

第 II 部 統計

4 集計と比較の基準

ポイント

- 量と順位と分類
- 一次元データと多次元データ

- 散布図と分割表
- 比較の規準

4.1 はじめに

- データはいろいろな情報を含みます。
- 分類し比較し新たな事柄を評価することができます。
- 評価基準は様々です。

4.2 データの種類

4.2.1 ^{きすう}基数データ

- 数の概念の内、量的側面をあらわす値。
- 物理量または数え上げ等によって測定される値。
- 単位が定義され、量によって測られる値。
- 値の大小に意味がある。
- 差、比に意味を持つことが多い。
- 大体の代表値に意味がある。

4.2.2 序数データ^{じょすう}

- 数の概念のうち、順序的側面をあらわす値。
- 何らかの基準により順序付けされた、順位データ。
- 上位・下位の概念はある。
- 上位・下位との差は不明。
- 連続しているかどうか不明。
- 代表値は最頻値のみ意味がある。

4.2.3 カテゴリカル・データ

- 対象を属性に基き、いくつかに分類するときに、どの分類先に属するかを表している値。
- 同質か異質かによって識別されるデータ。
- 分類先をコード化する為に数字が採用されることがある。
- この数字には記号としての意味しかなく、同じか異なるかだけで判断される。
- 数字を利用している場合でも大小といった概念は持っていない。
- 代表値は最頻値のみ意味を持つ。

4.2.4 一次元データと多次元データ

- i 番目の観測値を単一の変数 (x_i) とするデータを一次元データという。
- i 番目の観測値に対して m 個の属性 ($x_{i1}, x_{i2}, x_{i3}, \dots, x_{im}$) を一組として観測するデータを多次元データという。
- 多次元データの統計では一次元データの分析を複数回繰り返すのではない。
- 多次元データの統計とは多くの変数を一括して、その関係を取り扱うこと意味する。

4.3 相関

- ここで i をデータの組を表す添え字、 j を変数の属性を識別する添え字とし、変数の表記を x_{ij} とする。
- 2つの属性 (x_{i1}, x_{i2}) の関係において、 x_{i1} と x_{i2} の関係に因果の方向を考えず x_{i1} と x_{i2} の関係を対等に見る方法を相関という。
- 属性が m 個ある場合、2属性ずつの関係は $k = \frac{m(m-1)}{2}$ ある。

4.3.1 散布図

- i 番目の 2 属性からなる観測値の組を (x_{i1}, x_{i2}) とする。
- ここで、 x_{i1}, x_{i2} がともに基数データであった場合、 x_{ij} の一方を横軸にとり、もう一方を縦軸にとって観測値 (x_{i1}, x_{i2}) を平面上に散布した図を『散布図』とよぶ。
- 散布図により属性間の関係を視覚的に把握することができる。
- 統計学では直線に近い傾向がみられる場合『相関関係が有る』という。
- 右上がりの傾向がある場合『正の相関』があるといい右下がりの傾向がある場合『負の相関』があるという。

問題 II-4-1

表の値を使い、散布図を描画しなさい。

表1 小学生の身長と体重

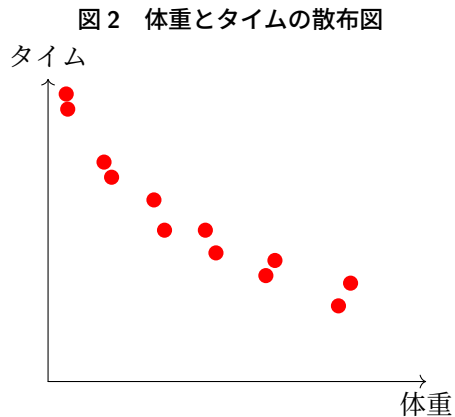
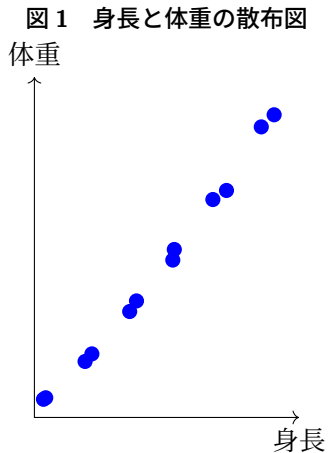
i	身長	体重	i	身長	体重
1	116.5	21.3	7	116.2	21.2
2	122.6	24.2	8	121.7	23.7
3	128.5	27.7	9	127.6	27.0
4	133.5	31.1	10	133.3	30.4
5	138.6	34.4	11	140.4	35.0
6	145.0	39.2	12	146.7	40.0

問題 II-4-2

表の値を使い、散布図を描画しなさい。

表2 小学生の体重と 50m 走タイム

i	体重	タイム	i	体重	タイム
1	21.3	11.6	7	21.2	11.8
2	24.2	10.7	8	23.7	10.9
3	27.7	10.0	9	27.0	10.4
4	31.1	9.7	10	30.4	10.0
5	34.4	9.4	11	35.0	9.6
6	39.2	9.0	12	40.0	9.3



4.3.2 分割表

- i 番目の 2 属性からなる観測値の組を (x_{i1}, x_{i2}) とする。
- ここで、 x_{ij} はカテゴリカル・データであり、 x_{ij} は属性 (j) において一つの区分にのみ属するとする。
- 一方の属性の区分を縦軸（表側）に、もう一方の属性の区分を横軸（表頭）にとり、それぞれの属性の組合わせに該当する度数を表頭・表側の項目の交差するマス目に記載したものを『分割表』とよぶ。
- 縦軸・横軸ともに最後または最初の欄に合計を記載する。
- 『相関表』『クロス表』ともよぶ。
- 基数データを階級分けすることで分割表を作ることができる。
- 序数データを利用しても分割表を作ることができる。

4.4 比較の基準

$$\text{構成比} = \frac{\text{部分}}{\text{全体}} \quad (1)$$

$$\text{変化量} = \text{変化後} - \text{変化前} \quad (2)$$

$$\text{倍率} = \frac{\text{変化後}}{\text{変化前}} \quad (3)$$

$$\text{変化率} = \frac{\text{変化量}}{\text{変化前}} = \frac{\text{変化後} - \text{変化前}}{\text{変化前}} = \frac{\text{変化後}}{\text{変化前}} - 1 = \text{倍率} - 1 \quad (4)$$

$$\text{変化の割合} = \frac{y \text{ の変化量}}{x \text{ の変化量}} \quad (5)$$

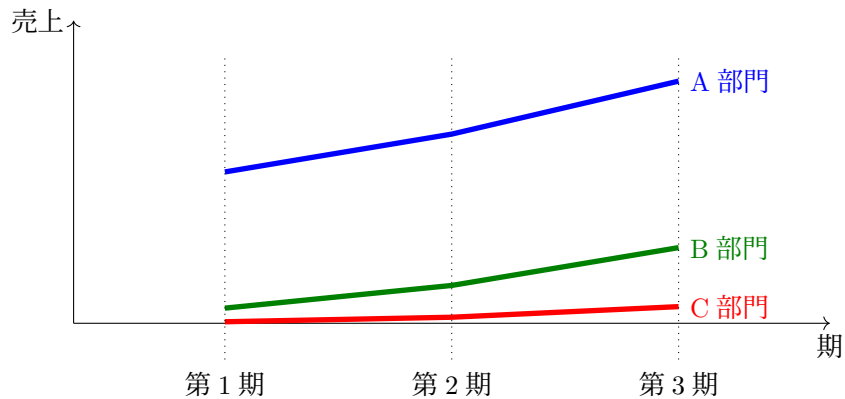
問題 II-4-3

表 3 を使い、各期の構成比, 対前期の構成比の推移, 変化量, 倍率, 変化率を求め、1 期当たりの適切な代表値を求めなさい。ただし、各期の期間は同一とする。

表 3 3つの部門の売上

部門	第 1 期	第 2 期	第 3 期	全期計
A 部門	100	125	160	385
B 部門	10	25	50	85
C 部門	1	4	11	16
全体	111	154	221	486

図3 売上の推移



4.4.1 構成比

表 4 期ごとの構成比

部門	第 1 期	第 2 期	第 3 期	代表値
A 部門	90.1 %	81.2 %	72.4 %	79.2 %
B 部門	9.0 %	16.2 %	22.6 %	17.5 %
C 部門	0.9 %	2.6 %	5.0 %	3.3 %
全体	100.0 %	100.0 %	100.0 %	100.0 %

※ 代表値は、全期計の構成比として算出

	第 1 期	第 2 期	...	第 n 期	全期計
A 部門	x_{11}	x_{21}	$\cdots$	x_{n1}	$\sum_{i=1}^n x_{i1}$
B 部門	x_{12}	x_{22}	$\cdots$	x_{n2}	$\sum_{i=1}^n x_{i2}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
第 m 部門	x_{1m}	x_{2m}	$\cdots$	x_{nm}	$\sum_{i=1}^n x_{im}$
全体	$\sum_{j=1}^m x_{1j}$	$\sum_{j=1}^m x_{2j}$	$\cdots$	$\sum_{j=1}^m x_{nj}$	$\sum_{i=1}^n \sum_{j=1}^m x_{ij}$

A 部門の各期の構成比は

$$\frac{x_{11}}{\sum_{j=1}^m x_{1j}}, \quad \frac{x_{21}}{\sum_{j=1}^m x_{2j}}, \quad \frac{x_{n1}}{\sum_{j=1}^m x_{nj}} \quad (6)$$

それぞれの分母の比率は

$$\frac{\sum_{j=1}^m x_{1j}}{\sum_{i=1}^n \sum_{j=1}^m x_{ij}}, \quad \frac{\sum_{j=1}^m x_{2j}}{\sum_{i=1}^n \sum_{j=1}^m x_{ij}}, \quad \frac{\sum_{j=1}^m x_{nj}}{\sum_{i=1}^n \sum_{j=1}^m x_{ij}} \quad (7)$$

(7) を用いて (6) を加重平均すると

$$\frac{\cancel{\sum_{j=1}^m x_{1j}}}{\cancel{\sum_{j=1}^m x_{1j}}} \frac{\sum_{j=1}^m x_{1j}}{\sum_{i=1}^n \sum_{j=1}^m x_{ij}} + \frac{\cancel{\sum_{j=1}^m x_{2j}}}{\cancel{\sum_{j=1}^m x_{2j}}} \frac{\sum_{j=1}^m x_{2j}}{\sum_{i=1}^n \sum_{j=1}^m x_{ij}} + \frac{\cancel{\sum_{j=1}^m x_{nj}}}{\cancel{\sum_{j=1}^m x_{nj}}} \frac{\sum_{j=1}^m x_{nj}}{\sum_{i=1}^n \sum_{j=1}^m x_{ij}} \quad (8)$$

$$= \frac{x_{11} + x_{21} + x_{n1}}{\sum_{i=1}^n \sum_{j=1}^m x_{ij}} = \frac{\sum_{i=1}^n x_{i1}}{\sum_{i=1}^n \sum_{j=1}^m x_{ij}} \quad (9)$$

この値は、全期計の構成比に一致する。

4.4.2 構成比の推移

表 5 前期からの構成比の推移

	1 期から 2 期	2 期から 3 期	代表値
A 部門	-8.9 pt	-8.8 pt	-8.8 pt
B 部門	7.2 pt	6.4 pt	6.8 pt
C 部門	1.7 pt	2.4 pt	2.0 pt

※ 代表値は算術平均で算出。

期をあらわす添え字を i 、部門をあらわす添え字を j とし、各部門の各期の構成比を x_{ij} とあらわす。すると部門 j の 1 期から 2 期への構成比の変化ポイントは

$$x_{2j} - x_{1j} \quad (10)$$

である。

2 期から 3 期への変化ポイントは

$$x_{3j} - x_{2j} \quad (11)$$

これを合わせると

$$(x_{3j} - x_{2j}) + (x_{2j} - x_{1j}) \quad (12)$$

$$= x_{3j} - x_{1j} \quad (13)$$

である。各期に同じだけ変化したと考え、2 で割り

$$\frac{x_{3j} - x_{1j}}{2} \quad (14)$$

4.4.3 変化量

表 6 前期からの変化量

	1 期から 2 期	2 期から 3 期	代表値
A 部門	25	35	30
B 部門	15	25	20
C 部門	3	7	5
全体	43	67	55

※ 代表値は算術平均で算出

期を表す添え字を i とし、各期の値を x_i すると 1 期から 2 期への変化量は

$$x_2 - x_1 \quad (15)$$

2 期から 3 期への変化量は

$$x_3 - x_2 \quad (16)$$

これらを足し合わせると

$$(x_3 - x_2) + (x_2 - x_1) \quad (17)$$

$$= x_3 - x_1 \quad (18)$$

2 期分の変化量なので 2 で割り 1 期当たりを求めると

$$\frac{x_3 - x_1}{2} \quad (19)$$

4.4.4 倍率

表 7 前期からの倍率

	1 期から 2 期	2 期から 3 期	代表値
A 部門	1.25 倍	1.28 倍	1.26 倍
B 部門	2.50 倍	2.00 倍	2.24 倍
C 部門	4.00 倍	2.75 倍	3.32 倍
全体	1.39 倍	1.44 倍	1.41 倍

※ 代表値は幾何平均で算出

期を表す添え字を i とし、各期の値を x_i すると 1 期から 2 期への倍率は

$$\frac{x_2}{x_1} \quad (20)$$

2 期から 3 期への倍率は

$$\frac{x_3}{x_2} \quad (21)$$

これらを掛け合わせると

$$\frac{x_2}{x_1} \frac{x_3}{x_2} = \frac{x_3}{x_1} \quad (22)$$

2 期分の倍率なので平方根をとると

$$\sqrt{\frac{x_2}{x_1} \frac{x_3}{x_2}} = \sqrt{\frac{x_3}{x_1}} \quad (23)$$

4.4.5 変化率

表 8 変化率

	1 期から 2 期	2 期から 3 期	代表値
A 部門	25%	28%	26%
B 部門	150%	100%	124%
C 部門	300%	175%	232%
全体	39%	44%	41%

※ 倍率から 1 を引くことで変化率を算出

4.5 カテゴリカル・データの集計

4.5.1 ダミー変数

- 0 または 1 しかとらない変数。
- 条件を満たしたときに 1、条件を満たしていない場合 0。
- カテゴリカル・データを扱うときに利用。

4.5.2 性別を識別するためのダミー変数の例

観測値が n 個あり i 番目の値を x_i と表す。ここで x_i を男性ならば 1 を取るダミー変数とする。

すると男性である観測値の数は

$$\sum_{i=1}^n x_i \quad (24)$$

であり、男性ではない観測値の数は

$$n - \sum_{i=1}^n x_i \quad (25)$$

表9 ダミー変数の度数分布と相対度数

ダミー変数	度数	相対度数
1	$\sum_{i=1}^n x_i$	$\frac{1}{n} \sum_{i=1}^n x_i$
0	$n - \sum_{i=1}^n x_i$	$1 - \frac{1}{n} \sum_{i=1}^n x_i$
合計	n	1

4.5.3 ダミー変数の相対度数の意味

- 観測値全体 (n) に対する男性の数
- 男性の比率
- 男性ダミーの出現率

問題 II-4-4

二つの工場 (F_1, F_2) がありそれぞれ同じ製品を二種類 (P_1, P_2) 製造している。表の値を基に不良品率を評価基準として二つの工場を比較せよ。

表 10 2つの工場における2種類の製品の製造数と不良品数

	P_1		P_2		合計	
	製造数	不良品数	製造数	不良品数	製造数	不良品数
F_1	1,000	650	500	225	1,500	875
F_2	500	350	1,000	500	1,500	850
合計	1,500	1,000	1,500	725	3,000	1,725

解例 II-4-4

表 11 2つの工場で作成した2種類の製品の不良品率

	P_1	P_2	全体
F_1	0.650	0.450	0.583
F_2	0.700	0.500	0.567
全体	0.667	0.483	0.575

問題 II-4-5

あるゲーム大会への参加者数と合計点数が表 12 であった。1 人当たりの平均点を規準として A チームと B チームを比較しなさい。

表 12 参加人数と合計得点

	A チーム		B チーム		合計	
	合計得点	人数	合計得点	人数	合計得点	人数
男性	6,500	100	3,500	50	10,000	150
女性	2,250	50	5,000	100	7,250	150
合計	8,750	150	8,500	150	17,250	300

解例 II-4-5

表 13 1人当たりの平均点

	A チーム	B チーム	全体
男性	65.00	70.00	66.67
女性	45.00	50.00	48.33
全体	58.33	56.67	57.50

4.6 まとめ

- 変数間の関係を把握するのが多次元データの統計。
- 比較の基準によって結果は異なる。
- 『適切な』 基準を選ばなくてはならない。