巻末付録

準1級
例題／解説

2015年6月から開始される準1級の例題です。
「選択問題及び部分記述問題」と「論述問題」からなります。
部分記述問題は，記述4 のように記載されているので，
解答用紙の指定されたスペースに解答を記入します。
論述問題は3題中1題を選択解答します。配点は全体の3割程度です。
準1級の合格水準は，この例題集において6割程度を想定しています。

※実際の問題冊子には統計数値表が掲載されます。

選択問題及び部分記述問題　例題

問 1　A, B, C の 3 つの事象について次のように確率が与えられている。

$$P(A) = 0.45, P(B) = 0.45, P(C) = 0.4, P(A \cap B) = 0.2,$$
$$P(B \cap C) = 0.1, P(A \cap C) = 0.15, P(A \cap B \cap C) = 0.05$$

〔1〕$P(A \cup B)$ として正しいものを次の ① ～ ⑤ のうちから一つ選べ。 **1**

① 0.5　② 0.6　③ 0.7　④ 0.8　⑤ 0.9

〔2〕$P(A \cup B \cup C)$ として正しいものを次の ① ～ ⑤ のうちから一つ選べ。 **2**

① 0.8　② 0.85　③ 0.9　④ 0.95　⑤ 1.0

問 2　次の表はある地域における 1 日の死亡者数の集計結果（500 日間）である。

死亡者数（人）	0	1	2	3	4	5	6 以上	計
件数（日数）	55	144	140	95	45	15	6	500

〔1〕1 日の死亡者数 X がパラメータ λ のポアソン分布に従うと仮定するとき，ある日の死亡者数が 3 人である確率を求める式として正しいものはどれか。次の ① ～ ⑤ のうちから適切なものを一つ選べ。 **3**

① λ　② $\lambda(1-\lambda)^3$　③ ${}_6C_3\lambda^3(1-\lambda)^3$

④ $\lambda e^{-3\lambda}$　⑤ $\dfrac{\lambda^3 e^{-\lambda}}{3!}$

〔2〕1 日の死亡者数 X がパラメータ λ のポアソン分布に従うと仮定するとき，X^2 の期待値 $E(X^2)$ とパラメータ λ の関係として正しいものはどれか。次の ① ～ ⑤ のうちから適切なものを一つ選べ。 **4**

① $E(X^2) = \dfrac{2}{\lambda^2}$　② $E(X^2) = \dfrac{1}{\lambda}$　③ $E(X^2) = \lambda$

④ $E(X^2) = \lambda^2 + \lambda$　⑤ $E(X^2) = \lambda^2$

〔3〕1 日の死亡者数 X がパラメータ λ のポアソン分布に従っているか否かの検定を行う。データの平均値をもとにパラメータを推定し，次の表の期待度数を得た。

死亡者数（人）	0	1	2	3	4	5	6 以上
期待度数（日）	67.7	135.3	135.3	90.2	45.1	18.0	8.3

このとき適合度検定の判断として，次の ① ～ ⑤ のうちから最も適切なものを一つ選べ。 **5**

① 検定統計量の χ^2 値 $=0.647$ が自由度 7 の χ^2 分布の上側 5 %点より大きいか比べる。

② 検定統計量の χ^2 値 $=0.647$ が自由度 6 の χ^2 分布の上側 5 %点より大きいか比べる。

③ 検定統計量の χ^2 値 $=4.498$ が自由度 7 の χ^2 分布の上側 5 %点より大きいか比べる。

④ 検定統計量の χ^2 値 $=4.498$ が自由度 6 の χ^2 分布の上側 5 %点より大きいか比べる。

⑤ 検定統計量の χ^2 値 $=4.498$ が自由度 5 の χ^2 分布の上側 5 %点より大きいか比べる。

問 3　$\begin{pmatrix} X \\ Y \end{pmatrix}$ を次の 2 変量正規分布に従う確率ベクトルとする。

$$N\left(\begin{pmatrix} 1.0 \\ 2.0 \end{pmatrix}, \begin{pmatrix} 3.0 & 2.0 \\ 2.0 & 4.0 \end{pmatrix}\right)$$

〔1〕 $\begin{pmatrix} X+Y \\ X-Y \end{pmatrix}$ が従う 2 変量正規分布を次の ① ～ ⑤ のうちから一つ選べ。 **6**

① $N\left(\begin{pmatrix} 3.0 \\ 1.0 \end{pmatrix}, \begin{pmatrix} 9.0 & 0.0 \\ 0.0 & 9.0 \end{pmatrix}\right)$　② $N\left(\begin{pmatrix} 3.0 \\ 1.0 \end{pmatrix}, \begin{pmatrix} 11.0 & 0.0 \\ 0.0 & 9.0 \end{pmatrix}\right)$

③ $N\left(\begin{pmatrix} 3.0 \\ -1.0 \end{pmatrix}, \begin{pmatrix} 11.0 & -1.0 \\ -1.0 & 3.0 \end{pmatrix}\right)$　④ $N\left(\begin{pmatrix} 3.0 \\ -1.0 \end{pmatrix}, \begin{pmatrix} 11.0 & -2.0 \\ -2.0 & 4.0 \end{pmatrix}\right)$

⑤ $N\left(\begin{pmatrix} 3.0 \\ -1.0 \end{pmatrix}, \begin{pmatrix} 11.0 & 0.0 \\ 0.0 & 3.0 \end{pmatrix}\right)$

〔2〕 X を与えたときの Y の条件つき分布を次の ① ～ ⑤ のうちから一つ選べ。 **7**

① $N(2.0, 4.0)$　② $N(2.0+X, 3.0)$　③ $N(1.33+X, 3.5)$

④ $N(1.33+0.67X, 2.67)$　⑤ $N(1.33+0.67X, 2.5)$

問 4　2013 年 6 月の NHK による政治意識月例調査に回答した 1008 人の内閣支持率は 62 %であった。

〔1〕母集団の内閣支持率の 95 %信頼区間を構成したい。最も適切な信頼区間を次の ① ～ ⑤ のうちから一つ選べ。 **8**

① $0.62 \pm 1.96\sqrt{0.62 \times 0.38}$

② $0.62 \pm 1.96\sqrt{0.62 \times 0.38/1008}$

③ $0.62 \pm 1.96\sqrt{0.62 \times 0.38}/1008$

④ $0.62 \pm 1.64\sqrt{0.62 \times 0.38}$

⑤ $0.62 \pm 1.64\sqrt{0.62 \times 0.38}/1008$

〔2〕次の記述 I ～ III は，信頼区間の幅を狭くするための方法に関する記述である。

I. 回答者数を増やす。
II. 回答者の若者の割合を増やす。
III. 信頼係数を大きくする。

これら記述の正誤の組合せとして，適切なものを次の ① ～ ⑤ のうちから一つ選べ。 **9**

① I のみが正しい。　② II のみが正しい。　③ III のみが正しい。
④ I と II が正しい。　⑤ I と III が正しい。

〔3〕真の内閣支持率が 60 %であるとき，内閣支持率の推定値が 62 %を超える確率は，およそいくらか。最も適切な値を，次の ① ～ ⑤ のうちから一つ選べ。 **10**

① 0.01　② 0.05　③ 0.1　④ 0.15　⑤ 0.2

〔4〕内閣支持率の 95 %信頼区間の幅が 4 %以内となるためには標本サイズは何人以上必要か。最も適切な値を，次の ① ～ ⑤ のうちから一つ選べ。 **11**

① 40　② 120　③ 600　④ 2400　⑤ 4800

問 5　ある 2 つのタイプ（A と B）の商品について，年齢層により利用傾向が異なるのではないかという意見が出された。この意見についてデータを収集して検討することとした。2 つの商品のタイプのどちらが好きかについて，2 つの年齢層（20 代と 40 代）に属する計 400 人からの回答を集計した結果，次のような分割表が得られた。これに関して下の問に答えよ。

年齢層と商品タイプの分割表

年齢層＼タイプ	A	B	計
20 代	130	110	240
40 代	70	90	160
計	200	200	400

〔1〕年齢層と商品タイプの関連について考える場合に，年齢層と商品タイプの選択が独立のとき，次の各セルの期待度数を表した分割表の（ア）と（エ）に入る値はいくつか。

（ア） **記述 1**　（エ） **記述 2**

分割表

年齢層＼タイプ	A	B	計
20 代	（ア）	（イ）	240
40 代	（ウ）	（エ）	160
計	200	200	400

〔2〕年齢層と商品タイプの独立性を検定するために，得られた分割表に対して以下のように有意水準を 5 % としてカイ 2 乗検定を行った。次の（オ）と（カ）に入る値はいくらか。

検定統計量の χ^2 値を計算すると χ^2 値は（オ）となった。カイ 2 乗分布表より得られる棄却点（カ）と比較して χ^2 値のほうが大きいので，年齢層により商品の好みに差があると結論する。

（オ） **記述 3**　（カ） **記述 4**

問 6　次のような 2×3 分割表を考える。行和及び列和を固定したもとで x_{11} のとりうる範囲として正しいものを下の ① ～ ⑤ のうちから一つ選べ。 **12**

x_{11}	x_{12}	x_{13}	12
x_{21}	x_{22}	x_{23}	8
10	2	8	20

① $0 \leq x_{11} \leq 10$　② $1 \leq x_{11} \leq 10$　③ $2 \leq x_{11} \leq 10$

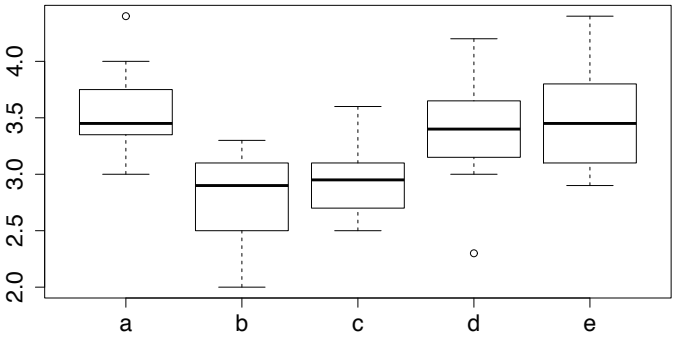
④ $3 \leq x_{11} \leq 10$　⑤ $4 \leq x_{11} \leq 10$

問 7　同じ植物の 2 つの品種 A と B がある。品種 A は 21 本，品種 B は 16 本についてある部位の長さ（cm）を測定した。下の表は，その平均値と分散である。

	平均値	分散
品種 A	2.78	0.145
品種 B	2.93	0.095

これより，2 つの品種のその部位の長さに差があるかどうかを検定したい。その部位の長さは正規分布にしたがい，分散は等しいと仮定できるものとして t 検定を行う。

〔1〕検定の前に，2 つの標本の分布の様子を箱ひげ図で確認しておく。次の図のうち，品種 A と品種 B はそれぞれどの箱ひげ図にあたるか。下の ① ～ ⑤ のうちから最も適切なものを一つ選べ。 **13**



① 品種 A：a，品種 B：e　② 品種 A：b，品種 B：c　③ 品種 A：a，品種 B：d

④ 品種 A：b，品種 B：e　⑤ 品種 A：c，品種 B：d

〔2〕2 つの標本をプールした分散 s^2 を求めよ。 **記述 5**

〔3〕t 統計量の値を求めよ。 **記述 6**

〔4〕この検定の結果として次のような結論を導いた。空欄（ア）～（ウ）に入る言葉として正しい組合せはどれか。下の ① ～ ⑤ のうちから適切なものを一つ選べ。

14

有意水準 5 %で両側検定を用いることとする。自由度 35 の t 分布の上

側 2.5 %点は 2.030 である。〔3〕で求めた t 統計量の値から,「2 つの品種のある部位の長さに（ア）」という帰無仮説は（イ）。よって「2 つの品種のある部位の長さに（ウ）」と結論する。

① （ア）差はない　（イ）棄却される　（ウ）差があるといえる
② （ア）差はある　（イ）棄却される　（ウ）差があるといえる
③ （ア）差はある　（イ）棄却されない　（ウ）差があるといえる
④ （ア）差はない　（イ）棄却されない　（ウ）差があるとはいえない
⑤ （ア）差はある　（イ）棄却される　（ウ）差があるとはいえない

問 8 R.A.Fisher の 1936 年の論文にある 3 種（setosa，versicolor，virginica）のあやめの「がく片の長さ」のデータを利用して分析した結果を考察する。このデータでは，それぞれ 50 ずつの個体が観測されている。

〔1〕 3 種の「がく片の長さ」の等分散性について有意水準 5 % の F 検定を行った。次の表は，その出力結果の一部である。num は分子，denom は分母の略語である。

出力結果の一部

```
・setosa と versicolor の結果（前者が分子，後者が分母，以下同様）
  F = 0.4663, num df = 49, denom df = 49, p-value = 0.008657
  95 percent confidence interval: 0.2646385 0.8217841
・versicolor と virginica の結果
  F = 0.6589, num df = 49, denom df = 49, p-value = 0.1478
  95 percent confidence interval: 0.3739257 1.1611546
・setosa と virginica の結果
  F = 0.3073, num df = 49, denom df = 49, p-value = 6.366e-05
  95 percent confidence interval: 0.1743776 0.5414962
```

この出力結果に関する説明として，最も適切なものを次の ① ～ ⑤ のうちから一つ選べ。 **15**

① 有意水準 5 % で有意でないのは，versicolor と virginica の分散の差異である。

② どの 95 % 信頼区間も 0 を含んでいないので，分散の差異がみられる種の組合せはない。

③ versicolor と virginica の分散については，95 % 信頼区間が 1 を含んでいるので，有意性を判断できない。

④ F 検定において，信頼区間と p-値に関係はない。

⑤ 一般に，F-値が小さいほうが分散の差異が大きいといえる。

〔2〕 setosa と versicolor の「がく片の長さ」について，平均値の差の検定を行った。次の表は，有意水準 5 % の Welch の t 検定を用いた出力結果の一部である。

出力結果の一部

```
・setosa と versicolor の結果
  t = -10.521, df = 86.538, p-value  < 2.2e-16
  mean of setosa mean of versicolor
  5.006          5.936
```

Welch の t 検定と出力結果に関する説明として，適切でないものを次の ① ～ ⑤ のうちから一つ選べ。 **16**

① このデータで分散の差異がまったくないなら，Welch の t 検定の自由度は 98 である。

② 個体数が同じなら，Welch の t 検定の自由度は分散の差が大きいほど小さくなる。

③ この結果から，Student の t 検定を用いたなら，有意差はないという結果になることがわかる。

④ 個体数が同じなら， Welch の t 検定の t-値と Student の t 検定の t-値は同じ値である。

⑤ 各種類において，個体数が異なる場合でも Welch の t 検定を利用できる。

〔3〕３種のあやめの「がく片の長さ」の平均値の差の検定を行うとき，単純に３回の t 検定を繰り返して判断してはいけない理由として，最も適切なものを次の ① ～ ⑤ のうちから一つ選べ。 **17**

① ３回のいずれか一つが棄却される確率が高くなるから

② ３回の検定を行う順番を変更すると結果が変わるから

③ 互いの分散が違うから

④ 検出力が悪くなるから

⑤ ３つの種類はたまたま選ばれただけで他にもあるから

問 9　確率変数 W は自由度 m のカイ二乗分布に従うとする。m が大きいとき，W 及び $\sqrt{W}$ の分布は正規分布で近似することができる。それぞれの正規分布の適切な組合せを次の ① ～ ⑤ のうちから一つ選べ。 **18**

① $W : N(m, m)$,　$\sqrt{W} : N(\sqrt{m}, \sqrt{m})$

② $W : N(m, m)$,　$\sqrt{W} : N(\sqrt{m}, 1)$

③ $W : N(m, 2m)$,　$\sqrt{W} : N(\sqrt{m}, \sqrt{2m})$

④ $W : N(m, 2m)$,　$\sqrt{W} : N(\sqrt{m}, 1)$

⑤ $W : N(m, 2m)$,　$\sqrt{W} : N(\sqrt{m}, 1/2)$

問 10　$W(t),\ 0 \leq t,$ を標準ブラウン運動とする。$W(1) = 1.0$ が与えられたもとで $W(0.5)$ の条件つき期待値（ア）と条件つき分散（イ）はいくらか。正しい組合せを次の ① ～ ⑤ のうちから一つ選べ。 **19**

① ア: 0.0　イ: 0.5　　② ア: 0.5　イ: 0.5　　③ ア: 1.0　イ: 0.5
④ ア: 0.5　イ: 0.25　　⑤ ア: 0.0　イ: 0.25

問 11　映画館に関する経済産業省「平成 22 年特定サービス産業実態調査（確報）」のデータでは，島根県と徳島県のスクリーン数が秘匿されている。これは，県内の映画館事業所が 2 つしかないための措置である。

徳島県のスクリーン数を回帰式で予測することを考える。被説明変数を都道府県別スクリーン数 y，説明変数を都道府県別映画館従業者数 x として，回帰式 $y = \beta_0 + \beta_1 x$ を想定する。これを最小二乗法で推定し，徳島県の映画館従業者数 48 人を代入してスクリーン数を予測する。

〔1〕下に示す回帰式の推定結果に基づく徳島県のスクリーン数の予測値を，次の ① ～ ⑤ のうちから一つ選べ。 **20**

① 20.5　　② 22.2　　③ 23.6　　④ 24.8　　⑤ 26.1

計算結果

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-54.566  -9.848  -4.421   6.684  82.728

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.465565   3.862580   4.781 2.07e-05 ***
x            0.106381   0.004791  22.204  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.41 on 43 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.9198,    Adjusted R-squared:  0.9179
F-statistic:   493 on 1 and 43 DF,  p-value: < 2.2e-16
```

〔2〕下の散布図（図中の点線は，徳島県の映画館従業者数をあらわす）及び回帰診断図から判断して，上の回帰分析及び徳島県のスクリーン数の予測の問題点を述べよ。 **記述 7**

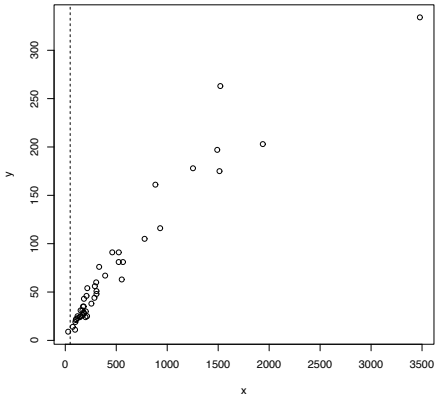


図 1：都道府県別映画館従業者数 x と映画館スクリーン数 y の散布図
資料：経済産業省「平成 22 年特定サービス産業実態調査」（確報）映画館

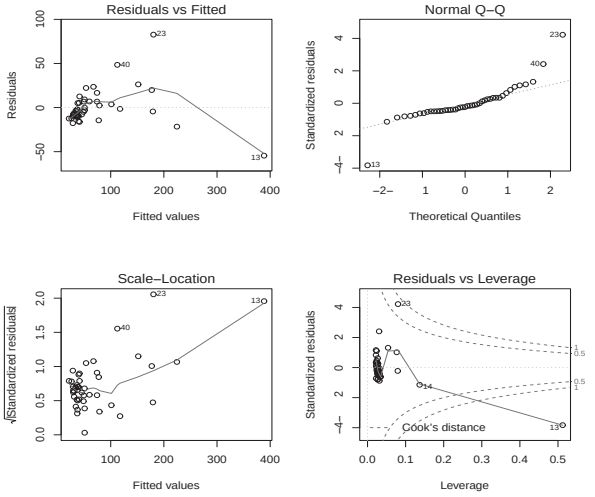
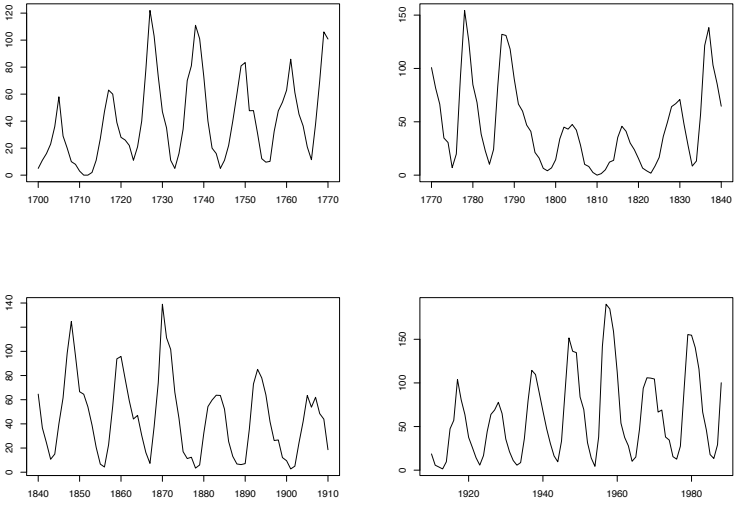


図 2：回帰診断図

問 12　ある 20 人のクラスで，勉強時間 x とテストの成績 y との相関係数を計算したら 0.50 であった。このクラスを無作為に 10 人ずつに分割して，一方で x の平均を，もう一方で y の平均を計算する。（無作為分割を反復した際の）x の平均と y の平均の相関係数はいくらか，次の ① ～ ⑤ のうちから一つ選べ。 **21**

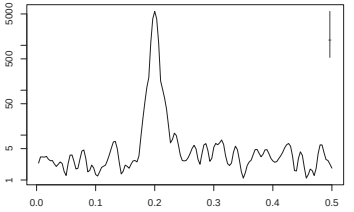
① -0.50　② -0.05　③ 0.0　④ 0.05　⑤ 0.50

問 13　次図の 4 つのグラフは太陽黒点の時系列の時系列プロットである（1700 年から 1988 年までをほぼ 70 年ごとに分割している）。

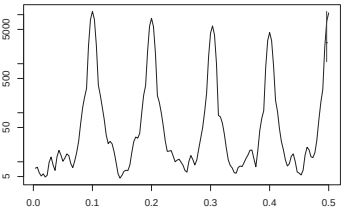


この時系列の平滑化したペリオドグラムとして適切なものを，次の ① ～ ⑤ のうちから一つ選べ。（各図中右上の縦棒は 95 %信頼区間の長さを示す）　**22**

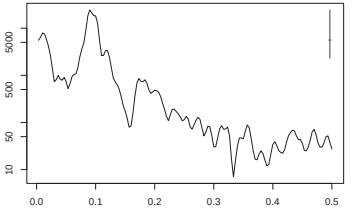
①



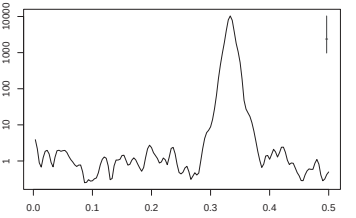
②



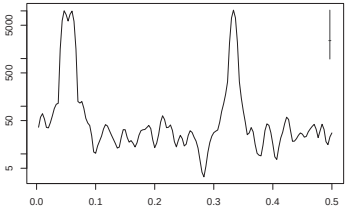
③



④



⑤



問 14　7 人が集まり麻雀（4 人ゲーム）を 7 ゲーム行うことになった。ゲームが公平になるように，どの 2 人の参加者のペアについても，7 ゲーム中で対戦する回数が等しくなるような組合せを考える。これは，処理（参加者）の数が 7，ブロックの大きさが 4，ブロックの数（対戦の数）が 7 であるブロック計画のひとつで，釣合い型不完備ブロック計画とよばれる。

〔1〕このブロック計画において，各参加者の参加するゲーム数（ア），および参加者の各ペアの対戦数（イ）の組合せとして正しいものを次の ① ～ ⑤ から一つ選べ。 **23**

① ア：3　イ：2　　② ア：3　イ：3　　③ ア：4　イ：1
④ ア：4　イ：2　　⑤ ア：4　イ：3

〔2〕この計画のブロックを部分的に以下で与えるとき，空欄のブロックとして適切な組合せを下の ① ～ ⑤ から一つ選べ。 **24**

1 回戦　(1, 4, 6, 7)
2 回戦　(2, 4, 5, 6)
3 回戦　(1, 3, 4, 5)
4 回戦　(1, 2, 5, 7)
5 回戦　(3, 5, 6, 7)
6 回戦　［　　］
7 回戦

① 6 回戦 (1, 2, 3, 4), 7 回戦 (2, 3, 6, 7)
② 6 回戦 (1, 2, 3, 4), 7 回戦 (2, 5, 6, 7)
③ 6 回戦 (2, 3, 4, 6), 7 回戦 (1, 2, 3, 5)
④ 6 回戦 (2, 3, 4, 6), 7 回戦 (1, 2, 3, 7)
⑤ 6 回戦 (2, 3, 4, 7), 7 回戦 (1, 2, 3, 6)

問 15　A 君は，ある路線の徒歩圏内にある一軒家の価格広告を利用しパス解析を試みた。物件の広告には価格（万円），新宿から最寄駅までの時間（分），最寄駅からの徒歩（分），土地面積（m^2），建物面積（m^2），築年数（年），部屋数（個）が示されている。

次の表は，A 君がはじめにこれらの関係を仮定し，パス解析した結果（AMOS の出力）である。ここで，「<---」は A 君が仮定した変数間の因果の関係であって，正しいかどうかはわからない。推定値，標準誤差はそれぞれ，変数間のパス係数に対するもので，検定統計量はパス係数が 0 であるという帰無仮説に対する t 検定の統計量である。確率は p-値を，***は p-値が 0.001 未満であることを示す。

	推定値	標準誤差	検定統計量	確率
建物面積 <--- 部屋数	13.186	2.942	4.483	***
建物面積 <--- 土地面積	0.446	0.069	6.467	***
価格 <--- 築年数	-52.317	13.540	-3.864	***
価格 <--- 新宿から	-44.407	6.763	-6.566	***
価格 <--- 土地面積	10.395	2.769	3.754	***
価格 <--- 部屋数	-214.767	105.267	-2.040	.041
価格 <--- 建物面積	17.763	3.926	4.525	***
価格 <--- 徒歩	-33.918	23.793	-1.426	.154

〔1〕この結果について，A君は考え方の間違いに気づき修正した。その間違いは何であって，どのように修正したか。最も適切なものを次の①～⑤のうちから一つ選べ。 **25**

① 部屋数が建物面積の要因と考えるのはおかしいので，因果の関係を逆にした。

② 部屋数から価格への標準誤差が他との比較で大きすぎるので削除した。

③ p-値から徒歩が価格に関係しないことがわかったので削除した。

④ 部屋数は住居者の好みがあるので削除した。

⑤ 価格への直接のパスが多すぎるのでいくつかを削除した。

〔2〕〔1〕で示した間違いを正し，出力結果を整理した結果，以下のような関係がモデルの適合度の観点から最もよいものとして評価された。この結果をもとにパス図を作成せよ。 **記述 8**

	推定値	標準誤差	検定統計量	確率
建物面積 <--- 土地面積	.612	.076	8.014	***
価格 <--- 築年数	-55.921	14.043	-3.982	***
価格 <--- 新宿から	-43.829	7.014	-6.249	***
価格 <--- 土地面積	8.922	3.164	2.820	.005
価格 <--- 建物面積	15.725	3.676	4.278	***
部屋数 <--- 建物面積	.017	.003	6.560	***

問 16　$X_1,\ldots,X_n$ は互いに独立に正規分布 $N(\mu,\sigma^2)$ に従うとする。(μ,σ^2) の同時事前密度関数を，

$$\pi(\mu,\sigma^2) \propto (\sigma^2)^{-2}\exp\left[-\frac{1}{2\sigma^2}(\mu^2+1)\right]$$

とすると，$x=(x_1,\ldots,x_n)'$ が観測されたとき (μ,σ^2) の同時事後密度関数は，

$$\pi(\mu,\sigma^2|x) \propto (\sigma^2)^{-(n/2+2)}\exp\left[-\frac{1}{2\sigma^2}\left\{\mu^2+1+\sum_{i=1}^{n}(x_i-\mu)^2\right\}\right]$$

となる。ただし $\bar{x}=\sum_{i=1}^{n}\frac{x_i}{n}$ は標本平均を，$\propto$ は両辺が比例的であることを，また $'$ はベクトルの転置を表す。

(μ,σ^2) の事後分布から，以下の初期化および**ステップ 1** ～ **ステップ 3** のようにギブス・サンプリングを用いて確率標本を発生させるとき，空欄 (A) に入る記述として適切なものを下の ① ～ ⑤ のうちから一つ選べ。 **26**

初期化. (μ,σ^2) の初期値を $(\mu^{(0)},\sigma^{2(0)})$ とし，$t=0$ とする。

ステップ 1. $\sigma^{2(t)},x$ を所与として，$\mu^{(t+1)}$ を (A) から発生させる。

ステップ 2. $\mu^{(t+1)},x$ を所与として，$\sigma^{2(t+1)}$ を以下の密度関数を持つ逆ガンマ分布から発生させる。

$$\pi(\sigma^2|\mu^{(t+1)},x) \propto (\sigma^2)^{-(n/2+2)}\exp\left[-\frac{1}{2\sigma^2}\left\{\left(\mu^{(t+1)}\right)^2+1+\sum_{i=1}^{n}\left(x_i-\mu^{(t+1)}\right)^2\right\}\right],$$

ステップ 3. t を $t+1$ として**ステップ 1** に戻る。

① 平均 $\mu^{(t-1)}$, 分散 $\dfrac{\sigma^{2(t)}}{n}$ の正規分布

② 平均 $\bar{x}$, 分散 $\dfrac{\sigma^{2(t)}}{n}$ の正規分布

③ 平均 $\bar{x}$, 分散 $\dfrac{\sigma^{2(t)}}{n+1}$ の正規分布

④ 平均 $\dfrac{n}{n+1}\bar{x}$, 分散 $\dfrac{\sigma^{2(t)}}{n}$ の正規分布

⑤ 平均 $\dfrac{n}{n+1}\bar{x}$, 分散 $\dfrac{\sigma^{2(t)}}{n+1}$ の正規分布

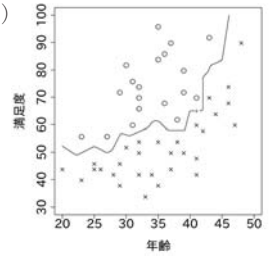
問 17　あるデパートで，商品購入者に対するアンケートを行い，年齢，満足度，今後も当デパートに再度来店したいかどうかについて回答してもらった。そして，年齢と満足度について基準化した変数を用い，今後も当デパートに来たいかどうかについて判別分析を行うこととした。

次の 5 つの図はアンケート結果のプロットと，線形判別，2 次判別，カーネル SVM，最近隣法，決定木のいずれかを用いた判別境界を示している（○印が「再度来店したい」，×印は「もう来店したくない」を表している）。ただし，線形判別と 2 次判別に関しては，事前確率はサンプルの比率を使うこととし，カーネル SVM ではガウシアンカーネルを用いている。

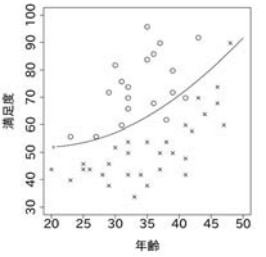
〔1〕以下の（ア）～（オ）の図の中で，2 次判別の判別境界を表しているものはどれか。次の ①～⑤ のうちから適切なものを一つ選べ。 **27**

① （ア）　② （イ）　③ （ウ）　④ （エ）　⑤ （オ）

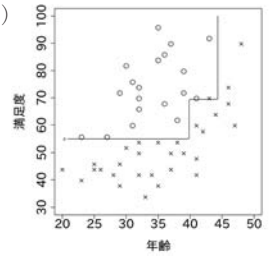
（ア）



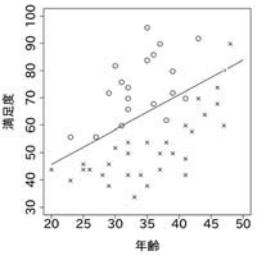
（イ）



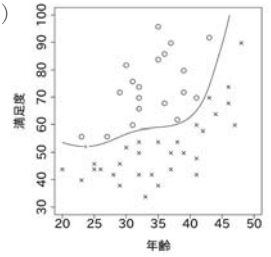
（ウ）



（エ）



（オ）



〔2〕5 種類の判別分析に関して述べた次の ① ～ ⑤ の意見のうちから，最も適切なものを一つ選べ。 **28**

① 線形判別は，分布が正規分布に従っていれば，2 群の分散共分散行列によらずに適用してよい。

② 2 次判別は，2 群の分散共分散行列が異なっていても適用できるが，分布は正規分布に従っていなければならない。

③ カーネル SVM は，指定すべきパラメータがいくつかあるが，分析結果にほとんど影響しないので，適当なパラメータを設定してよい。

④ 最近隣法は，データが特定の分布に従っているときにしか適用できない。

⑤ 決定木は，変数間に相関があるデータの分析に適している。

統計検定準１級　例題　正解一覧

選択問題及び部分記述問題の正解一覧です。次ページ以降に解説を掲載しています。問題の趣旨やその考え方を理解するために活用してください。

論述問題の問題文，解答例は208ページに掲載しています。

問		解答番号	正解
問１	〔1〕	1	③
	〔2〕	2	③
問２	〔1〕	3	⑤
	〔2〕	4	④
	〔3〕	5	⑤
問３	〔1〕	6	③
	〔2〕	7	④
問４	〔1〕	8	②
	〔2〕	9	①
	〔3〕	10	③
	〔4〕	11	④
問５	〔1〕	記述 1	120
		記述 2	80
	〔2〕	記述 3	4.17
		記述 4	3.84
問６		12	③
問７	〔1〕	13	②
	〔2〕	記述 5	0.124
	〔3〕	記述 6	− 1.284
	〔4〕	14	④

問		解答番号	正解
問８	〔1〕	15	①
	〔2〕	16	③
	〔3〕	17	①
問９		18	⑤
問10		19	④
問11	〔1〕	20	③
	〔2〕	記述 7	※
問12		21	①
問13		22	③
問14	〔1〕	23	④
	〔2〕	24	⑤
問15	〔1〕	25	①
	〔2〕	記述 8	※
問16		26	⑤
問17	〔1〕	27	②
	〔2〕	28	②

※は次ページ以降を参照。

選択問題及び部分記述問題　解説

問 1

〔1〕 **1** …… 正解 ③

包除原理により $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ が成り立つので，

$$P(A \cup B) = 0.45 + 0.45 - 0.2 = 0.7$$

となる。

〔2〕 **2** …… 正解 ③

事象が 3 つの場合，包除原理より，

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

が成り立つ（ベン図から容易に確認できる）。この結果を用いると，

$$P(A \cup B \cup C) = 0.45 + 0.45 + 0.4 - 0.2 - 0.15 - 0.1 + 0.05 = 0.9$$

となる。

問 2

〔1〕 **3** …… 正解 ⑤

パラメータ λ のポアソン分布の確率関数は，

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

であり，$k = 3$ を代入するとよい。

〔2〕 **4** …… 正解 ④

死亡者数 X がパラメータ λ のポアソン分布に従うとき，期待値 $E(X) = \lambda$，分散 $V(X) = \lambda$ であり，

$$V(X) = E(X^2) - \{E(X)\}^2$$

であることから，$E(X^2) = \lambda^2 + \lambda$ である。

〔3〕 **5** …… 正解 ⑤

ポアソン分布ではパラメータ λ をデータの平均値より推定する。死亡者数 6 人以上を 6 人とすると，死亡者数の平均値は，

$$0 \times \frac{55}{500} + 1 \times \frac{144}{500} + 2 \times \frac{140}{500} + 3 \times \frac{95}{500} + 4 \times \frac{45}{500} + 5 \times \frac{15}{500} + 6 \times \frac{6}{500} = 2.00$$

となり，表にある期待度数は λ=2.00 として求められている（6 人以上についてどのよう

にするかによって多少値が変わる)。

適合度検定は期待度数に対して，実際の度数のあてはまりのよさを検定するものであり，χ^2 検定量を用いる。表から，死亡者数は 0 人から 6 人以上までの 7 区分であるが，データをもとにパラメータを推定しているため，

検定統計量の自由度は $5(=7-2)$

$$\chi^2 = \sum \frac{(\text{件数} - \text{期待度数})^2}{\text{期待度数}} = 4.498$$

である。

この χ^2 値と，自由度 5 の χ^2 分布の上側 5 %点（11.07）を比較する。$4.498 < 11.07$ なので，平均値 2.0 のポアソン分布に従っていることは棄却できない。

問 3

〔1〕 **6** …… 正解 ③

期待値の線形性から，

$$E[X+Y] = E[X] + E[Y] = 3.0$$
$$E[X-Y] = E[X] - E[Y] = -1.0$$

である。また，分散の定義から，

$$V(X+Y) = V(X) + 2Cov(X,Y) + V(Y) = 11.0$$

であり，同様にして，

$$V(X-Y) = V(X) - 2Cov(X,Y) + V(Y) = 3.0$$
$$Cov(X+Y, X-Y) = V(X) - V(Y) = -1.0$$

である。

【別解】

線形変換 $\begin{pmatrix} X+Y \\ X-Y \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$ に注意すると，ベクトル $\begin{pmatrix} X+Y \\ X-Y \end{pmatrix}$ の平均は，

$$E\left[\begin{pmatrix} X+Y \\ X-Y \end{pmatrix}\right] = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} E\left[\begin{pmatrix} X \\ Y \end{pmatrix}\right] = \begin{pmatrix} 3.0 \\ -1.0 \end{pmatrix}$$

分散共分散行列は，

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3.0 & 2.0 \\ 2.0 & 4.0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 11.0 & -1.0 \\ -1.0 & 3.0 \end{pmatrix}$$

である。

〔2〕 **7** ····· 正解 ④

確率変数 $\begin{pmatrix} X \\ Y \end{pmatrix}$ が 2 変量正規分布

$$N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}\right)$$

に従うとき，X が与えられたときの Y の条件つき分布が，

$$N(\mu_Y + (\sigma_{XY}/\sigma_X^2)(X - \mu_X), \sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2)$$

であることから，

$$N(2.0 + (2.0/3.0)(X - 1.0), 4.0 - 2.0^2/3.0)$$

となり，

$$N(1.33 + 0.67X, 2.67)$$

となる。

条件つき分布 $f_{Y|X}(y|x)$ を求めるには，X と Y の同時分布 $f_{X,Y}(x, y)$ と X の周辺分布 $f_X(x)$ を用いて，$f_{Y|X}(y|x) = f_{X,Y}(x, y)/f_X(x)$ を計算すればよい。なお，X の周辺分布 $f_X(x)$ は平均 μ_X，分散 σ_X^2 の正規分布である。

問４

〔1〕 **8** ····· 正解 ②

二項分布 $X \sim B(n,p)$ の平均および分散は，$\mu = np$, $\sigma^2 = np(1-p)$ である。n が十分大きければ $\dfrac{X}{n}$ の分布は，平均 p，分散 $\dfrac{p(1-p)}{n}$ の正規分布を用いて近似できる。題意より，$\dfrac{X}{n}$ の平均の推定値は 0.62，分散の推定値は $\dfrac{0.62 \times (1-0.62)}{1008} = \dfrac{0.62 \times 0.38}{1008}$ となる。

一般に，比率に関する $(1-\alpha) \times 100$ % 信頼区間については，標準正規分布の近似を利用し，標準正規分布表より上側 $\dfrac{\alpha}{2} \times 100$ % 点 z_0 を求める。ここでは，95 %信頼区間であるため，上側 2.5 %点を求めると $z_0 = 1.96$ を得る。

95 %信頼区間は $\left(0.62 - z_0\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}, 0.62 + z_0\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$ で構成され，p に 0.62 を，n に 1008 を代入して近似的に用いる。

〔2〕 **9** ····· 正解 ①

信頼区間の構成の方法より，信頼区間の幅を狭くするのは，回答者数 n を大きくする（Ⅰは正しい）か，z_0 を小さくするかである。信頼係数 $(1-\alpha)$ を大きくすると z_0 は大きくなり，信頼区間の幅は広くなる（Ⅲは誤り）。回答者が若者であるか否かは関係しない（Ⅱは誤り）。

〔3〕 **10** ····· 正解 ③

真の内閣支持率が 0.6 であるので，X/n の平均は 0.6，分散は $(0.6 \times 0.4)/1008$ となる。また，

$$(0.62 - 0.6)/\sqrt{(0.6 \times 0.4)/1008} = 1.296$$

となることから，標準正規分布表を用いて，

$$P(X/n > 0.62) = P(z_0 > 1.296) \fallingdotseq 0.0968$$

となる。

〔4〕 **11** ····· 正解 ④

回答者数 n に対し，95 %信頼区間の幅は，

$$2 \times 1.96\sqrt{p(1-p)/n}$$

である。この幅が 0.04 以内となるためには，$1.96\sqrt{p(1-p)/n} \leq 0.02$ を満たすよう n を定める。$p(1-p)$ が最大となるのは $p = 0.5$ のときなので，代入して解くと，

$$(1.96/0.02)^2 \times 0.5(1-0.5) = 2401$$

となる。また，真の支持率 $p = 0.6$ を用いて解くと，

$$(1.96/0.02)^2 \times 0.6(1-0.6) = 2305$$

となる。これらのことから 2400 人以上を調査することが好ましい。

問5

〔1〕 記述 **1**（ア） …… 正解 120

記述 **2**（エ） …… 正解 80

クロス表を用いた独立性の検定に関する問題である。

年齢層と商品のタイプ A と B の好みが独立であることを仮定すると，各年代の数は，A と B の計の比，つまり 200 : 200=1 : 1 となる。つまり，（ア）と（イ）は 120，（ウ）と（エ）は 80 である。

〔2〕 記述 **3**（オ） …… 正解 4.17

記述 **4**（カ） …… 正解 3.84

このクロス表を用いた独立性の検定の検定統計量は，

$$\chi^2 = \sum_{\text{全セル}} \frac{\{(\text{セルの度数}) - (\text{セルの期待度数})\}^2}{(\text{セルの期待度数})}$$

で計算され，自由度 1 の χ^2 分布に従う。自由度が 1 であるのは，（ア），（イ），（ウ），（エ）のうち，1 つが決まればすべて決まるためである。

（オ）：χ^2 値は 0.833+0.833+1.25+1.25=4.166 である。

（カ）：自由度 1 の χ^2 分布の上側 5 %点は 3.84 である。

問6

12 …… 正解 ③

まず，x_{11} を含む第 1 列の列和 $x_{11} + x_{21} = 10$ から x_{11} は 10 以下であることがわかる。

さらに，x_{21} は第 2 行の行和より 8 以下であることから，x_{11} は 2 以上であることもいえる。

同様に，x_{11} を含む第 1 行の行和 $x_{11} + x_{12} + x_{13} = 12$ から x_{11} は 12 以下であり，x_{12}，x_{13} は第 2,3 列の列和よりそれぞれ 2 以下，8 以下であることから，ここでも x_{11} は 2 以上であることがいえる。

以上をまとめると，x_{11} の範囲は 2 以上 10 以下となる。

実際に，$(x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23})$ に対して，上限と下限は $(10, 2, 0, 0, 0, 8)$ と $(2, 2, 8, 8, 0, 0)$ で実現する。

問 7

〔1〕 **13** …… 正解 ②

箱ひげ図 a：最小値は約 3.0 であり，データの平均値は品種 A の平均値 2.78 または品種 B の平均値 2.93 になることはない。したがって，どちらにもあてはまらない。

箱ひげ図 b と c：どちらも四分位範囲の中に品種 A と品種 B の平均値があり，b では c に比べて中央値が小さく四分位範囲は大きい。正規分布を仮定しているので，b の方が c より平均値は小さく分散は大きいと判断できる。したがって，**品種 A が b，品種 B が c** であるとしても矛盾がない。

箱ひげ図 d：最小値の外れ値が 1 つあるが 2.0 以上であり，ヒゲの下限が約 3.0 であるので，データの平均値は品種 A の平均値 2.78 または品種 B の平均値 2.93 になることはない。したがって，どちらにもあてはまらない。

箱ひげ図 e：最小値は約 3.0 であり，データの平均値は品種 A の平均値 2.78 または品種 B の平均値 2.93 になることはない。したがって，どちらにもあてはまらない。

〔2〕 **記述 5** …… 正解 0.124

サンプルサイズ 21 の品種 A の分散と，サンプルサイズ 16 の品種 B の分散より，プールした分散 s^2 は，

$$s^2 = \frac{20 \times 0.145 + 15 \times 0.095}{21 + 16 - 2} = 0.1235 \cdots \fallingdotseq 0.124$$

と計算できる。

〔3〕 **記述 6** …… 正解 -1.284

品種 A のサンプルサイズ n_A，品種 B のサンプルサイズ n_B，品種 A の平均値 $\bar{x}_A$，分散 s_A^2，品種 B の平均値 $\bar{x}_B$，分散 s_B^2，プールした分散 s^2 とすると，検定統計量である t 統計量は以下のように求められる。

$$t = \frac{\bar{x}_A - \bar{x}_B}{s\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}} = \frac{2.78 - 2.93}{s\sqrt{\dfrac{1}{21} + \dfrac{1}{16}}}$$

これより，$t \fallingdotseq -1.284$ となる。

〔4〕 **14** …… 正解 ④

部位の長さに差があるかについて検定したいので，帰無仮説は「2 つの品種のある部位の長さに**差はない**（ア）」，対立仮説は「2 つの品種のある部位の長さに差がある」という両側検定を行う。棄却限界は，-2.030 および 2.030 である。t 統計量の値が -1.284 であり，この値が棄却限界の -2.030 より小さくないので，帰無仮説は**棄却されない**（イ）。したがって，「2 つの品種のある部位の長さに**差があるとはいえない**（ウ）」と結論する。

問8

〔1〕 **15** ····· 正解 ①

2 群の等分散性についての F 検定の検定統計量は s_1^2/s_2^2 で定義される。ここで，s_i^2 は第 i 群の標本分散である。1 群のサンプルサイズが n，2 群のサンプルサイズが m とすると，この検定統計量は 2 群の等分散性の仮定の下，自由度 $(n-1, m-1)$ の F 分布に従う。自由度 $(n-1, m-1)$ の F 分布の上側 $\alpha \times 100$ %点を $F_\alpha(n-1, m-1)$ と表すと，2 群の分散比 σ_1^2/σ_2^2 の 95 %信頼区間は，

$$\frac{1}{F_{0.025}(n-1, m-1)}\frac{s_1^2}{s_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1}{F_{0.975}(n-1, m-1)}\frac{s_1^2}{s_2^2}$$

となる。F 検定において有意差があるか否かは，p-値で判断するか，信頼区間が 1 を含むか否かで判断できる。

①：正しい。分散比の信頼区間に 1 が含まれ，p-値 > 0.05 であるのは，versicolor と virginica についてだけである。このことから，versicolor と virginica の分散の差異のみが有意水準 5 %で有意でないため，正しい。

②：誤り。分散比の信頼区間は標本分散が 0 とならない限り 0 を含むことはないため，誤り。

③：誤り。F 検定において仮説が棄却されないことと，信頼区間が 1 を含むことは同値であるため，誤り。

④：誤り。上の説明にあるように信頼区間と p-値には関係があるため，誤り。

⑤：誤り。F-値は 1 から離れるほど分散の差異が大きいので，F-値が小さい方が分散の差異が大きいとはいえないため，誤り。

〔2〕 **16** ····· 正解 ③

$\bar{y}_1$ を第 1 群の標本平均，$\bar{y}_2$ を第 2 群の標本平均とすると，各群の標本数がともに n の場合の Welch の t 検定は，

$$(\bar{y}_1 - \bar{y}_2)/\sqrt{(s_1^2 + s_2^2)/n}$$

と定義される（これは，各群の分散が等しいとした場合の Student の t 検定と同じ形である）。この自由度は，

$$\frac{(n-1)(s_1^2 + s_2^2)^2}{s_1^4 + s_2^4}$$

であり，これは Student の t 検定の自由度 $2n-2$ 以下となる（2 群の標本分散が一致する場合に限り，これらの自由度は一致する）。

①：正しい。2 群の標本分散が一致する場合，Welch の t 検定の自由度は Student の t 検定の自由度 $2n-2=98$ と同じになるため，正しい。

②：正しい。各群の分散比を r とすると，Welch の t 検定の自由度は，

$$\frac{(n-1)(1+r^2)^2}{1+r^4} = (n-1)\left\{1+\frac{2r^2}{1+r^4}\right\} = (n-1)\left\{1+\frac{2}{r^2+1/r^2}\right\}$$

と表され，$r=1$ のときに最大となり，r が 1 から離れるほど小さくなっていくため，正しい。

③：誤り。t 分布は自由度が小さいほど上側 $\alpha \times 100$ %点が大きくなる。各群の分散が異なる場合，（Welch の検定と Student の t 検定の検定統計量は等しいが，）Welch の t 検定の自由度は Student の t 検定の自由度より小さいので，Welch の検定で有意であれば，Student の t 検定でも有意となるため，誤り。

④：正しい。上の説明にあるように，個体数が同じなら Welch の検定の t-値は Student の検定の t-値と同じになるため，正しい。

⑤：正しい。個体数が異なる場合でも Welch の検定は（自由度は少し複雑となるが）定義できるため，正しい。

よって，③が正解である。

〔3〕 **17** …… 正解 ①

すべての平均に差がないという状況の下，平均の差の検定を有意水準 5 %で独立に 3 回行うと，少なくとも 1 つが棄却される確率は $1-(0.95)^3 = 0.143$ となり，有意となる確率が高くなる。

①：正しい。上の説明にあるように，少なくとも 1 つが棄却される確率は高くなるため，正しい。

②：t 検定を行う順序は結果に影響を与えないため，誤り。

③：互いの分散が同じであっても，違っていても繰り返してはいけないため，誤り。

④：帰無仮説が棄却される確率が高くなることは，検出力もよくなるため，誤り。

⑤：複数回検定を行うことと関係のないことであるため，誤り。

問 9

18 …… 正解 ⑤

自由度 m のカイ二乗分布の平均は m，分散は $2m$ であり，m が大きいとき，W の分布は中心極限定理により正規分布で近似できる。

また，$\sqrt{W}$ の分布を求めるためにはデルタ法を利用すればよい。

デルタ法は，確率変数 X が分散 σ^2 の小さな正規分布 $N(\mu,\sigma^2)$ に近似的に従うとき，関数 f により変換された確率変数 $f(X)$ の分布を $N(f(\mu), f'(\mu)^2\sigma^2)$ で近似するものである。

$\sqrt{W} = \sqrt{m}\sqrt{\dfrac{W}{m}}$ に注意して，$X = \dfrac{W}{m}$, $f(x) = \sqrt{x}$, $\mu = 1$, $\sigma^2 = \dfrac{2}{m}$ とおいてデルタ法を適用する。中心極限定理より $\dfrac{W}{m}$ は正規分布 $N(1, \dfrac{2}{m})$ に近似的に従うので，$f(1) = 1$ と $f'(1) = \dfrac{1}{2}$ に注意すると，$f(\dfrac{W}{m}) = \sqrt{\dfrac{W}{m}}$ の分布は $N(1, \dfrac{1}{2m})$ で近似される。したがって，$\sqrt{W}$ の分布は $N(\sqrt{m}, \dfrac{1}{2})$ で近似される。

問 10

19 …… 正解 ④

標準ブラウン運動 $W(1)$ の従う分布は平均 0，分散 1 の正規分布である。また，標準ブラウン運動について，$W(0.5)$ と $W(1) - W(0.5)$ は独立であり，それぞれ平均 0，分散 0.5 の正規分布に従う。$W(0.5)$ と $W(1) = W(0.5) + \{W(1) - W(0.5)\}$ の共分散は 0.5 となる。以上から，$(W(0.5), W(1))$ の従う同時分布は平均 $(0, 0)$，分散共分散行列 $\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ の 2 変量正規分布となり，したがって，$W(1)$ が与えられたもとでの $W(0.5)$ が従う条件つき分布は，平均 $W(1)/2$，分散 0.25 の正規分布である（問 3〔2〕を参照）。これより，$W(1) = 1.0$ が与えられると平均は 0.5 となる。

したがって，求める条件つき期待値は 0.5，条件つき分散は 0.25 である。

問 11

〔1〕 **20** …… 正解 ③

回帰分析の結果より，(スクリーン数) $= 18.465565 + 0.106381 \times$ (映画館従業員数) なので，映画従業員数 48 人の徳島県のスクリーン数は $23.57185 \fallingdotseq 23.6$ と予測できる。

〔2〕 記述 7 ····· 正解 下記参照

解答例 自由度調整済み決定係数や，従業者数に対する回帰係数が有意であることから，一見，導出された回帰式が有効と判断されるが，これは外れ値（都道府県番号 13）の影響が大きいためである。右下の診断図（てこ比と標準化した残差の図）から Cook 距離が 1 を超えていることがわかり，この外れ値の影響を考慮する要請がある。左上と左下の診断図（それぞれ，予測値に対する残差の図と標準化した残差の平方根の図）から従業員数が多いほど誤差が大きくなる傾向があることがわかる。また，右上の診断図（正規 Q-Q プロット）から裾が広いことがわかり正規性も疑問である。そのため，誤差項の不均一分散および非正規性に対する処理が必要である。また，緩やかな曲線に沿っていることが見られるため，非線形の回帰式も考慮しなければならない。

問 12

21 ····· 正解 ①

無作為に分割された 10 人ずつのグループを，それぞれ A, B とする。また，A グループの x の平均を $\bar{X}_A$ と表し，他の記法も同様とする。20 人全体の平均を $\bar{x}, \bar{y}$ と表す。このとき，$\bar{Y}_B = 2\bar{y} - \bar{Y}_A$ となることに注意すれば，

（無作為分割の反復に関して）

$V(\bar{Y}_B) = V(2\bar{y} - \bar{Y}_A) = V(-\bar{Y}_A) = V(\bar{Y}_A)$

$Cov(\bar{X}_A, \bar{Y}_B) = Cov(\bar{X}_A, 2\bar{y} - \bar{Y}_A) = Cov(\bar{X}_A, -\bar{Y}_A) = -Cov(\bar{X}_A, \bar{Y}_A)$

が成り立つ。ここで，定数 $2\bar{y}$ を加えても分散や共分散が変化しないことを利用した。

一方，A グループにおける $\bar{X}_A$ と $\bar{Y}_A$ との相関係数は，20 人全体で計算した x と y との相関係数 $\rho_{x,y}$ と等しい。したがって，$\bar{X}_A$ と $\bar{Y}_B$ との相関係数 $\rho_{\bar{X}_A, \bar{Y}_B}$ は，

$$\begin{aligned}\rho_{\bar{X}_A, \bar{Y}_B} &= Cov(\bar{X}_A, \bar{Y}_B)/\sqrt{V(\bar{X}_A)\,V(\bar{Y}_B)} \\ &= -Cov(\bar{X}_A, \bar{Y}_A)/\sqrt{V(\bar{X}_A)\,V(\bar{Y}_A)} \\ &= -\rho_{x,y}\end{aligned}$$

となる。したがって，$\rho_{\bar{X}_A, \bar{Y}_B} = -\rho_{x,y} = -0.50$ となる。

なお，A グループにおける $\bar{X}_A$ と $\bar{Y}_A$ の分散や共分散は，それぞれ，20 人全体で計算した x と y の分散や共分散の $1/19$ 倍である。相関係数の場合とは異なり，全体での値と A グループでの値とが一致しない。

問 13

22 …… 正解 ③

ペリオドグラムとは，時系列の標本自己共分散関数の離散フーリエ変換であり，時系列の各周波数（1/周期）成分の大きさ（パワースペクトル密度）を表している。問題の太陽黒点の時系列では，10 年前後の明らかな周期が見られるため，ペリオドグラムは周波数 $1/10 = 0.1$ 付近でピークを取る。

①：誤り。ピークの周波数が 0.2 付近にあるため，誤り。

②：誤り。0.1 の倍数の周波数で複数のピークが見られるが，元の時系列では単一の周期しか見られないため，誤り。

③：正しい。周波数 0.1 のみにピークが見られるため，正しい。

④：誤り。ピークの周波数が 0.3 から 0.4 の間にあるため，誤り。

⑤：誤り。ピークの周波数が 0.1 から離れたところに 2 つあるため，誤り。

問 14

〔1〕 **23** …… 正解 ④

このブロック計画では，

［ブロックの数 7］×［ブロックのサイズ 4］÷［参加者の数 7］= 4

なので，どの参加者も 4 回対戦を行う。したがって，**（ア）は4となる**。

会合数（任意の 2 人が対戦する回数）を λ とおく。ある参加者，たとえば参加者 1 に注目すると，参加者 1 が含まれる 4 つの対戦（ブロック）では，参加者 1 以外の 3 人の対戦者が全部で $4 \times 3 = 12$〔人〕必要である。この 12 人は参加者 2 から 7 の 6 人から選ばれるが，会合数の定義より参加者 2 から 7 のどの参加者も，参加者 1 との対戦数が等しいことから $12 = 6 \times \lambda$ となる。したがって，会合数は $\lambda = 2$ と定まる。したがって，**（イ）は2となる**。

〔2〕 **24** …… 正解 ⑤

すでに定まっている 1 回戦 ～5 回戦を見ると，参加者 1 とまだ 2 回対戦していないのは，参加者 2，3，6 の 3 名であり，参加者 4，5，7 の 3 名はすでに 2 回対戦している。したがって，6 回戦と 7 回戦のいずれかは，$(1, 2, 3, 6)$ でなければならない。

残りの 1 回戦に関しては，参加者 2 がまだ 2 回対戦していないのは，参加者 3，4，7 の 3 名なので，$(2, 3, 4, 7)$ である。

これにより，任意のペアが対戦する回数が 2 回となる。

問 15

〔1〕 **25** …… 正解 ①

①：正しい。パス図を描くときは因果関係を正しく表現する必要があるため，正しい。

②：誤り。標準誤差は推定値に比べると大きくないため削除要因とはならないため，誤り。

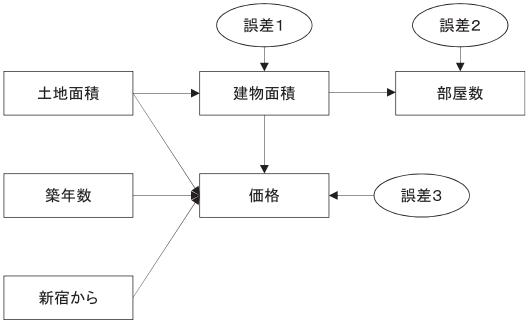
③：誤り。徒歩から価格へ第 3 の変数を介して関係していることがあるため，誤り。

④：誤り。部屋数が居住者の好みによることはあっても，それにより部屋数を削除する理由にはならないため，誤り。

⑤：誤り。価格への直接のパスが多いことが必ずしも間違いにはならないため，誤り（直接のパスの推定値に多重共線性が見られる場合などは削除を検討することも考えられる）。

〔2〕 **記述 8** …… 正解 下記参照

解答例 以下の図のように，出力結果に従って変数間を矢印で結び，矢印が向けられた各変数に誤差の矢印を加えたものがパス図となる。ここでは，観測変数を四角で囲み，誤差を楕円で囲むこととした。誤差を囲まずに表記する書き方もある。また，パス係数の値がわかる場合には，矢印にパス係数を書くことがある。



本問では，適合度によるモデルの評価がすでになされている。共分散構造分析（構造方程式モデリング）では，適用するモデル（仮説）を利用者自身が決定し，検証・分析できることがひとつの利点である。適合度によるモデル評価はその一環であり，実際には利用者自らが必要に応じてモデルの妥当性を検証することになる。また，パス解析においては，潜在変数を考慮した分析を行うことも多い。

問 16

26 ····· 正解 ⑤

ギブスサンプリングのステップ 1. において $\mu^{(t+1)}$ を発生させる分布は，$\sigma^{2(t)}, x$ を所与とする μ の周辺事後分布である。この周辺事後密度関数は，

$$\begin{aligned}\pi(\mu|\sigma^{2(t)}, x) &= \frac{\pi(\mu,\sigma^{2(t)}|x)}{\int_{-\infty}^{\infty}\pi(\mu,\sigma^{2(t)}|x)d\mu} \\ &\propto \pi(\mu,\sigma^{2(t)}|x) \\ &\propto (\sigma^{2(t)})^{-(n/2+2)}\exp\left[-\frac{1}{2\sigma^{2(t)}}\left\{\mu^2+1+\sum_{i=1}^{n}(x_i-\mu)^2\right\}\right] \\ &\propto \exp\left[-\frac{n+1}{2\sigma^{2(t)}}\left(\mu-\frac{1}{n+1}\sum_{i=1}^{n}x_i\right)^2\right]\end{aligned}$$

となり，これは規格化定数を踏まえれば平均 $\dfrac{n}{n+1}\bar{x}$，分散 $\dfrac{\sigma^{2(t)}}{n+1}$ の正規分布の密度関数となる。

問 17

〔1〕 **27** …… 正解 ②

2 次判別の判別境界は一般に 2 次曲線となるため，境界線が明らかに 2 次曲線でない（ア），（ウ），（オ）は該当しない。また，線形判別は明らかに（エ）であるため，2 次判別は（イ）である。

ちなみに，（ア）は最近隣法，（ウ）は決定木，（オ）は SVM である。

〔2〕 **28** …… 正解 ②

①：誤り。線形判別は，各群のデータが正規分布に従い，分散共分散行列が等しいことを仮定しているため，誤り。

②：正しい。2 次判別は，各群のデータは正規分布に従うことを仮定しているが，分散共分散行列が異なっていてもかまわないため，正しい。

③：誤り。カーネル SVM は，分布に関する仮定は必要ないが，パラメータによって分析結果が異なるため，適切なパラメータを設定する必要があるため，誤り。

④：誤り。最近隣法は，分布を問わず使える手法であるため，誤り。
なお，両群のデータが混在している場合には，決定境界が複雑になる（そのような場合には，k 最近隣法を使う方が好ましい）。

⑤：誤り。決定木は，各変数に対し場合分けを行い，その組合せで判別を行う手法であることから，変数間の相関が強いデータの分析には適さないため，誤り。

論述問題　例題と解答例

問 1

あるチョコレート菓子の箱には，おまけとしてシールが 1 枚ずつ入っている。シールは全部で K 種類あり，菓子の箱に各シールが入っている確率は等しく $\frac{1}{K}$ ずつである。以下の各問に答えよ。

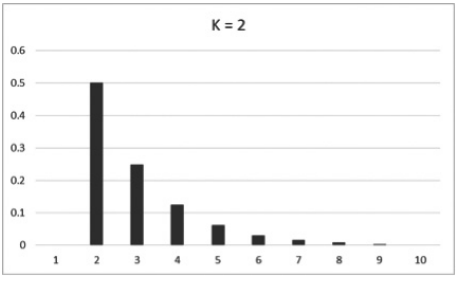
〔1〕$K = 2$ とする。2 種類のシールの両方を手に入れるまでに購入する菓子の個数を X としたとき，確率 $P(X = x)$ を x の式として求めよ。また，X の期待値 $E[X]$ はいくらか。

〔2〕$K = 3$ とする。3 種類のシールすべてを手に入れるまでに購入する菓子の個数を X としたとき，確率 $P(X = x)$ を x の式として求めよ。また，X の期待値 $E[X]$ はいくらか。

〔3〕$K = 20$ のとき，すべての種類のシールを集めるためには，平均何個の菓子を買えばよいか。

解答例

〔1〕$P(X = 1) = 0$ であり，$x \geq 2$ に対しては $P(X = x) = (0.5)^{x-1}$ である（幾何分布）。確率のグラフは次図のようである。



$K = 2$ の場合の確率

期待値は，

$$
\begin{aligned}
E[X] &= \sum_{x=2}^{\infty} x(0.5)^{x-1} = 2 \times 0.5 + 3 \times (0.5)^2 + 4 \times (0.5)^3 + \cdots \\
&= 2 \times \{0.5 + (0.5)^2 + (0.5)^3 + \cdots\} + (0.5)^2 + 2 \times (0.5)^3 + 3 \times (0.5)^4 + \cdots \\
&= 2 \times \frac{0.5}{1 - 0.5} + \frac{(0.5)^2}{1 - 0.5} + \frac{(0.5)^3}{1 - 0.5} + \cdots \\
&= 2 + 2\{(0.5)^2 + (0.5)^3 + \cdots\} \\
&= 2 + 2 \times \frac{(0.5)^2}{1 - 0.5} = 2 + 2 \times 0.5 = 3
\end{aligned}
$$

となる。

〔2〕$P(X = x) = 0$ $(x = 1, 2)$ である。$x \geq 3$ に関しては，$X = x$ となるためには，$(x-1)$ 回目までに 2 種類のみのシールが選ばれていて，x 回目に 3 種類目のシールが選ばれればよい。1 回目と x 回目を異なるシール（たとえば A と C）で固定すると，それらの間の $(x-2)$ 回では，B が 1 回以上で残りは A となればよい。すなわち，すべてが A の場合を排除するので，全部で $(2^{x-2} - 1)$ 通りとなる。1 回目のシールの選び方が 3 通り，x 回目のシールの選び方が 2 通りであるので，求める確率は，

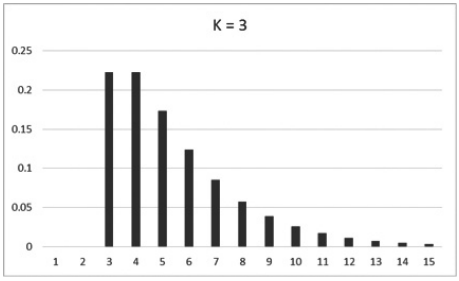
$$
P(X = x) = \frac{6(2^{x-2} - 1)}{3^x} = \frac{2(2^{x-2} - 1)}{3^{x-1}} = \left(\frac{2}{3}\right)^{x-1} - 2\left(\frac{1}{3}\right)^{x-1}
$$

となる。

あるいは，$x - 1$ 回目まで 2 種類のシールが出続け（1 種類のみが出続ける確率を引く），x 回目に 3 種類目のシールが出る確率であるので，

$$
P(X = x) = 3\left\{\left(\frac{2}{3}\right)^{x-1} - 2\left(\frac{1}{3}\right)^{x-1}\right\} \times \frac{1}{3} = \left(\frac{2}{3}\right)^{x-1} - 2\left(\frac{1}{3}\right)^{x-1}
$$

となる，という考察によっても答えが導かれる。確率のグラフは次図のようである。



$K = 3$ の場合の確率

期待値は，

$$E[X]=\sum_{x=3}^{\infty}x\left\{\left(\frac{2}{3}\right)^{x-1}-2\left(\frac{1}{3}\right)^{x-1}\right\}=\sum_{x=3}^{\infty}x\left(\frac{2}{3}\right)^{x-1}-2\sum_{x=3}^{\infty}x\left(\frac{1}{3}\right)^{x-1}$$

であるが，一般に，確率 p に対して，

$$\begin{aligned}\sum_{x=3}^{\infty}xp^{x-1}&=3p^2+4p^3+5p^4+\cdots\\&=3(p^2+p^3+p^4+\cdots)+p^3(1+2p+3p^2+\cdots)\\&=3\times\frac{p^2}{1-p}+p^3\times\frac{1}{1-p}(1+p+p^2+\cdots)\\&=\frac{3p^2}{1-p}+\frac{p^3}{(1-p)^2}=\frac{p^2}{1-p}\left(3+\frac{p}{1-p}\right)\end{aligned}$$

であるので，$p=2/3$ および $p=1/3$ として，

$$\begin{aligned}E[X]&=\sum_{x=3}^{\infty}x\left(\frac{2}{3}\right)^{x-1}-2\sum_{x=3}^{\infty}x\left(\frac{1}{3}\right)^{x-1}\\&=\frac{(2/3)^2}{1/3}\left(3+\frac{2/3}{1/3}\right)-2\frac{(1/3)^2}{2/3}\left(3+\frac{1/3}{2/3}\right)\\&=\frac{4}{3}\times5-2\times\frac{1}{6}\times3.5=\frac{20}{3}-\frac{7}{6}=\frac{33}{6}=\frac{11}{2}=5.5\end{aligned}$$

と求められる。

〔3〕 $K=2$ では $E[X]=1+\dfrac{1}{1/2}=3$，$K=3$ では $E[X]=1+\dfrac{1}{2/3}+\dfrac{1}{1/3}=5.5$ の計算により求められる。この類推で $K=20$ では，

$$\begin{aligned}E[X]&=1+\frac{1}{19/20}+\frac{1}{18/20}+\cdots+\frac{1}{1/20}=1+20\left(\frac{1}{19}+\frac{1}{18}+\cdots+1\right)\\&=1+20\times3.54774\fallingdotseq71.955\quad\cdots\cdots①\end{aligned}$$

であるので，おおよそ 72 個となる。このことは，次のように証明できる。

一般の K で証明する。まず，確率 p の幾何分布（確率関数が $p(x)=p(1-p)^{x-1}$，$x=1,2,\ldots$ で与えられる分布）の期待値は $1/p$ となることに注意する。1 個目の菓子でのシールの種類を A_1 とする。A_1 でないシール A_2 を得るまでの菓子の個数を X_2 とすると，X_2 は確率 $(K-1)/K$ の幾何分布に従う。よって，

$$E[X_2]=1/\{(K-1)/K\}=K/(K-1)$$

となる。A_1 および A_2 でないシール A_3 を得るまでの菓子の個数を X_3 とすると，X_3 は確率 $(K-2)/K$ の幾何分布に従い，その期待値は，

$$E[X_3] = 1/\{(K-2)/K\} = K/(K-2)$$

である。以下これを繰り返すと，すべてのシールを得るまでの菓子の個数 X の期待値は，

$$\begin{aligned} E[X] &= 1 + \frac{1}{(K-1)/K} + \frac{1}{(K-2)/K} + \cdots + \frac{1}{1/K} \\ &= 1 + K\left(\frac{1}{K-1} + \frac{1}{K-2} + \cdots + 1\right) \end{aligned}$$

と求められる。$K = 20$ とすると①が得られる。一般に，

$$\sum_{r=1}^{k} \frac{1}{r} \fallingdotseq \log k \quad \cdots\cdots ②$$

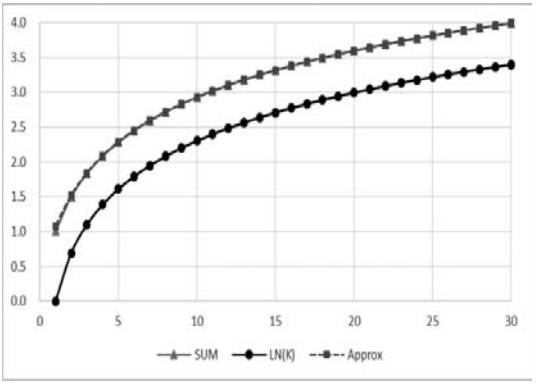
なる近似が成り立つが，これは，

$$\sum_{r=1}^{k} \frac{1}{r} \fallingdotseq \log k + \gamma + \frac{1}{2k} + O\left(\frac{1}{k^2}\right) \quad \cdots\cdots ③$$

とすると（ここで $\gamma \fallingdotseq 0.57721\cdots$ はオイラーの定数），より精確な近似となることが知られている。③の近似では，

$$E[X] = 1 + 20(\log 19 + \gamma + 1/38) = 1 + 20 \times 3.54796 \fallingdotseq 71.959$$

と①とほぼ同じ結果を得る。なお，②および ③の近似は次図のようである。



期待値の近似

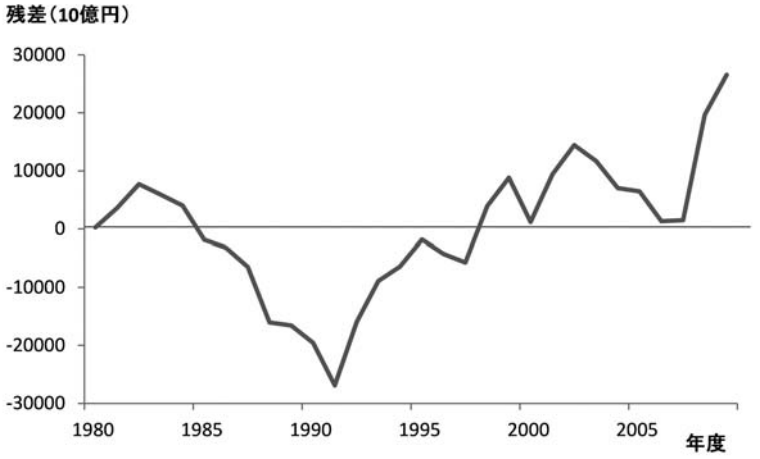
問2

1980 年度から 2009 年度までの実質民間最終消費支出を Y_t，実質国民可処分所得を X_t（いずれも単位 10 億円。$t = 1, 2, \ldots, 30$）として，被説明変数 Y を説明変数 X で単回帰したところ，次の推定結果が得られた。

$$\hat{Y}_t = \underset{(-3.66)}{-53800} + \underset{(19.82)}{0.8308}\ X_t,\quad R^2 = 0.9335, \bar{R}^2 = 0.9311, s = 11896, AIC = 565.0$$

ただし，( ) 内は t 値，R^2 は決定係数，$\bar{R}^2$ は自由度修正済決定係数，s は誤差項の標準偏差の推定値，AIC は赤池の情報量規準である。この結果について，以下の〔1〕～〔3〕の問に答えよ。

〔1〕上の推定結果に基づいて各年度の残差 $e_t = Y_t - \hat{Y}_t$ を計算したところ，次の図のようになった。



この残差をもとに算出した 1 次の自己相関係数 $\hat{\rho}$（e_t と e_{t-1} の相関係数）とダービンワトソン統計量 $DW = \sum_{t=2}^{30}(e_t - e_{t-1})^2 \Big/ \sum_{t=1}^{30} e_t^2$ の値は $\hat{\rho} = 0.85$，$DW = 0.29$ であった。回帰モデルの通常の仮定に照らして，このような残差の検討の意味について述べよ。

〔2〕〔1〕の単回帰に人口 N（単位千人）を説明変数に加えて重回帰分析を行ったところ，次の推定結果が得られた。

$$\hat{Y}_t = \underset{(-12.55)}{-1119473} + \underset{(3.63)}{0.2013}\, X_t + \underset{(11.93)}{10.4302}\, N_t,$$

$$R^2 = 0.9894,\ \bar{R}^2 = 0.9886,\ s = 4838,\ AIC = 511.1$$

この推定結果について，それぞれの回帰係数の解釈について述べよ。また消費の所得弾力性を平均値で評価する方法について述べよ。

〔3〕実質可処分所得のみを説明変数とする単回帰分析と，これに人口を加えた重回帰分析のどちらのモデルがよりすぐれているかについて，回帰係数の検定およびモデル選択の観点から説明せよ。

解答例

〔1〕回帰モデル $Y_t = \alpha + \beta X_t + \epsilon_t$ において，「誤差項は自己相関していない」という仮定が，標準的な仮定の 1 つとして設定される。実証的には，自己相関のうち 1 次の自己相関，すなわち ϵ_t と ϵ_{t-1} の相関が問題になることが多い。ϵ_t と ϵ_{t-1} の相関係数 ρ は，残差 e_t と e_{t-1} の 1 次の自己相関係数 $\hat{\rho}$ によって推定され，$\hat{\rho}$ の絶対値が 1 に近いほど，自己相関なしの仮定（$\rho = 0$）は成立しないことになる。

そして，$\rho = 0$，すなわち自己相関なしの仮定を検定する代表的な方法が，ダービンワトソン統計量 に基づく検定である。DW の定義から，

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} = \frac{\sum_{t=2}^{T} e_t^2}{\sum_{t=1}^{T} e_t^2} + \frac{\sum_{t=2}^{T} e_{t-1}^2}{\sum_{t=1}^{T} e_t^2} - 2\frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2} \fallingdotseq 2(1-\hat{\rho})$$

となり，$-1 \leq \hat{\rho} \leq 1$ より，DW は $0 \leq DW \leq 4$ の範囲をとる。よって，誤差項に正の自己相関があれば DW は 0 に近い値を，負の自己相関があれば DW は 4 に近い値をとり，DW が 2 に近い値をとれば自己相関がないことになる（DW による実際の検定は，説明変数の個数とデータの個数によって決まる DW 統計量の下限分布と上限分布のパーセント点によって判定される）。

ここで，$\hat{\rho} = 0.85$, $DW = 0.29$ であり，DW はかなり 0 に近く，誤差項の自己相関なしの仮説 $\rho = 0$ は棄却される。また，残差のグラフをみると，残差が負の値をとる期間が連続したり（1986–1998 年度），正の値をとる期間が連続したり（1999 年度以降），1 期前の残差と当期の残差に相関がある場合の典型的なグラフとなっている。以上により，誤差項には正の自己相関があると判断できる。

誤差項に自己相関がある場合は，係数の推定量は最良線形不偏推定量とはならず，係数の

推定量の分散を過小に推定するなどの問題が生じるため，自己相関を除去するためにコクラン・オーカット法など適切な対応が必要となる。

〔2〕実質国民可処分所得の係数は 0.2013 であり，これは人口を一定とした場合，可処分所得が 10 億円増加すると，実質民間最終消費支出が約 2 億円増加することを意味する。この値は限界消費性向で，経済理論的には 0 から 1 の間をとり，推定値 0.2013 はこの条件を満たしている。しかしながら，所得が 1 万円増えると消費が約 2000 円しか増えないというのは，やや小さい値であると考えられる。特に，単回帰では係数が 0.8308 であったが，重回帰では 0.6 以上減少している。これは実質国民可処分所得と人口の間に強い正の相関が存在するためである。

一方，人口の係数は 10.4302 であり，これは実質国民可処分所得を一定とすると，人口が 1000 人増加すると，消費が 104.302 億円増加することを表す。理論的には人口が増加すれば消費も増加するので，人口の係数がプラスであることは，理論と整合的である。

消費の所得弾力性 η は，所得が 1 %増加すると消費が何%増加するかを表し，

$$\eta = \frac{dY/Y}{dX/X} = \beta \div \frac{Y}{X}$$

と定義できる（β は所得の係数）。したがって，この弾力性を平均値で評価するには，$\beta \div \dfrac{\bar{Y}}{\bar{X}}$ とすればよい。$\dfrac{\bar{Y}}{\bar{X}}$ は，実質民間最終消費支出の平均値を実質国民可処分所得の平均値で除した値であるから，これは平均値で評価した可処分所得に占める消費の割合，すなわち平均消費性向である。したがって，β の推定値である 0.2013 を，実質民間最終消費支出と実質国民可処分所得のそれぞれの平均値から算出した平均消費性向で除すれば，平均値で評価した消費の所得弾力性 η を算出できる。

〔3〕単回帰とそれに人口 N を加えた重回帰の決定係数 R^2 を比較すると，0.9335 から 0.9894 に増加しているが，説明変数を追加すると R^2 の値は必ず大きくなる（少なくとも減少しない）ので，R^2 で両者のあてはまりの程度を比較することはできない（説明変数の追加による誤差項の標準偏差 s の減少についても同様）。

この問題を修正したのが自由度修正済決定係数 $\bar{R}^2$ で，$\bar{R}^2$ は 0.9311 から 0.9886 に上昇しており，これは人口 N を加えたことにより説明力が上昇したことを意味している。また，AIC をみると 565.0 から 511.1 に低下している。AIC はモデル選択の基準（説明変数の個数と誤差項の標準偏差を総合的に評価する指標）で，AIC が小さい方が望ましいモデルであると判断できる。よって，$\bar{R}^2$ と AIC からみると，人口 N を加えた重回帰の方が望ましい。

さらに，追加した人口 N の係数の有意性を検定すると，追加した N の係数の t 値が 11.9279 と十分大きいので，N の係数が 0 であるという仮説は棄却される（データの個数が 30 なので，正規分布による検定を利用しても問題ないため，5 %点は 1.645 である）。また，実質国民可処分所得 X の係数の t 値も 1.645 を上回っており，単回帰の場合と同様に有意である。

以上から，人口 N を加えた重回帰の方がすぐれているといえる（ただし，〔2〕でも示した

ように，実質国民可処分所得の係数が単回帰と重回帰で大きく異なっており，しかも重回帰での係数の大きさはかなり小さい。これは説明変数間の強い相関，すなわち多重共線性によるものであり，重回帰分析の結果の解釈は十分な注意が必要である）。

問3

ある薬剤の有害事象の発生率を薬剤の投与群と非投与群で比較する際，発生率の評価ではリスク比やオッズ比が用いられることが多い。投与群での事象の発生率を p_1 とし，非投与群での事象の発生率を p_0 とすると，リスク比 (Risk Ratio) は $RR = p_1/p_0$ で定義され，オッズ比 (Odds Ratio) は $OR = \{p_1/(1-p_1)\}/\{p_0/(1-p_0)\}$ で定義される。以下の各問に答えよ。

〔1〕リスク比 RR と非投与群での事象の発生率 p_0 を用いてオッズ比 OR を求める式を導け。

〔2〕観測データが，

	発生	非発生	計
投与群	a	b	m
非投与群	c	d	n
計	s	t	N

と表されるとき，オッズ比は通常 $OR^* = (ad)/(bc)$ と推定され，対数オッズ比 $\log OR^*$ の標準誤差は，

$$SE[\log OR^*] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

で与えられる。これを参考に，リスク比 RR の推定値 RR^* と対数リスク比 $\log RR^*$ の標準誤差を求めよ。

〔3〕ある研究では，上問〔2〕の記号を用いると $m = 51$，$n = 50$ で $OR^* = 3.38$，$\hat{p}_0 = c/n = 0.58$ であったという（数値は小数第 3 位を四捨五入）。このとき，リスク比の推定値 RR^*，およびその 95 %信頼区間を求めよ。

〔4〕上問〔3〕の結果を踏まえ，リスク比をオッズ比で近似する場合の注意点を述べよ。

解答例

〔1〕オッズ比の式 $OR = p_1(1-p_0)/\{p_0(1-p_1)\}$ を p_1 について解くと，

$$p_1 = \frac{p_0 OR}{1 - p_0 + p_0 OR}$$

となるので，これより，

$$RR = \frac{p_1}{p_0} = \frac{OR}{1 - p_0 + p_0 OR} \quad \cdots\cdots ①$$

となり，

$$OR = \frac{p_1}{p_0} \times \frac{1-p_0}{1-p_1} = RR \times \frac{1-p_0}{1-p_0 RR}$$

を得る。

〔2〕リスク比は，

$$RR^* = (a/m)/(c/n) \quad \cdots\cdots ②$$

で推定され，標準誤差は，

$$SE[\log RR^*] = \sqrt{\frac{1}{a} - \frac{1}{m} + \frac{1}{c} - \frac{1}{n}} \quad \cdots\cdots ③$$

となる。これは以下のように示される。

一般に，X を二項分布 $B(n,\theta)$ に従う確率変数とする。$E[X/n] = \theta$，$V[X/n] = \theta(1-\theta)/n$ である。$\log(X/n) = \log X - \log n$ の分散の近似値をデルタ法により求める。$\log x$ の微分は $(\log x)' = 1/x$ であるので，$V[g(X)] \fallingdotseq \{g'(\theta)\}^2 V[X]$ の式より，

$$V\left[\log\left(\frac{X}{n}\right)\right] = \frac{1}{\theta^2} \times \frac{\theta(1-\theta)}{n} = \frac{1-\theta}{n\theta} = \frac{1}{n\theta} - \frac{1}{n} \quad \cdots\cdots ④$$

を得る。$n\theta$ の推定値は x であるので，

$$SE[\log(X/n)] = \sqrt{(1/x) - (1/n)}$$

となる。問題文の表の記号では，a は二項分布 $B(m, p_1)$ に従い，c は二項分布 $B(n, p_0)$ に従っていて，それらは互いに独立であるので，対数リスク比

$$\log\{(a/m)/(c/n)\} = \log(a/m) - \log(c/n)$$

の標準誤差は④の組合せにより，③となる。

〔3〕$OR^* = 3.38$，$\hat{p}_0 = 0.58$ であるので $c = \hat{p}_0 n = 0.58 \times 50 = 29$ である。そして①より $RR^* = 1.42$ を得る。$\log RR^* = 0.35$ であり，$\hat{p}_1 = \hat{p}_0 RR^* = 0.58 \times 1.42 = 0.824$ となるので，$a = \hat{p}_1 m = 42$ と求められる。よって，$\log RR^*$ の標準誤差は，

$$SE[\log RR^*] = \sqrt{\frac{1}{42} - \frac{1}{51} + \frac{1}{29} - \frac{1}{50}} \fallingdotseq 0.137$$

となり，$\log RR$ の 95 %信頼区間は $0.351 \pm 1.96 \times 0.137 = (0.083, 0.618)$ であるので，RR の 95 %信頼区間は $(\exp[0.083], \exp[0.618]) = (1.09, 1.86)$ となる。

〔4〕事象の発生率が極めて小さい場合には，オッズ比はリスク比のよい近似を与える。しかし，事象の発生率がある程度大きい場合はその限りではなく，〔3〕で見るように，オッズ比はリスク比をかなりの程度過大評価する。