Activity 7

有価証券報告書にある「天候による影響」

このActivityでわかること

- 相関係数
  偏差，偏差積，標準偏差
  正の相関，負の相関
  無相関，完全相関
- 見せかけの相関
- 外れ値
- 修正箱ひげ図
- 2変数の統計量
- 1次回帰式と2次関数の最大・最小
- 1次回帰式と相関係数

コラム
相関係数と因果関係
統計のウソ

Activity 7

「江崎グリコ株式会社の有価証券報告書の内容です。」

(1) 天候による影響

当社グループが展開している事業の中には，菓子・アイスクリーム・ヨーグルト・飲料等，気温の高低や晴雨という天候状況によって消費者の購買行動が影響を受けやすい商品があり，春秋の低温，猛暑，多雨をはじめとする天候不順の場合は当社グループの業績に悪影響を及ぼす可能性があります。

アイスクリームの製造量が天候に左右されとのこと，気温との関係で具体的な数字をもとに調べてみます。

表 7.1　アイスクリームの生産量と東京の平均気温

	2008		2009	
月	平均気温	生産量（Kl）	平均気温	生産量（Kl）
month	temp_2008	pro_2008	temp_2009	pro_2009
1	5.9	8,706	6.8	7,183
2	5.5	8,153	7.8	9,640
3	10.7	12,114	10.0	10,207
4	14.7	11,180	15.7	11,208
5	18.5	11,814	20.1	11,178
6	21.3	12,192	22.5	11,403
7	27.0	11,923	26.3	12,510
8	26.8	13,223	26.6	12,573
9	24.4	11,857	23.0	11,395
10	19.4	12,082	19.0	10,534
11	13.1	10,966	13.5	9,217
12	9.8	9,825	9.0	9,131

列名

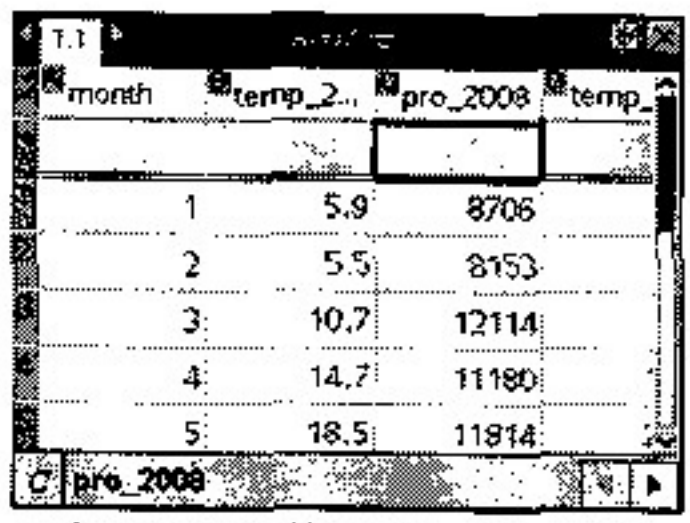
乳製品の生産量 アイスクリーム（乳脂肪分８％以上のもの)
牛乳乳製品統計(農林水産省統計情報部)
東京の気温：気象庁

データを俯瞰する

散布図を描いてデータの分布を見ます。

1 データの入力

ctrl +page で「List & Spreadsheet」のページを追加します。

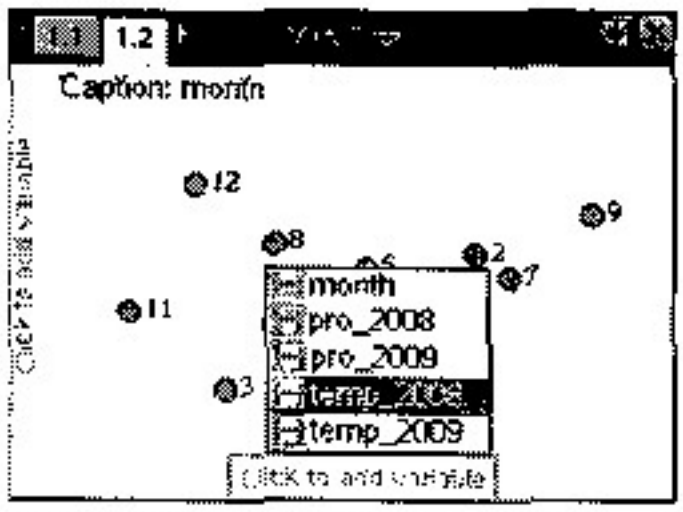


列名を付けて数値を入力します。
＿は，ctrl で入力できます。

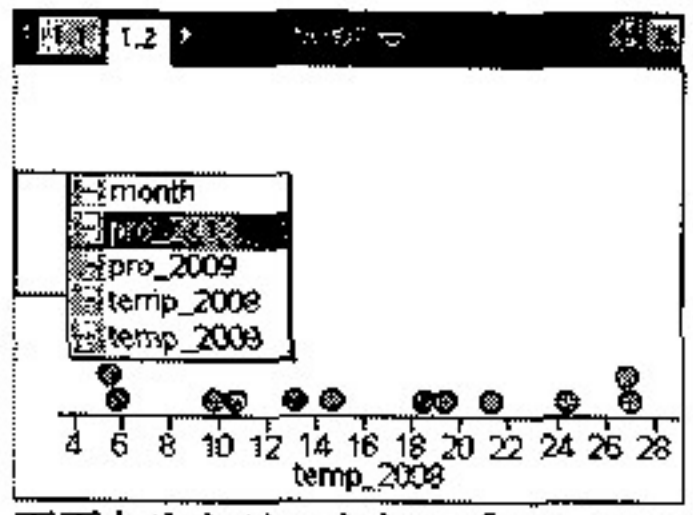
2 散布図を描く

x 軸に平均気温，y 軸に生産量をとった散布図を描きます。

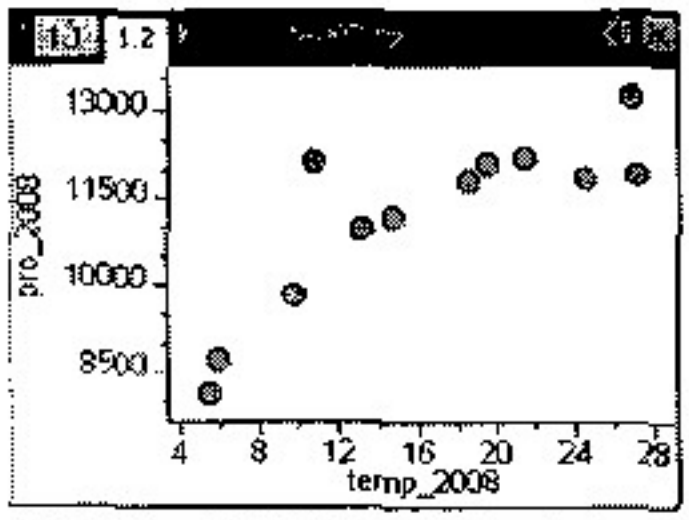
ctrl +page で「Data & Statistics」のページを追加します。



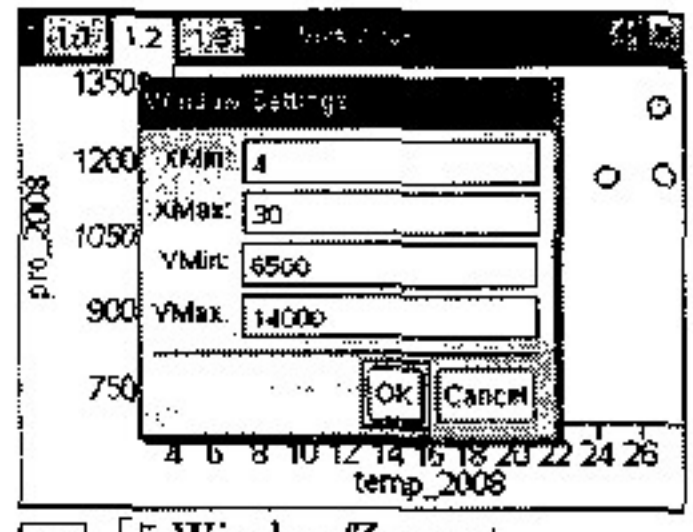
画面下をクリックすると変数一覧が表示されるので，「temp_2008」を選択して enter を押します。



画面左をクリックして「pro_2008」を選択して enter を押します。

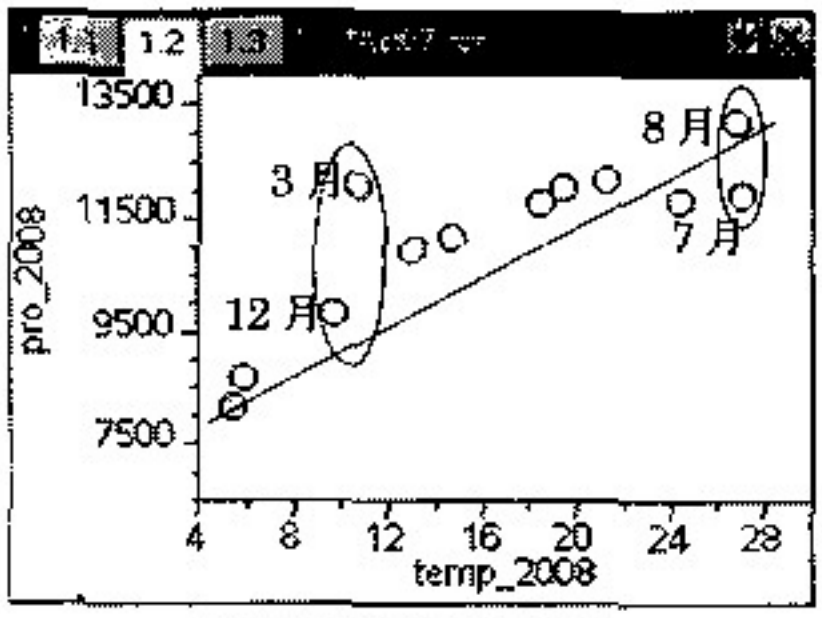


2008 年の散布図が描かれます。

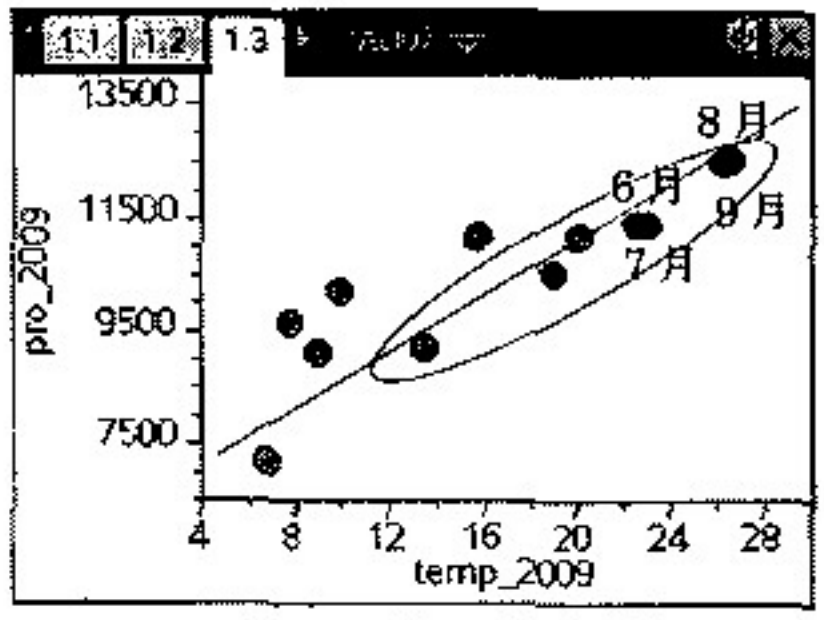


menu 「5:Window/Zoom」
「1:Window　Setting..」で，
X と Y の範囲を修正します。

同じように 2009 年の散布図も描きます。



2008 年の散布図



2009 年の散布図

色を付けたいときは，●にカーソルを持っていく，カーソルが✎となった時に ctrl menu を押し，「3:Color」「2:Fill Color」で好きな色を選択します。

散布図からわかることをまとめます

- 2008 年も 2009 年も気温が高くなる概ね生産量が増えている。
- 2008 年の散布図は，中ほどのデータは一直線上に並んでいるが，7 月と 8 月，3 月と 12 月は，気温が概ね同じにもかかわらず生産量は随分違う。
- 2009 年の散布図は概ね一直線上に並んでいるように見える。
  7 月と 8 月，6 月と 9 月は気温も生産量もほとんど同じである。

気温が決まれば生産量が決まるというような 1 次関数で表現できるわけではないが，どちらの散布図も気温と生産量との間に関連がありそうです。2 つを比べてみると，2009 年のほうがシャープで，2008 年の散布図はそれに比べると少し“ばらけている”感じがします。
これは，気温と生産量の関係は 2009 年の方が強くて，2008 年の方が弱かったと言えます。
では，

◇ 関連性が強い，弱いを数字にできないか。

相関係数

相関係数 r (correlation coefficient)は，気温と生産量，身長と体重，広告費と売上などといった 2 つの変数 x, y の間に，どの程度の関連があるかを測るための指標です。相関係数を計算することによって，x と y の間に，どの程度の直線的な関係があるか（＝データが直線の近くにどのくらい集中しているか）を知ることができます。
相関係数は以下の式で求めます。

$$\text{相関係数 } r = \frac{\text{偏差積の平均}}{(x\text{の標準偏差}) \times (y\text{の標準偏差})}$$

以下のデータを使って計算の手順とそれを図式化したものを示します。

	x	y	C 偏差 x	D $(\text{偏差 } x)^2$	F 偏差 y	G $(\text{偏差 } y)^2$	偏差積
1	2	15	-3	9	-2	4	6
2	3	18	-2	4	1	1	-2
3	6	15	1	1	-2	4	-2
4	6	19	1	1	2	4	2
5	8	18	3	9	1	1	3
合計	25	85		24		14	7
平均	5 A	17 B		4.8 D		2.8 G	1.4 I
標準偏差				2.19 E		1.673 H	

❑ 計算の手順

A　x の合計を計算し，平均($\overline{x}$)を求めます。

B　y の合計を計算し，平均($\overline{y}$)を求めます。

C　偏差 $x = x - \overline{x}$　を求めます。

D　$(\text{偏差 } x)^2$ を求め，その合計を計算し平均を求めます。

E　D で求めた値の平方根（ルート）を求めます。

この値が「x の標準偏差」です。

F　偏差 $y = y - \overline{y}$　を求めます。

G　$(\text{偏差 } y)^2$ を求め，それの合計を計算し平均を求めます。

H　G で求めた値の平方根（ルート）を求めます。

この値が「y の標準偏差」です。

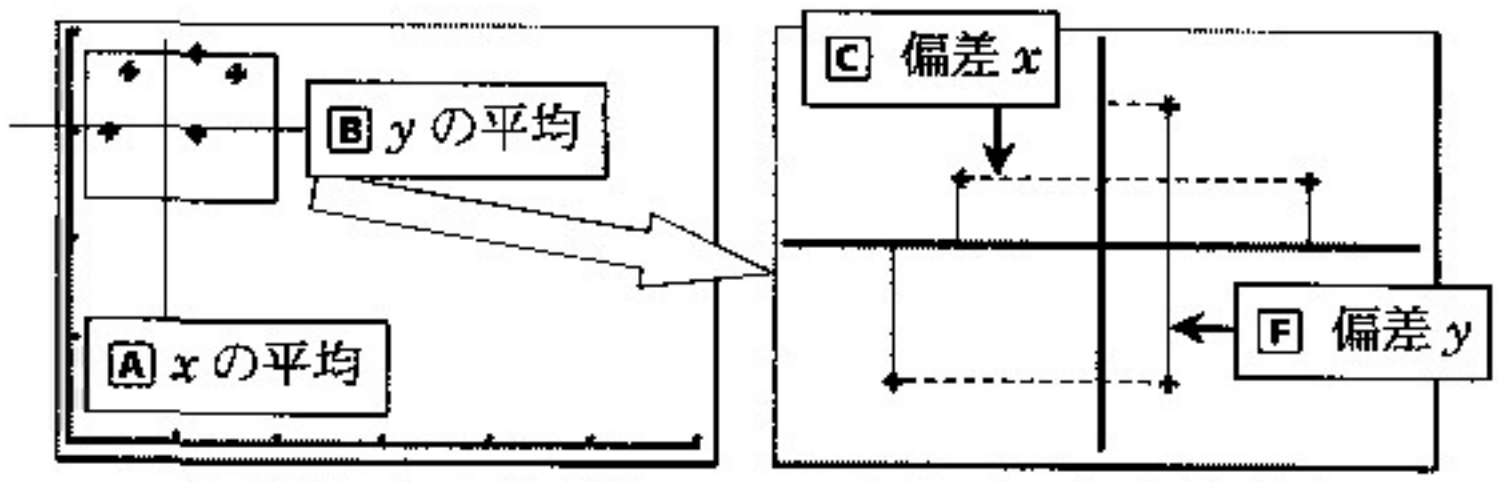
I　偏差積＝(偏差 x) ×(偏差 y)を求め，その合計を計算し平均を求めます。この値が「偏差積の平均」です。

J　相関係数を求めます。

$$\text{相関係数 } r = \frac{\text{偏差積の平均}}{(x\text{の標準偏差}) \times (y\text{の標準偏差})} = \frac{1.4}{2.19 \times 1.673} = 0.382$$

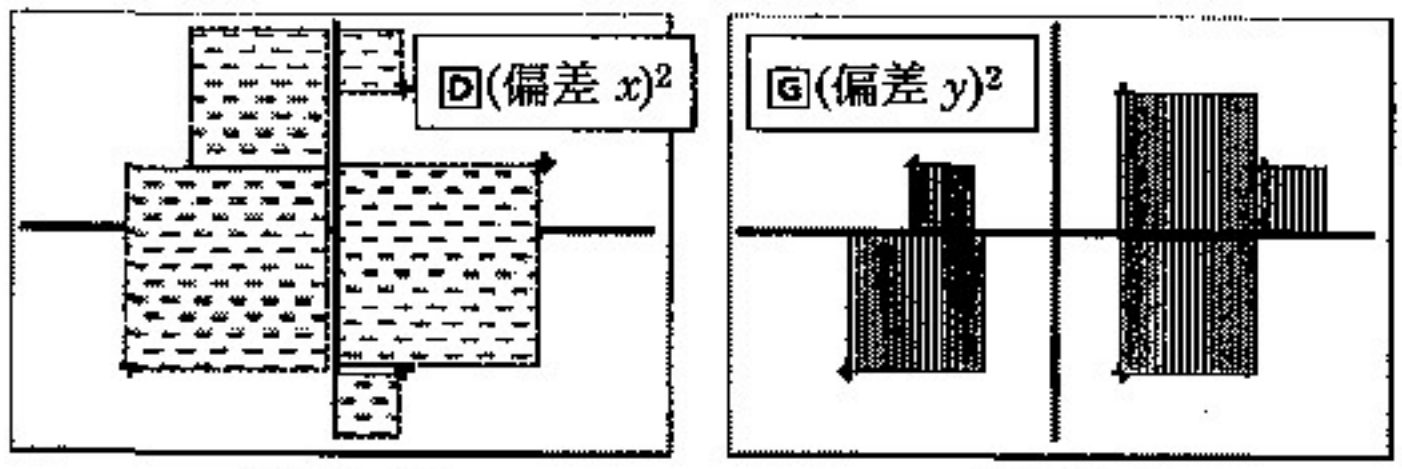
□ 計算の図式化

x と y の平均を求め，それを 0 とした（それぞれの値からから平均を引いた）値で散布図を作ります。



元のデータの散布図　　平均を 0 にした散布図

偏差を 2 乗することは，偏差の正方形を作ることを意味します。



(偏差 x)^2　　(偏差 y)^2

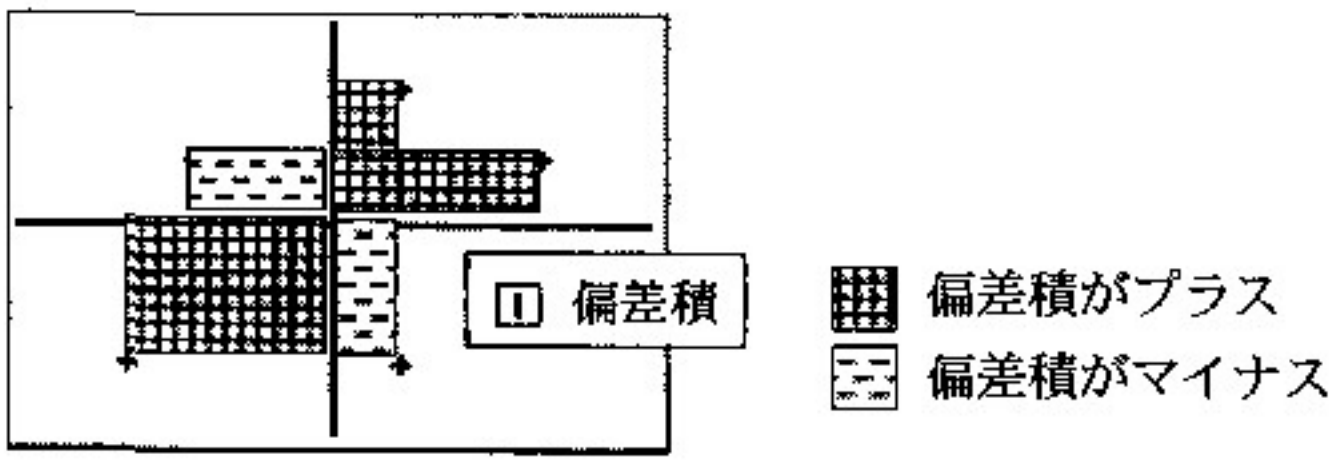
正方形の面積を足してその平均の平方根（ルート）をとります。

(偏差 x)^2 の平均のルート　　(偏差 y)^2 の平均ルート

⇩　　⇩

x の標準偏差　　y の標準偏差

偏差積は，(偏差 x) ×(偏差 y)を掛けたものです。



偏差積

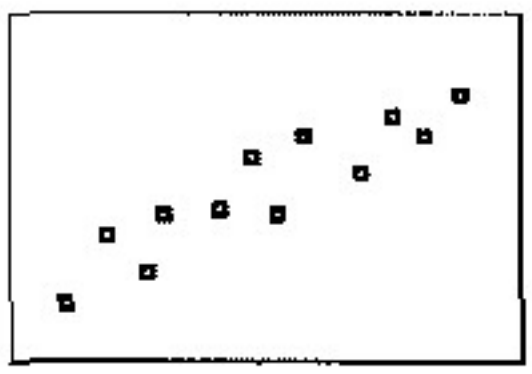
この偏差積の面積の平均を求めます。

偏差積の図から明らかなように，

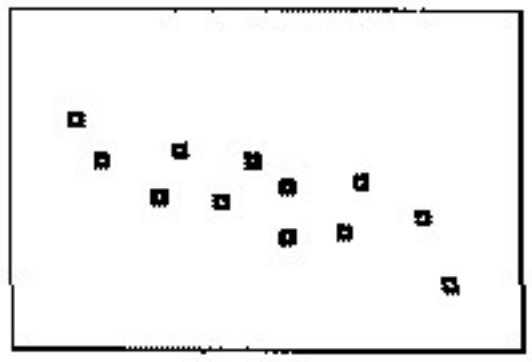
A：偏差積の合計の値がプラスであれば，相関係数はプラス。
　これは，▦の面積が▨の面積より大きい状態であり，
　散布図は右肩上がりになります。これを「**正の相関**」といいます。

B：偏差積の合計の値がマイナスであれば，相関係数はマイナス。
　これは，▦の面積が▨の面積より小さい状態であり，
　散布図は右肩下がりになります。これを「**負の相関**」といいます。

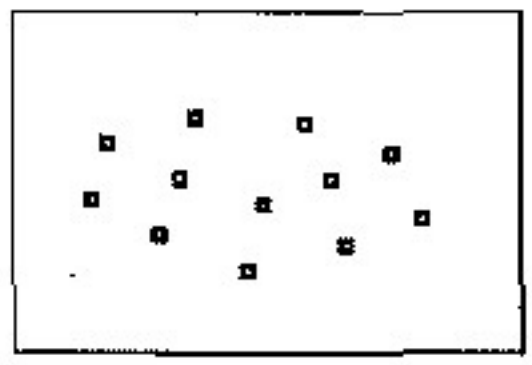
C：偏差積の合計の値が０のとき，「**相関ゼロ＝無相関**」となります。



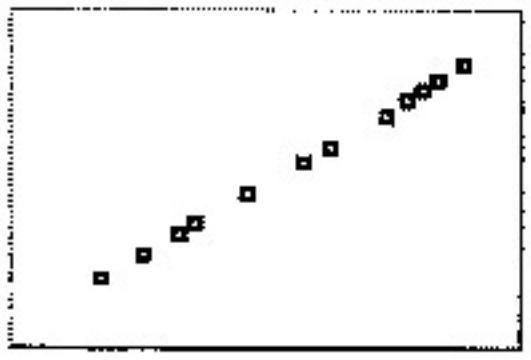
「正の相関」



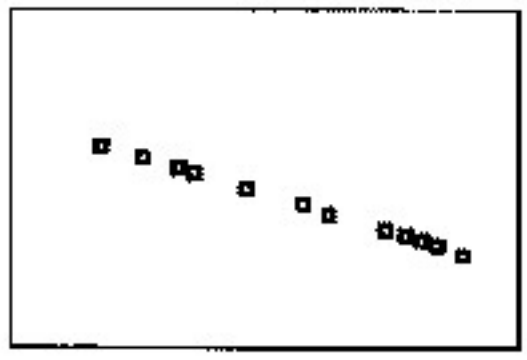
「負の相関」



「相関が小さい」



「正の完全相関（$r=1$）」



「負の完全相関（$r=-1$）」

完全相関のとき，すべてのデータは一直線上にあります。

一般に相関係数の値により，以下のように表現します。

±0.7～±1	強い相関がある
±0.4～±0.7	中程度の相関がある
±0.2～±0.4	弱い相関がある
±0～±0.2	ほとんど相関がない

※偏差積和を平均して標準偏差で割るわけ

偏差積和を平均するということは，データの個数の影響を取り除くことです。 偏差積和だけだと，データの数が大きくなるとそれにつれて値が大きくなってしまいます（負の相関の時は小さくなってしまう）。したがって，偏差積和をデータの個数で割ります。

※xの標準偏差とyの標準偏差で割るわけ

xの標準偏差とyの標準偏差で割るということは，xについても，yについても，標準偏差を1に揃えるということになります。

2008年の「平均気温」と「生産量」の相関係数を求める

相関係数の意味を理解したところで，2008年の相関係数を，手順を追って求めます。

	2008		2009		2008				
月	平均気温	生産量(Kl)	平均気温	生産量(Kl)	偏差 temp	偏差 $temp^2$	偏差 pro	偏差 pro^2	偏差積
A	B	C	D	E	F	G	H	I	J
month	temp_2008	pro_2008	temp_2009	pro_2009	t_he_2008	t_he2_2008	p_he_2008	p_he2_2008	tp_he_2008
1	5.9	8,706	6.8	7,183					
2	5.5	8,153	7.8	9,640					
3	10.7	12,114	10.0	10,207					
4	14.7	11,180	15.7	11,208					
5	18.5	11,814	20.1	11,178					
6	21.3	12,192	22.5	11,403	A	B	C	D	E
7	27.0	11,923	26.3	12,510					
8	26.8	13,223	26.6	12,573					
9	24.4	11,857	23.0	11,395					
10	19.4	12,082	19.0	10,534					
11	13.1	10,966	13.5	9,217					
12	9.8	9,825	9.0	9,131					
平均						F-1		G-1	H
標準偏差						F-2		G-2	

上記の A から H までの計算の仕方を以下に示します。

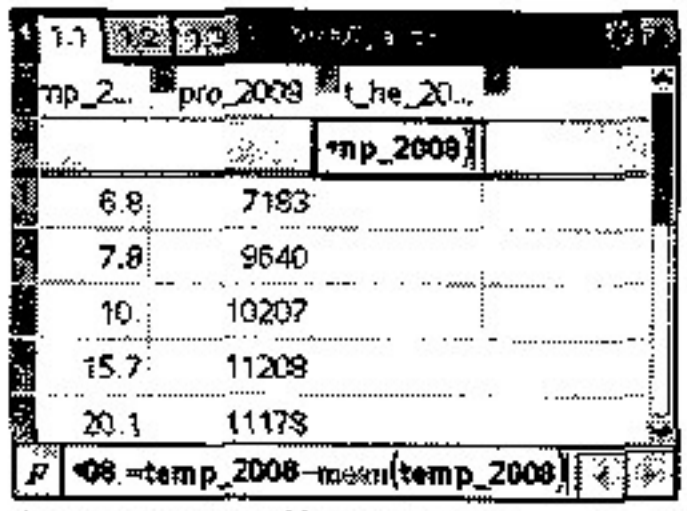
① 平均気温の偏差と偏差の 2 乗を計算する

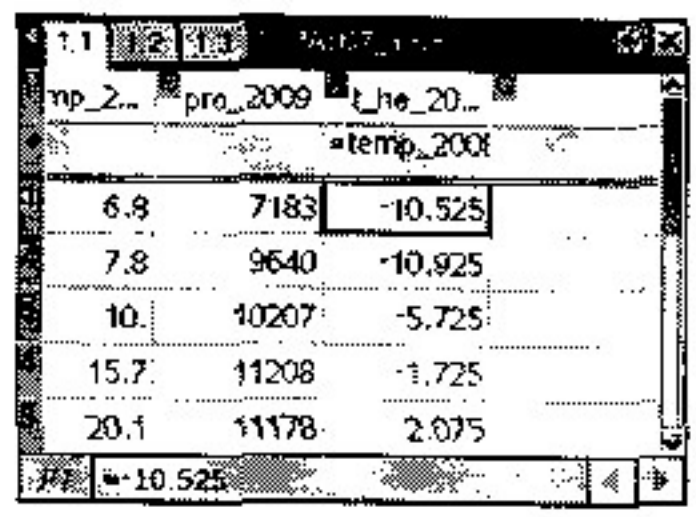
Ⓐ 平均気温（temp_2008）の偏差（t_he_2008）を求める。

t_he_2008 = temp_2008 − mean(temp_2008)

mean は平均を求める関数で,

mean(temp_2008)は，temp_2008 の平均を計算します。

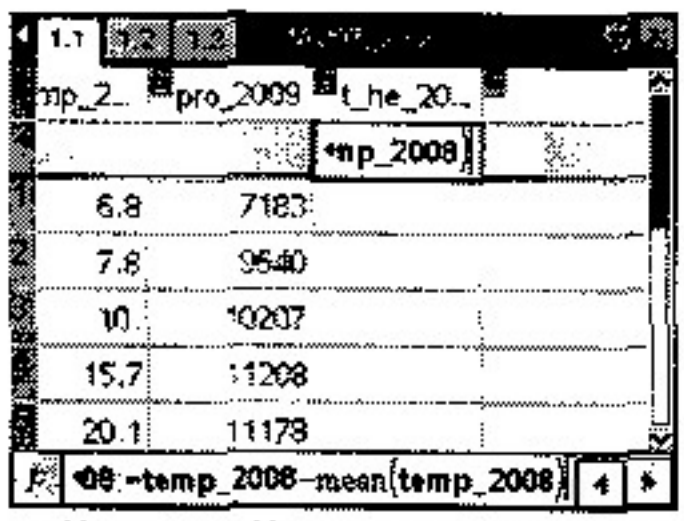


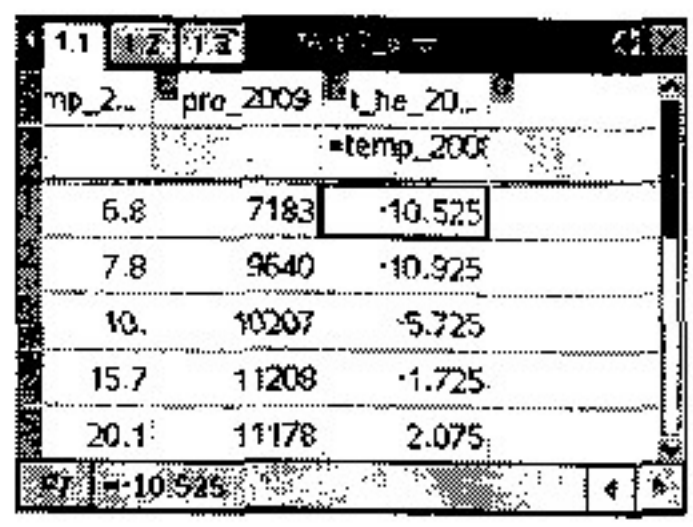


数式セルに計算式を入力し，enter を押します。

Ⓑ 平均気温の偏差（t_he_2008）の 2 乗を求める。

t_he2_2008 = （t_he_2008）^{2}





計算式を定義フィールドに計算式を入力し，enter を押します。

② 生産量の偏差と偏差の 2 乗を計算する

Ⓐ，Ⓑ と同様に計算します。

Ⓒ 生産量（pro_2008）の偏差（p_he_2008）を求める。

p_he_2008 = pro_2008 − mean(pro_2008)

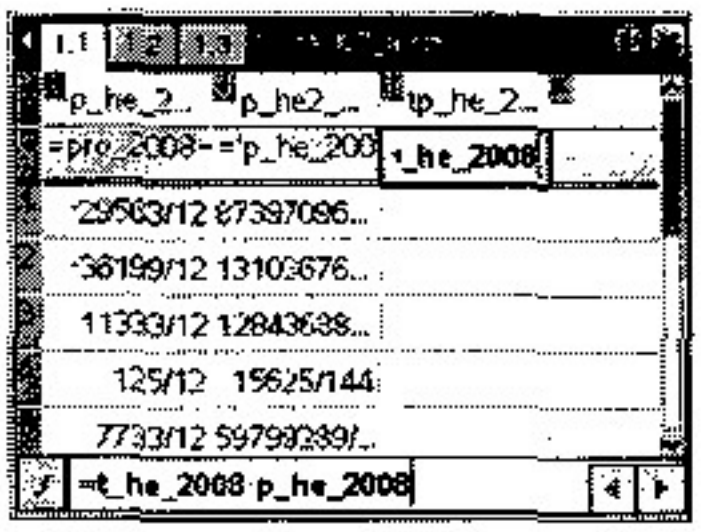
Ⓓ 平均気温の偏差（p_he_2008）の 2 乗を求める。

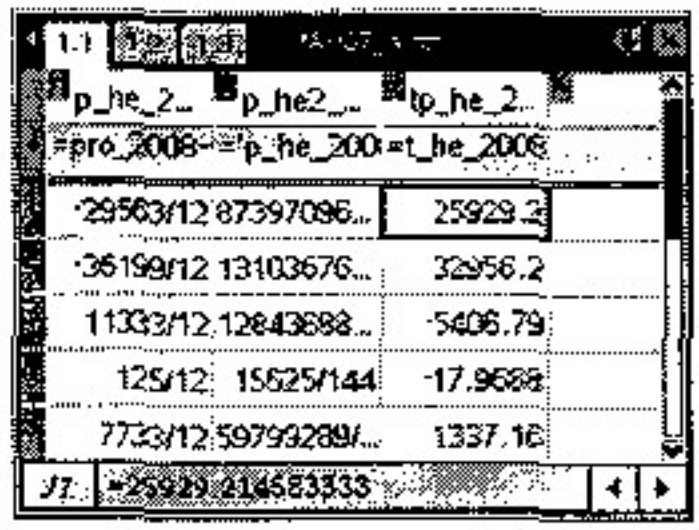
p_he2_2008 = （p_he_2008）^{2}

3 偏差積を求める

E 偏差積（tp_he_2008）を求める。

$$tp\_he\_2008 = t\_he\_2008 \times p\_he\_2008$$





計算式を定義フィールドに計算式を入力し，enterを押します。

4 分散と標準偏差を求める

ctrl+pageで「Calculator」のページを追加します。

F 平均気温の分散と標準偏差を求める。

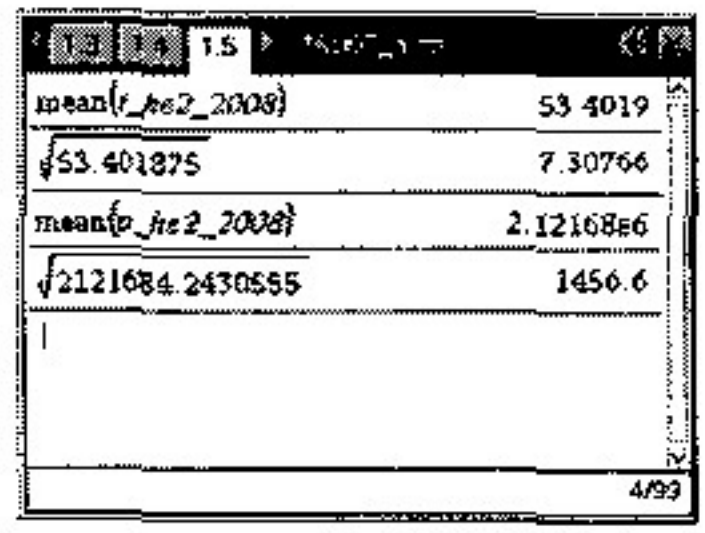
F -1　分散＝mean(t_he2_2008)

F -2　標準偏差＝分散の平方（ルート）

G 生産量の分散と標準偏差を求める。

G -1　分散＝mean(t_he2_2008)

G -2　標準偏差＝分散の平方（ルート）



5 偏差積の平均を求める

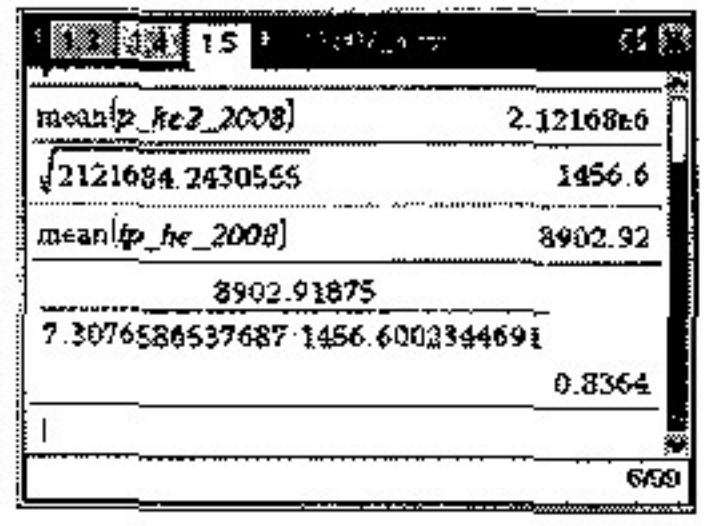
H 偏差積の平均を求める。

偏差積の平均＝mean(tp_he_2008)

6 相関係数を求める

相関係数を求める。

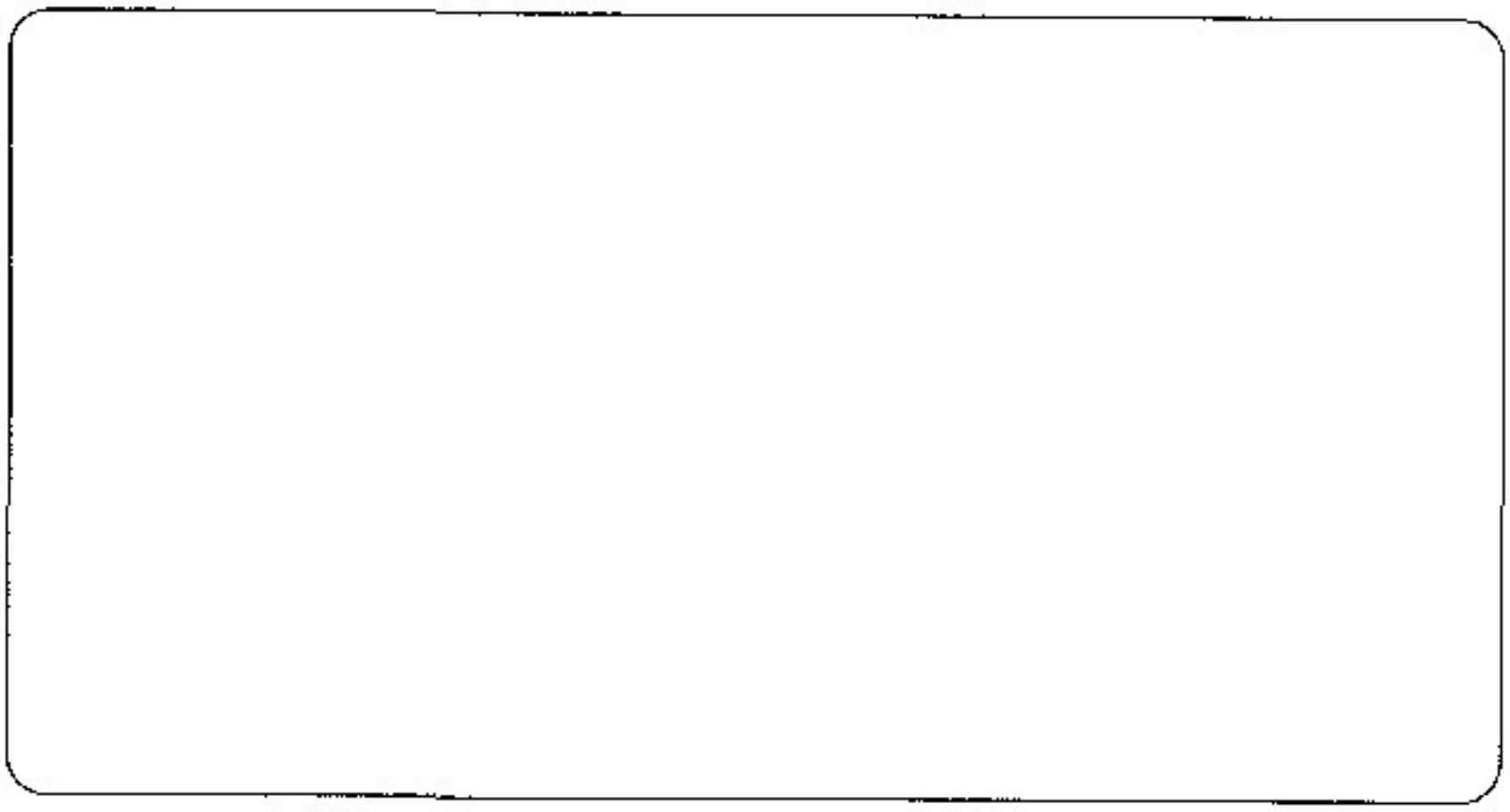
$$\frac{\text{偏差積の平均}}{(\text{気温の標準偏差})\times(\text{生産量の標準偏差})}$$



❑ 結果

2008 年の平均気温と生産量の相関係数は，0.836 となります。

この 2 つの間には，「強い正の相関」があります。



課題 7.1

「表 7.1　アイスクリームの生産量と東京の平均気温」のデータをもとに，2009 年の以下の統計値を計算しなさい。

(1) 相関係数を計算するとき必要な統計量を求めなさい。

平均気温の		生産量の	
平均 ……	(　　)	平均 ……	(　　)
分散 ……	(　　)	分散 ……	(　　)
標準偏差 …	(　　)	標準偏差 …	(　　)

(2) 相関係数を求めなさい。

偏差積の平均 …	(　　)	相関係数 …	(　　)

課題7.2

相関係数が0となる6組のデータを1セットとして，2セットのデータを作成しなさい。そのとき，どのような考えに基づいてデータを作成したかを記述しなさい。

（注）作成したデータの相関係数を計算して確かめなさい。

セット1

No.	x	y
1		
2		
3		
4		
5		
6		

セット2

No.	x	y
1		
2		
3		
4		
5		
6		

コラム
Column

相関関係と因果関係

相関とは，2つのできごとの間の関係ということです。相関について調べるのは，2つのできごとの間に何らかの因果関係があり，その強さを知りたいからです。

この Activity6 では，“気温の高低などの天候不順がアイスクリームなどの販売に影響を与えて業績に悪影響を及ぼす可能性がある。”との決算書の内容について，どの程度の相関があるかを，実際の気温と生産量の相関で調べたわけです。その背景には気温が原因になって，生産量が結果として決まるのではないかということを想定しています。このように2つの出来事の相関の強さがわかれば，これからのことを予測する上で有益な情報になります。

このように，相関関係は因果関係を含みますが，相関関係があるからといって，それにより因果関係があるとは限りません。

次ページのポスターで表現されていることは，

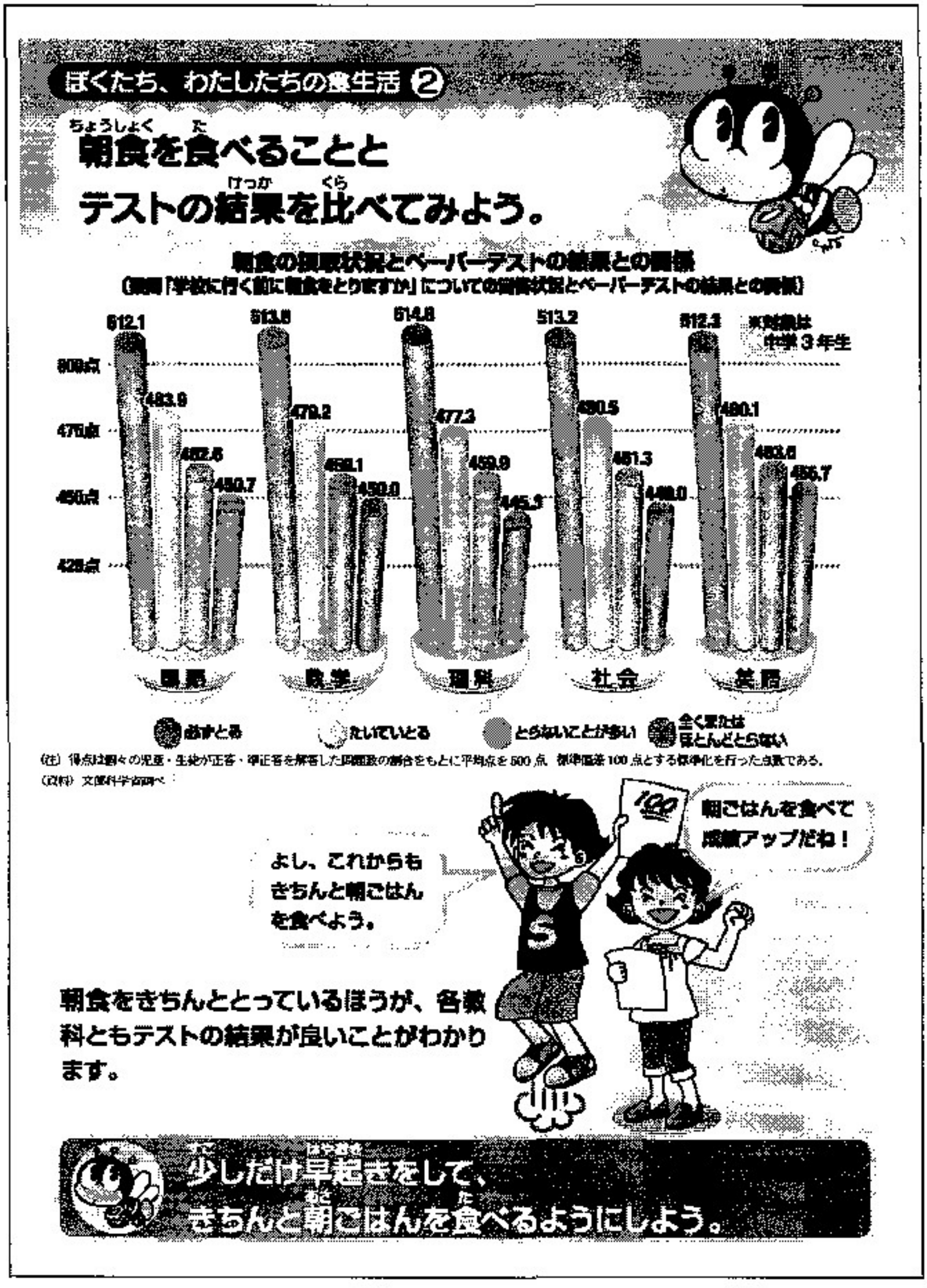
- 中学生の各科目の成績と朝食を食べる頻度の相関を棒グラフにしている。
- 全ての科目において，朝食を食べる頻度が高いグループほど成績の平均値も高い。

これだけであれば，事実をそのままグラフ化したもので，何も問題はありません。しかしながら，この図で子供にしゃべらせているセリフが，相関関係と因果関係についての認識ができていないことを示しています。

「よし、これからもきちんと朝ごはんを食べよう。」
「朝ごはんを食べて成績アップだね！」

ここまで読んで，何が悪いのかわからないと言う人。あなたは統計のウソに騙されやすい要素を持っているかもしれません。

■　全国一斉の学力調査の結果から，以下のようなポスターを文部科学省が作成しました。



この図の棒グラフが意味するものは，成績と朝ごはん摂取率との間に相関関係が見られるということのみです。

「朝ごはんを食べたら成績が上がる」という因果関係は，このグラフからは決して言うことはできません。それを言うには，これまで朝ごはんを食べなかった群が，朝ごはんを食べるようになると成績が上がることを示すデータ，あるいは朝ごはんを抜くと成績が下がるデータを示さなければなりません。

相関関係があることと，因果関係があることは，
まったく次元が異なる話です。

故に，このグラフだけを元に，朝ごはんを食べる・食べないで成績が決まるかのような発言をするのはナンセンスです。

ポスターに記されている（注）について，考えてみます。

> 注）得点は個々の児童・生徒が正答・準正答を解答した問題数の割合をもとに平均点を 500 点，標準偏差 100 点とする標準化を行った点数である。

これを端的に言い表すと，

得点は，平均点を 500 点，標準偏差 100 点とする標準化した点数。

変数の尺度を変換して，平均値や標準偏差が特定の値になるようにすることを標準化といい，そこで得られる得点は標準得点・Z 得点を呼ばれます。Z 得点の算出方法の考え方に基づいて，平均点を 50，標準偏差を 10 になるように設定した標準得点のことを T 得点，一般に偏差値を呼ばれます。すなわち，このポスターの数値は，平均点を 500，標準偏差を 100 とした偏差値です。「偏差値」というと何かと問題があるので，このような表現をしたのでしょう。

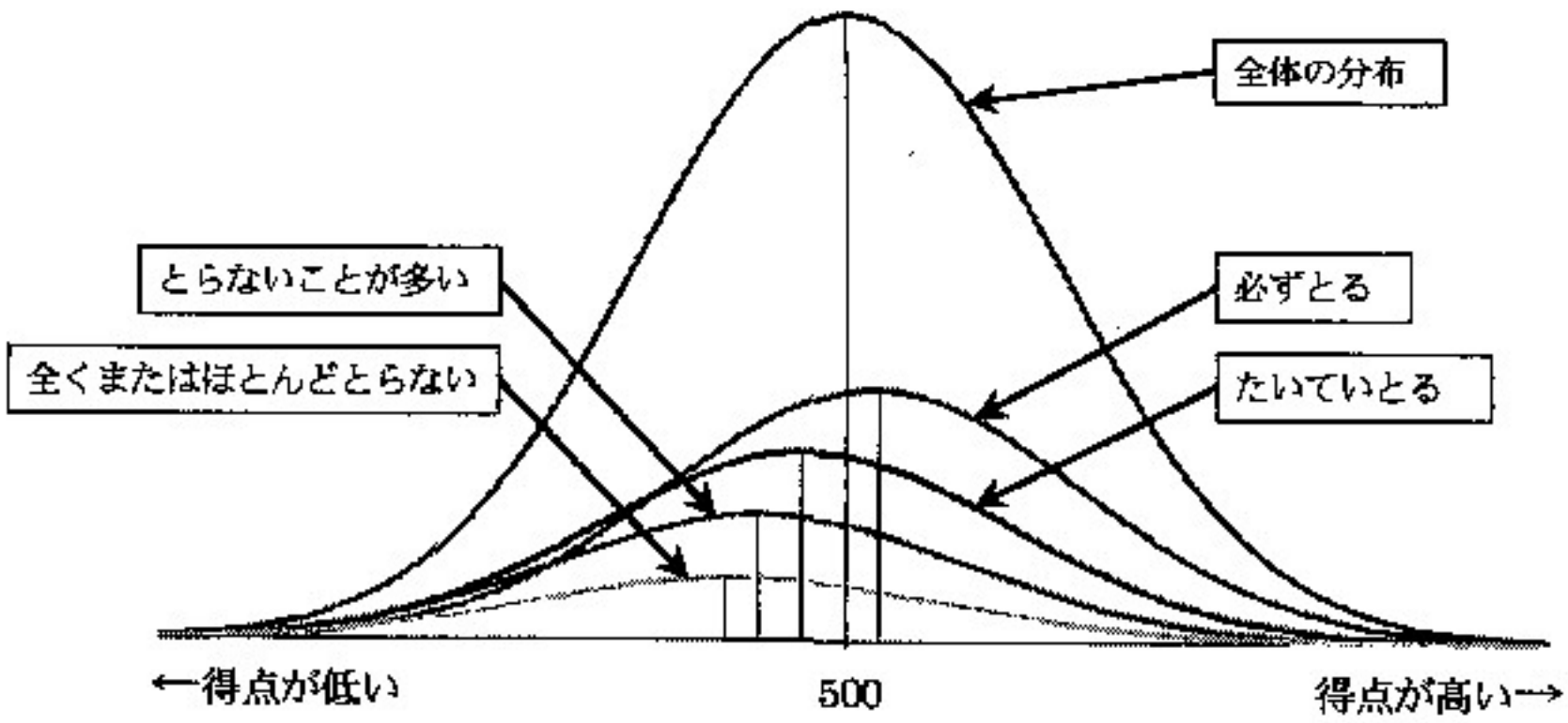
ここで留意したことは。

ある対象を平均という 1 つの代表値で語るとき，それが持つ**その他の情報はすべて捨て去られている**ことを意識しなければなりません。

例えば 20 人の生徒がいて，数学の試験で 10 人の生徒が 100 点，残り 10 人

の生徒が 0 点のとき，その平均点は 50 点。20 人全員が 50 点であるときも平均点は 50 点です。平均では点数の分布はわかりません。すなわち，実際の点数の分布は捨てさられています。また，このポスターのように標準化されたデータであれば，その段階で元データ（ローデータ）の分布は捨て去られています。

朝食を「必ずとる」という生徒が 4 割くらい，「たいていとる」という生徒が 3 割くらい，「とらないことが多い」という生徒が 2 割くらい，「全くまたはほとんどとらない」という生徒が 1 割くらい，と想定したとき，標準化した得点の分布は以下のようになっていると考えることができます。



「必ずとる」生徒の得点の分布は右寄り（得点が高い方に），「たいていとる」生徒の分布は左寄り（得点が低い方に）に，とらない割合が高くなる程，分布は左によって行きます。これは，得点の平均が低くなっていくことに合致します。このように想定すると，「全くまたはほとんどとらない」という生徒でも，得点の高い生徒がいますし，「必ずとる」という生徒でも得点の低い生徒がいることになります。これが実感をあっていませんか。

そうであればやはり，

「よし、これからもきちんと朝ごはんを食べよう。」

「朝ごはんを食べて成績アップだね！」

という，生徒の一言には違和感を持ちます。

見せかけの相関

生ビールの売り上げとアイスクリームの売り上げの相関は強いと考えられます。これは，両方の変数に気温という変数が共通しているからと考えられます。つまり，

「気温が高いから，生ビールの売り上げが増える」

「気温が高いから，アイスクリームの売り上げが増える」

という因果関係が同時に成立しているので，見かけ上，2つの間の相関が強くなっています。これを見かけの相関と呼びます。

外れ値

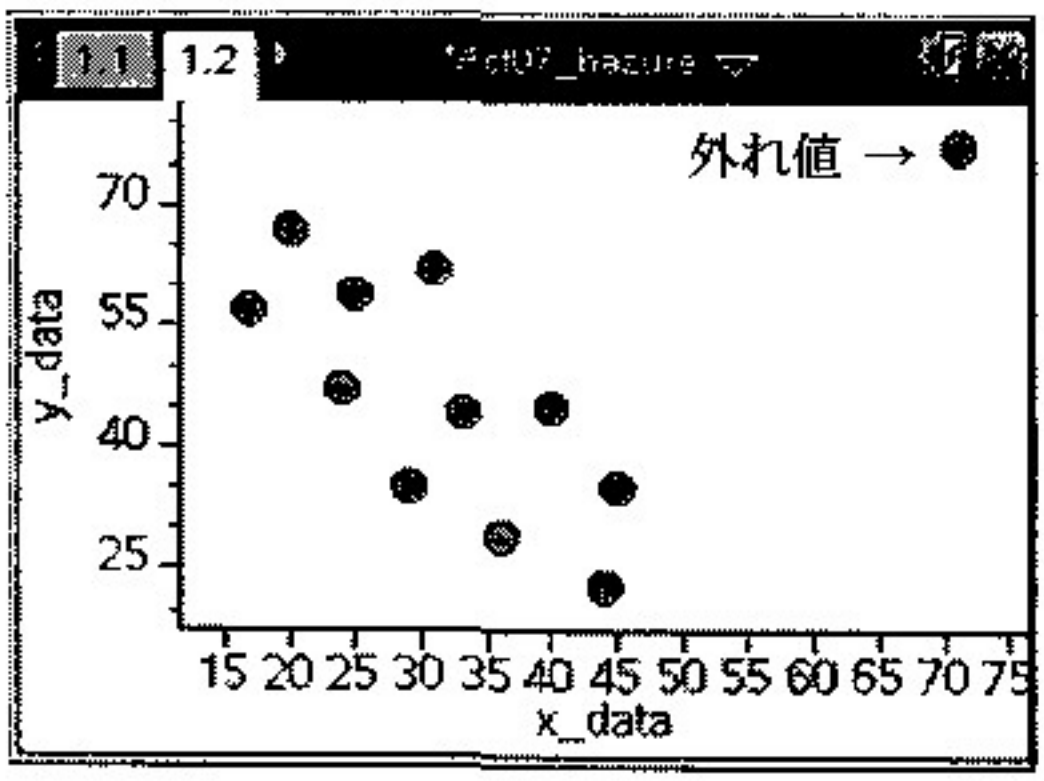
相関の強さを表す相関係数は2つの間の関係を見るのに有効なものですが，「外れ値（Outliers）」が大きく相関係数の値に影響を及ぼし本来の関係を見誤る可能性があります。

外れ値とは，全体の分布の中心から極端に外れた値のことです。データの入力ミスであれば修正すればいいのですが，そうでない場合は，外れ値の扱いを慎重に検討しなければいけません。

右の表7.2の散布図を描くと下記のようになります。

表7.2　外れ値

No.	x-data	y-data
1	25	59
2	17	57
3	**71**	**77**
4	29	35
5	33	44
6	20	67
7	40	45
8	45	35
9	44	23
10	36	29
11	31	62
12	24	47

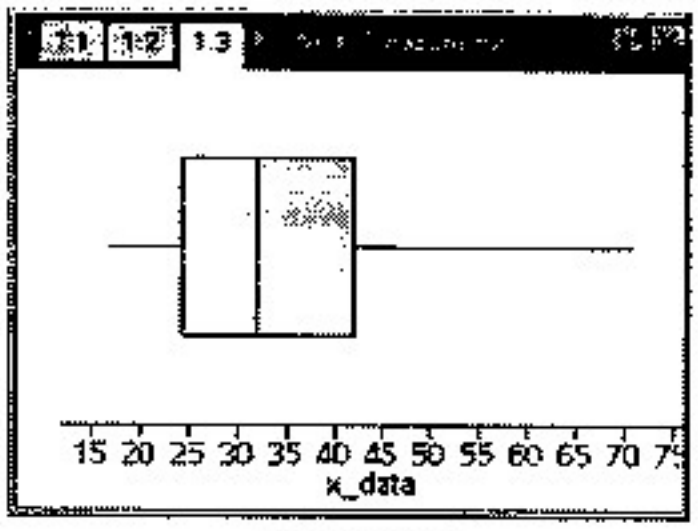


右上に１つだけ離れたデータがあります。このようなデータを外れ値と言います。上記のデータの相関係数を求めると 0.054 となります。No.3 の（71,77）の外れ値を除外して相関係数を求めると－0.743 となります。このように，外れ値は相関係数に大きな影響を与えることがあります。データから除外するかどうかは，目的と照らし合わせて慎重に検討する必要があります。

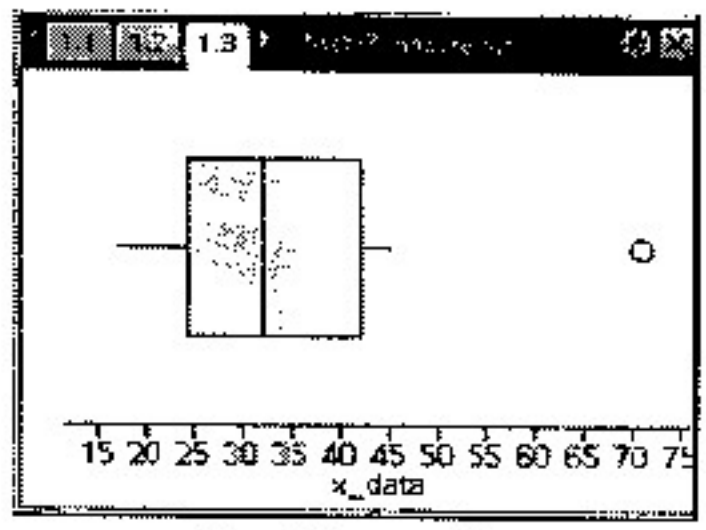
修正箱ひげ図

修正ひげ図では，区間（Q1−X，Q3+X）の外にある点を除外します。ここで，X=1.5(Q3−Q1)です。外れ値といわれるこれらの点は，ひげの外側にそれぞれプロットされます。

下の図は，表 7.2 の x-data について，箱ひげ図と修正箱ひげ図を描いたものです。散布図で外れ値を探す方法と併用するとよいでしょう。

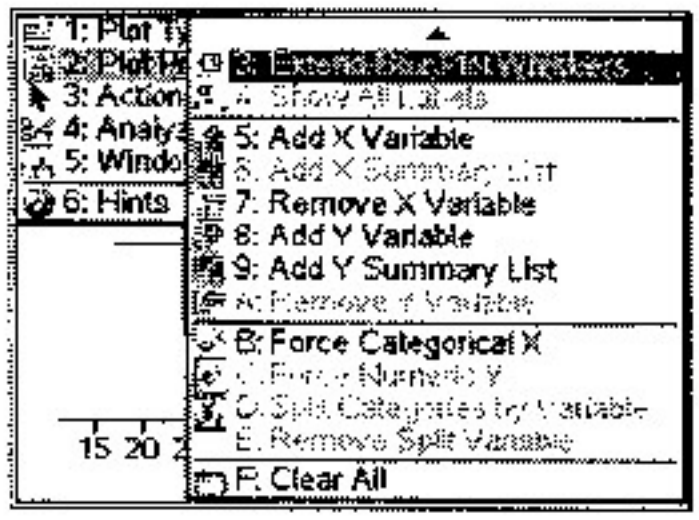


箱ひげ図



修正箱ひげ図

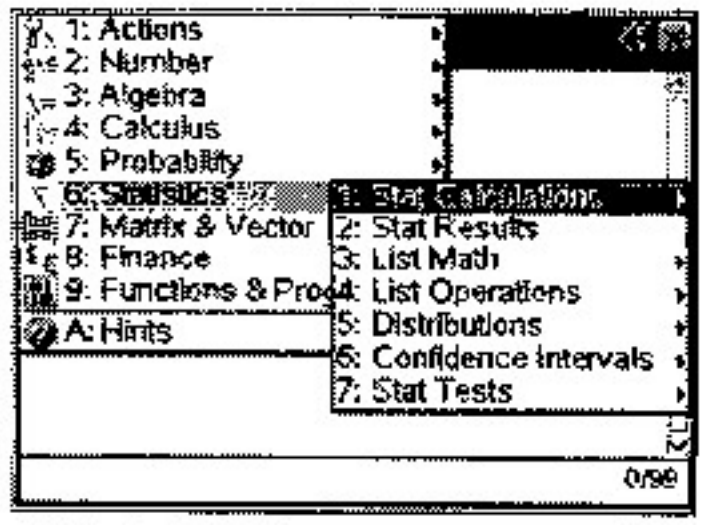
TI-Nspire では，修正箱ひげ図が規定値となっています。箱ひげ図に変更するには，menu を押して，「2:Plot Properties▶」「3:Extend Box Plot Whiskers」で enter を押します。



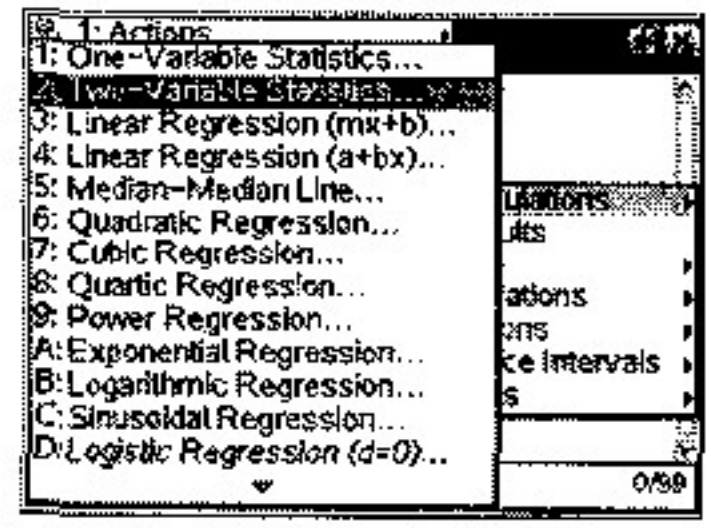
統計変数の意味，計算方法（手順は）が理解できれば，実際に統計変数がほしいときは，下記の方法で簡単に得ることができます。答え合わせにも利用できます。

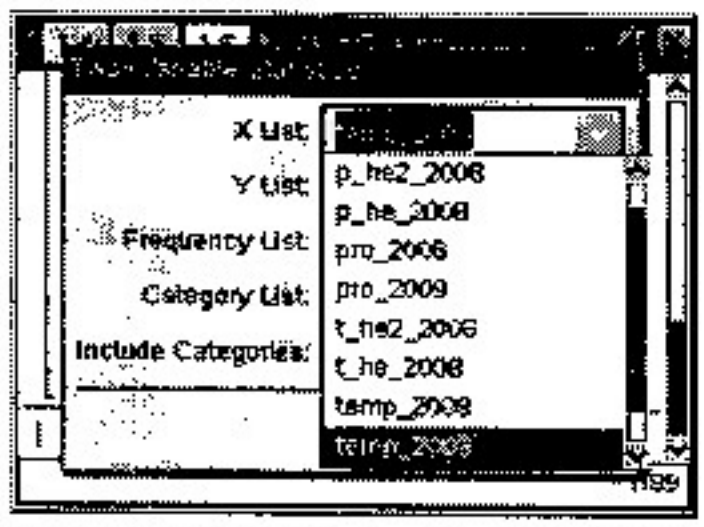
2 変数の統計量を求める

2009 年の平均気温(temp_2009)と生産量 (pro_2009)のデータについて統計変数を計算します。

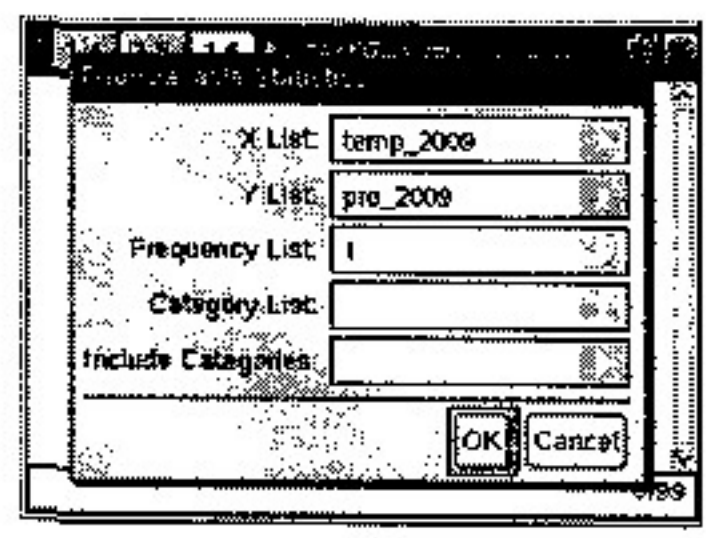


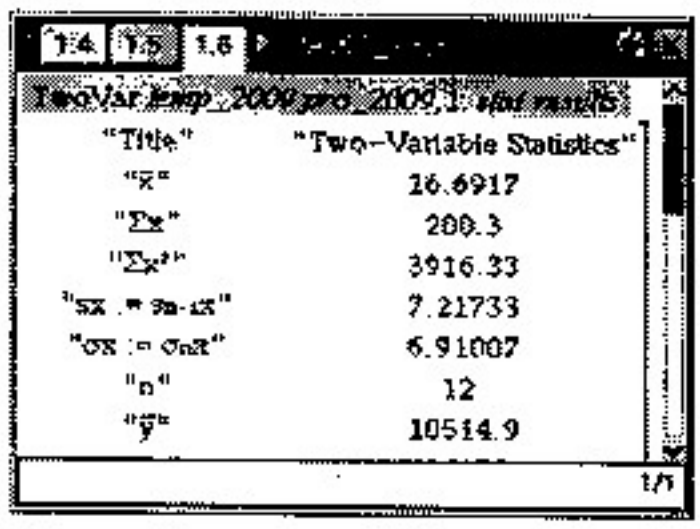
計算式を定義フィールドに計算式を入力し，enterを押します。



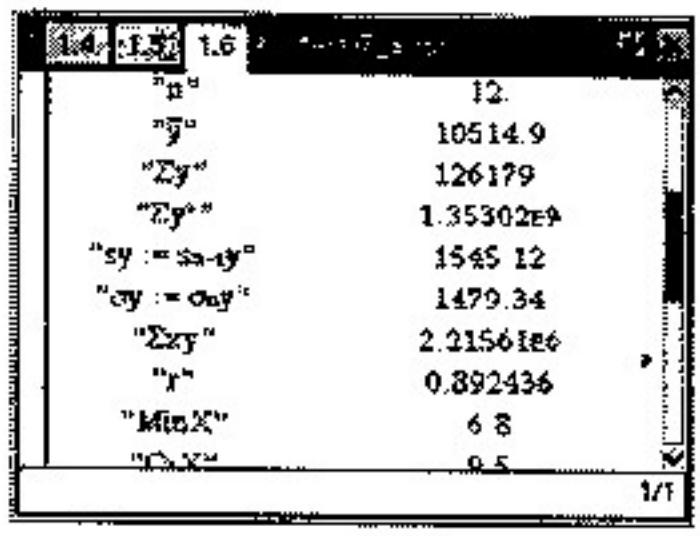


計算式を定義フィールドに計算式を入力し，enterを押します。





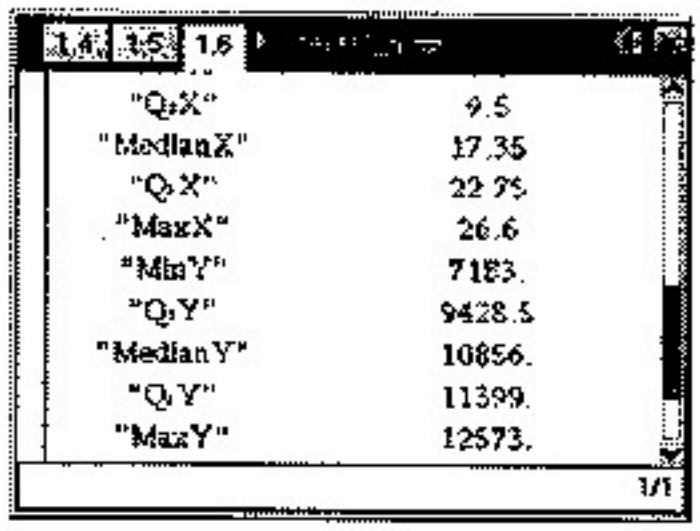
*x*の平均
*x*の総和
*x*の2乗の総和
*x*の標本標準偏差
*x*の母集団標準偏差
*x*のデータ数



"n"	12.
"$\bar{y}$"	10514.9
"Σy"	126179
"Σy²"	1.35302e9
"sy := Sn-1y"	1545.12
"σy := σny"	1479.34
"Σxy"	2.21561e6
"r"	0.892436
"MinX"	6.8
"Q1X"	9.5

1/1

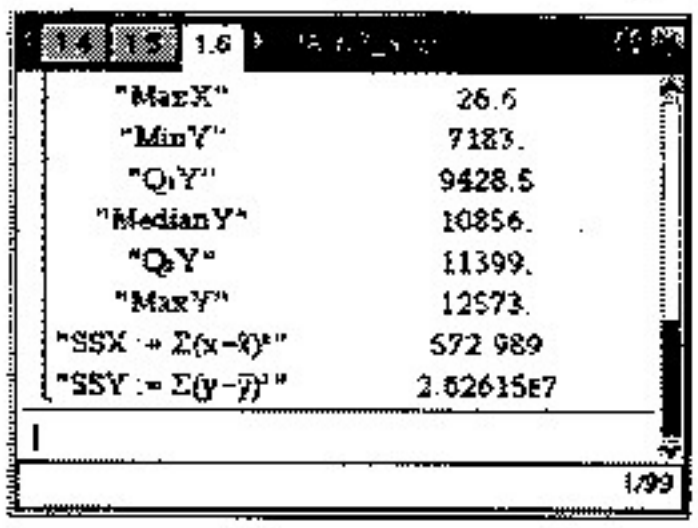
yの平均
yの総和
yの2乗の総和
yの標本標準偏差
yの母集団標準偏差
xy積の総和
相関係数
xの最小値



"Q1X"	9.5
"MedianX"	17.35
"Q3X"	22.75
"MaxX"	26.6
"MinY"	7183.
"Q1Y"	9428.5
"MedianY"	10856.
"Q3Y"	11399.
"MaxY"	12573.

1/1

xの第1四分位数
xの中央値
xの第3四分位数
xの最大値
yの最小値
yの第1四分位数
yの中央値
yの第3四分位数
yの最大値



"MaxX"	26.6
"MinY"	7183.
"Q1Y"	9428.5
"MedianY"	10856.
"Q3Y"	11399.
"MaxY"	12573.
"SSX := Σ(x-x̄)²"	572.989
"SSY := Σ(y-ȳ)²"	2.62615e7

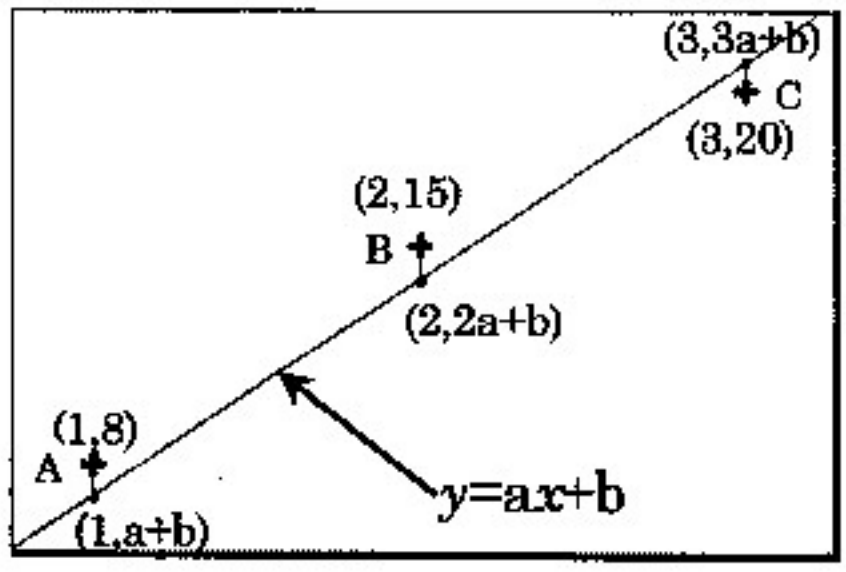
1/99

xの偏差の2乗の総和
yの偏差の2乗の総和

1次回帰式を求める 「2次関数の最大・最小の応用」

1次回帰式は，与えられた散布図上にある“すべて点を満足するような直線”です。「すべての点を満足する」ということを以下のように考えます。求めるべき1次関数を $y=ax+b$ とする，各データからその直線に下した足の長さが最も小さくなるような a，b を求めます。

ここでは，3点で考えます。



A(1, 8)，　B(2, 15)，　C(3, 20)の3点から直線 $y=ax+b$ に下した直線との交点は，それぞれ

$(1,\ a+b)$，　$(2,\ 2a+b)$，$(3,\ 3a+b)$となります。ここで，A から直線に下した足の長さは，$|a+b-8|$ で表すことができます。同様にB，Cから下した足の長さを表し，それらを足した値Dは，以下のように表すことができます。

$$D=|a+b-8|+|2a+b-15|+|3a+b-20|$$

このDが最も小さくなる a，b を求めればよいことになります。

計算を簡単にするために2乗します。これが「最小二乗法」の考え方です。

$$D=(a+b-8)^2+(2a+b-15)^2+(3a+b-20)^2$$

以下，Dが最小となる a，b を求めます。

$$D=14a^2+12ab-196a+3b^2-86b+689$$

$$=14\left(a^2+\frac{12b-196}{14}a\right)+3b^2-86b+689$$

$$=14\left(a+\frac{3b-49}{7}\right)^2-14\left(\frac{3b-49}{7}\right)^2+3b^2-86b+689$$

$$=14\left(a+\frac{3b-49}{7}\right)^2-\frac{3}{7}b^2-2b+3$$

$$=14\left(a+\frac{3b-49}{7}\right)^2+\frac{3}{7}\left(b-\frac{7}{3}\right)^2+\frac{2}{3}$$

Dを最小にするには，$(a+\frac{3b-49}{7})$ が0になり，$\frac{3}{7}(b-\frac{7}{3})$ が0になればよいので，

$\frac{3}{7}(b-\frac{7}{3})=0$ から $b=\frac{7}{3}$

$a+\frac{3b-49}{7}=0$ に b を代入して，$a=6$

よって，Dは，$a=6$　　$b=\frac{7}{3}$ のとき最小になる。

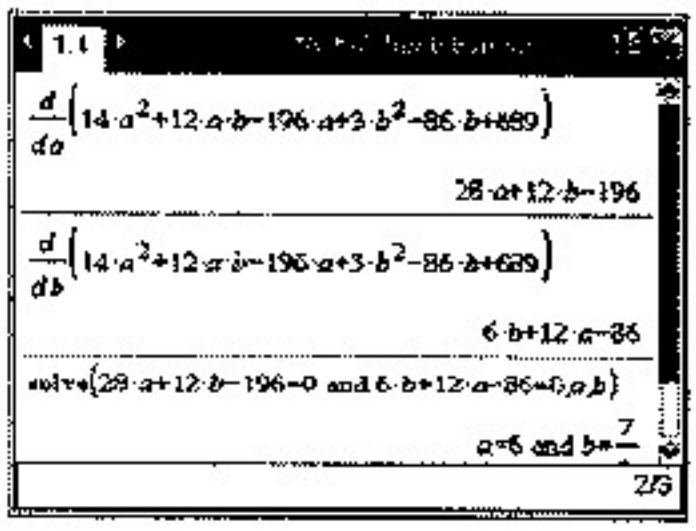
一般化して考えれば，

$\sum(y_k-ax_k-b)^2$ を最小にする a，b を求めることです。

参考

$D=14a^2+12ab-196a+3b^2-86b+689$

の最小値を偏微分で解くと以下のようになります。



a について微分 (a が変数，b は定数)

b について微分 (b が変数，a は定数)

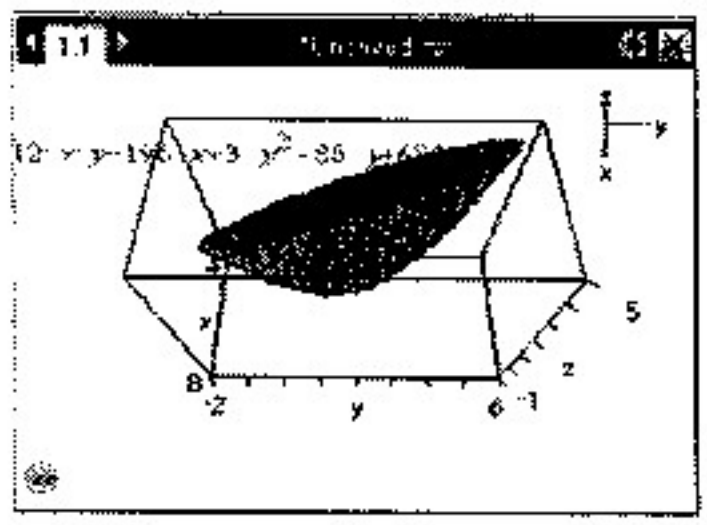
それぞれで求められた式の連立方程式を解く。

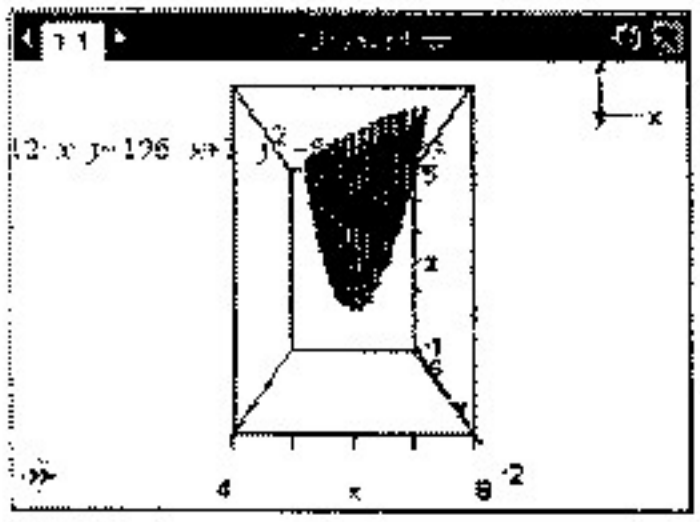
偏微分で解いていることを視覚的に確認するために，前記の式を下記のようにして，3次元のグラフを描きます。

$Z=14x^2+12xy-196x+3y^2-86y+689$

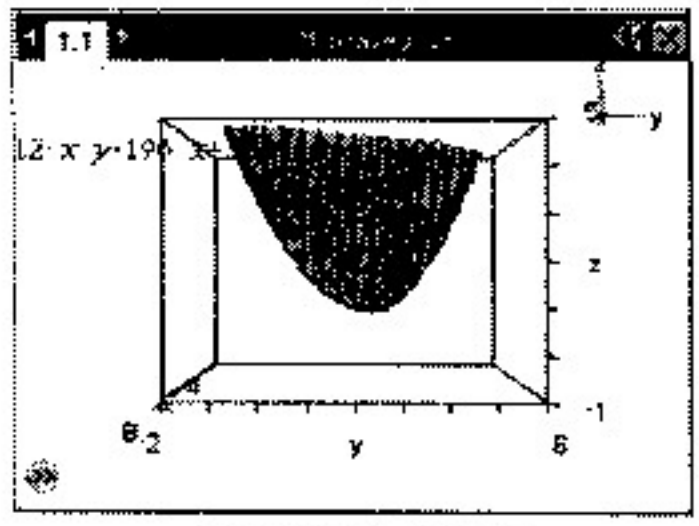
Z の最小値を求めることは，右の図のような楕円形のお椀の底になる x，y を求めていることになり。

この図を x 方向ら，y 方向から見た図が以下のものです。





x 方向から見た図



y 方向から見た図

共に一方を定数としていますから 2 次関数であり，それぞれが最小値をとる交点がお椀の形をした関数の底を求めていることがわかります。

1 次回帰式と相関係数の関係

1 次回帰式　$y = ax + b$ の

傾き a と切片 b は，以下のように定義されます。

● 傾き a＝ 相関係数 $\times \dfrac{(y\text{の標準偏差})}{(x\text{の標準偏差})}$ ……①

$$= \frac{\text{偏差積の平均}}{(x\text{の標準偏差}) \times (y\text{の標準偏差})} \times \frac{(y\text{の標準偏差})}{(x\text{の標準偏差})}$$

$$= \frac{\text{偏差積の平均}}{(x\text{の標準偏差})^2} \quad \cdots\cdots ②$$

● 切片 b＝y の平均－傾き a×x の平均

課題 7.3

前の例で用いた，A(1, 8)，　B(2, 15)，　C(3, 20)をもとに，下記の統計値を求め，その値を使って回帰式を求めなさい。

(1) 必要な統計量を求めなさい。

xの平均 …… (　　)

xの標準偏差 …… (　　)

yの平均 …… (　　)

yの標準偏差 …… (　　)

偏差積の平均 …… (　　)

相関係数 …… (　　)

(2) 各統計量を使って傾きaと切片bを求めなさい。

$a=$

$b=$

コラム
Column

統計のウソ

「嘘には三種類ある。

ただの嘘、大嘘、そして統計である。

by ベンジャミン・ディズレーリ」

昔のイギリスの政治家がこんな名言を残しているくらいで，統計というものはよく嘘に利用される。というよりも，統計データは簡単に自分に都合のいいように曲解できると言うべきでしょうか。また，都合のいいように改ざんすることもできます。

下記のニュースは，相関関係と因果関係，データの意図した？改ざんについて考えさせられるニュースです。

猪口（少子化男女共同参画担当大臣）が “改竄図” 使い説明

ＴＶで「女性就業率と出生率は正比例」

平成 17 年 11 月 12 日

猪口邦子・少子化男女共同参画担当大臣は六日、民放テレビに出演し男女共同参画局が作成した図を基に、経済協力開発機構（OECD）加盟国では女性の労働力率が高い国は出生率も高い、と説明。だが、この図は、同局が設けた恣意（しい）的尺度に基づき、それに全く合致しない加盟国、トルコ、メキシコが除外されており、「国民に向けてデマを流したに等しい」と批判の声が上がっている。

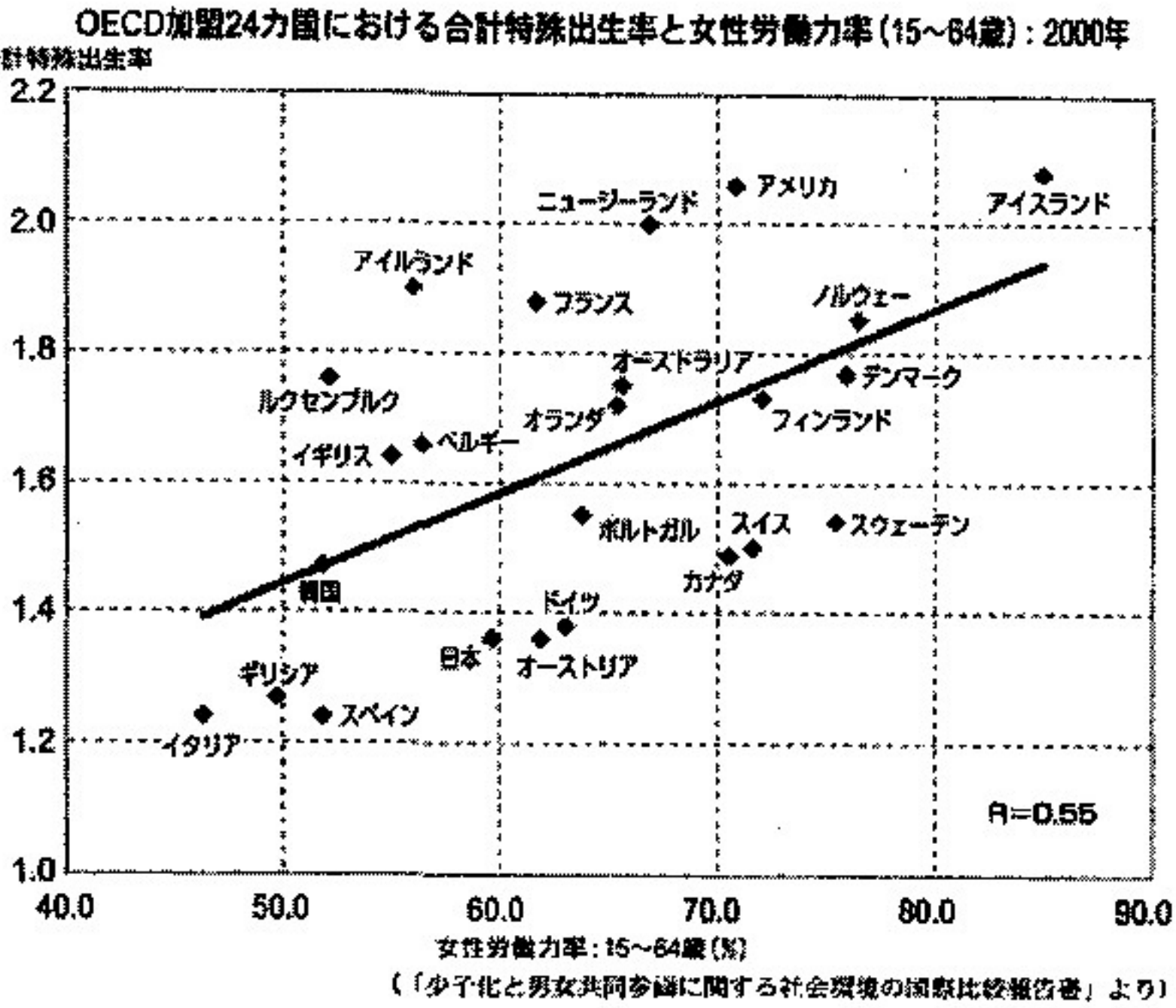
（山本　彰）

http://www.worldtimes.co.jp/special2/inoguti/051112.html

OECDで合致せぬ国を除外

「所得水準低い国外した」と男女共同参画局

「国民にデマ流した」と批判も



（「少子化と男女共同参画に関する社会環境の国際比較報告書」より）

この図は、「少子化と男女共同参画に関する専門調査会」（会長・佐藤博樹東大教授）が今年の九月にまとめた「少子化と男女共同参画に関する社会環境の国際比較報告書」に掲載されているもの。

猪口大臣は、フジテレビの報道番組『報道2001』でこの図を掲げ「『女性が外で働くから子供を生まなくなる』という人がいるが、この（グラフに見る）ように、労働力率が上がれば、合計特殊出生率も上がっている。日本の場合は、労働力率も出生率も低い」と訴えた。

この説明は、大沢真理東大教授らジェンダーフリー推進論者が、繰り返し主

張してきたもの。少子化には晩婚化、非婚化などさまざまな要因があり、この説明には異論も多い。

男女共同参画局は、女性の労働力率は低いがOECDの中で出生率が一位と二位のメキシコとトルコを除外した理由について、本紙の問い合わせに、「一人当たりのGDP（国内総生産）が一万ドル以下の国は除いている」とし、「一万ドル以上の国だけにすると所得水準で比較的日本に近いため」と説明した。メキシコ、トルコの一人当たりGDPは、それぞれ約六千二百ドルと二千八百九十ドル（二〇〇三年現在）。

これに対し、林道義・前東京女子大学教授は、「全体だってバラバラなのだから、GDP一万ドルで線引きするのは恣意的な操作と言える」と批判。日本は一人当たりのGDPが三・九万ドルで、それが一万ドル強のギリシャ、ポルトガルとの格差も大きい。

林氏は近著『家族を蔑む人々』で、同局ホームページ掲載の猪口大臣が示したものとほぼ同じ図を取り上げ、トルコ、メキシコ両国が除外されている点を問題にし、「出生率と女性労働力率の『相関図』は改竄（かいざん）」と、この図の欺瞞（ぎまん）性を厳しく指摘。

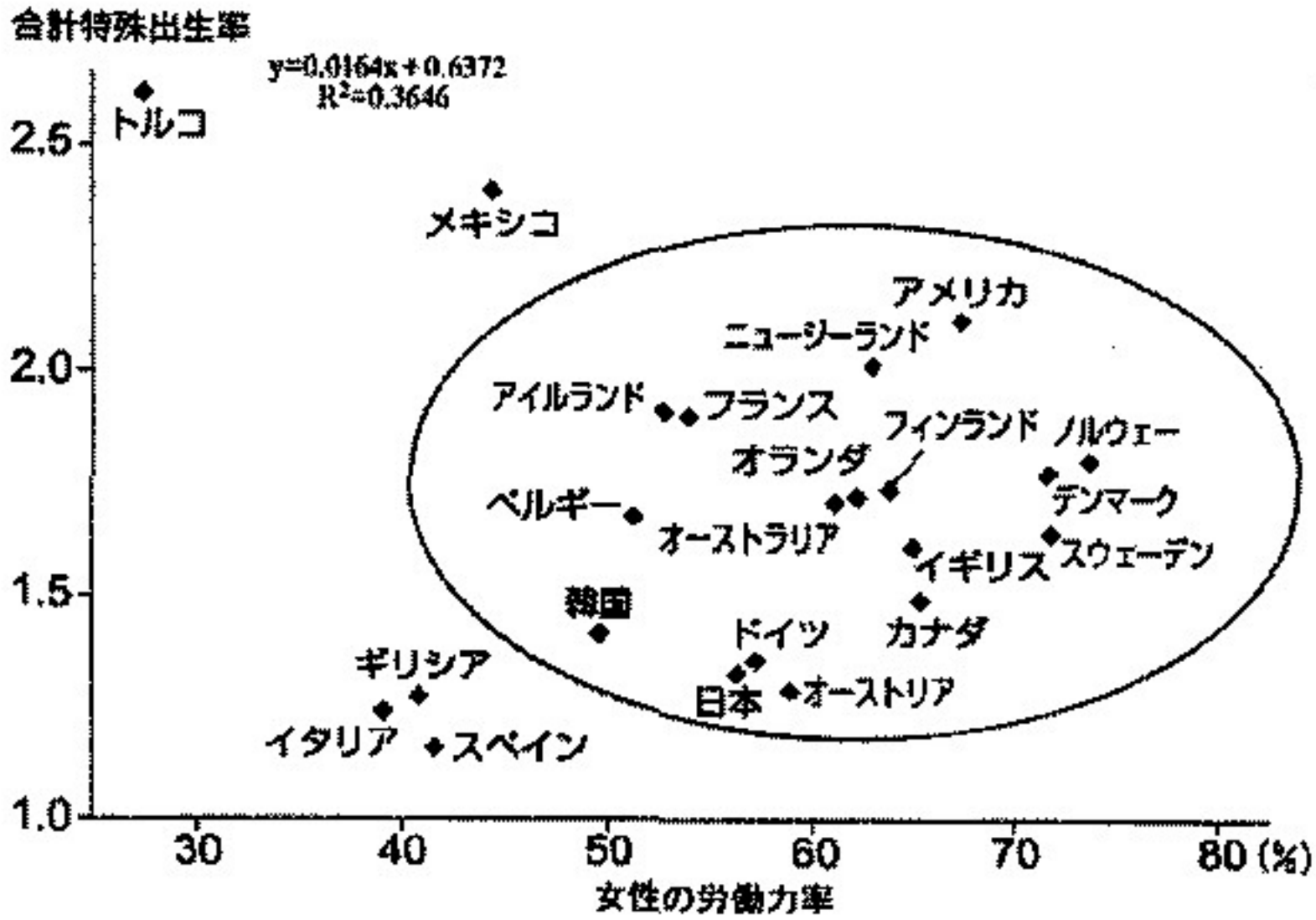
その上で、この両国を加え、左下のギリシャなど三カ国をはずして、他の国の塊を円で囲めば、「ランダムなばらつきが見られ、何の相関関係も見られない」（同著）と書いたばかり。

図にはまた、女性の労働力率と出生率に相関関係があるかのように、右上がりの補助線が引かれている。同局は、これに関して「バラツキの中にも一定の傾向を示す線としては許容範囲」としながらも、「どの程度が許容範囲か明確な基準はない」と説明している。

猪口大臣が言う傾向を最も示しているアイスランドは、人口が二十九万人。わが国と人口規模がよく似たメキシコ、トルコを外し、所得水準だけを尺度に、それが全く違う人口小国を書き込んだ図が参考になるかどうかは疑問が残る。

メキシコ、トルコの出生率の高さはカトリック、イスラム教という宗教上の要因が考えられる。人口に加えて宗教の視点も除外したことになる。

猪口大臣が、問題の多い図を使って、影響力のあるテレビ番組でジェンダーフリー論者と同じアピールをしたことに関して、林前教授は「大臣は中立でなければいけないのに、これでは国民にデマを流したようなもの。学者出身で厳密な人かと思ったが……」と語っている。



(林道義著『家族を壊す人々』より)

みなさんのご感想は如何でしょうか。

女性が働いている人が多いほど，子供が多く生まれる。この 2 つを比べることに意味があるのでしょうか。何か意図があってこのような図を作成しているのでしょうか。ほんとかな？ おかしいのでは？ と思う感覚を養う必要があるようです。