大学評価・IR 担当者のための統計入門（補足資料）

2017/08/24

統計学は，実際に得られたデータ（**標本**）を理解して，そのデータが抽出された元の集団（**母集団**）の性質についての推測を行なうための方法論を扱う学問である．本章では，まずデータの形式について確認し，数値やグラフによるデータの要約方法について紹介する．得られたデータの要約を行なう方法論は，統計学の中でも特に**記述統計学**と呼ばれる．記述統計学は，大量のデータを扱うことの多い現代社会において，その重要性が高まっている．

1　統計学で扱うデータの形式

以下の数値は，平成 21 年に国土交通省に新型届出のあった普通/小型自動車のうち，一部の車種についての燃費値 [km/ℓ] を並べたものである．

平成 21 年新車燃費データ

32.6 20.8 19.8 18.0 14.0 9.7 19.8 17.4 13.2 11.0 18.0 18.0
18.0 17.2 12.8 12.0 8.4 14.0 12.4 11.6 8.7 15.6 13.4 12.6
12.4 8.8 26.0 25.8 14.4 12.4 11.6 11.2 9.9 8.6 8.2

統計学では，このようなデータ（数値や記号）の集合を，興味のある集団全体からランダムに抽出された個体の特定の項目（**変数**）を観測して得られた値（**観測値**）と考える．抽出された個体の集まりを**標本**と呼び，興味のある集団全体のことを**母集団**と呼ぶ（図 1）．標本に含まれる観測値を一般に以下のように表す．

$$x_1, x_2, \ldots, x_n$$

ここで，n は標本に含まれる個体の数（＝観測値の数）で，特に，標本であるということを強調する場合，**標本サイズ**もしくは**サンプルサイズ**と呼ぶ．燃費のデータの場合，$x_1 = 32.6, x_2 = 20.8, \ldots, x_{35} = 8.2$ であり，サンプルサイズ $n = 35$ である．

1 つの個体につき，2 項目の値を測定した場合の標本は **2 変数のデータ**と呼ばれる．例えば，燃費のデータにおいて，排気量 [cc] も同時に記録した場合には，

$$(32.6, 1797), (20.8, 1329), \ldots, (8.2, 3471)$$

のようになる．一般には

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

のように表現できる．$x_1, x_2, \ldots, x_n$ を変数 x で，$y_1, y_2, \ldots, y_n$ を変数 y で代表させる．この例の場合は，変数 x が燃費，変数 y が排気量を示している．

さらに，1 つの個体につき，多くの項目を測定した場合の標本は**多変量データ**と呼ばれる．表 2 は，多変量データの例である．サンプルサイズ n で p 変数の多変量データは一般には以下

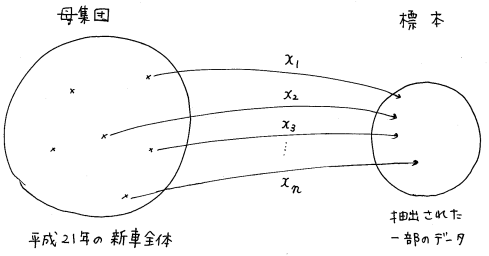

図 1　母集団と標本

のような行列の形式で表現できる．

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

表 2 のデータにおいては，$n = 35, p = 10$ である．$(x_{11}, x_{21}, \ldots, x_{n1})$ が車名に，$(x_{1p}, x_{2p}, \ldots, x_{np})$ が駆動方式に対応している．

2　尺度水準

表 2 のデータでは，総排気量，車両重量，燃費値などの変数の観測値は数値である一方，車名，通称名，駆動方式などの変数の観測値はは文字となっている．文字は例えば，駆動方式においては $\mathrm{F} = 1, \mathrm{A} = 2, \mathrm{R} = 3$ のように，便宜的に数値に置き換えることができるが，その数値の尺度は本質的に数値として得られたデータとは異なる．ここでは，とりうる値の尺度による変数の分類について説明する．

■名義尺度　駆動方式の例で利用した 1,2,3 という数値は駆動方式を分類するためだけに用いられる尺度で，数値の大小比較や演算を行なうことに意味はない．このような尺度を**名義尺度**と呼ぶ．割り当てる数値は，1,2,3 とする必要もなく，10,11,20 としても不都合はない．典型的な例としては，性別や血液型に割り当てられた数値や，ユーザー ID などがある．

■順序尺度　名義尺度に，大小比較ができる性質が加わったものが**順序尺度**である．表 2 のデータにおいて，低排ガス認定レベルは 3 と 4 のいずれかの値を取っている．レベル 4 の車はレベル 3 の車よりも排気ガスをより低減できていると解釈できるため，この数値には順序関係が存在する．典型的な例としては，成績評価（S=4,A=3,B=2,C=1,D=0 など）やアンケート調査における選択項目（特にそう思う=5, そう思う=4, どちらでもない=3, そう思わない=2, まったくそう思わない=1）などがあげられる．

■間隔尺度　順序尺度に，差が等しければ間隔も等しいという性質が加わったものが**間隔尺度**である．順序尺度であげたアンケート調査における選択項目の例では，5 と 4 の差と 4 と 3 の差が実質的に等しいかどうかは明確でないため，これは間隔尺

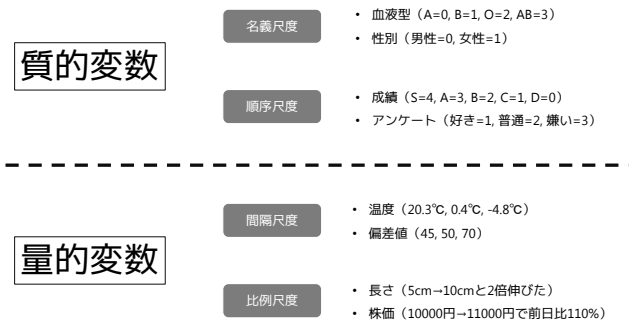

図 2　変数の分類と尺度水準

度であるとはいえない．間隔尺度の典型的な例としては，温度（摂氏）などがあげられる．

■比例尺度　間隔尺度に，数値同士の比に実質的な意味があるという性質が加わったものが**比例尺度**である．表 2 のデータにおいては，総排気量，車両重量，燃費値，CO2 排出量などが該当する．温度の場合，10 ℃から 20 ℃に気温が上昇したからといって，温度が 2 倍になったとはいわない．したがって，温度は比例尺度とは考えられない．比例尺度の場合は，「ゼロ」が存在しないことを意味する．典型的な例としては，重さや長さなどがあげられる．

名義尺度もしくは順序尺度の値をとる変数を**質的変数**，間隔尺度もしくは比例尺度の値をとる変数を**量的変数**と呼ぶ．データに含まれる変数がどのようなタイプのものであるかによって，統計学におけるどの分析手法を用いるかや，データをどのように表現するかも変わってくる．統計学の体系を理解するためにも，変数の分類や尺度水準について理解しておくことが重要である．

問 表 2 のデータにおける各変数について，それが質的変数であるか量的変数であるか，さらにどの尺度水準であるかを検討しなさい．

3　度数分布表

これまで，量的変数の観測値の特徴をいくつかの数値で表現してきた．より詳細に量的変数の観測値の分布を検討するために，**度数分布表**を作成する．度数分布表とは，量的変数の観測値の範囲をいくつかの区間に分割し，それらの区間（**階級**）ごとに含まれる観測値の個数（**度数**）をカウントして表にしたものである．表 2 のデータにおける，車両重量の度数分布表の例を表 1 に示す．データ全体の観測値の個数に対するその階級の度数の割合を**相対度数**という．また，その階級以下の階級に含まれる観測値の個数を**累積度数**という．データ全体の観測値の個数に対するその階級の累積度数の割合を**累積相対度数**という．**階級値**は各階級の区間のちょうど中央，すなわち区間の上限と下限を足して 2 で割った値である．

表 1　車両重量の度数分布表

階級	階級値	度数	相対度数	累積度数	累積相対度数
$800_{以上} \sim 1000_{未満}$	900	1	0.03	1	0.03
$1000_{以上} \sim 1200_{未満}$	1100	5	0.14	6	0.17
$1200_{以上} \sim 1400_{未満}$	1300	7	0.20	13	0.37
$1400_{以上} \sim 1600_{未満}$	1500	5	0.14	18	0.51
$1600_{以上} \sim 1800_{未満}$	1700	7	0.20	25	0.71
$1800_{以上} \sim 2000_{未満}$	1900	8	0.23	33	0.94
$2000_{以上} \sim 2200_{未満}$	2100	1	0.03	34	0.97
$2200_{以上} \sim 2400_{未満}$	2300	1	0.03	35	1.00
合計		35	1.00		

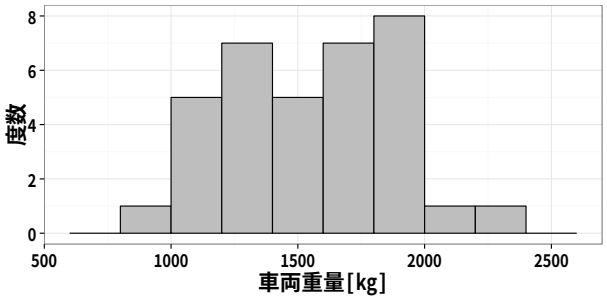

図 3　車両重量のヒストグラム

4　ヒストグラム

度数分布表を棒グラフのように表現したものが**ヒストグラム**である（図 3）．ヒストグラムによって，観測値の分布の特徴を直感的に理解することができる．ヒストグラムを観察する際には，ピークの数や対称性を確認しておくとよいだろう．2 つ以上のピークは，異質のデータが混在している可能性を示す．非対称なデータは，平均値を利用することへの危険性を示す．

■階級幅と階級数について　度数分布表の階級幅は，必ずしも等間隔である必要はない．例えば年間所得のデータに対する階級を，1000 万円までは 100 万円刻みで作成して，1000 万円以降は 1000 万円刻みで作成するような場合がある．このような度数分布表に対して，ヒストグラムの棒の高さを度数としてしまうと実際とは異なった印象を与えてしまう．階級幅が異なる場合には，ヒストグラムの棒の面積を度数に比例させるように描くべきである．特に，棒の面積を相対度数とした場合の棒の高さ（相対度数÷区間幅）のことを **密度**と呼ぶことがある．このような意味で，ヒストグラムと棒グラフは異なるものであるため，混同しないよう気をつけるべきである．

同一のデータに対して階級数の異なるヒストグラムを作成することができる．階級数が少なすぎたり，多すぎたりすると母集団におけるデータの分布を適切に反映したものとならないので，階級数の取りかたもヒストグラムを作成する際に注意すべき点のひとつである．

問 表 2 のデータで燃費値などの量的変数についてヒストグラムを作成しなさい．階級数を極端に少なくしたり多くしたりした場合のヒストグラムを作成し，比較してみなさい．また，度数分布表の一部の隣接した階級を併合した場合のヒストグラムを作成してみなさい．

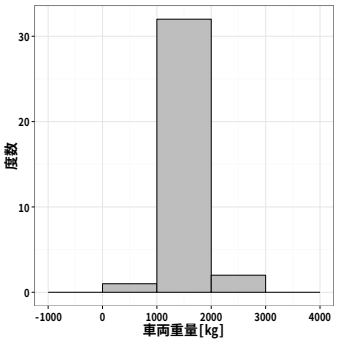

図 4　階級数が少ない場合の例

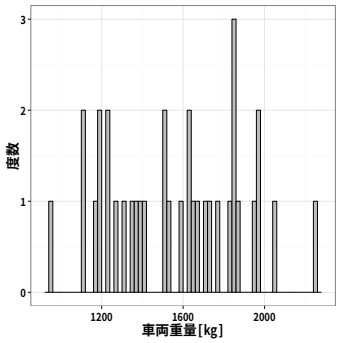

図 5　階級数が多い場合の例

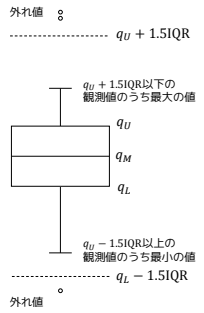

図 6　箱ひげ図の描き方

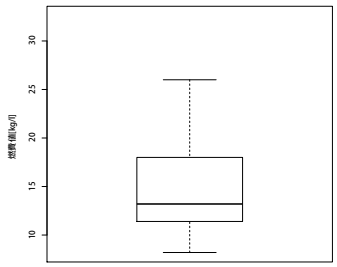

図 7　燃費値の箱ひげ図

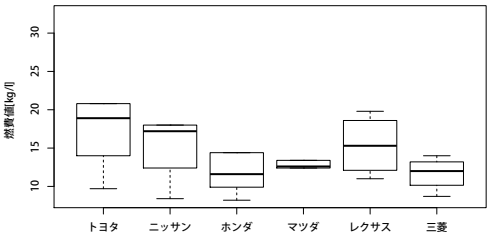

図 8　箱ひげ図による燃費値の分布の車名ごとの比較

5　四分位数

量的変数の分布を示す指標として**四分位数**がある．四分位数は昇順に並べた観測値を，それぞれの区間に含まれる観測値の個数が等しくなるように 4 分割する点である．四分位数は簡易的には以下のように求めることができる．中央値により観測値を 2 分割できるので，まず中央値を求めてから，中央値より大きい観測値の中央値と，中央値より小さい観測値の中央値を求めることにより全ての四分位数を得ることができる．四分位数の小さいほうから順に**第 1（下側）四分位数**，**第 2 四分位数**（＝中央値），**第 3（上側）四分位数**と呼び，それぞれ記号で q_L, q_M, q_U とあらわす．

パーセント点（百分位点）が用いられることもある．第 1 四分位数は 25 パーセント点，第 2 四分位数は 50 パーセント点，第 3 四分位数は 75 パーセント点（＝上側 25 パーセント点）に対応する．

四分位数を用いて散布度を表現することもある．上側四分位数と下側四分位数の差

$$\text{IQR} = q_U - q_L$$

を**四分位範囲**と呼ぶ．また，IQR の半分の値を**四分位偏差**と呼ぶ．

例 表 2 のデータにおいて，車名が「F 社」である車の CO2 排出量を昇順に並べると

$$89, 90, 161, 187, 200, 207, 235, 270, 283$$

となる．観測値の個数は 9 個であるから，中央値 $q_2 (= m)$ は 5 番目の 200 となる．第 1 四分位数 q_L は中央値より小さい 4 個の観測値の中央値であるから，$(90 + 161)/2 = 125.5$ となる．第 3 四分位数 q_U は中央値より大きい 4 個の観測値の中央値であるから，$(235 + 270)/2 = 252.5$ となる．四分位範囲 IQR は $252.5 - 125.5 = 127$ である．

6　箱ひげ図

四分位数を用いて，量的変数の分布を**箱ひげ図**により可視化することができる．図 6 に箱ひげ図の描き方を示している．箱の部分は箱の底の y 座標が第 1 四分位数，箱の上部の y 座標が第 3 四分位数，中央の水平線の y 座標が中央値となるように描かれる．すなわち，箱の y 座標の範囲に観測値の半分が含まれることになる．箱から上下に伸びた「ひげ」の部分は，上端が $q_U + 1.5\text{IQR}$ 以下の観測値のうち最大の値，下端が $q_L - 1.5\text{IQR}$ 以上の観測値のうち最小の値となるように描かれる．$[q_L - 1.5\text{IQR}, q_U + 1.5\text{IQR}]$ の範囲外の観測値を外れ値と定義し，観測値をプロットする．

例 表 2 のデータの燃費値についての箱ひげ図は，図 7 のようになる．第 1 四分位数 q_L は 11.4，中央値 q_M は 13.2，第 3 四分位数 q_U は 18.0 である．四分位範囲は

$$\text{IQR} = 18.0 - 11.4 = 6.6$$

であるから，

$$q_L - 1.5 \cdot \text{IQR} = 11.4 - 1.5 \cdot 6.6 = 1.5$$
$$q_U + 1.5 \cdot \text{IQR} = 18.0 + 1.5 \cdot 6.6 = 27.9$$

となる．区間 $[1.5, 27.9]$ に含まれる観測値のうち，最小値は 1.5 で最大値は 26.0 であり，これらがそれぞれ「ひげ」の下端と上端となる．範囲外の観測値 32.6 は，外れ値となるため，個別にプロットされる．

箱ひげ図もヒストグラムと同様に，量的変数の分布を理解するために有効である．箱ひげ図は，分布を理解することに加え，複数のグループ間での分布の比較をするために用いられる．図 6 は，燃費値の分布を車名ごとに箱ひげ図として表現したものである．

問 表 2 のデータで燃費値などの量的変数について箱ひげ図を作成しなさい．車名ごとの比較も行なってみなさい．

7　平均値の性質

n 個の観測値 $x_1, x_2, \ldots, x_n$ に対する，平均値は

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

で定義される．この定義を変形すると

$$\sum_{i=1}^{n} x_i = n\bar{x}$$

となり，「**観測値の合計は，平均値の n 倍に等しい**」ことが言える．また，

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \tag{1}$$

であることも導ける．

つまり，「**個々の観測値と平均値との差の合計は常にゼロ**」であることが言える．

■データ変換と平均値　$x_1, x_2, \ldots, x_n$ を

$$y_i = ax_i + b, \quad (i = 1, 2, \ldots, n)$$

により変換した場合の $y_1, y_2, \ldots, y_n$ の平均値 $\bar{y}$ は

$$\bar{y} = a\bar{x} + b \tag{2}$$

により求められる．

例題 ここ 3 日間の最高気温 [C°] は，

`24.5，24.6，24.4`

であった．摂氏 t[C°] から華氏 θ[F°] への変換式は

$$\theta = \frac{9}{5}t + 32$$

で与えられる．3 日間の最高気温の平均値を摂氏 [C°] で求め，その後，華氏での平均値 [F°] を求めなさい．

8　散らばりの指標（分散・標準偏差）

n 個の観測値 $x_1, x_2, \ldots, x_n$ に対する**分散**は

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

で定義される．分散の正の平方根 s を**標準偏差**と呼ぶ．分散は，以下の式によっても計算できる．

$$s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

■データ変換と分散　$x_1, x_2, \ldots, x_n$ を

$$y_i = ax_i + b, \quad (i = 1, 2, \ldots, n)$$

により変換した場合の $y_1, y_2, \ldots, y_n$ の分散 s_y^2 は

$$s_y^2 = a^2 s_x^2$$

により求められる．ただし，s_x^2 は $x_1, x_2, \ldots, x_n$ の分散である．

階級値	x_1	x_2	$\cdots$	x_K
度数	f_1	f_2	$\cdots$	f_K

9　度数分布表と平均・分散

n 個の観測値に対する度数分布表が以下のように与えられたとする．

このとき，平均値の近似値は

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{K} x_k f_k$$

で計算できる．各階級の相対度数を $p_1, p_2, \ldots, p_K$ とすると，

$$\bar{x} = \sum_{i=1}^{K} x_k p_k$$

とも表せる．分散の近似値は，

$$s^2 = \frac{1}{n}\sum_{k=1}^{K}(x_k - \bar{x})^2 f_k$$

もしくは，

$$s^2 = \frac{1}{n}\left(\sum_{k=1}^{K} x_k^2 f_k - \bar{x}^2\right)$$

により計算できる．

10　外れ値と箱ひげ図

第 1，第 2，第 3 四分位数をそれぞれ q_1, q_2, q_3 とする．また，**四分位偏差**を $\mathrm{IQR} = q_3 - q_1$ とする．このとき，区間

$$[q_1 - 1.5\mathrm{IQR}, q_3 + 1.5\mathrm{IQR}]$$

の外側にある観測値のことを，**外れ値**と呼ぶ．

例題 以下のデータに対して四分位数を計算し，外れ値があればそれを特定し，それらを反映させた箱ひげ図を作成しなさい．

`150.0，154.9，156.8，157.5，157.8，157.8，158.4，158.6，158.9，159.4，159.9，160.0，160.6，160.8，160.9，162.1，162.9，163.0，163.1，170.0`

11　散布図

2 つの量的変数の n 組の観測値 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ を，座標平面上の点としてプロットしたものが**散布図**である．図 9 は，表 2 のデータにおける車両重量と総排気量の散布図である．車両重量が重くなるにつれて，総排気量も増える傾向にあることが見てとれる．

12　共分散

散布図により，2 変数間の関連性を直観的に理解することができる．2 変数の関連性の強さの指標のひとつが**共分散**である．共分散は，

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{x})$$

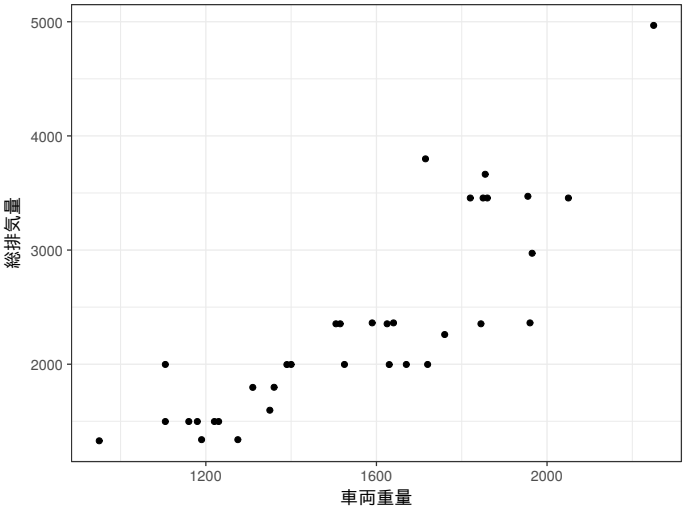

図 9　車両重量と総排気量の散布図

で定義される．散布図を $y = \bar{y}$ と $x = \bar{x}$ の 2 直線で 4 分割し，右上から半時計回りにそれぞれ第 I 象限，第 II 象限，第 III 象限，第 IV 象限とする．第 I 象限に含まれる観測値については，$x_i - \bar{x}$ がいずれも正となるため，それらの積 $(x_i - \bar{x})(y_i - \bar{y})$ も正となる．同様の考え方で，第 III 象限については $(x_i - \bar{x})(y_i - \bar{y}) > 0$，第 II，IV 象限については $(x_i - \bar{x})(y_i - \bar{y}) < 0$ となる．したがって，第 I，第 III 象限に含まれる観測値の数が多い場合，つまり散布図が右上がりの傾向を示す場合，共分散は正の値を示す．逆に，第 II，第 IV 象限に含まれる観測値の数が多い場合，つまり散布図が右下がりの傾向を示す場合，共分散は負の値を示す．第 I，第 III 象限と第 II，第 IV 象限に含まれる観測値の数がおおよそ等しい場合，共分散は 0 に近い値を示す．つまり，共分散は「2 つの変数の直線的な関係の強さを示す指標」であるといえる．

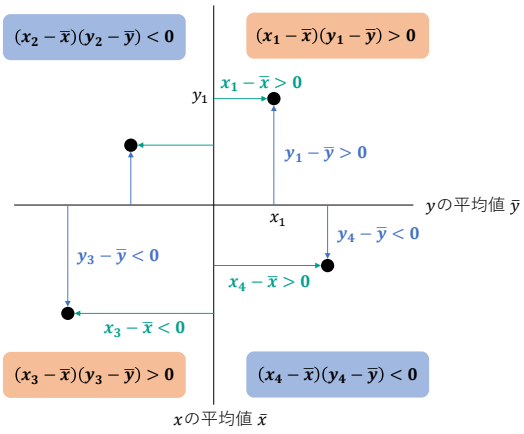

図 10　共分散の概念図 (1)

13　相関係数

共分散は，異なるグループについて同じ 2 変数の直線的な関係の強さを比較する場合などには有効である．例えば，車両重量と総排気量の共分散を乗車定員が 5 人以下のグループと 6 人以上のグループで比較したい場合などである．しかし，グループ間で値のばらつきが異なったり，単位が異なったりする場合，共分散の値のみで直線的な関係の強さを判断することはできない．共分散を基準化して，直線的な関係の強さの指標としたものが**相関係数**である．相関係数は，

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

で定義される．s_x, s_y はそれぞれ x と y の標準偏差である．相関係数は，

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

を満たし，r が 1 に近いほど観測値は散布図上で右上がりの直線に近い傾向を示す．r が-1 に近いほど，右下がりの直線に近い傾向を示す．$r = \pm 1$ の場合，すべての観測値は直線上に並ぶ．r が 0 に近い場合は，共分散 s_{xy} が 0 に近い場合と同様である．

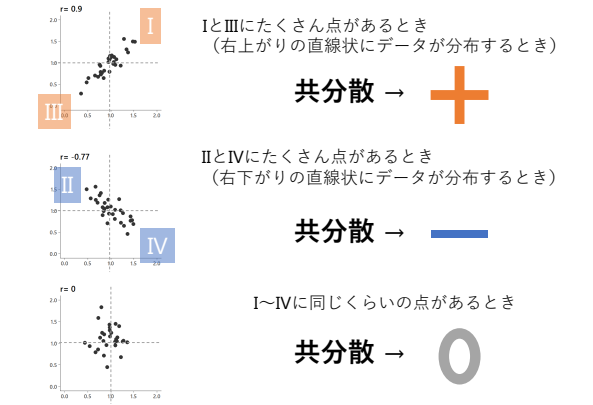

図 11　共分散の概念図 (2)

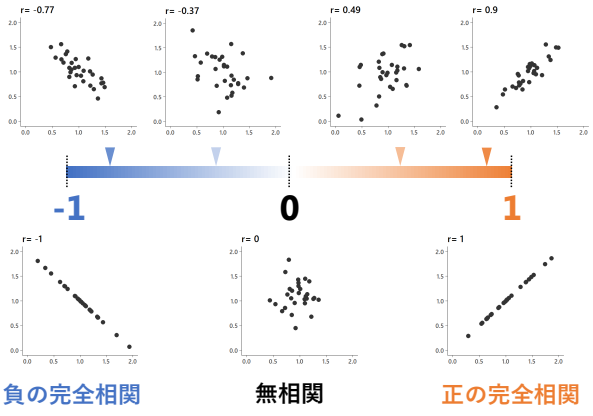

図 12　相関係数の概念図

表 2　燃費性能データ（平成 21 年末までに新型届出のあった普通/小型自動車）

車名	総排気量 [cc]	変速装置	車両重量 [kg]	乗車定員	燃費値 [kg/ℓ]	CO2 排出量 [g-CO2/km]	駆動形式	低排ガス認定レベル
A 社	1797	CVT	1310	5	32.6	71	F	4
A 社	1329	CVT	950	4	20.8	112	F	4
A 社	2362	CVT	1590	5	19.8	117	F	4
A 社	2362	CVT	1960	8	18	129	A	4
A 社	3456	CVT	1850	5	14	166	R	4
A 社	3456	6AT	1820	8	9.7	239	F	4
B 社	2362	CVT	1640	5	19.8	117	F	4
B 社	3456	CVT	2050	5	17.4	133	F	4
B 社	3456	CVT	1860	5	13.2	176	R	4
B 社	4968	CVT	2250	5	11	211	A	4
C 社	1498	CVT	1105	5	18	129	F	4
C 社	1498	CVT	1160	5	18	129	F	4
C 社	1498	CVT	1180	5	18	129	F	4
C 社	1498	CVT	1220	5	17.2	135	F	4
C 社	1597	4AT	1350	7	12.8	181	F	4
C 社	1997	CVT	1630	8	12	193	F	4
C 社	3799	6AT	1715	4	8.4	276	A	3
D 社	1798	CVT	1360	5	14	166	F	4
D 社	1998	CVT	1525	7	12.4	187	F	4
D 社	1998	CVT	1720	8	11.6	200	F	4
D 社	2972	5AT	1965	5	8.7	267	A	4
E 社	1498	CVT	1230	5	15.6	149	F	4
E 社	1998	5AT	1400	5	13.4	173	F	4
E 社	1998	5MT	1105	2	12.6	184	R	4
E 社	1998	5AT	1670	8	12.4	187	F	4
E 社	2260	6AT	1760	5	8.8	264	A	3
F 社	1339	CVT	1190	5	26	89	F	4
F 社	1339	CVT	1275	5	25.8	90	F	4
F 社	1997	CVT	1390	7	14.4	161	F	4
F 社	2354	CVT	1625	7	12.4	187	F	4
F 社	2354	5AT	1505	5	11.6	200	F	4
F 社	2354	5AT	1515	5	11.2	207	F	4
F 社	2354	5AT	1845	8	9.9	235	F	4
F 社	3664	5AT	1855	5	8.6	270	A	4
F 社	3471	5AT	1955	8	8.2	283	F	4