7 1　【データサイエンス学部】

総　合　問　題

2024（令和 6 ）年度

【注意事項】

1. 試験時間は 90 分である。
2. 試験開始の合図まで，この問題冊子を開いてはいけない。ただし，表紙はあらかじめよく読んでおくこと。
3. この問題冊子の印刷は 1 ページから 12 ページまである。
4. 解答用紙は問題冊子中央に 2 枚はさみこんである。
5. 問題冊子に落丁，乱丁，印刷不鮮明な箇所等があった場合および解答用紙が不足している場合は，手をあげて監督者に申し出ること。
6. 試験開始後，2 枚ある解答用紙の所定の欄に，受験番号と氏名を記入すること（1 枚につき受験番号は 2 箇所，氏名は 1 箇所）。
7. 解答は必ず解答用紙の指定された箇所に記入すること。解答用紙の裏面に記入してはいけない。
8. 問題番号に対応した解答用紙に解答していない場合は，採点されない場合もあるので注意すること。
9. 解答用紙を切り離したり，持ち帰ってはいけない。
10. 問題冊子の中の白紙部分は下書き等に使用してよい。
11. 試験終了時刻まで退室を認めない。試験中の気分不快やトイレ等，やむを得ない場合には，手をあげて監督者を呼び，指示に従うこと。
12. 試験終了後は問題冊子を持ち帰ること。

〔Ⅰ〕 人が住んでいない住宅である「空き家」に関する統計データについて，以下の問いに答えなさい。

（1） 表1は，総務省の「住宅・土地統計調査(2018年)」に基づき，住宅数・空き家数(およびその内訳)・空き家率・その他の住宅の空き家率を，日本全国について整理したものである。なお，小数点以下第1位を四捨五入している関係で，空き家数の内訳の数値を合計しても空き家数の数値と一致していない場合がある。

ここで，人が住んでいない住宅である**空き家**は次の3種類に分類される。

- **二次的住宅**：ふだんは人が住んでいない別荘などの住宅。
- **賃貸用または売却用の住宅**：賃貸または売却のために空き家となっている住宅。
- **その他の住宅**：上記以外の人が住んでいない住宅。例えば，転勤・入院などのため居住世帯が長期にわたって不在の住宅や，建て替えなどのために取り壊すことになっている住宅など。

また，次のように，空き家率・その他の住宅の空き家率を定義する。

- **空き家率**：全ての住宅数に対して空き家数が占める割合として，「空き家率(%) ＝ 100 × 空き家数／住宅数」と定義する。
- **その他の住宅の空き家率**：全ての住宅数に対してその他の住宅数が占める割合として，「その他の住宅の空き家率(%) ＝ 100 × その他の住宅数／住宅数」と定義する。

表1：住宅数・空き家数(およびその内訳)・空き家率・その他の住宅の空き家率の推移(全国)

年次		1988	1993	1998	2003	2008	2013	2018
住宅数(万戸)		4201	4588	5025	5389	5759	6063	－
空き家数(万戸)		394	448	576	659	757	820	849
(空き家数の内訳)	二次的住宅数(万戸)	30	37	42	50	41	41	38
	賃貸用または売却用の住宅数(万戸)	234	262	352	398	448	460	462
	その他の住宅数(万戸)	131	149	182	212	268	318	349
空き家率(%)		9.4	9.8	11.5	a	13.1	13.5	13.6
その他の住宅の空き家率(%)		3.1	3.2	3.6	3.9	4.7	5.3	b

（ア） 表1の空欄a・bに入る数値を計算しなさい。ただし，数値に小数点以下の値が出たときは，小数点以下第2位を四捨五入して小数点以下第1位まで示しなさい。なお，2018年の住宅数については表示を省略している。

（イ） 変化率は，基準時点の値を分母とし，基準時点から比較時点までの増減分を分子とした比率であり，数値がAからBに変化したときの変化率は，$(B-A)/A$と定義される。次の文章は，表1について記述したものである。文章中の空欄 c ・ d に入る数値を答えなさい。ただし，数値に小数点以下の値が出たときは，小数点以下第1位を四捨五入して整数値で答えなさい。

> 1988年から2018年にかけて，空き家数は c %増加し，その他の住宅数は d %増加した。

（ウ） t年における空き家数をN_tと表す。2013年から2018年にかけての空き家数の変化率が，2008年から2013年にかけての空き家数の変化率$(N_{2013}-N_{2008})/N_{2008}$に一致すると仮定し，2008年，2013年の空き家数N_{2008}，N_{2013}をもとに，2018年の空き家数N_{2018}を予測することを考える。このとき，以下の問いに答えなさい。

(a) 2018年の空き家数N_{2018}の予測式を，N_{2008}，N_{2013}の記号を用いて表現しなさい。

(b) (a)で求めた予測式と表1の値を用いて，2018年の空き家数の予測値を「万戸」を単位として求めなさい。ただし，数値に小数点以下の値が出たときは，小数点以下第1位を四捨五入して整数値で答えなさい。

(c) (b)で求めた予測値と，表1における実際の観測値との大小関係を比較しなさい。また，このような大小関係となる数学的理由について，空き家数の変化率の観点から50字程度で考察しなさい。

（2） 総務省の「住宅・土地統計調査(2018年)」をもとに，図1は都道府県別の空き家率・その他の住宅の空き家率を，図2は都道府県別の空き家数の内訳(二次的住宅，賃貸用または売却用の住宅，その他の住宅の構成比)を，図3は都道府県別の住宅数を示したものである。

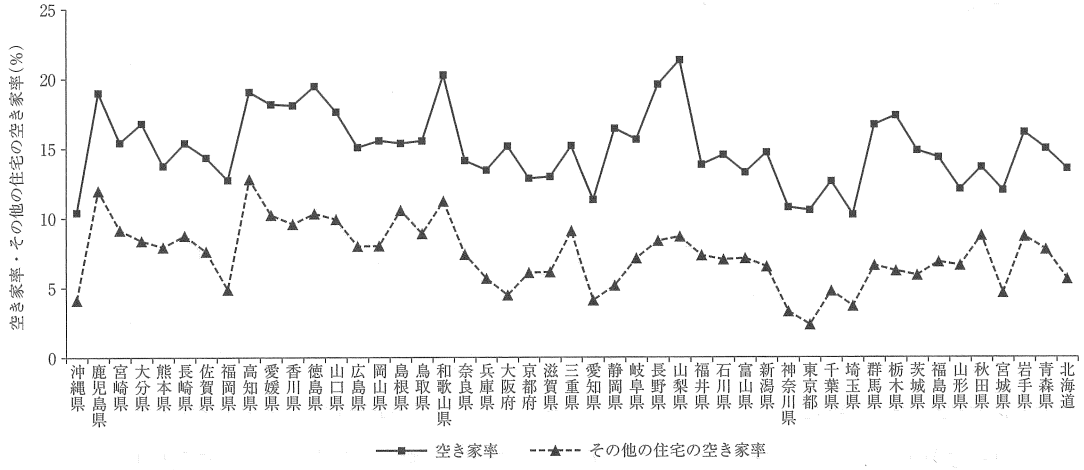

図1：都道府県別の空き家率・その他の住宅の空き家率

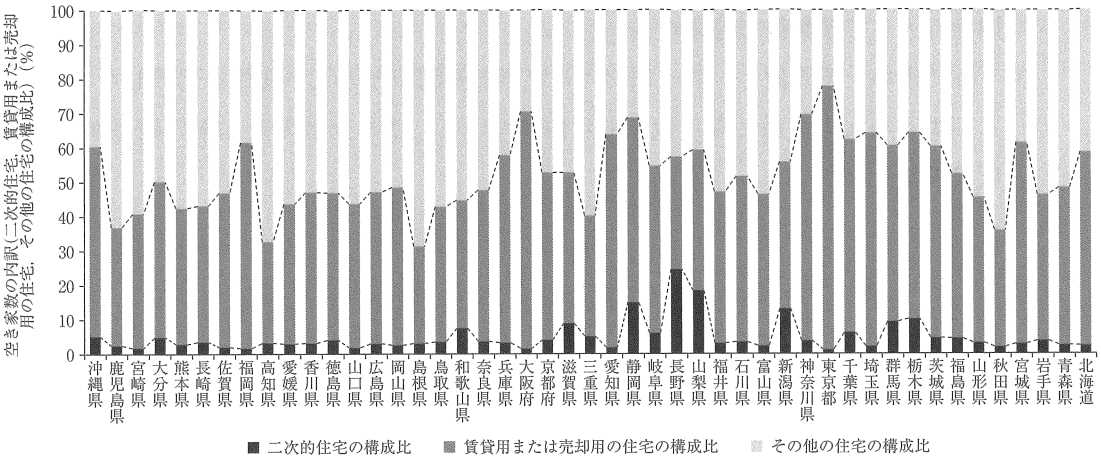

図2：都道府県別の空き家数の内訳(二次的住宅，賃貸用または売却用の住宅，その他の住宅の構成比)

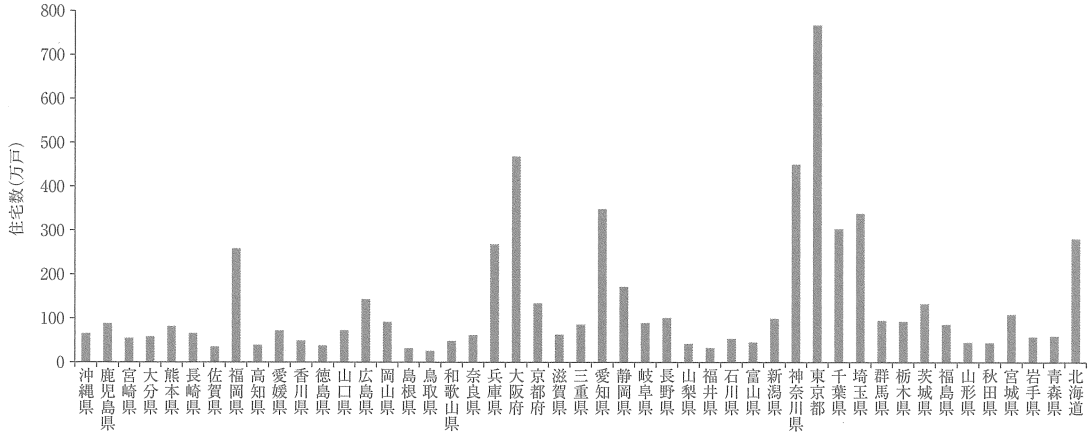

図３：都道府県別の住宅数

（ア） 図１を見ると，空き家率が高い都道府県として山梨県が含まれている。この背景として考えられることを，図２をもとに30字程度で考察しなさい。

（イ） 次の文章のうち，図１～３から読み取れる内容を全て選び，①～③の記号で答えなさい。

① 神奈川県では，山口県に比べ，その他の住宅の空き家率は小さい。
② 大阪府では，賃貸用または売却用の住宅が，空き家の過半数を占めている。
③ 東京都では，高知県に比べ，その他の住宅数は小さい。

（3） 図4は，総務省の「住宅・土地統計調査(2018年)」，総務省の「人口推計(2018年10月1日現在)」をもとに，高齢化率と空き家率との関係を示したものである。各プロットは47都道府県を表しており，特に秋田県を「■」で表している。

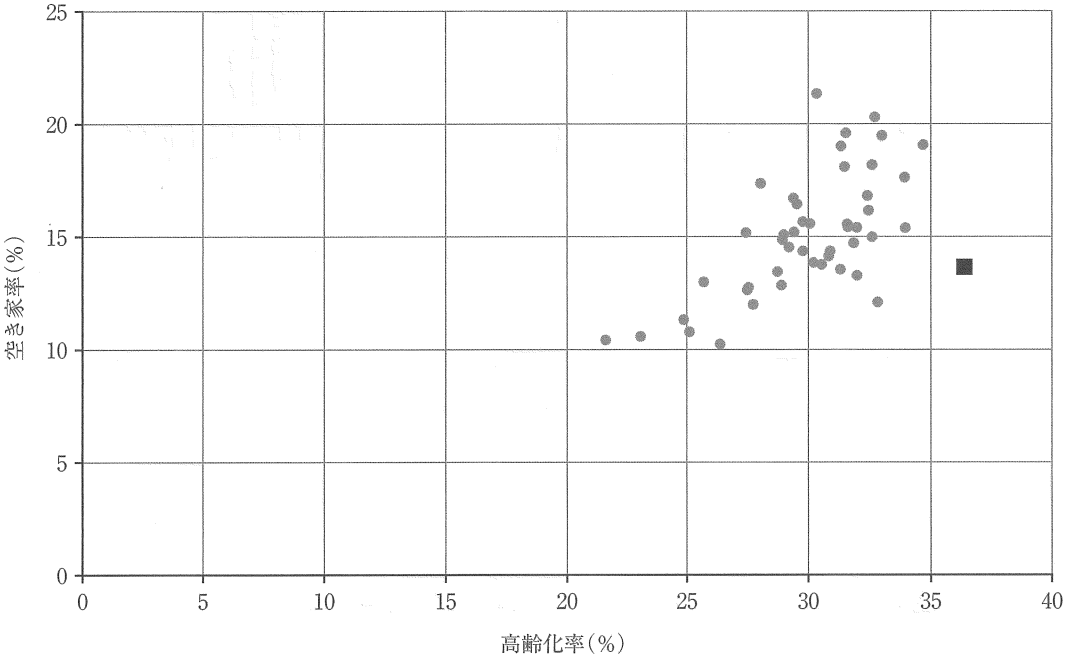

図4：高齢化率と空き家率との関係(都道府県別，2018年)

ここで，高齢化率 x と空き家率 y の相関係数 r_{xy} は，次の式(＊)で表される。ただし，$i=1$，…，n の都道府県の高齢化率と空き家率をそれぞれ x_i，y_i とし，x，y の平均値はそれぞれ $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ である。

$$r_{xy}=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}} \qquad (*)$$

（ア） 図4の都道府県別の高齢化率と空き家率との関係について，47都道府県を対象とした相関係数は0.605であった。ここで，図5のように，$(\overline{x}, \overline{y})$ を原点とし，横軸を x，縦軸を y とした4象限において，各都道府県の高齢化率 x_i と空き家率 y_i をプロットすることを考える。

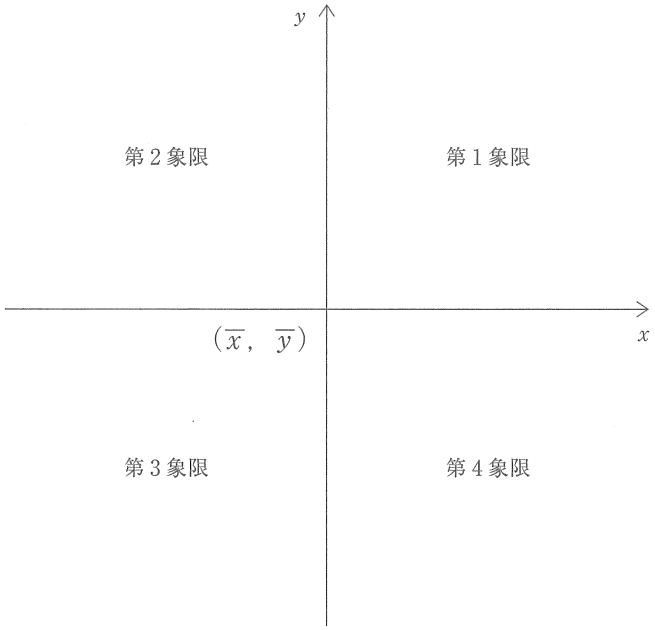

図５：$(\overline{x},\ \overline{y})$を原点とした４象限

このとき，次の文章中の空欄 [a] ・ [b] ・ [c] に入る語句として適切なものを選択肢から選び，記号で答えなさい。

相関係数は [a] の値をとる指標であり，図４では相関係数が0.605であることから，都道府県別の高齢化率と空き家率との間には [b] の相関がみられるといえる。

図４に示すデータについて，第１象限と第３象限における$(x_i - \overline{x})(y_i - \overline{y})$の和は，第２象限と第４象限における$(x_i - \overline{x})(y_i - \overline{y})$の和と比べて [c] 。

［空欄ａの解答選択肢］

① $-\infty$ から $+\infty$　　② -1 から 1

③ 0 から 1　　④ 0 から $+\infty$

［空欄ｂの解答選択肢］

① 正　　② 負

［空欄ｃの解答選択肢］

① 大きい　　② 小さい

（イ） 秋田県の高齢化率は 36.4%，空き家率は 13.6%である。また，47 都道府県について，高齢化率の平均は 30.1%，空き家率の平均は 15.0%である。このとき，次の文章中の空欄 [d] ・ [f] に入る語句として適切なものを選択肢から選び，記号で答えなさい。また，空欄 [e] に入る数値を答えなさい。ただし，数値に小数点以下の値が出たときは，小数点以下第 1 位を四捨五入して整数値で答えなさい。

> 秋田県は図 5 において第 [d] 象限に位置しており，秋田県について $(x_i - \overline{x})(y_i - \overline{y})$ の値を計算すると [e] $(\%)^2$ である。
>
> 式（1）の分母に登場する $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ の値について，47 都道府県について計算した値に比べ，秋田県を除いた 46 都道府県について計算した値は [f] なる。

［空欄 d の解答選択肢］

① 1　② 2　③ 3　④ 4

［空欄 f の解答選択肢］

① 大きく　② 小さく

（ウ） 仮に，全ての都道府県の高齢化率が 3 %ずつ等しく増加し，空き家率が 1 %ずつ等しく増加したとする。このとき，相関係数がどのように変化するか，式（*）中の表現を用いながら 100 字程度で考察しなさい。

〔Ⅱ〕 ある工場では，製品Aを作成している。工場長は，作成した製品の中に不良品が混在していることに悩まされていた。人手で各製品を検査すれば，良品と不良品を高精度に分類することができるが，人件費と効率の面から現実的ではない。そこで，容易に取得できるセンサ値を用い，製品Aの良品と不良品を自動で分類することを検討している。このとき，以下の問いに答えなさい。

（1） まずは，ランダムに選んだ製品25個に対し，2種類のセンサ値x，yおよび人手による検査結果をデータとして収集した。25個の製品に1～25までの製品番号を割り当て，A_1，...，A_{25}と記す。また，A_1，...，A_{25}の各センサ値を$s_1 = (x_1, y_1)$，...，$s_{25} = (x_{25}, y_{25})$と記す。$s_1$，...，$s_{25}$を$x$，$y$に関する二次元座標上のデータ点とみなし，散布図として描写したものが図1である。なお，白い点は良品を表し，黒い点は不良品を表している。また，検査結果およびセンサ値を表にしたものが表1である。以下の問い(ア)，(イ)に答えなさい。

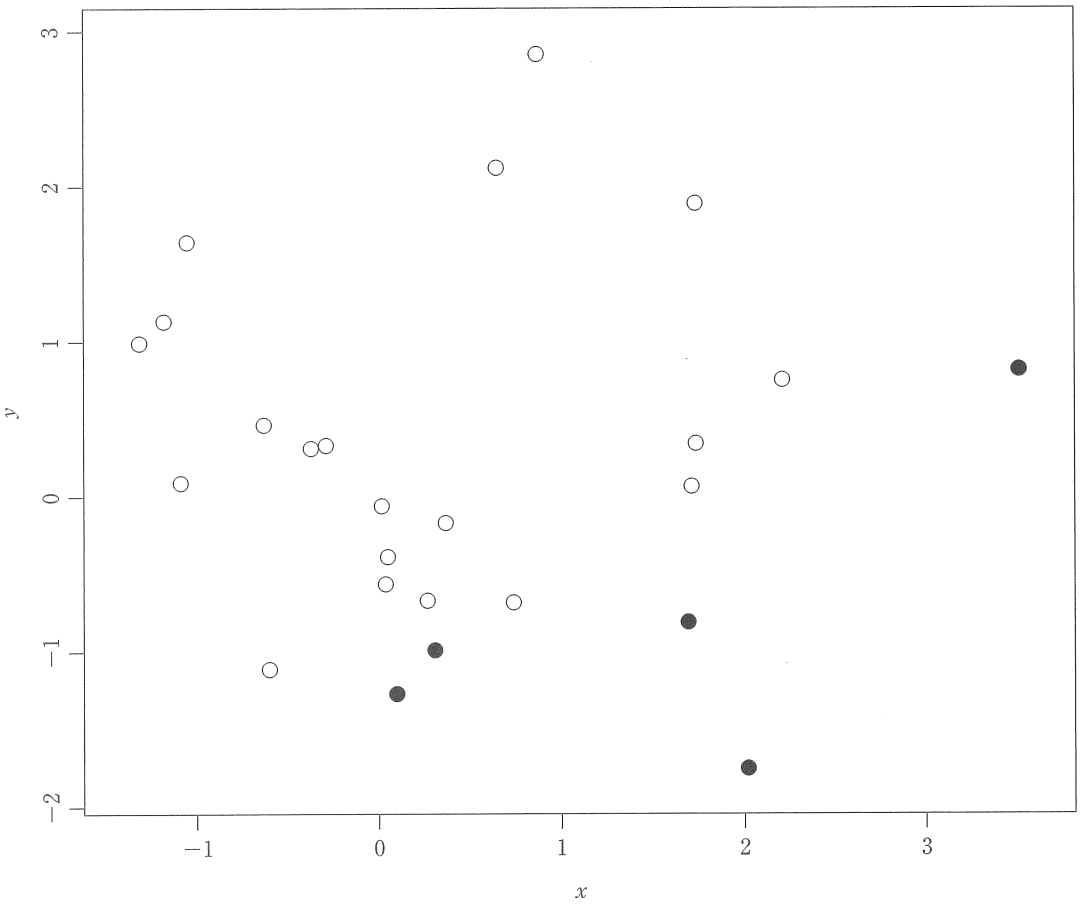

図1：製品A_1，...，A_{25}のセンサ値

表1：各製品の検査結果およびセンサ値

製品番号	検査結果	x	y
1	良品	0.27	−0.67
2	良品	−0.63	0.46
3	良品	0.87	2.85
4	良品	1.73	1.89
5	良品	0.02	−0.06
6	良品	0.37	−0.17
7	良品	−1.31	0.99
8	良品	0.74	−0.68
9	良品	0.04	−0.56
10	良品	−1.05	1.64
11	良品	1.73	0.34
12	良品	−1.18	1.13
13	良品	0.65	2.12
14	良品	−0.37	0.31
15	良品	−0.60	−1.11
16	良品	0.05	−0.39
17	良品	1.71	0.07
18	良品	−1.09	0.08
19	良品	−0.29	0.33
20	良品	2.21	0.75
21	不良品	2.02	−1.75
22	不良品	0.10	−1.27
23	不良品	3.51	0.82
24	不良品	0.31	−0.99
25	不良品	1.69	−0.81

（ア） 図１と表１から考察できることとして，誤っていることを以下の①～④から１つ選びなさい。

① ランダムに選んだ製品中の不良品におけるセンサ値 y は，1.00 より小さい。

② ランダムに選んだ製品中の不良品におけるセンサ値 x は，0.00 より大きい。

③ ランダムに選んだ製品中の良品におけるセンサ値 y は，-1.5 より大きい。

④ ランダムに選んだ製品中の良品におけるセンサ値 x は，2.0 より小さい。

（イ） 工場長は，この検査結果が既知の製品に関するデータを利用し，良品か不良品かが未知の製品 A に対し，そのセンサ値 $s=(x,\ y)$ から良品か不良品かを予測することを考えた。工場長は，以下のプロセス P により予測を行う。

プロセス P

(i) 収集した各データ点 s_1，...，s_{25} と s との距離 $d_{s_1,s}$，...，$d_{s_{25},s}$ を全て算出する。

(ii) 収集したデータ点 s_1，...，s_{25} の中から，s と最も距離が小さいデータ点 s_i を算出する。

(iii) データ点 s_i に対応する製品 A_i が良品であれば製品 A を良品と予測し，データ点 s_i に対応する製品 A_i が不良品であれば製品 A を不良品と予測する。

ここで，２つのデータ点 $s=(x,\ y)$ と $s'=(x',\ y')$ の距離 $d_{s,s'}$ は，以下の式により求める。

$$d_{s,s'}=\sqrt{(x-x')^2+(y-y')^2}$$

良品か不良品かが未知の製品 A に対し，センサ値を取得したところ，$s=(2.85,\ 0.78)$ であった。このとき，以下の問い(a)，(b)に答えなさい。

(a) s_1，...，s_{25} の中で，s と最も距離が小さいデータ点 s_i はどれか。製品番号 i と，そのときの**距離の二乗値** $d^2_{s_i,s}$ を，**小数点第三桁を四捨五入し，第二桁**まで求めなさい。

(b) プロセス P を用いると，製品 A は良品，不良品どちらと予測されるか答えなさい。

（2） 工場長は，新たにデータを追加し，計10000個の製品に関するセンサ値と検査結果を用意した。良品か不良品かが未知の製品に対し予測を行うときの性能を見積もるため，データをランダムに7500個と2500個に分け，7500個を検査結果が既知のデータ，2500個を検査結果が未知のデータと想定し，どの程度正しく予測ができるかを算出してみることにした。（1）におけるプロセスPを用い，7500個のデータを利用して2500個のデータに対し予測を行った結果をまとめたのが表2である。このとき，以下の問い(ア)～(エ)に答えなさい。

表2：プロセスPによる予測結果

実際＼予測	良品	不良品
良品	1986	38
不良品	87	389

（ア） 以下の［ a ］，［ b ］に入る数を答えなさい。

表2の結果から，実際は良品である製品に対し不良品と予測してしまった製品の数は［ a ］個であり，実際は不良品である製品に対し良品と予測してしまった製品の数は［ b ］個である。

（イ） 工場長は，以下の５つの指標を用いて評価することにした。表２から，［指標１］～［指標５］を，**小数点第三桁を四捨五入し，小数点第二桁**まで求めなさい。

［指標１］：$\dfrac{\text{（予測を正しく行うことができたデータの数）}}{\text{（予測を行ったデータの数）}}$

［指標２］：$\dfrac{\text{（実際に不良品である製品に対し，不良品と予測した製品の数）}}{\text{（不良品と予測した製品の数）}}$

［指標３］：$\dfrac{\text{（実際に良品である製品に対し，良品と予測した製品の数）}}{\text{（良品と予測した製品の数）}}$

［指標４］：$\dfrac{\text{（実際に不良品である製品に対し，不良品と予測した製品の数）}}{\text{（実際の不良品の数）}}$

［指標５］：$\dfrac{\text{（実際に良品である製品に対し，良品と予測した製品の数）}}{\text{（実際の良品の数）}}$

（ウ） 工場長は「不良品と予測された製品は良品も含めて廃棄せざるを得ない。しかし，良品を廃棄してでも，不良品を見逃さないことを最優先したい。そのためには［ c ］が大きいプロセスを開発する必要がある」と考えた。［ c ］に入る指標のうち最も適切なものを，［指標１］～［指標５］の中から１つ選びなさい。

（エ） （ウ）における工場長の考えに対して，工場長補佐から以下のような指摘があった。［ d ］～［ f ］に入る**整数**を答えなさい。

工場長補佐の指摘：
「コストも考慮する必要があると思います。工場長のおっしゃるとおり，不良品と予測されたものは良品も含めて廃棄することとします。例えば，この 2500 個の製品を出荷するときの，不良品を１個出荷してしまったときのコストを 1000 円とし，良品を１個廃棄してしまったときのコストを 400 円とします。もしプロセスＰを使わず，2500 個全てを良品とみなして出荷してしまった場合，かかるコストは［ d ］円ですが，プロセスＰを用いて良品と予測した製品のみ出荷した場合，かかるコストは［ e ］円と大きく低減します。一方で，不良品を一切見逃さずに予測できるプロセスが見つかったとしても，良品の廃棄個数が［ f ］個以上になると，プロセスＰよりコストは大きくなってしまいます。」