実践統計学

データサイエンス研究所

概　要

データの可視化
基本統計量
データの評価の方法
統計的検定
2つの平均値の差の検定
相関分析
回帰分析

データの可視化

1.箱ひげ図

データの可視化

データを分析するには、まず全体を視覚的に把握することが重要。

各データ群の把握

①外れ値の有無の確認

②分布の偏り

箱ひげ図

ヒストグラム

1.箱ひげ図

中央値（MEDIAN）

230	287	252	313	401	331	292	851	229

平均値＝354

平均値は極端な値の影響を受け易い

1	2	3	4	5	6	7	8	9
229	230	252	287	292	313	331	401	851

大きさの順番に並び替えて、中央に位置する値を代表値

四分位数

・データを小さい順に並べる。

範囲＝最大値－最小値

・4分の1ずつの場所にある値

Q1：第1四分位数（25%点）
Q2：第2四分位数（50%点・中央値）
Q3：第3四分位数（75%点）

四分位範囲（IQR）　：　Q3－Q1

230	287	252	313	401	331	292	851	229

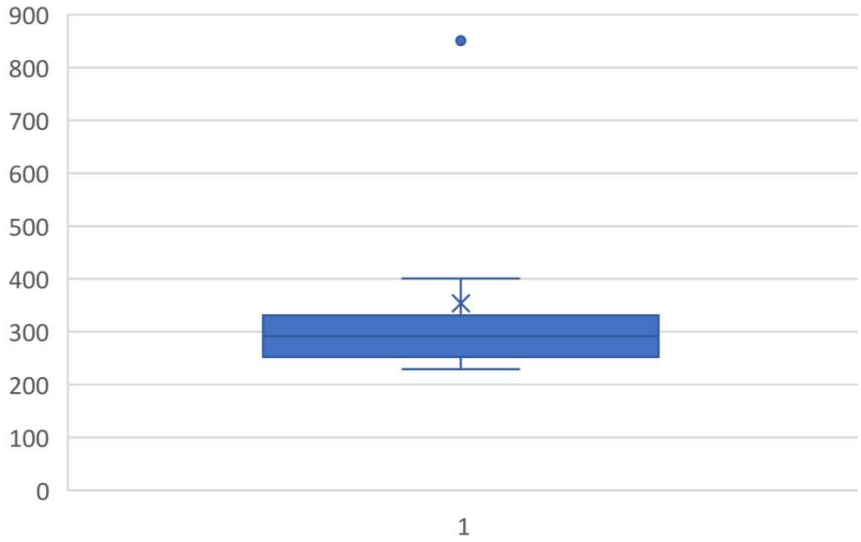
1	2	3	4	5	6	7	8	9
229	230	252	287	292	313	331	401	851

平均　　：354
Q1 第1四分位数　：252
Q2 中央値　　：292
Q3 第3四分位数　：331

四分位範囲　Q3－Q1＝331－252＝79

箱ひげ図（EXCEL)

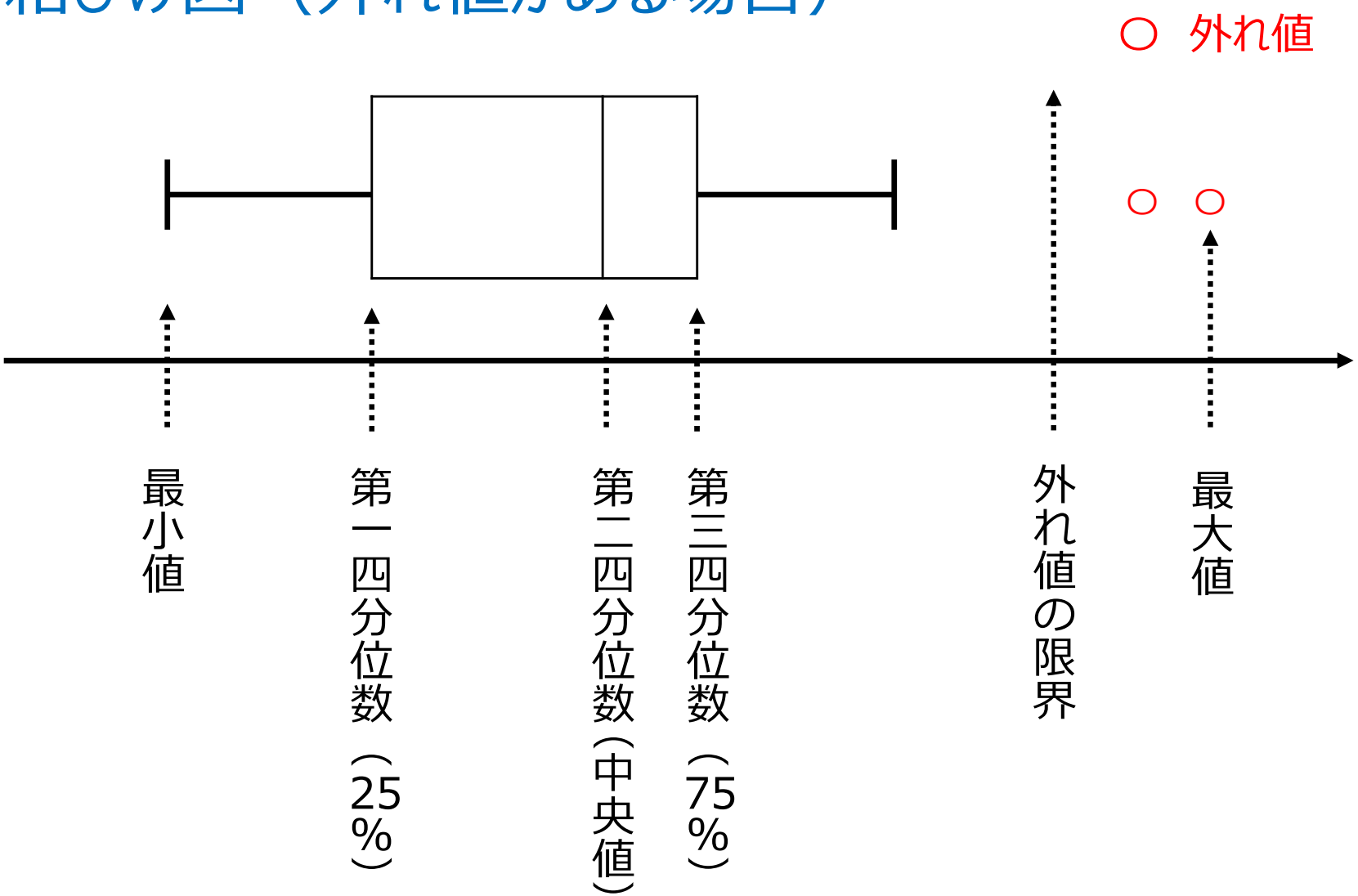
1）「挿入」、「グラフ」、「箱ひげ図」、「OK」をクリックする。
2）表示された箱ひげ図の箱の中を右クリック、「データ系列の書式設定」をクリックする。
3）「四分位数計算」、「包括的な中央値」を選択する。



×：算術平均

注）排他的な中央値：四分位数の計算時に中央値を使用しない。

箱ひげ図（外れ値がある場合）



外れ値

Q1 四分位点(25%)　：252
Q2 中央値　(50%)　：292
Q3 四分位点(75%)　：331

四分位範囲　Q3－Q1＝331－252＝79

◇四分位範囲（IQR）の1.5倍を基準

下の限界	上の限界
Q1－1.5×IQR以下	Q3＋1.5×IQR以上
252－(1.5×79)	331+(1.5×79)
外れ値:133.5以下	外れ値:449.5以上

データの可視化

2.ヒストグラム

ヒストグラムの作成（EXCEL）

1）「挿入」、「グラフ」、「ヒストグラム」、「OK」をクリックする。

2）横軸をクリックし、「軸の書式設定」をクリックする。

3）「ビンの数（階級数）」を設定し、「ビンの幅」を調整する。

＜ビンの数（階級数）の設定法＞

スタージェスの公式

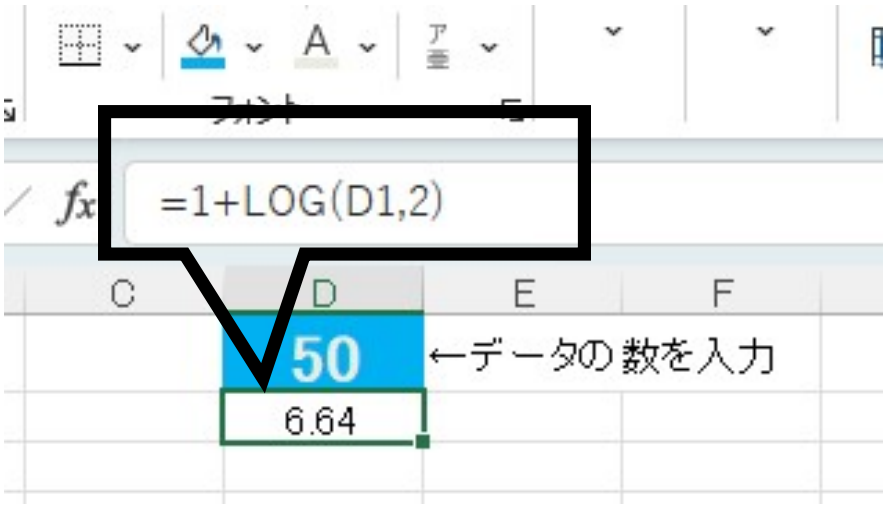
階級の数 ＝ $1 + \log_2(N)$　　N：データ数

＊スタージェスの公式

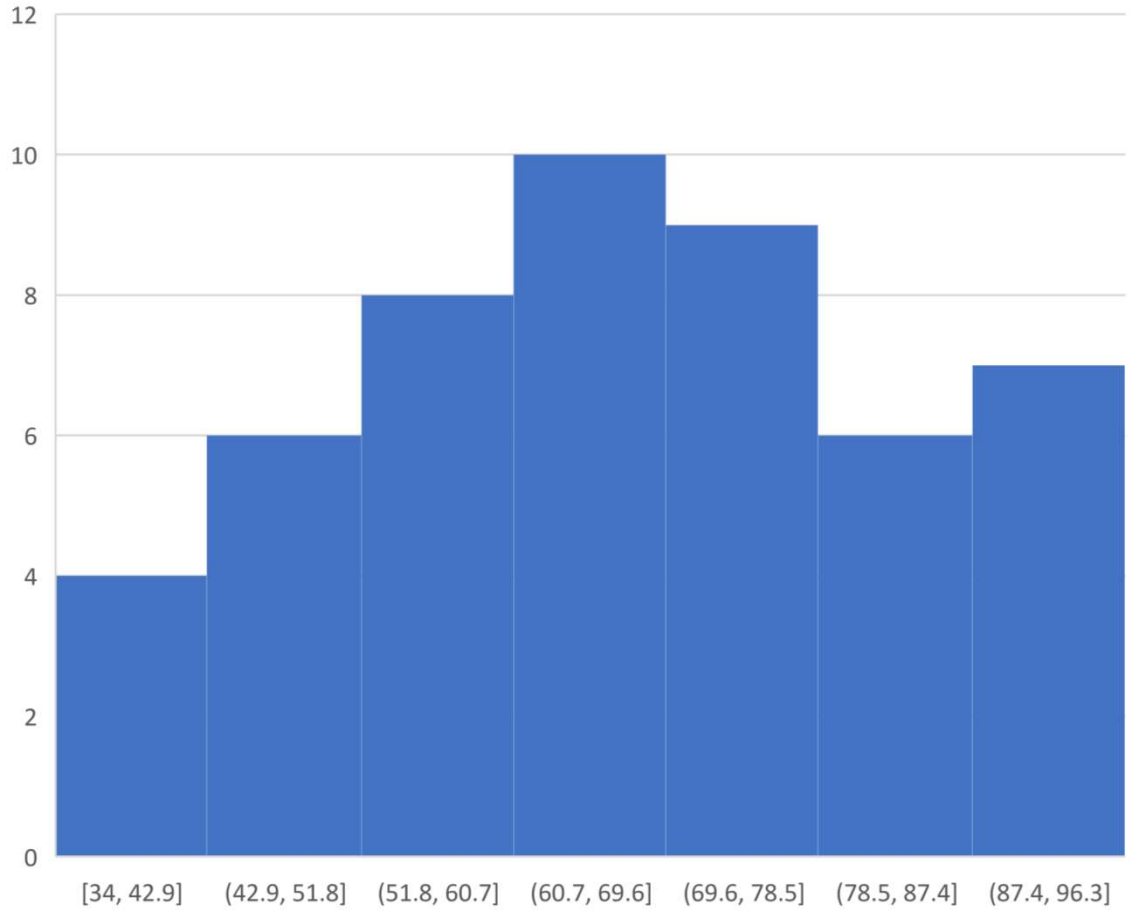
$1+\log_2 50=6.64\cdots$

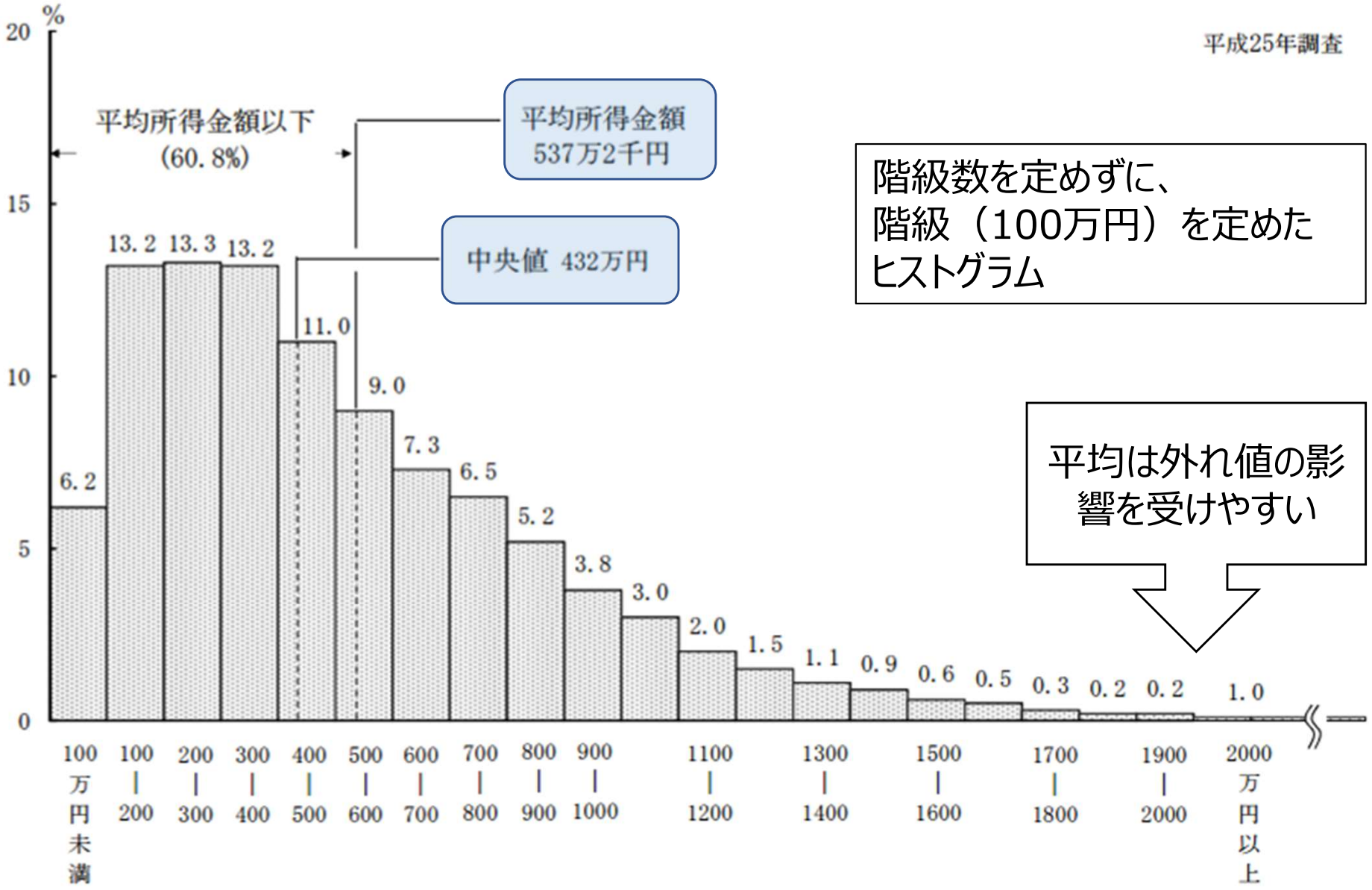
N→データの数
2→定数

EXCEL： 1+log(N,2)



ビンの数（階級の数）＝7
ビンの幅＝8.9





厚生労働省　1世帯当たりの所得金額階級別相対度数分布

基本統計量

3.いろいろな平均

①算術平均

Q. 1,2,3,4,5 の平均は？
（1+2+3+4+5）/ 5＝3

AVERAGE

	A	B	C	D
1		単位:億円		
2	支店	売上高		
3	北海道	121		
4	東北	213		
5	関東	1,452		
6	中部	417		
7	関西	912		
8	九州	332		
9	沖縄	28		
10	合計	3,475		
11	平均	496		
12				
13				
14				

②幾何平均

◇ある会社の年度別売上高。平均伸び率は？

（億円）

年度	売上高
2020	100
2021	120
2022	168

2020～2021　20% ↑　（1.2倍）
2021～2022　40% ↑　（1.4倍）

（20%+40%）/2＝30%？

100×1.30×1.30＝169.00

比率の平均に算術平均は使用できない

（億円）

年度	売上高
2020	100
2021	120
2022	168

1.2と1.4の幾何平均は、

$\sqrt{1.2\times1.4} = 1.2961...$

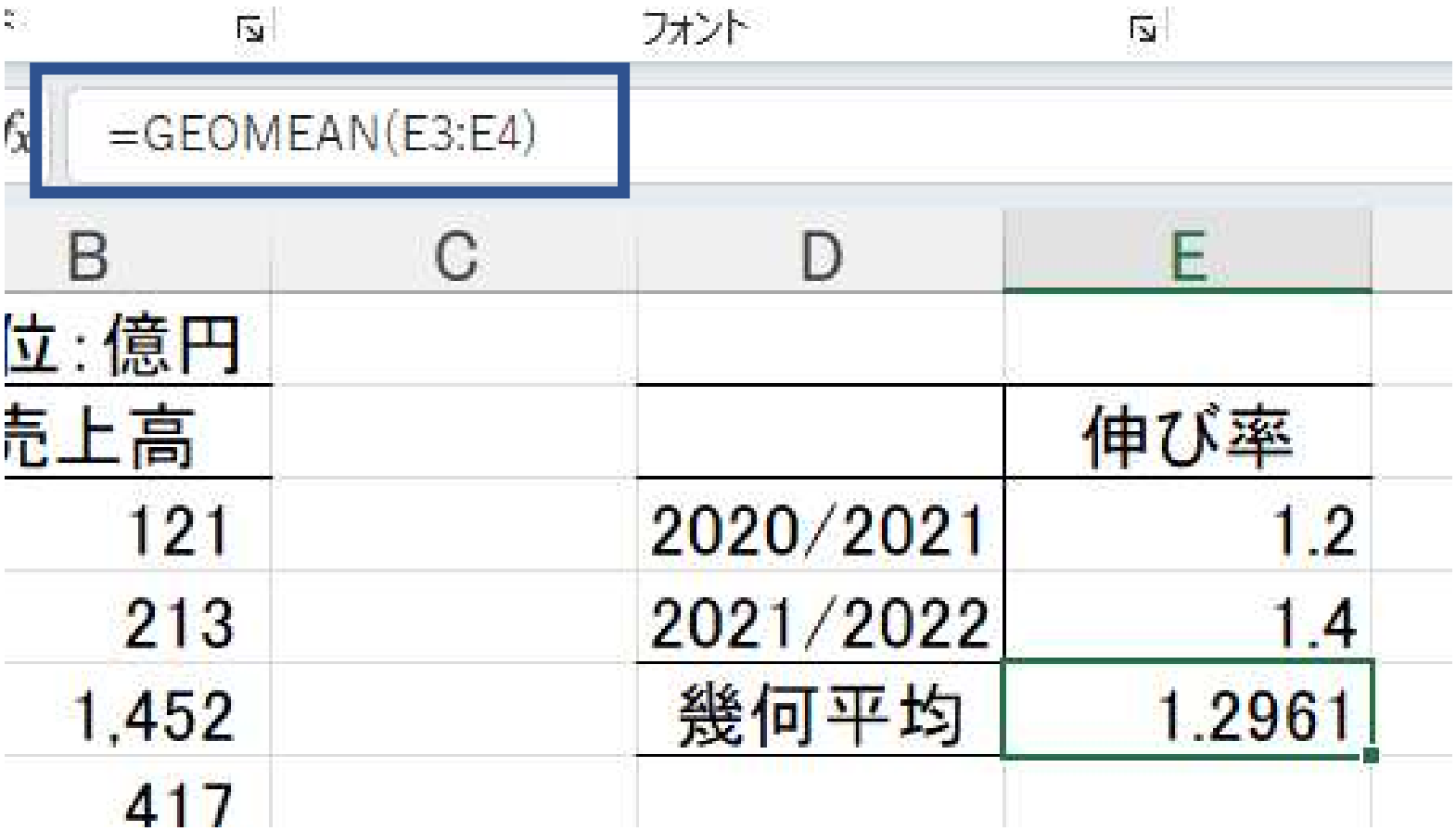
$100\times1.2961...\times1.2961... = 167.99$

2つの幾何平均　$\sqrt{a\times b}$

3つの幾何平均　$\sqrt[3]{a\times b\times c}$

4つの幾何平均　$\sqrt[4]{a\times b\times c\times d}$

GEOMEAN（幾何平均）



基本統計量

4.分散と標準偏差

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

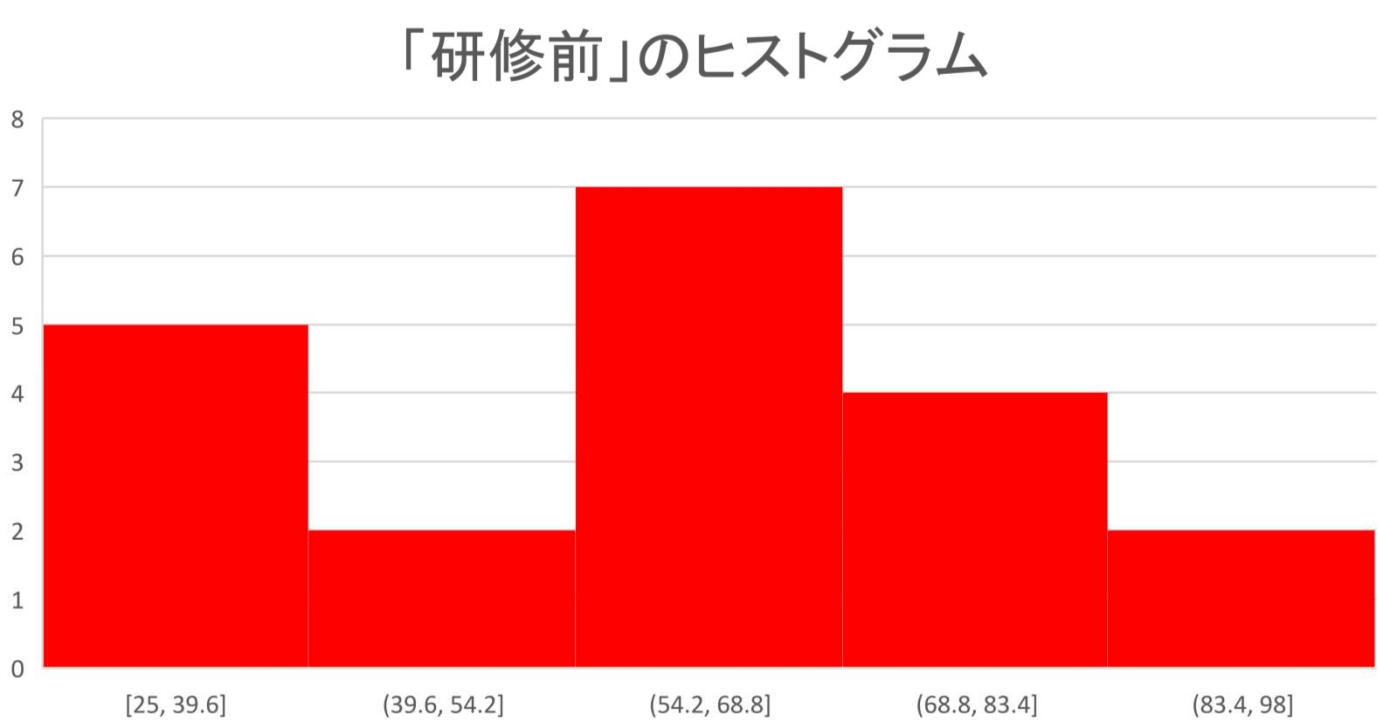
68、25、65、67、33、68

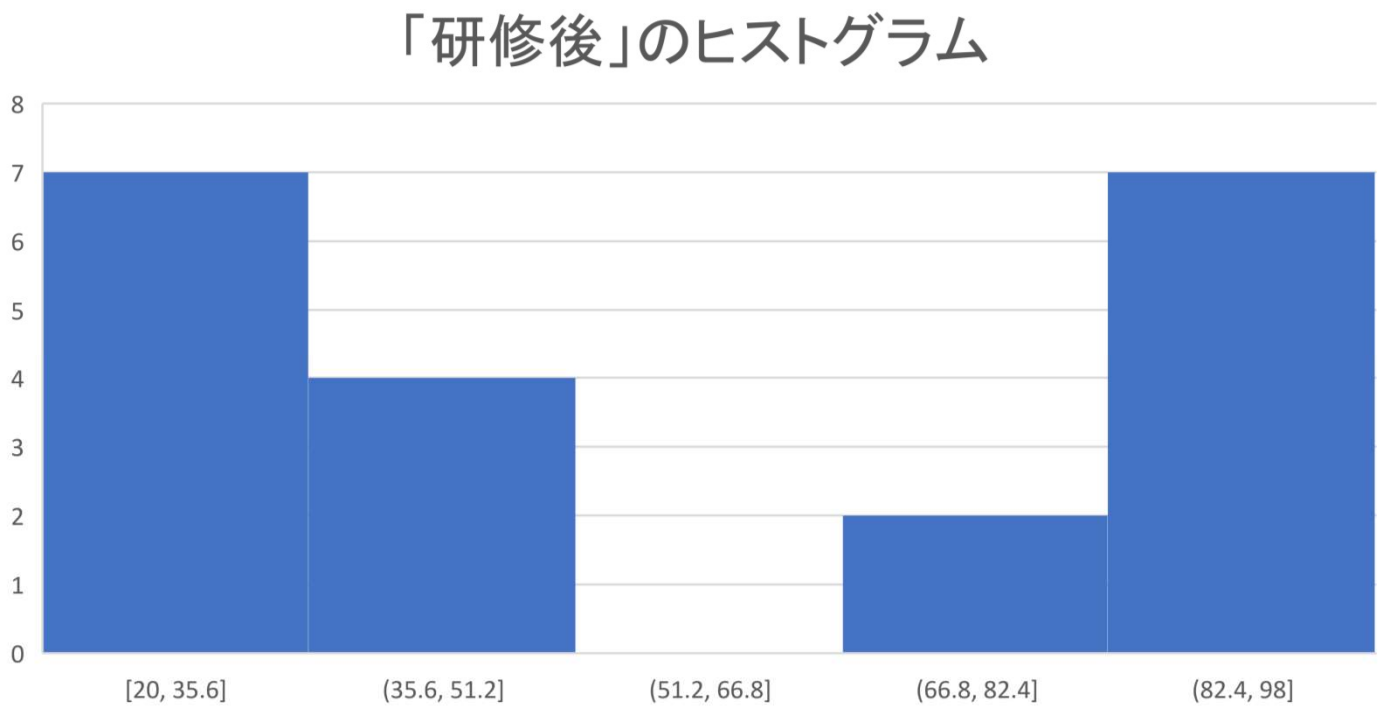
＜研修後＞

72、31、95、36、89、88、89

76、28、47、23、28、96、48

51、20、33、91、27、98





どちらも平均値は
58.3

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）

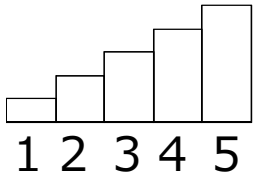
研修前より研修後のバラツキが大きい

成績を比較するためには各データ集団の
バラツキを表す指標が必要！

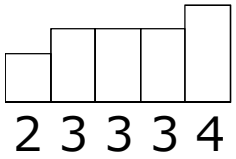
平均値が同じデータ群

A群～D群の平均値はすべて同じ値3

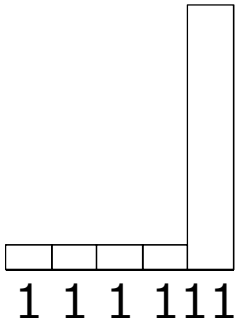
A: {1,2,3,4,5}



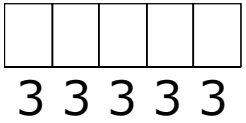
B: {2,3,3,3,4}



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　（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。

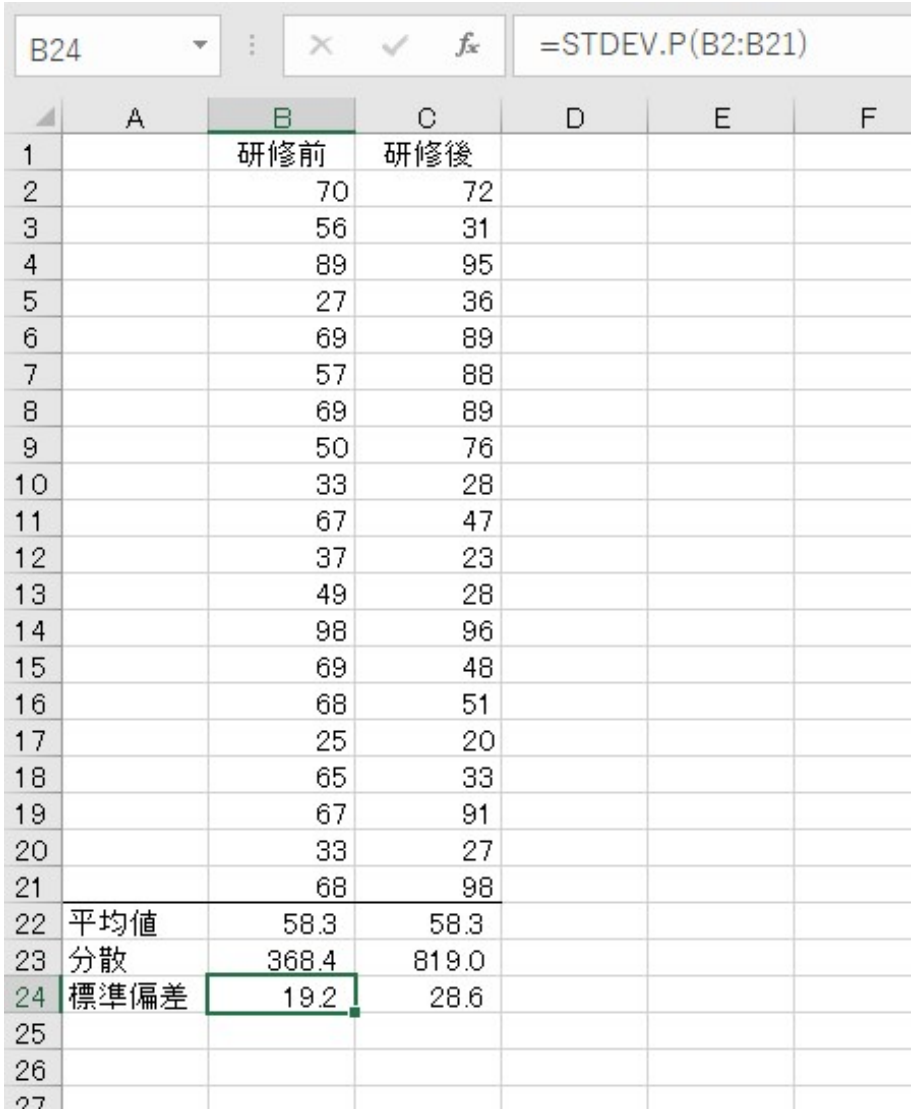
A群：$\sqrt{2}$ ≒ 1.414万円

F群：$\sqrt{200}$ ≒ 14.14千円

⇒ バラツキの代表値②「標準偏差」

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						
27						

データの評価の方法

5.標準化（Z値と偏差値）

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

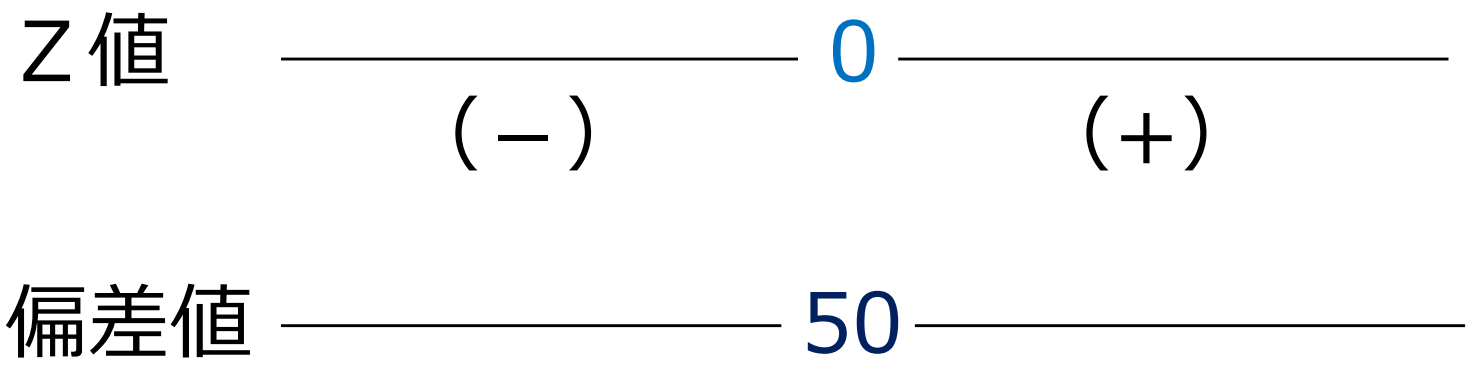
Z 値×10＋50

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

Z値の特性

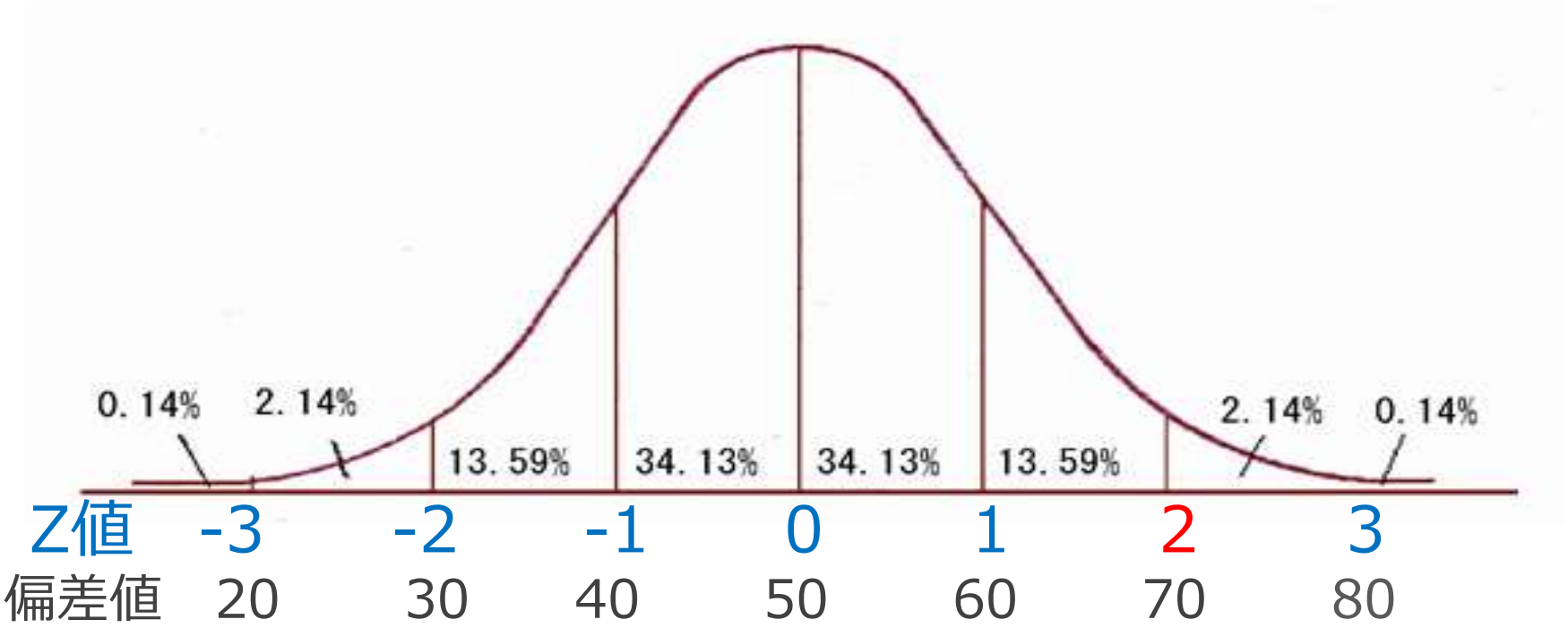
	データ	Z値	偏差値
	1	-1.414	35.858
	2	-0.707	42.929
	3	0	50
	4	0.707	57.071
	5	1.414	64.142
平均値	3	0	50
分散	2	1	100
標準偏差	1.414	1	10

Z値の平均値＝0　標準偏差＝1

データの評価の方法

6.標準正規分布

標準正規分布（standard normal distribution）



◇平均を中心に左右対称（平均0、標準偏差1）

Z値＝1（偏差値60）以上の割合は全体の約15.87%
Z値＝2（偏差値70）以上の割合は全体の約2.28%

標準正規分布の活用

■日本の成年男子の身長が平均170cm 標準偏差6cm、
182cm以上の人は、全体の何%を占めるか？

$$Z = \frac{182-170}{6} = 2$$

182cm以上の人は約2.28%

■全国で10万人が参加したある資格試験
A君の成績は、88点、平均点62.4点、標準偏差が12.8点
A君の成績は上位何%？

$$Z = \frac{88-62.4}{12.8} = 2$$

88点以上の人は約2.28%

■通信販売会社における顧客売上高の平均5,568円、
標準偏差2,216円、10,000円以上購入した顧客は、
全体の何%を占めるか？

$$Z = \frac{10{,}000\text{-}5{,}568}{2{,}216} = 2$$

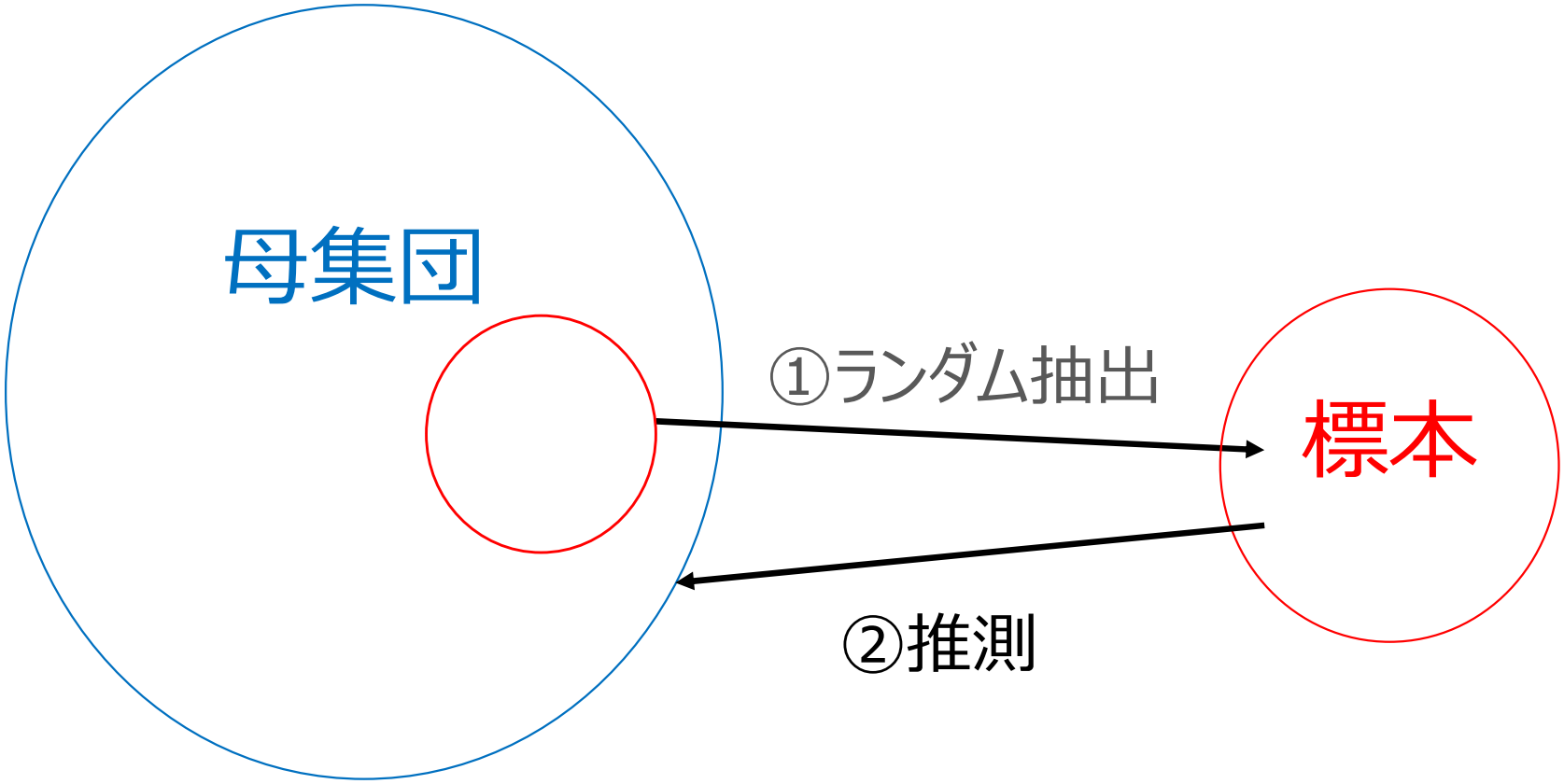
10,000円以上：約2.28%

Z値は単位とは無関係。

Z値は0を基準としたときの違いの大きさ

7.統計的検定とは

推測統計の基本的な考え方



一部（標本）から全体（母集団）を推測する

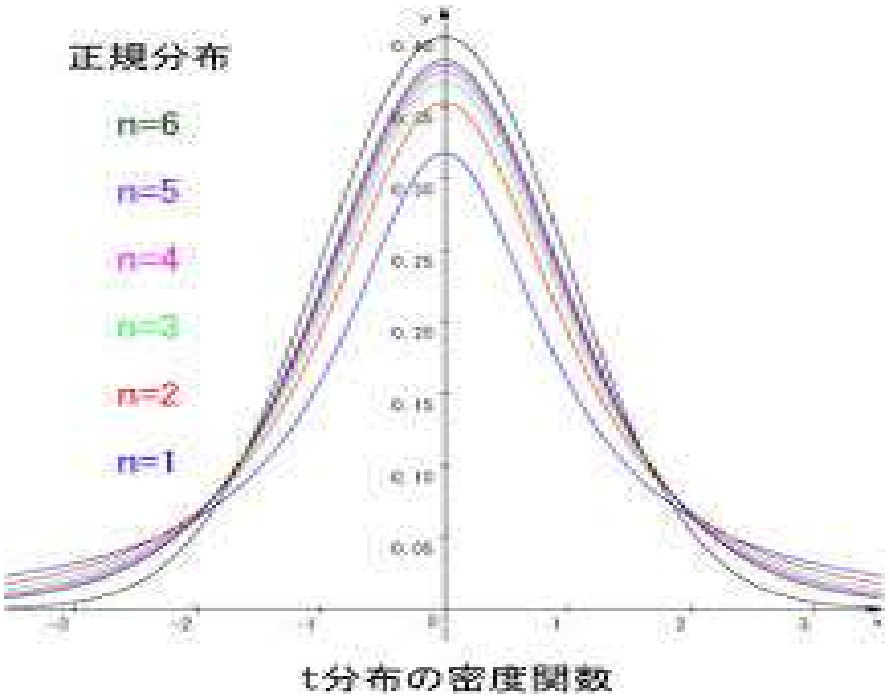
対応のないデータと対応のあるデータ

対応のない	回答者を1条件だけに割り当てた場合	回答者を全ての条件に割り当てた場合
	データを対応づける理由がなく独立	１人の回答者に2個以上のデータがある
	男女の差 小学生と中学生の差	右手と左手で握力を測定 ＣＭの視聴前後の企業イメージ

2つの平均値の差の検定

8.対応のないt検定

t 分布



t分布の密度関数

注）1908年、W.S.ゴセット（キネスビールの研究員）が考案。

サンプルサイズが小さいときに精度を上げるために t 分布を考案。

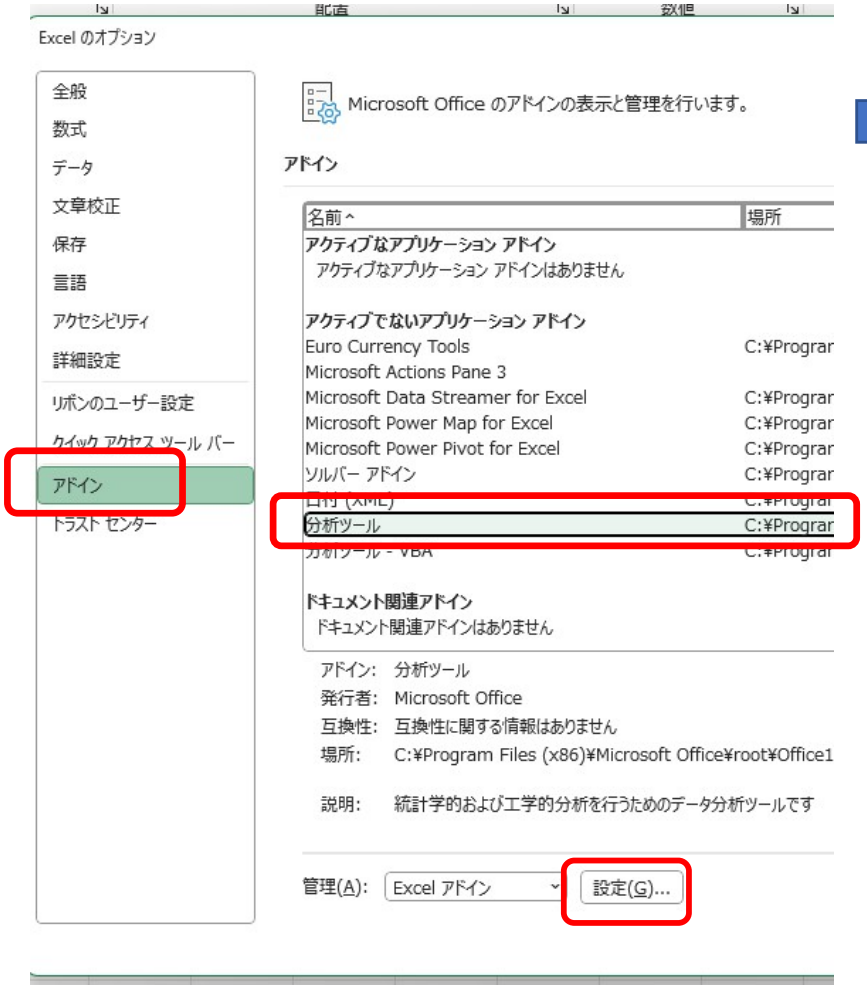
t分布を使う分析方法を　t 検定という

(A)商品のパッケージの好感度について
男女各10人に10点満点にて調査した。
男女の平均値の差に違いは見られるか？

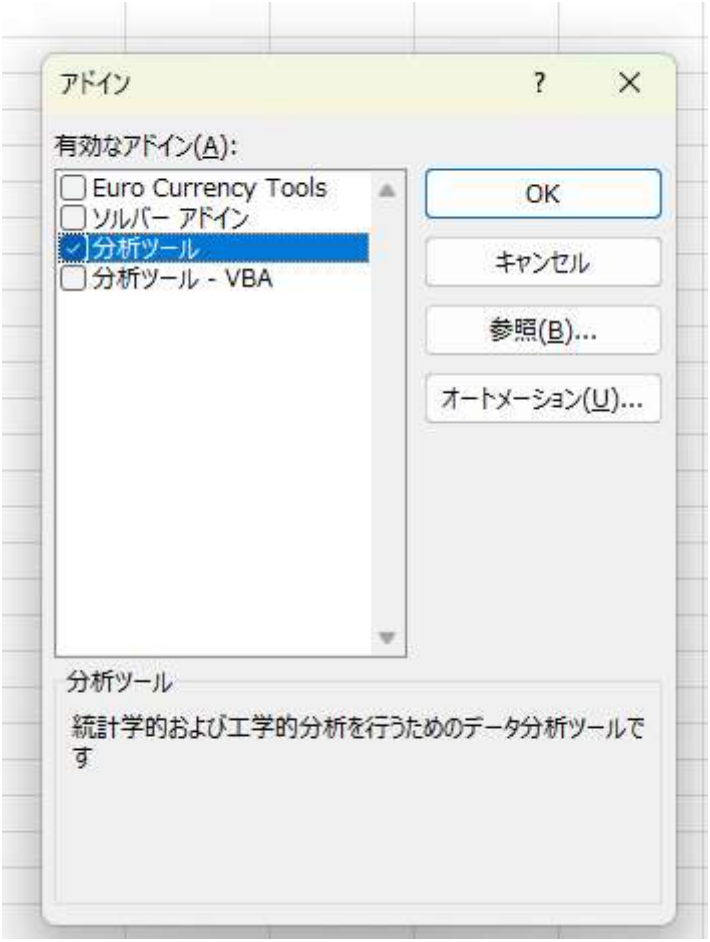
											平均
男性	6	4	5	5	6	5	6	6	4	6	5.3
女性	7	6	7	5	6	5	6	7	6	6	6.1

分析ツールのインストール

① 「ファイル」－「オプション」－「アドイン」－「分析ツール」を選択し、「設定」をクリックする。

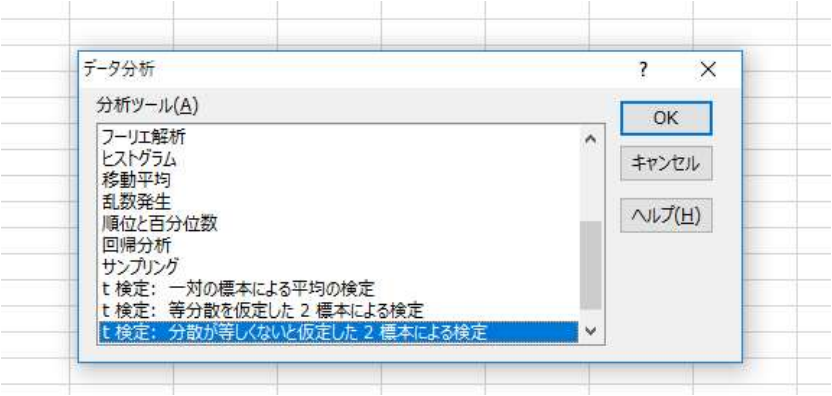


② 「分析ツール」を選択、「OK」をクリックする。

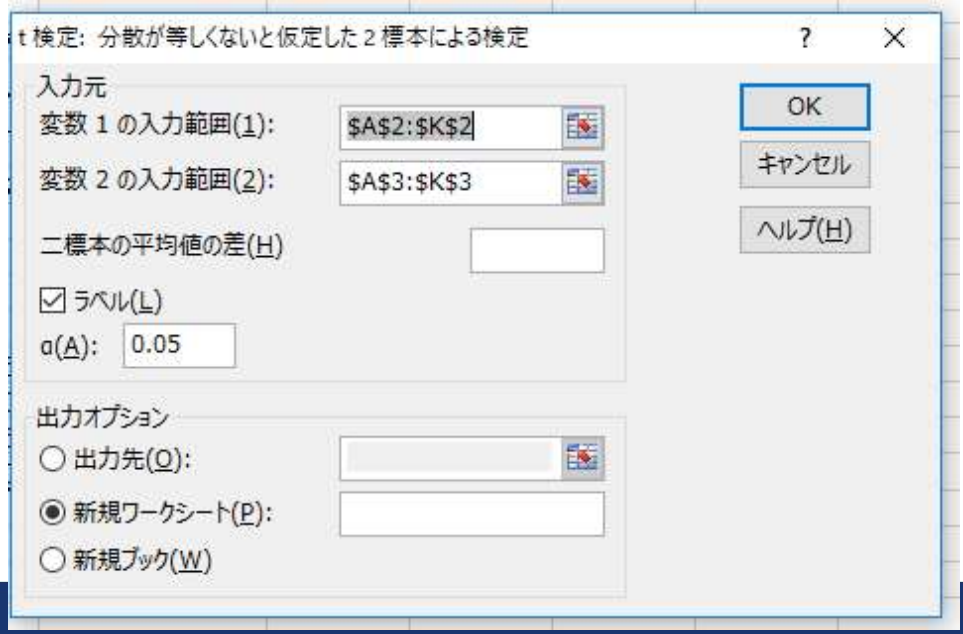


1）メニューバーの「データ」ー「データ分析」をクリックする。

2）「ｔ検定：分散が等しくないと仮定した」を選択し、「OK」をクリック。



3）「変数1の入力範囲」、「変数2の入力範囲」を指定、「ラベル」をチェック、出力先を指定し、「OK」をクリックする。



分析ツールによる出力結果

t-検定: 分散が等しくないと仮定した２標本による検定

	男性	女性
平均	5.3	6.1
分散	0.678	0.544
観測数	10	10
仮説平均との差異	0	
自由度	18	
t	-2.288	
P(T<=t) 片側	0.0172	
t 境界値 片側	1.734	
P(T<=t) 両側	0.0344	
t 境界値 両側	2.100	

t 値＝2.288　P値＝0.0344

p値(有意確率、危険率)

◇p値の判定

p値が**5%**より小さいので、「違いがある」と判断する

p値＝0.034 ＜ **0.05（有意水準）**

t(18)=2.288, p=0.034
有意水準5%において、有意である
（男性平均と女性平均に違いが見られる）。

（B）平均値の差が拡大したとき

◇商品のパッケージの好感度について
男女各10人に10点満点にて調査した。
男女の平均値の差に違いは見られるか？

											平均値
男性	6	4	5	5	6	5	6	6	4	6	5.3
女性	7	6	7	6	6	6	6	7	6	6	6.3

（C）データのバラツキが拡大したとき

◇商品のパッケージの好感度について
男女各10人に10点満点にて調査した。
男女の平均値の差に違いは見られるか？

											平均値
男性	7	3	8	3	7	2	2	6	6	9	5.3
女性	7	4	9	3	8	3	5	8	7	9	6.3

（D）サンプルサイズが拡大したとき

◇商品のパッケージの好感度について
男女各10人に10点満点にて調査した。
男女の平均値の差に違いは見られるか？

											平均値
男性	7	3	8	3	7	2	2	6	6	9	5.3
	7	3	8	3	7	2	2	6	6	9	
女性	7	4	9	3	8	3	5	8	7	9	6.3
	7	4	9	3	8	3	5	8	7	9	

男女各10人から男女各20人に増加

効果量（Cohen’s d）

t値とp値はサンプルサイズの影響を受ける

$$\text{効果量(d)} = \text{t値} \div \left(\frac{\sqrt{n_1 + n_2}}{2} \right)$$

注）効果量（Cohen’s d）の大きさの評価

0.2	0.5	0.8
小	中	大

効果量　$d = t \div \left(\frac{\sqrt{n_1+n_2}}{2}\right)$

(C) $d = 0.880 \div \left(\frac{\sqrt{10+10}}{2}\right) = 0.394$ (p=0.390)

(D) $d = 1.313 \div \left(\frac{\sqrt{20+20}}{2}\right) = 0.415$ (p=0.197)

サンプルサイズが変わっても、
効果量はほとんど変わらない

2つの平均値の差の検定

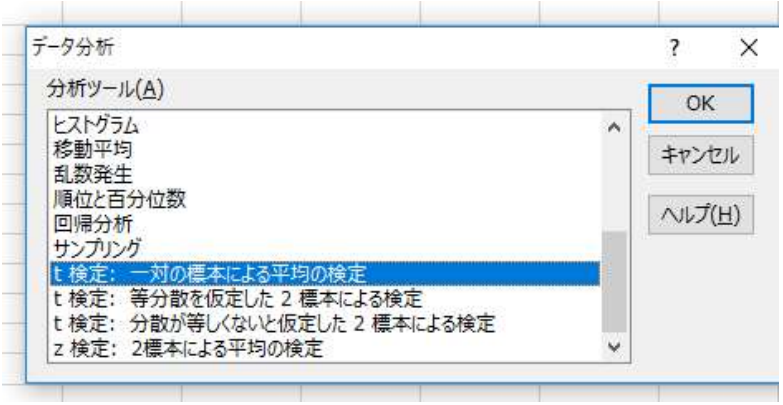
9.対応のあるt検定

商品の理解度について、10人に商品説明前後に、
それぞれ10点満点にて調査した。
説明前後による理解度に差は見られるか。

											平均値
説明前	7	3	8	3	7	2	2	6	6	9	5.3
説明後	7	4	9	3	8	3	5	8	7	9	6.3

データ数＝20
回答者10人のデータ

「 t 検定：一対の標本による平均の検定」を選択し、「OK」をクリックする。



	説明前	説明後
平均	5.3	6.3
分散	6.677	5.567
観測数	10	10
ピアソン相関	0.931	
仮説平均との差異	0	
自由度	9	
t	-3.354	
P(T<=t) 片側	0.0042	
t 境界値 片側	1.833	
P(T<=t) 両側	0.0084	
t 境界値 両側	2.262	

回答者が各自2回ずつ回答
各回答者の差に着目

対応のある2標本 t 検定

t 値 ＝ 3.354　P値＝0.008

効果量（Cohen’s d）

$$効果量(d) = t値 \div \sqrt{n}$$

注）効果量（Cohen’s d）の大きさの評価

0.2	0.5	0.8
小	中	大

対応のあるデータを対応のないデータとして検定した場合

	説明前	説明後
平均	5.3	6.3
分散	6.678	5.567
観測数	10	10
仮説平均との差異	0	
自由度	18	
t	-0.904	
P(T<=t) 片側	0.189	
t 境界値 片側	1.734	
P(T<=t) 両側	0.378	
t 境界値 両側	2.101	

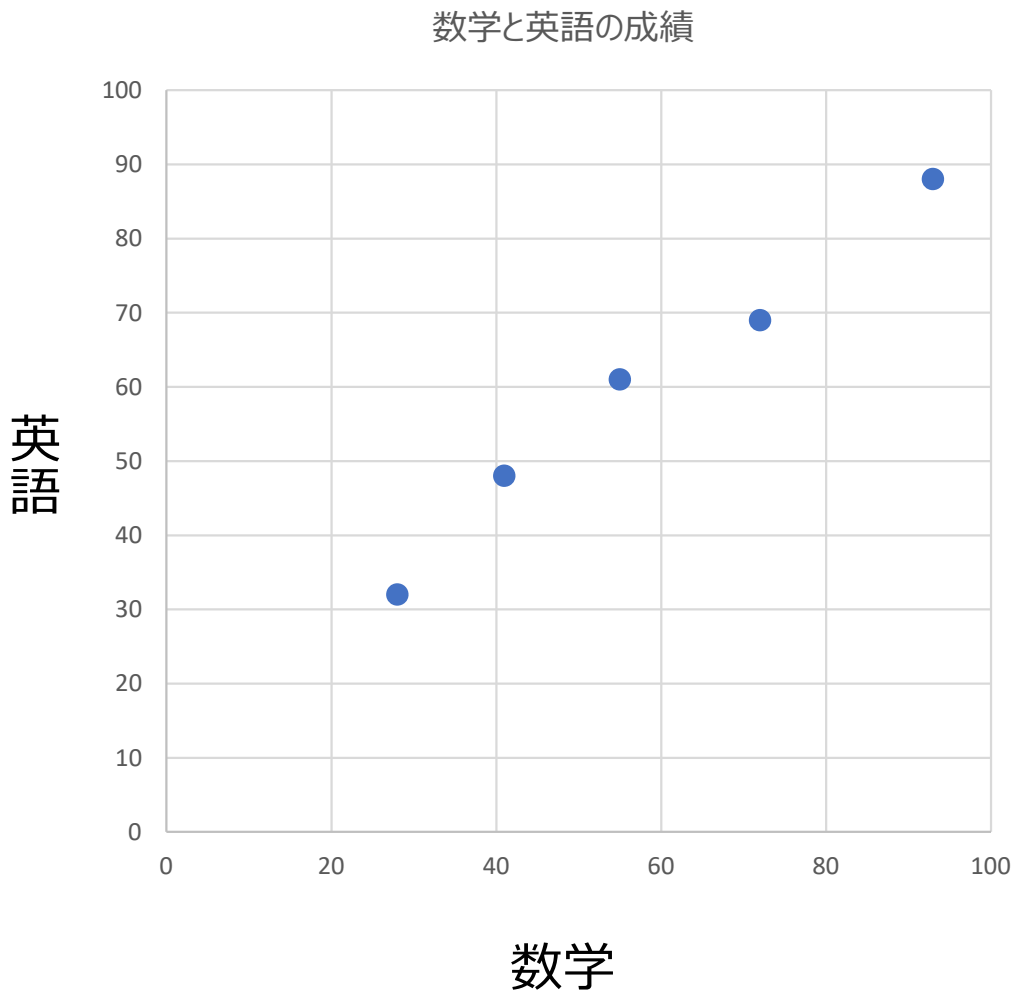
	P値	t 値	効果量
対応のある場合	0.008	3.354	1.060
対応のない場合	0.378	0.904	0.404

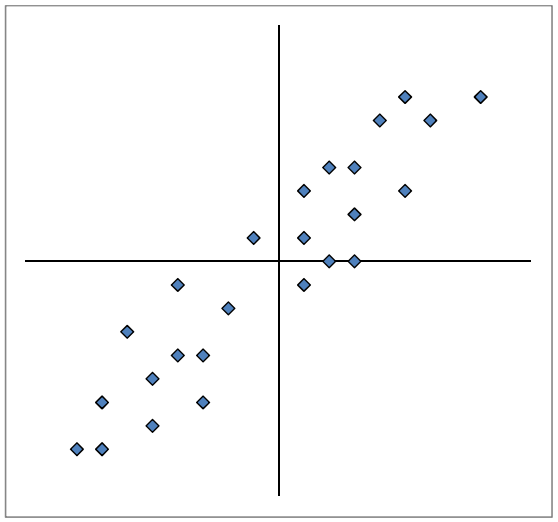
相関分析

10.散布図と積率相関係数

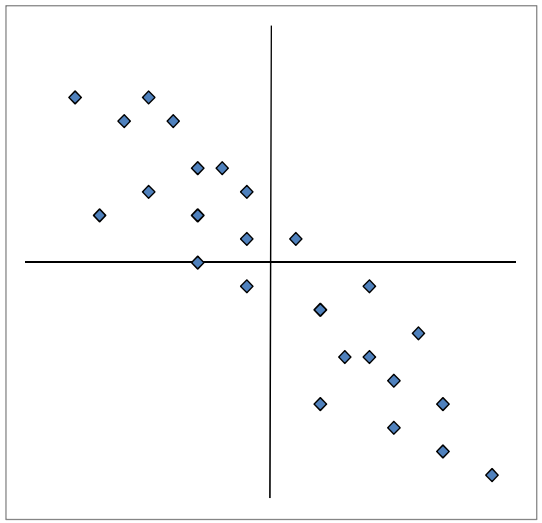
散布図

	数学	英語
Aさん	28	32
Bさん	55	61
Cさん	93	88
Dさん	72	69
Eさん	41	48

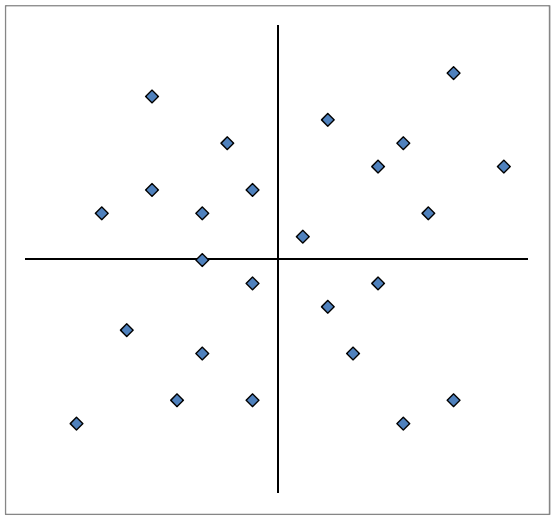




正の相関



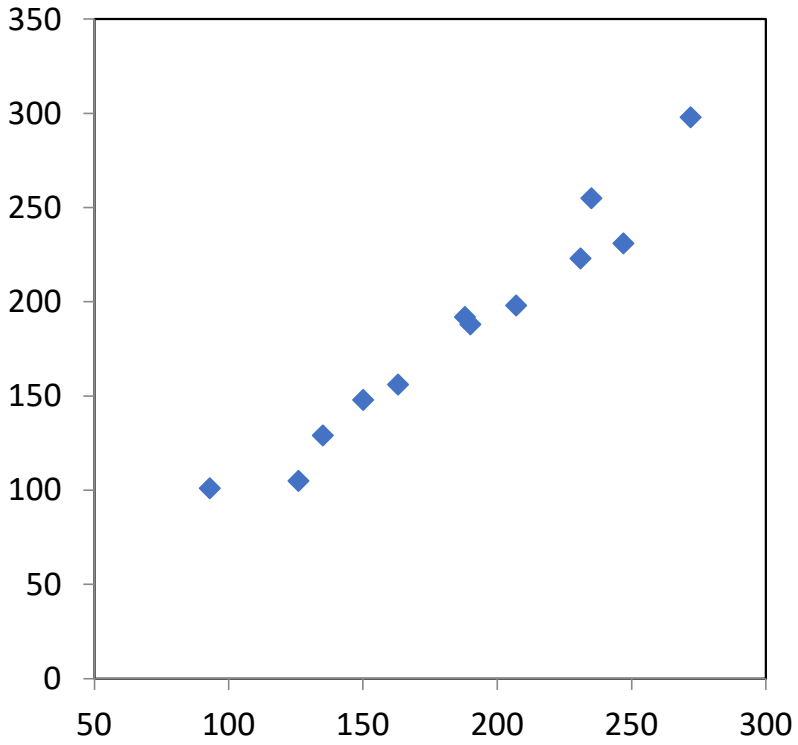
負の相関



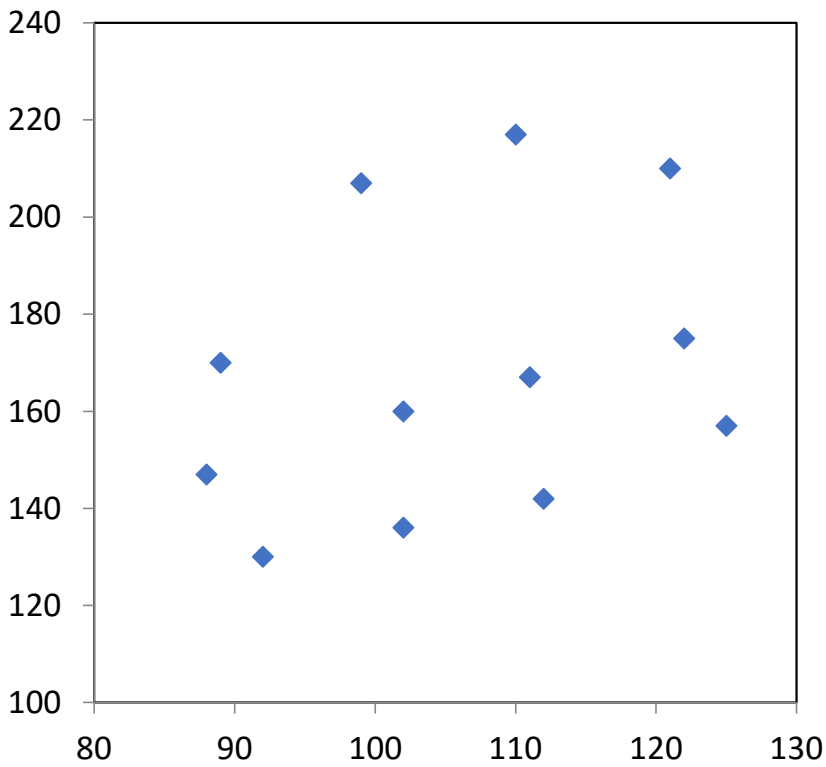
無相関

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。

積率相関係数（r）



r＝0.97



r＝0.32

$-1 \leqq r \leqq 1$

◇支店別広告費と売上高

CORREL（積率相関係数）

	A	B	C	D
1	支店	広告費	売上高	
2	北海道	92	44	
3	東北	93	102	
4	関東	332	288	
5	北陸	78	54	
6	中部	181	118	
7	近畿	108	138	
8	中国	113	138	
9	四国	72	86	
10	九州	243	152	
11	沖縄	13	22	
12	合計	1,325	1,142	
13				
14	相関係数	0.902359		
15				

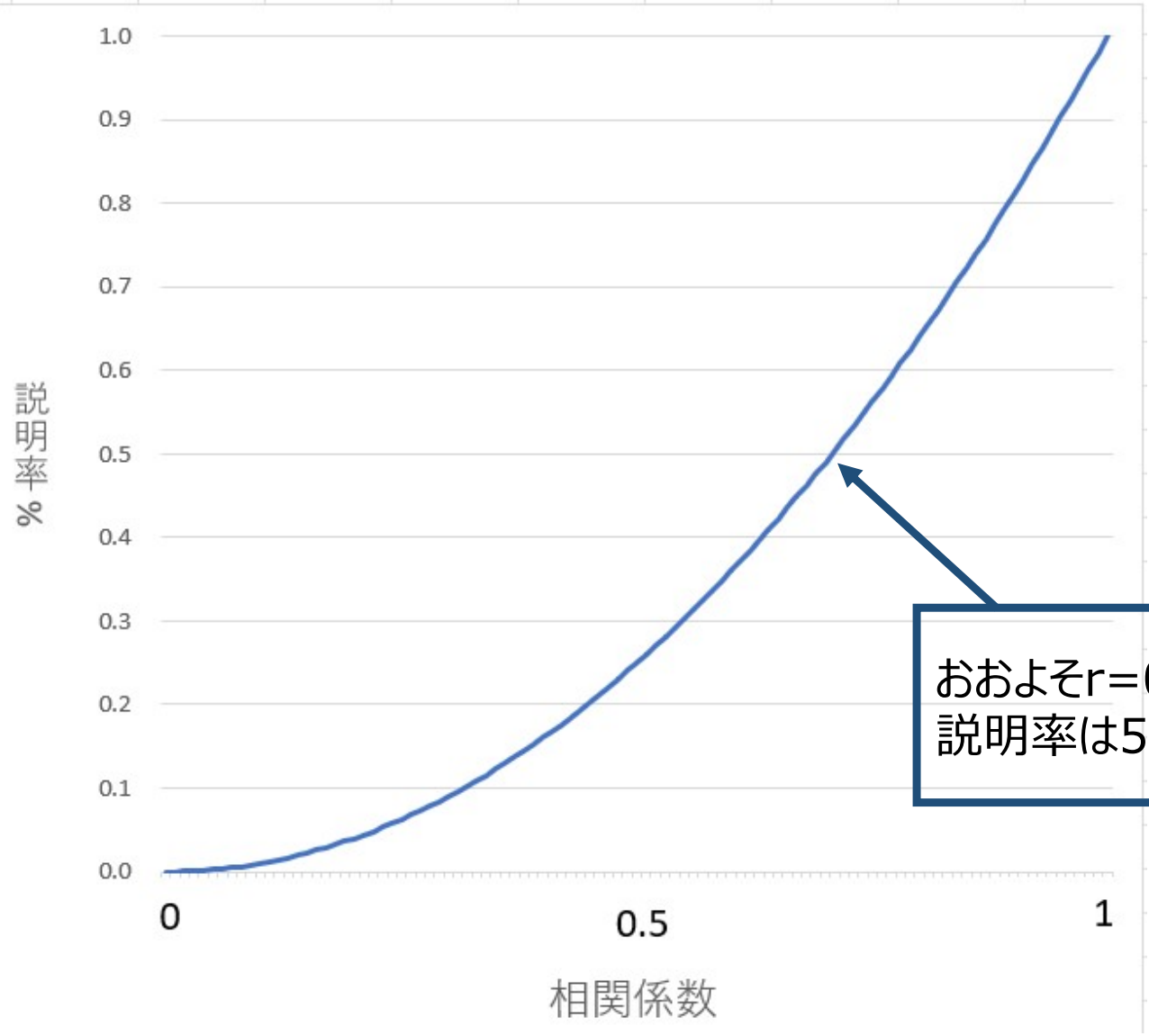
r=0.902

説明力は２乗する

$0.902^2 = 0.814$

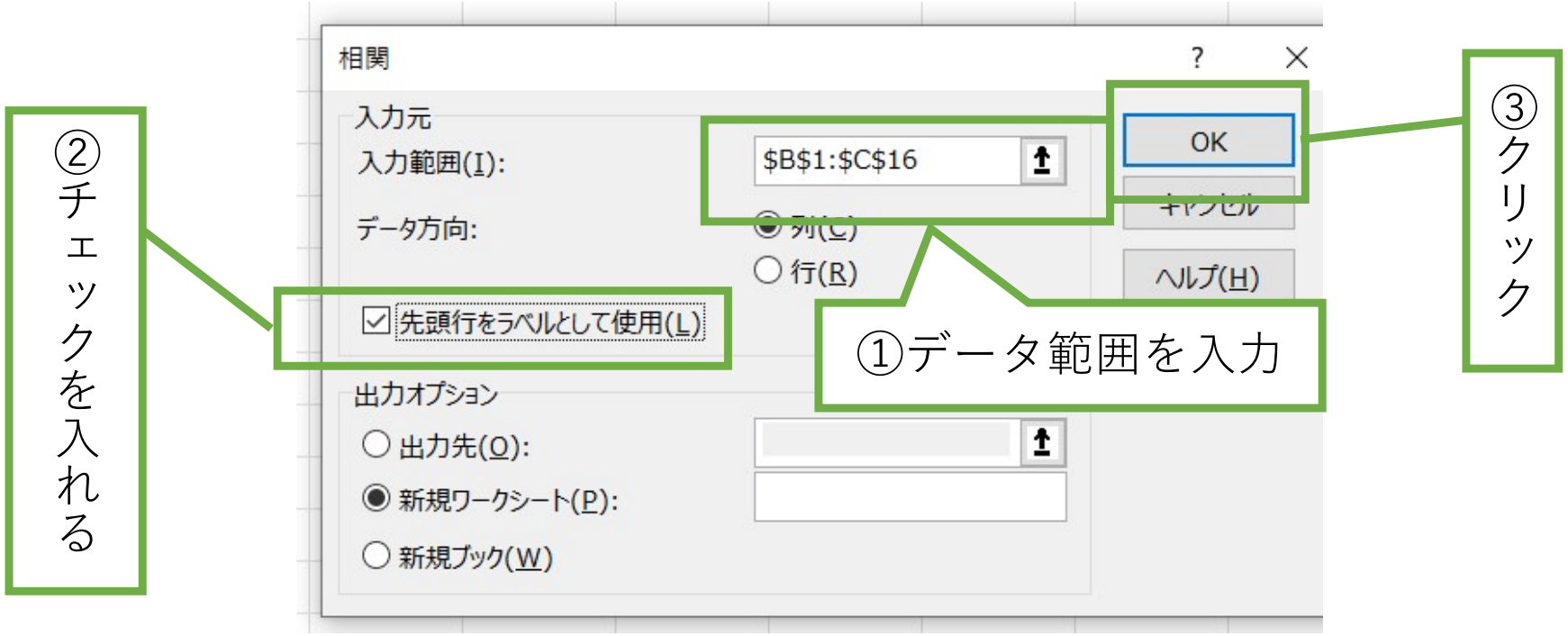
→81.4%

相関の「強さ」


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
説明率%
0
0.5
1
相関係数
おおよそr=0.7以上で
説明率は50%超える

分析ツールでもできる

- 「データ」タブをクリック　→　分析グループの「データ分析」をクリック　→　分析ツールの「相関」をクリック　→　「OK」をクリック



	A	B	C
1		広告費	売上高
2	広告費	1	
3	売上高	0.902359	1
4			

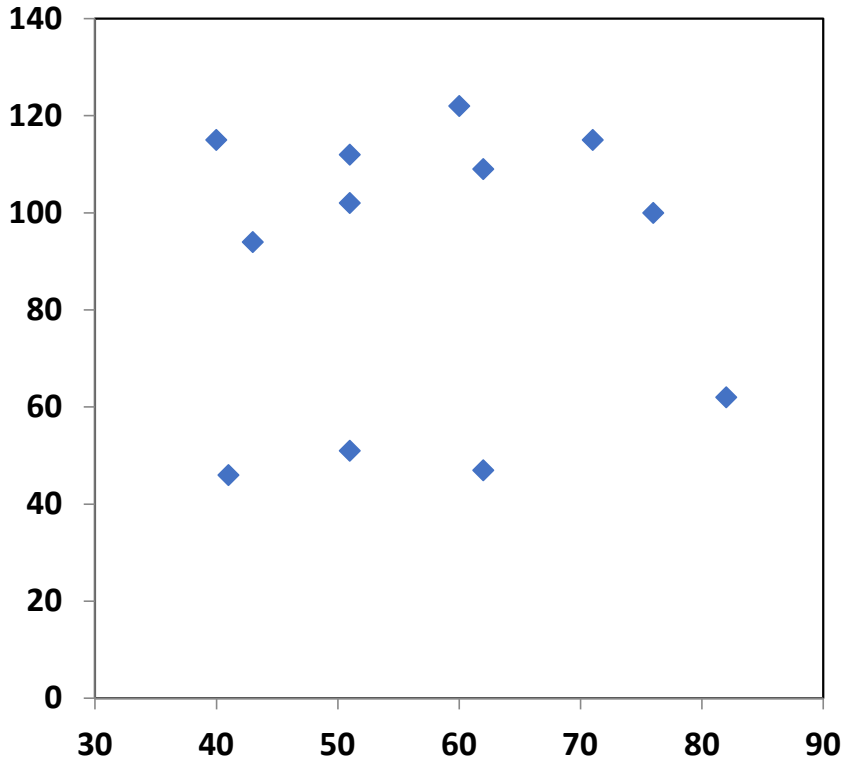
◇A支店の広告費と売上高

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！



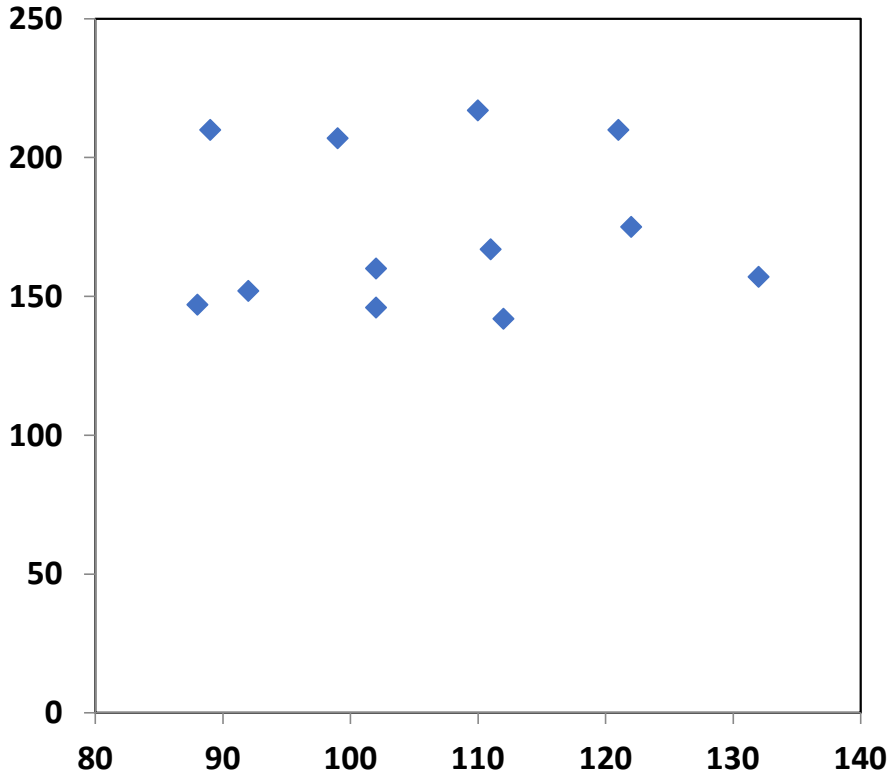
◇B支店の広告費と売上高

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

⇩

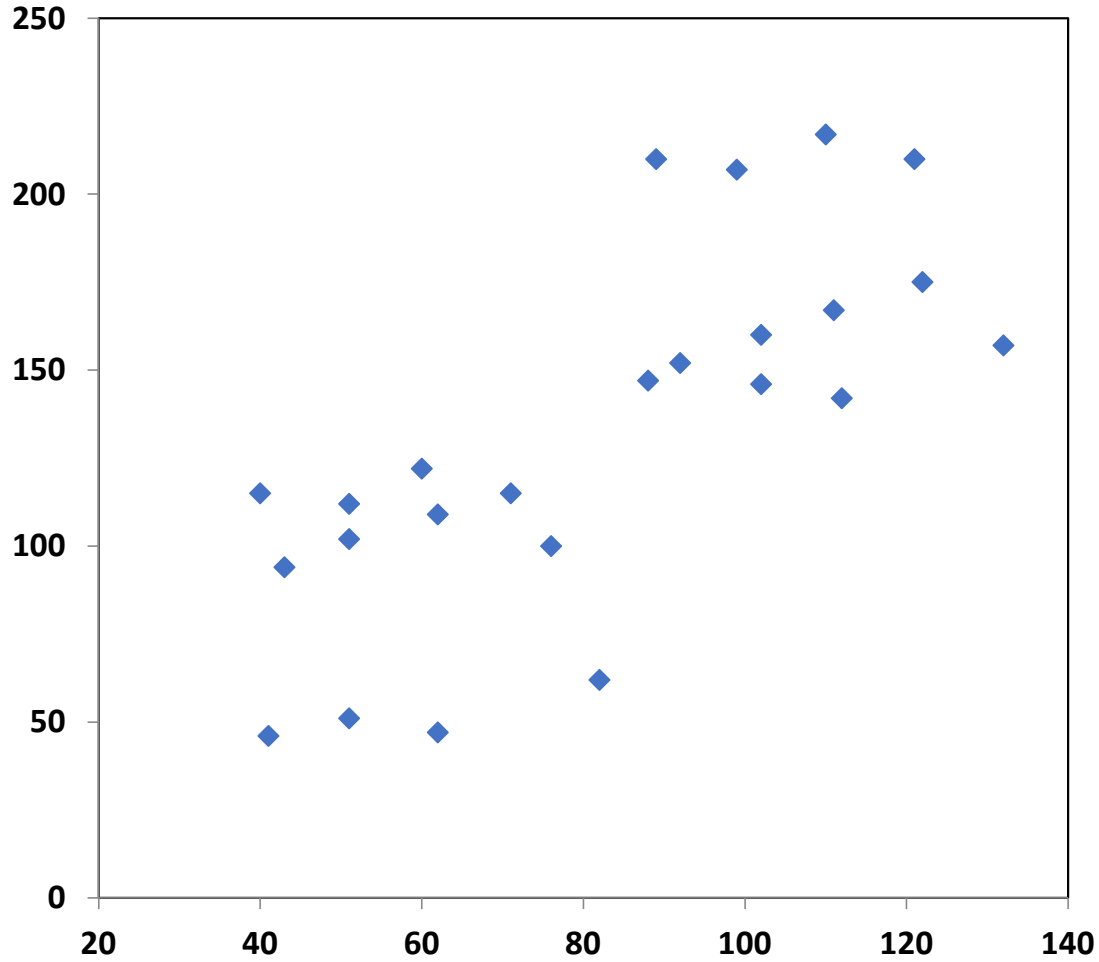
無相関！

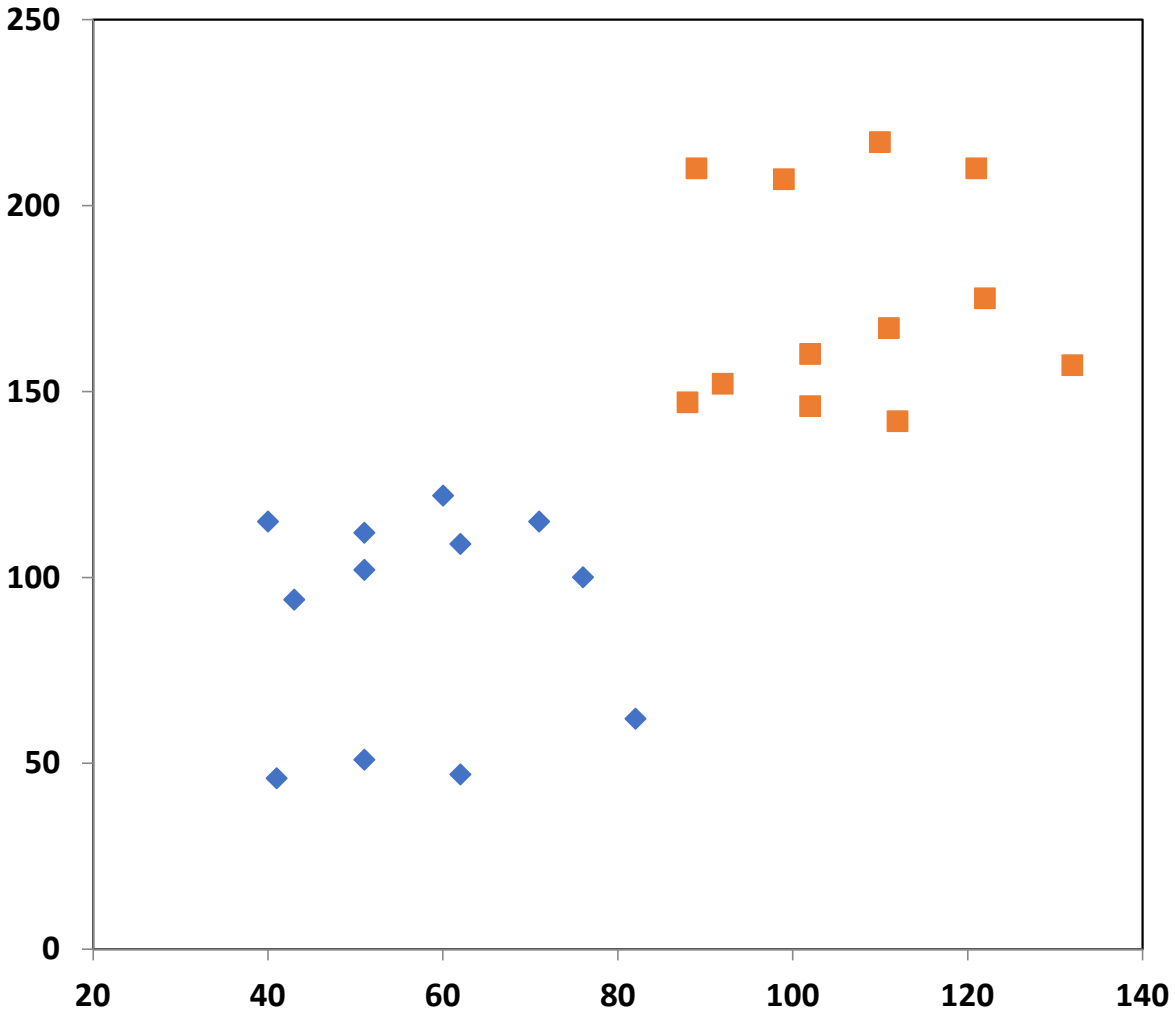


（A支店＋B支店）の広告費と売上高

r ＝0.823

???





擬似相関

相関分析

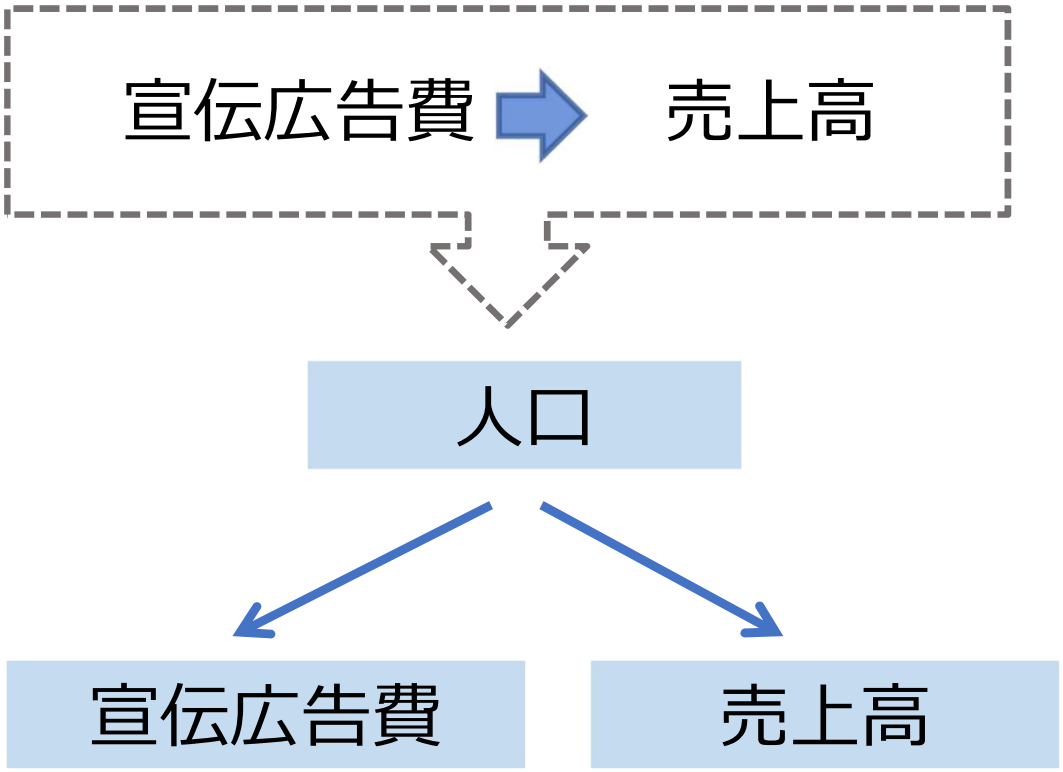
11.交絡要因と偏相関係数

◇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?

相関関係と因果関係



交絡要因（人口）に注意が必要！

◇支店別広告費と売上高

	A	B	C	D
1	支店	広告費	売上高	人口
2	北海道	92	44	5506
3	東北	93	102	9335
4	関東	332	288	42604
5	北陸	78	54	5443
6	中部	181	118	18127
7	近畿	108	138	12912
8	中国	113	138	15554
9	四国	72	86	3976
10	九州	243	152	13204
11	沖縄	13	22	1393

	広告費	売上高	人口
広告費	1		
売上高	0.902	1	
人口	0.895	0.951	1

人口と広告費の相関係数	r=0.895
人口と売上高の相関係数	r=0.951
広告費と売上高の相関係数	r=0.902

人口は交絡要因

人口の影響を除いたときの広告費と売上高の積率相関係数

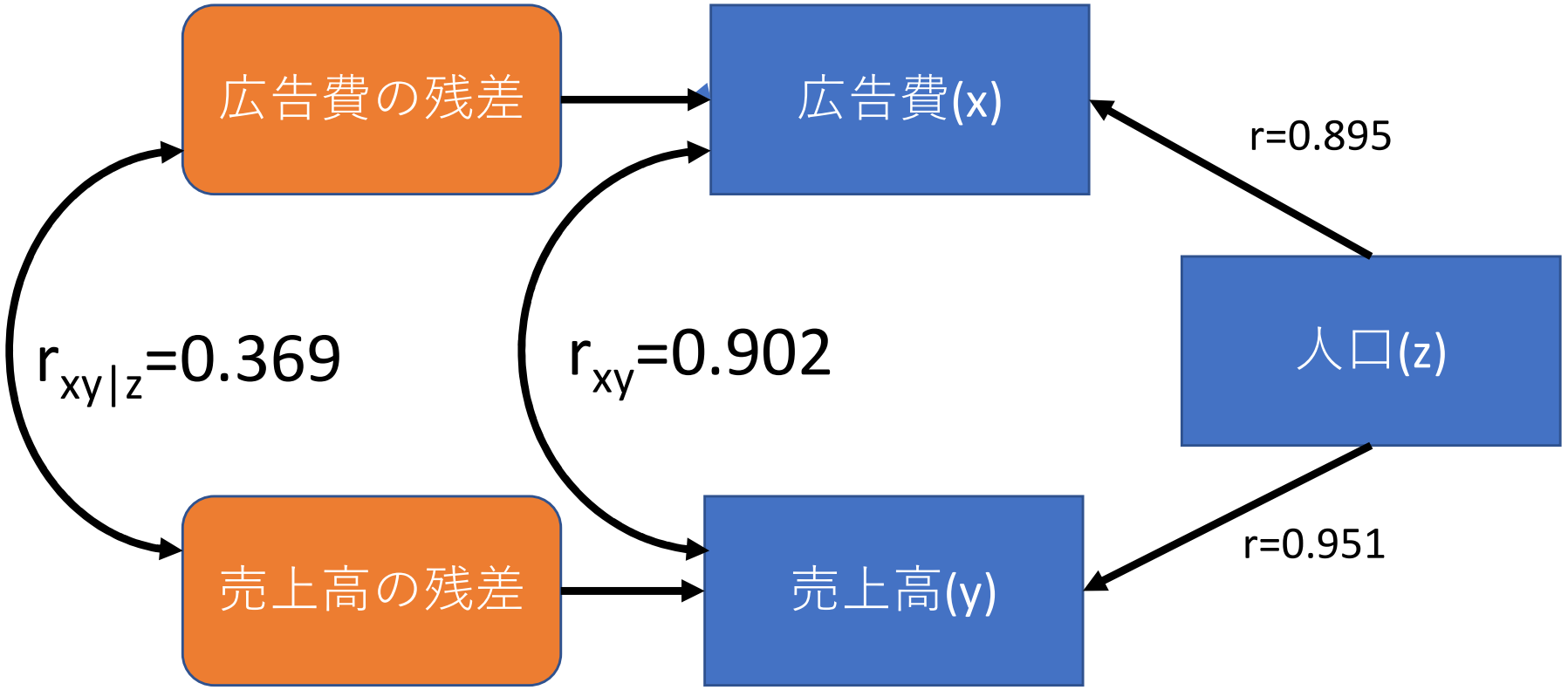
⇩

偏相関係数

偏相関係数の計算

人口の影響を除いたときの広告費と売上高の
偏相関係数　：　0.369

広告費と売上高の関係に人口が及ぼす影響



人口の影響を除くと、広告費と売上高の相関は弱い

回帰分析

12.回帰分析のしくみ

例）売上高と売上高に影響を与える要因との関係

売上高⇐広告宣伝費、人口、セールスマン数、・・・

従属変数(y) ⇐ 説明変数(x)（独立変数）

説明変数が1つ：単回帰
説明変数が2つ以上：重回帰

目的1：回帰式を求め、予測する。

回帰式（ y =a+b x +…）を求める。

a・・・ 切片（定数）
b・・・ 偏回帰係数

回帰分析

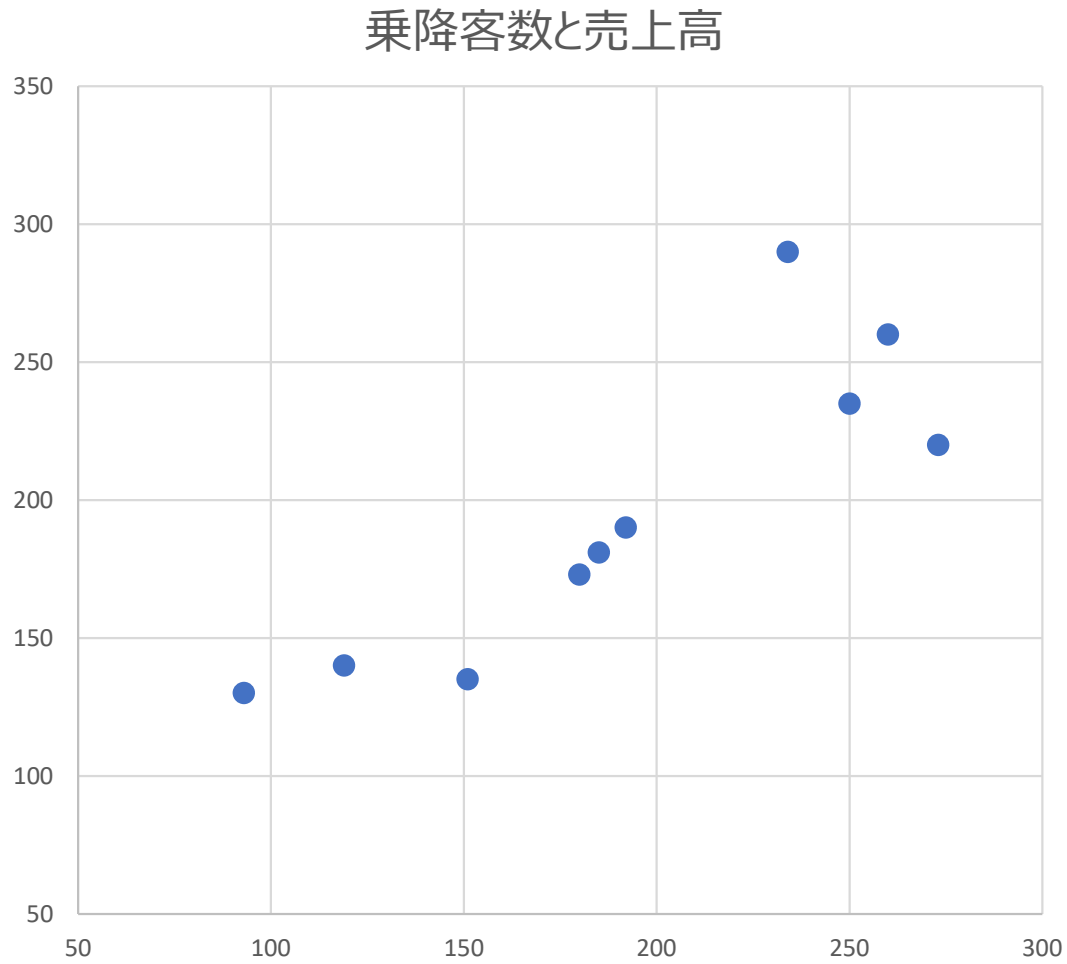
13.単回帰分析

◇駅前コンビニの売上高と乗降客数

売上高⇐乗降客数

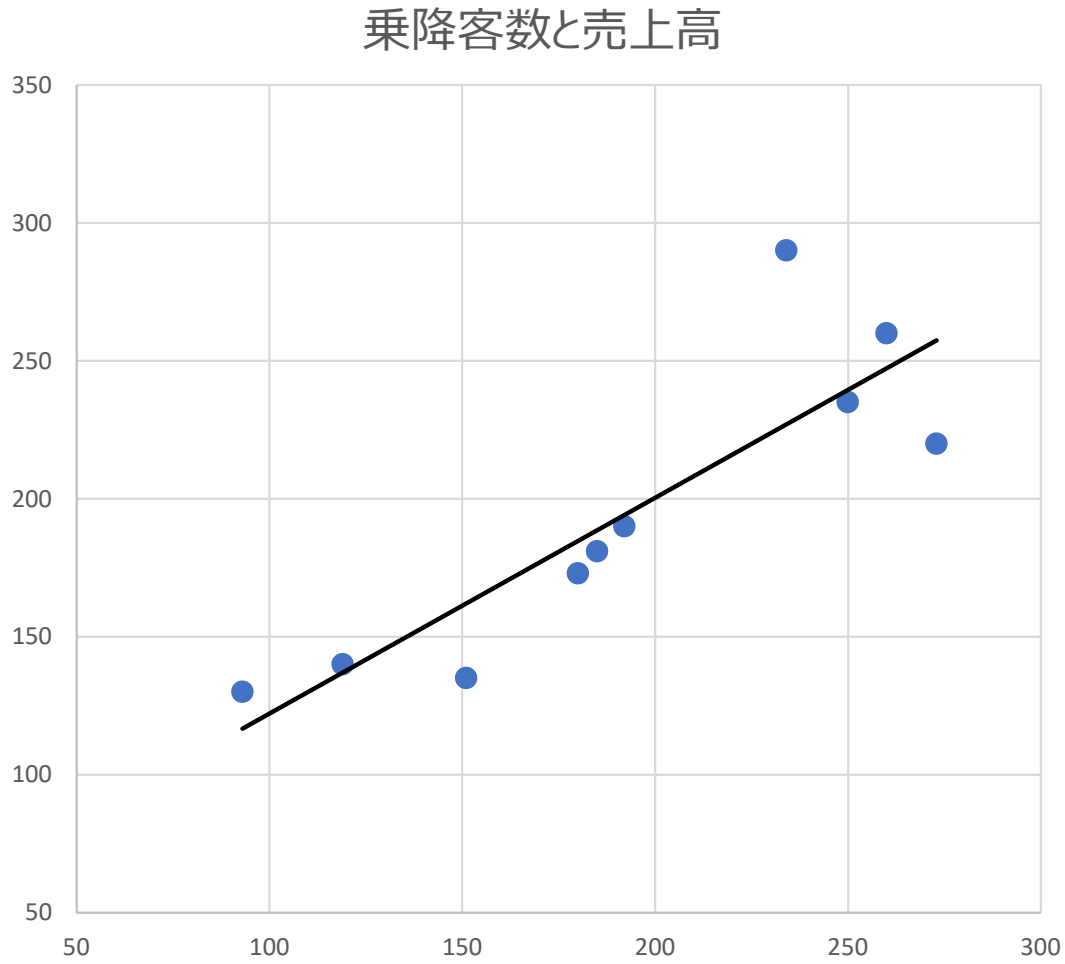
	売上高 （百万円/月）	乗降客数 （百人/日）
1	130	93
2	290	234
3	235	250
4	260	260
5	140	119
6	173	180
7	135	151
8	190	192
9	220	273
10	181	185

相関分析



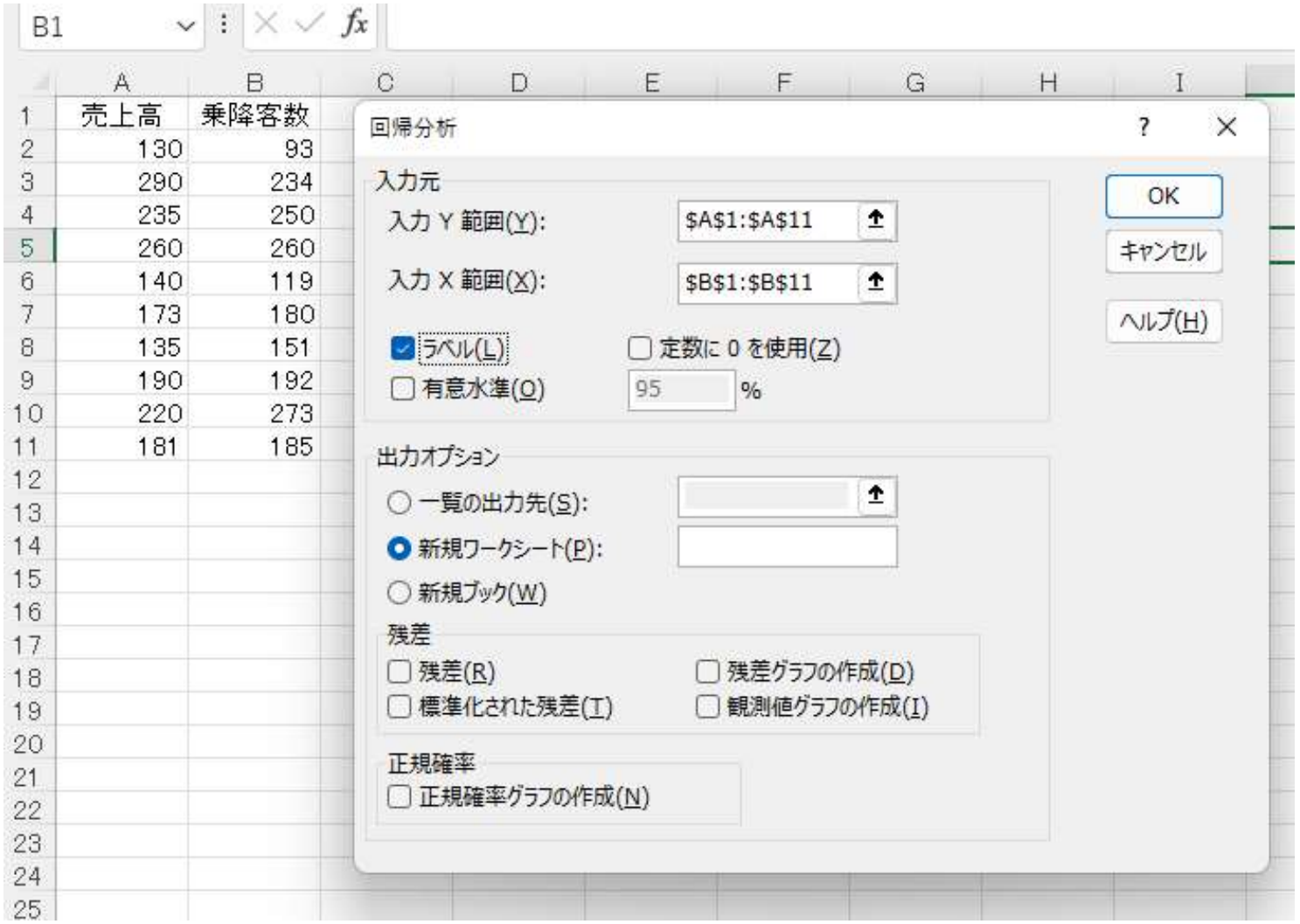
r＝0.867

回帰分析



回帰式（ｙ＝a+b ｘ）を求める。

1）分析ツールメニューから「回帰分析」を選択し、「OK」をクリックする。
2）従属変数、説明変数の範囲を入力する。
3）「ラベル」をチェック、一覧の出力先を指定し「OK」をクリックする。



回帰統計	
重相関 R	0.8675
重決定 R2	0.7525
補正 R2	0.7216
標準誤差	28.916
観測数	10

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	43.992	32.03	1.373	0.2069	-29.88	117.9
乗降客数	0.7817	0.159	4.932	0.0011	0.416	1.147

回帰式：売上高＝43.992＋0.782×乗降客数

回帰式の信頼性

1）決定係数の大きさ（補正R2）

自由度調整済み決定係数＝0.722

⇒　回帰式により約72.2%説明できる

2）偏回帰係数の t 検定

P値の確認

乗降客数の偏回帰係数のP値

0.0011

回帰分析

14.重回帰分析

説明変数に「取扱品目数」を追加

	売上高 （百万円/月）	乗降客数 （百人/日）	取扱品目数 （品）
1	130	93	150
2	290	234	311
3	235	250	182
4	260	260	245
5	140	119	149
6	173	180	160
7	135	151	98
8	190	192	180
9	220	273	113
10	181	185	105

回帰統計	
重相関 R	0.9868
重決定 R2	0.9738
補正 R2	0.9664
標準誤差	10.049
観測数	10

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	7.1538	12.12	0.590	0.573	-21.50	35.82
乗降客数	0.6015	0.060	10.05	0.000	0.460	0.743
取扱品目数	0.4237	0.055	7.697	0.000	0.294	0.554

自由度調整済み決定係数＝ 0.966 ＞ 0.722

重回帰分析の目的1

売上高の予測

売上高＝7.15＋0.602×乗降客数＋0.424×品目数

＊乗降客数=200 取扱品目数が180のとき、売上高の予測

売上高=7.15＋0.602×200＋0.424×180=203.87

◇自由度調整済決定係数

$R^2 = 0.966$

約96.6%説明できる。

回帰分析

15.標準偏回帰係数

重回帰分析の目的2

説明変数の影響力の比較

乗降客数と取扱品目数のどちらの方が影響力が強い？

偏回帰係数の比較

乗降客数：0.602 ＞　取扱品目数：0.424

乗降客数の方が売上高に与える影響は強い !?

入力単位の変更

乗降客数の単位 ： 百人/日 ⇒ 千人/日

	売上高 （百万円/月）	乗降客数 （千人/日）	取扱品目数 （品）
1	130	9.3	150
2	290	23.4	311
3	235	25.0	182
4	260	26.0	245
5	140	11.9	149
6	173	18.0	160
7	135	15.1	98
8	190	19.2	180
9	220	27.3	113
10	181	18.5	105

偏回帰係数の比較

	変更前	変更後
乗降客数	0.60153	6.01533
取扱品目数	0.42368	0.42368

乗降客数の偏回帰係数の値が10倍

t 検定・決定係数の値は同じ。

偏回帰係数はデータの単位の影響を受ける。

偏回帰係数の単純比較は無意味

偏回帰係数の標準化

１）従属変数、説明変数のZ値を求める。

２）Z値を用いて回帰分析を行う。

偏回帰係数の比較

乗降客数　　：0.667

取扱品目数　：0.511

乗降客数の方が影響力は強い。

回帰分析

16.説明変数同士に相関がある場合

説明変数に世帯数（半径500m以内）を追加

	売上高 （百万円/月）	乗降客数 （千人/日）	取扱品目数 （品）	世帯数 （世帯）
1	130	93	150	143
2	290	234	311	284
3	235	250	182	320
4	260	260	245	302
5	140	119	149	182
6	173	180	160	225
7	135	151	98	190
8	190	192	180	242
9	220	273	113	320
10	181	185	105	235

回帰統計	
重相関 R	0.9873
重決定 R2	0.9748
補正 R2	0.9622
標準誤差	10.659
観測数	10

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-1.8266	23.00	-0.079	0.9393	-58.11	54.46
乗降客数	0.4231	0.3843	1.101	0.3131	-0.517	1.363
取扱品目数	0.4214	0.0586	7.194	0.0004	0.278	0.565
世帯数	0.1798	0.3819	0.471	0.6544	-0.755	1.114

◇偏回帰係数のt検定

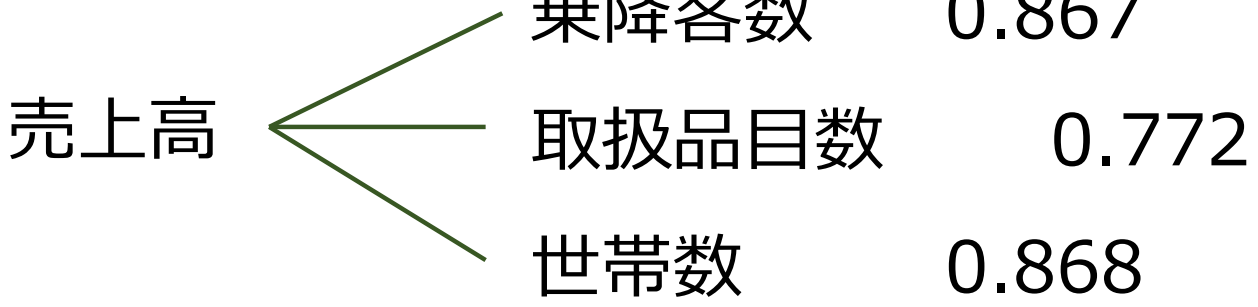
乗降客数　　P値 ＝ 0.3131
取扱品目数　P値 ＝ 0.0004
世帯数　　P値 ＝ 0.6544

乗降客数、世帯数のP値が大きい！

積率相関係数行列

1）分析ツールメニューから「相関」を選択し、「OK」をクリックする。
2）入力範囲を入力する。
3）「ラベル」をチェック、一覧の出力先を指定し「OK」をクリックする。

	売上高	乗降客数	取扱品目数	世帯数
売上高	1			
乗降客数	0.8675	1		
取扱品目数	0.7722	0.3911	1	
世帯数	0.8679	0.9884	0.3979	1



いずれも高い値⇒売上高を説明する説明変数として妥当

乗降客数	取扱品目数	0.391
乗降客数	世帯数	0.988
取扱品目数	世帯数	0.398

乗降客数と世帯数の値0.988が高い

説明変数相互の積率相関係数は低い方が良い！

説明変数⇒独立変数

回帰分析

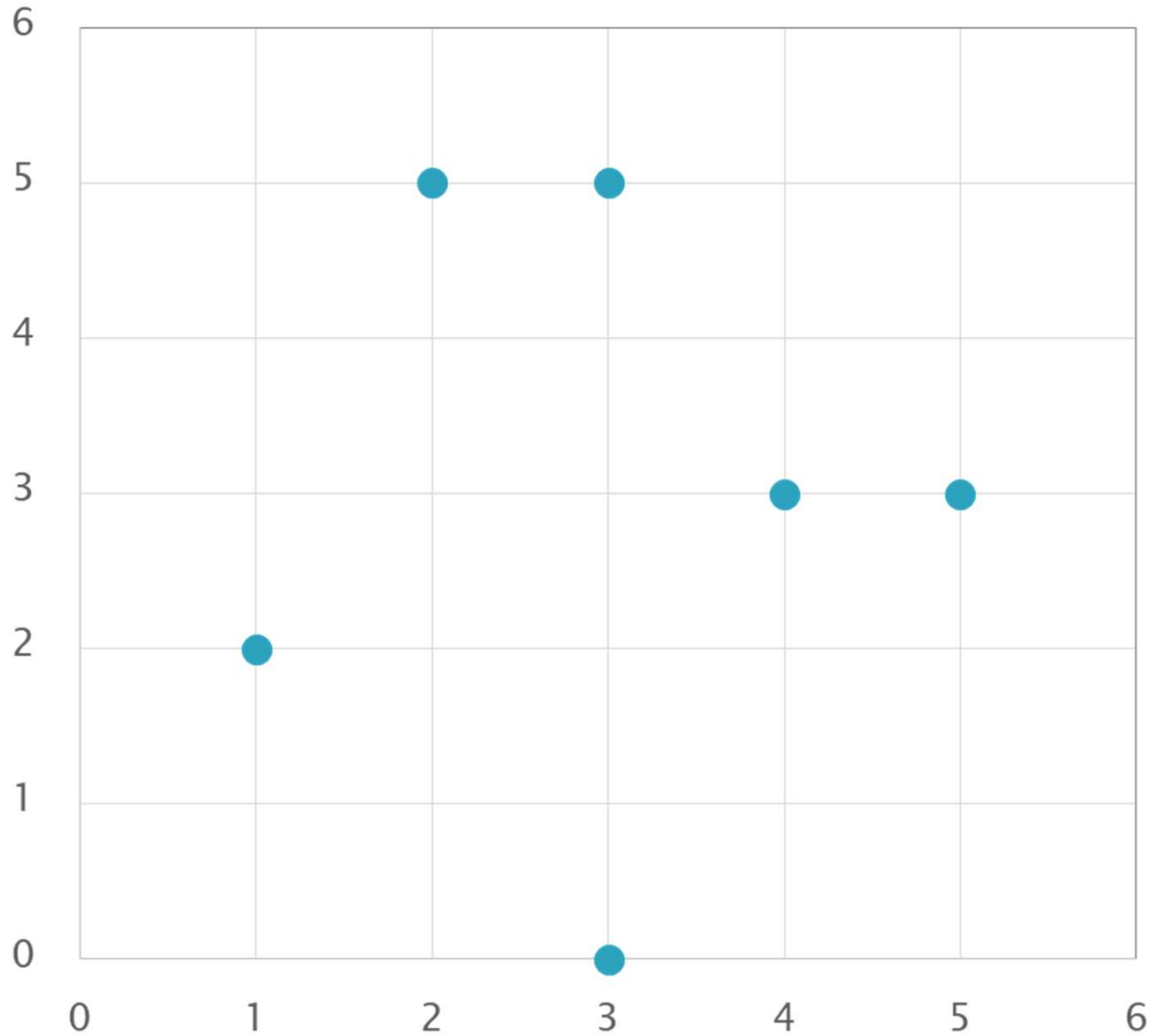
17.ダミー変数を用いた分析

◇訪問回数と成約件数（1）

成約件数	訪問回数
2	1
5	2
5	3
0	3
3	4
3	5

訪問回数と成約件数の関係は？

散布図



積率相関係数＝0

◇訪問回数と成約件数（2）

成約件数	訪問回数	性別
2	1	女性
5	2	女性
5	3	女性
0	3	男性
3	4	男性
3	5	男性

成約件数と、訪問回数、性別の関係は？

性別データの数値化  男性＝1、女性＝0

成約件数	訪問回数	ダミー
2	1	0
5	2	0
5	3	0
0	3	1
3	4	1
3	5	1

回帰統計	
重相関 R	0.9128
重決定 R2	0.8333
補正 R2	0.7222
標準誤差	1
観測数	6

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	1.0	1.155	0.866	0.4502	-2.675	4.675
訪問回数	1.5	0.500	3.000	0.0577	-0.091	3.091
ダミー	-5.0	1.290	-3.873	0.0305	-9.109	-0.891

成約件数＝1.0＋1.5×訪問回数−5×ダミー

ダミー変数の解釈

○ダミーの偏回帰係数＝ −5

訪問回数が同じであれば、男性は女性に比べて
平均的に成約件数が５件少ない。

○訪問回数の偏回帰係数＝1.5

性別が同じであれば、訪問回数が1回増えるごとに
平均的に成約件数が1.5件ずつ多い。

ダミー変数の作成方法

（2区分）	ダミー1
有	1
無	0

区分数 − 1　⇒　ダミー変数

（3区分）	ダミー1	ダミー2
大	1	0
中	0	1
小	0	0

（4区分）	ダミー1	ダミー2	ダミー3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

売上数 ⇐ 曜日、温度

売上数	曜日	温度
356	月	28
245	火	21
128	水	15
189	木	15
215	金	28
412	土	24
388	日	22
312	月	19
301	火	22
355	水	25
…	..	..

曜日：7区分

ダミー変数 ： 6個

（6個のダミー変数を
ワンセットで用いる）

入力データ

曜日	売上数	ダミー1	ダミー2	ダミー3	ダミー4	ダミー5	ダミー6	温度
月	356	1	0	0	0	0	0	28
火	245	0	1	0	0	0	0	21
水	128	0	0	1	0	0	0	15
木	189	0	0	0	1	0	0	15
金	215	0	0	0	0	1	0	28
土	412	0	0	0	0	0	1	24
日	388	0	0	0	0	0	0	22
月	312	1	0	0	0	0	0	19
火	301	0	1	0	0	0	0	22
水	355	0	0	1	0	0	0	25
	・・・	・・	・・	・・	・・	・・	・・	・・

（日曜日をベースラインとする）ダミー変数

■■　実践統計学　■■

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

2023年9月1日　第8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

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝