２０２０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
（２０１９年８月１９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，全員が解答すること．第２問及び第３問は記述式問題であり，この２問の中から１問を選択して解答すること．
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２枚ある．解答した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となる．
6. 解答用紙の上方の欄に，解答した問題の番号及び受験番号を必ず記入すること．問題番号及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，メモ用紙，及び分布表が印刷された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 (必須問題)

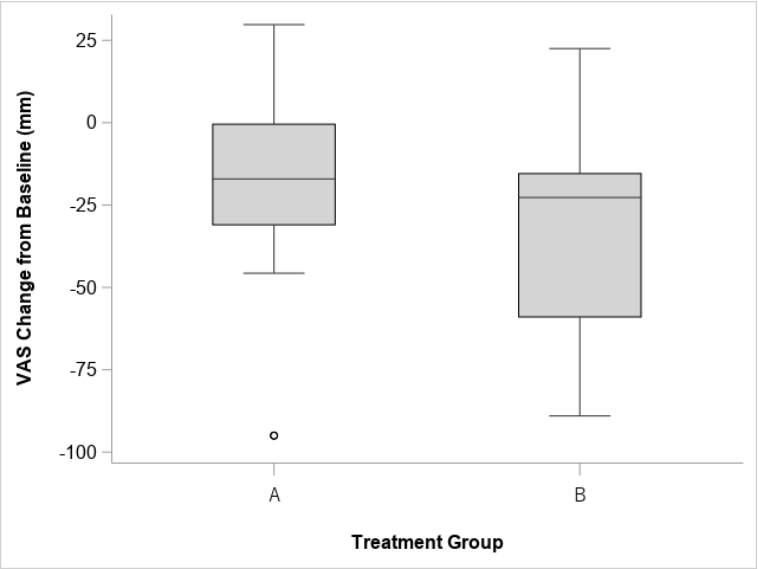
以下の問い **(1-1)**～**(1-20)** に答えよ．解答用紙には問いの番号と解答のみを，問いの番号の順序に従って記載せよ．なお，別途配布した分布表は適宜参照してよい．

(1-1) 次の幹葉図は，あるクラスの数学のテストの得点の分布を示している．このクラスの数学のテストの得点の中央値はいくらか．以下のア～オのうちから正しいものを一つ選べ．

度数	幹	葉
1	3	3
2	4	5 7
9	5	2 3 4 4 5 5 5 7 8
7	6	1 2 2 3 5 5 5
3	7	3 6 7
2	8	2 8
1	9	4

ア．60 点　　イ．61 点　　ウ．62 点　　エ．63 点　　オ．64 点

(1-2) ある疾患の症状を軽減する薬剤の有効性を比較するためのランダム化 2 群比較試験を行った．各患者の症状は Visual Analog Scale (VAS) によりベースライン及び治療後の 2 時点で測定され，次の図が得られた．



各群の VAS 変化量に関する記述として次の (a)〜(d) のうちから正しいものをすべて選んだ組み合わせを，以下のア〜オのうちから一つ選べ．

(a) A 群の分布は右に歪んでいる．

(b) B 群の分布は左に歪んでいる．

(c) A 群の第三四分位点は B 群の第三四分位点よりも小さい．

(d) A 群の四分位範囲は B 群の四分位範囲よりも広い．

(e) A 群の範囲は B 群の範囲よりも広い．

ア．(a), (c)　イ．(a), (d)　ウ．(b), (c)　エ．(b), (d)　オ．(b), (e)

(1-3) 確率変数 X の平均と標準偏差はそれぞれ 2, 4 で与えられているとする．確率変数 Y が $Y = 2X^2$ と定義されているとき，確率変数 Y の平均はいくらか．次のア〜オのうちから正しいものを一つ選べ．

ア．8　イ．16　ウ．32　エ．40　オ．80

【次頁に続く】

(1-4) ある母集団における身長 X の分布が，平均 160 cm，標準偏差 10 cm の正規分布であるとする．この集団からランダムに 1 人選ぶとき，その人の身長が 170 cm 以上である確率はいくらか．次のア～オのうちから最も近い値を一つ選べ．

ア．0.025　　イ．0.16　　ウ．0.32　　エ．0.5　　オ．0.84

(1-5) 二つの確率変数 X_1, Y_1 の相関係数を ρ と表す．新たに確率変数 $X_2 = X_1/3$ と $Y_2 = Y_1 + 5$ を考えるとき，X_2 と Y_2 の相関係数として正しいものを次のア～オのうちから一つ選べ．

ア．$\dfrac{\rho}{3}$　　イ．$\dfrac{\rho}{\sqrt{3}}$　　ウ．ρ　　エ．$\sqrt{3}\rho$　　オ．3ρ

(1-6) ある疾患の症状スコア X が，$1, 2, 3$ と離散的な値をとり，確率分布が次のように与えられているとする：

$$\Pr(X = 1) = 0.2, \quad \Pr(X = 2) = 0.5, \quad \Pr(X = 3) = 0.3.$$

100 人の患者の症状スコアが独立に X の確率分布に従うとき，症状スコアの平均値が 2 以下となる確率はいくらか．次のア～オのうちから最も近い値を一つ選べ．

ア．0.1　　イ．0.2　　ウ．0.3　　エ．0.4　　オ．0.5

(1-7) 4 例の患者を割り付け比率 1 : 1 で単純ランダム化により 2 群へ割り付けたとき，群間で人数が等しくならない割り付け結果が得られる確率はいくらか．次のア～オのうちから正しいものを一つ選べ．

ア．25%　　イ．37.5%　　ウ．50%　　エ．62.5%　　オ．75%

(1-8) あるアンケート調査において，各対象者の回答確率を 0.99 と想定したとき，すべての対象者が回答する確率 0.5 と算出された．このアンケート調査の対象者数はいくらか．次のア～オのうちから最も近い値を一つ選べ．ただし，$p \approx 1$ のとき，$\ln p \approx p - 1$，$\ln 0.5 \approx -0.69$ とする．

ア．50　　イ．70　　ウ．90　　エ．110　　オ．130

(1-9) ある年の全国の都道府県別の観光客数を調査したところ，平均は 50 万人，標準偏差は 50 万人であった．翌年の都道府県ごとの観光客数が 2 倍に増加したとき，変動係数は何倍になるか．次のア～オのうちから正しいものを一つ選べ．

ア．$\frac{1}{4}$ 倍　　イ．$\frac{1}{2}$ 倍　　ウ．1 倍　　エ．2 倍　　オ．4 倍

(1-10) 確率変数 X と Y の同時確率分布が次の表で与えられている．

	$Y = 1$	$Y = 0$
$X = 1$	p_{11}	p_{10}
$X = 0$	p_{01}	p_{00}

確率変数 X と Y に関する記述として，次の (a)～(e) のうちから適切なものをすべて選んだ組み合わせを，以下のア～オのうちから一つ選べ．

(a) $X = 1$ かつ $Y = 0$ となる同時確率は p_{01} である．

(b) $X = 0$ となる周辺確率は $p_{01} + p_{00}$ である．

(c) $Y = 0$ となる周辺確率は $p_{10} + p_{01} + p_{00}$ である．

(d) $X = 1$ を与えた下で $Y = 1$ となる条件付き確率は $\frac{p_{11}}{p_{11} + p_{10}}$ である．

(e) $X = 1$ を与えた下で $Y = 1$ となる条件付き確率は $\frac{p_{11}}{p_{10}}$ である．

ア．(a), (c)　　イ．(b), (d)　　ウ．(b), (e)　　エ．(c), (d)　　オ．(c), (e)

【次頁に続く】

(1-11) 第一種の過誤確率が 5%の独立な検定を 2 回繰り返した．帰無仮説の下で少なくとも一つの検定で有意になる確率はいくらか．次のア～オのうちから正しいものを一つ選べ．

ア．0.0025　　イ．0.025　　ウ．0.05　　エ．0.1　　オ．0.0975

(1-12) 正規近似に基づく信頼区間を算出する際に，信頼係数を両側 95%から 90%に変更すると，信頼区間の幅は何倍になるか．次のア～オのうちから最も近い値を一つ選べ．

ア．0.7 倍　　イ．0.75 倍　　ウ．0.8 倍　　エ．0.85 倍　　オ．0.9 倍

(1-13) ある薬剤の効果を調べるために，実薬，プラセボの順に投与される群とプラセボ，実薬の順に投与される群へ対象者をランダムに割り付ける臨床試験を行い，実薬とプラセボで症状の有無を比較した．

		プラセボ投与		合計
		症状あり	症状なし	
実薬投与	症状あり	A_1	B_1	N_1
	症状なし	A_0	B_0	N_0
	合計	M_A	M_B	N

薬剤の効果を調べる検定の帰無仮説として，次の (a)～(d) のうちから正しいものをすべて選んだ組み合わせを以下のア～オのうちから一つ選べ．

(a) $A_1 = B_1$　　(b) $A_1 = B_0$　　(c) $B_1 = A_0$　　(d) $\dfrac{N_1}{N} = \dfrac{M_A}{N}$

ア．(a)　　イ．(a), (b)　　ウ．(c)　　エ．(c), (d)　　オ．(d)

(1-14) ある疾患のスクリーニング検査を 5,000 人の地域住民に対して行った．この検査の感度と特異度がそれぞれいずれも 0.9 とし，この疾患を有する患者が 40 人いた場合に，検査で陽性となる人は何人となることが期待されるか．次のア～オのうちから最も近い値を一つ選べ．

ア．35 人　　イ．50 人　　ウ．100 人　　エ．500 人　　オ．535 人

(1-15) ある希少肺疾患患者 500 例の肺がん発症確率を調べたところ，20 例が肺がんを合併していることがわかった．同様の年齢，性別，喫煙歴をもつ一般集団における肺がんの発症確率が 0.1%であった．一般集団と比較し，この希少肺疾患患者の肺がん発症確率が高いかどうかの検定を行う際の計算式について，次の (a)〜(e) のうちから適切な組み合わせを，以下のア〜オのうちから一つ選べ．

(a) 検定統計量の分子は $\dfrac{20}{500}$ である．

(b) 検定統計量の分子は $\dfrac{20}{500} - 0.001$ である．

(c) 検定統計量の分母は $\sqrt{500\dfrac{20}{500}\left(1 - \dfrac{20}{500}\right)}$ である．

(d) 検定統計量の分母は $\sqrt{\dfrac{20}{500}\left(1 - \dfrac{20}{500}\right)}$ である．

(e) 検定統計量の分母は $\sqrt{\dfrac{0.001}{500}(1 - 0.001)}$ である．

ア．(a), (c)　　イ．(a), (d)　　ウ．(b), (c)　　エ．(b), (d)　　オ．(b), (e)

(1-16) ある疾患の発症の有無に対し，新規治療と標準治療の有効性を比較するランダム化臨床試験を計画した．治療間の有効性の差を評価するための仮説検定の検出力の計算に必要になる要素として適切ではないものを一つ選べ．

ア．症例数

イ．登録期間

ウ．有意水準

エ．片方の群の疾患発生割合

オ．疾患発生割合の群間差

(1-17) 1 年目の累積生存確率が 60%，2 年目の累積生存確率が 36%であった．1 年間あたりの死亡確率は何%と推定されるか．次のア〜オのうちから正しいものを一つ選べ．

ア．24%　　イ．32%　　ウ．40%　　エ．48%　　オ．52%

【次頁に続く】

(1-18) ワイブル分布に従う生存時間 T の累積分布関数 $F(t)$ が次で与えられているとする：

$$F(t) = 1 - \exp\left[-(\lambda t)^{\gamma}\right].$$

生存時間 T の確率密度関数として正しいものを次のア～オのうちから一つ選べ.

ア. $f(t) = \lambda\gamma(\lambda t)^{\gamma} \exp[-(\lambda t)^{\gamma-1}]$

イ. $f(t) = \gamma(\lambda t)^{\gamma-1} \exp\left[-(\lambda t)^{\gamma}\right]$

ウ. $f(t) = \lambda\gamma(\lambda t)^{\gamma-1} \exp\left[-(\lambda t)^{\gamma}\right]$

エ. $f(t) = \gamma \exp\left[-(\lambda t)^{\gamma}\right]$

オ. $f(t) = \lambda\gamma \exp[-(\lambda t)^{\gamma-1}]$

(1-19) ある疾患発症の有無に対する環境汚染物質への曝露の影響を調べる研究を行い，次の分割表が得られた.

	疾患発症あり	疾患発症なし	合計
曝露あり	A_1	B_1	N_1
曝露なし	A_0	B_0	N_0
合計	M_A	M_B	N

オッズ比 $\widehat{\mathrm{OR}}$ の 95%信頼区間として，最も近いものを以下のア～オのうちから一つ選べ．ただし，対数オッズ比 $\ln\widehat{\mathrm{OR}}$ の分布は正規近似ができ，対数オッズ比の分散 $\mathrm{Var}[\ln\widehat{\mathrm{OR}}]$ は次の式で計算できるものとする：

$$\mathrm{Var}[\ln\widehat{\mathrm{OR}}] = \frac{1}{A_1} + \frac{1}{B_1} + \frac{1}{A_0} + \frac{1}{B_0}.$$

ア. $\left[\widehat{\mathrm{OR}} - \left(\frac{1}{A_1} + \frac{1}{B_1} + \frac{1}{A_0} + \frac{1}{B_0}\right), \widehat{\mathrm{OR}} + \left(\frac{1}{A_1} + \frac{1}{B_1} + \frac{1}{A_0} + \frac{1}{B_0}\right)\right]$

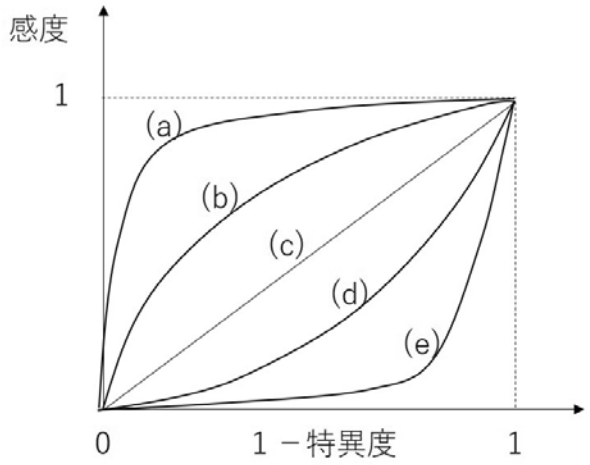
イ. $\left[\widehat{\mathrm{OR}} - 1.96\left(\frac{1}{A_1} + \frac{1}{B_1} + \frac{1}{A_0} + \frac{1}{B_0}\right), \widehat{\mathrm{OR}} + 1.96\left(\frac{1}{A_1} + \frac{1}{B_1} + \frac{1}{A_0} + \frac{1}{B_0}\right)\right]$

ウ. $\left[\exp\left[\widehat{\mathrm{OR}} - 1.96\sqrt{\left(\frac{1}{A_1} + \frac{1}{B_1} + \frac{1}{A_0} + \frac{1}{B_0}\right)}\right], \exp\left[\widehat{\mathrm{OR}} + 1.96\sqrt{\left(\frac{1}{A_1} + \frac{1}{B_1} + \frac{1}{A_0} + \frac{1}{B_0}\right)}\right]\right]$

エ. $\left[\exp\left[\ln(\widehat{\mathrm{OR}}) - 1.96\sqrt{\left(\frac{1}{A_1} + \frac{1}{B_1} + \frac{1}{A_0} + \frac{1}{B_0}\right)}\right], \exp\left[\ln(\widehat{\mathrm{OR}}) + 1.96\sqrt{\left(\frac{1}{A_1} + \frac{1}{B_1} + \frac{1}{A_0} + \frac{1}{B_0}\right)}\right]\right]$

オ. $\left[\exp\left[\ln(\widehat{\mathrm{OR}}) \div 1.96\left(\frac{1}{A_1} + \frac{1}{B_1} + \frac{1}{A_0} + \frac{1}{B_0}\right)\right], \exp\left[\ln(\widehat{\mathrm{OR}}) \times 1.96\left(\frac{1}{A_1} + \frac{1}{B_1} + \frac{1}{A_0} + \frac{1}{B_0}\right)\right]\right]$

(1-20) ある疾患の診断を補助するための検査薬の性能を考える．検査薬の検査結果は連続量で与えられ，検査結果の値があるカットオフ値を超えると陽性，カットオフ値以下であれば陰性とする．検査性能の異なる 5 つの検査薬各々について，カットオフ値を変化させ，縦軸に検査の特異度，横軸に検査の 1 – 感度 をプロットしたとき，次の図の (a)〜(e) のような曲線が得られた．検査性能が最も高いと考えられる検査薬を表す曲線はどれか．以下のア〜オのうちから正しいものを一つ選べ．



ア．(a)　　イ．(b)　　ウ．(c)　　エ．(d)　　オ．(e)

生物統計情報学 第2問 (選択問題)

パラメータ $\gamma > 0$ の定める確率分布 P_γ の確率密度関数が次のように与えられている：

$$f(x;\gamma) = \begin{cases} \frac{1}{\gamma} e^{-x/\gamma} & (x > 0) \\ 0 & (\text{otherwise}) \end{cases}$$

確率変数 $X_1, \ldots, X_n$ が P_γ に独立に従うと仮定し，$X_1, \ldots, X_n$ を昇順に並べ替えたものを $X_{(1)}, \ldots, X_{(n)}$ と表す．このとき，以下の問いに答えよ．

(2-1) 確率変数 X_1 の期待値 $E[X_1]$ を求めよ．

(2-2) 確率変数 $X_{(1)}$ の確率密度関数 $f_{(1)}(x;\gamma)$ と期待値 $E[X_{(1)}]$ を求めよ．

(2-3) 確率変数 X_1 および $X_{(1)}$ の歪度をそれぞれ求めよ．ただし，確率変数 Y の期待値と分散が μ_Y および σ_Y^2 であるとき，Y の歪度は $E[(Y-\mu_Y)^3]/\sigma_Y^3$ で与えられる．

(2-4) 標本 $X_1, \ldots, X_n$ に基づく統計量として，$X_1, \ldots, X_n$ から $X_{(1)}$ を除いたものの平均 $T_n = \sum_{i=2}^{n} X_{(i)}/(n-1)$ を考える．統計量 T_n をパラメータ γ の推定量とみなしたときのバイアス $E[T_n] - \gamma$ を求めよ．

生物統計情報学 第3問 (選択問題)

ある会社の定期健診において，収縮期血圧が繰り返し測定されている．健診対象者全員を測定した場合の 1 回目の収縮期血圧を X_1，2 回目の収縮期血圧を X_2 としたとき，以下の問いに答えよ．

(3-1) 収縮期血圧の測定値 X_1, X_2 が，それぞれ $X_1 \sim N(\mu_1, \sigma^2)$，$X_2 \sim N(\mu_2, \sigma_2^2)$ であり，平均 $\mu_1 = \mu_2$，分散 $\sigma_1^2 = \sigma_2^2$，共分散 σ_{12} の二変量正規分布に従うとする．X_2 を応答変数，X_1 を説明変数とした場合の母回帰直線の切片，および傾きを示せ．

(3-2) **(3-1)** において示した傾きは，変数間の相関係数 ρ が 1 でない限り，1 にはならない．この理由を，相関係数 ρ を用いて説明せよ．

(3-3) 次の表は，実際に観察された収縮期血圧の平均と標準偏差である．ただし，この会社では，定期健診において収縮期血圧が 140 mmHg 以上であった人に対してのみ，1ヶ月後に 2 回目の血圧測定を行うことにしている．

	1 回目平均 (標準偏差)	2 回目平均 (標準偏差)
健診対象者全体	120 mmHg (20 mmHg)	
1 回目 140 mmHg 以上のグループ	150 mmHg (9 mmHg)	141 mmHg (16 mmHg)

この情報をもとに，初回の収縮期血圧値が 140 mmHg であった人の 2 回目の収縮期血圧の期待値を算出せよ．ただし，**(3-1)** と同様に，1 回目と 2 回目の測定値は同一の分布に従うものとする．

(3-4) 後日，健診の現場では，1 回目の測定値が 140 mmHg 以上のグループに対して，食生活の改善を含む保健指導が実施されていたことが明らかとなった．保健指導を実施した保健師は，**(3-3)** の表から，「収縮期血圧 140 mmHg 以上の集団に対して保健指導を実施した結果，血圧を平均 9 mmHg 下げる効果が確認できた」と主張している．この主張は正しいか．誤っている場合には，その理由と，どのような方法をとればこのような保健指導の効果を確認することができるかについて，あなたの考えを述べよ．