受験番号	

2025 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

2024 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～12 時 30 分。

２.問題は本文８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、〇で囲みなさい。１枚の解答用紙で２つ以上の問題名を〇で囲んでいる場合や、問題名を〇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、〇で囲みなさい。 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、〇で囲みなさい。 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. クロスヴァリデーション(交差検証)とはどのような手法か説明せよ．
2. マルコフの不等式を説明せよ．
3. ベイズの定理を説明せよ．
4. 中心極限定理とは何か説明せよ．定義式も記述すること．
5. 線形回帰における，分散均一性の仮定について説明せよ．
6. XとYは独立な確率変数で，それぞれの平均と分散をμ_x, μ_y, σ_x^2, σ_y^2とする．このとき，XYとYの共分散を計算せよ．
7. 検定における検出力関数とは何か説明せよ．
8. 尤度比検定とは何か説明せよ．

問 2

線形重回帰モデルを考える．サンプルサイズ$n = 30$のシミュレーションデータを以下の式より生成する．

$$
\begin{aligned}
&y_i = 3 + 2x_{1i} + 0.5x_{2i} + \epsilon_i, \quad \epsilon_i \sim_{i.i.d.} N(0,1) \\
&x_{1i}, x_{2i} \sim_{i.i.d.} N(0,1), \quad i = 1, 2, \ldots, n
\end{aligned}
$$

追加的に$x_{3i} \sim N(0.8x_{1i}, 0.04),\ i = 1, 2, \ldots, n$を生成し，以下の2つのモデルを推定する．

$$
\begin{aligned}
&\text{Model1} : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \\
&\text{Model2} : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)
\end{aligned}
$$

Model1とModel2をそれぞれ最小二乗法により推定した結果は表1の通り． このとき，以下の設問に答えよ．

表 1　線形回帰モデルの最小二乗法による推定結果

Model1	推定値	標準誤差	Model2	推定値	標準誤差
β_0	3.113	0.168	β_0	3.128	0.171
β_1	2.001	0.183	β_1	2.684	0.994
β_2	0.500	0.176	β_2	0.564	0.200
			β_3	-0.847	1.210

1. 各回帰係数の両側検定を考える．帰無仮説と対立仮説を明記し，検定統計量tを定義せよ．
2. 標準誤差，p値の定義を明記せよ．
3. x_1とx_3の相関は0.96であった． この場合，Model2の回帰係数の推定結果にどのような影響が生じうるか理由を含めて議論せよ．
4. Model1とModel2を最小二乗法により推定した決定係数と自由度調整済み決定係数は表2の通り．決定係数と自由度調整済み決定係数のそれぞれの定義を明記し，違いを説明せよ．また，決定係数はModel2が高く，自由度調整済み決定係数はModel1が高い理由を説明せよ．

表 2 決定係数と自由度調整済み決定係数

	決定係数	自由度調整済み決定係数
Model1	0.825	0.812
Model2	0.828	0.808

統計学②

以下の問1，問2，問3より2つ選択し解答せよ．全て解答した場合は，問1，問2を採点対象とする．

問 1

確率変数(X, Y)の同時分布は以下で定義される．

$$f(x,y) \propto \exp\left\{-\frac{1}{2}[Ax^2y^2 + x^2 + y^2 - 2Bxy - 2Cx - 2Dy]\right\},$$
$$A > 0, \quad -\infty < x < \infty, \quad -\infty < y < \infty$$

ただし，A, B, C, Dは定数とする．$\propto$は比例記号であり「左辺は右辺に比例する」ことを意味する．以下の設問に答えよ．

1. $X \mid Y = y \sim N(\frac{By+C}{Ay^2+1}, \frac{1}{Ay^2+1})$を示せ．
2. 同様に$Y \mid X = x$が従う分布を導出せよ．
3. $A = 1, B = 0, C = D = 4$のとき，同時分布は二峰であることを示せ．

問 2

すべての$n \in \mathbb{N}$について$\mathbb{E}\, X_n = \mu$かつ$\mathrm{Var}(X_n) = \sigma^2$である確率変数列$\{X_n\}$について考える．$n \to \infty$のとき，以下の設問に答えよ．

1. 確率収束 $X_n \xrightarrow{p} \mu$ の定義を述べよ．
2. $\mathrm{Var}\,(X_n) \to 0$ ならば $X_n \xrightarrow{p} \mu$ であることを示せ．
3. $\{X_n\}$ は互いに独立であるとする．次を示せ．

$$\frac{2}{n(n+1)} \sum_{j=1}^{n} jX_j \xrightarrow{p} \mu$$

問 3

互いに独立な2次元確率ベクトルの列

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_i \\ \mu_i \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right), \quad i = 1, \ldots, n$$

を考える．ただし，$\mu_1, \ldots, \mu_n$, σ^2 は未知のパラメータである．以下の設問に答えよ．

1. μ_i, σ^2 の最尤推定量 $\hat{\mu}_i, \hat{\sigma}^2$ を求めよ．
2. 最尤推定量 $\hat{\sigma}^2$ は σ^2 の一致推定量でないことを示したうえで，一致性をもつように $\hat{\sigma}^2$ を修正せよ．

3. 次の推定量

$$\tilde{\sigma}^2 = \frac{1}{2n}\left\{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2\right\}$$

が σ^2 に対して一致性をもつための条件について議論せよ．ただし，$\bar{X}_n, \bar{Y}_n$ は標本平均で，$\bar{\mu}_n = n^{-1}\sum_{i=1}^{n}\mu_i$, $s_n^2 = n^{-1}\sum_{i=1}^{n}(\mu_i - \bar{\mu}_n)^2$ とし，極限 $\lim_{n\to\infty}\bar{\mu}_n = \mu_*$, $\lim_{n\to\infty} s_n^2 = s_*^2$ は実数の範囲で存在すると仮定する．

情報学①

以下の問いにすべて答えよ．サンプルプログラムは Python を前提として書かれているが，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． Python では，ダブルクォート3つ（"""　"""）で囲まれた部分はコメントである．

ほとんどの Python の実装で用いられているソーティングアルゴリズムはティムソート (Timsort) と呼ばれているものである．Timsort はマージソート（merge sort）と挿入ソート（insertion sort）を組み合わせた安定ソート（stable sort）アルゴリズムである．

問 1

安定ソートとは何か説明せよ．また，安定ソートが必要となる状況と，必要ではない状況について，具体例を挙げて説明せよ．

問 2

in-place アルゴリズムとは何か説明せよ．また，in-place アルゴリズムが必要となる状況と，必要ではない状況について，具体例を挙げて説明せよ．

問 3

以下に示すマージソートのマージ部分のプログラムを作成せよ．また，このマージ操作に必要な時間計算量（time complexity）および空間計算量（space complexity）について説明せよ．

プログラム 1　マージソートのマージ部分

```
1 def merge(left, right):
2     """
3     Given that left and right are lists sorted in ascending order with integer elements,
4     return a sorted list consisting of the same elements as the union of left and right.
5     """
6     Please fill in this part.
```

問 4

Timsort では単純な挿入ソートではなく，2分挿入ソートが用いられている．2分挿入ソートは挿入する箇所を2分探索する挿入ソートの改良である．以下に示す2分挿入ソートの2分探索部分のプログラムを作成せよ．また，この操作の時間計算量について説明せよ．

プログラム 2　2分挿入ソートの2分探索部分

```
1 def bsearch(L, x, low, high):
2     """
3     Let L be a list sorted in ascending order with integer elements. Low and high
4     represent the indices that define the current range within the search is being
5     performed. The bsearch function returns the index where x is found in L, or -1
6     if it is not found.
7     """
8     Please fill in this part.
```

問 5

Timsort では通常のマージソートとは異なり，あらかじめ入力の配列をある程度のサイズの配列に分割し，分割された配列内は挿入ソートを用いてソートを行い，その後ソート済みの配列に対してマージソートを行う．このように2つのソートアルゴリズムを組み合わせる理由について，それぞれのソートアルゴリズムの性質を明らかにしつつ，説明せよ．

情報学②

以下の問いに全て答えよ．

ある検査の値xから感染症感染の有無 t（$t=0$であれば感染なし，$t=1$であれば感染あり）をロジスティック回帰を用いて識別するケースを考える． ロジスティック回帰では，検査の値xの患者が感染している確率を

$$P(t=1|x) = y = \sigma(w_0 + w_1 x)$$

でモデル化する．ここで，関数σはシグモイド関数と呼ばれ， $\sigma(z) = 1/(1+\exp(-z))$ で与えられる．

問 1

シグモイド関数の微分$\sigma'(z)$が$\sigma(z)(1-\sigma(z))$となることを示せ．

問 2

上記のロジスティック回帰モデルにおけるパラメータw_1は， 「xが1単位増加すると○○がw_1単位増加する」と解釈できる． ○○に当たる量について説明せよ．

問 3

患者N人分の検査の値と感染の有無のデータをそれぞれ$\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$と$\mathbf{T} = \{t_1, t_2, \ldots, t_N\}$で表す．このとき，交差エントロピー型誤差関数は負の対数尤度を用いて下記のように表される．

$$E(w_0, w_1) = -\log P(\mathbf{T}|\mathbf{X}) = -\sum_{n=1}^{N} (t_n \log y_n + (1-t_n)\log(1-y_n)).$$

なお，$y_n = \sigma(w_0 + w_1 x_n)$であり，logは自然対数を表す．

交差エントロピー型誤差関数の勾配ベクトルが下記のように与えられることを示せ．

$$\nabla E(w_0, w_1) = \left[\sum_{n=1}^{N}(y_n - t_n), \quad \sum_{n=1}^{N}(y_n - t_n)x_n\right].$$

問 4

上記の交差エントロピー型誤差関数が最小になるw_0とw_1を勾配法（最急降下法）で求めるアルゴリズムを疑似コードを用いて説明せよ．なお，処理内容が判別可能であれば，疑似コードの形式は問わない．

問 5

データを基に識別を行い，以下のような結果を得たとする．

- 本当に感染しており，感染していると識別された患者がN_1人いた．
- 本当は感染しているが，感染していないと識別された患者がN_2人いた．
- 本当は感染していないが，感染していると識別された患者がN_3人いた．
- 本当に感染しておらず，感染していないと識別された患者がN_4人いた．

「この識別手法の有用性をどのように評価するのか?」という問題を考える．様々な評価指標があり，例えば，適合率（precision）と呼ばれる指標は$N_1/(N_1+N_3)$で与えられる． また，感度（sensitivity）と呼ばれる指標は$N_1/(N_1+N_2)$で与えられる．

適合率（precision）で有用性を評価することが適切なケースを具体例と理由を挙げて説明せよ（具体例は感染症と関係なくても良い）．

感度（sensitivity）で有用性を評価することが適切なケースを具体例と理由を挙げて説明せよ（具体例は感染症と関係なくても良い）．