

数学 I 第 5 章 データの分析 No. 1

学習のねらい

平均値、中央値、最頻値について理解しよう！

1. 平均値と中央値

例) 5 人のテストの点数が 42、38、40、44、96(単位は点)としたとき、平均値を求めよ。

平均値は、 $\frac{1}{5}(42 + 38 + 40 + 44 + 96) = 52$ となり、52 点

変量 x についてのデータが n 個の値 $x_1, x_2, \dots, x_n$ であるとき、それらの総和を n で割ったものを、データの 平均値 といい、 $\bar{x}$ と表す。

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$$

また、先の例を小さい順に並べると、38、40、42、44、96 となる。よって、中央値は 42 点である。

データを大きさの順に並べたとき、中央の位置にくる値をデータの 中央値 または メジアン という。

先の例でいうと、平均が 52 点で中央値が 42 点である。つまり 44 点の生徒は真ん中よりは上、平均よりは下ということ。つまり「自分は上位だ！」と言っていいわけだ。(いやそれでも点数低い気はするけど。)

2. 仮平均

先の例で、 $(42 + 38 + 40 + 44 + 96)$ の計算がダルい。そこで、「だいたいこのくらいが平均になりそう(仮平均)」→「そこからのズレを計算する」という手法がある。具体的に計算しよう。

仮平均が 50 点だとしよう(何点でも良い)。すると仮平均との差は、

実際の点数	42	38	40	44	96	合計
仮平均との差	-8	-12	-10	-6	46	10

これの平均は、 $10 \div 5 = 2$ となるので、実際の平均は仮平均の 50 に 2 を加えた 52 点となる。

3. 最頻値

漢字のままの意味で、「最も頻繁に出てくる値」という意味で、データの中で最も個数の多い値を 最頻値 または モード という。

例) 次の表はテストの点数を分類した表である。

点数(点)	4	5	6	7	8	9
人数(人)	1	6	4	2	3	2

このとき、最頻値は 5 点である。

データって、ダルいなあって思うよね。

でも AI 時代を生きる今、

大切なんだよね・・・。



◇問題

1. 次のデータは、A 班 5 人、B 班 6 人の 10 点満点の小テストの結果である。

A 班：5、7、8、4、9 B 班：7、10、9、4、8、6 （単位は点）

(1) A 班のデータの平均値と B 班のデータの平均値をそれぞれ求めよ。ただし小数第 2 位を四捨五入せよ。

(2) A 班と B 班を合わせた 11 人のデータの平均値を求めよ。

(3) A 班のデータの中央値と B 班のデータの中央値をそれぞれ求めよ。

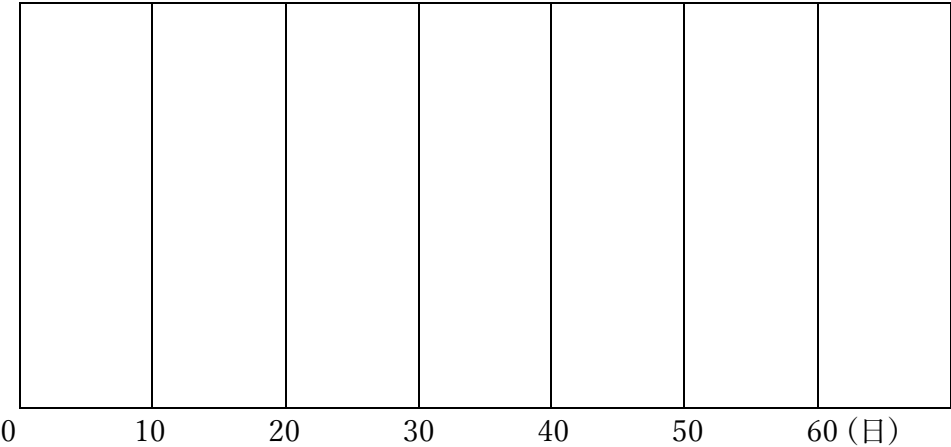
◇問題

1. 次のデータは、2つの都市 A、B について、最低気温が 25℃以上であつた日の日数を 1 年ごとに集計した 12 年分の結果である。(単位は日)。

A	19	31	46	41	25	33	17	41	31	20	31	25
B	25	46	37	57	48	48	30	51	46	44	44	42

このデータから、下の表を埋め、箱ひげ図を完成させよ。

	最小値	Q_1	中央値	Q_3	最大値
A					
B					



数学 I 第 5 章 データの分析 No.3

学習のねらい

分散と標準偏差を理解して公式を覚えよう！

1. 分散

例えば A 組と B 組、どちらのクラスの点数の散らばりが大きい？

出席番号	1	2	3	4	5
A 組	48	49	50	51	52
B 組	20	40	50	60	80

いや、そりゃ B 組でしょ！！！！となると思う。しかしどちらも平均は 50 点である。つまり平均だけだと、散らばり具合が分からない。これだとデータをうまく活用できない。

そこで、平均からの離れ具合(つまり、「(データ)-(平均)」の足し算)を見てみようという発想になったわけだ。このような $x - \bar{x}$ のことを **偏差** という。

しかし、これを計算すると、それぞれのデータから 50 を引いて、

A 組	-2	-1	0	1	2
-----	----	----	---	---	---

となる。これを足すと $-2 - 1 + 0 + 1 + 2 = 0$ となる。あちゃあ・・・。
ということでボツ！

そこで、全部 2 乗してしまえ！！という発想になった。

A 組	4	1	0	1	4
-----	---	---	---	---	---

これを足すと、 $4 + 1 + 0 + 1 + 4 = 10$ となる。しかしこのままだとデ

ータの数が多ければ多いほど、値が大きくなってしまう。そこで、「個数で割ろう！！」ということにしたのだ。

よって、A 組の分散は $10 \div 5 = 2$ で 2 となる。ちなみに同様に B 組の分散を求めると、400 となる。確かに B 組の方が散らばっているよね、と分かるわけだ。偏差の 2 乗の平均値の値を **分散** といい、 s^2 で表す。

$$s^2 = \frac{1}{n} \{ (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \}$$

2. 標準偏差

しかし、分散にも問題がある。「B 組の分散は 400 である！！！」と言ったけど、単位はなんだろう。点数を 2 乗しているから、イメージで言うと、(点数)² となるわけだ。

「えーと、その・・・。(点数)² ってなんすか??」
となるよね。その通り。そう思います。そこで、「分散の値に $\sqrt{\quad}$ を付けちゃえば良いじゃん！！」ってなったわけだ。この値のことを **標準偏差** という。

先の例なら、A 組の標準偏差は $\sqrt{2} = 1.41 \cdots$ 点、B 組の標準偏差は $\sqrt{400} = 20$ 点となる。実際に表を見てみると A 組、B 組の散らばり具合は標準偏差の点数くらいになっている。(平均ジャストの点数の人(3 番の人)がいるので、ちょっと散らばり具合がギュツとなっている感じ。)

公式としてまとめておこう。標準偏差は s で表す。

$$s = \sqrt{\frac{1}{n} \{ (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \}}$$

◇問題

1. 次のデータは、ある商品 A、B の 5 日間の売り上げ個数である。

A : 5、7、4、3、6 B : 4、6、8、3、9 (単位は個)

A、B の変量をそれぞれ x 、 y とするとき、次の問いに答えよ。

- (1) x 、 y のデータの平均値、分散、標準偏差を求めよ。ただし標準偏差については小数第 2 位を四捨五入せよ。
- (2) x 、 y のデータについて、標準偏差によってどちらの方が散らばり度合いが大きいか判定せよ。

数学 I 第 5 章 データの分析 No.4

学習のねらい

分散と平均の関係について理解しよう！

変数の変換について理解しよう！

1. 分散と平均の関係

分散 s^2 を変形してみよう。

$$\begin{aligned}s^2 &= \frac{1}{n}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\} \\&= \frac{1}{n}\{x_1^2 + x_2^2 + \cdots + x_n^2 - 2\bar{x}(x_1 + x_2 + \cdots + x_n) + n(\bar{x})^2\} \\&= \frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2) - 2\bar{x} \cdot \frac{1}{n}(x_1 + x_2 + \cdots + x_n) + (\bar{x})^2 \\&= \overline{x^2} - 2\bar{x} \cdot \bar{x} + (\bar{x})^2 \\&= \overline{x^2} - (\bar{x})^2 \quad (\overline{x^2} \text{とは、} x_1^2, x_2^2, \dots, x_n^2 \text{の平均のことである。})\end{aligned}$$

よって、以下の公式が得られる。

$$s^2 = \overline{x^2} - (\bar{x})^2$$

実際の計算では、こちらの公式を使うと楽になることが多い。

2. 変数の変換

テストとかで、全員の点数を0.8倍するとか、全員の点数に+10点するっていう場面に出会ったことはないかな？

ここでは、データの各値に一齐に同じ数を加えたり、一齐に同じ数を掛けたりする時、平均値、分散、標準偏差がどのようなになるかを考えよう。

変数 x についてのデータが n 個の値 $x_1, x_2, \dots, x_n$ であるとして、 x のデータの平均値を $\bar{x}$ 、分散を s_x^2 、標準偏差を s_x とする。そして、 a, b を定数として、式 $y = ax + b$ で新しい変数 y を作る。この y の平均値を $\bar{y}$ 、分散を s_y^2 、標準偏差を s_y とすると以下が成り立つ。(証明は◇問題で。)

$$\bar{y} = a\bar{x} + b \quad s_y^2 = a^2 s_x^2 \quad s_y = |a|s_x$$

つまり、分散、標準偏差には $+b$ の部分は影響しないということ。考えてみたら当たり前で、分散と標準偏差は、散らばり具合だから、 $+b$ したところで、散らばり具合は変わらない。

数学 I あと少しで終わり！！！！

最後まで駆け抜けよう！！



◇問題

1. 25 個の値からなるデータがあり、そのうちの 10 個の値の平均値は 4、分散は 14、残りの 15 個の値の平均値は 9、分散は 19 である。
- (1) このデータの平均値を求めよ。
- (2) このデータの分散を求めよ。

2. a 、 b は定数とする。変数 x のデータから $y = ax + b$ によって新しい変数 y のデータが得られるとき、 x 、 y のデータの平均値を $\bar{x}$ 、 $\bar{y}$ 、分散を s_x^2 、 s_y^2 、標準偏差を s_x 、 s_y とするとき (1)～(3) の式が成り立つことを示せ。

$$(1) \bar{y} = a\bar{x} + b \quad (2) s_y^2 = a^2 s_x^2 \quad (3) s_y = |a|s_x$$

数学 I 第 5 章 データの分析 No.5

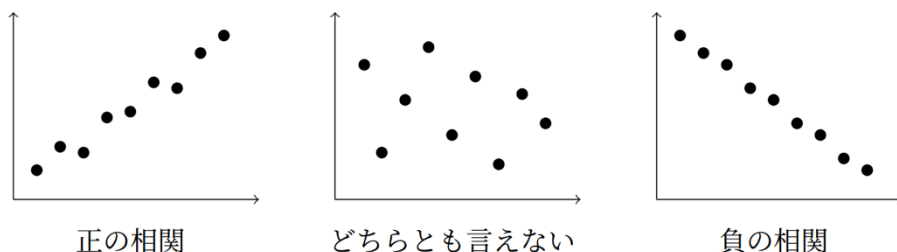
学習のねらい

散布図、データの相関関係について理解しよう！
共分散について理解して公式を覚えよう！

1. 散布図、データの相関関係

下の図は、何かしらのデータを座標平面上に、表したものである。
(例えば、数学の点数： x 点、英語の点数： y 点など。)このような図を
散布図という。ここでは形式的に全て横軸を x 軸、縦軸を y 軸と呼ぶことに
する。

例えば、左の図は、 x が増加するにつれて y も増加している。このよう
な関係を正の相関関係があるという。一方、右の図は、 x が増加するに
つれて y は減少している。このような関係を負の相関関係があるとい
う。真ん中の図のようにどちらの傾向も見られない場合、相関関係がな
いという。



2. 共分散

しかし、散布図を見たときに相関関係の判断に迷った場合……。人
によって意見が割れてしまうかもしれない。しかも、毎回散布図を見る
のがダルい。つまり「相関関係の正 or 負」が分かれば良いと思うわけ
だ。そこで用意したのが、共分散だ。変数 x 、 y の共分散を s_{xy} という記
号で表す。

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

例) 以下はお友達の勉強時間とテストの点数をまとめた表である。

お友達	A	B	C	D	E
勉強時間(x 時間)	1	2	3	4	5
点数(y 点)	40	50	65	70	75

このデータの共分散を求める。

まず、 $\bar{x} = 3$ 、 $\bar{y} = 60$

生徒	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
A	1	40	-2	-20	40
B	2	50	-1	-10	10
C	3	65	0	5	0
D	4	70	1	10	10
E	5	75	2	15	30

$(x - \bar{x})(y - \bar{y})$ の平均は $\frac{1}{5}(40 + 10 + 0 + 10 + 30) = 18$ 。よって、共分
散が正なので、勉強時間と点数には正の相関があると考えられる。

◇問題

1. 次の表は、ある地域における月ごとのアイスクリーム売り上げと熱中症患者数を表したものである。

月	アイス売り上げ	熱中症患者数
6月	少	少
7月	中	中
8月	多	多

この表を見ると、アイスの売り上げが多い月ほど、熱中症患者数も多いことがわかる。

- (1) この2つのデータには、正の相関・相関なし・負の相関のうちどれが見られるか
- (2) 「アイスを食べると熱中症になる」と結論づけて良いか。理由も説明せよ。

数学 I 第 5 章 データの分析 No.6

学習のねらい

相関係数について理解しよう！

変量の変換を行った場合の相関係数について理解しよう！

1. 相関係数

共分散で、「相関関係の正 or 負」が分かる。しかし問題がある。まず強弱が分からない。また、比較ができない。というのも、No.5 の例を

例) 以下はお友達の勉強時間とテストの点数をまとめた表である。

お友達	A	B	C	D	E
勉強時間(x時間)	1	2	3	4	5
点数(y点)	40	50	65	70	75

10 点満点にしたら、共分散の値は 100 分の 1 になってしまう。本質は変わらないのに・・・。

そこで、個数や単位に関わらず、上手に-1から1までの範囲を取るようなものを考えた。それが相関係数である。相関係数は r で表す。

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n} \{(x_1 - \bar{x})(y_1 - \bar{y}) + \dots (x_n - \bar{x})(y_n - \bar{y})\}}{\sqrt{\frac{1}{n} \{(x_1 - \bar{x})^2 + \dots (x_n - \bar{x})^2\}} \sqrt{\frac{1}{n} \{(y_1 - \bar{y})^2 + \dots (y_n - \bar{y})^2\}}}$$

ただし $-1 \leq r \leq 1$ である。

r が1(or-1)に近いとき、強い正(or 負)の相関があるということ。

相関係数を求めるときは、表を書くの良い。先の例で書いてみよう。

生徒	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
A	1	40	-2	-20	4	400	40
B	2	50	-1	-10	1	100	10
C	3	65	0	5	0	25	0
D	4	70	1	10	1	100	10
E	5	75	2	15	4	225	30
計	15	300	0	0	10	850	90

よって、分散 $s_x^2 = \frac{10}{5} = 2$ 、 $s_y^2 = \frac{850}{5} = 170$ となるので、標準偏差 $s_x =$

$\sqrt{2}$ 、 $s_y = \sqrt{170}$ となる。また、共分散 $s_{xy} = \frac{90}{5} = 18$ である。

よって、 $r = \frac{18}{\sqrt{2} \times \sqrt{170}} = 0.976 \dots$ となる。

ゆえに、強い正の相関があると言える。

2. 変量の変換を行った場合の相関係数

$u = ax + b$ 、 $v = cy + d$ で変換するとする。No.4 の公式を使って、

$s_u = |a|s_x$ 、 $s_v = |c|s_y$ 。また、共分散は $s_{uv} = acs_{xy}$ となるので、

$r_{uv} = \frac{ac}{|ac|} \cdot r_{xy}$ となる。 $ac > 0$ のとき、 $r_{uv} = r_{xy}$ 、 $ac < 0$ のとき、 $r_{uv} =$

$-r_{xy}$ 。つまり、変換をしても場合によって符号しか変わらない。

◇問題

1. 次の表は学生 5 名の身長 x (cm)と体重 y (kg)を測定した結果である。
 x と y の相関係数 r を求めよ。

	A	B	C	D	E
身長 x (cm)	181	167	173	169	165
体重 y (kg)	75	59	63	67	61

数学 I 第 5 章 データの分析 No. 7

学習のねらい

仮説検定の考え方を理解しよう！

ちょっと自分の感覚を確認してみよう。(表も裏も $\frac{1}{2}$ で出る)100 枚コインを投げて、60 枚以上表が出る確率はどのくらいだろう？実際に、コイントスができるページに行って試してみたい。(Web コイントスツール(確率任意設定、300 枚まで同時振対応) <https://noveblo.com/coin/>)



すると、思ったよりも出ないと思う。実は、確率を計算すると、2.8%しか出ないのだ。あらびっくり。

もちろん、今回、 $\frac{1}{2}$ で設定したけれど、それが分かっていたなかった場合。100 枚投げて 60 枚以上表が出たら、このコインはイカサマか？を判断したくなるでしょう。

そこで登場するのが仮説検定である！まず自分の主張したい仮説と自分の主張と逆の仮説の 2 つを立てる。今回の例なら、

[A] このコインはイカサマをしている。(主張したいこと)

[B] このコインはイカサマをしていない。(逆の主張)

B を否定することで、A が妥当か判断するということだ。つまり、「このコインはイカサマをしていない」＝「このコインは公平」という仮説のもと確率を計算するわけだ。

そして、もう一つ。「何%以下」なら「B でない」というか」という基準を決める。よく用いられるのは、5%と 1%である。

つまり、先の例なら、

Step. 1

仮説の設定

[A] このコインはイカサマをしている。

[B] このコインはイカサマをしていない。

Step. 2

基準となる確率の設定

60 枚以上表が出る確率が 5%を切ったら B を否定できるとする。

Step. 3

データや計算から判定

計算した結果 2.8%で 5%より小さいから、仮説 B は棄却！

よって A を採用する！このコインはイカサマをしているだろう！

注意点があって、例えば、58 枚以上表が出る確率は 6.7%となる。このようになると、仮説 B が棄却できない。だからと言って、「B だ！」＝「このコインはイカサマをしていない」とも言えない。

例えば裁判で、有罪を主張したいとする。つまり[A] 被告人は有罪 [B] 被告人は無罪となる。証拠を集めた結果、仮説 B が棄却できないとする。だからと言って、本当に無罪かは分からない。有罪にするための証拠が足りなかっただけである。仮説検定も同じである。

◇問題

1. ある人がコインを見せて、
「このコインは普通のコインで、表も裏も同じ確率で出るよ！」
と言ってきた。
そこで、このコインが本当にそのようなコインかを確かめるために、このコインを 10 回投げたところ、表が 9 回出た。
(1) まず、このコインについてどのような仮説を置くと自然か書け。
(2) (1)で立てた仮説のもとで、10 回投げたとき表が 9 回以上出る確率を求めなさい。
(3) (2)の結果から、このコインについてどのように考えられるか説明しなさい。
(ただし基準とする確率は 5%とする。)
(この問題には数学 A の反復試行の知識が必要。(数学 A 第 2 章 No.3))

数学 I 終わり！！！！
お疲れ様！！！！！！！！

