.9937	1.9	0.2788	6.9	0.8388
0.20	1.2214	0.70	2.0138	2.0	0.3010	7.0	0.8451
0.21	1.2337	0.71	2.0340	2.1	0.3222	7.1	0.8513
0.22	1.2461	0.72	2.0544	2.2	0.3424	7.2	0.8573
0.23	1.2586	0.73	2.0751	2.3	0.3617	7.3	0.8633
0.24	1.2712	0.74	2.0959	2.4	0.3802	7.4	0.8692
0.25	1.2840	0.75	2.1170	2.5	0.3979	7.5	0.8751
0.26	1.2969	0.76	2.1383	2.6	0.4150	7.6	0.8808
0.27	1.3100	0.77	2.1598	2.7	0.4314	7.7	0.8865
0.28	1.3231	0.78	2.1815	2.8	0.4472	7.8	0.8921
0.29	1.3364	0.79	2.2034	2.9	0.4624	7.9	0.8976
0.30	1.3499	0.80	2.2255	3.0	0.4771	8.0	0.9031
0.31	1.3634	0.81	2.2479	3.1	0.4914	8.1	0.9085
0.32	1.3771	0.82	2.2705	3.2	0.5051	8.2	0.9138
0.33	1.3910	0.83	2.2933	3.3	0.5185	8.3	0.9191
0.34	1.4049	0.84	2.3164	3.4	0.5315	8.4	0.9243
0.35	1.4191	0.85	2.3396	3.5	0.5441	8.5	0.9294
0.36	1.4333	0.86	2.3632	3.6	0.5563	8.6	0.9345
0.37	1.4477	0.87	2.3869	3.7	0.5682	8.7	0.9395
0.38	1.4623	0.88	2.4109	3.8	0.5798	8.8	0.9445
0.39	1.4770	0.89	2.4351	3.9	0.5911	8.9	0.9494
0.40	1.4918	0.90	2.4596	4.0	0.6021	9.0	0.9542
0.41	1.5068	0.91	2.4843	4.1	0.6128	9.1	0.9590
0.42	1.5220	0.92	2.5093	4.2	0.6232	9.2	0.9638
0.43	1.5373	0.93	2.5345	4.3	0.6335	9.3	0.9685
0.44	1.5527	0.94	2.5600	4.4	0.6435	9.4	0.9731
0.45	1.5683	0.95	2.5857	4.5	0.6532	9.5	0.9777
0.46	1.5841	0.96	2.6117	4.6	0.6628	9.6	0.9823
0.47	1.6000	0.97	2.6379	4.7	0.6721	9.7	0.9868
0.48	1.6161	0.98	2.6645	4.8	0.6812	9.8	0.9912
0.49	1.6323	0.99	2.6912	4.9	0.6902	9.9	0.9956
0.50	1.6487	1.00	2.7183	5.0	0.6990	10.0	1.0000

注: 常用対数を自然対数に直すには 2.3026 をかけねばよい。

【解答用紙記入例】

- マークシートの記入例

(例) **10** と表示のある間にに対して**③**と解答する場合

次のように解答番号**10**の解答の**③**にマークすること。

解答番号	解 答
10	① ② <input checked="" type="radio"/> ④ ⑤

- 論述問題解答面のページ先頭の記入例

(例) 論述問題問**1**を解答する場合

次のように問題番号に解答する問題番号を記入し、得点の欄には何も書かないこと。

統計検定 準1級 論述問題 解答面

問題番号
1

※選択した問題番号を記入すること
※2問以上解答した場合は採点対象としない
※以下は自由に使ってよい（裏面も可）

得点3

著作権法により、本冊子の無断での複製・転載等は禁止されています。

一般財団法人 統計質保証推進協会

統計検定センター

〒101-0051 東京都千代田区神田神保町3丁目6番

URL <http://www.toukei-kentei.jp>

2021.6