古谷数学教室第 18 回

データの分析

2024 年 8 月 25 日

1　基礎事項

1.1　データの整理

人の身長、体重や運動の記録などのように、ある特性を表す数量を**変量**という。数学では、ある変量の測定値や観測値の集まりを**データ**という。

表 1 は、妻からの仕事終わりの連絡が届いた時刻[1)] を日付ごとにまとめた表である。 表 1 の

表 1　妻からの仕事終わりの連絡が届いた時刻と日付

9/9	9/8	9/6	9/5	9/4	9/1	8/31	8/30	8/29	8/28
16:31	19:03	18:25	18:50	16:49	17:24	17:43	18:44	17:34	17:26

データは、表 2 のように**度数分布表**と呼ばれる表で整理できる。度数分布表において、区切られた各区間を**階級**、区間の幅を**階級の幅**、各階級に含まれる値の個数を**度数**という。また、各階級の中央の値を **階級値** という。

表 2　表 1 の度数分布表

階級（時刻）	度数
16:30 〜 17:00	2
17:00 〜 17:30	2
17:30 〜 18:00	2
18:00 〜 18:30	1
18:30 〜 19:00	2
19:00 〜 19:30	1
計	10

1)　時刻をデータとして用いるのは少し注意が必要である。例えば、不等式で評価しようとすると注意すべきことが分かる。

図 1 は、表 2 を**ヒストグラム**で表したものである。

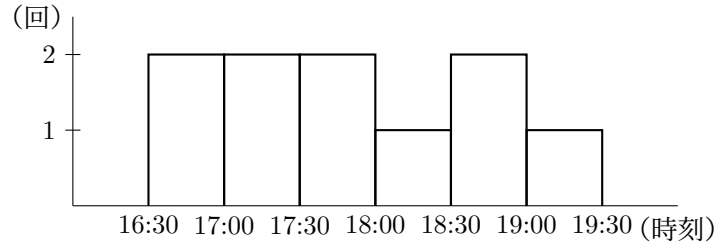


図 1　表 2 のヒストグラム

1.2　データの代表値

データ全体の特徴を適当な 1 つの数値で表せると便利である。そのような値をデータの**代表値**という。ここでは、データの代表値として、平均値、最頻値、中央値を定義する。

データにおける測定値や観測値の個数を、そのデータの**大きさ**という。

変量 x についてのあるデータの大きさが n であるとき、このデータの個々の値を

$$x_1,\ x_2,\ x_3,\ \cdots,\ x_n$$

で表す[2] ことがある。

一般に、変量 x について、大きさ n のデータの値の総和を n で割った値を、このデータの**平均値**といい、$\bar{x}$ で表す：

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

データにおいて、最も個数の多い値を、そのデータの**最頻値**または**モード**という。度数分布表に整理したときは、度数が最も大きい階級の階級値を最頻値とする。

データを値の大きさの順に並べたとき、中央の位置にくる値を**中央値**または**メジアン**という。データの大きさが偶数のときは、中央に 2 つ値が並ぶが、その場合は 2 つの値の平均をとって中央値とする。

1.3　データの散らばりと四分位数

データの散らばりの度合いを表す値として、データの最大値から最小値を引いた差が考えられる。この差をデータの**範囲**という。しかし、データの中に極端に飛び離れた値があると、データの範囲が散らばりの度合いを表すとは考えにくい。そこで、データを値の大きさの順に並べたとき、4 等分する位置の値を考える。これを**四分位数**（しぶんいすう）という。四分位数は、小さい方から順に、**第 1 四分位数**、**第 2 四分位数**、**第 3 四分位数**という。第 2 四分位数はデータの中央値にほかならない。四分位数にはさまざまな定義があるが、高校数学では、次のルールによって四分位数が定義されている。

2)　データの値を並べたとき、k 番目の値を x_k と表している。

四分位数

1. データを値の大きさの順に並べ、中央値を求める。これを、第 2 四分位数とする。

2. 中央値を境としてデータの個数を 2 等分し、値が中央値以下の組（組 a とする）と値が中央値以上の組（組 b とする）に分ける。ただし、データの大きさが奇数だった場合、中央値は組 a と組 b の両方に含めない。

3. 組 a の中央値を第 1 四分位数、組 b の中央値を第 3 四分位数とする。

第 3 四分位数と第 1 四分位数の差（の絶対値）を**四分位範囲**という。また、四分位範囲の半分を、**四分位偏差**という[3)]。

データの分布は、**箱ひげ図**と呼ばれる図で表すことがある。箱ひげ図は次のようなルールで描かれる。

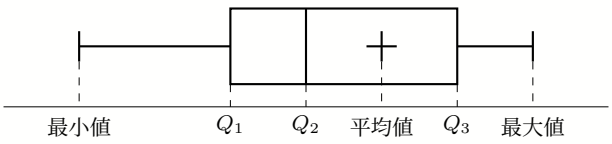
箱ひげ図

1. 横軸にデータの値の目盛りをとるとする。

2. 第 1 四分位数（Q_1 とする）を左端、第 3 四分位数（Q_3 とする）を右端とする箱を描き、箱の中に中央値（Q_2 とする）を示す縦線を描く。

3. 箱の左端から最小値まで、箱の右端から最大値まで、線分を引く。

次の箱ひげ図平均値を記しているが、省略されることもある。



1.4 分散と標準偏差

四分位数は、データの中のいくつかの代表的な値を用いて散らばりの度合いを表す値であった。ここでは、データの値をすべて使って散らばりの度合いを表す値について考える。

変量 x のデータの値 $x_1,\ x_2,\ \cdots,\ x_n$ の平均値を $\bar{x}$ とするとき、各値と平均値との差 $x_1-\bar{x},\ x_2-\bar{x},\ \cdots,\ x_n-\bar{x}$ を、それぞれ平均値からの**偏差**といい、$x-\bar{x}$ で表す。偏差の総和は 0

3) なぜ四分位偏差という量をわざわざ定義したのか私には分からない。

になるから、偏差の平均値も 0 である。よって、偏差の平均値では、散らばりの度合いを表すことはできない。そこで、偏差をそのまま用いずに、偏差の 2 乗の平均値を考える。

データにおいて、偏差の 2 乗の平均値**分散**という。さらに、分散の正の平方根**標準偏差**といい、s で表す：

$$s^2 = \sum_{k=1}^{n} \frac{(x_k - \bar{x})}{n},$$
$$s = \sqrt{s^2}.$$

データの値が平均値の周りに集中しているほど、それぞれの偏差の絶対値は小さくなり、分散、標準偏差も小さくなる傾向にある。

1.5　データの相関

表 3 は、よめとだんなの小学校 5 年生用の漢字テストと算数テストの得点をまとめたものである。このように 2 つの変量からなるデータの間に、一方が増加すればそれに従って他方も増加する、または他方が減少するという傾向がみられるとき、2 つの変量の間に**相関**がある、または**相関関係**があるという。

2 つの変量からなるデータにおいて、一方が増加すると、他方も増加する傾向がみられるとき、2 つの変量には**正の相関**があるという。また、一方が増加すると、他方が減少する傾向がみられるとき、2 つの変量には**負の相関**があるという。どちらの傾向もみられないとき、**相関がない**または**相関関係がない**という。

表 3　よめとだんなの小学校 5 年生用の漢字テストと算数テストの得点

	漢字テスト x	算数テスト y
だんな	5	20
よめ	11	0

表 3 を平面上に図示した図 2 のような図を**散布図**という。散布図は、データにおける 2 つの変量の間の関連性を視覚的にとらえるのに役立つ。

2 つの変量 x、y からなるデータとして n 個の値の組

$$(x_1,\ y_1),\quad (x_2,\ y_2),\quad \cdots,\quad (x_n,\ y_n)$$

が得られているとする。以下では、変量 $x_1,\ x_2,\ \cdots,\ x_n$、$y_1,\ y_2,\ \cdots,\ y_n$ の平均値をそれぞれ $\bar{x}$、$\bar{y}$ とし、標準偏差をそれぞれ s_x、s_y とする。x の偏差と y の偏差の積の平均値である、

$$s_{xy} = \frac{1}{n}\big\{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})\big\}$$

を**共分散**という。共分散と 2 つの標準偏差を用いて、次のように表せる r を**相関係数**という。

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

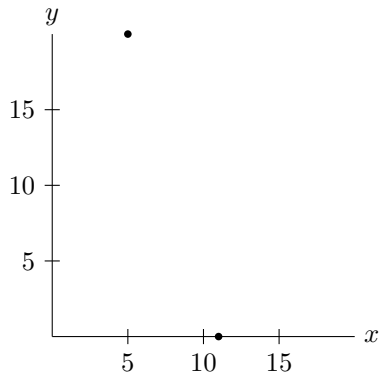


図 2　よめとだんなの小学校 5 年生用の漢字テストと算数テストの得点の散布図

相関係数について、$-1 \leq r \leq 1$ であることが知られている。相関係数が 1 に近づけば近づくほど、正の相関があると評価できる。また、相関係数が -1 に近づけば近づくほど、負の相関があると評価できる。また、相関係数が 0 に近づけば近づくほど、相関がないと評価できる。

1.6　仮説検定の考え方

仮説検定の授業は第 20 回で詳しく扱う[4] が、ここでは簡単に問題を解くのに最低限必要な説明[5] をする。

仮説検定の考え方を利用して、主張 A が正しいと判断できるかどうか調べる。

1. 主張 A と反する仮定を立てる。これを主張 B とする。

2. 主張 B のもので、実際に起こった出来事が起こりにくい出来事がどうか調べる。

3. **2.** で調べた結果、実際に起こった出来事は起こりにくいと判断するとき[6] 、主張 B の仮定は正しくないと判断できる。

4. 主張 A は正しいと判断してもよいと考えられる。

3. の「実際に起こった出来事が起こりにくい出来事かどうか」を調べるためには、ある基準が必要だが、「基準となる確率」を与えられ、さらに公正なコインなどを用いて、実際に統計データを取

4)　仮説検定は数学 B の「統計的な推測」で学習する。しかし、文部科学省の指導要領解説によれば、数学 I のみを学習する学生さん向けに、直感的に「仮説検定」を学習できるような分野「仮説検定の考え方」を作った。~~極めて迷惑である。~~

5)　この立場はあまり好みじゃない。数学はやはり「頭を使って考える」ことに意味がある。「この問題はこのようにしたら解ける」を教えるのは、それを終えたあとである。

6)　実際に起こった出来事が十分起こりにくいと判断しないときは、主張 B の仮定は否定できず、主張 A は正しいと判断できない。このとき、主張 B が正しいと判断できるわけではないことに注意する。

得したものが与えられる。その実際の統計データの相対度数[7]を計算し、「基準となる確率」と比較することにより、**3.** が可能となる。

2 演習問題

1. 次のデータは、H 市のある月の日ごとの最低気温（単位は℃）である：

7.1	10.7	8.9	7.5	11.0	12.6	17.0	18.6	16.5	13.9
10.1	12.6	14.1	17.6	14.0	11.7	16.9	16.3	13.7	13.5
12.2	13.3	11.4	12.5	12.2	4.9	5.0	8.6	5.6	4.4.

(1) 階級の幅を 2℃ として、度数分布表を作れ。ただし、階級は 4℃ から区切り始めるものとする。

(2) (1) で作った度数分布表からヒストグラムを作れ。

2. 次のデータは、ある高校生 20 人の小テストの得点（単位は点）である：

3	4	9	7	6	10	5	5	7	9
6	8	1	5	7	10	8	6	3	7.

(1) 平均点を求めよ。

(2) 中央値を求めよ。

(3) 最頻値を求めよ。

3. 次のデータ A、B の範囲を求め、データの散らばりの度合いを比較せよ[8]：

A:	9	12	10	11	8	13	7	12,
B:	9	15	6	12	21	12	18	12.

4. 次のデータの第 1 四分位数、第 2 四分位数、第 3 四分位数を求めよ：

12　35　47　59　68　73　74　79　87　97.

5. 次のデータは、ある商店における A 弁当と B 弁当の 10 日間の販売数（単位は個）である：

---

7) 例えば、公正なコインなら、表が出る確率は 1/2 である。しかし、これを 30 回投げたからといって必ず 15 回表になるわけではない。なので、公正なコインを 30 回投げて表の出た回数を記録する実験を、たとえば 200 セット行って統計をとる。このとき、200 セットのうち、表の回数がでた回数で度数分布表を作る。200 セットのうち、表の回数が 15 回のセットの数が最も多いと期待できる。（もちろん、必ずそうとは限らない。）

例えば 30 回コインを投げたとき、15 回表が出る（34 セットだったとする）ときの相対度数とは、34/200 である。

8) 散らばりの度合いが大きいのはどちらと考えられるか答えよ。という問いだと思えばよい。

A 弁当: 22 28 16 25 33 27 17 21 23 40,
B 弁当: 18 24 40 20 17 15 28 35 32 16.

6. 次のデータは、5 人の生徒の通学にかかる時間 x（分）である：

25 15 35 20 30.

(1) このデータの平均値 $\bar{x}$ を求めよ。

(2) 分散 s^2 を求めよ。

(3) 標準偏差 s を求めよ。ただし、少数第 2 位を四捨五入せよ。

7. 次のような 2 つの変量 x、y についてのデータがある。

x	78	63	86	54	92	57	95	69	81	73
y	32	29	4	48	2	37	13	41	26	15

これらについて、散布図をかき、x と y の間に相関関係があるかどうかを調べよ。また、相関関係がある場合には、正、負のどちらであるか答えよ。

8. ある市の市長選挙に X、Y の 2 人が立候補した。有権者の中から無作為に 30 人を選んで X、Y のどちらを支援しているかを調査したところ 21 人が X を支持していることがわかった。この調査から、X の方が支持者が多いと判断してよいか。仮説検定の考え方を用い、基準となる確率を 0.05 として考察せよ。ただし、公正なコインを 30 回投げて表の出た回数記録する実験を 200 セット行ったところ、次の表のようになったとし、この結果を用いよ。

表の回数	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	計
度数	1	2	2	12	20	23	24	34	25	18	17	9	7	4	1	1	200