受験番号	

2024年度　一橋大学大学院ソーシャル・データサイエンス研究科（修士課程）
二次選考（筆記試験）

統計学・情報学　入試問題

2023年9月実施

【 注 意 事 項 】

1．「解答はじめ」というまで開いてはいけない。試験時間：10時30分～11時30分。

2．問題は本文11ページ、解答用紙は２枚である。下書き用紙２枚は自由に使ってよい。
　試験開始後、直ちに確認し、ページ数・枚数が異なる場合は、手を挙げなさい。

3．試験開始後、問題の表紙、解答用紙、下書き用紙に受験番号を正確に記入しなさい。氏名は記入しないこと。
　問題、解答用紙、下書き用紙は一切持ち帰らないこと。

4．統計学①　統計学②　情報学①　情報学②の４つの問題から２つを選んで、それぞれ別々の解答用紙に日本語または英語で解答を記入しなさい。
　解答用紙上段の「問題名」の中から、各解答用紙で解答する問題名１つを選んで、○で囲みなさい。１枚の解答用紙で２つ以上の問題名を○で囲んでいる場合や、問題名を○で囲んでいない場合は得点を与えないので、以下の記入例をよく確認すること。

《記入例》　統計学①　及び　情報学①　を解答する場合

解答用紙1枚目

問題名	以下の中から解答する問題名1つを選んで、○で囲みなさい。 もう一方の解答用紙で選択した問題を選ぶことはできません。
	統計学① ／ 統計学② ／ 情報学① ／ 情報学②

解答用紙2枚目

問題名	以下の中から解答する問題名1つを選んで、○で囲みなさい。 もう一方の解答用紙で選択した問題を選ぶことはできません。
	統計学① ／ 統計学② ／ **情報学①** ／ 情報学②

5．解答を記入する際は、解答する問の番号（問１、問１.(1)等）を必ず記入すること。

統計学①

以下の問1，問2の両方に解答せよ．

問1

以下の8つの小問のうち4問選択し解答せよ．5問以上選択した場合は，問題番号が小さい順に4問を採点対象とする．

1. 決定係数とは何か説明せよ．定義式も記述すること．
2. ガウスマルコフの定理とは何か説明せよ．
3. 二値変数に対する一般化線形モデルの例を挙げよ．
4. 推定量の不偏性，一致性，有効性とは何か説明せよ．
5. 繰り返し期待値の法則とは何か説明せよ．
6. 分布の無記憶性とは何か，具体例を挙げて説明せよ．
7. 中心極限定理とは何か説明せよ．定義式も記述すること．
8. 統計的仮説検定における第一種，第二種の誤りとは何か説明せよ．

問2

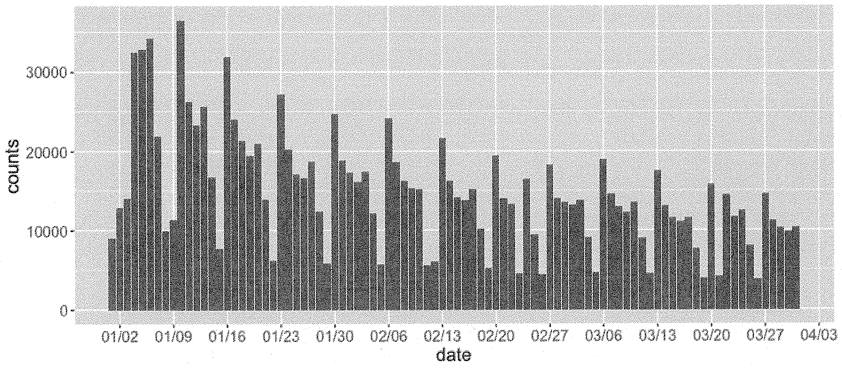


図 1　2023/1/1から2023/3/31までの東京都コロナ検査実施件数(東京都オープンデータサイトよりダウンロード，https://portal.data.metro.tokyo.lg.jp/)

図1は2023/1/1から2023/3/31までの東京都コロナ検査実施件数のグラフである．y_tは東京都における時点tでのコロナ検査実施数とする．y_tは強度$\lambda_t > 0$のポアソン分布に従うとする．つまり，

$$y_t \sim \text{Poisson}(\lambda_t), \quad t = 1, \ldots, T$$

このとき，強度関数の対数値を説明変数に回帰させるモデルを考える．つまり，

$$\log \lambda_t = \mathbf{x}_t^\top \boldsymbol{\beta}$$

$\mathbf{x}_t$は時点tでの$p \times 1$の説明変数ベクトルとする．以下の小問に答えよ．

1. 図1より，平日に比べて休日の検査実施件数が少ない傾向があることがわかる．この状況を説明するための説明変数を1つ設定せよ．
2. 図1より，検査実施件数が時間に関して単調に減少する傾向があることがわかる．この状況を説明するための説明変数を1つ設定せよ．
3. 図1より，検査実施件数は前日の検査実施件数と強い相関があると考えられる．この状況を説明するための説明変数を1つ設定せよ．
4. 上記で設定した説明変数を含めたモデルを推定したい．$\boldsymbol{\beta}$をどのように推定すれば良いか議論せよ．

統計学②

以下の問1，問2，問3より1つ選択し解答せよ．2つ以上解答した場合は，問題番号が小さい問を採点対象とする．

問1

n個の標本とm個のグループを考え，グループjはn_j $(j = 1, \ldots, m)$個の標本を持つ$(\sum_{j=1}^{m} n_j = n)$．各標本はいずれかのグループに属し，グループの個数と各標本がどのグループに属しているかは既知とする．この時，以下の階層モデルを考える．以下の小問に解答せよ．

$$\theta_1, \ldots, \theta_m | \mu, \tau^2 \sim_{\text{i.i.d}} N(\mu, \tau^2)$$
$$y_{1,j}, \ldots, y_{n_j,j} | \theta_j, \sigma^2 \sim_{\text{i.i.d}} N(\theta_j, \sigma^2), \quad j = 1, \ldots, m$$

1 $\mathrm{Var}[y_{i,j}|\theta_j, \sigma^2]$と$\mathrm{Var}[y_{i,j}|\mu, \tau^2, \sigma^2]$を計算せよ．また，それぞれはどのように解釈できるか議論せよ．ただし，$\mathrm{Var}[y|x]$はxが与えられた下での，yの条件付き分散を意味する．

2 $\mathrm{Cov}[y_{i_1,j}, y_{i_2,j}|\theta_j, \sigma^2]$は負，正，0のどれか．また，$\mathrm{Cov}[y_{i_1,j}, y_{i_2,j}|\mu, \tau^2, \sigma^2]$についても同様に答えよ．ただし，$\mathrm{Cov}[y_1, y_2|x]$は$x$が与えられた下での，$y_1$と$y_2$の条件付き共分散を意味する．

3 θ_jに関する事後分布$p(\theta_j|\mu, \tau^2, \sigma^2, \mathbf{y}_1, \ldots, \mathbf{y}_m)$を求めよ．ただし，$\mathbf{y}_j = (y_{1,j}, \ldots, y_{n_j,j})$ $(j = 1, \ldots, m)$とする．

4 μに対する事前分布を$p(\mu)$とする．この時，以下を示せ．

$$p(\mu|\theta_1, \ldots, \theta_m, \tau^2, \sigma^2, \mathbf{y}_1, \ldots, \mathbf{y}_m) = p(\mu|\theta_1, \ldots, \theta_m, \tau^2)$$

また，この結果の意味を解釈せよ．

問2

1. 互いに排反な事象列$B_1, \ldots, B_n$について，$\min_{k=1,\ldots,n} P(B_k) > 0$ かつ $\bigcup_{k=1}^{n} B_k = \Omega$を満たすとする．ただし Ω は標本空間である．

1. ベイズの定理

$$P(B_j \mid A) = \frac{P(A \mid B_j) P(B_j)}{\sum_{k=1}^{n} P(A \mid B_k) P(B_k)}$$

を証明せよ．

2. ある病気の罹患率が0.2%であるとする．この病気を発見するための検査法は，罹患している人の99%に陽性反応を示し，また，罹患していない人にも4%の陽性反応を

示す．Aさんがこの検査を受けて陽性反応が出たとき，Aさんが実際にこの病気に罹患している確率を求めよ．

2. 正の値をとる離散型確率変数Xは$E[X] < \infty$を満たすとする．

   1. マルコフの不等式

$$P(X \geq x) \leq \frac{E[X]}{x}, \qquad x > 0$$

   を証明せよ．

   2. 大数の弱法則について，どのような定理か述べ，適当な仮定を置いて証明せよ．

問3

確率変数Xはパラメータ(n, p)の二項分布に従うとする．ただし，nは既知の自然数，pは$(0, 1)$に値をとる未知定数とする．

1. pの対数尤度関数を定義し，pの最尤推定量 $\hat{p}$を導出せよ．
2. $\hat{p}$ は pの一致推定量であることを示せ．
3. フィッシャー情報量を導出せよ．
4. $\hat{p}$ の漸近分布を導出し，それを用いて仮説 $H_0 : p = 1/2$ vs. $H_1 : p \neq 1/2$ を検定する手順を説明せよ．

情報学①

以下の問いにすべて答えよ．プログラムのソースコードを記述する際には，処理内容が判別可能であれば，その形式（疑似コード，プログラミング言語の種類）は問わない．

問1

基本的なデータ構造のひとつにキューまたは待ち行列と呼ばれるものがある．キューQにデータを加えることを，enqueueと呼び，Q.enqueue(x)でデータxをQに加えたとする．同様にQ.dequeue()で，先頭のデータを取り出すことを示すとする． 以下の例と回答例にしたがって，以下のプログラム1 Queueに示された操作を行ったときの出力を出力される順に書け．

以下のプログラムで最初の行のQ.init()はキューQを空の状態に初期化することを，print()はカッコ内のデータを出力することを示すこととする．

例

```
1  Q.init()
2  Q.enqueue(3)
3  print(Q)
4  Q.enqueue(5)
5  print(Q)
6  Q.dequeue()
7  Q.dequeue()
8  print(Q)
```

解答例

3
3, 5
（空）

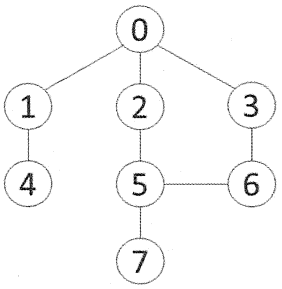
プログラム 1：キュー

```
1  Q.init()
2  Q.enqueue(15)
3  Q.enqueue(8)
4  Q.dequeue()
5  print(Q)
6  Q.enqueue(20)
7  Q.dequeue()
8  print(Q)
9  Q.enqueue(11)
10 Q.dequeue()
11 print(Q)
```

問2

幅優先探索は，グラフ上の頂点を始点から連結なすべての頂点を発見順に訪問するアルゴリズムである．

例えば，図1のグラフにおいて，頂点0を始点とすると，最初に頂点0から連結な頂点1,2,3を発見し，次にこれらの頂点を任意の順番で訪問する．以後，頂点1を訪問したときに頂点4を，頂点2を訪問したときに頂点5を，頂点3を訪問したときに頂点6を発見し，さらに，頂点5を訪問後に頂点7を発見し，発見順にこれらを訪問する（可能な訪問順序の一つとして，0→1→2→3→4→5→6→7の順）．このように，幅優先探索は，始点から近い順にすべての訪問可能な頂点を訪問するアルゴリズムである．



左のグラフをあらわす隣接行列

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

図 1　グラフと隣接行列

幅優先探索は，キューを用いて実装することができる．次ページの幅優先探索のプログラム2は各ノードに対して，何らかの値を計算している (問3を参照)．このプログラムが幅優先探索であることに留意して，空欄 ① に入るプログラムを書け．空欄 ① 内のプログラムは直前のif節の中にあるとする．Gは探索するグラフの隣接行列を表し，Gが表すグラフの各頂点は0からNの番号が付けられ，頂点0を始点とする．また，プログラム中の各行の#以下はコメントであり，処理に関与しない．

隣接行列: グラフを表す正方行列で，各要素は対応する頂点間の辺の有無をあらわす．隣接行列Gのi行j列の要素G[i][j]は，頂点i, j間に辺が存在するとき1，辺が存在しないとき0である．プログラムでは隣接行列は2次元配列で表されている (隣接行列の例は，図1を参照)．

プログラム 2：幅優先探索

```
1   def bfs(G):
2     N = len(G)   # len()は配列の要素数を返す。ここでは、NはGが表すグラフの頂点数となる
3     D = [-1 for  i in range(N)] # Dは要素数Nで、すべての要素が-1である配列
4     cur = 0  # 現在訪問中の頂点を始点0とする
5     print(cur)
6     D[0]=0 # 頂点0を訪問済みにしておく
7     Q.init()
8     Q.enqueue(0)
9     while len(Q) != 0 : # Qが空でない間繰り返す
10      cur = Q.dequeue()
11      print(cur)
12      for dst in range(N): # dstを0からN-1まで1ずつ増やしながら、以下の節をN回繰り返す
13        if G[cur][dst] == 1 and D[dst] == -1: # 頂点curとdstに辺があり、dstが未訪問ならば、以下の処理を行う
14          [          ①          ]
15          D[dst] = D[cur] + 1
```

問 3

幅優先探索終了後の問2のプログラム2：幅優先探索における配列Dのそれぞれの要素は何を表しているか，簡潔に書け.

問 4

幅優先探索によって，図2のグラフを探索するときのプログラム2の出力を書け.

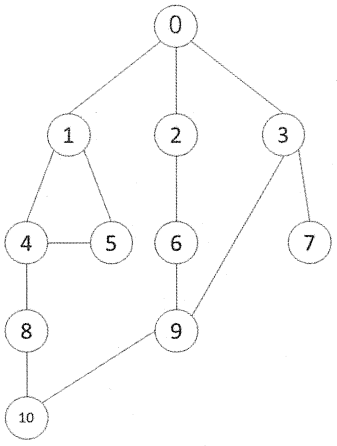


図 2　グラフ

問5

X(旧Twitter)やInstagramのようなソーシャルネットワーキングサービス (SNS)) では，他者とつながりを持つことができ（followなど），各ユーザーを頂点，つながりを辺とみなすことで，つながりの関係をグラフとみなすことができる．そこで，SNS上での友達関係のつながりの広がりを調べるため，ある頂点（あるユーザー）を始点に幅優先探索を利用することを考える．プログラム2：幅優先探索と同様の幅優先探索アルゴリズムはすべての連結な頂点を訪問する手続きであるため，非常に大量のユーザーを訪問することになり，現実的な実行時間では終了しないことがある．このような場合，K次のつながり（始点から直接つながっているユーザーを1次のつながり，そのユーザーからつながっているユーザーを2次のつながり...とする）までに幅優先探索の訪問先を限定することで，実行時間を短縮できる．

例を参考に，プログラム2：幅優先探索を訪問先が始点0からのK次のつながりまでに限定されるように書き換えよ．

例) 6行目を削除，7行目の下に新たな行を追加，9行目を書き換えたい場合，

1. 6行目を削除
2. 7行目の下に以下のコードを挿入

   Q.enqueue(K)
3. 9行目を以下のコードに変更

   while len(Q)>0 and K>0:

情報学②

主成分分析とニューラルネットワークに関する以下の問1から問3にすべて答えよ.

問1

主成分分析に関する問題：
今，表1のような成績データが与えられたとする.

表 1 成績表（10点満点）

名前＼科目	国語	数学	理科
矢部	2	5	10
武田	2	1	2
谷垣	2	3	6
福山	7	2	4
一条	7	4	8

この成績データから，各科目の得点間の（標本）分散共分散行列は以下で与えられる.

$$\mathbf{\Sigma} = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 2 & 4 \\ 0 & 4 & 8 \end{pmatrix}$$

(1) 分散共分散行列$\mathbf{\Sigma}$を対角化し，第一主成分と第二主成分に対応する２つの固有値とノルムが1の固有ベクトルの組を算出せよ．また，得られた固有ベクトル同士が直交することを示せ.

(2) (1)で得られた第一主成分と第二主成分に対応する固有値・固有ベクトルの意味について，「寄与率」「理系」「文系」という単語を用いて論ぜよ.

問2

ニューラルネットワークに関する問題：

(1) 等式によって表現される制約付き関数極大化・極小化問題を解く方法としてラグランジュの未定乗数法がある．今，ある等式制約$g(\mathbf{x}) = 0$の元で関数$f(\mathbf{x})$が極値をとるd次元ベクトル$\mathbf{x}$を求めることを考える．ラグランジュの未定乗数法で，この制約条件付き関数極大化・極小化問題を解くには，任意定数であるラグランジュ乗数λを導入した関

数$L(\mathbf{x},\lambda)=f(\mathbf{x})+\lambda g(\mathbf{x})$を導入した上で，

$$\begin{cases} \dfrac{\partial L(\mathbf{x},\lambda)}{\partial \mathbf{x}}=\mathbf{0} \\ \dfrac{\partial L(\mathbf{x},\lambda)}{\partial \lambda}=0 \end{cases}$$

という連立微分方程式を解くと極大値・極小値が得られることが知られている．ここで，$\frac{\partial L(\mathbf{x},\lambda)}{\partial \mathbf{x}}=\left(\frac{L(\mathbf{x},\lambda)}{\partial x_1},\frac{L(\mathbf{x},\lambda)}{\partial x_2},\ldots,\frac{L(\mathbf{x},\lambda)}{\partial x_d}\right)^\top$，$\top$は転置行列を表す．
これらの前提を用いて，$x_1^2+x_2^2=1$という制約の元で，$f(x_1,x_2)=2x_1+x_2$を最小化するx_1,x_2及び，その最小値を求めよ．

(2) オートエンコーダ（自己符号化器）型ニューラルネットワーク（以下，オートエンコーダ）とは，与えられたデータセットをできるだけ復元できる次元圧縮を目的としたニューラルネットワークモデルの呼称である．以下，図1のようなオートエンコーダによって次元圧縮を行う場合を考える．
今，n個のd次元実数ベクトル$\mathbf{x}$によって構成されるデータセット$\{\mathbf{x}_i\}_{i=1}^n$が与えられたとする．ここで，説明を簡単にするために$\mathbf{x}$のサンプル平均は$\mathbf{0}$であるとする．この時，d次元ベクトル$\mathbf{x}$を，d_h次元実数ベクトル$\mathbf{h}$に圧縮し，それをさらに元のd次元実数ベクトル$\mathbf{y}$に復元することを考える．つまり，図1のオートエンコーダの入力次元をd_in，出力次元をd_outとすると，$d_\mathrm{in}=d_\mathrm{out}=d>d_\mathrm{h}$であるとする．オートエンコーダの入出力関係は$d\times d_\mathrm{h}$実行列$\mathbf{W}^{(1)}$及び，$d_\mathrm{h}\times d$実行列$\mathbf{W}^{(2)}$を用いて，$y_l=\sum_{k=1}^{d_\mathrm{h}} w_{kl}^{(2)}h_k$，$h_k=f(\sum_{j=1}^{d} w_{jk}^{(1)}x_j)$で与えられるものとする．ここで$f$はニューラルネットワークの用語でよく活性化関数と呼称される．今，活性化関数fが恒等写像$f(x)=x$であり，変換行列が層間で共有されている，即ち$\mathbf{W}^{(1)}=\mathbf{W}$，$\mathbf{W}^{(2)}=\mathbf{W}^\top$であるとする．この場合，オートエンコーダの出力の入力との一致度である再構成誤差$L_\mathrm{ae}(\mathbf{W})$は，平均二乗誤差を用いて以下のように与えられる．

$$L_\mathrm{ae}(\mathbf{W})=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{y}_i)^\top(\mathbf{x}_i-\mathbf{y}_i)=\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i-\mathbf{W}\mathbf{W}^\top\mathbf{x}_i\right)^\top\left(\mathbf{x}_i-\mathbf{W}\mathbf{W}^\top\mathbf{x}_i\right)$$

$\mathbf{W}$が直交行列となる制約（$\mathbf{W}^\top\mathbf{W}=\mathbf{I}$，$\mathbf{I}$は$d_\mathrm{h}\times d_\mathrm{h}$の単位行列）の元で，再構成誤差$L_\mathrm{ae}(\mathbf{W})$を最小化する$\mathbf{W}$を求めることを考える．つまり，行列$(\mathbf{W}^\top\mathbf{W}-\mathbf{I})$の全ての要素が0となるという複数制約の元で，$L_\mathrm{ae}(\mathbf{W})$を最小化する．この複数制約を表現するために，互いに異なるラグランジュ未定乗数λ_{ij}を要素とする$d_\mathrm{h}\times d_\mathrm{h}$の実行列$\mathbf{\Lambda}$を導入する．このような$\mathbf{\Lambda}$と(2)で導出した$L_\mathrm{ae}(\mathbf{W})$を用いることで，$\mathbf{W}$が直交行列となる制約の元での$L_\mathrm{ae}(\mathbf{W})$の最小化問題が，以下の関数の最小化問題として与えられることを示せ．

$$\begin{aligned} L(\mathbf{W},\mathbf{\Lambda}) &= -\mathrm{tr}\left(\mathbf{W}^\top\mathbf{S}\mathbf{W}\right)+\mathrm{tr}\left[\mathbf{\Lambda}\left(\mathbf{W}^\top\mathbf{W}-\mathbf{I}\right)\right], \\ \mathbf{S} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top. \end{aligned}$$

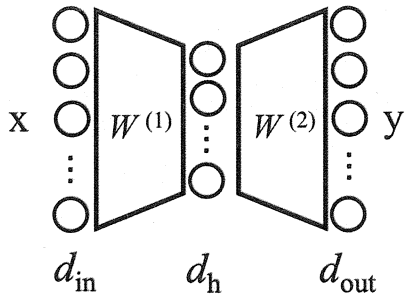


図 1　オートエンコーダ型ニューラルネットワーク

ここで，trは行列のトレース（対角和）を表すものとし，証明においては，問題文中の任意のベクトル$\mathbf{v}$に対して，$\mathrm{tr}(\mathbf{v}\mathbf{v}^\top) = \mathbf{v}^\top\mathbf{v}$という関係式が成り立つとして良いものとする．

(3) $L(\mathbf{W}, \mathbf{\Lambda})$を最小化するための必要条件$\frac{\partial L(\mathbf{W},\mathbf{\Lambda})}{\partial \mathbf{W}} = \mathbf{O}$，$\frac{\partial L(\mathbf{W},\mathbf{\Lambda})}{\partial \mathbf{\Lambda}} = \mathbf{O}$から，$L(\mathbf{W}, \mathbf{\Lambda})$を最小化する$\mathbf{W}$，$\mathbf{\Lambda}$に対して，以下の関係式が成り立つことを示せ．

$$\mathbf{S}\mathbf{W} = \mathbf{W}\mathbf{\Lambda},$$
$$\mathbf{W}^\top\mathbf{W} = \mathbf{I}.$$

さらに，この関係式が満たされる場合の関数$L(\mathbf{W}, \mathbf{\Lambda})$の値$\tilde{L}$が以下で与えられることを示せ．

$$\tilde{L} = -\mathrm{tr}(\mathbf{\Lambda})$$

証明においては，問題文中の任意の行列$\mathbf{X}$,$\mathbf{A}$に対して，次の関係式$\frac{\partial \mathrm{tr}(\mathbf{X}\mathbf{A}\mathbf{X}^\top)}{\partial \mathbf{X}} = \mathbf{X}(\mathbf{A} + \mathbf{A}^\top)$，$\frac{\partial \mathrm{tr}(\mathbf{X}\mathbf{A})}{\partial \mathbf{X}} = \mathbf{A}^\top$が成立するとして良いものとする．

問 3

問 1，問 2 で得られた知見を元に，主成分分析とオートエンコーダの関係性を論ぜよ．（ヒント：問2(3)から得られる最小解のうち，$\mathbf{\Lambda}$が対角行列となる場合の解に着目せよ．）