MDBセミナー

統計分析力養成講座

データサイエンス研究所
伊藤嘉朗

本日の内容

- 統計学の考え方

  データのバラツキ、評価方法

- 実践的な相関分析

  相関係数、偏相関係数、散布図

- 機械学習のしくみと活用例

  教師あり学習、教師なし学習
  クラスター分析

- マーケティングデータサイエンス養成講座のご案内

ビジネスとデータサイエンス

ビジネスにデータサイエンスを活用する！

■ 自分の仕事にどう役立つのか分からない。
■ ビジネスに必要な範囲が分からない。

◆数式を理解する必要があるのか。
◆分かり易い参考書が無い。

統計学の考え方

英語の研修前後の成績（20人）

	A君の得点	全体の平均点	得点-平均点
研修前	70	58.3	+11.7
研修後	72	58.3	+13.7

A君の成績の評価は？

全員の成績

＜研修前＞

70、56、89、27、69、57、69
50、33、67、37、49、98、69
68、25、65、67、33、68

＜研修後＞

72、31、95、36、89、88、89
76、28、47、23、28、96、48
51、20、33、91、27、98

データを見る視点は？

全員の成績

＜研修前＞

70、56、89、27、69、57、69

50、33、67、37、49、98、69

68、25、65、67、33、68

＜研修後＞

72、31、95、36、89、88、89

76、28、47、23、28、96、48

51、20、33、91、27、98

平均の信頼性

（例）暑い日に暑い場所で待ち合わせをした。
いつも遅れてくる人が何分後に来るのか予測。

＜過去10回の遅刻データ＞

											平均
①	21	46	8	28	19	34	13	33	19	31	25.2分
②	25	26	23	24	26	27	26	25	24	26	25.2分

①、②それぞれにおける待ち時間の行動は？

<u>①のデータは「バラツキ」が大きい。</u>

「バラツキ」の大きいデータの平均は信頼できない。

A君の成績の評価

全体の平均点は同じ。
自分の得点は上がった。

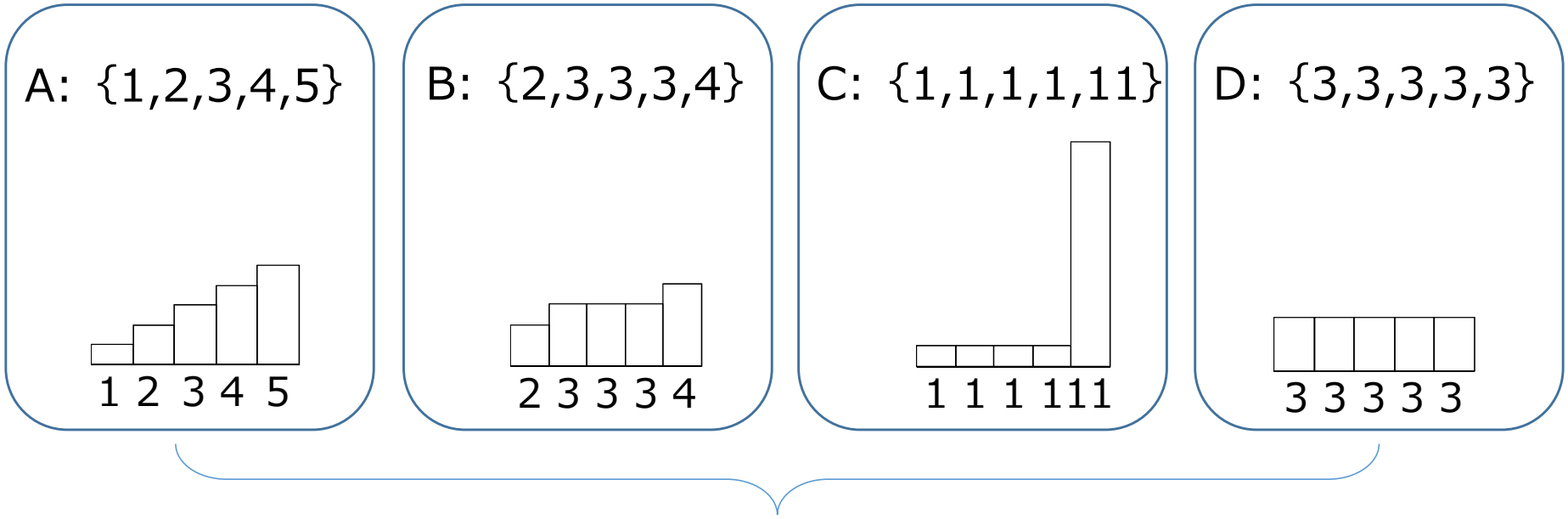
研修前　3番（20人中）
研修後　9番（20人中）

研修前より研修後のバラツキが大きい

成績を比較するためには各データ集団の
バラツキを表す代表値が必要！！

平均値が同じデータ群

A群～D群の平均値はすべて同じ値3



平均値は同じであるので同類集団？

各群のバラツキが異なる。　D<B<A<C

バラツキの計算方法の検討

A群のバラツキの部分を抽出する。

A群		平均値		値－平均値
1	－	3	＝	-2
2	－	3	＝	-1
3	－	3	＝	0
4	－	3	＝	1
5	－	3	＝	2
		計	＝	0

合計すると0！

バラツキの部分を2乗し、合計！

A群のバラツキの部分を2乗して合計

	値－平均値	（値－平均値）^{2}
	-2	4
	-1	1
	0	0
	1	1
	2	4
計	0	10

➡ A群のバラツキ＝10

同様に、B、C、D群について計算する。

B群	平均	差	$(差)^2$
2	3 =	-1	1
3	3 =	0	0
3	3 =	0	0
3	3 =	0	0
4	3 =	1	1
計		0	2

C群	平均	差	$(差)^2$
1	3 =	-2	4
1	3 =	-2	4
1	3 =	-2	4
1	3 =	-2	4
11	3 =	8	64
計		0	80

D群	平均	差	$(差)^2$
3	3 =	0	0
3	3 =	0	0
3	3 =	0	0
3	3 =	0	0
3	3 =	0	0
計		0	0

D<B<A<C　　0<2<10<80

「偏差平方和」

E群: {1,1,2,2,3,3,4,4,5,5} の偏差平方和

E群		平均値		差	$(差)^2$
1	-	3	=	-2	4
1	-	3	=	-2	4
2	-	3	=	-1	1
2	-	3	=	-1	1
3	-	3	=	0	0
3	-	3	=	0	0
4	-	3	=	1	1
4	-	3	=	1	1
5	-	3	=	2	4
5	-	3	=	2	4
		計		0	20

E群の偏差平方和=20

E群: {1,1,2,2,3,3,4,4,5,5} の偏差平方和＝20

A群: {1,2,3,4,5} の偏差平方和＝10

A群とE群の構造は同じであるが、
偏差平方和の値はE群の方が大きい。

偏差平方和をデータ数で割ると同じ代表値となる。

E群：20÷10＝2　　A群：10÷5＝2

バラツキの代表値①
「分散」

F群: {10,20,30,40,50}　　A群: {1,2,3,4,5} の10倍

F群の単位：千円　A群の単位：万円の場合、全く同じデータ。

偏差平方和を計算すると

F群	平均		差	$(差)^2$
10	− 30	=	-20	400
20	− 30	=	-10	100
30	− 30	=	0	0
40	− 30	=	10	100
50	− 30	=	20	400
計			0	1000

F群の偏差平方和： 1000

F群の分散： 200千円^2

A群の分散： 2万円^2

分散の値の比較は困難

分散の平方根は、同じ値となる。

F群：$\sqrt{200}$ ≒ 14.14千円

A群：$\sqrt{2}$ ≒ 1.414万円

⇒ バラツキの代表値②「標準偏差」

EXCEL

VAR.P（分散）

STDEV.P（標準偏差）

B24 =STDEV.P(B2:B21)

	A	B	C	D	E	F
1		研修前	研修後			
2		70	72			
3		56	31			
4		89	95			
5		27	36			
6		69	89			
7		57	88			
8		69	89			
9		50	76			
10		33	28			
11		67	47			
12		37	23			
13		49	28			
14		98	96			
15		69	48			
16		68	51			
17		25	20			
18		65	33			
19		67	91			
20		33	27			
21		68	98			
22	平均値	58.3	58.3			
23	分散	368.4	819.0			
24	標準偏差	19.2	28.6			
25						
26						
27						

A君の成績の評価

	成績	平均	成績－平均	標準偏差
研修前	70	58.3	11.7	19.2
研修後	72	58.3	13.7	28.6

研修前　　研修後

$$\frac{70-58.3}{19.2} = 0.609 > \frac{72-58.3}{28.6} = 0.479$$

$$\frac{\text{成績と平均値の差}}{\text{標準偏差}}$$ ⇨ Z値

Ｚ値の比較によりデータの評価・比較が可能

Ｚ値×10＋50

偏差値

研修前の偏差値＝0.609×10＋50＝56.09

研修後の偏差値＝0.479×10＋50＝54.79

Ｚ値と偏差値

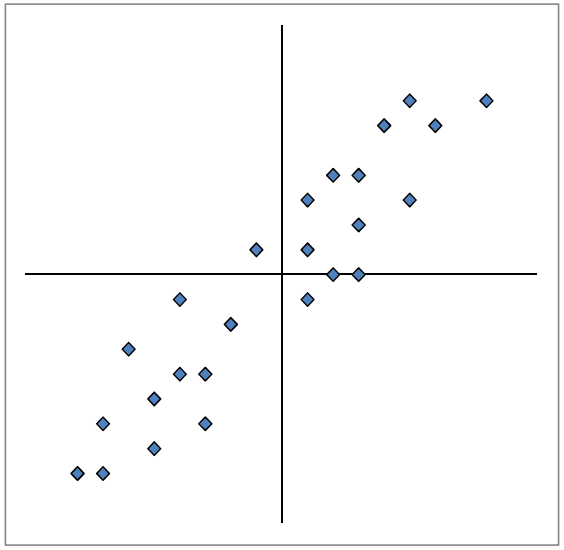
Ｚ値 ―――――――――― 0 ――――――――――
（－）　　（＋）

偏差値 ―――――――――― 50 ――――――――――

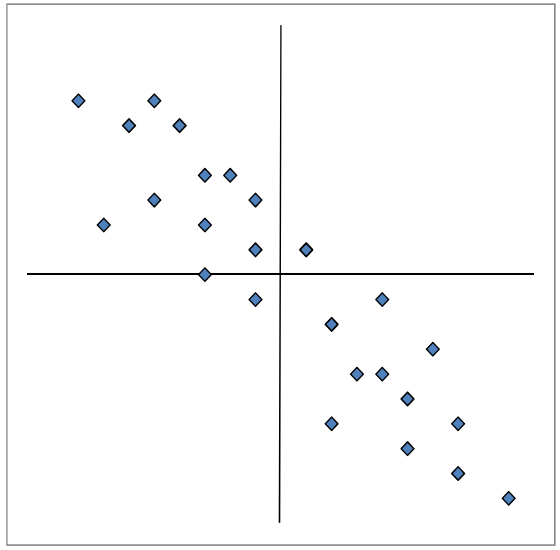
Ｚ値は、平均値＝0
平均値より大のときプラスの値、小のときマイナスの値

偏差値は、平均値＝50
平均値より大のとき50より大きな値、平均値より小のとき50より小さな値

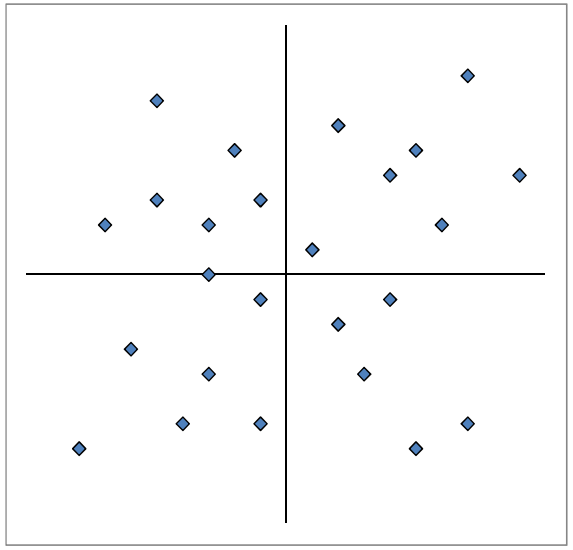
相関関係



正の相関



負の相関

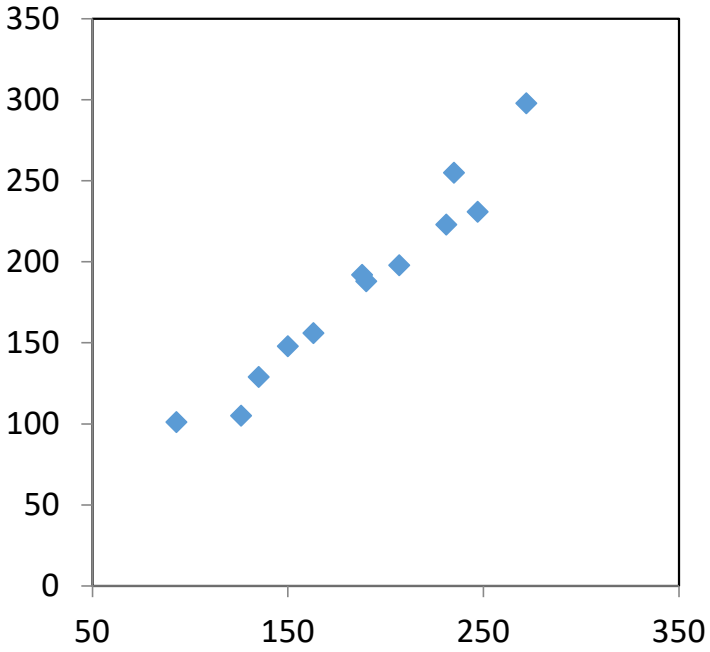


無相関

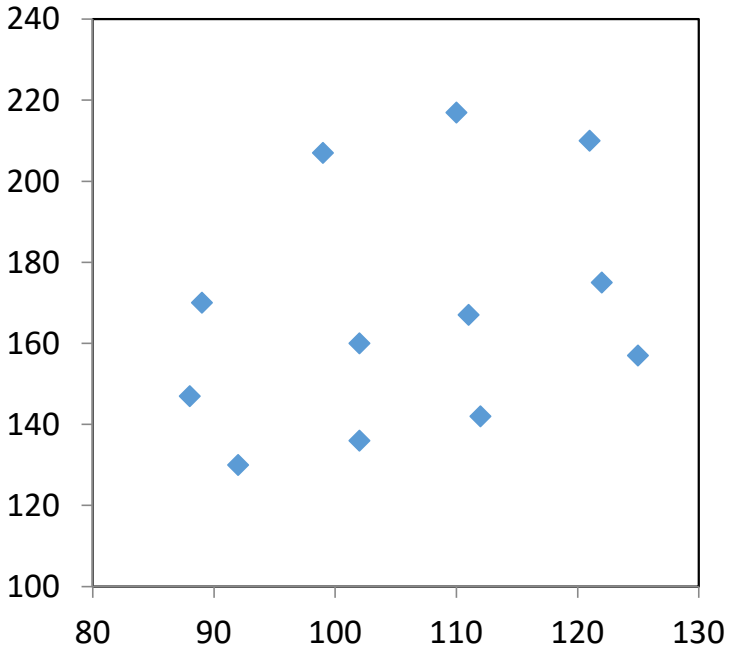
相関関係には正の相関、負の相関、無相関。
点の集中度が関係の強さを測定する手がかり。

相関係数（ｒ）

$$-1 \leqq r \leqq 1$$



ｒ＝0.97



ｒ＝0.32

相関係数（ｒ）は相関関係の強さ

相関係数（ｒ）の解釈

ｒ＝0　⇒　説明力は0%

ｒ＝1　⇒　説明力は100%

ｒ＝0.32　⇒　説明力は？%

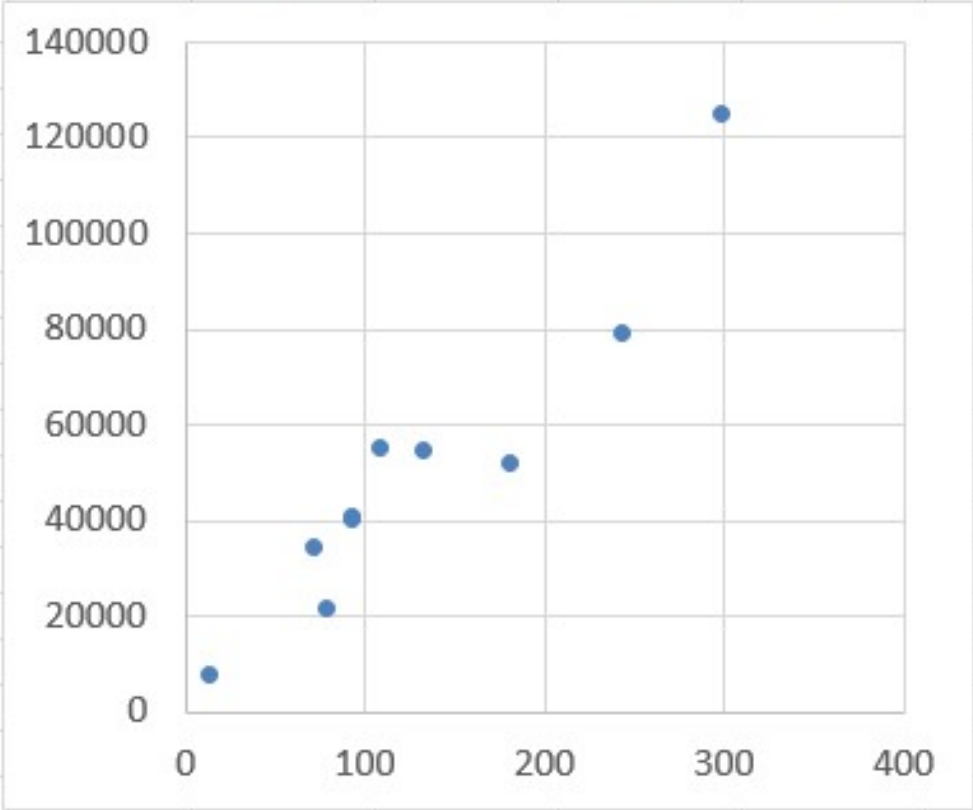
0.32×0.32＝0.1024
説明力は、10.24%

相関係数について

EXCEL ： CORREL

B13　=CORREL(A2:A11,B2:B11)

	A	B
1	DM発送数	売上高
2	92	40212
3	93	40801
4	298	125198
5	78	21772
6	181	52122
7	108	55200
8	133	54801
9	72	34411
10	243	79221
11	13	7801
12		
13	相関係数	0.9492499

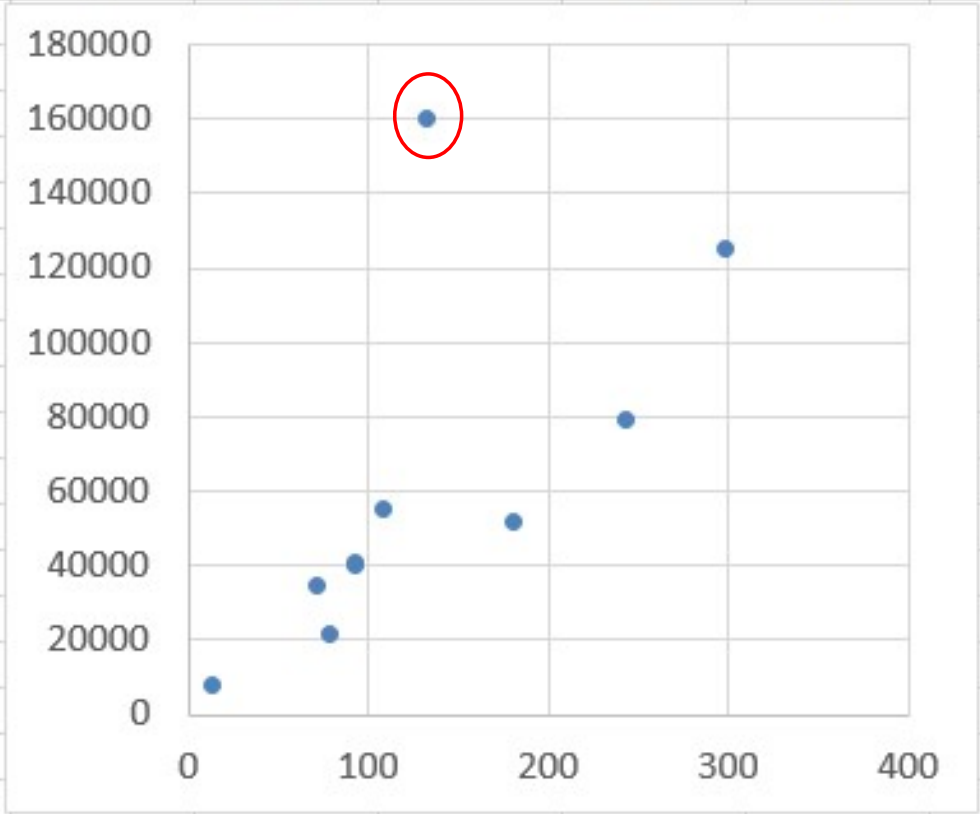


r ＝0.949

相関係数について（外れ値を含む場合）

B13　=CORREL(A2:A11,B2:B11)

	A	B
1	DM発送数	売上高
2	92	40212
3	93	40801
4	298	125198
5	78	21772
6	181	52122
7	108	55200
8	133	160000
9	72	34411
10	243	79221
11	13	7801
12		
13	相関係数	0.656717



r ＝0.657　　外れ値の影響を受ける

◇都道府県別広告費と売上高実績

都道府県	売上高	広告費
北海道	245	26
青森	123	14
・・・	・・・	・・・
	・・・	・・・
東京	5,672	492
・・・	・・・	・・・
・・・	・・・	・・・
沖縄	59	6

地域別の売上高と広告費の関係？

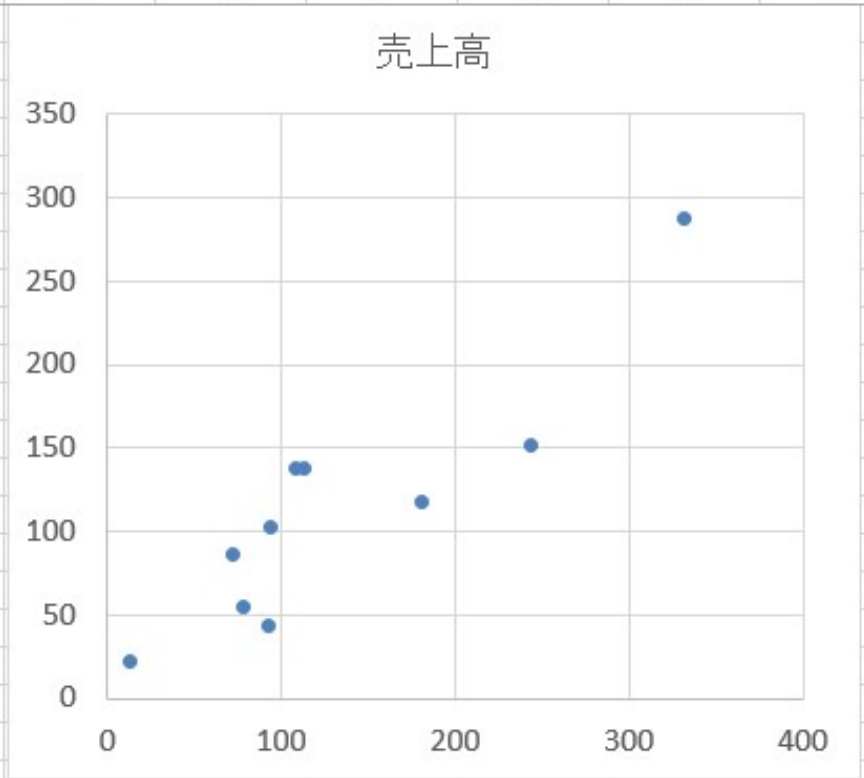
◇支店別広告費と売上高

支店	広告費	売上高
北海道	92	44
東北	93	102
関東	332	288
北陸	78	54
中部	181	118
近畿	108	138
中国	113	138
四国	72	86
九州	243	152
沖縄	13	22

相関係数

B13 =CORREL(A2:A11,B2:B11)

	A	B
1	広告費	売上高
2	92	44
3	93	102
4	332	288
5	78	54
6	181	118
7	108	138
8	113	138
9	72	86
10	243	152
11	13	22
12		
13	相関係数	0.902405



r ＝0.902

広告宣伝費は売上高に貢献？

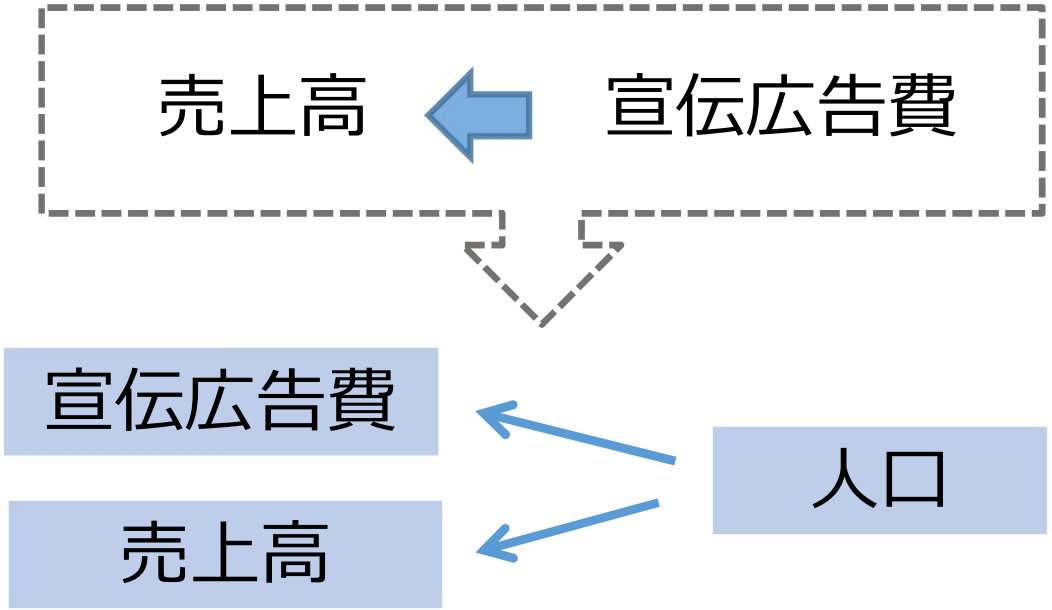
◇支店別広告費、売上高、人口

支店	広告費	売上高	人口
北海道	92	44	5,506
東北	93	102	9,335
関東	332	288	42,604
北陸	78	54	5,443
中部	181	118	18,127
近畿	108	138	12,912
中国	113	138	15,554
四国	72	86	3,976
九州	243	152	13,204
沖縄	13	22	1,393

相関係数

広告費と売上高　r ＝0.902

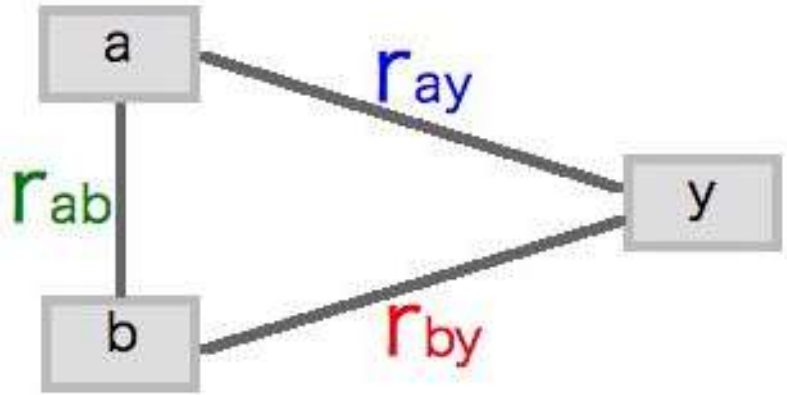
人口と広告費　r ＝0.895
人口と売上高　r ＝0.951



人口は交絡要因　⇒　疑似相関

人口の影響を除いたときの広告費と売上高の相関係数　⇨　偏相関係数

偏相関係数の求め方（EXCEL）



aの影響を除いた、bとyの偏相関係数

$$\frac{r_{by}-(r_{ay}\times r_{ab})}{\sqrt{1-r_{ay}^2}\times\sqrt{1-r_{ab}^2}}$$

偏相関係数の計算

$$\frac{0.902 - (0.951 \times 0.895)}{\sqrt{1 - 0.951^2} \times \sqrt{1 - 0.895^2}}$$

$$= 0.368$$

人口の影響を除いたときの広告費と売上高の
相関係数　：　0.368

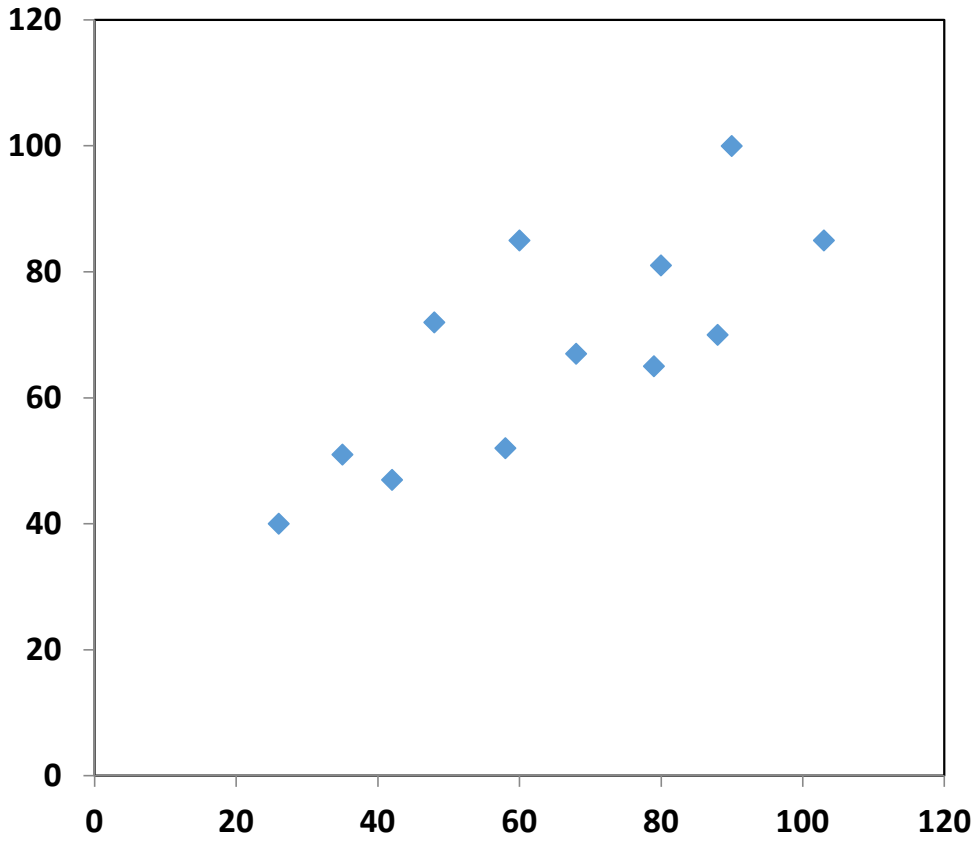
A支店の広告費と売上高

月	4	5	6	7	8	9	10	11	12	1	2	3
広告費	42	26	35	68	103	90	88	58	79	80	48	60
売上高	47	40	51	67	85	100	70	52	65	81	72	85

r＝0.775

⇩

強い正の相関



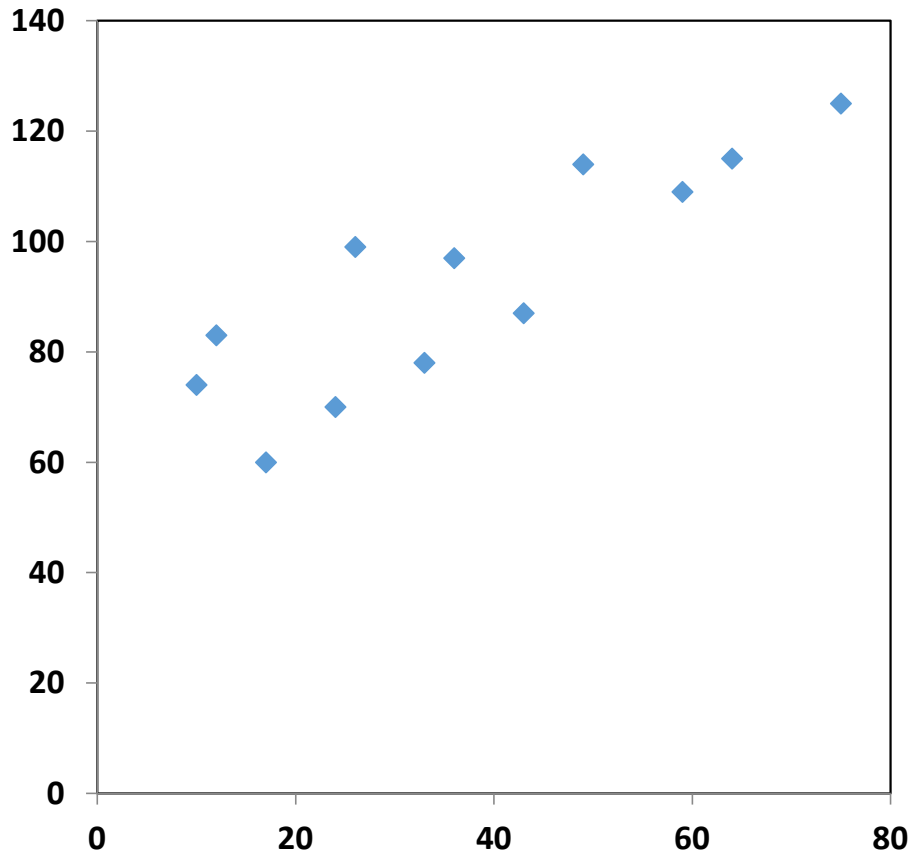
B支店の広告費と売上高

月	4	5	6	7	8	9	10	11	12	1	2	3
広告費	12	10	17	59	75	64	33	24	49	43	36	26
売上高	83	74	60	109	125	115	78	70	114	87	97	99

r＝0.863

⇩

強い正の相関

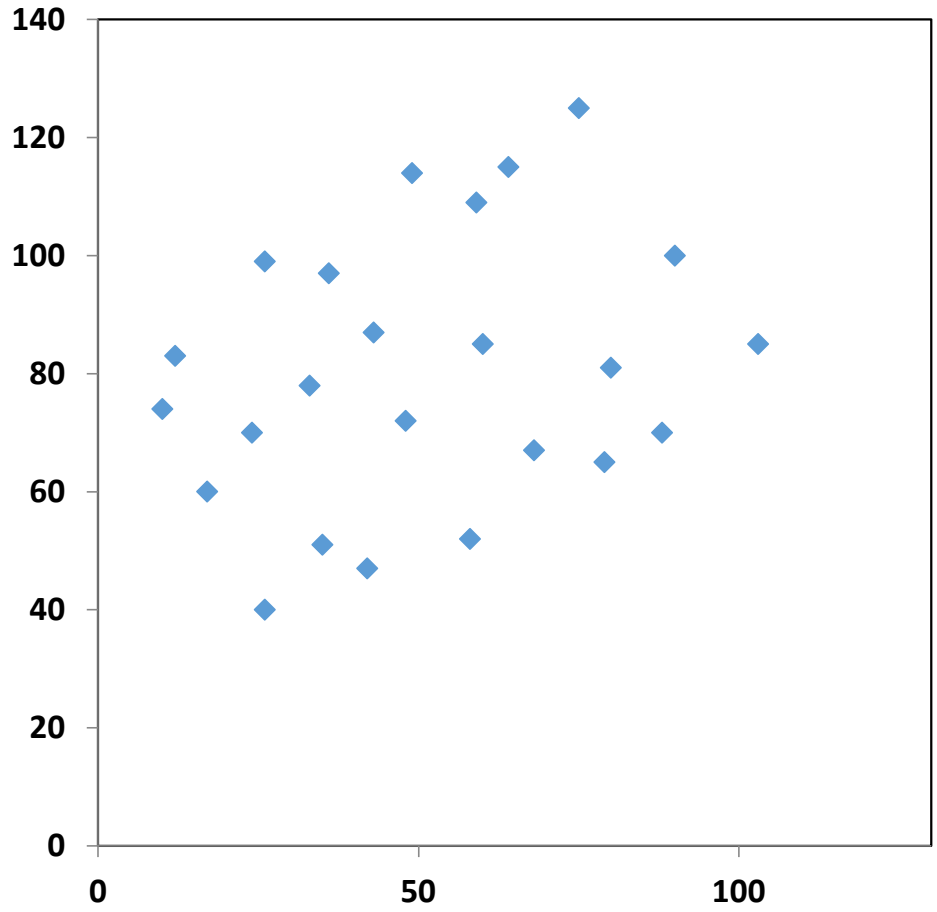


（A支店＋B支店）の広告費と売上高

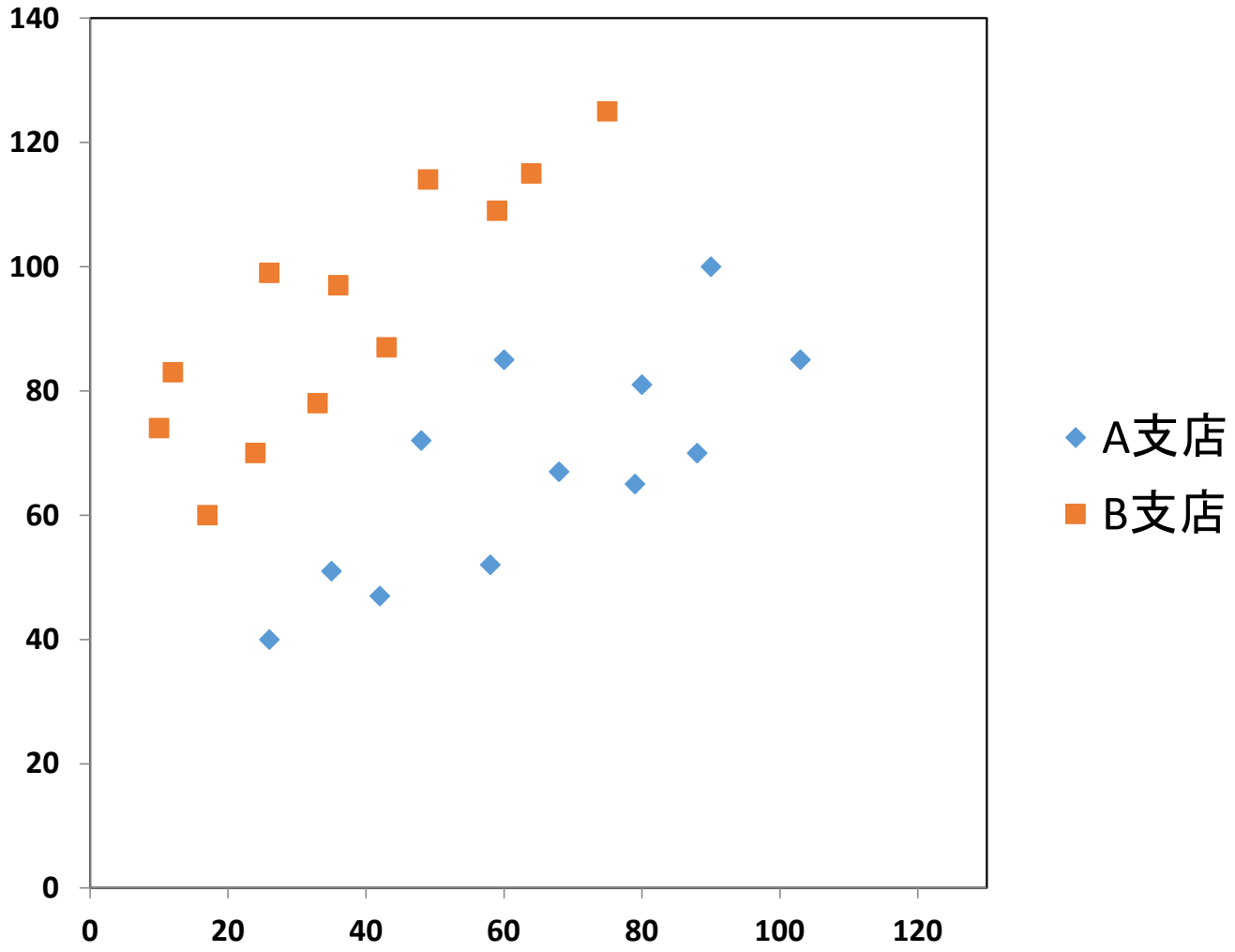
r＝0.272

⇩

???



（A支店＋B支店）の広告費と売上高



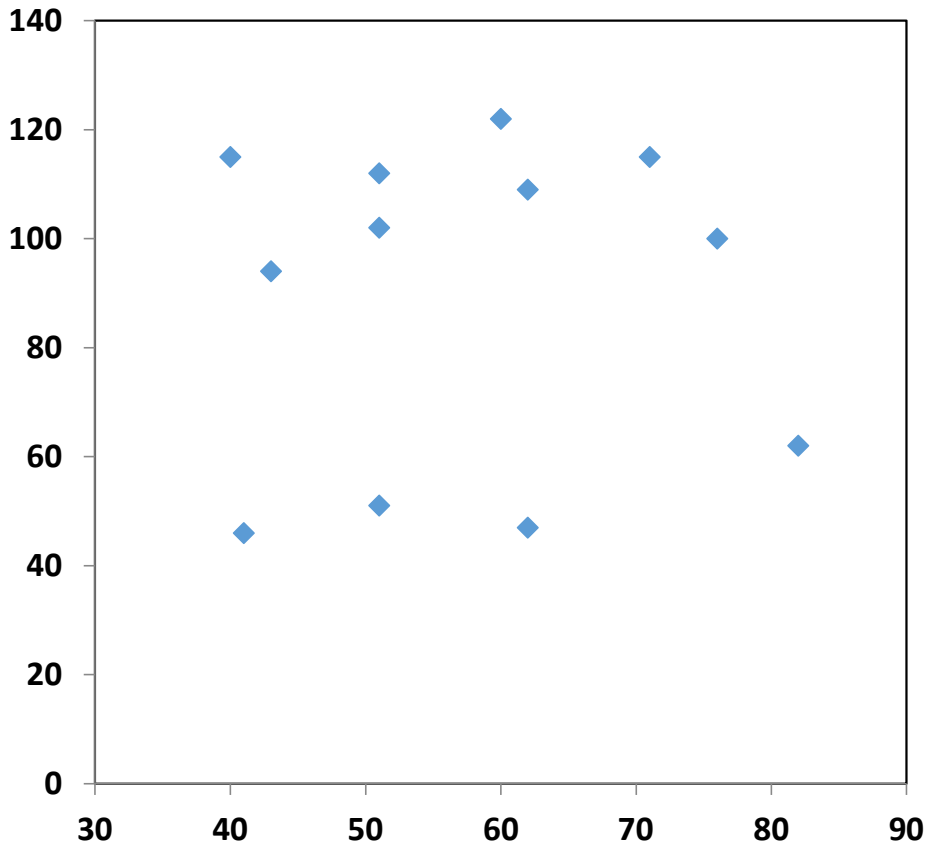
A支店の広告費と売上高

月	4	5	6	7	8	9	10	11	12	1	2	3
広告費	43	51	62	76	41	51	62	82	40	51	60	71
売上高	94	102	109	100	46	51	47	62	115	112	122	115

r＝0.016

⇩

無相関



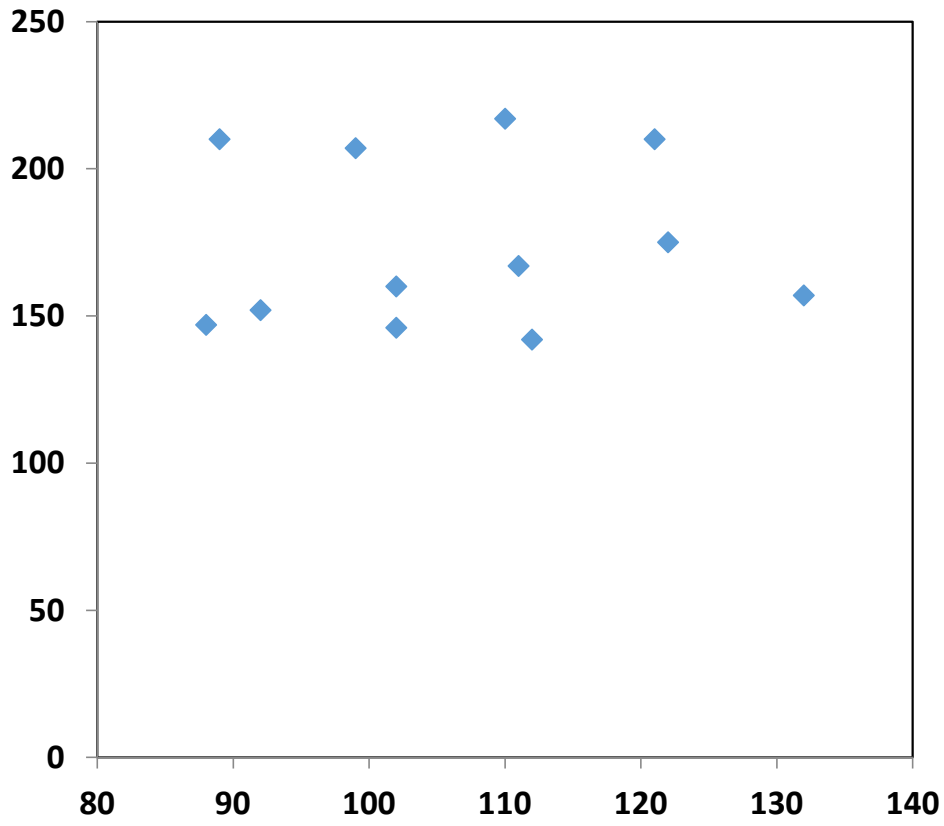
B支店の広告費と売上高

月	4	5	6	7	8	9	10	11	12	1	2	3
広告費	92	102	111	122	88	102	112	132	89	99	110	121
売上高	152	160	167	175	147	146	142	157	210	207	217	210

r＝0.040

⇩

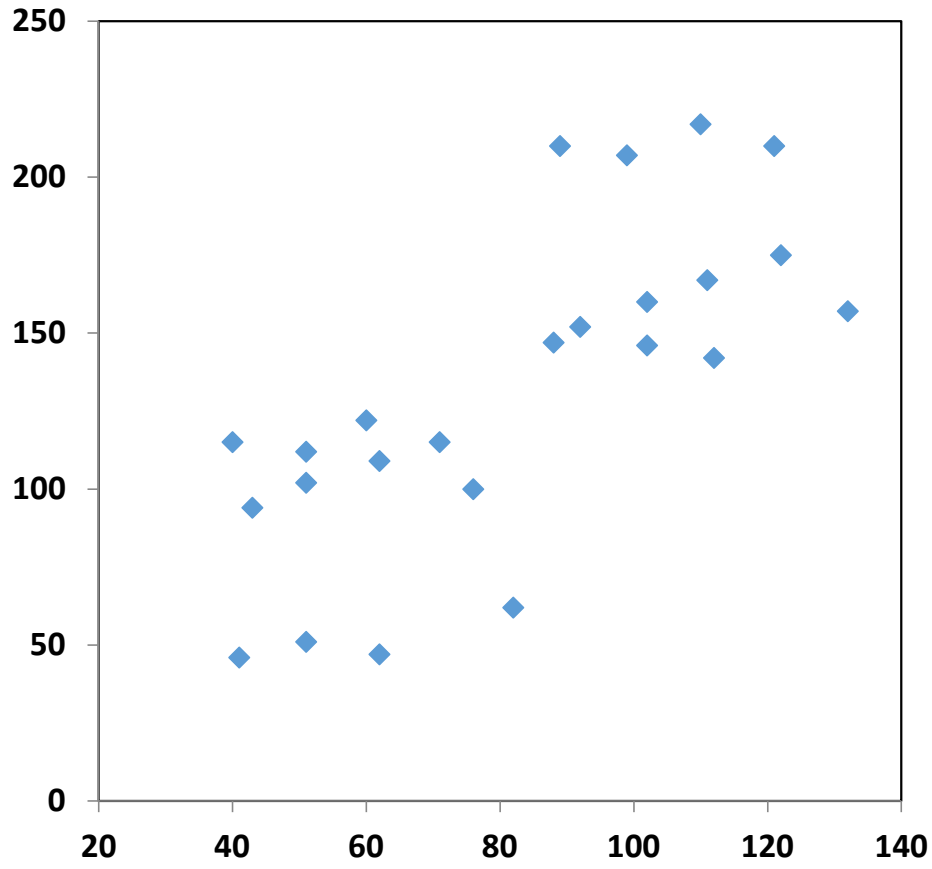
無相関



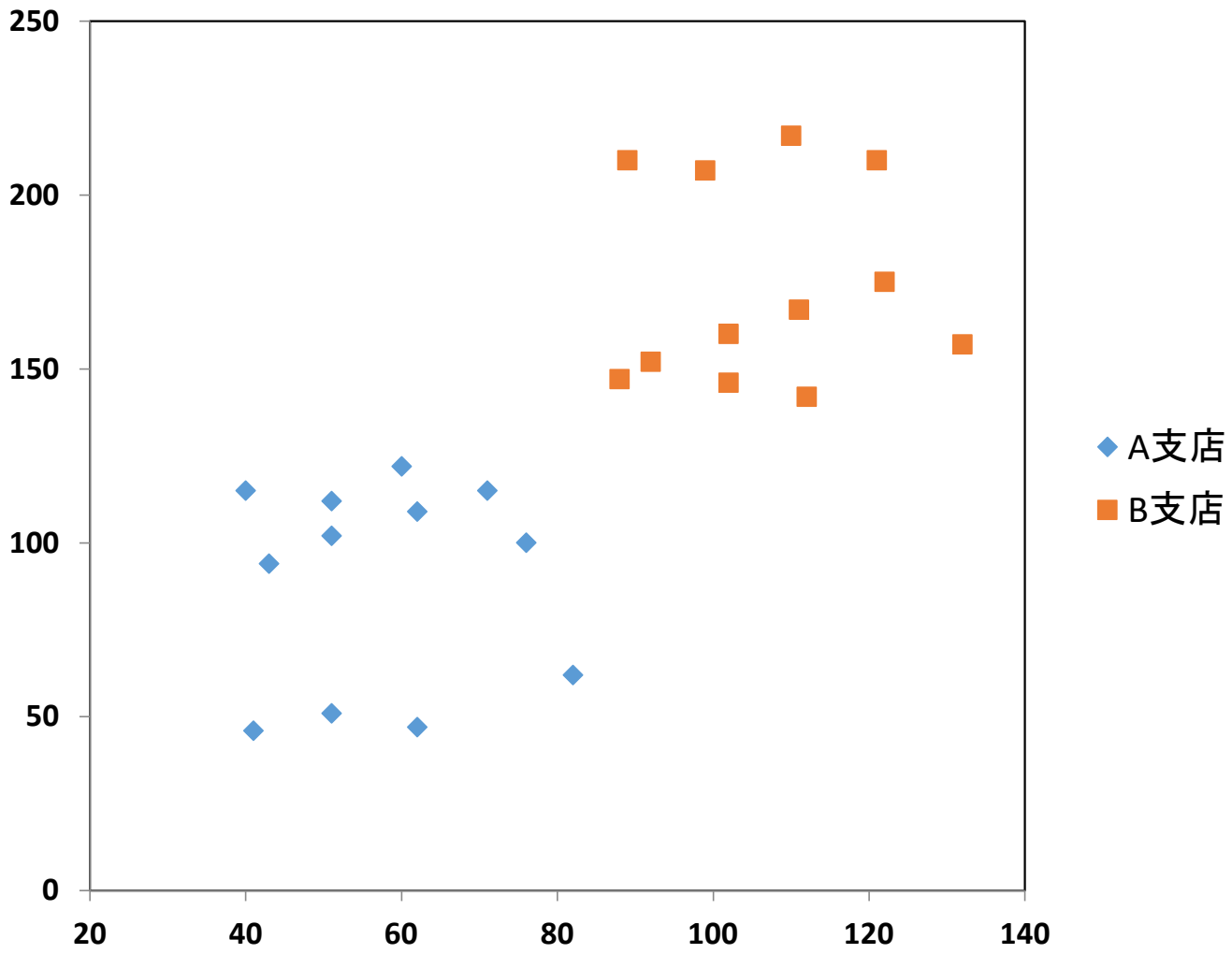
（A支店＋B支店）の広告費と売上高

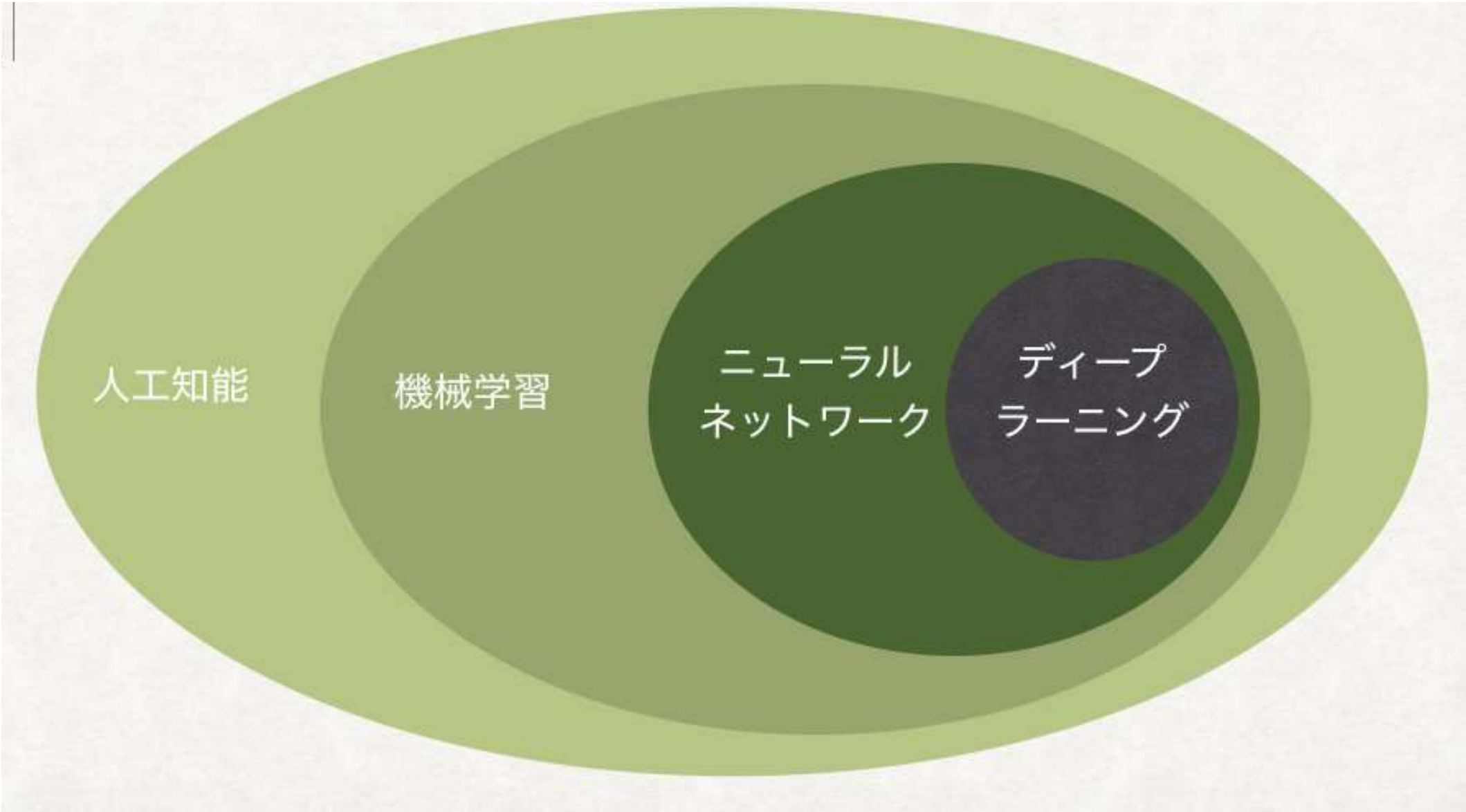
r＝0.823

???



（A支店＋B支店）の広告費と売上高




人工知能
機械学習
ニューラル
ネットワーク
ディープ
ラーニング

機械学習とは

◇コンピューターやロボットなどの機械に自動的に概念や行動プログラムを学習させる研究分野。
（世界大百科辞典）

◇コンピューターによる学習。人工知能の一分野であり、人間がもつ学習能力と同じく、コンピューターも経験から学習し、将来予測や意思決定を実現できるようにする技術や手法を指す。（大辞泉）

(1)分類方法は？

A	B
うさぎ	ねずみ
はがねのつるぎ	てつのおの
きんとき	あんこ
たわし	わたし

では…

- 「ししゃも」は？
- 「ほっけ」は？
- 「しゃけ」は？

A	B
うさぎ	ねずみ
はがねのつるぎ	てつのおの
きんとき	あんこ
たわし	わたし

- 「ししゃも」は？ ⇒ A
- 「ほっけ」は？ ⇒ B
- 「しゃけ」は？ ⇒ B

（2）仲間はずれは？

くも
やどかり
たこ
いか
たらばがに
毛がに
えび

「足の数」と「かたさ」で分類すると…

		足の数	
		8本	10本
かたさ	やわらかい	くも たこ	いか
	かたい	たらばがに やどかり	毛ガニ えび

◆（1）分類の場合：

・区別するルールを与えられた事例から見つける

・未知の対象に対してルールを適用し分類する

→ 教師（学習データ）あり

◆（2）仲間はずれ探しの場合：

・ある視点から対象をグループ分けする

・それぞれのメンバーを評価

→ 教師なし

【機械学習の2大タスク】

「教師あり学習」＝予測

「教師なし学習」＝発見

教師あり

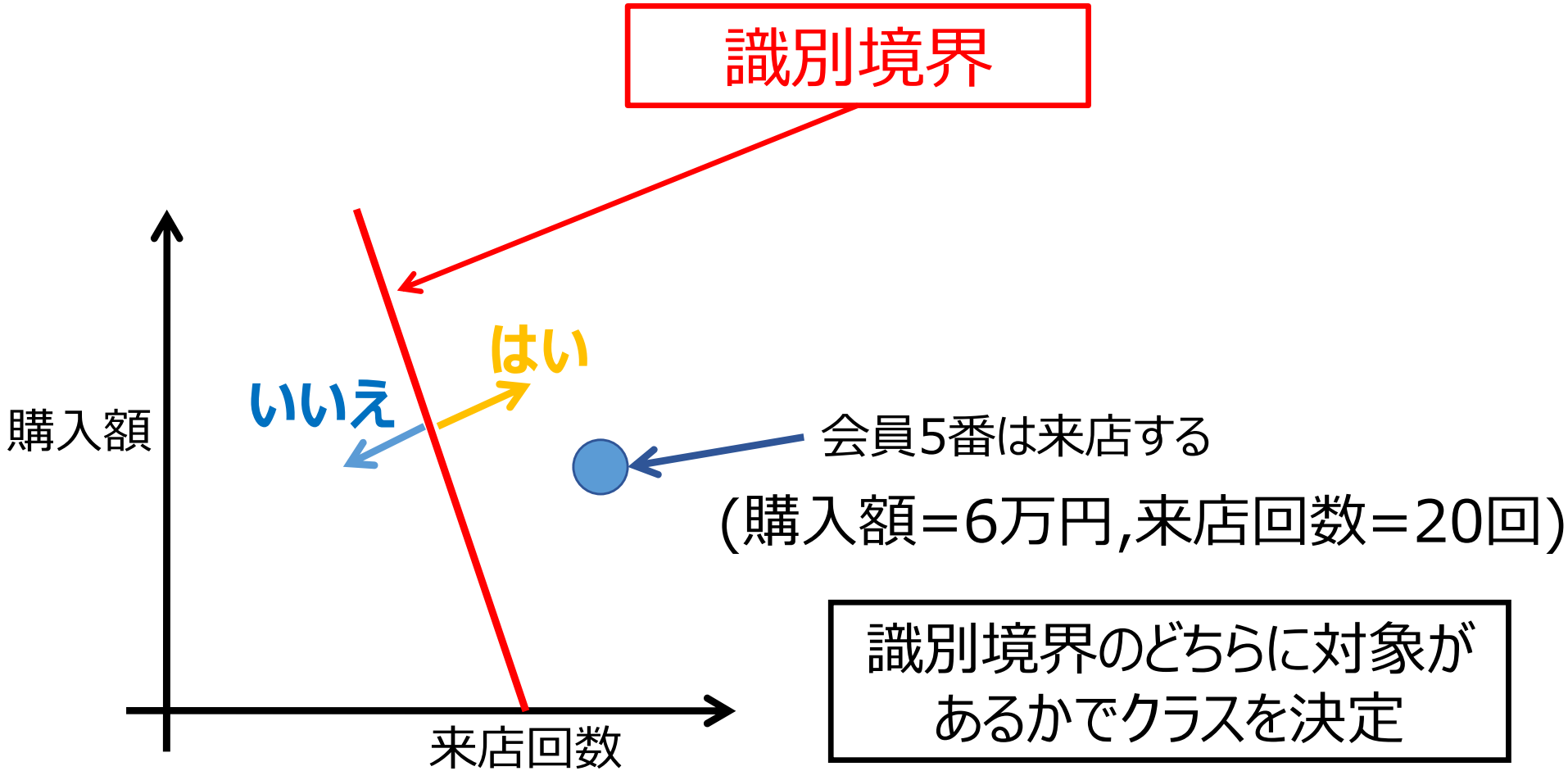
（例）ポイント会員にバーゲンの案内状を送る。
バーゲンに来ない客に送ると余分な郵送費がかかる。
できるだけ来店しそうな顧客を選びたい。

会員のバーゲンに来る可能性に影響しそうな要因を選ぶ

先月の購入額
先月の来店回数
店舗から住所までの距離

会員3番 ：〈5万円、20回、1.5km〉

会員3（はい,〈5万円、20回、1.5km〉）

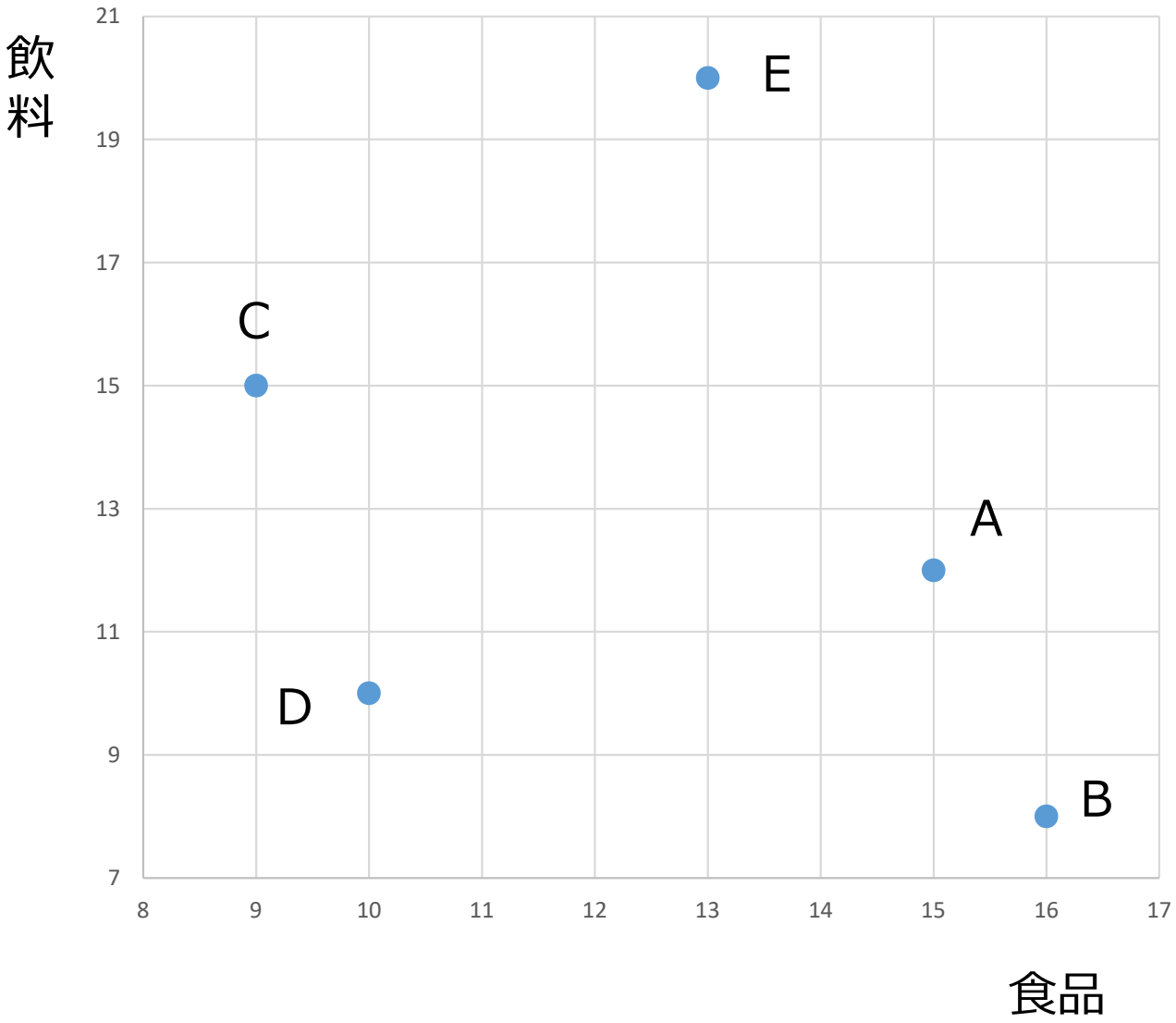

識別境界
はい
いいえ
購入額
会員5番は来店する
(購入額=6万円,来店回数=20回)
来店回数
識別境界のどちらに対象が
あるかでクラスを決定

教師なし

階層的クラスター分析

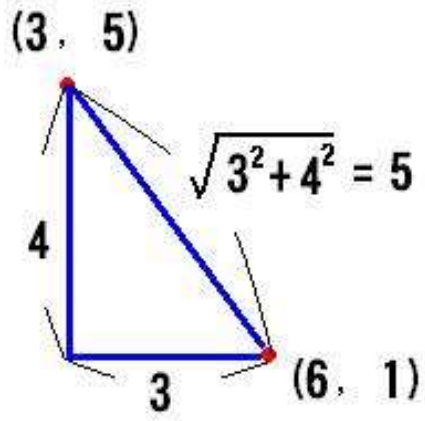
◇AからEの顧客の購入金額
クラスター分析により分類する。

顧客	食品	飲料
A	15	12
B	16	8
C	9	15
D	10	10
E	13	20


飲料
21
19
17
15
13
11
9
7
8
9
10
11
12
13
14
15
16
17
E
C
A
D
B
食品

距離の計算法（例）

・2点の距離は三平方の定理で計算



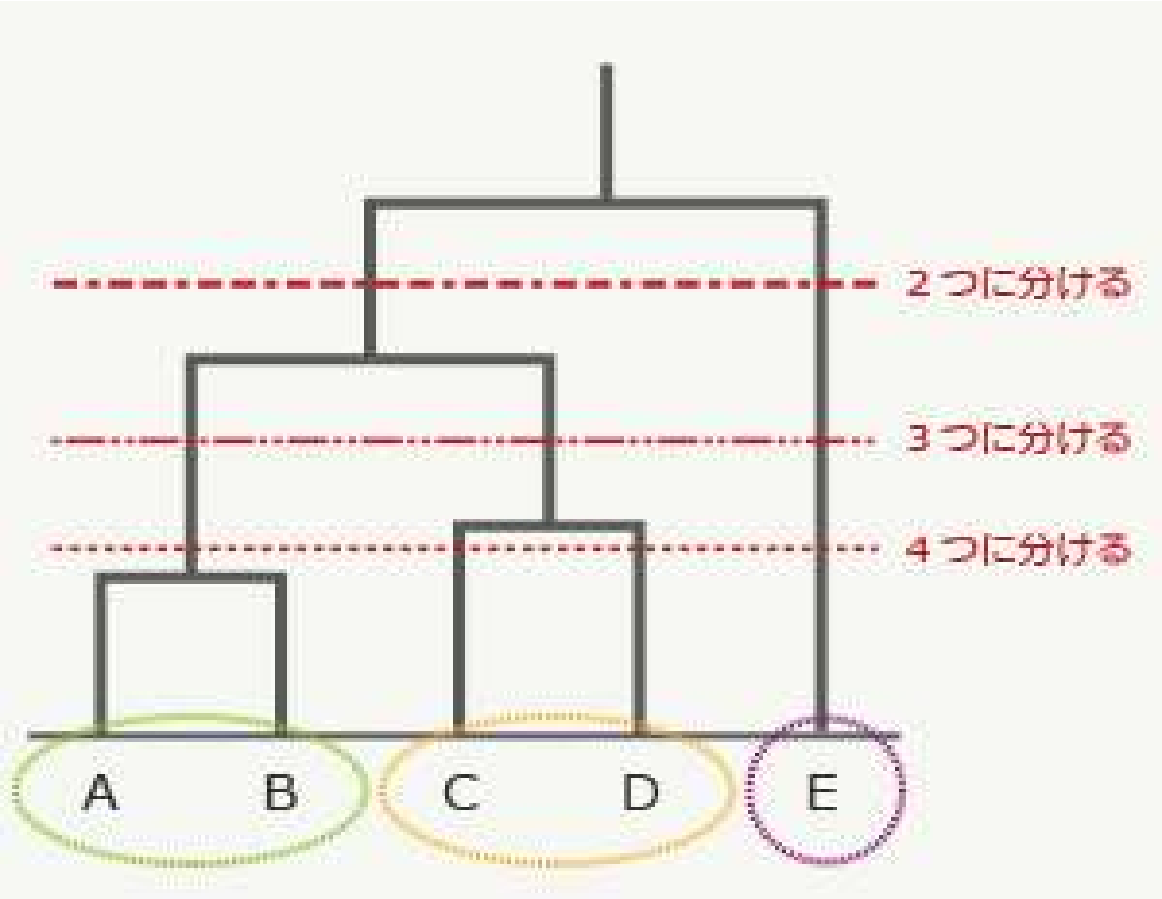
・クラスター分析は複数の点の集合同士の距離を計算する必要。

＜5つの距離の計算＞

	A	B	C	D
B	4.12			
C	5.39	6.32		
D	6.71	9.90	5.10	
E	8.25	12.37	10.44	6.40

一番距離が短いのはAとBの4.12であるので、
この2つをまとめて1つのクラスターにする。

樹形図



最も似ている顧客から順にクラスターにまとめる

クラスタの例

第1クラスタ：主に昼食品を購入する顧客
第2クラスタ：主に日用品を購入する顧客

クラスタの利用例

第1クラスタの顧客はその場で購入を決めると予想
　⇒ セットメニューによる客単価の向上
第2クラスタの顧客はまとめ買いをすると予想
　⇒ DMは、第2クラスタの顧客を中心に送る

相関の発見と利用例　Association Rule

{インスタントラーメン} ⇒ {チャーシュー}

インスタントラーメン売り場にチャーシューを並べると
同時購入が増えると予想される。

{牛乳} ⇒ {パン}

牛乳の特売をすると牛乳の販売数が増え、パンの販売数も
増えると予想されるのでパンの仕入を増やす。

{レトルトカレー}⇒{洗濯用洗剤}

✕ カレーを食べると服が汚れるから

○ たまたま，カレーと洗剤を同じ日に特売した

見つけた知識の合理性は必ず検証

分析から得た知識は、無条件に受け入れずに必ず検証し、
より妥当な知識を見つけるよう試行錯誤

データを活用するために

・正しい分析方法を習得する。
　分析手法：統計分析、機械学習等
・データの吟味が重要。（外れ値、カテゴリ）

・数字だけを分析していても、正確に把握することは困難。
・データを活用する上で定性的なデータも非常に重要。

・目的や仮説を持ち、擬似相関ではなく
因果関係 を見極める。

第四次産業革命スキル習得講座

分野別データサイエンティスト養成講座

（マーケティング、マニュファクチュアリング、メディカル）

株式会社データサイエンス研究所

■経済産業省：第四次産業革命スキル習得講座

IT・データを中心とした将来の成長が強く見込まれる分野において、社会人が高度な専門性を身に付けてキャリアアップを図る、専的・実践的な教育訓練講座を、経済産業大臣が認定する制度

◎厚生労働省

【個人向け助成金】 専門実践教育訓練指定講座

助成額 70%（自己負担30%）

【事業主向け助成金】 2022年4月開始

「高度デジタル人材訓練助成金」対象講座

訓練経費助成額：60%～75%

賃金助成額 ：480円～960円/時

https://www.mhlw.go.jp/content/11600000/000807259.pdf

■分野別データサイエンティスト養成講座の受講対象

・データサイエンティストを目指す方
・分析実務担当者
・実践力・問題解決力を習得したい方

・分析初心者の方
・プログラミング未経験の方

・研修室受講、オンライン受講、アーカイブ受講
・平日昼間、平日夜間、土曜日開講

■分野別データサイエンティスト養成講座の特徴

◎実践力（課題解決力）の養成

① データの取捨選択、特性を把握した分析手法の選択
② 各分析手法のしくみの理解
③ 分析結果を正しく解釈、分析目的との整合

分野別データの特性に合わせたトレーニング

マーケティングデータサイエンティスト養成講座

マーケター必須のデータサイエンスを実践的・効率的に習得

マーケティング分野において用いられる解析手法のしくみ・活用方法の理解と共に、マーケティング分野の特性に合わせたデータを用いた具体的演習により、実践的解析力、問題解決力の習得を目指す。

- ■研修時間　54時間（4カ月）
- ■受講費用　50万円（税込）

昼間コース（９：30～17：30）　6時間×　9日＝54時間
夜間コース（18：30～21：30）　3時間×18日＝54時間
＊オンライン、アーカイブによる受講が可能です。

■修了認定要件

修了認定要件

要件 1

全科目受講
（各科目 90%以上出席）

要件 2

総合テスト 70 点以上

要件 3

課題解決プレゼンテーションに合格

合格の場合

修了認定証を発行します。

＊統計検定 2 級取得を無料サポートします。

不合格の場合

1. 受講科目不足の場合は、オンラインによるアーカイブ受講をして頂きます。
2. 総合テスト不合格の場合は、不合格分野の再受講をお勧めし、再試験を受けて頂きます。
3. 課題解決プレゼンテーションに不合格の場合は、課題レポートを提出して頂きます。
4. 総合的に審査し、合格と認められた場合は修了認定証を発行します。

【受講お申込み】

https://datascience.co.jp/reskill/

■ご質問は下記にお願い致します。

que@datascience.co.jp

■■　**分析力養成講座**　■■
＝＝＝