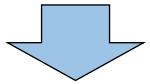
初心者のためのデータ分析法入門

データサイエンス研究所

ビジネスと統計学

ビジネスに統計学を活用する！



■ 統計学が自分の仕事にどう役立つのか分からない。
■ ビジネスに必要な統計学の範囲が分からない。

◆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

＜データの評価方法＞

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？

①のデータは「バラツキ」が大きい。

「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）

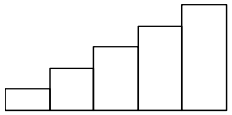
研修前より研修後のバラツキが大きい

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！！

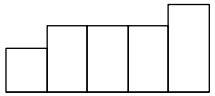
平均値が同じデータ群

A群～D群の平均値はすべて同じ値3

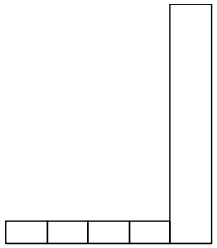
A: {1,2,3,4,5}



B: {2,3,3,3,4}



C: {1,1,1,1,11}



D: {3,3,3,3,3}

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

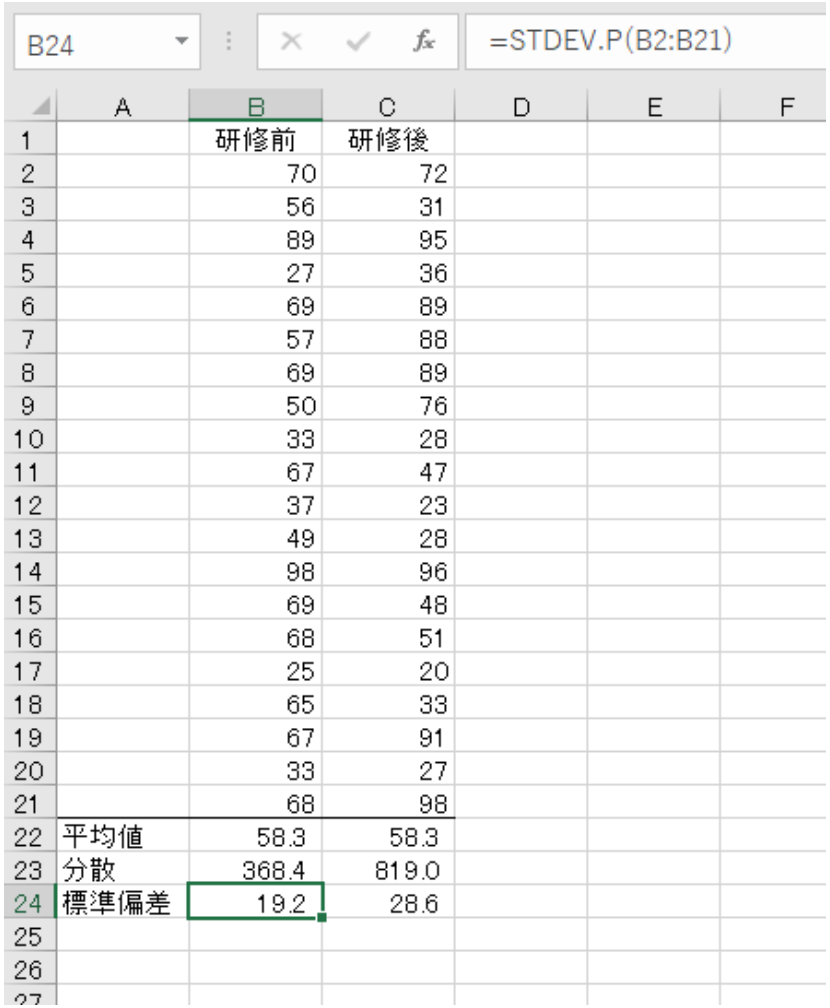
A群：$\sqrt{2}$ ≒ 1.414万円

⇒ バラツキの代表値②「標準偏差」

EXCEL

VAR.P（分散）

STDEV.P（標準偏差）



	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						

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

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

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

Z値の比較によりデータの評価・比較が可能

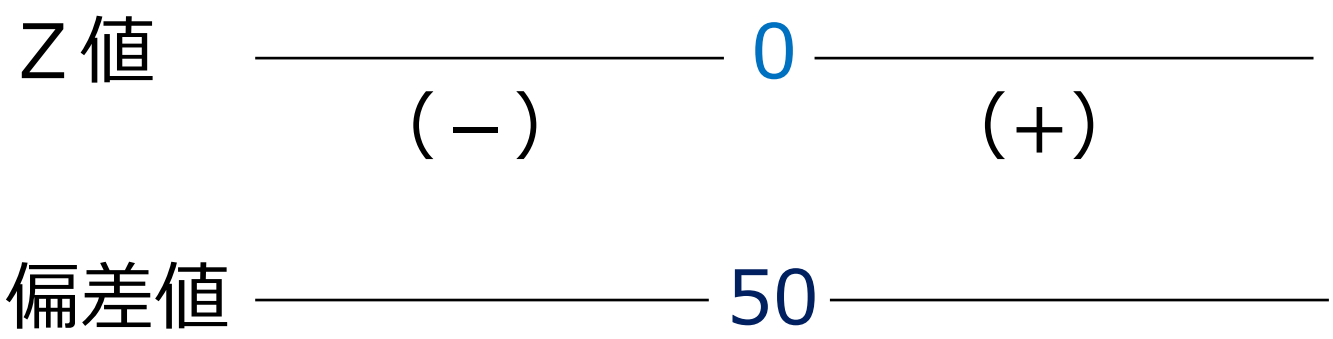
Z値×10＋50

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



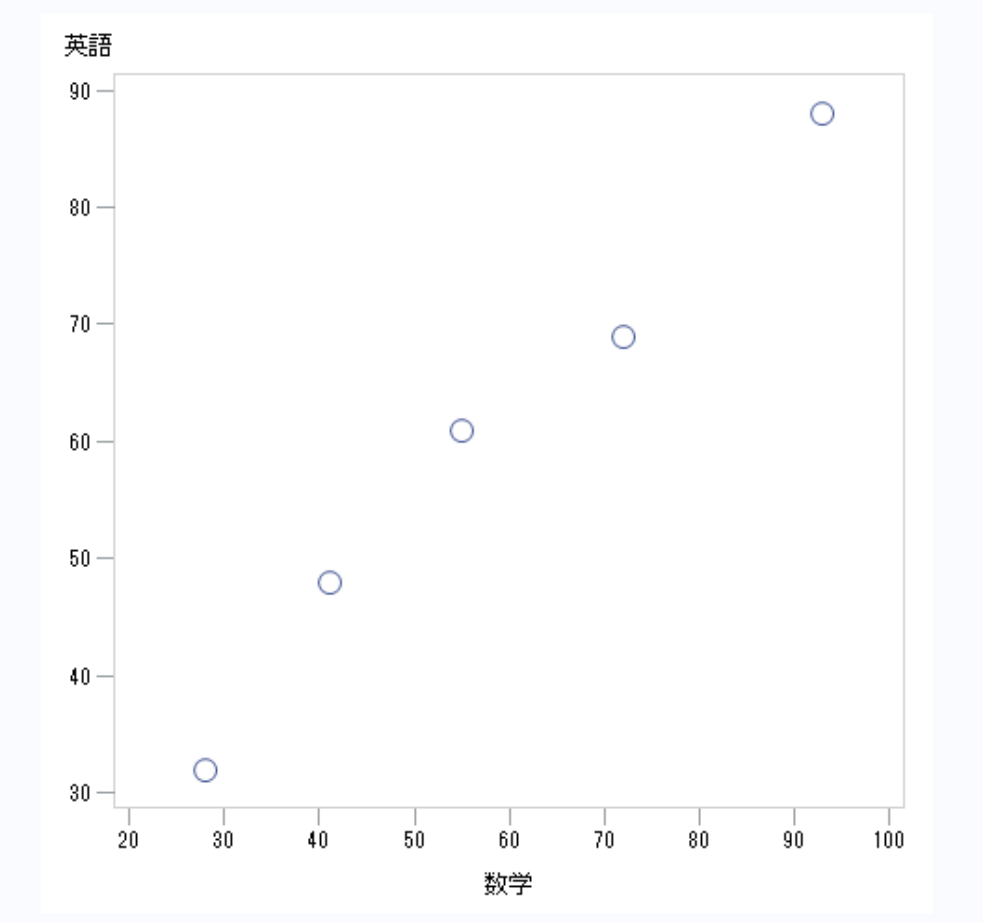
Z値は、平均値＝0
平均値より大のときプラスの値、小のときマイナスの値

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

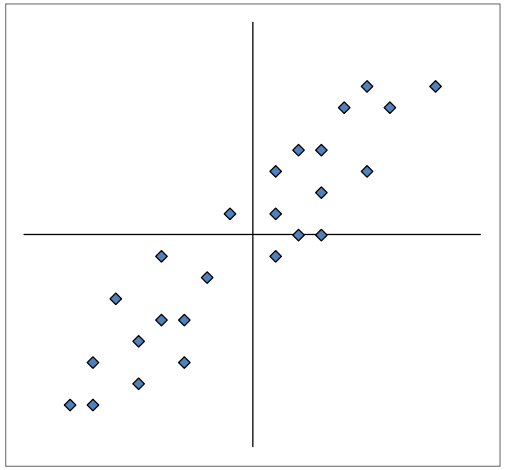
2群の関係を調べる（相関分析）

数学	英語
28	32
55	61
93	88
72	69
41	48

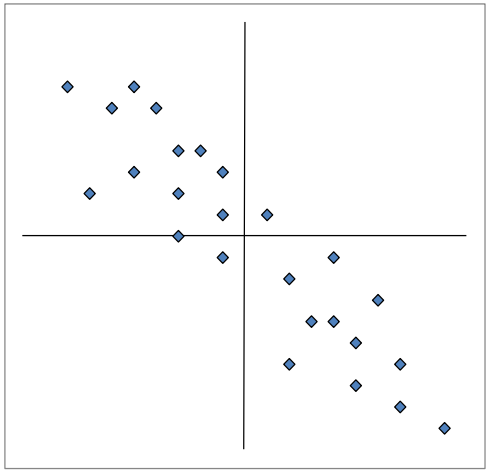
散布図



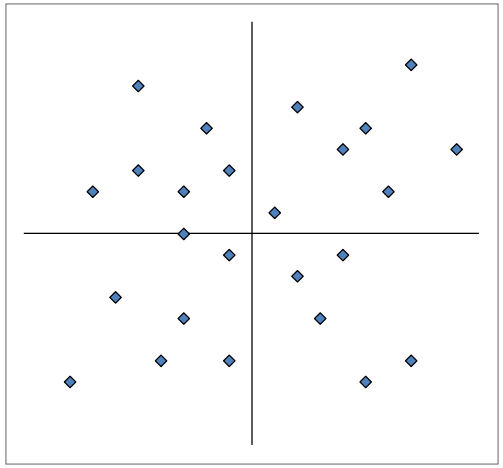
相関分析



正の相関



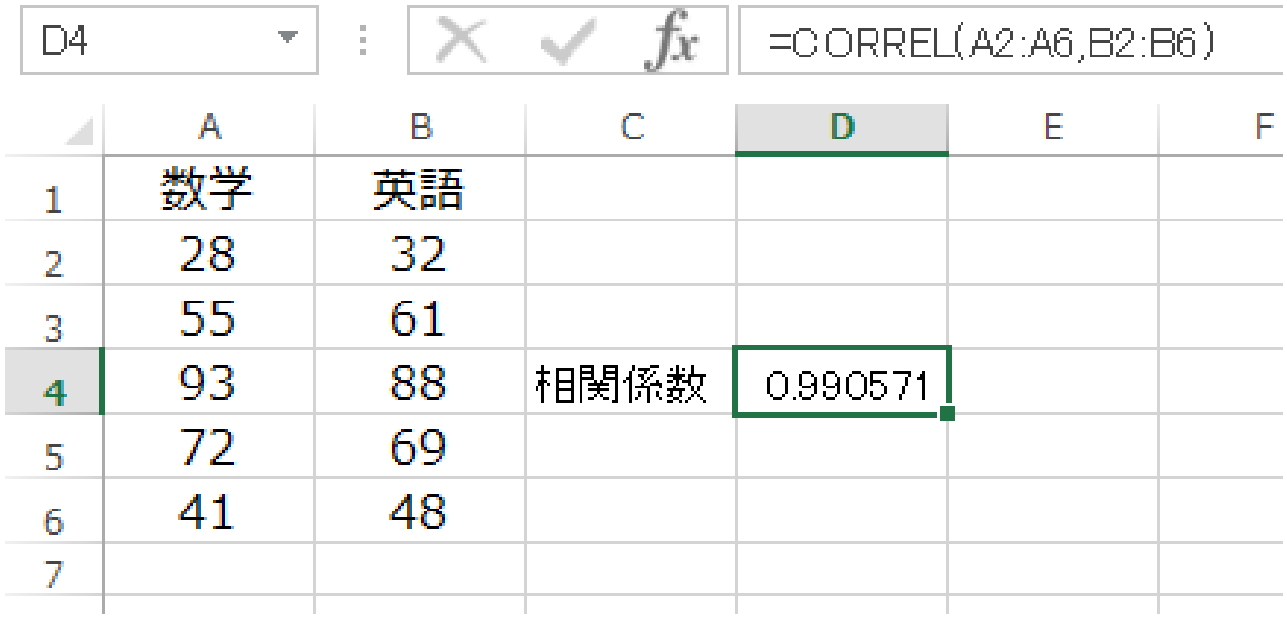
負の相関



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。

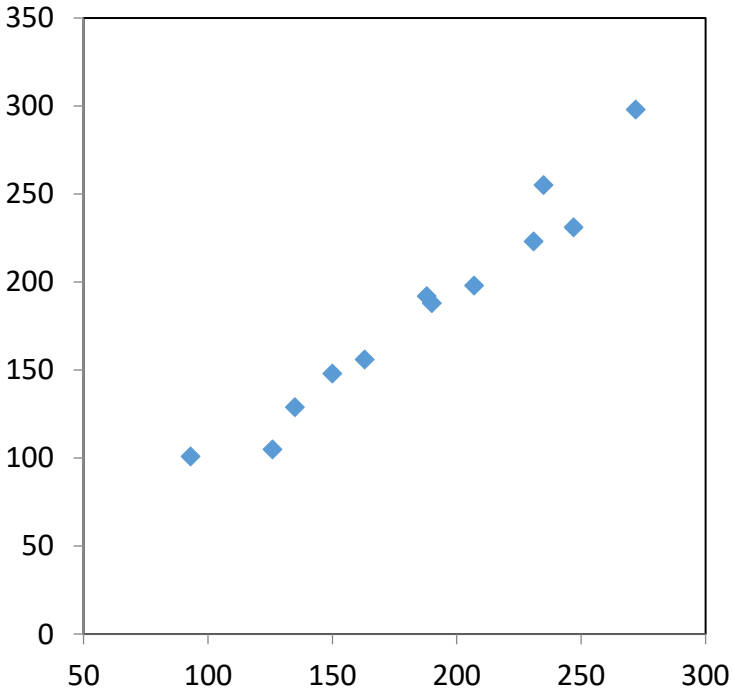
EXCEL



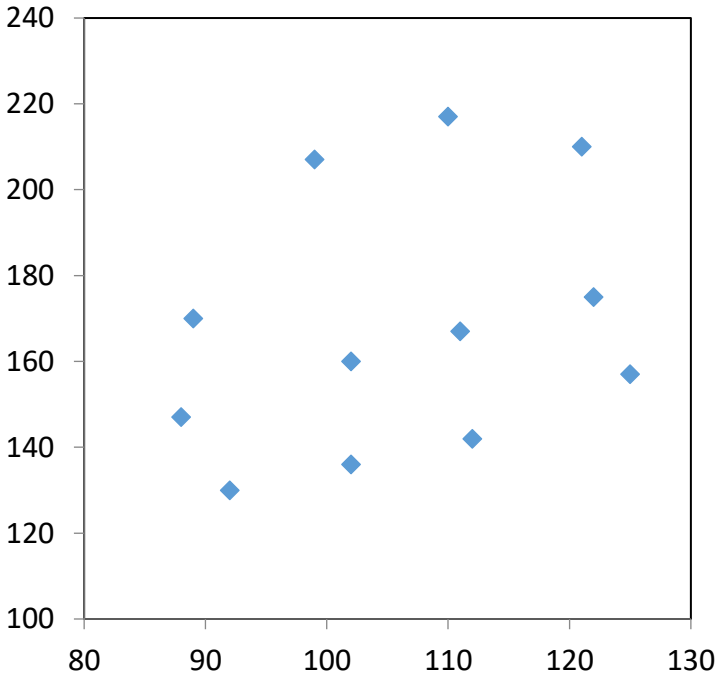
相関係数　CORREL

相関関係の強さ

相関係数（r）　　$-1 \leqq r \leqq 1$



r＝0.97



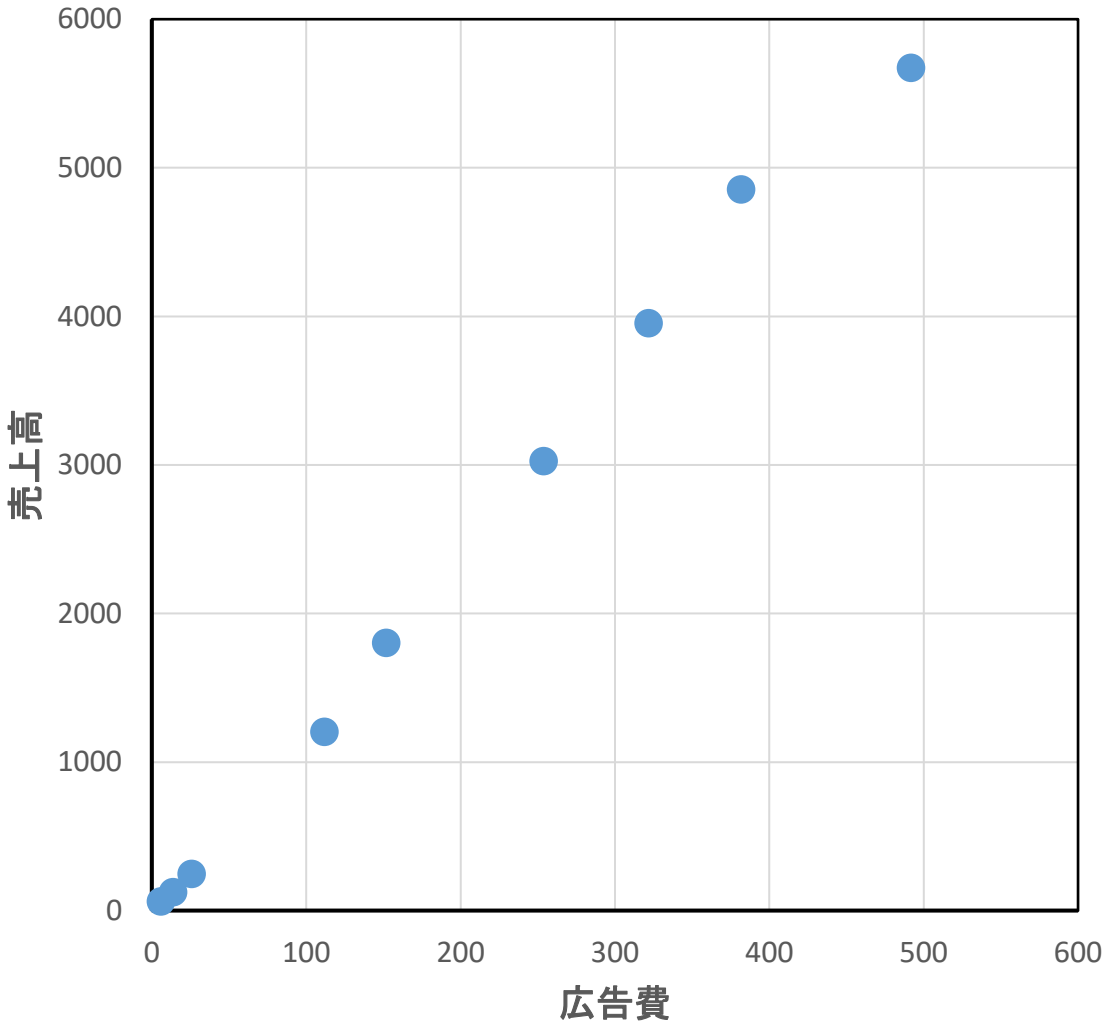
r＝0.32

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

広告費と売上高の関係は？

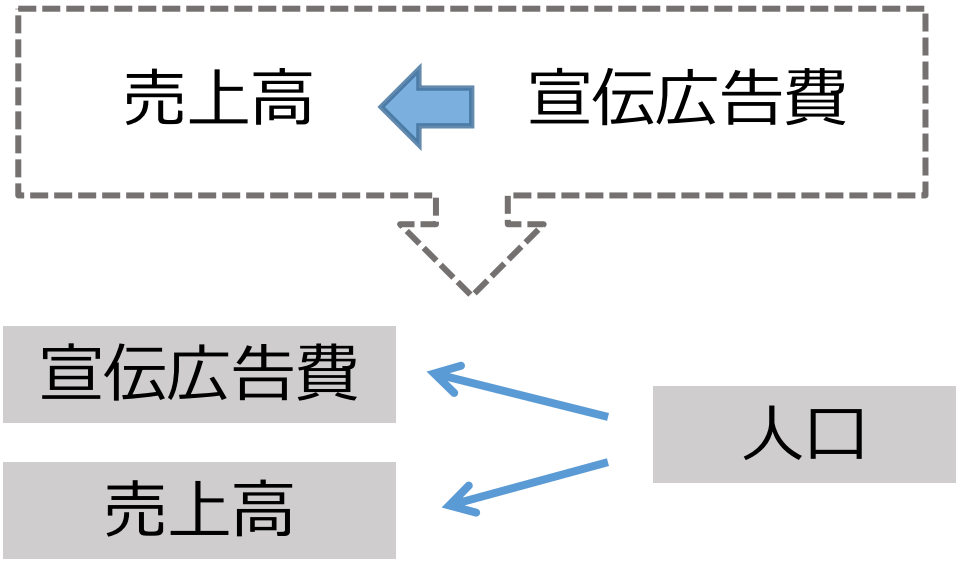
都道府県別広告費と売上高



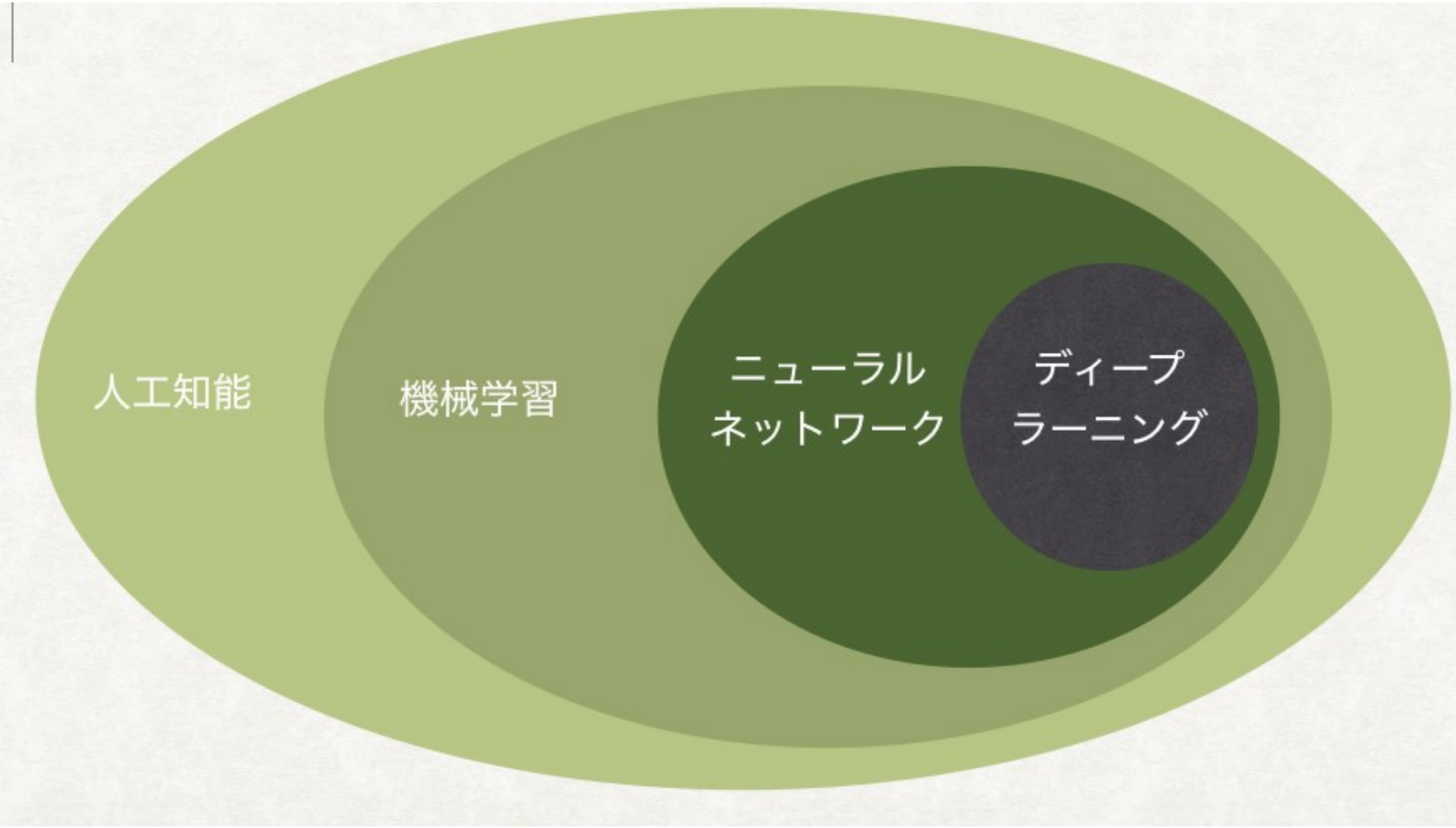
r = 0.991

相関関係と因果関係

広告宣伝費は売上高に貢献？



交絡変数（人口）に注意が必要！


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）

◇データから反復的に学習し、そこに潜むパターンを見つけ出すこと。学習した結果を新たなデータにあてはめ、将来を予測することができる。（SAS）

(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

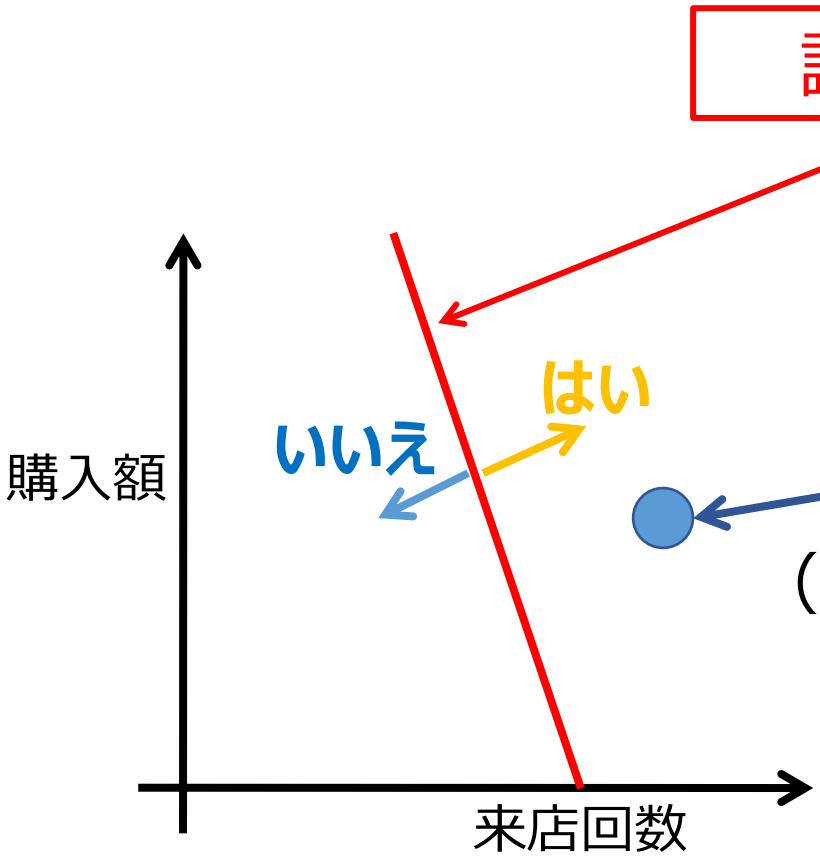
(例) ポイント会員にバーゲンの案内状を送る。
バーゲンに来ない客に送ると余分な郵送費がかかる。
できるだけ来店しそうな顧客を選びたい。

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〉）



会員5番は来店する

(購入額＝6万円,来店回数＝20回)

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

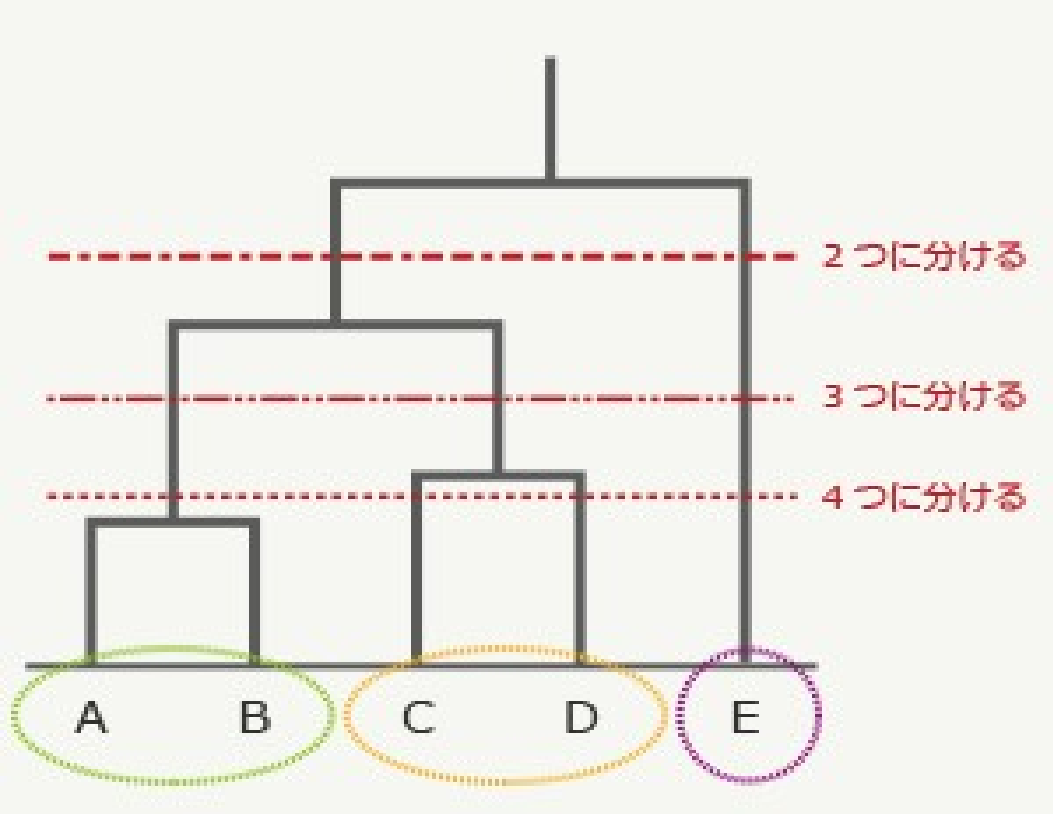
ポイント会員を、その購買の傾向が似ているクラスタに分割する。

会員を属性で表現。

第1属性　：弁当の購入額
第2属性　：清涼飲料の購入額
第・・属性：野菜類の購入額

会員3　：〈1000円、200円、・・250円〉

クラスター分析



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

まとめ

バラツキが重要である。

仮説（シナリオ）と検証が重要である。

有用なデータ＝数値データ＋背景である。

データの活用 ⇒ 分析力（基本知識）× 現場実践力

大学における統計学科の設置数

米国　　主だった大学はすべて設置

中国　　161大学（2005）

韓国　　75大学（2011）

1dayセミナー

セミナー名称	料金（税込）	内容	日程
初心者のためのデータ分析法入門（１Ｈ無料セミナー）	無料	内容	日程
EXCELによるデータ分析入門	49,500円	内容	日程
実践統計学	49,500円	内容	日程
実践統計手法	49,500円	内容	日程
管理者のための実践統計学	49,500円	内容	日程
医療データ解析入門	49,500円	内容	日程
製造現場のための統計解析法（基礎編）	49,500円	内容	日程
製造現場のための統計解析法（手法編）	49,500円	内容	日程
アンケート調査法入門	49,500円	内容	日程
販売予測・需要予測入門	49,500円	内容	日程
ビッグデータのためのデータ解析法（入門編）	49,500円	内容	日程
Rによる統計学（入門編）	49,500円	内容	日程
Rによる統計学（実践編）	49,500円	内容	日程
Pythonによるデータ解析入門	49,500円	内容	日程
Pythonによるビッグデータ解析	49,500円	内容	日程
統計検定２級対応講座（２日間コース）	49,500円	内容	日程

分野別データサイエンティスト養成講座のご案内

■第四次産業革命スキル習得講座

■専門実践教育訓練講座
（受講料の50%給付）

◎昼夜開講
◎完全オンラインによる受講
（ライブ＋アーカイブ）

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

■■　初心者のためのデータ分析法入門　■■
＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝
2020年8月10日　第4刷

発行元：　株式会社 データサイエンス研究所

本社　〒152-0021 東京都千代田区平河町2-5-5 全国旅館会館
tel ：03-3265-3908　mail：info@datascience.co.jp
＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝