平成27年度　数学科リレー講座

統計入門

3日目「相関」

2015年8月26日
平山 裕之

目 次

1　いくつかの変量の関係

1.1　はじめに

1 つの対象に身長，体重，視力，…など複数の変量 (変数) が与えられていることがあります。これらの変量の間に何かの関係があるかどうかを調べてみましょう。これらの変量の間に何かしらの関係があれば，一方の変量から他の変量を予測することができるようになります。ここでは，2 変量の関係について調べることにします。

1.2　散布図

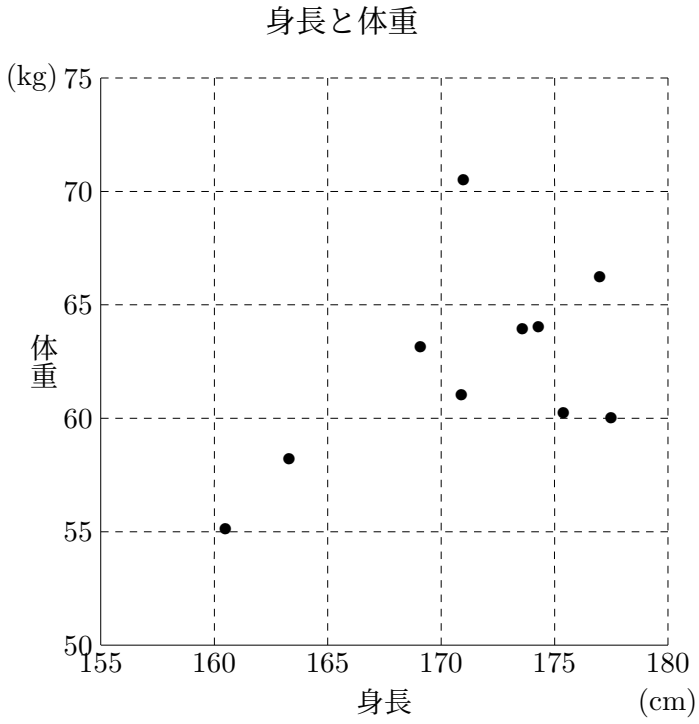
下の表は 10 人の人たちの身長と体重を測定した結果です。

表 1: 10 人の身長と体重を調べた資料

番号	身長	体重
1	177.0	66.2
2	177.5	60.0
3	163.3	58.2
4	170.9	61.0
5	169.1	63.1
6	171.0	70.5
7	173.6	63.9
8	174.3	64.0
9	160.5	55.1
10	175.4	60.2

資料の表だけでは 2 変量の関係をつかみにくいので，身長を x 座標，体重を y 座標として，各人のデータを平面上の点で表すと，2 変量の間にある大体の関係を知ることができます。このような図を散布図または相関図といいます。

身長と体重はどちらも体の大きさをあらわしているので，体が大きければ身長が高く体重も重いだろうということは，既に判っていますが，散布図をかけば，その関係が視覚的により明瞭になります。



散布図を見る時には,

- 全体的な傾向をつかみ，その関係と強さを見る。
- 全体的な傾向から著しく離れたデータがあるかどうかを見る。

というポイントに注意をします。

1.3　正の相関・負の相関

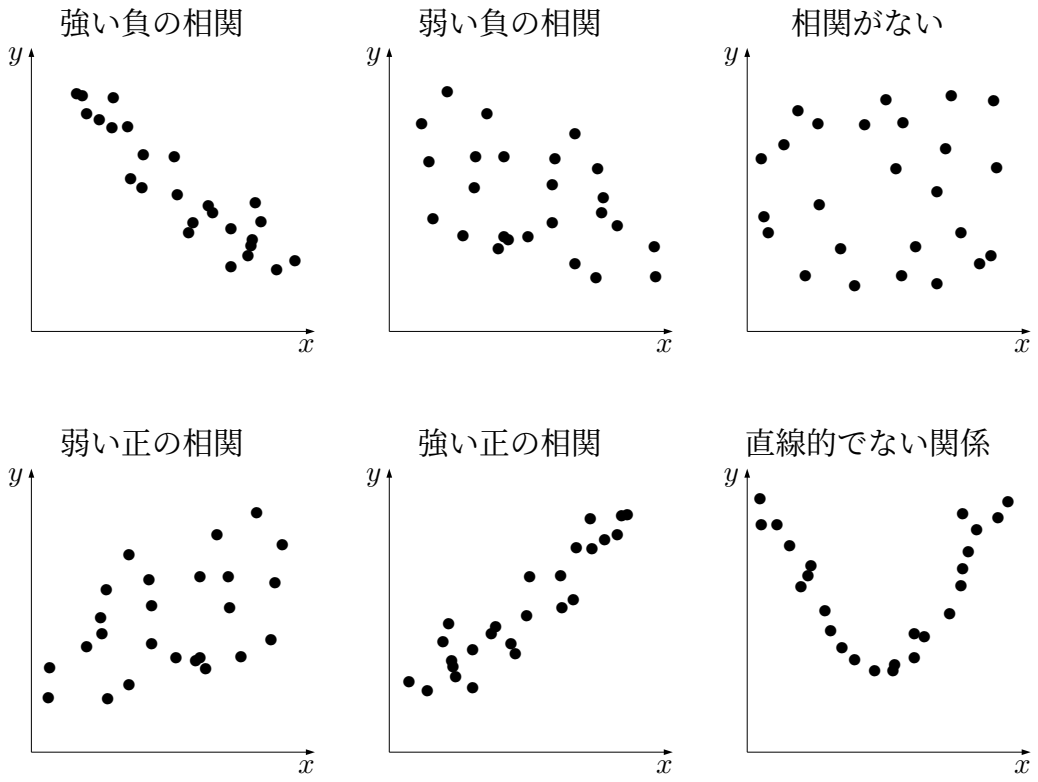
散布図において,

(1) 一方が増加すると他方が直線的に増加する傾向があるとき，正の相関があるといいます。

(2) 一方が増加すると他方が直線的に減少する傾向があるとき，負の相関があるといいます。

(3) どちらでもないとき，相関関係はない (無相関) といいます。

次の図は散布図の典型的な形です。



直線上に並ぶ度合いによって，強い相関，弱い相関という言い方をします。最後の散布図のように，直線的ではなくても曲線的な関係がある場合もあります。相関という用語は，直線的な関係を表しているときに使われます。

2　相関係数

散布図から 2 つの変量 x, y の間に直線的な相関が認められるとき，この関係の度合いを数値で表してみましょう。

2.1　1 変量の統計値

次のような変量の資料が与えられたものとします。

番号	x	y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

まず，それぞれの変量について，変量 x の平均, 分散, 標準偏差を $\bar{x}, s_{xx}, s_x$，変量 y の平均, 分散, 標準偏差を $\bar{y}, s_{yy}, s_y$ とします。

計算式は次のように与えられます。
平均

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},\ \bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

分散

$$s_{xx} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

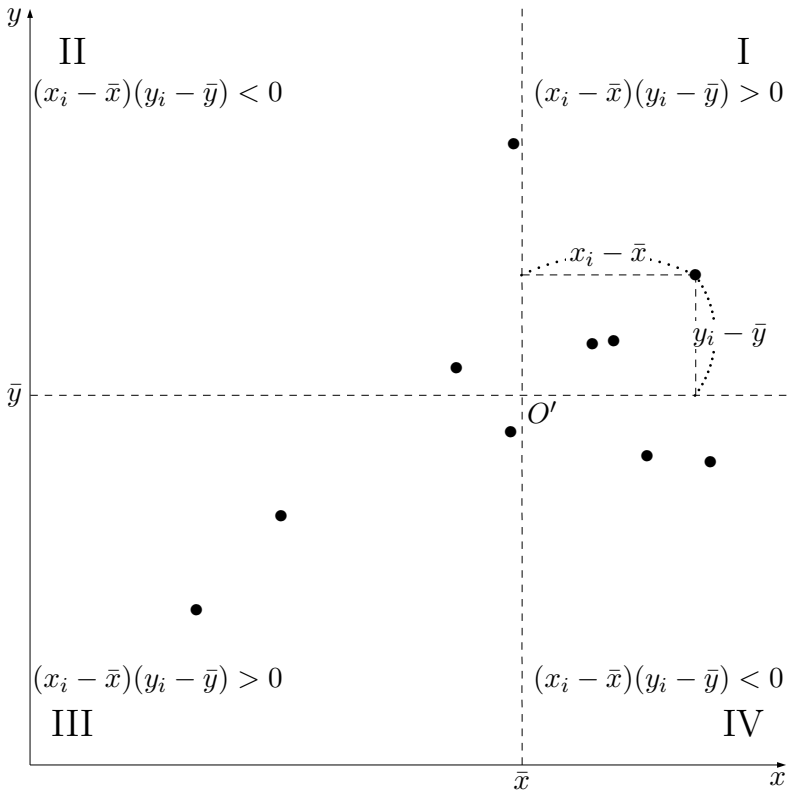
$$s_{yy} = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n}$$

標準偏差

$$s_x = \sqrt{s_{xx}},\ s_y = \sqrt{s_{yy}}$$

2.2 共分散

続いて散布図上に，点 $O'(\bar{x}, \bar{y})$ を原点とする座標軸を考えて，平面を I,II,III,IV の 4 つの象限に分けます。散布図上の n 個の点は，この 4 つの象限のいずれかに属します。



x と y の間に正の相関があるときは，第 I と第 III 象限に点が多く，第 II と第 IV 象限に点が少ないと考えられます。逆に，x と y の間に負の相関があるときは，第 I と第 III 象限に点が少なく，第 II と第 IV 象限に点が多いと考えられます。

データ (x_i, y_i) に対して，平均 $(\bar{x}, \bar{y})$ からの偏差の積 $(x_i - \bar{x})(y_i - \bar{y})$ をとると，その符号は

(1) 第 I 象限に点があるとき，$(x_i - \bar{x}) > 0,\ (y_i - \bar{y}) > 0$ だから $(x_i - \bar{x})(y_i - \bar{y}) > 0$

(2) 第 II 象限に点があるとき，$(x_i - \bar{x}) < 0,\ (y_i - \bar{y}) > 0$ だから $(x_i - \bar{x})(y_i - \bar{y}) < 0$

(3) 第 III 象限に点があるとき，$(x_i - \bar{x}) < 0,\ (y_i - \bar{y}) < 0$ だから $(x_i - \bar{x})(y_i - \bar{y}) > 0$

(4) 第 IV 象限に点があるとき，$(x_i - \bar{x}) > 0,\ (y_i - \bar{y}) < 0$ だから $(x_i - \bar{x})(y_i - \bar{y}) < 0$

になります。

したがって, x と y の間に正の相関があるときは, 第 I と第 III 象限にある $(x_i - \bar{x})(y_i - \bar{y}) > 0$ となる点が多いので，偏差積の平均

$$s_{xy} = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

は正の値をとります。

逆に，x と y の間に負の相関があるとき，s_{xy} は負の値をとります。相関関係がないときは，点が 4 つの象限に均等にばらつくので，s_{xy} は 0 に近い値をとります。

このように，s_{xy} は相関の有無を調べる指標になり，x と y の共分散とよばれます。

2.3 相関係数の定義

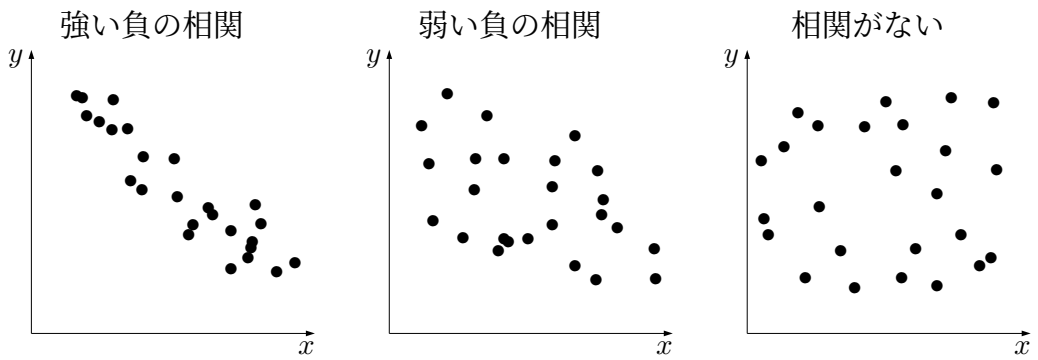
共分散は便利な指標ではありますが，身長の単位を cm から m に変えると，相関関係の強さは変わらないのに，s_{xy} の値は変わってしまいます。そこで単位のとり方に関係しないようにするために，x と y の標準偏差で割ります。

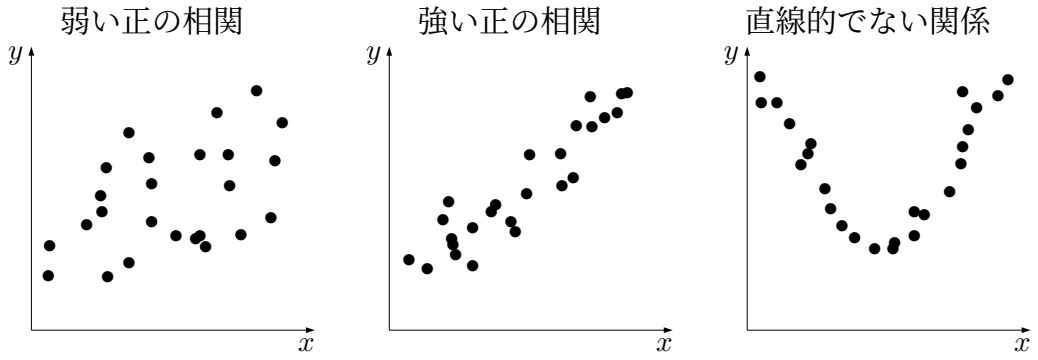
$$r = \frac{s_{xy}}{s_x s_y}$$

この値を，相関係数 (ピアソンの相関係数) とよびます。相関係数は $-1 \leqq r \leqq 1$ の値をとり，$|r|$ が 1 に近いほど相関は強くなります。$r = \pm 1$ のときは，散布図上のデータの各点は一直線上に並びます。(完全相関)

2.4 相関の強弱

先の相関図における相関係数を求めると次のようになります。





直線的関係がない相関図における相関係数は $r = 0.03$ ですが，2 次式的な関係があります。相関係数だけから判断すると，他の関係を見過ごすことがあるので，相関図等も用いて調べることが必要です。身長と体重の相関係数は $r = 0.52$ で，弱い正の相関があります。

2.5 相関関係と因果関係

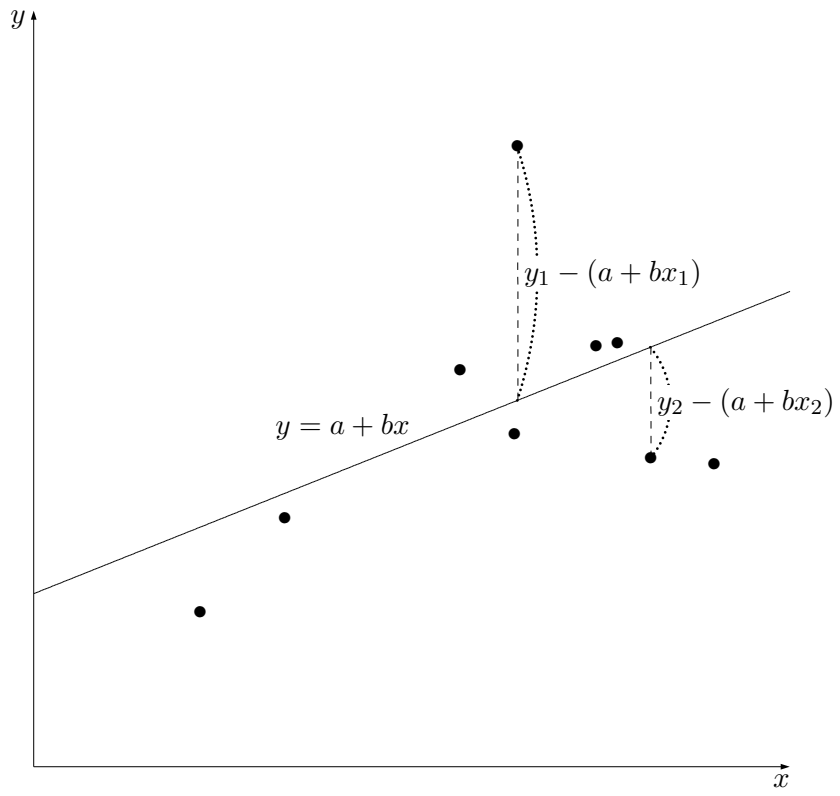
相関関係は 2 つの変量間の直線的な関係の強さを表していますが，原因と結果というような因果関係を表しているわけではありません。相関関係が強い場合でも，2 つの変量間に共通の別な変量が関係して，見かけ上，相関関係がみられることがあります。これを，擬似相関または偽相関といいます。都道府県別のごみ総排出量と交通事故件数では，ごみ総排出量が多い県は交通事故件数も多く，相関関係が認められますが，事故の原因がごみであるとは言えません。人口という共通の原因があって，見かけ上の相関関係が生じていると考えられます。(赤いポストと火事件数もそうでした。)

3 回帰直線

2 つの変量 x, y の間に直線的な関係が認められるとき，x の値から y の値を求める 1 次式を求めてみましょう。最もあてはまりのよい直線の式はどのように考えたらよいでしょうか。

3.1 直線関係を近似する直線

求める直線を $y = a + bx$ とおきます。変量 x, y が完全に直線の式 $y = a + bx$ に従うとすれば，$x = x_i$ のとき y は計算上は $a + bx_i$ になるはずですが，資料の y の値は y_i であり，実測値と計算値の間に差があります。この差 $\{y_i - (a + bx_i)\}$ を残差とよびます。残差ができるだけ小さくなるような直線が，最もあてはまりのよい直線と考えられるでしょう。



3.2 回帰直線の決定

そこで，残差の平方和

$$Q = \{y_1 - (a + bx_1)\}^2 + \{y_2 - (a + bx_2)\}^2 + \cdots + \{y_n - (a + bx_n)\}^2$$

が最小になるになるように a, b の値を決定します。このような方法を最小二乗法といいます。

a, b を決定するために Q の式を a, b で微分するのですが，ここでは，結果だけ示します。

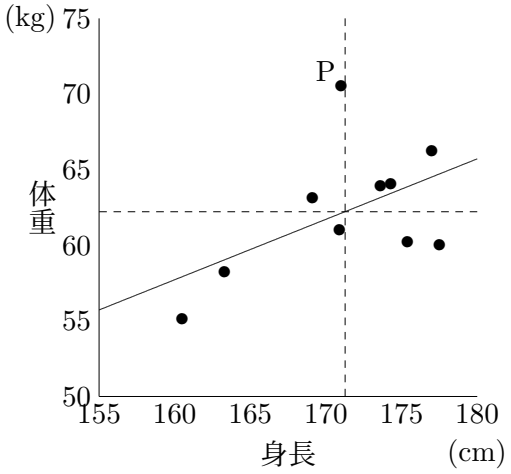
$$b = \frac{s_{xy}}{s_{xx}} \text{ (回帰係数)}, \ a = \bar{y} - b\bar{x}$$

と，x の分散と x, y の共分散から計算できます。

身長と体重の相関図における回帰直線の式は

$$y = 0.40x - 6.21, \ r = 0.52$$

と計算されます。

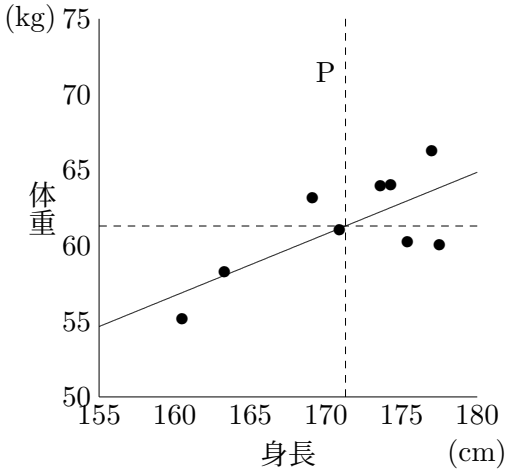


3.3 外れ値

ところで，相関図において，点 P は回帰直線との残差が大きな点です。このように，全体の傾向から外れた点がある場合，相関係数や回帰直線がこの点の影響をうけ，相関関係を適切に表していないことがあるので注意が必要です。この点を除外して相関係数および回帰直線を求めると，

$$y = 0.41x - 8.58,\ r = 0.72$$

と変化します。



4　回帰分析

4.1　回帰分析してみよう

都道府県別の人口とごみ総排出量の関係を調べ，回帰分析してみましょう。
下の表は 5 県の人口とごみ総排出量の資料です。

(1) 人口を x 軸に，ごみ総排出量を y 軸にとって散布図をかく。

(2) 平均, 分散, 相関係数, 回帰直線を作業表の各欄をうめながら求める。

(3) 資料の空欄の値を予測する。

表 2: 人口とごみ総排出量

都道府県	人口 (千人)	ごみ総排出量 (千トン)
岩手	1303	449
山形	1152	377
石川	1163	425
滋賀	1415	454
長崎	1408	497
東京	13230	?
富山	?	402

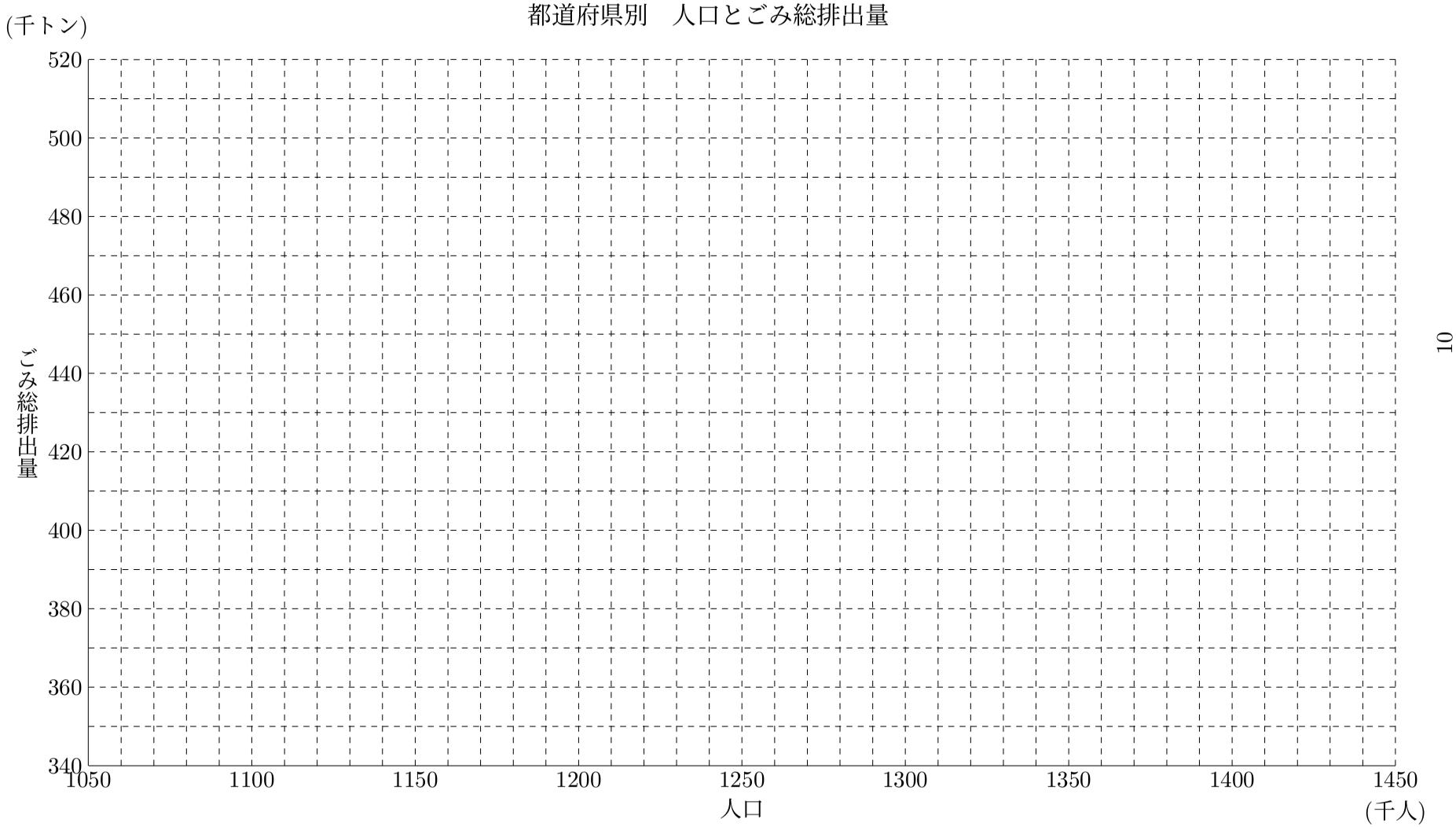
統計学習サイト「なるほど統計学園」より

参考文献

[1] 教科書『詳説 数学 I』, 啓林館

[2] 鷲尾泰俊『日常のなかの統計学』(新装版数学入門シリーズ), 岩波書店,2015

[3] 湧井良幸・湧井貞美『統計学の図鑑』, 技術評論社,2015


都道府県別　人口とごみ総排出量
(千トン)
520
500
480
460
440
420
400
380
360
340
ごみ総排出量
1050
1100
1150
1200
1250
1300
1350
1400
1450
人口
(千人)

都道府県	x	y	x^2	y^2	xy
岩手	1303	449			
山形	1152	377			
石川	1163	425			
滋賀	1415	454			
長崎	1408	497			
合計	①	②	③	④	⑤

$$ns_{xx} = \sum_i^n x_i^2 - \frac{\left(\sum_i^n x_i\right)^2}{n} = ③\boxed{} - \frac{①\boxed{}^2}{5} = ⑥\boxed{}$$

$$ns_{yy} = \sum_i^n y_i^2 - \frac{\left(\sum_i^n y_i\right)^2}{n} = ④\boxed{} - \frac{②\boxed{}^2}{5} = ⑦\boxed{}$$

$$ns_{xy} = \sum_i^n x_i y_i - \frac{\left(\sum_i^n x_i\right)\left(\sum_i^n y_i\right)}{n}$$

$$= ⑤\boxed{} - \frac{①\boxed{} \times ②\boxed{}}{5} = ⑧\boxed{}$$

$$\text{相関係数 } r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{⑧\boxed{}}{\sqrt{⑥\boxed{} \times ⑦\boxed{}}} = ⓡ\boxed{}$$

$$b = \frac{s_{xy}}{s_{xx}} = \frac{⑧\boxed{}}{⑥\boxed{}} = ⑨\boxed{}$$

$$a = \bar{y} - b\bar{x} = \frac{②\boxed{} - ⑨\boxed{} \times ①\boxed{}}{5} = ⑩\boxed{}$$

$$\text{回帰直線 } y = ⑩\boxed{} + ⑨\boxed{}x$$