２０２５年度
東京大学大学院学際情報学府学際情報学専攻
（生物統計情報学コース）
入学試験問題
専門科目
（２０２４年８月１日１４：００～１６：００）

試験開始の合図があるまで問題冊子を開いてはいけません．開始の合図があるまで，下記の注意事項をよく読んでください．

1. 本冊子は，生物統計情報学コースの受験者のためのものである．
2. 本冊子の本文は１０ページである．落丁，乱丁，印刷不鮮明の箇所などがあった場合には申し出ること．
3. 本冊子には，第１問から第３問までの計３問の問題が収録されている．第１問は択一式問題であり，第２問および第３問は記述式問題である．いずれも必須問題であり，3問すべてに解答すること．
4. 本冊子の問題は，日本語文で記述されている．
5. 解答用紙は３枚ある．問題ごとに解答用紙１枚を使用すること．このほかにメモ用紙が１枚ある．なお，解答用紙のみが採点の対象になる．
6. 解答用紙の上方の欄に，問題の番号および受験番号を必ず記入すること．問題番号および受験番号を記入していない答案は無効である．
7. 解答には必ず黒色鉛筆（または黒色シャープペンシル）を使用すること．
8. 解答は日本語によるものとする．
9. 試験開始後は，中途退場を認めない．
10. 本冊子，解答用紙，およびメモ用紙は持ち帰ってはならない．
11. 次の欄に受験番号と氏名を記入せよ．

受験番号	
氏　名	

生物統計情報学 第１問

以下の問い(1-1)～(1-15)に答えよ．解答用紙には問いの番号と解答のみを，問いの番号の順序に従って記載せよ．

(1-1) あるクラスの学生 5 名の身長のデータを測定したとき，以下の測定値のデータが得られた．このデータに関する記述として，次のア～オのうちから正しいものを一つ選べ．

	A さん	B さん	C さん	D さん	E さん
身長(cm)	171	172	173	175	180

ア．平均値は 173cm である．

イ．中央値は 173cm よりも大きい．

ウ．平均値は中央値よりも大きく，分布は右に裾を引いている．

エ．平均値は中央値よりも大きく，分布は左に裾を引いている．

オ．平均値は中央値よりも小さく，分布は左に裾を引いている．

(1-2) あるテストを受験した集団における得点の標準偏差が 15 点であったとき，A さんと B さんの偏差値の差が 4 と算出された．A さんと B さんの得点の差はいくらか．次のア～オのうちから正しいものを一つ選べ．

ア．2

イ．4

ウ．6

エ．8

オ．10

(1-3) 標準偏差と標準誤差に関する記述として，次のア～オのうちから適切なものをすべて選んだ組合せを一つ選べ．

> (a) 標準偏差は，ある集団における個人間のデータのばらつきを意味する指標である．
> (b) 標準偏差は，分散から導くことはできない．
> (c) 標準誤差は，仮にある研究を反復して実施した際に観察されるであろう，平均などの統計量のばらつきを意味する指標である．
> (d) 標準誤差は，1 回の研究データから推定することは不可能である．

ア．(a), (b)

イ．(a), (c)

ウ．(b), (c)

エ．(b), (d)

オ．(c), (d)

(1-4) ある確率変数Xの確率関数が以下で与えられるとき，Xの分散はいくらか．次のア～オのうちから正しいものを一つ選べ．

	$x = 0$	$x = 0.5$
$\Pr(X = x)$	$1 - p$	p

ア．$\frac{1}{2}p$

イ．$\frac{1}{4}p$

ウ．$\frac{1}{4}p^2$

エ．$\frac{1}{4}p(1-p)$

オ．$p(1-p)$

【次頁に続く】

(1-5) ある連続型確率変数Xの確率密度関数が$f(x) = \frac{5}{8}(1 - x^4), -1 \leq x \leq 1$で与えられている．$Y = 2X$により得られる連続型確率変数$Y$の確率密度関数として，次のア～オのうちから適切なものを一つ選べ．

ア．$f(y) = \frac{5}{8}[1 - (2y)^4],\quad -1 \leq y \leq 1$

イ．$f(y) = \frac{5}{8}[1 - (2y)^4],\quad -2 \leq y \leq 2$

ウ．$f(y) = \frac{5}{8}\left[1 - (\frac{y}{2})^4\right],\quad -\frac{1}{2} \leq y \leq \frac{1}{2}$

エ．$f(y) = \frac{5}{8}\left[1 - (\frac{y}{2})^4\right],\quad -1 \leq y \leq 1$

オ．$f(y) = \frac{5}{16}\left[1 - (\frac{y}{2})^4\right],\quad -2 \leq y \leq 2$

(1-6) ある変数xの関数$\log(1 + x^2), -\infty < x < \infty$ のxに関する１階微分として，次のア～オのうちから正しいものを一つ選べ．

ア．$\log(1 + x^2)$

イ．$\exp(1 + x^2)$

ウ．$\frac{1}{\log(1+x^2)}$

エ．$\frac{2x}{\log(1+x^2)}$

オ．$\frac{2x}{1+x^2}$

(1-7) ある離散型確率変数XとYの同時確率関数の値が以下で与えられるとき，条件付き確率$\Pr(Y=3|X\geq 2)$はいくらか．次のア～オのうちから正しいものを一つ選べ．

	$y=1$	$y=2$	$y=3$
$x=1$	$\frac{1}{10}$	$\frac{3}{20}$	$\frac{3}{20}$
$x=2$	$\frac{1}{10}$	0	$\frac{1}{5}$
$x=3$	0	$\frac{1}{5}$	$\frac{1}{10}$

ア．$\frac{3}{10}$

イ．$\frac{9}{20}$

ウ．$\frac{2}{3}$

エ．$\frac{1}{2}$

オ．$\frac{1}{3}$

(1-8) ある確率変数Xは期待値μ，分散σ^2をもち，ある確率変数Yは期待値2μ，分散 $3\sigma^2$をもつ．XとYが独立のとき，$2X-3Y$の期待値$\mathrm{E}(2X-3Y)$との$2X-3Y$の分散$\mathrm{V}(2X-3Y)$はそれぞれいくらか．次のア～オのうちから正しいものを一つ選べ．

ア．$\mathrm{E}(2X-3Y)=-4\mu$,　　$\mathrm{V}(2X-3Y)=-\sigma^2$

イ．$\mathrm{E}(2X-3Y)=-\mu$,　　$\mathrm{V}(2X-3Y)=-7\sigma^2$

ウ．$\mathrm{E}(2X-3Y)=5\mu$,　　$\mathrm{V}(2X-3Y)=11\sigma^2$

エ．$\mathrm{E}(2X-3Y)=-4\mu$,　　$\mathrm{V}(2X-3Y)=31\sigma^2$

オ．$\mathrm{E}(2X-3Y)=8\mu$,　　$\mathrm{V}(2X-3Y)=-24\sigma^2$

(1-9) 10 個の異なる番号が書かれたくじをグループ A とグループ B に 5 個ずつにランダムに分ける．当たりくじが 3 個含まれているとき，2 グループ間の当たりくじの個数の差が 1 個以上となる分け方は全部で何通りあるか．次のア～オのうちから正しいものを一つ選べ．

ア．105

イ．210

ウ．126

エ．252

オ．42

(1-10) ある離散型確率変数XとYの同時確率関数の値が以下で与えられるとき，XとYの共分散はいくらか．次のア～オのうちから正しいものを一つ選べ．

	$y=1$	$y=2$	$y=3$	合計
$x=1$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{5}{12}$
$x=2$	0	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{4}$
$x=3$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{3}$
合計	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

ア．$\frac{1}{2}$

イ．$\frac{1}{3}$

ウ．$\frac{1}{4}$

エ．$\frac{1}{5}$

オ．$\frac{1}{6}$

(1-11) ある確率変数XとYの相関係数がρで与えられているとする．$2X$と$2Y$の相関係数はいくらか．次のア～オのうちから正しいものを一つ選べ．

ア．4ρ

イ．2ρ

ウ．$\sqrt{2}\rho$

エ．ρ

オ．$\frac{1}{2}\rho$

(1-12) ある前向き観察研究によって疾患の発生と曝露の関係を調べたところ，以下のようなデータが得られた．曝露なしに対する曝露ありの，疾患の発生率比はいくらか．次のア～オのうちから最も近いものを一つ選べ．

	曝露あり	曝露なし	合計
疾患あり	45 人	15 人	60 人
疾患なし	2,755 人	5,585 人	8,340 人
合計	2,800 人	5,600 人	8,400 人
観測人年	28,014 人年	13,970 人年	41,984 人年

ア．6.0

イ．4.5

ウ．3.0

エ．1.5

オ．1.0

【次頁に続く】

(1-13) ある 10 個の確率変数$X_1, \ldots, X_{10}$が独立に期待値 5，分散が 100 の正規分布に従っているとする．$X_1, \ldots, X_{10}$の平均$\frac{X_1+\cdots+X_{10}}{10}$はどのような確率分布に従うか．次のア～オのうちから正しいものを一つ選べ．

ア．期待値が 50，分散が 1000 の正規分布

イ．期待値が 5，　分散が 1000 の正規分布

ウ．期待値が 50，分散が 100 の正規分布

エ．期待値が 50，分散が 10 の正規分布

オ．期待値が 5，　分散が 10 の正規分布

(1-14) 2 群のランダム化比較試験において平均値の差の t 検定を行ったところ，平均値の差が 5，両側 p 値が 0.05 であった．平均値の差の両側 95%信頼区間として，次のア～オのうちから最も適切なものを一つ選べ．

ア．[0, 10]

イ．[1, 9]

ウ．[2, 8]

エ．[3, 7]

オ．[4, 6]

(1-15) 仮説検定に関する記述として，次のア～オのうちから最も適切なものを一つ選べ.

ア. 第一種の過誤確率とは，対立仮説が正しい場合に誤って帰無仮説を棄却してしまう確率を意味する.

イ. 仮説検定を繰り返しても，いずれかの正しい帰無仮説を誤って棄却してしまう確率は変化しない.

ウ. 第一種の過誤確率に対する名目水準は，サンプルサイズにより変化する.

エ. 第一種の過誤確率はその名目水準に一致するとは限らない.

オ. 第一種の過誤確率は実際に観察された p 値に一致する.

生物統計情報学 第2問

正の実数値をとる二つの独立な確率変数X, Yを考える．Xの周辺分布はパラメータ$\lambda > 0$の定める指数分布であると仮定し，その確率密度関数は

$$p_X(x) = \lambda e^{-\lambda x}, \quad x > 0$$

である．一方，Yの周辺分布は次の確率密度関数によって定まると仮定する：

$$p_Y(y) = c(\lambda) y e^{-\lambda y}, \quad y > 0.$$

ただし，$c(\lambda)$は正規化定数を表す．このとき，以下の問いに答えよ．

(2-1) Xの累積分布関数を求めよ．

(2-2) Xの期待値と中央値をそれぞれ求めて，その大小を比較せよ．

(2-3) $c(\lambda)$を求めよ．

(2-4) XがY以下である確率$\Pr(X \leq Y)$を求めよ．

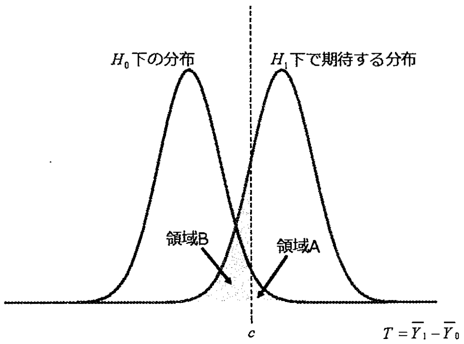
生物統計情報学 第3問

プラセボ群と新薬群からなるランダム化比較試験を計画している。この試験の主要変数は、治療開始から 6 ヶ月経過した時点における quality of life（QOL）スコアである。QOL スコアは高値であるほど、QOL が高い状態にある。以下の問いに答えよ。以降では標準正規分布の上側 100γ%点を$z_{1-\gamma}$と書く。

(3-1) 治療群i（プラセボ群：$i=0$、新薬群：$i=1$）に割り付けられたj番目（$j=1,\dots,n_i$）の対象者の QOL スコアを$Y_{ij}\sim \mathrm{N}(\mu_i,\sigma_i^2)$とする。治療群$i$の平均値$\bar{Y}_i=n_i^{-1}\sum_{j=1}^{n_i}Y_{ij}$がしたがう分布を答えよ。

(3-2) QOL スコアの平均値の群間比較に興味がある。帰無仮説$H_0\colon \mu_0=\mu_1$、対立仮説$H_1\colon \mu_0<\mu_1$とする。分散σ_i^2が既知のとき、標準化統計量
$Z=(\bar{Y}_1-\bar{Y}_0)\sqrt{\left(\frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1}\right)^{-1}}$を用いて片側有意水準$\alpha$の仮説検定を行う手順を説明せよ。

(3-3) (3-2)の検定を行う状況を考える。検出力を$1-\beta$とするために必要な解析対象者数の導出過程を次に示す。空欄①～④に入る数式を求めよ。ただし、$n_0=n_1=n$、分散σ_i^2は群間で共通の値σ^2で既知とする。

> 新薬群とプラセボ群の平均値の差$\mu_1-\mu_0$を$\bar{Y}_1-\bar{Y}_0$で推定する。H_1で$\mu_1-\mu_0$に期待する値（正の値）をΔとする。$\frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1}=\frac{2\sigma^2}{n}=\sigma_*^2$とおくと、$T=\bar{Y}_1-\bar{Y}_0$は$H_0$の下（つまり、$\mu_1-\mu_0=0$の下）で$\boxed{①}$、$H_1$における$\mu_1-\mu_0=\Delta$の下で$\boxed{②}$にしたがう。①と②で導出した$T$の確率密度関数はそれぞれ右下図のようになる。
> いま、$T>c$を満たす場合に統計学的に有意とする検定方式を考える。右図の領域 A の面積がαの場合に第 1 種の過誤確率がαになる。この関係から$c=\sigma_* z_{1-\alpha}$が導かれる。同様にして右図の領域 B がβの場合に検出力が$1-\beta$になるので、$c=\boxed{③}$が導かれる。棄却域cに関する上述の 2 つの等式をnについて解くと、必要解析対象者数は各群で$n=\boxed{④}$と導出される。
>
> 
>