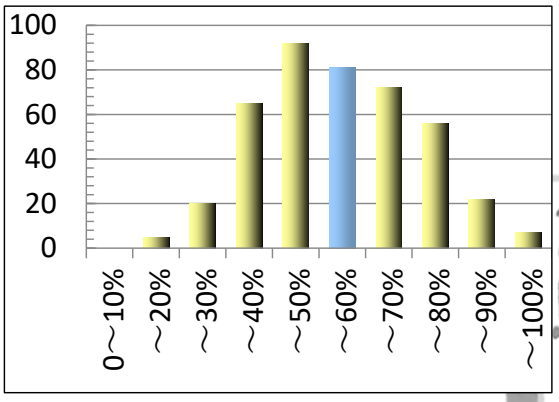
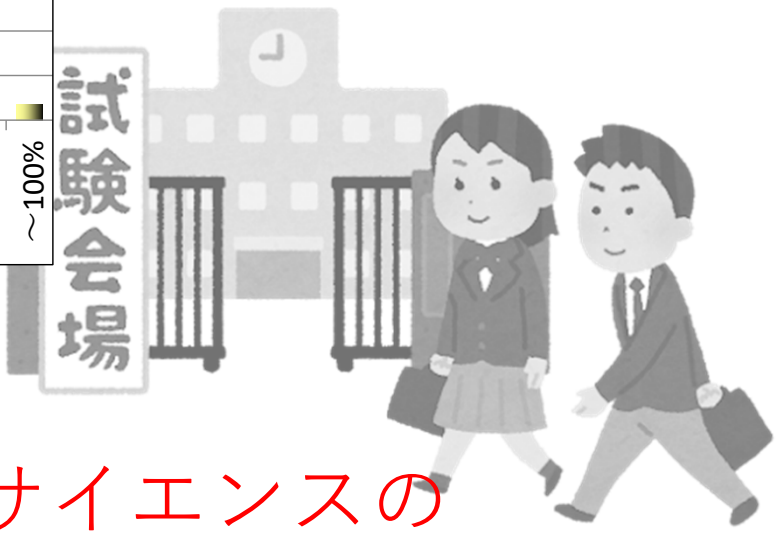
情報I 試作等のデータサイエンスの定番問題





データサイエンスの定番問題を解いて確実な得点アップ

ドラフト版 2024年12月

太田 剛

学習を開始する前に読んでください

この教材について

この教材は2025年より始まる大学入学共通テストの「情報I」の学習を支援するために作成されました。2024年時点ではまだ「情報I」の試験は始まっていませんが、大学入試センターや多くの情報系の一般入試を行う私立大学では情報Iに対応した試作問題を公開しています。

これらの試験から予想される出題には、統計、データサイエンス分野が含まれています。この分野は過去問題の蓄積は少ないですが、出題パターンも限られているので、その解き方を理解すれば必ず点数がとれるものであり、この教材自体の問題数もそれほど多くないため、短時間で統計、データサイエンス分野の学習ができるものだと考えています。

この教材は統計、データサイエンス分野の各分野ごとに各大学の試作問題等を整理しています。各問題は以下の表に内容を示していますが、平均の検定分野については出題の可能性が低いと想定し、扱っていません。

大問	問題(元問題)	基本統計	グラフ	移動平均	標準偏差	度数分布	散布図と相関係数	クロス集計
1	若年層の生活時間調査 (共通テスト試作:2022年11月)	○	○		○		○	
2	,天気と売り上げ (広島市立大学模擬(2024年9月	○			○	○	○	
3	エアコンとアイスクリームの売り上げ (共通テスト試作参考:2022年11月)		○	○			○	
4	相関係数(日本大学サンプル:2024年3月)					○	○	
5	ロボットの生産品質 (京都産業大学模擬問題:2024年)							○
6	サッカーのチーム分析 (共通テストサンプル:2021年3月)	○				○	○	○

学習の進め方について

本教材では問題と回答のみ記載しています。基本的な考え方の解説は随時動画で公開していいく予定です。

定番問題

大問1.若年層の生活時間調査(共通テスト試作:2022年11月改)

15歳以上19歳以下の若年履について,都道府県別に平日1日の中で各生活行動に費やした時間(分)の平均値を,スマートフォン・パソコンなどの使用時間をもとにグループに分けてまとめたものの一部である。ここでは,1日のスマートフォン・パソコンなどの使用時間が1時間未漂の人を表1-A,3時間以上6時間未満の人を表1-Bとしている。

表１−Ａ：スマートフォン・パソコンなどの使用時間が

１時間未満の人の生活行動時間に関する都道府県別平均値

都道府県	睡眠（分）	身の回りの用事（分）	食事（分）	通学（分）	学業（分）	趣味・娯楽（分）
北海道	439	74	79	60	465	8
青森県	411	74	73	98	480	13
茨城県	407	61	80	79	552	11
栃木県	433	76	113	50	445	57

表１−Ｂ：スマートフォン・パソコンなどの使用時間が

３時間以上６時間未満の人の生活行動時間に関する都道府県別平均値

都道府県	睡眠（分）	身の回りの用事（分）	食事（分）	通学（分）	学業（分）	趣味・娯楽（分）
北海道	436	74	88	63	411	64
青森県	461	57	83	55	269	44
茨城県	443	80	81	82	423	63
栃木県	386	120	79	77	504	33

問1　花子さんたちは,これらのデータから次のような仮説を考えた。表1-A,表1−Bのデータだけからは分析できない仮説を,次の⓪~③のうちから1つ選べ[1]。

⓪若年層でスマートフォン・パソコンなどの使用時間が長いグループは,使用時間が短いダループよりも食事の時間が短くなる傾向があるのではないか。
①若年層でスマートフォン・パソコンなどの使用時間が長いグループに注目すると,スマートフォン・パソコンなどを朝よりも夜に長く使っている傾向があるのではないか。
②若年層でスマートフォン・パソコンなどの使用時間が長いグループに注目すると,学業の時間が長い都道府県は趣味・娯楽の時間が短くなる傾向があるのではないか。
③若年層でスマートフォン・パソコンなどの使用時間と通学の時間の長さは関係ないのではないか。

問2　花子さんたちは表1ーA,表1ーBのデータから睡眠の時間と学業の時間に注目し,それぞれを図1と図2の箱ひげ図にまとめた。これらから読み取ることができる最も適当なものを一つ選べ。[2]

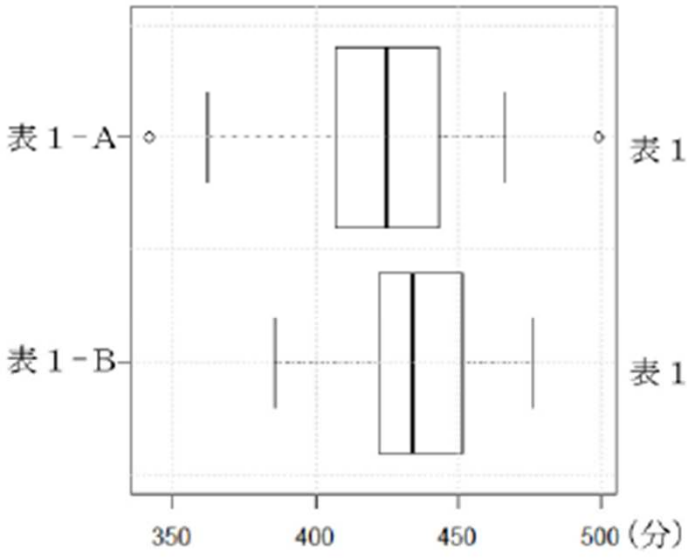


図１　睡眠の時間の分布

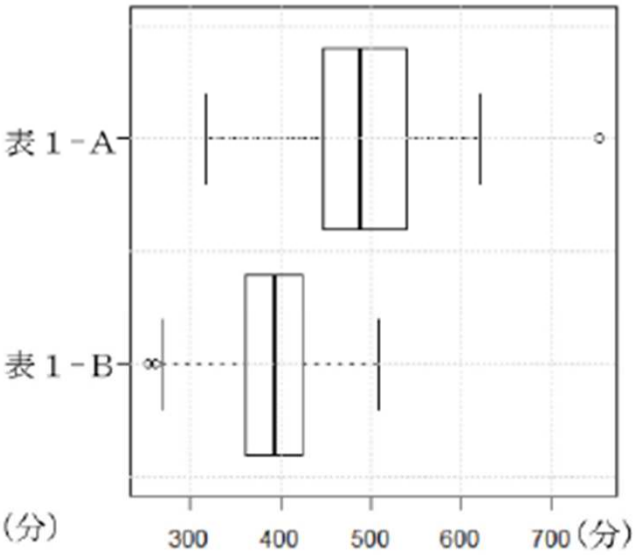
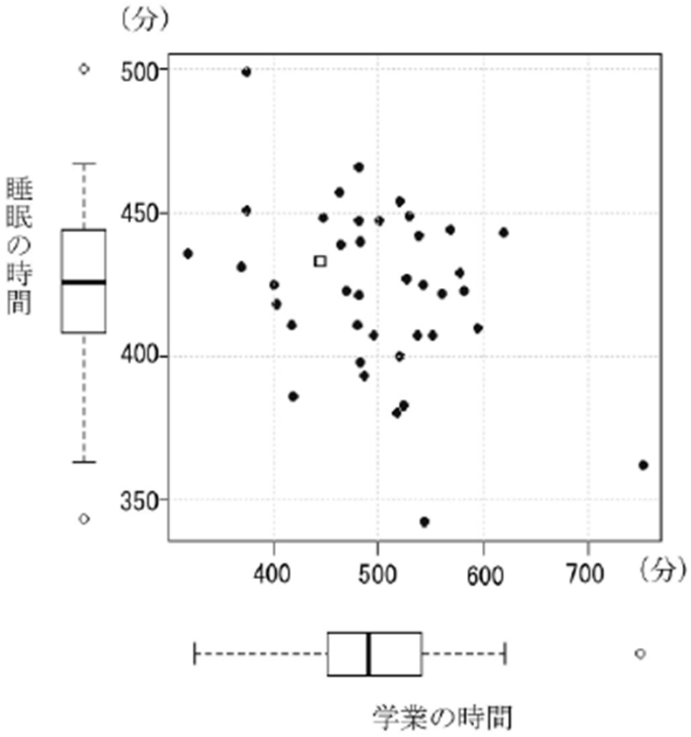


図２　学業の時間の分布

⓪ 唾眠の時間が420分以上である都道府県の数を見たとき,表1-A方が表1ーBよりも多い。
① 学業の時間が550分以上の都道府県は,表1-Aにおいては全体の半数以上あり,表1-Bにおいては一つもない。
② 学業の時間が450分未満の都道府県は,表1-Bにおいては全体の75%以上であり,表1-Aにおいては50%未満である。
③ 都道府県別の睡眠の時間と学業の時間を比較したとき,表1-Aと表1-Bの中央値の差の絶対値が大きいのは睡眠の時間の方である。

問3 花子さんたちは,表1-Aについて,睡眠の時間と学業の時間の関連を調べることとした。次の図3は,表1-Aについて学業の時間と睡眠の時間を散布図で表したものである。ただし,2個の点が重なって区別できない場合はロで示している。



都道府県単位でみたとき,学業の時間と睡眠の時間の間には,全体的には弱い負の相関があることが分かった。この場合の負の相関の解釈として最も適当なものを一つ選べ。[3]なお,ここでは,データの範囲を散らぽりの度合いとして考えることとする。
⓪ 睡眠の時間の方が,学業の時間より散らばりの序合いが大きいと考えられる。
① 睡眠の時間の方が,学業の時間より散らばりの序合いが小さいと考えられる。
② 学業の時間が長い都道府県ほど睡眠の時間が短くなる傾向がみられる。
③ 学業の時間が長い都道府県ほど睡眠の時間が長くなる傾向がみられる。

問4 花子さんたちは都道府県別になかたときの睡眠の時間を学業の時間で説明する回帰直線を求め,図3の散布図にかき加えた(図4)。すると回帰直線から大きく離れている県が多いことが分かったため,自分たちの住むP県がどの税度外れているのかを調べようと考え,実際の睡眠の時間から回帰直線により推定される睡眠の時間を引いた差(残差)の程度を考えることとした。そのために,残差を比較しやすいように,回帰直線の式をもとに学業の時間から推定される腰眠の時間(推定値)を横軸に,残差を平均値0,標準偏差1に変換した値(変換値)を縦軸にしてグラフ図5を作成した。参考にG県がそれぞれの図でどこに配置されているかを示している。また,図5の口で示した点については,問題の都合上黒丸で示している。

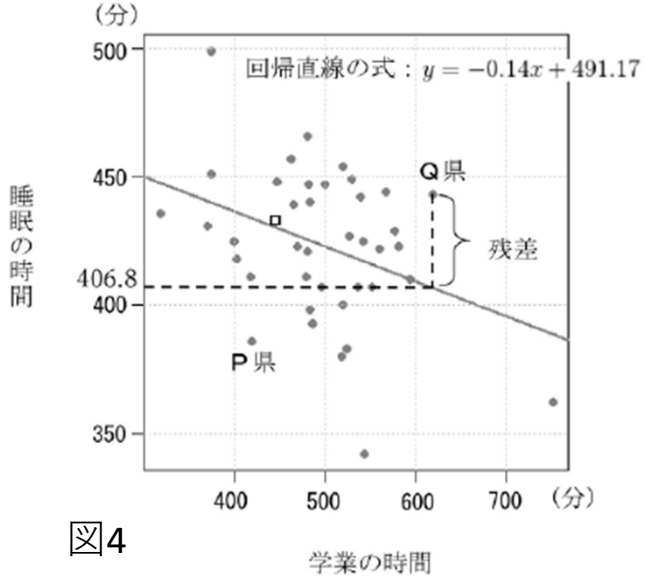


図4

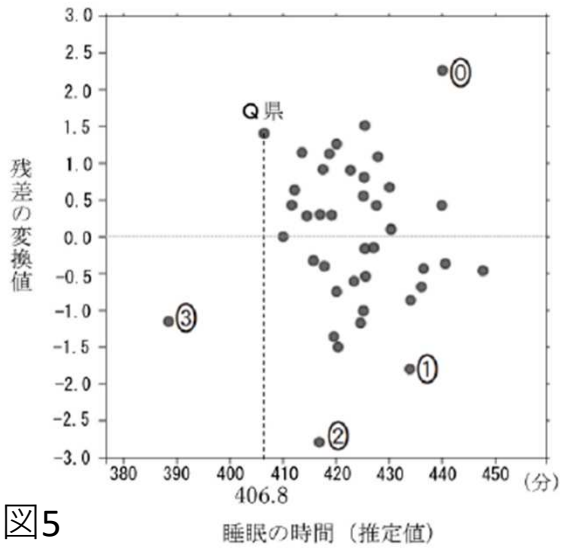


図5

図4と図5から読み取ることができることとして,平均値から標準偏差の2倍以上離れた値を外れ値とする基準で考えれば,外れ値となる都道府県の数は[4]個である。図5中のP県については,図6中の0~3のうち[5]に対応しており,花子さんたちはこの基準に従いP県は[6]と判断した。

[6]の選択肢
⓪ 外れ値となっている ① 外れ値となっていない
② 外れ値かそうでないかどちらともいえない

大問2. 天気と売り上げ(広島市立大学模擬(2024年9月改 マークシート化)

表は,ある店舗の3月3日から3月12日までの10日間の気象情報(天気と最高気温),来客数および売り上げのデータである。このとき,以下の問いに答えよ。

表１　ある店舗の売上データ

No.	日付	天気	最高気温（℃）	来客数（人）	売上高（千円）
1	3月3日	晴れ	10	5	900
2	3月4日	雨	8	4	300
3	3月5日	晴れ	15	9	2,000
4	3月6日	雨	6	8	600
5	3月7日	曇り	9	10	950
6	3月8日	晴れ	12	9	1,800
7	3月9日	晴れ	10	12	2,500
8	3月10日	曇り	8	8	800
9	3月11日	晴れ	12	10	1,800
10	3月12日	雨	5	15	－

問1 天気は[１]的データで[２]である。他の3つのデータは[３]的データであり、気温は[４], 来客数は[５]、売上高は[６]である。

[１]～[５]の選択肢
⓪ 名義尺度 ① 順序尺度 ② 間隔尺度 ③ 比例尺度 ④ 質 ⑤ 量

問2　表は天気ごとの売上高の平均値と標準偏差である。これからから読み取れることを一つ選べ。[７]

天気	合計(千円)	平均(千円)	標準偏差
晴れ	9000	1800	511.7
曇り	1750	875	75.0
雨	900	450	150.0

⓪ 曇りの日が最も、売り上げの日によるバラつきが少ない。
① 雨の日が最も、売り上げの日によるバラつきが少ない。
② 晴れと雨の日を比較した場合、日による晴れの方がバラつきが少ない。
③ 晴れの日が最も、売り上げのひによるバラつきが少ない。

問3 1月1日から3月31日までの売上データから,天気が晴れのときの来客数と売上高との相関係数を求めたところ0.90であった。この相関係数が意味することを一つ選べ。[８]

⓪ 晴れの日は日ごとに来客人数に関係なく、売り上げが大きい。
① 晴れの日は日ごとに来客数が多いほど比例して、売り上げが大きくなる。
② 晴れの日は日ごとに来客数が多くなると、相対的に売り上げが減っていく。
③ 晴れの日は来客数が多い。

大問3. エアコンとアイスクリームの売り上げ(共通テスト試作参考改:2022年11月)

次のデータは,2016年1月から2020年12月までの全国のエアコンの売上台数(単位は千台)とK市のアイスクリームの売上個数(単位は個)を表している。

表1　エアコンとアイスクリームの売上データ

年月	エアコン(千台)	アイス(個)
2016年 1月	434	464
2016年 2月	504	397
2016年 3月	769	493
2016年 4月	420	617
2016年 5月	759	890
2016年 6月	1470	883
2016年 7月	1542	1292
2020年12月	635	599

花子さんは,これら二つの売上数の関係を調べるためにこのデータを,次の図1のようなグラフで表した。このグラフでは,横軸は期間を月ごとに表し,縦軸はエアコンの売上台数(単位は千台)とアイスクリームの売上個数(単位は個)を同じ場所に表している。破線はエアコン,実線はアイスクリームの売上数を表している。

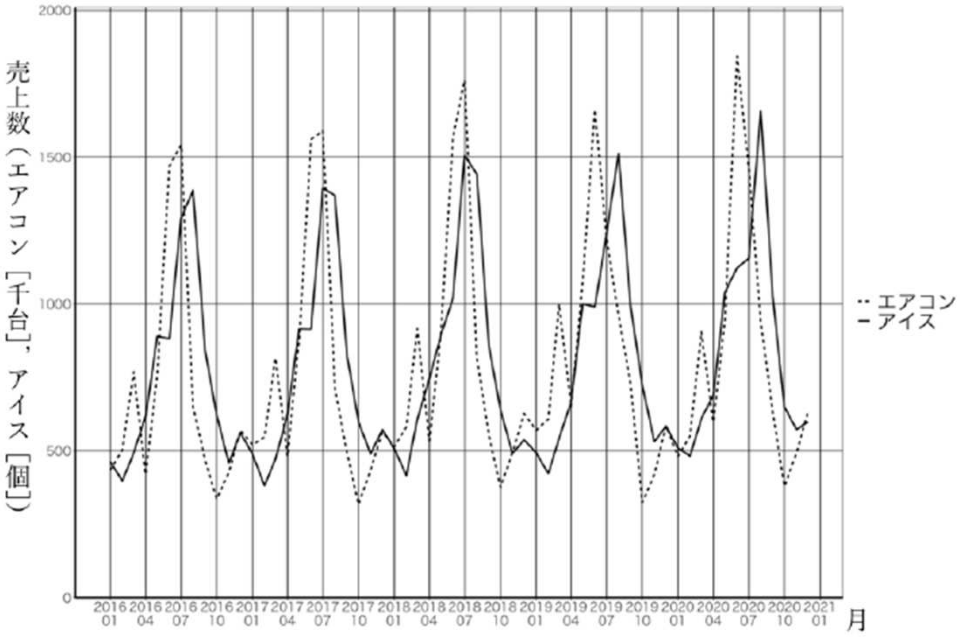


図1　エアコンとアイスクリームの売上数のグラフ

問1 図1のグラフを見て読み取れることを,次のうちから一つ選べ。[1]

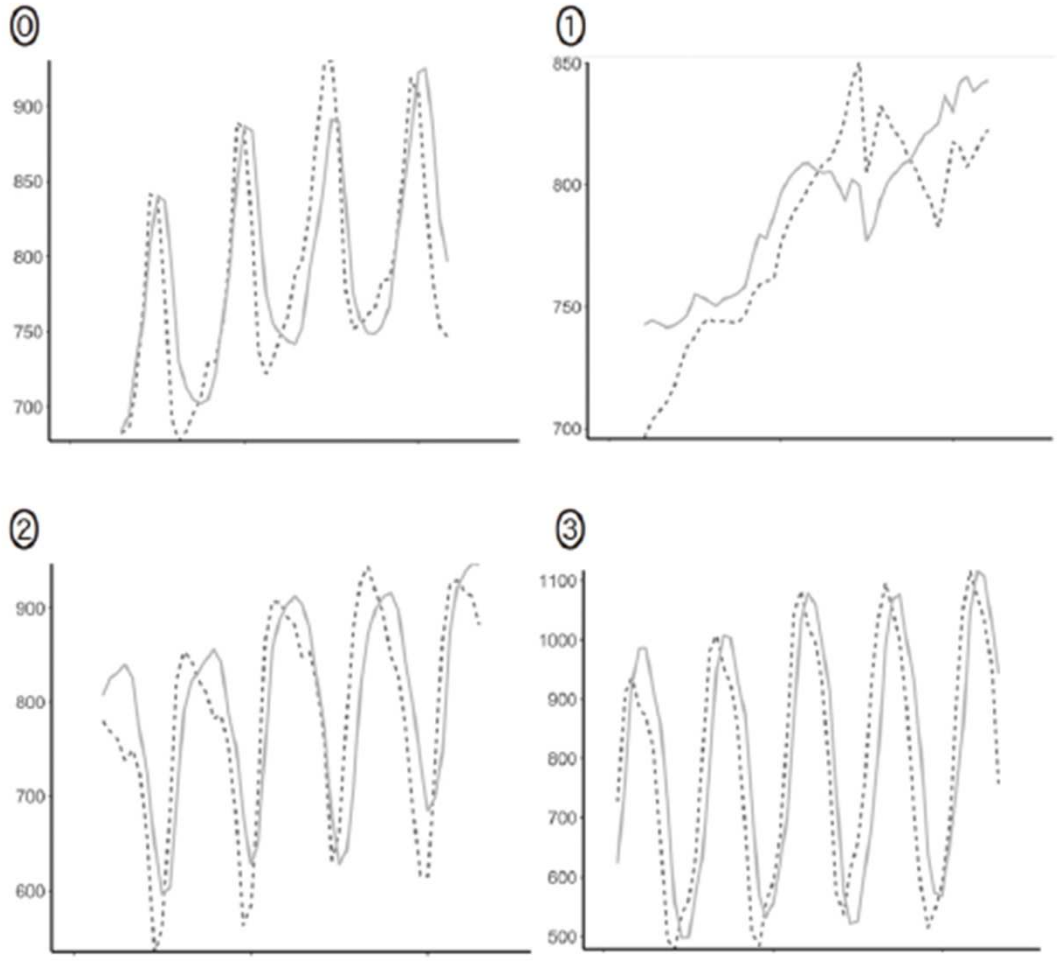
⓪ アイククリームの売上個数は毎月増加している。
① エアコンの売上台数は年々減少している。
② 年ごとの最もよく売れる時期についてはエアコンの方がアイスクリームよりもやや早い傾向がある。
③ 2016年10月は,エアコンの売上台数よりもアイスクリームの売上個数の方が多い。

問2　エアコンやアイスクリームの売り上げが年々増加しているのかどうかを調べたいと考えた花子さんは,月ごとの変動が大きいので,数か月のまとまりの増減を調べるためにその月の前後数か月分の平均値(これを移動平均という)を考えてみることにした。

表２　エアコンの移動平均を計算するシート

年月	エアコン(千台)	６か月移動平均
2016 年　1 月	434	
2016 年　2 月	504	
2016 年　3 月	769	
2016 年　4 月	420	726.0
2016 年　5 月	759	910.7
2016 年　6 月	1470	935.2
2016 年　7 月	1542	885.2
2016 年　8 月	651	871.2

例えば,表2は6か月ごとのまとまりの平均を計算している例である。`6か月移動平均」の列について,2016年1月から6月までの6か月の平均値である726.0%2016年4月の行に記載している。このようにエアコンとアイスクリームの売上数について6か月,9か月,12か月,15か月の移動平均を求め,それらの一部をグラフに描いたものが次の⓪~③である。これらのグラフはそれぞれ順不同である。この中から,12か月移動平均の増港を表していると考えられるグラフを,次のうちから一つ選べ。[２]



問3 次の文章を読み,空欄[3]に入れるのに最も適当なものを,後の解答群のうちから選べ。

より詳細な増減について調べることにした。図1では,エアコンやアイスクリームの売上数は,ある一定期間ごとの繰返しでほぼ変化している傾向があるのではないかという仮説を立て,これが正しいかどうかを確かめるために,まずエアコンの売上台数のデータと,そのデータをnか月だけずらしたデータとの相関係数を求めてみることにした。そのデータについて統計ソフトウェアを用いてグラフにたものが次の図2である。

横軸は`ずれの月数(nで)あり,縦軸は相関係数を表している。例えば,横軸の0のときの値は,エアコンの同じデータ同士の相関係数であるので,明らかに1を示していることが分かる。

図2から,エアコンの売上台数の増減は,おおよそ[3]月ごとにほぼ同じように変化していると考えることができる。同様のグラフを作成すると,アイスクリームの売上個数もエアコンと同じ月数ごとで変動していることが分かった。

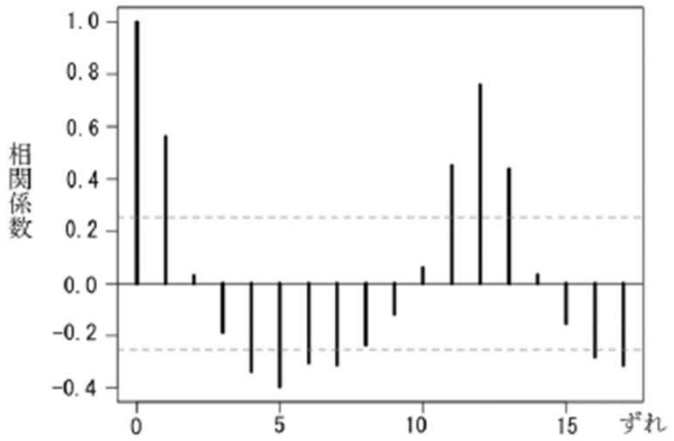


図２　エアコンの売上データをずらした月数とその相関係数

[3]の選択肢
⓪ 2　① 5　② 12　③ 14

問4　次にエアコンとアイスクリームの売上数の関係を調べようと考えて,その相関係数を求めると,約0.62であった。しかし,図1を見て,売上のピークが多少ずれていると考え,試しに次の表3のようにエアコンの売上台数のデータを1か月あとにずらして考えてみた。例えば,2016年1月のエアコンの売上台数である434(千台)を2016年2月にずらし,以降の月についても順次1か月ずらしている。このデータをもとに,相関係数を求めてかみたところ約0.86となった。同様に,エアコンの売上台数のデータをnか月後にすらしたデータとの相関係数を求めてみたところ,次の表4のような結果になった。

このことから考えられる,最も適当なものを,次の⓪~④のうちから選べ。[4]

表４　エアコンとアイスクリームの売上数のずらした月数と相関係数

ずれ(n)	-3	-2	-1	0	1	2	3
相関係数	-0.45	-0.17	0.21	0.62	0.86	0.70	0.17

⓪ アイスクリームの売上個数のピークの方が,エアコンの売上台数のピークより1か月早く訪れる。
① エアコンを買った人は,翌月に必ずアイスクリームを購入している。
② アイスクリームが売れたので,その1か月後にエアコンが売れることが分かる。
③ 気温が高いほどエアコンもアイスクリームも売れる。
④ ある月のアイスクリームの売上個数の予測をするとき,その前月のエアコンの売上台数から,ある程度の予測ができる。

問5 次の文章を読み,空欄[5]～[7]に入れるのに最も適当なものを,後の解答群のうちから選べ。

売上数と他の要素との関係も調べという意見をもらった。そこで,K市の同じ期間の月別平均気温と平均湿度のデータを収集し,それらのデータを合わせて,図3のような図を作成した。(これを散布図・相関行列という。)図3の左下の部分は相関係数,右上の部分は散布図,左上から有下への対角線の部分はそれぞれの項目のヒストダラムを表している。

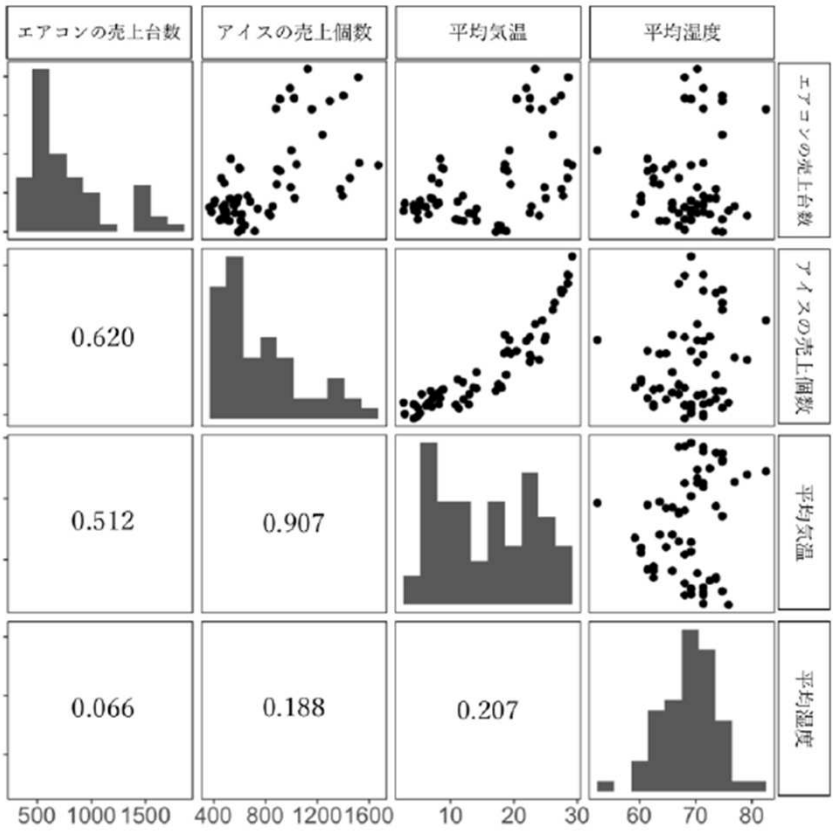


図３　散布図・相関行列

図3から花子さんは,次の図4のような関係図を作成した。図中の実線の矢印の向きは,ある項目への影響を表している。また,矢印の線の太さは相関係数の絶対値が0.7以上を太い縄で,0.7未満を細い線で表し,その相関の強さを示している。

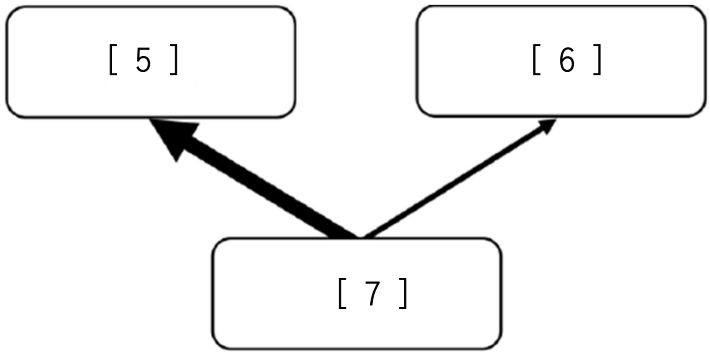
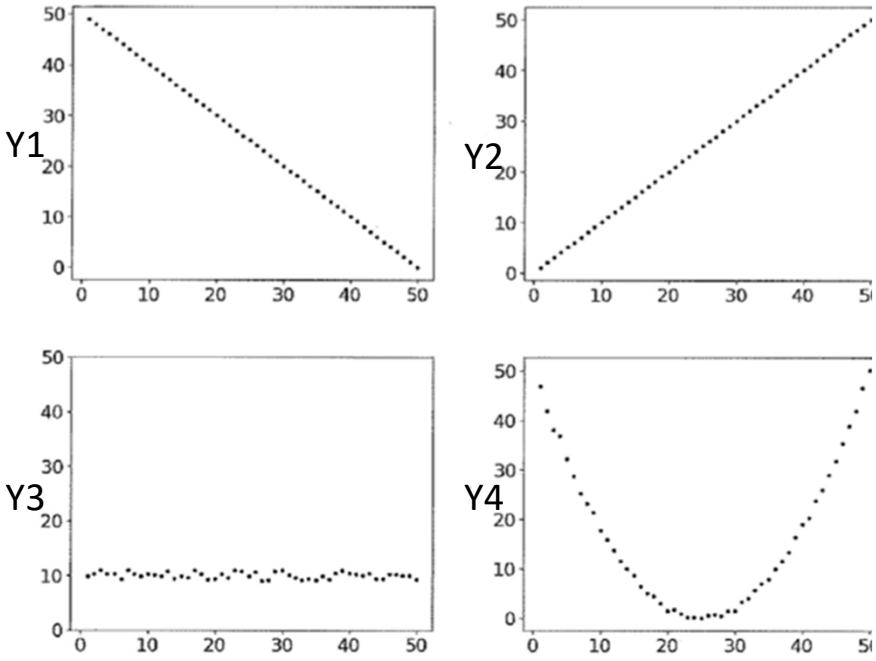


図４　項目間の相関と影響を表した図

[5] ～ [7]の選択肢
⓪ エアコンの売上台数　① アイスクリームの売上個数
② 平均気温　③ 平均湿度

大問4. 相関係数(日本大学サンプル改:2024年3月)

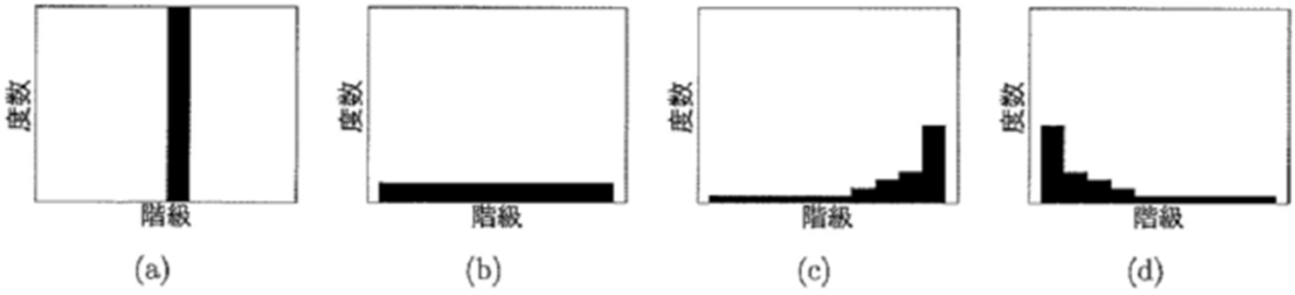
図 1に。5つの属性(カラム)X,Y1Y2,Y3,Y4から構成成されるデータ5個のデータセットに関する散布図を示す、とのデータセットに関し,空欄[1] 〜[4]入れるのに最も適当なものをそれぞれの解答群から選びなさい.



問1 Y1〜Y4のうち,Xと相関関係が強いのは[1]のみである.
問2 Y1とY2の間には,Xを[2]とする擬似相関が疑われる。[3]。.

[1]の選択肢
① Y1　② Y2　③ Y1とY2　④ Y1とY4　⑤Y2とY4　⑥Y1,Y2, Y4　⑦Y3とY4
[2]の選択肢
①交換因子　②連結因子　③ 相関因子　④交絡因子
[3]の選択肢
① 加えて,Y1の値が減少するとY2の値が増社するという因果関係も認めあられる
② そのため,両者には相関関係はないものと考えるべきである
③ しかしながら,強い相関関係は認められる
④ また,両者の問で分散が大きく異なる

問3 新たな属性として,Y5=X+Y1およびY6=X+Y4を考える、とのとき,Y5とY6のヒストグラムの概形はそれぞれ図3の[4]となる.



[4]の選択肢
① (a)と(c)　② (a)と(d)　③ (b)と(c)　④ (b)と(d)

大問5.ロボットの生産品質(京都産業大学模擬問題改:2024年)

あるロボット工場では、製造したロボットの出荷前の評価を2段階に分けて行っている。1つ目は、パーツごとの基本性能を数値的に評価する性能試験で、2つ目は、ロボットが全体として不良品でないかを最終的に判定する動作チェックである。工場で製造したロボットには全て、これら2種類の評価をそれぞれ実施し、そのデータを製造月ごとにまとめている。

下の表1は、この工場で製造したロボットの性能試験の結果を表している。この表は、工場で製造した全ロボットのうち、性能試験で基準値未満でかあったロボットの個数と、基値以上であった個数を、4月から7月までの月ごとにまとめたものである。一方、表2は、工場で製造したロボットのうち最終動作チェックにおいて不良品と判定されたロボットだけに絞って、表1と同様に、性能試験で基準値未満であった個数と基準値以上であった個数を月ごとにまとめたものである。例えば、4月に製造したロボットで不良品と判定されたもののうち性能試験で基準値未満であったロボットは24個、1つ目の性能試験で基準値以上であっても不良品と判断されたロボットは6個である。これらを合わせて、4月の不良品の総数は30個である。

以下の設問では、これら性能試験と最終動作チェックの結果の関係性について考える。

表 1. ロボットの性能試験における基準値未満の個数と基準値以上の個数（個）

		月				
		4 月	5 月	6 月	7 月	計
性能試験	基準値未満	96	142	160	242	640
	基準値以上	654	858	1090	1758	4360
	計	750	1000	1250	2000	5000

表 2. 不良品のロボットが性能試験で基準値未満であった個数と基準値以上であった個数（個）

		月				
		4 月	5 月	6 月	7 月	計
性能試験	基準値未満	24	48	40	48	160
	基準値以上	6	12	10	12	40
	計	30	60	50	60	200

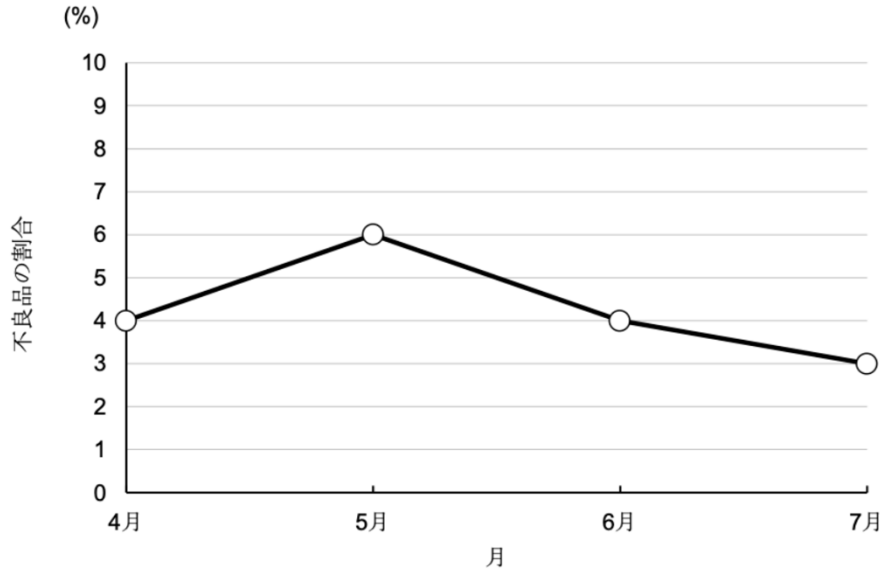
問1　まず初めに、4月から7月までの月を合算して、性能試験の結果(基準値未満、基準値以上)と最終動作チェックの結果(不良品、不良品でない)とを掛け合わせたロボットの個数の表を作成した。これを示す下の表3において、例えば、性能試験が基準値未満で最終動作チェックで不良品と判定された個数は、表中の①の箇所に記載され、その数は160個である。同様にして、表内の[１]から[７]までの箇所全てに、適切な整値(該当する個数)を記入せよ。

表3. 4月から7月までを合算した､性能試験と最終動作チェックの結果を掛け合わせた表(個)

		最終動作チェック		
		不良品	不良品でない	計
性能試験	基準値未満	① 160	[１]	[２]
	基準値以上	[３]	[４]	[５]
	計	[６]	[７]	⑨ 5000

問2 作成した表3を用いて、まず4月から7月までの月を合算して、不良品のロボットのうち性能試験で基準値未満であった割合を調べるろことを考える。この割合は百分率で[8](%)である。
　同様に表3を用いて次に、ロボットが性能試験で基準値未満であった場合に不良品である割合を見積もることを考える。4月から7月までの月を合算して、この割合を百分率で[9](%)である。

問3 次に、4月から7月にかけてこの工場で製造したロボットに占める不良品の割合がどのように変化したかを考える。そこで、冒頭に示した表1と表2のデータから、各月に工場で製造したロボットの中の不良品の割合を百分率(%)でそれぞれ求め、その値の4月から7月までの変化を表すダラフを作成した。



　この作成したグラフと、冒頭に示した表1および表2から分かることとして、下に挙げたaーeを考えた。これらのうち、正しいものにはOを、正しくないものにはXをつけよ。
a.[10]この工場で製造したロボットの不良品の割合は、4月から7月まで毎月増え続けてい
る
b. [11]この工場で製造したロボットの不良品の割合は、4月から5月にかけて増加したが、それ以降6月から7月にかけて減少した
c. [12]不良品のロボットのうち性能試験で基準値未満であった割合は、4月から7月まで変化せず一定である
d. [13]不良品のロボットの総数が6月から7月にかけて増えていることは、この工場で製造したロボットに占める不良品の割合が6月から7月にかけて増えたことを示している
e. [14]不良品のロボットの総数が6月から7月にかけて増えていることは、この工場で製造したロボットに占める不良品の割合が増えたためではなく、製造したロボットの個数が6月から7月にかけて増えたことと関係している

大問6.サッカーのチーム分析(共通テストサンプル改:2021年3月)

S高等学校サッカー部のマネージャーをしている鈴木さんは,「強いサッカーチームと弱いサッカーチームの違いはどこにあるのか」というテーマについて研究している。鈴木さんは,ある年のサッカーのワールドカップにおいて,予選で敗退したチーム(子選敗退チーム)と,予選を通過し,決勝トーナメントに進出したチーム(決勝進出チーム)との違いを,データに基づいて分析することにした。このデータで各国の代表の32チームの中で,決勝進出チームは16チーム,予選敗退チナームは16チームであった。

分析対象となるデータは,各チームについて,以下のとおりである。

試合数‥大会期間中に行った試合数

総得点‥大会で行った試合すべてで獲得した得点の合計

ショートパス本数…全試合で行った短い距離のパスのうち成功した本数の合計

ロングパス本数…全試合で行った長い距離のパスのうち成功した本数の合計

反則回数‥全試合において審判から取られた反則回数の合計

鈴木さんは,決勝進出チームと予選敗退チームの違いについて,このデータを基に,各項目間の関係を調べることにし、,表1のデータシートを作成した。決勝進出チームと予選政退チームの違いを調べるために,決勝進出の有無は,決勝進出であれば1。予選敗退であれば0とした。また,チームごとに試合数が異なるので,各項目を1試合当たりの数値に変換した。

表1　ある年のサッカーのワールドカップのデータの一部（データシート）

	A	B	C	D	E	F	G	H	I	J	K
1	チームID	試合数	総得点	ショートパス本数	ロングパス本数	反則回数	決勝進出の有無	1試合当たりの得点	1試合当たりのショートパス本数	1試合当たりのロングパス本数	1試合当たりの反則回数
2	T01	3	1	834	328	5	0	0.33	278.00	109.33	1.67
3	T02	5	11	1923	510	12	1	2.20	384.60	102.00	2.40

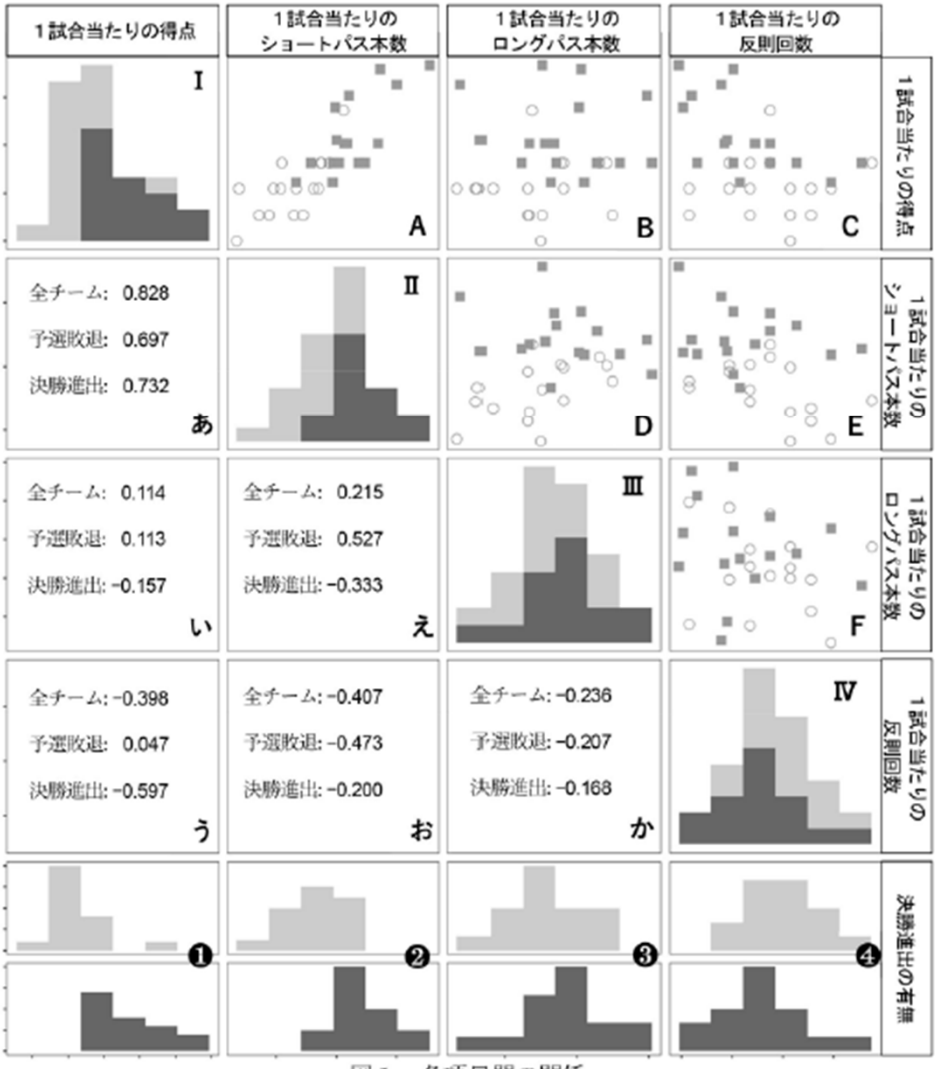


図1　各項目間の関係

図1のI〜IVは,それぞれの項目の全参加チームのヒストグラムを決勝進出チームと子選敗退チームとで色分けしたものであり,❶—❹は決勝進出チームと予選敗退チームに分けて作成したヒストグラムである。あ〜かは,それぞれのつの項目の全参加チームと決勝進出チーム,予選政退チームのそれぞれに限定した相関係数である。またA〜Fは,それぞれのつの項目の散布図を決勝進出チームと予選敗退チームをマークで区別して描いている。例えば,図1のAは縦軸を`1試合当たりの得点」,横軸を`1誠合当たりのショートバス本数とした散布図であり,それに対応した相関係数はあで表されている。

問1　図1を見ると,予選政退チームにおいてはほとんど相関がないが,決勝進出チームについて負の相関がある項目の組合せは,一試合当たりの[1]と[2]である〟また,決勝進出チームと予選敗退チームとで,相関係数の符合が逆符号であり,その差が最も大きくなっている関係を表している散布図は[3]である〟したがって,散布図のニつの記号のどちらが決勝進出チームを表しているかが分かった。

[1]と[2]の選択肢
⓪ 得点　①ショートバス本数　②ロングバス本数　③ 反則回数
[3]の選択肢
⓪A　①B　②C　③D　④E　⑤F

問2 図1から読み取れることとして誤っているものを一つ選べ。[4]
⓪ それぞれの敢布図の中で,決勝進出チームは黒い四角形(■),予選敗退チームは白い円(〇)で表されている。
① 全参加チームを対象としてみたとき,最も強い相関がある項目の組合せは1試合あたりの得点と1試合あたりのショートバス本数である。
② 全参加チームについて正の相関がある項目の組合せの中には,決勝進出チーム,予選敗退チームのいずれも負の相関となっているものがある。
③ 1試合当たりのショートパス本数の分布を表すダラフ❷で,下の段は決勝進出チームのヒストグラムである。

問3 次の文書の[5]〜[7]の数値を求めよ(小数第二位までに四捨五入する)
　鈴木さんは,図1から,1試合当たりの得炭とショートパス本数の関係に着目し,さらに詳しく調べるために,1試合当たりの得点をショートバス本数で予測する回帰直線を,決勝進出チームと予選政退チームとに分けて図2のように作成した。

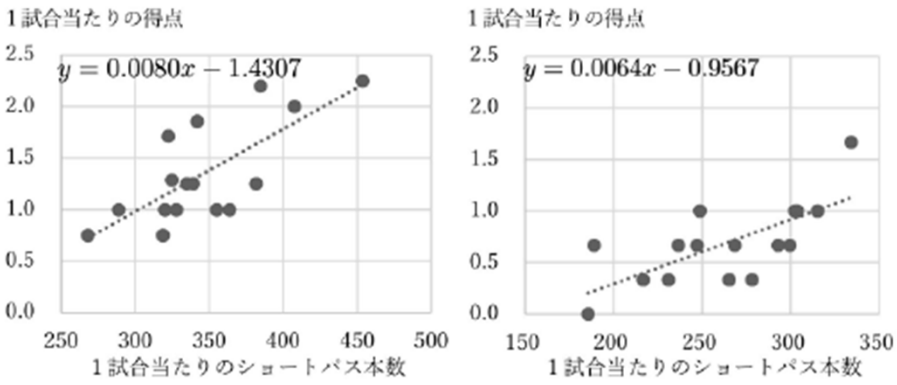


図２　決勝進出チーム(左)と予選敗退チーム(右)の
１試合当たりの得点とショートパス本数の回帰直線

この結果からショートバス100本につき,1試合当たりの得点増加数を決勝進出チームと予選敗退チームで比べた場合,[5]点の差があり,ショートバスの数に対する得点の増加量は決勝進出チームの方が大きいと考えた。また,1試合当たりのショートパスが320本0とき,回帰直線から予測できる得点の差は,決勝進出チームと予選敗退チームで,[6]点差があることが分かった。鈴木さんは,グラフからは傾きに大きな差が見られないこの二つの回帰直線について,実際に計算してみると差を見つけられることが実感できた。
さらに,ある決勝進出チームは,1試合当たりのショートバス本数が384.2本で,1試合当たりの得点が2.20点であったが,実際の1試合当たりの得点と回帰直線による子測値との差は,[7]点であった。

問4　鈴木さんは,さらに分析を進めるために,データシートを基に,決勝進出チームと予選敗退チームに分けて平均値や四分位数などの基本的な統計量を算出し,表2を作成した。

表２　１試合当たりのデータに関する基本的な統計量（分析シート）

	A	B	C	D	E	F	G	H	I
1		決勝進出チーム				予選敗退チーム			
2	統計量	１試合当たりの得点	１試合当たりのショートパス本数	１試合当たりのロングパス本数	１試合当たりの反則回数	１試合当たりの得点	１試合当たりのショートパス本数	１試合当たりのロングパス本数	１試合当たりの反則回数
3	合計	21.56	5532.21	1564.19	41.30	11.00	4213.33	1474.33	48.00
4	最小値	0.75	268.00	74.40	1.50	0.00	185.67	73.67	1.67
5	第１四分位数	1.00	321.82	92.25	2.10	0.33	235.25	87.67	2.58
6	第２四分位数	1.25	336.88	96.02	2.40	0.67	266.83	91.67	3.00
7	第３四分位数	1.75	368.33	103.50	3.00	1.00	300.08	98.00	3.42
8	最大値	2.25	453.50	118.40	4.50	1.67	334.00	109.33	4.67
9	分散	0.23	1926.74	137.79	0.67	0.15	1824.08	106.61	0.61
10	標準偏差	0.48	43.89	11.74	0.82	0.38	42.71	10.33	0.78
11	平均値	1.35	345.76	97.76	2.58	0.69	263.33	92.15	3.00

　次の中から正しいものを二つ選べ。[8] と [9]

⓪ 1試合当たりのロングバス本数のデータの散らばりを四分位範囲の視点で見ると,決勝進出チームよりも予選政退チームの方が小さい。

① 1試合当たりのショートパス本数は,決勝進出チームと予選敗退チームともに中央値より平均値の方が小さい。

② 1試合当たりのショートパス本数を見ると,決勝進出チームの第1四分位数は予選敗退チームの中央値より小さい。

③ 1試合当たりの反則回数の標準側差を比べると,決勝進出チームの方が予選敗退チームよりも散らばりが大きい。

④ 1試合当たりの反則回数の予選敗退チームの第1四分位数は,決勝進出チームの中央値より小さい。

問5　次の[10]に入る適切な文章を選び、[11] と [12]に入る数値を求めなさい。

　鈴木さんは,作成した図1と表2の両方から[10]ことに気づき,決勝進出の有無と1試合当たりの反則団数の関係に着目した。そこで,全参加チームにおける1試合当たりの反則回数の第1四分位数(Q1)未満のもの,第3四分位数(Q3)を超えるもの,Q1以上Q3以下の範囲のものの三つに分け,それと決勝進出の有無で,次の表3のクロス集計表に全参加チームを分類した。ただし,※の箇所は値を隠してある。この表から,決勝進出チームと予選敗退チナームの傾向が異なることに気づいた鈴木さんは,割合に着目してみようと考えた。決勝進出チームのうち1試合当たりの反則回数が全参加チームにおける第3内分位数を超えるチームの割合は19%であった。また,1試合当たりの反則回数がその第1四分位数より小さいチームの中で決勝進出したチームの割合は[12]％でぁった。

表３　決勝進出の有無と１試合当たりの反則回数に基づくクロス集計表

	１試合当たりの反則回数			
	Q1 未満	Q1 以上 Q3 以下	Q3 を超える	計
決勝進出チーム	※	※	※	16
予選敗退チーム	2	※	[11]	16
全参加チーム	8	※	7	32

[10]の選択肢

⓪ 1試合当たりの反則回数が最も多いチームは,決勝進出チームである

① 1試合当たりの反則回数と1試合当たりの得点の間には,全参加チームにおいて正の相関がある

② 1試合当たりの反則回数と①試合当たりの得点の間には,決勝進出チームと予選敗退チームのそれぞれで負の相関がある

③ 図1の❹のヒストダラムでは決勝進出チームの方が予選敗退チームより分布が左にすれている

解答

大問1

問1. [1] ①
問2. [2] ②
問3. [3] ②
問4. [4] 2 [5] ① [6] ①

大問2

[1] ④ [2] ⓪ [3] ⑤ [4] ② [5] ③ [6] ③ [7] ⓪ [8] ①

大問3

問1. [1] ②
問2. [2] ①
問3. [3] ②
問4. [4] ④
問5. [5] ① [6] ⓪ [7] ②

大問4

問1. [1] ③
問2. [2] ④
問3. [3] ③
問4. [4] ②

大問5

問1. [1] 480 [2] 640 [3] 40 [4] 4320 [5] 4360 [6] 200 [7] 4800
問2. [8] 80 [9] 25
問3. [10] × [11] ○ [12] ○ [13] × [14] ○

大問6

問1. [1] ⓪ [2] ③ ([1][2]順不同) [3] ③
問2. [4] ②
問3. [5] 0.16 [6] 0.04 [7] 0.56
問4 .[8] ⓪ [9] ③ ([8][9]順不同)
問5. [10] ③ [11] 4 [12] 75