ビジネスパーソン必須の統計的思考法-2

2022年9月15日

Copyright © SAS Institute Inc. All rights reserved.

本日の内容

・統計学の基本的な考え方

■データ群の代表値

売上高の平均、研修成績の平均・・・

■データ群の違いの大きさ

男女の好感度、CM前後の知名度・・・

■データ群の関係

広告費と売上高、DM発送数と売上高・・・

・SAS Studioのご紹介

Copyright © SAS Institute Inc. All rights reserved.

データ群の代表値

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	75	58.3	+16.7

A君の成績の評価は？

平均を評価基準とする合理性？

Copyright © SAS Institute Inc. All rights reserved.

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（分）＞

											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。

Copyright © SAS Institute Inc. All rights reserved.

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＞

  75、31、95、36、89、88、89
  76、28、47、23、28、96、48
  51、20、30、91、27、98

研修後のバラツキが大きい！

Copyright © SAS Institute Inc. All rights reserved.

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＞

  75、31、95、36、89、88、89
  76、28、47、23、28、96、48
  51、20、30、91、27、98

研修後の成績順位は下がった！

Copyright © SAS Institute Inc. All rights reserved.

成績の評価

＜平均を基準とした場合＞

研修前（+11.7） ＜ 研修後（+16.7）

＜順位＞

研修前（3番） ＞ 研修後（9番）

バラツキ

研修前 ＜ 研修後

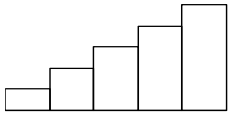
「バラツキ」の大きいデータの平均は信頼できない。

Copyright © SAS Institute Inc. All rights reserved.

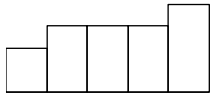
「バラツキ」の大きさを求める

A群～D群の平均値はすべて同じ値3

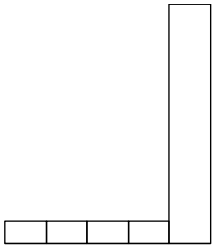
A: {1,2,3,4,5}



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

Copyright © SAS Institute Inc. All rights reserved.

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！

Copyright © SAS Institute Inc. All rights reserved.

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。

Copyright © SAS Institute Inc. All rights reserved.

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」

Copyright © SAS Institute Inc. All rights reserved.

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	=		20

E群の偏差平方和=20

Copyright © SAS Institute Inc. All rights reserved.

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」

Copyright © SAS Institute Inc. All rights reserved.

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	400千円^2
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：1000千円^2

F群の分散：200千円^2

A群の分散：2万円^2

分散の比較は困難

分散の平方根は、同じ値となる。

F群：$\sqrt{200}$ ≒ 14.14千円

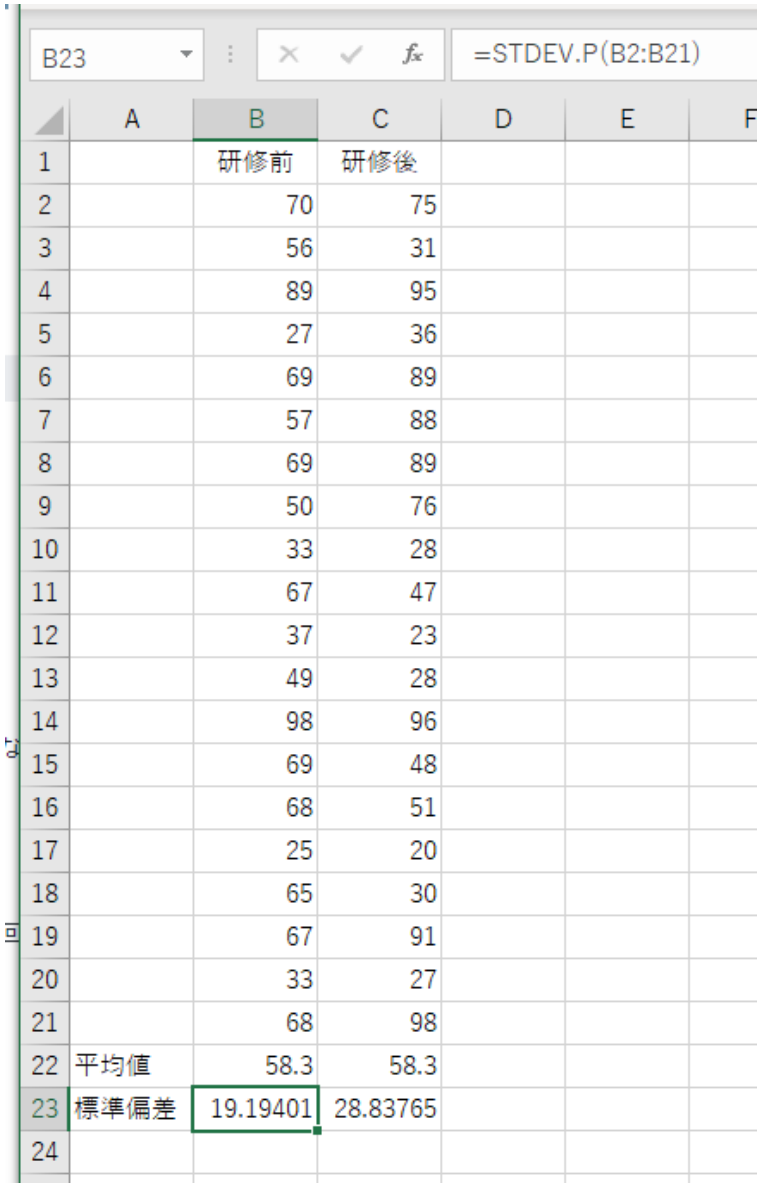
A群：$\sqrt{2}$ ≒ 1.414万円

バラツキの代表値②
「標準偏差」

Copyright © SAS Institute Inc. All rights reserved.

EXCEL

STDEV.P（標準偏差）

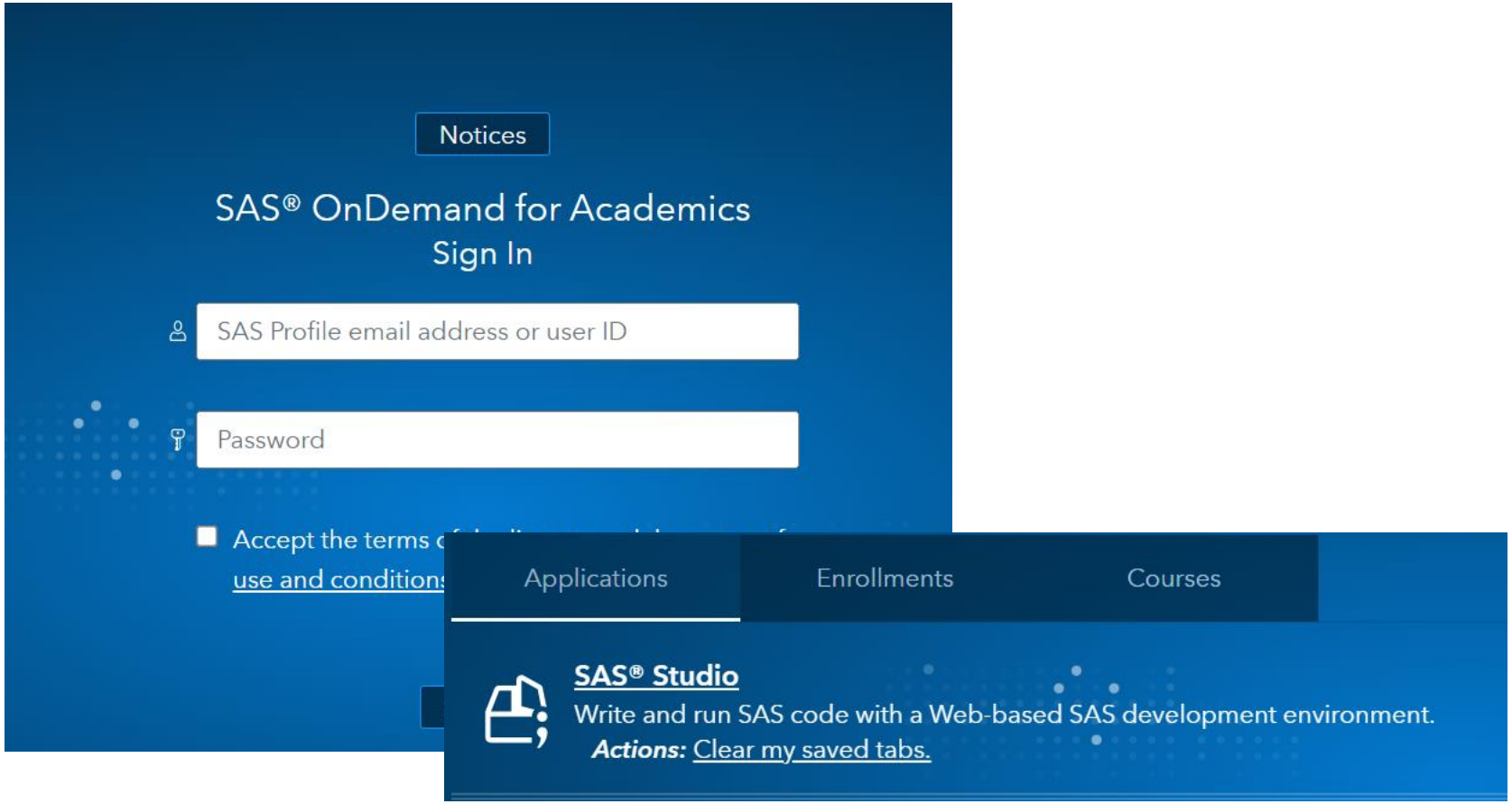


	A	B	C	D	E	F
1		研修前	研修後			
2		70	75			
3		56	31			
4		89	95			
5		27	36			
6		69	89			
7		57	88			
8		69	89			
9		50	76			
10		33	28			
11		67	47			
12		37	23			
13		49	28			
14		98	96			
15		69	48			
16		68	51			
17		25	20			
18		65	30			
19		67	91			
20		33	27			
21		68	98			
22	平均値	58.3	58.3			
23	標準偏差	19.19401	28.83765			
24						

Copyright © SAS Institute Inc. All rights reserved.

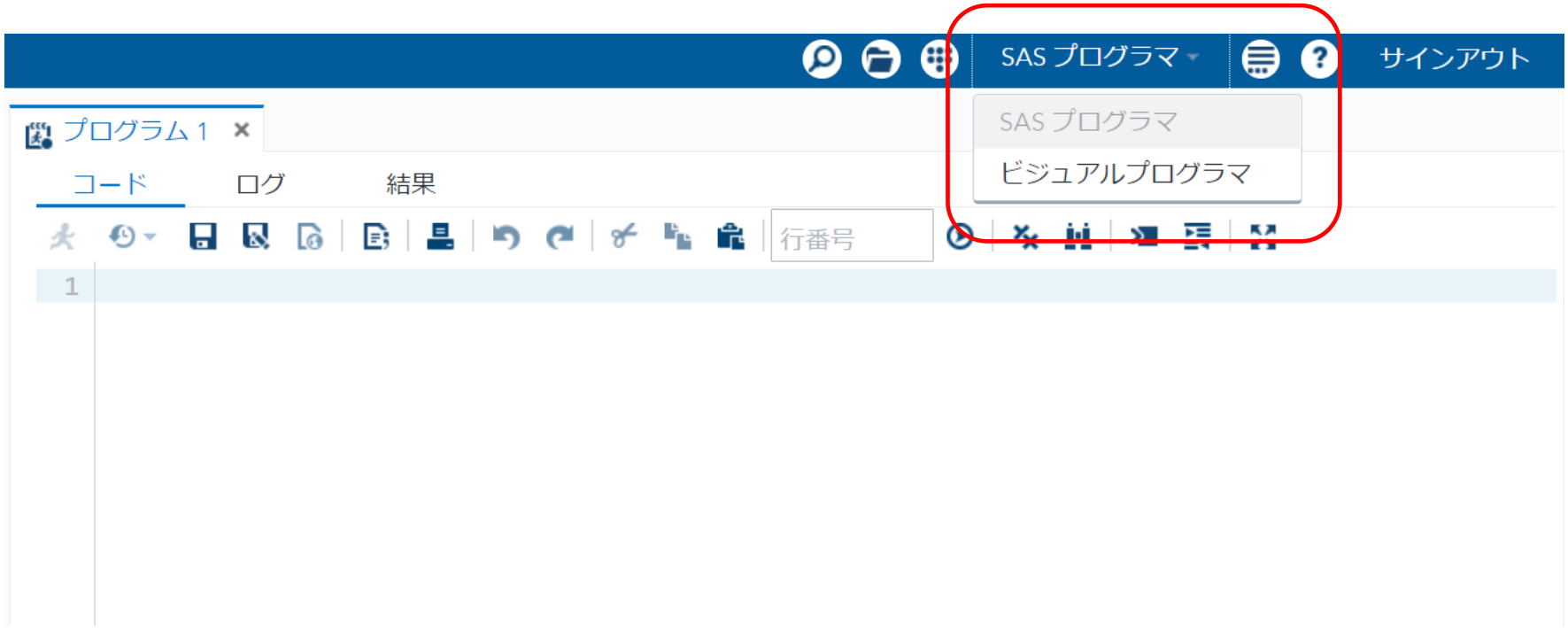
SAS Studio

1.SAS Studio にログインする。
2.SAS®Studio をクリックする。



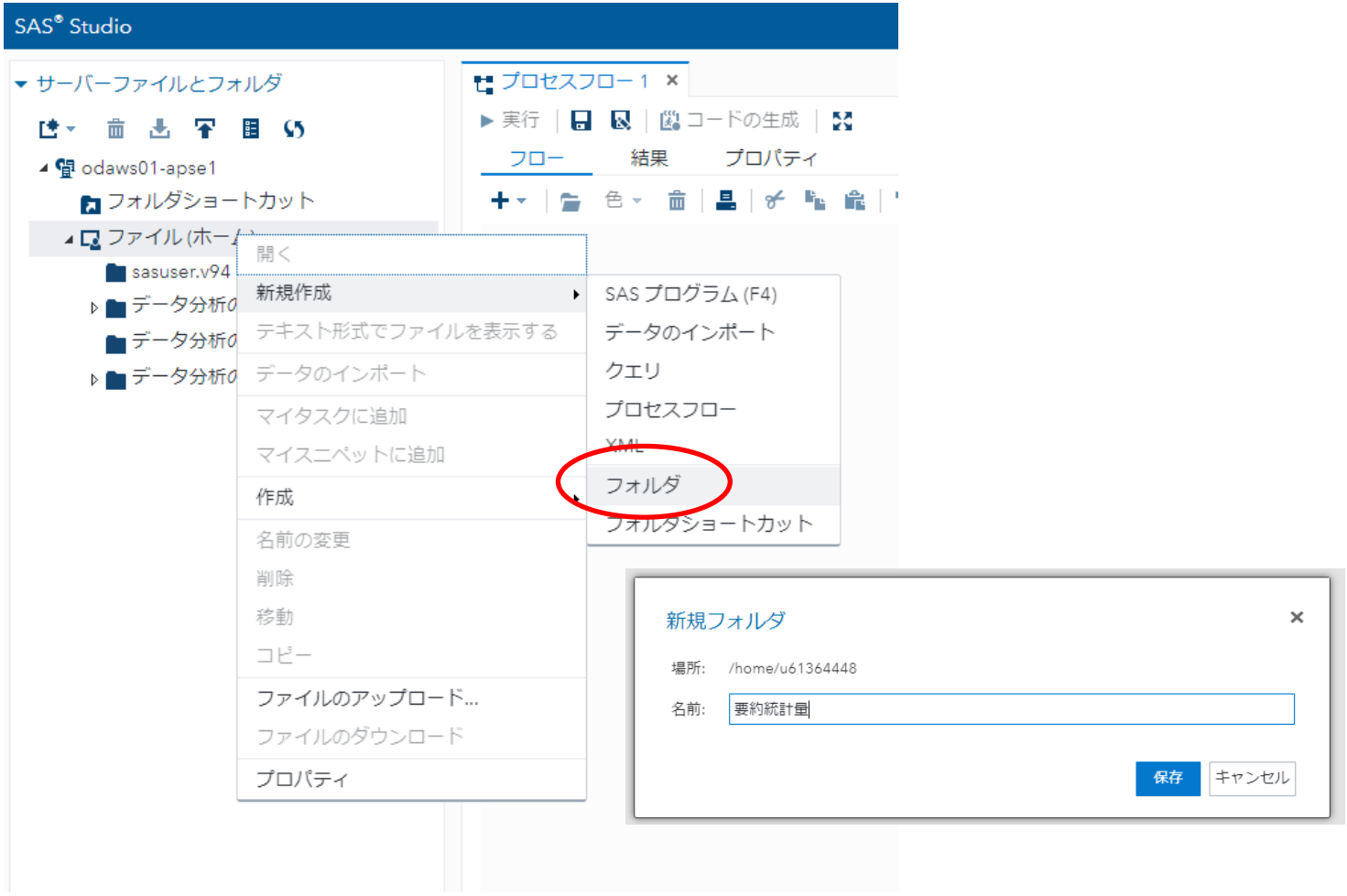
Copyright © SAS Institute Inc. All rights reserved.

3.「SASプログラマ」をクリックし、「ビジュアルプログラマ」を選択する。



Copyright © SAS Institute Inc. All rights reserved.

4. 新規フォルダを作成する。
「ファイル（ホーム）」を右クリックー「新規作成」ー「フォルダ」をクリック、
「新規フォルダ名（要約統計量）」を入力し、「保存」をクリックする。

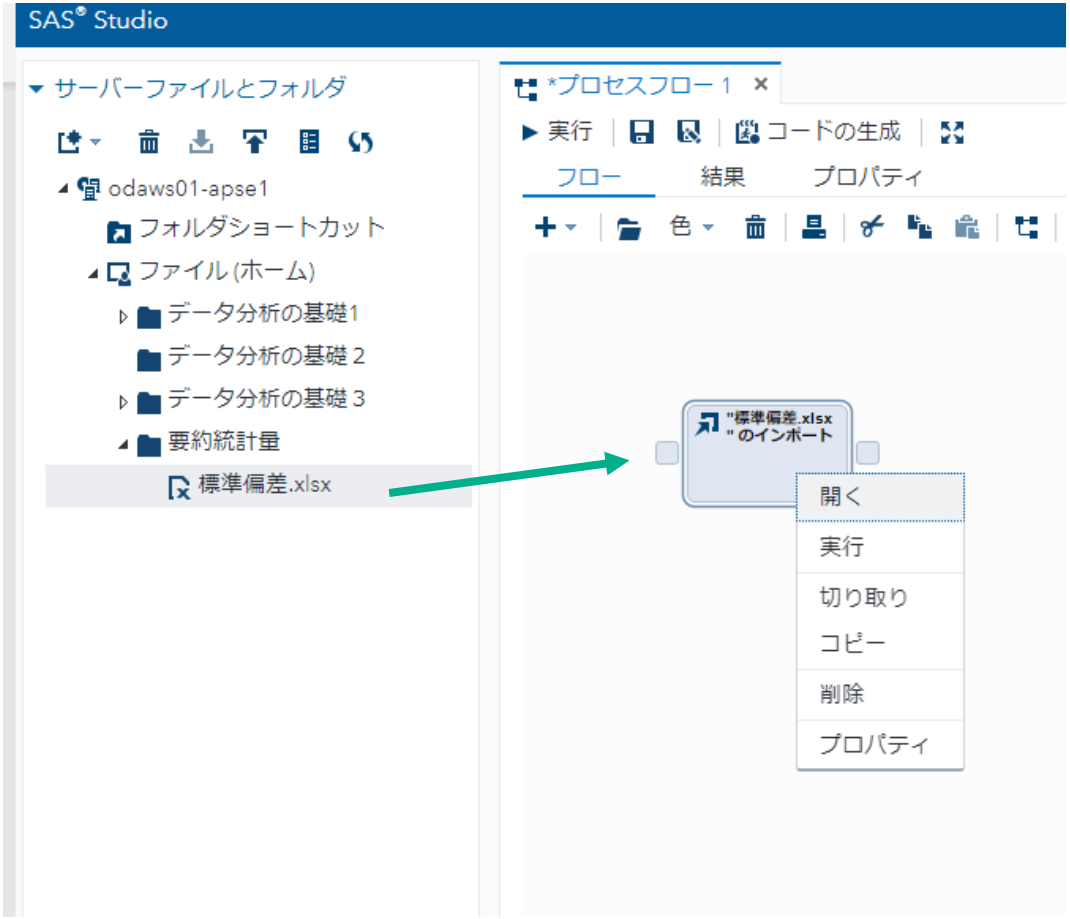


Copyright © SAS Institute Inc. All rights reserved.

5. ファイルをアップロードする。
「要約統計量」フォルダを右クリック、「ファイルのアップロード」をクリックする。
「ファイルの選択」から、ファイルを選択し「アップロード」をクリックする。

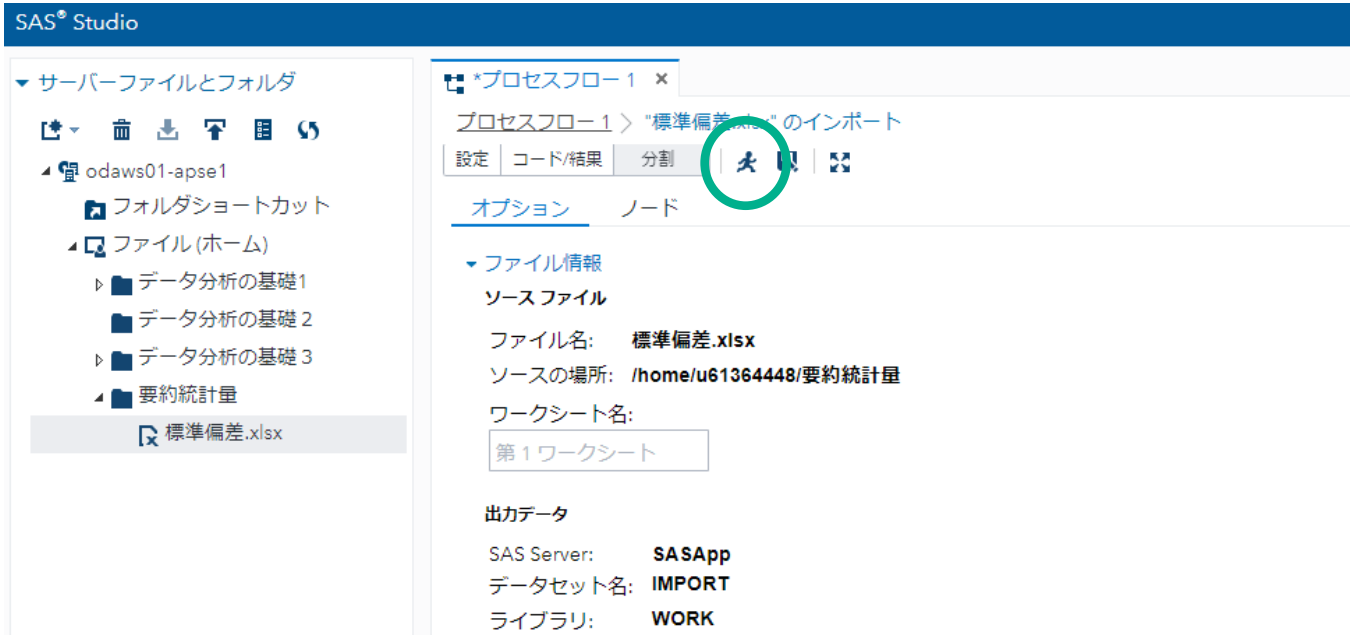
Copyright © SAS Institute Inc. All rights reserved.

6. フォルダを開きファイルを右側のプロセスフロー画面にドラッグし、右クリックして「開く」を選択する。



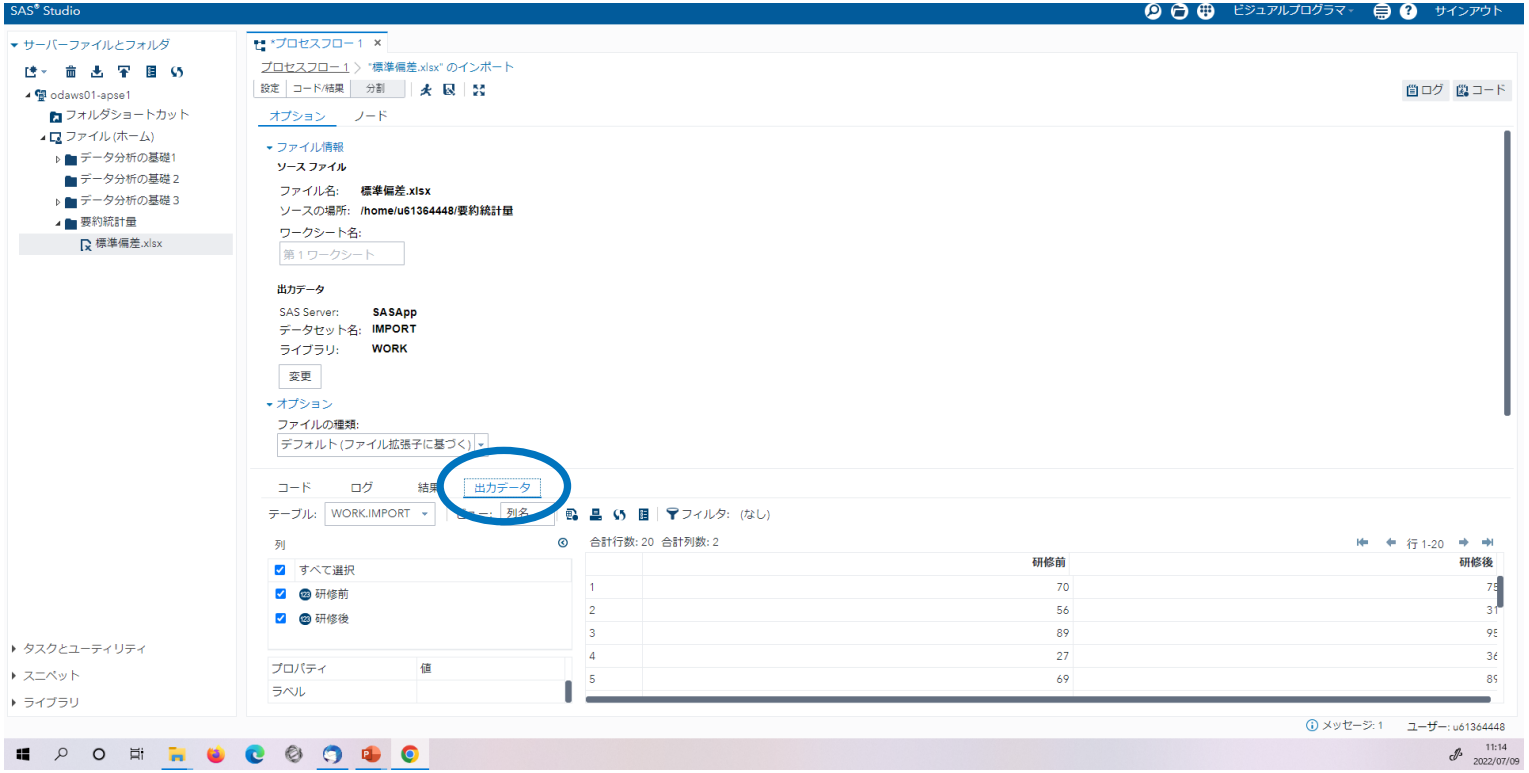
Copyright © SAS Institute Inc. All rights reserved.

7. 実行ボタンをクリックすると、データが出力される。



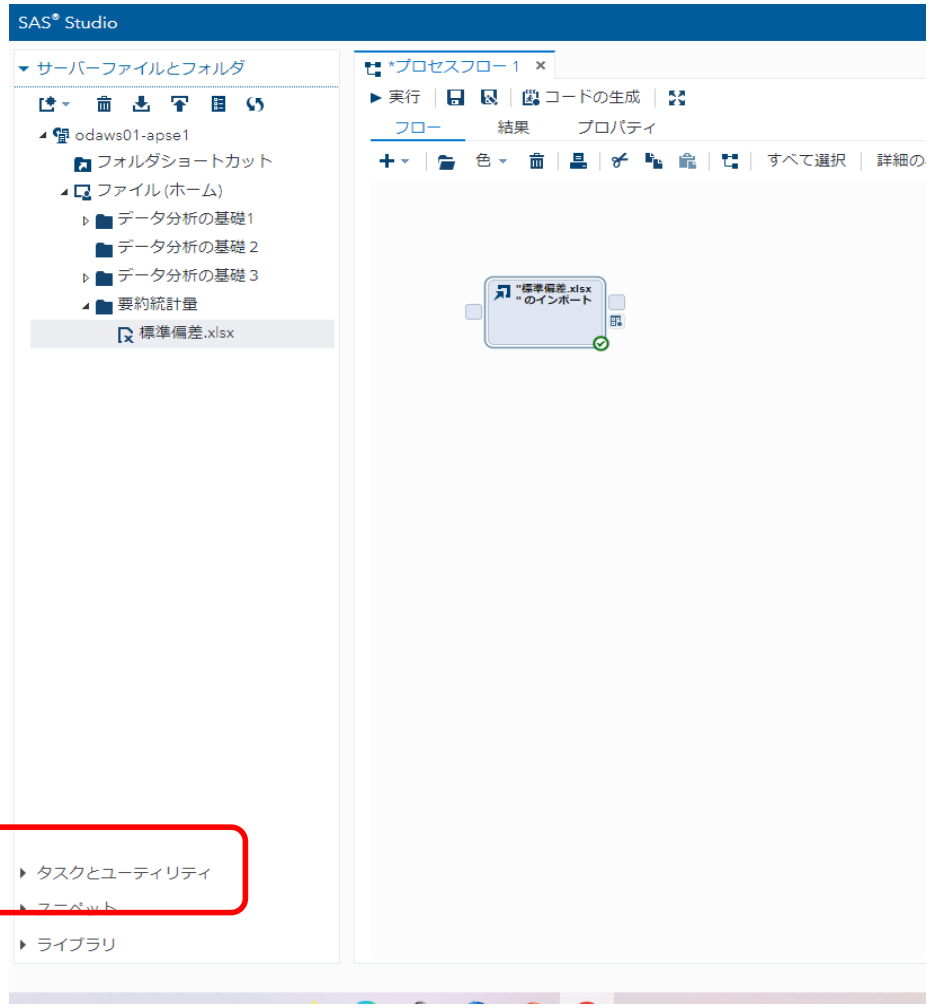
Copyright © SAS Institute Inc. All rights reserved.

8.「出力データ」をクリックし、「データセット名」、データの内容を確認する。



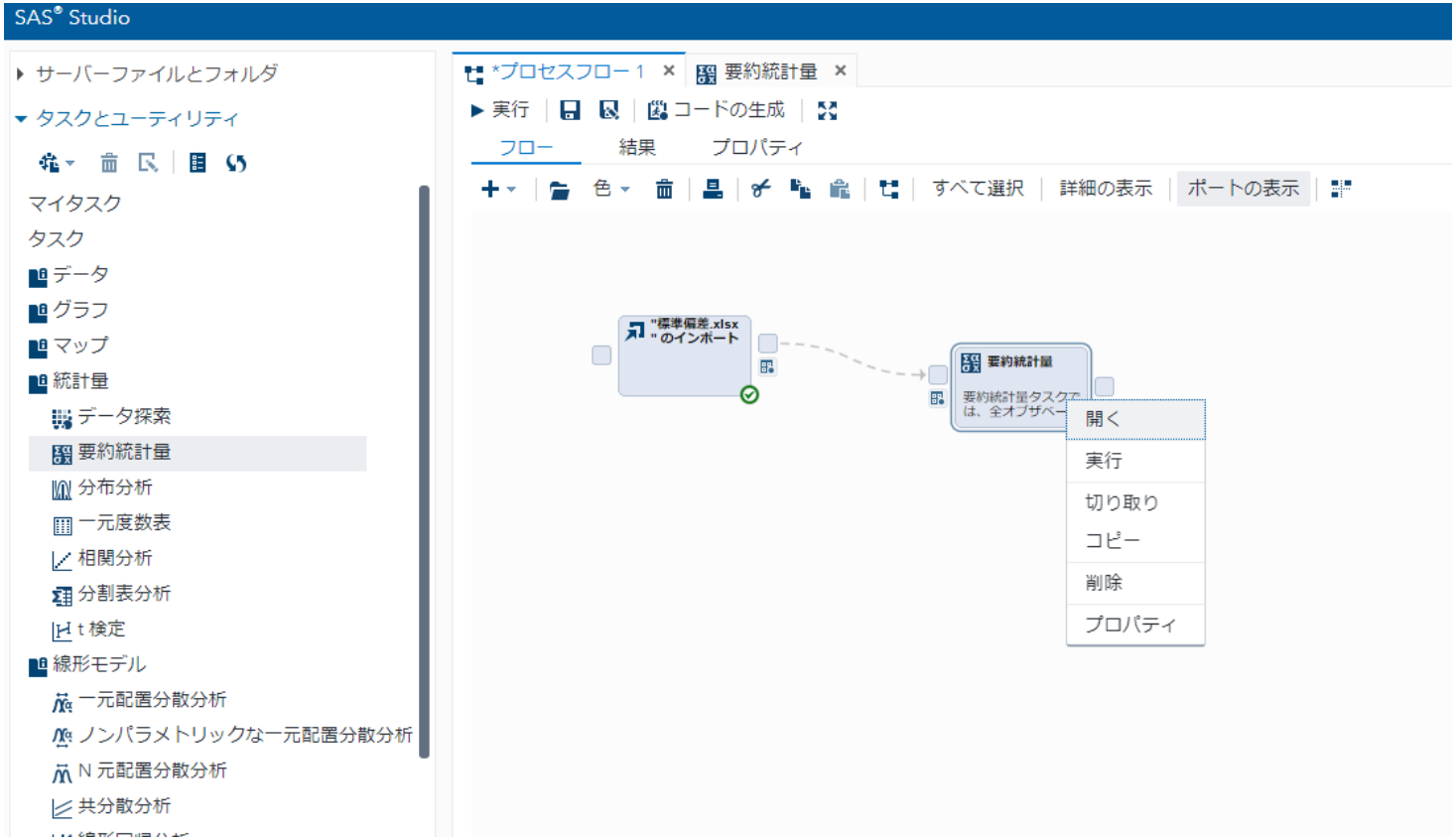
Copyright © SAS Institute Inc. All rights reserved.

9. 「プロセスフロー」をクリックしてフロー画面に戻り、「タスクとユーティリティ」を開く。



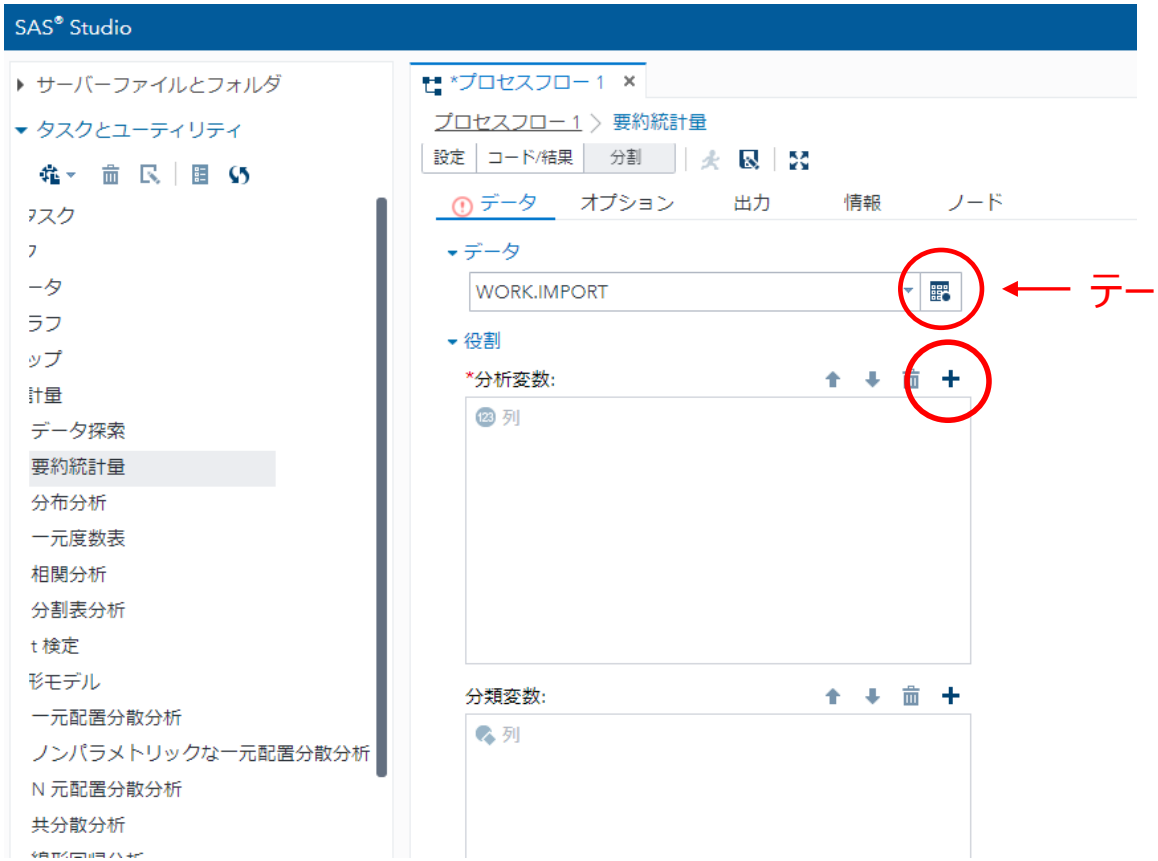
Copyright © SAS Institute Inc. All rights reserved.

10.「タスク」–「統計量」をクリックし、「要約統計量」をフロー画面にドラッグし「標準偏差・・」と結合、右クリック–「開く」をクリックする。



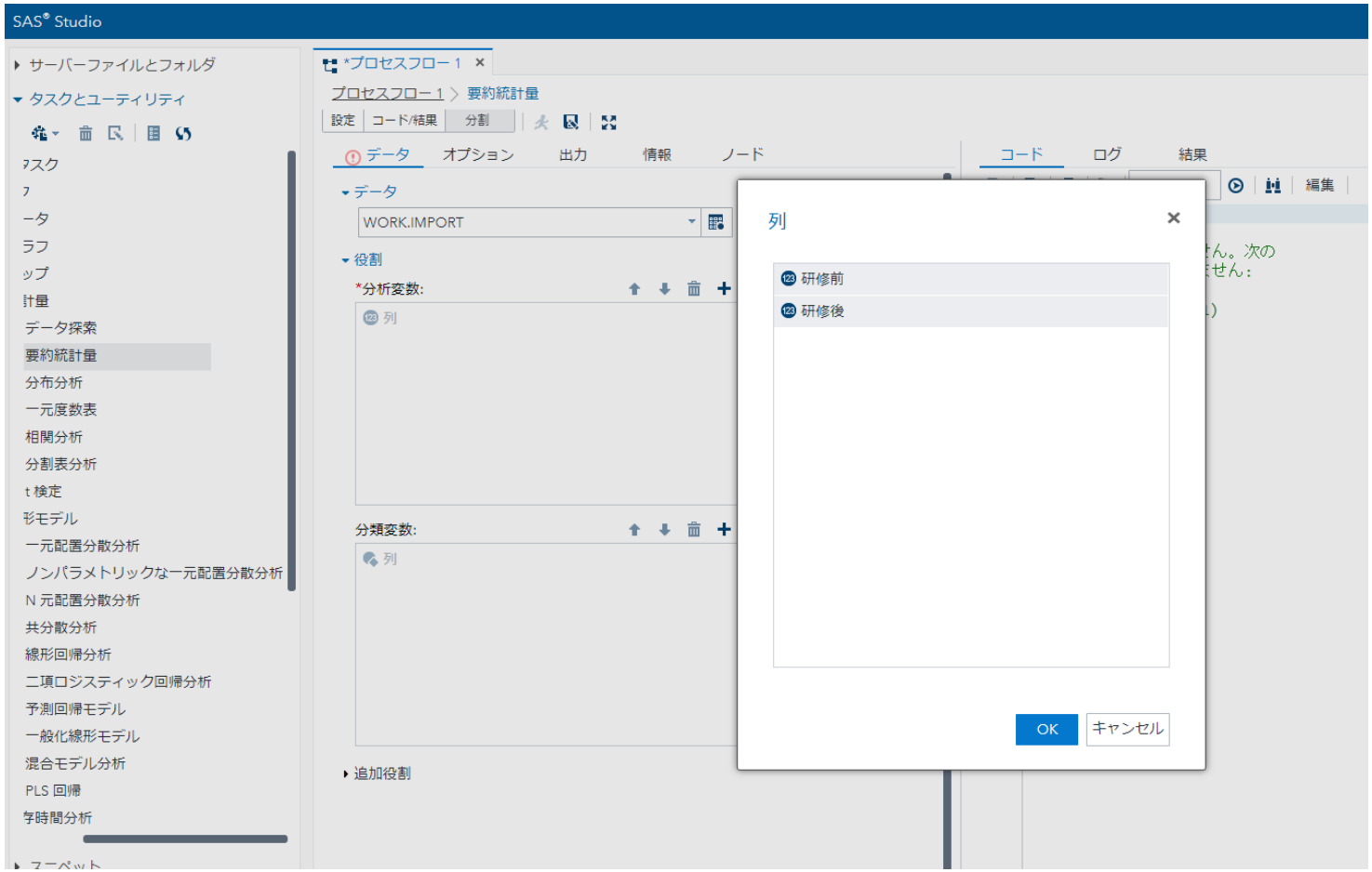
Copyright © SAS Institute Inc. All rights reserved.

11. 「データ」の「データセット名」を確認する。変更する場合は「テーブルの選択」から「データセット」を選択する。
12. 「分析変数」をセットする。（右の＋マークをクリック）



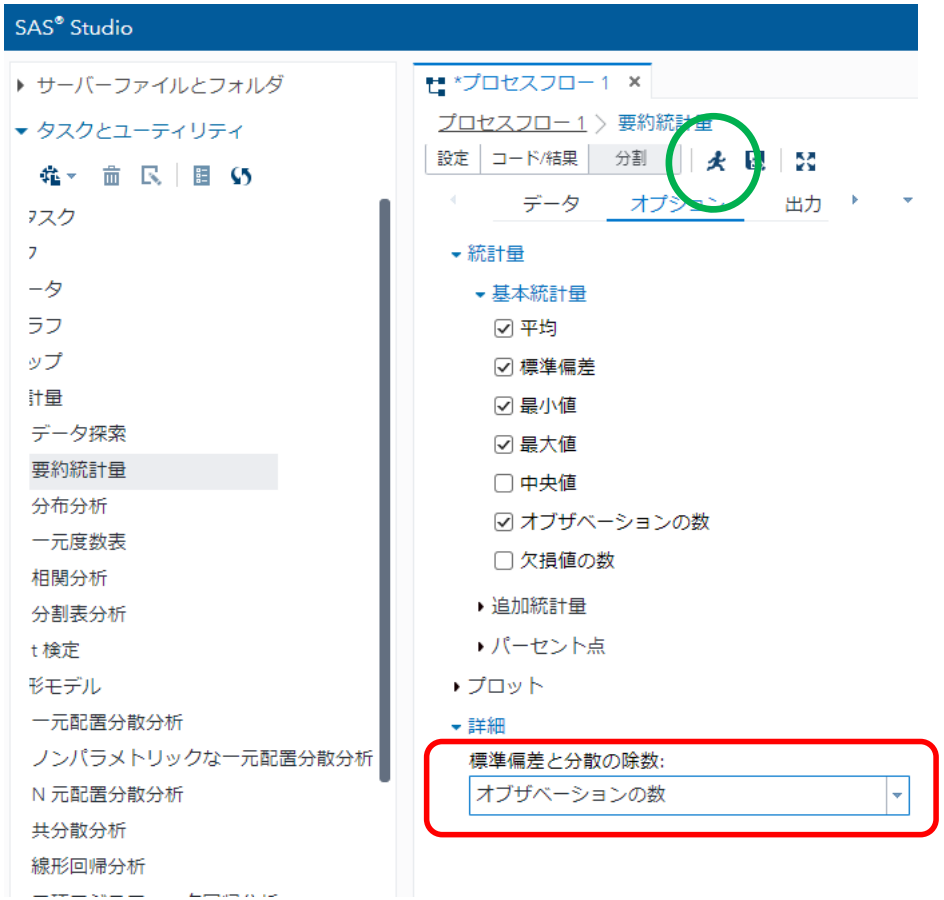
Copyright © SAS Institute Inc. All rights reserved.

13.「研修前」「研修後」をシフトで両方選択し、OKボタンをクリックする。



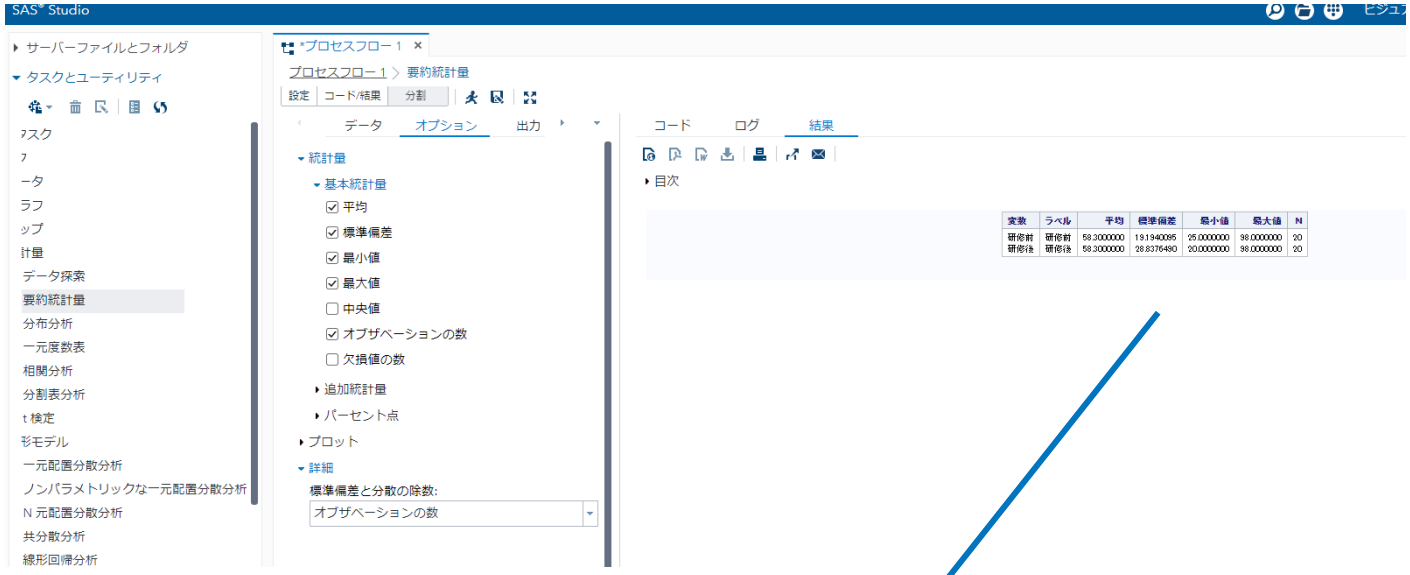
Copyright © SAS Institute Inc. All rights reserved.

14. 「オプション」をクリックし、「統計量」→「基本統計量」の出力項目にチェックを入れ選択する。
「詳細」をクリックし、「標準偏差と分散の除数」から「オブザベーションの数」を選択、実行する。



Copyright © SAS Institute Inc. All rights reserved.

15. 結果が表示される。



変数	ラベル	平均	標準偏差	最小値	最大値	N
研修前	研修前	58.3000000	19.1940095	25.0000000	98.0000000	20
研修後	研修後	58.3000000	28.8376490	20.0000000	98.0000000	20

Copyright © SAS Institute Inc. All rights reserved.

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	19.19
研修後	75	58.3	16.7	28.84

研修前　　研修後

$$\frac{70-58.3}{19.19} = 0.610 > \frac{75-58.3}{28.84} = 0.579$$

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

Copyright © SAS Institute Inc. All rights reserved.

Ｚ値の比較によりデータの評価・比較が可能

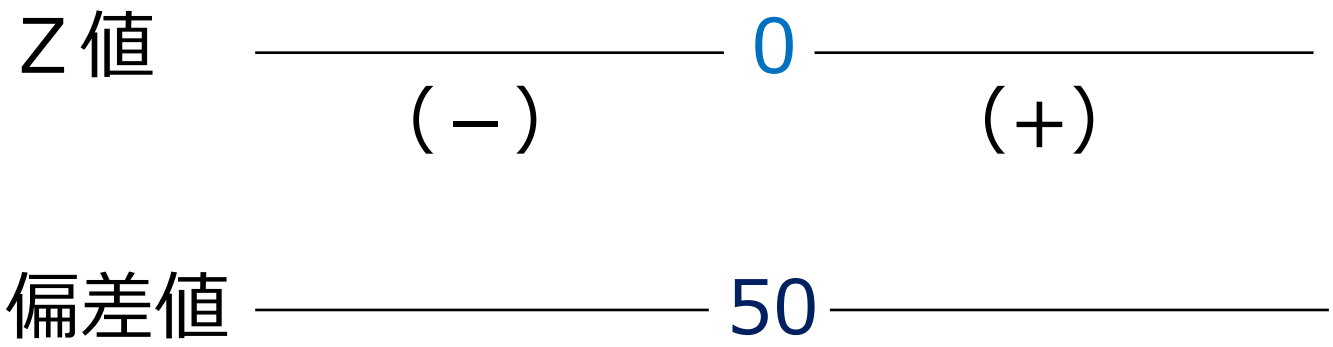
Ｚ値×10＋50

偏差値

研修前の偏差値＝0.610×10＋50＝56.10

研修後の偏差値＝0.579×10＋50＝55.79

Copyright © SAS Institute Inc. All rights reserved.



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、小のとき50より小さな値

Copyright © SAS Institute Inc. All rights reserved.

違いの大きさの判断方法

◇新製品の好感度について、男女別に10点満点にて調査した。
男女間の評価に違いは見られるか？

											平均
2020年	男性	7	6	7	5	6	5	6	7	6	6	6.1
	女性	6	4	5	5	6	5	6	6	4	6	5.3

平均値の差 0.8（6.1－5.3）は大きい？

平均値（6.1、5.3）は信頼できる？

Copyright © SAS Institute Inc. All rights reserved.

◇新製品の好感度について、男女別に10点満点にて調査した。
男女間の評価に違いは見られるか？

											平均
2021年	男性	7	3	8	5	9	2	5	7	6	9	6.1
	女性	6	4	7	3	6	3	6	6	4	8	5.3

2021年のデータは2020年に比較して
全体的にバラツキが大きい。

Copyright © SAS Institute Inc. All rights reserved.

平均の信頼度

2020年　バラツキが小　⇒　平均値は信頼できる

差（0.8）も信頼できる。

2021年　バラツキが大　⇒　平均値は信頼できない

差（0.8）も信頼できない。

差（0.8）が同じであっても
男女間の違いの大きさ　：　2020年　＞　2021年

Copyright © SAS Institute Inc. All rights reserved.

違いの大きさの判断方法

①男性と女性の平均値の差の大きさの検討

男女の平均値の「差」が大きい程、
男女間の好感度は違うと判断できる。

②平均値の信頼性の検討

男女の各データの「バラツキ」が大きい程、
男女間の好感度は違うと判断できない。

Copyright © SAS Institute Inc. All rights reserved.

$$\frac{\text{①男女間の平均値の「差」の大きさ}}{\text{②男女各データの「バラツキ」の大きさ}}$$

上記の値が大きいほど違うと判定できる。

Copyright © SAS Institute Inc. All rights reserved.

判断に対する調査人数の影響

男女各10人の調査、男女各100人の調査
どちらが違いに対する信頼性が高い？

100人の調査！

$$\text{違いの大きさ} = \frac{\text{①平均値の「差」の大きさ}}{\text{②データの「バラツキ」の大きさ}} \times \text{③(調査人数)}$$

①が大きい程、違いは大きい。
②が小さい程、違いは大きい。
③が大きい程、違いは大きい。

Copyright © SAS Institute Inc. All rights reserved.

◇主力製品におけるＣＭ好感度調査

① 昨年度100人について調査を実施。

違いがあるという判断ができなかった。

② 今年度は10万人について調査を実施。

違いがあるという判断ができた！

CMの効果は？

Copyright © SAS Institute Inc. All rights reserved.

「違い」と「効果」

「違いの大きさ」は調査人数の影響を受ける。

ビッグデータにおいては
「違いがある！」という分析結果が出やすい。

違いがある　≠　効果が大きい

効果の大きさを重視する！

Copyright © SAS Institute Inc. All rights reserved.

統計学における違いの大きさ

P値（有意確率）

⇨ 違いがあると判断したときの危険率

■サンプルサイズ（調査人数）が大きいとき

違いの大きさが大きい＝P値が小さい

⇒ 違いがあると判断できる

違いがあっても効果が無い可能性！

■サンプルサイズ（調査人数）が小さいとき

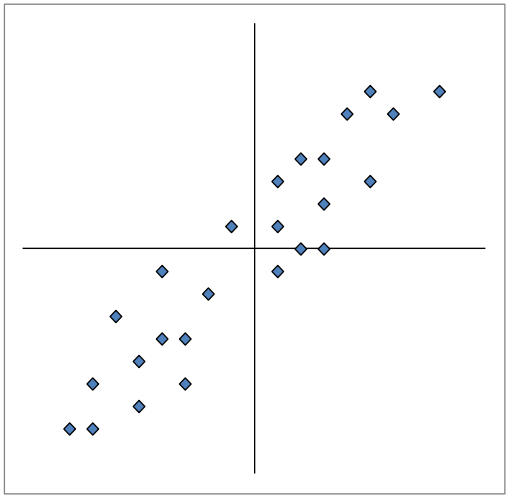
違いの大きさが小さい＝P値が大きい

⇒ 違いがあると判断できない

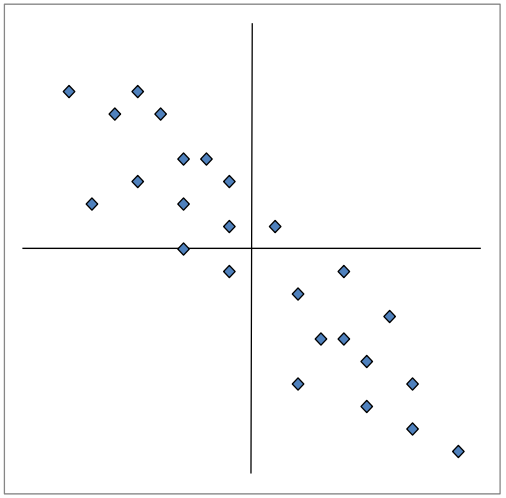
違いが小さくとも効果がある可能性！

Copyright © SAS Institute Inc. All rights reserved.

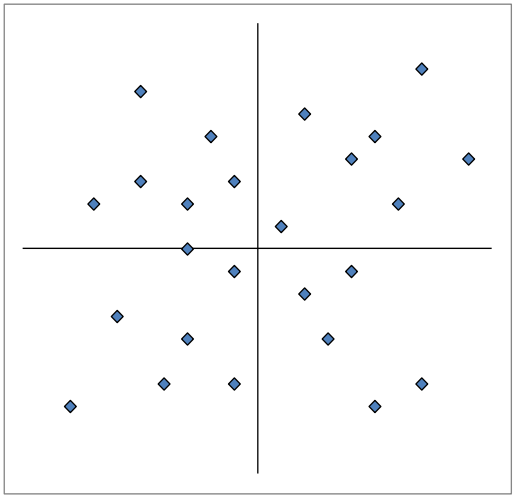
データ群の関係（相関関係）



正の相関



負の相関

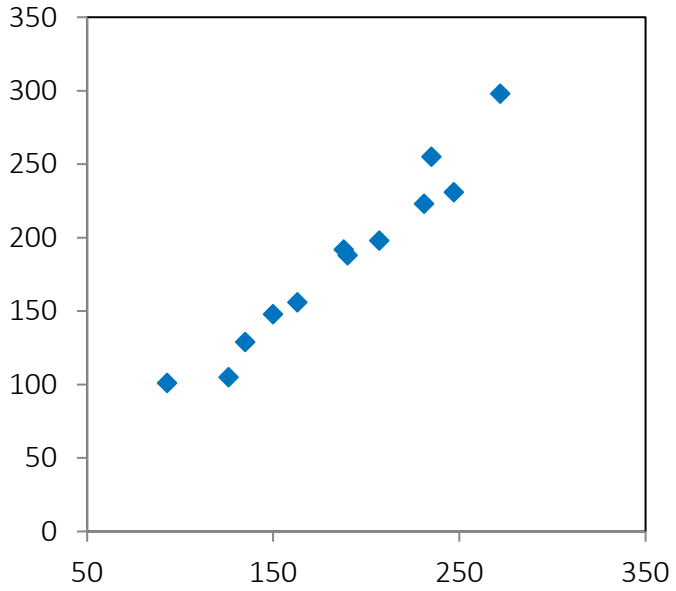


無相関

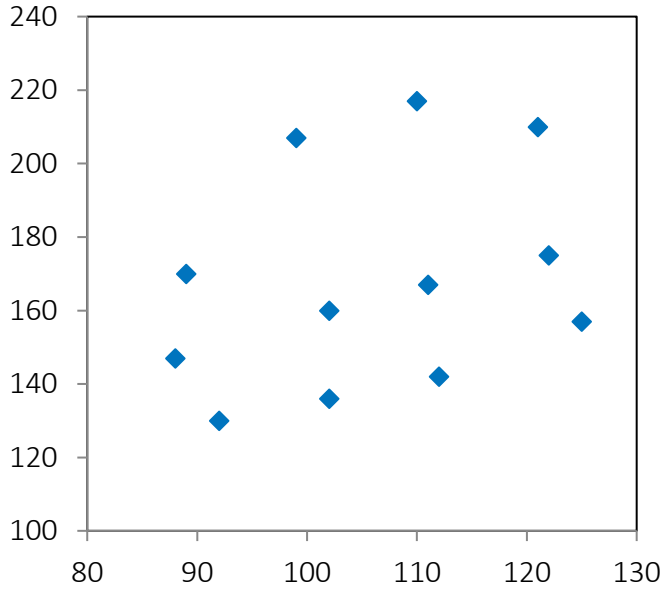
相関関係には正の相関、負の相関、無相関。
点の集中度が関係の強さを測定する手がかり。

Copyright © SAS Institute Inc. All rights reserved.

積率相関係数（ｒ）



r ＝0.97



r ＝0.32

積率相関係数（ｒ）は相関関係の強さ

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

Copyright © SAS Institute Inc. All rights reserved.

積率相関係数（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%

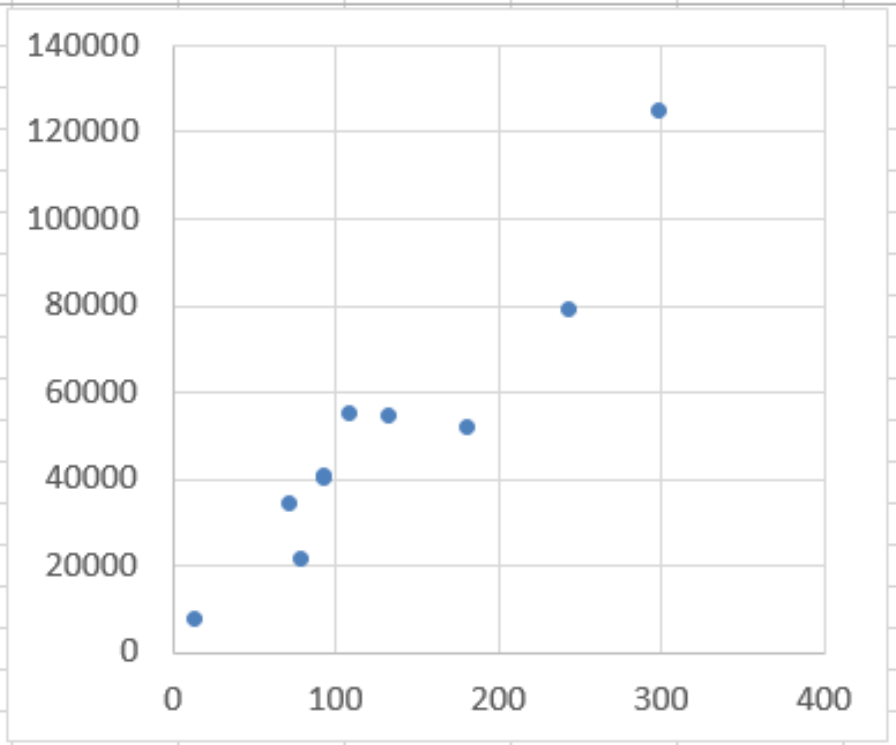
Copyright © SAS Institute Inc. All rights reserved.

◇DM発送数と売上高

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



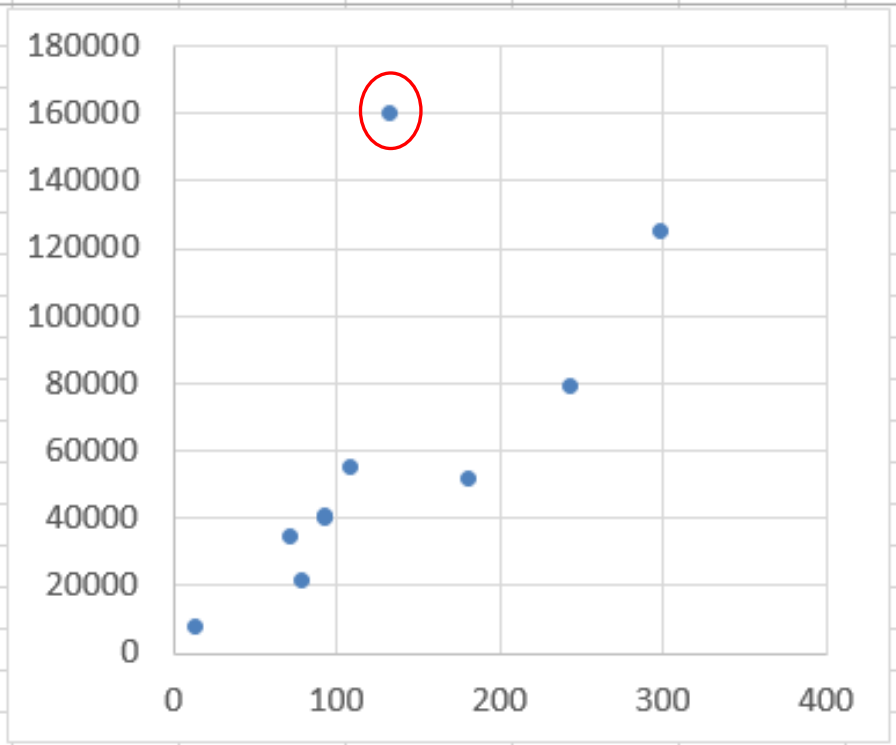
r ＝0.949

Copyright © SAS Institute Inc. All rights reserved.

◇DM発送数と売上高（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

外れ値の有無の検討は必須

Copyright © SAS Institute Inc. All rights reserved.

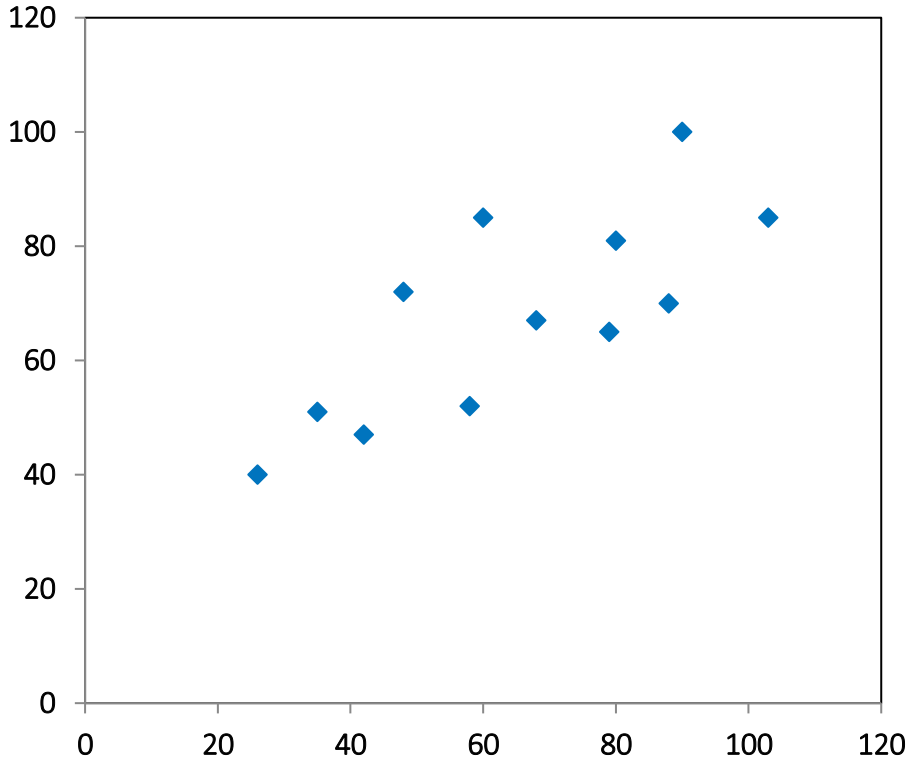
A支店の広告費と売上高

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

⇩

正の相関関係！



Copyright © SAS Institute Inc. All rights reserved.

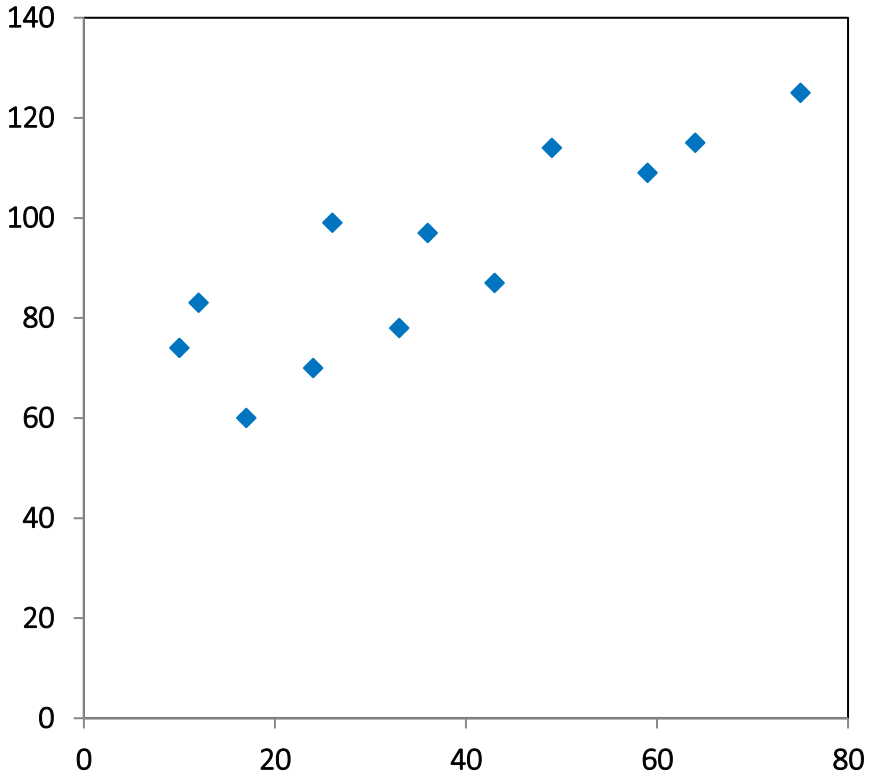
B支店の広告費と売上高

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

⇩

正の相関関係！

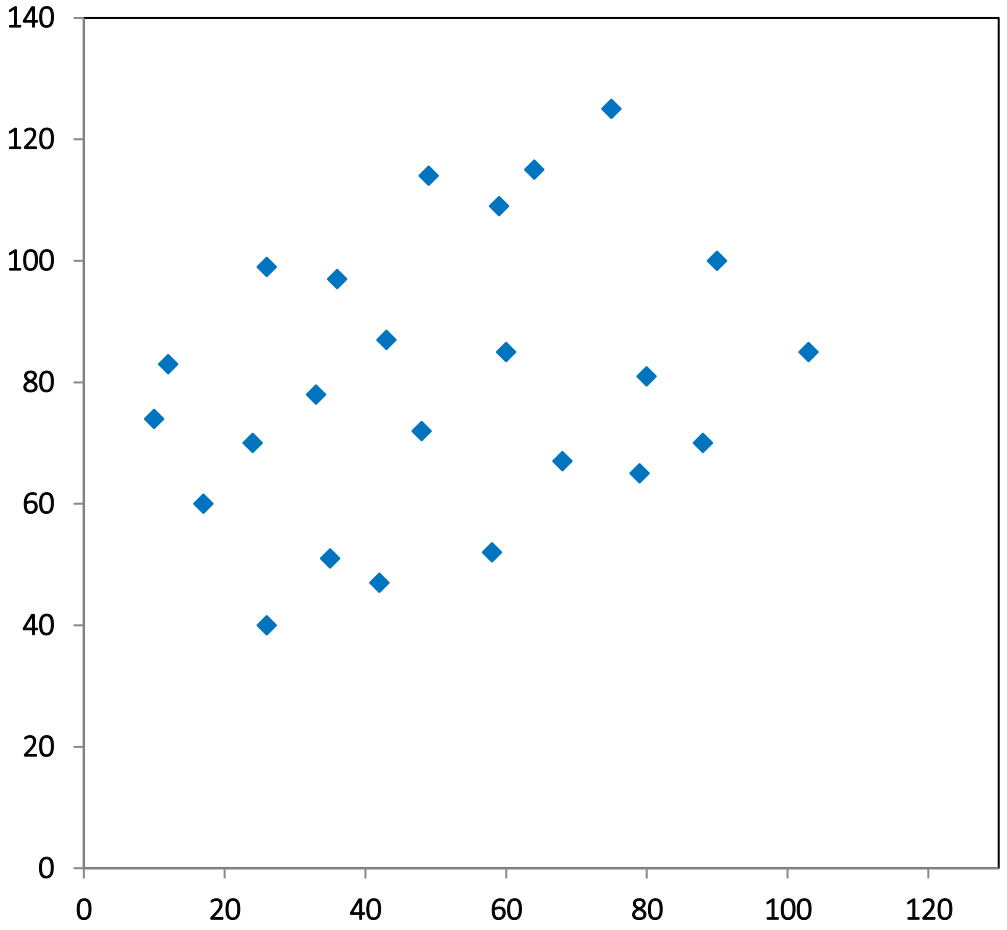


Copyright © SAS Institute Inc. All rights reserved.

（A支店＋B支店）の広告費と売上高

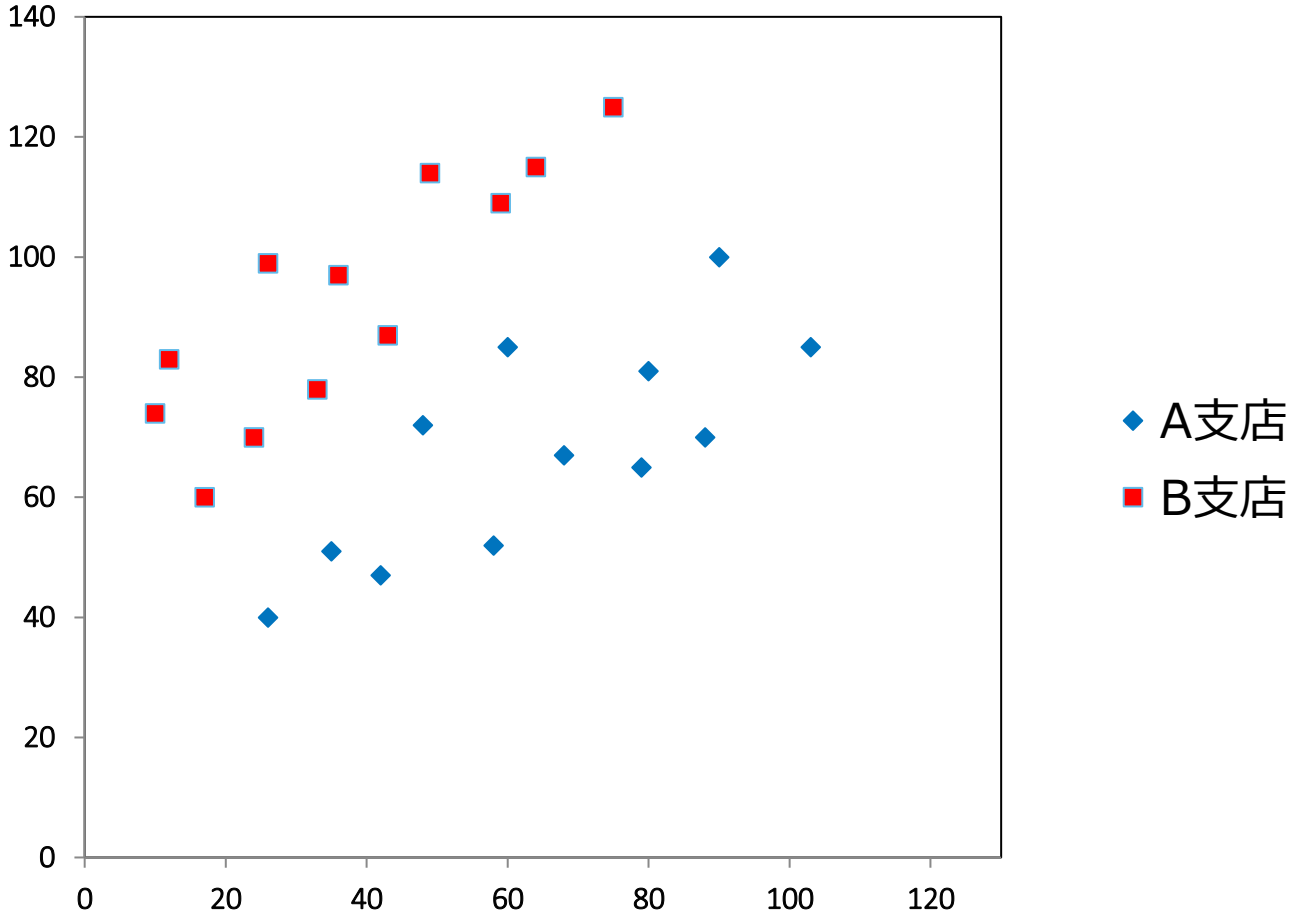
r ＝0.272

???



Copyright © SAS Institute Inc. All rights reserved.

（A支店＋B支店）の広告費と売上高



Copyright © SAS Institute Inc. All rights reserved.

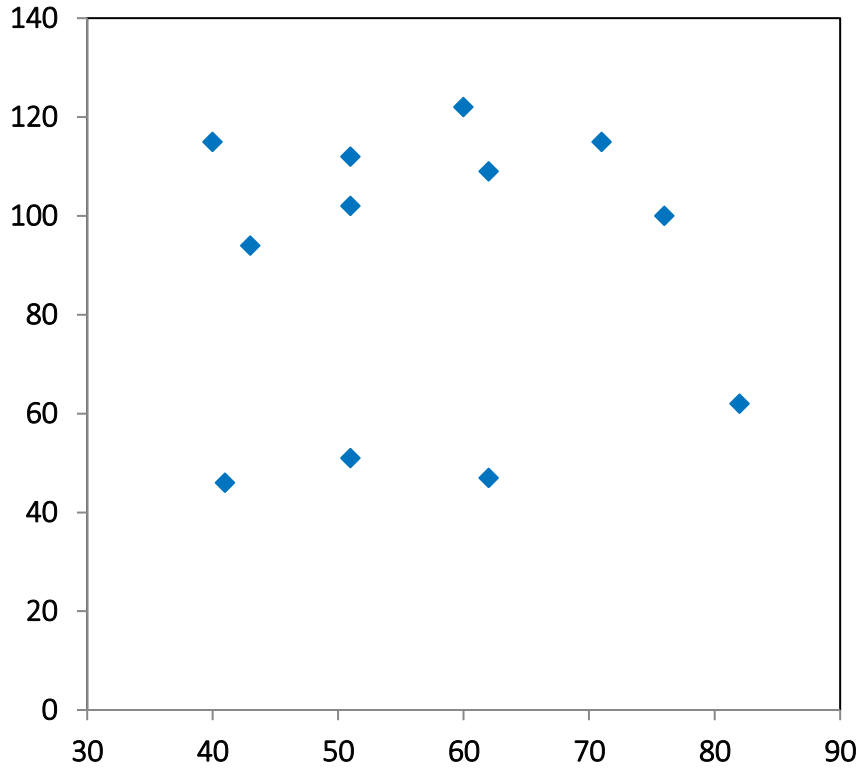
A支店の広告費と売上高

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！



Copyright © SAS Institute Inc. All rights reserved.

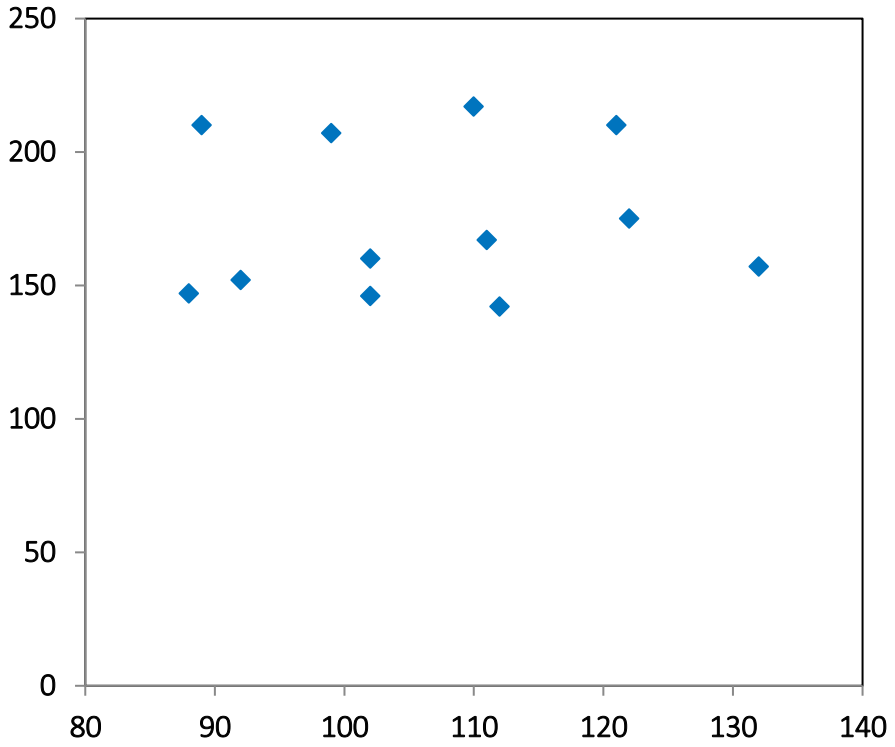
B支店の広告費と売上高

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！

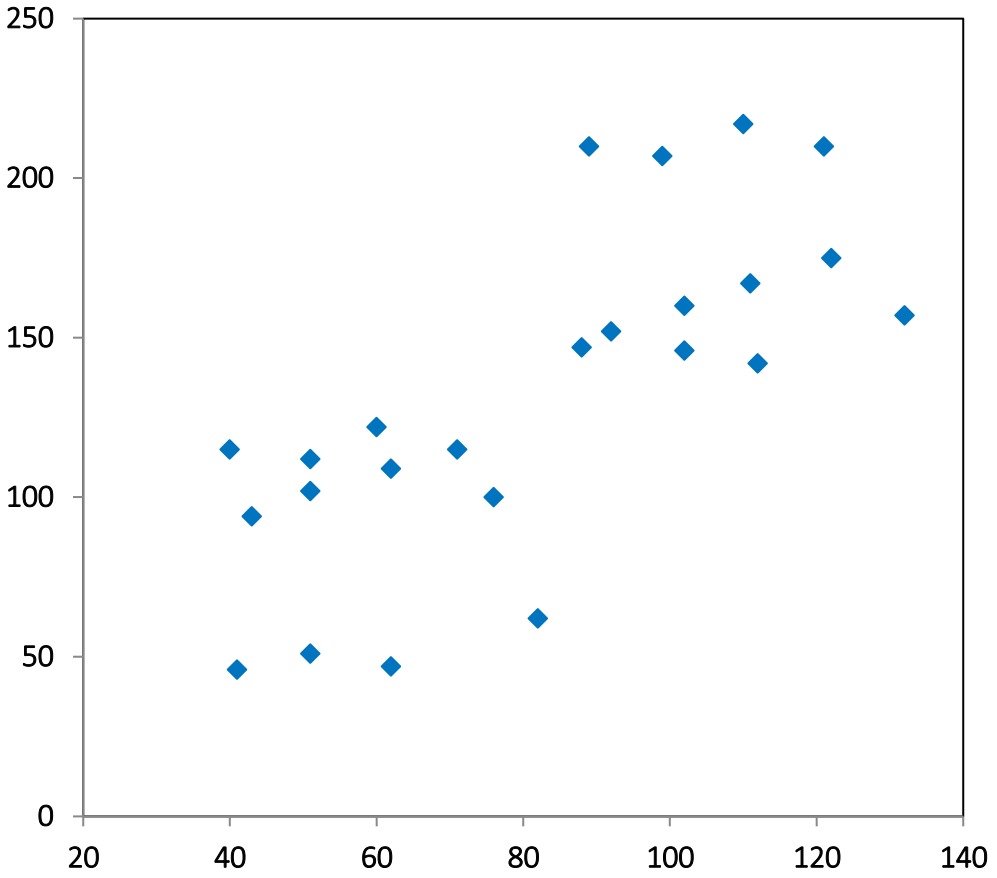


Copyright © SAS Institute Inc. All rights reserved.

（A支店＋B支店）の広告費と売上高

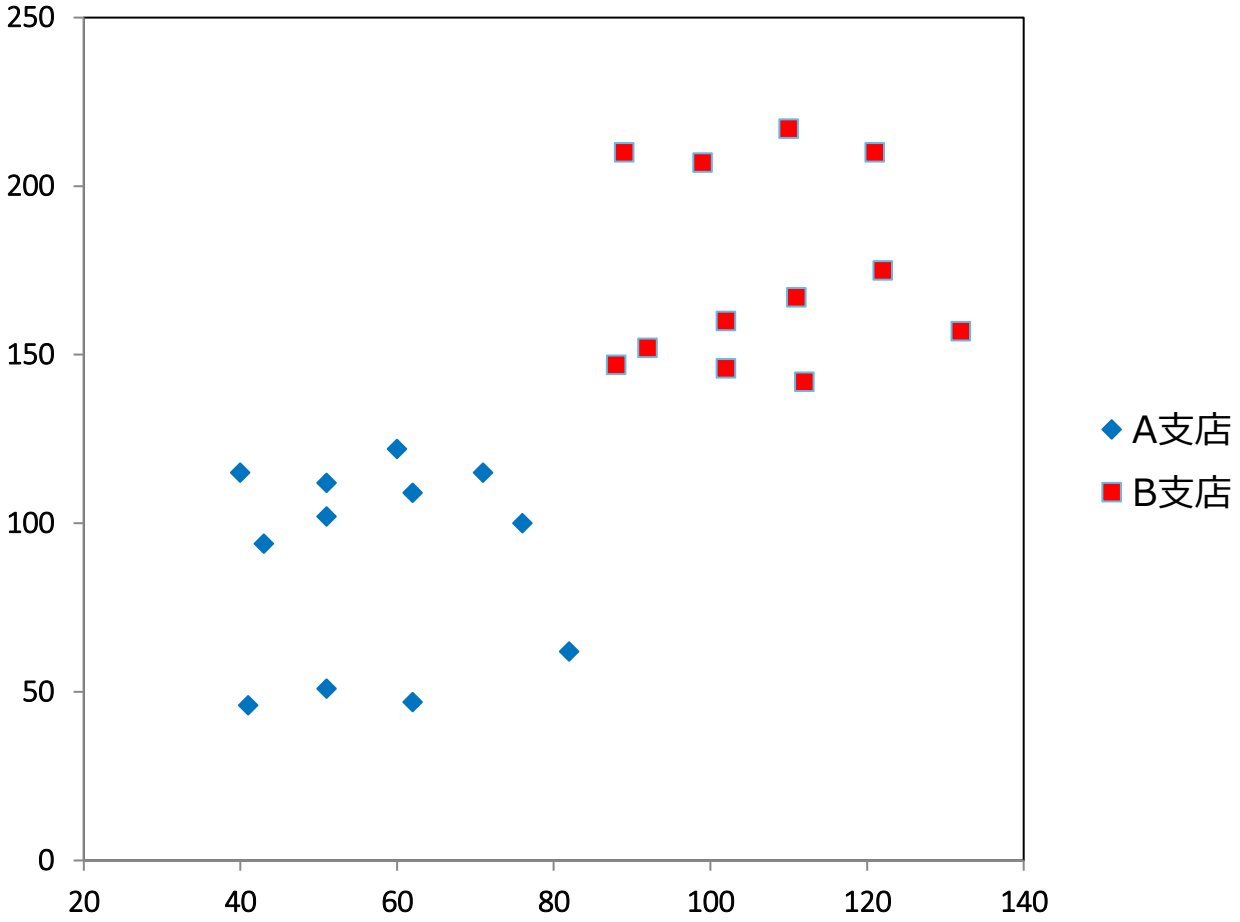
r ＝0.823

???



Copyright © SAS Institute Inc. All rights reserved.

（A支店＋B支店）の広告費と売上高



Copyright © SAS Institute Inc. All rights reserved.

◇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

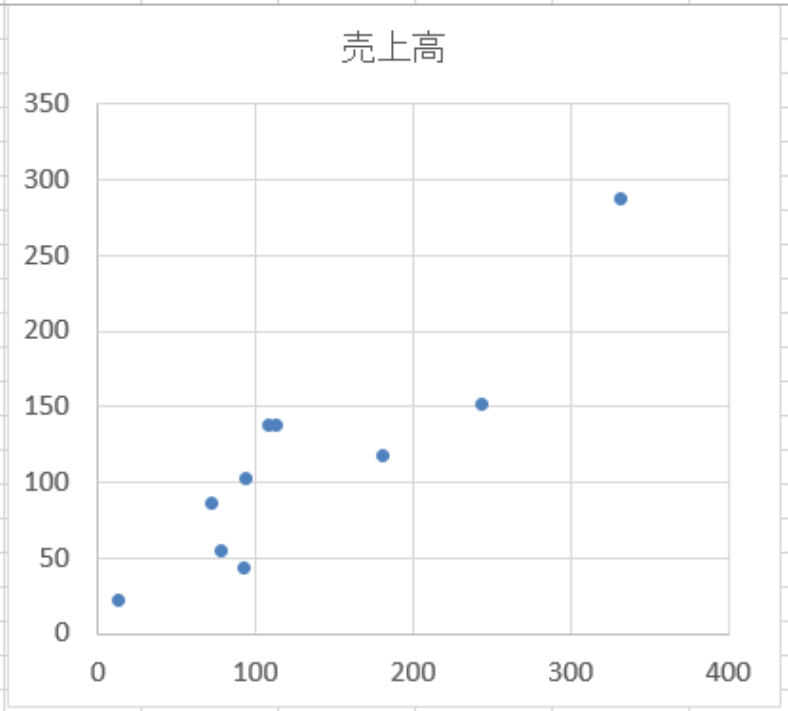
売上高と広告費の関係？

Copyright © SAS Institute Inc. All rights reserved.

◇広告費と売上高

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？

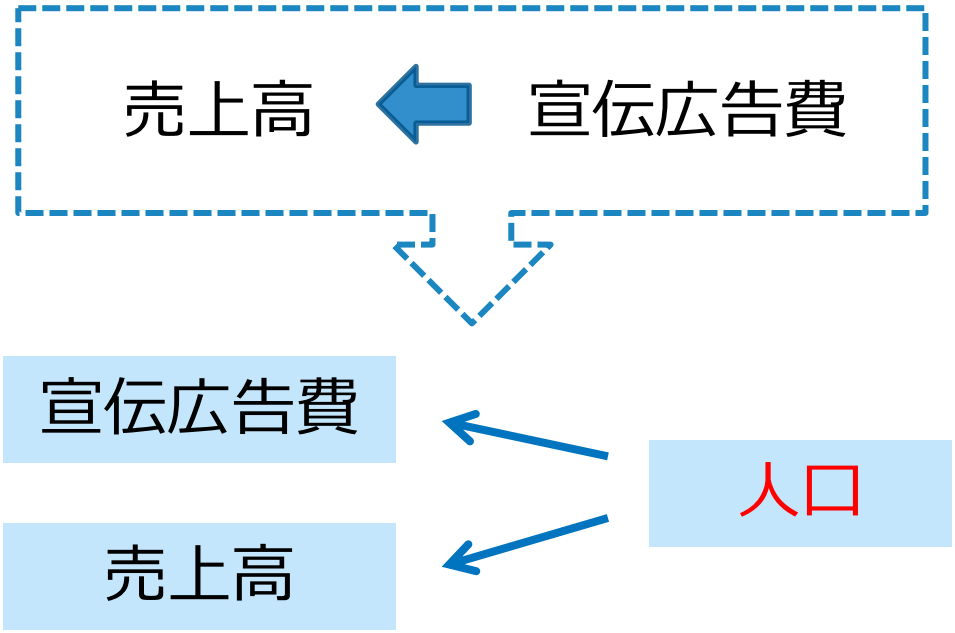
Copyright © SAS Institute Inc. All rights reserved.

◇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	5506
東北	93	102	9335
関東	332	288	42604
北陸	78	54	5443
中部	181	118	18127
近畿	108	138	12912
中国	113	138	15554
四国	72	86	3976
九州	243	152	13204
沖縄	13	22	1393

Copyright © SAS Institute Inc. All rights reserved.

相関関係と因果関係



交絡要因（人口）に注意が必要！

Copyright © SAS Institute Inc. All rights reserved.

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

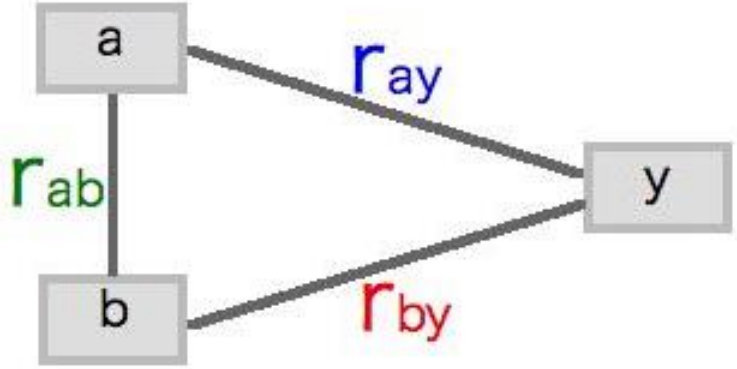
人口は交絡要因

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

Copyright © SAS Institute Inc. All rights reserved.

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

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

Copyright © SAS Institute Inc. All rights reserved.

偏相関係数の計算

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

$$= 0.373$$

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　：　0.373

Copyright © SAS Institute Inc. All rights reserved.

相関関係の検討

- アイスクリームの売り上げと熱中症
- ビールの売り上げと水難事故
- 収入と血圧
- 警察官定員と犯罪件数
- 図書館の数と犯罪件数
- 都道府県別コンビニの数と甲子園勝率
- 少子化と温暖化

Copyright © SAS Institute Inc. All rights reserved.

まとめ

統計学の基本的な考え方

・代表値（平均、標準偏差、Z値）
・違いの大きさ
　違いの大きさと効果の大きさ
・関係の把握方法
　積率相関係数、偏相関係数（交絡要因）

分析結果の合理性の検証
有用なデータ＝数値データ＋背景

データの活用 ⇒ 分析基本知識（統計的思考法）
× 現場力

Copyright © SAS Institute Inc. All rights reserved.