

富士未来学 I

ーデータ分析講座 I

データ分析講座 I でできるようになること

平均値や中央値、最頻値などの代表値の長所と短所を理解し、データの特徴を表現することができる。
箱ひげ図や分散、標準偏差からデータの散らばりなどの特徴を判断することができる。

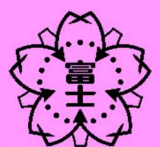
データ分析講座 I で学ぶこと

平均値や中央値、最頻値などの代表値の意味や求め方を学ぶ。箱ひげ図の作成方法、分散と標準偏差の計算方法と、その意味を学ぶ。

月 日 ()

東京都立富士高等学校

東京都立富士高等学校附属中学校



中学1年 組 番 氏名



データ分析講座 I ①

ループリックを見返しながら、自己の到達度を確認し、改善を図りましょう。

育成したい資質・能力	グランドデザインの観点	評価の観点	評価の対象	高度に達成されている	達成されている	一部に課題あり	自己評価	教員による評価
				A	B	C		
挑戦力	試行錯誤	主体的に学習に取り組む態度	2の記述	2において、中央値の短所について、理由や具体例とともに記述している。	2において、中央値の短所について記述しているが、理由や具体例がないなどの不備がある。	2において、記述していない。		
理数的解決力	分析解析	思考・判断・表現	4の記述	4において、ヒストグラムで変化したことと、それにより分かったことを記述している。	4において、ヒストグラムで変化したことを記述している。	4において、記述していない。		
理数的解決力	分析解析	思考・判断・表現	5の記述	5において、ヒストグラムを確率と見なす考え方について理解し、案を理由とともに記述している。	5において、ヒストグラムを確率と見なす考え方について理解し、案を記述している。	5において、案を記述していない。		

1 代表値

データの分布の特徴を表す数値として、代表値があります。代表値は分布全体を1つの値で代表するものです。平均値、中央値、最頻値などがあります。最大値や最小値なども代表値の1つです。

(1) 平均値

全データの和を、データの総数で割った値のことです。

(2) 中央値（メジアン）

全データを小さい順または大きい順に並べたときに、ちょうど真ん中にくる値のことです。全データの総数が偶数の場合は、中央の2つの値の平均値を中央値とします。

(3) 最頻値（モード）

全データの中で最も多く現れるデータの数のことです。

次のデータに対して、代表値をそれぞれ求めましょう。

1 2 2 2 4 5 5 6 10 100

①平均値

②中央値

③最頻値



2 代表値の長所と短所

分布の全体的な位置から極端に離れた位置にあるデータを、**外れ値**といいます。先ほどのデータでは、100が外れ値にあたります。が実験などにおいて、計器の故障や計測ミスなどで外れ値が生じることがあります。

また、例のように、平均値は外れ値の影響を受けやすいという特徴があります。それに対して、中央値や最頻値は外れ値の影響を受けにくいという特徴があります。では、平均値よりも中央値や最頻値が代表値として優れているのでしょうか。

表1 代表値の長所と短所

	長所	短所
平均値	全データから求めているので、全データを考慮しているといえる	外れ値の影響を受けやすい
中央値	外れ値の影響を受けにくい	中央のデータのみから求めているので、それ以外のデータを考慮していないといえる
最頻値	外れ値の影響を受けにくい	一部のデータから求めているので、それ以外のデータを考慮していないといえる データの個数が少ない場合はあまり意味がない

表1のように、代表値には、それぞれの長所と短所があります。このことを理解しながら、代表値を扱う必要があります。

表1以外の、中央値の短所を挙げ、理由や具体例とともに書きましょう。

スポーツ競技などの採点では、外れ値などの極端なデータを除いてから平均値を求める方法を採用しているものもあります。



3 度数分布表

次のデータは、ある中学 1 学年 1 0 0 人分のハンドボール投げの記録です。単位は〔m〕です。

1 6	1 2	2 7	1 8	1 8	2 3	2 2	2 4	1 5	1 3
2 6	1 2	2 4	2 4	1 5	1 0	1 8	1 5	1 8	1 8
1 8	1 8	1 5	1 6	2 1	1 1	1 2	2 0	2 6	2 7
1 6	2 0	2 5	2 1	1 8	1 8	2 3	1 6	1 8	2 4
1 6	1 8	1 4	1 8	1 4	1 4	1 8	1 5	1 4	1 8
2 3	2 3	2 3	1 4	1 4	2 1	2 1	2 7	2 5	2 3
2 0	2 2	2 7	1 8	1 8	1 4	1 8	1 8	2 7	2 4
1 5	2 5	1 5	2 4	2 3	2 1	2 5	2 5	1 5	1 6
2 4	1 1	2 5	2 3	1 3	1 3	2 0	1 5	2 0	2 6
1 8	2 0	2 5	2 2	2 3	2 3	2 1	2 2	1 6	2 2

上のデータを用いて、**度数分布表**（データをある幅ごとに区切ってその中に含まれるデータの個数を表にしたもの）を作成すると表 2 のようになります。ここで、階級、階級の幅、階級値、度数、累積度数、相対度数について言葉の意味を確認します。

(1) **階級**

データの区間のことで、区間の幅を**階級の幅**といいます。表 2 の階級の幅は 3 です。

(2) **階級値**

階級の中央の値のことです。度数の一番大きい階級の階級値を、最頻値とします。

(3) **度数**

階級に含まれるデータの個数のことです。累積度数は、度数を累積したものです。

(4) **相対度数**

各階級の度数が、全体の中でどれだけの割合にあたるかを示す値のことです。

表 2 度数分布表

階級〔m〕	階級値	度数〔人〕	累積度数	相対度数	累積相対度数
以上～未満					
6～ 9	7.5	0	0	0.0 0	0.0 0
9～1 2	1 0.5	3	3	0.0 3	0.0 3
1 2～1 5	1 3.5	1 3	1 6	0.1 3	0.1 6
1 5～1 8	1 6.5	1 6	3 2	0.1 6	0.3 2
1 8～2 1	1 9.5	2 5	5 7	0.2 5	0.5 7
2 1～2 4	2 2.5	2 1	7 8	0.2 1	0.7 8
2 4～2 7	2 5.5	1 7	9 5	0.1 7	0.9 5
2 7～3 0	2 8.5	5	1 0 0	0.0 5	1.0 0
3 0～3 3	3 1.5	0	1 0 0	0.0 0	1.0 0
合計		1 0 0		1.0 0	

度数分布表を作成することで、データの分布がわかりやすくなります。階級を変えると度数分布表が変わることに注意が必要です。



4 ヒストグラム

度数分布表をグラフで表したものを**ヒストグラム**といいます。表2の度数分布表をヒストグラムで表してみます。

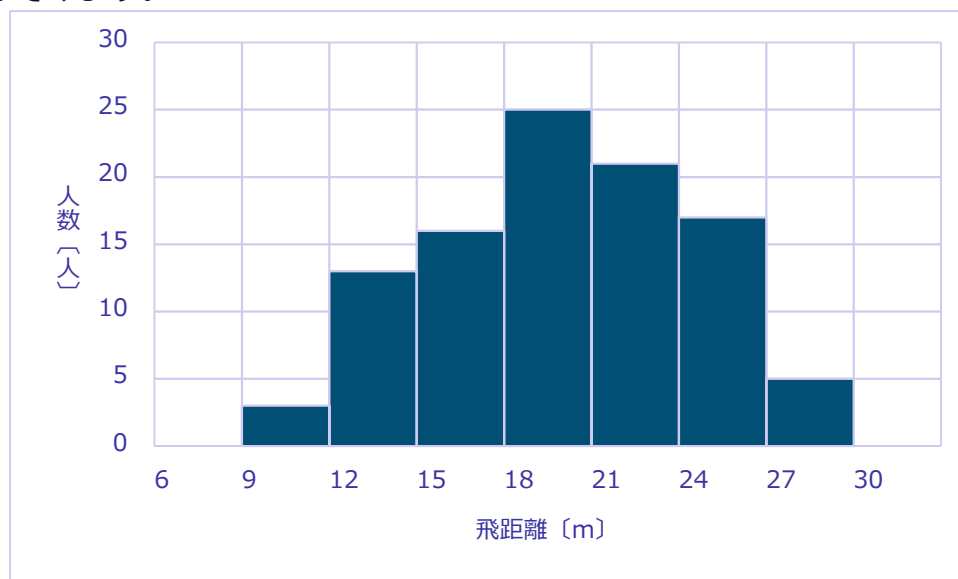


図1 ヒストグラム1

同じデータでも、図2のようなヒストグラムもあります。

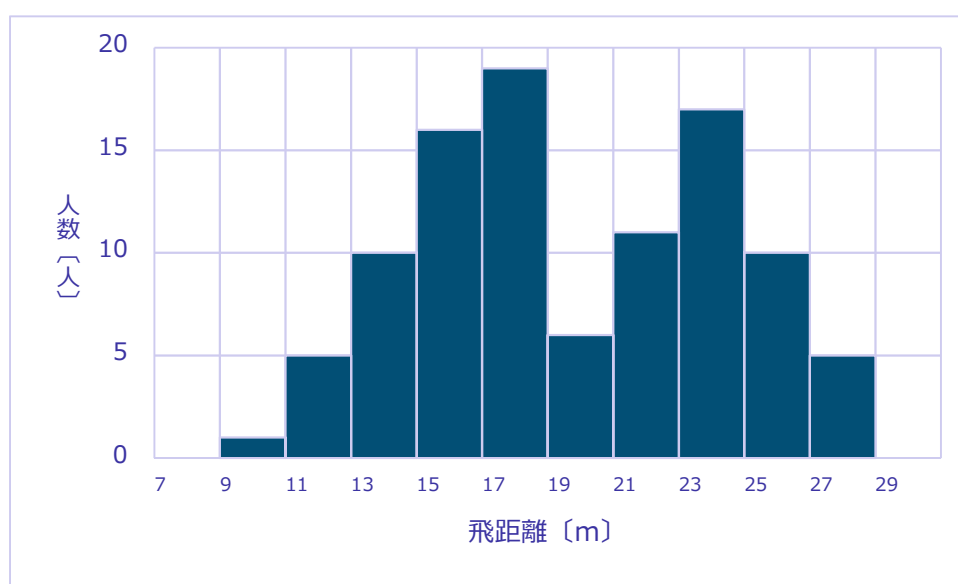


図2 ヒストグラム2

図1と図2を比較して、変化したところは何でしょうか。また、このことから分かることを書きましょう。



5 相対度数を確率とみなす

多数回の観察や試行の結果を基に、不確定な事象について考察する際には、相対度数を確率（ある事柄が起こると期待される程度）と見なして用いることが考えられます。ハンドボール投げを例に考えると、100人分のデータで求めた相対度数は、ある中学1学年の生徒のハンドボール投げの飛距離を予測する際に、活用できます。また、相対度数を確率とみなすことは、ヒストグラムをある事象の起こりやすさを視覚化したものと見なすことにもなります。

ハンドボール投げの飛距離を計測する係2名の配置を考えます。図2のヒストグラムから

考えて、何m地点に計測係を配置すれば、負担が少なく効率よく計測できるでしょうか。

案を理由とともに書きましょう。

6 ルーブリックによる自己評価

育成したい資質・能力	グランドデザインの観点	評価の観点	評価の対象	高度に達成されている	達成されている	一部に課題あり	自己評価	教員による評価
				A	B	C		
挑戦力	試行錯誤	主体的に学習に取り組む態度	2の記述	2において、中央値の短所について、理由や具体例とともに記述している。	2において、中央値の短所について記述しているが、理由や具体例がないなどの不備がある。	2において、記述していない。		
理数的解決力	分析解析	思考・判断・表現	4の記述	4において、ヒストグラムで変化したことと、それにより分かったことを記述している。	4において、ヒストグラムで変化したことを記述している。	4において、記述していない。		
理数的解決力	分析解析	思考・判断・表現	5の記述	5において、ヒストグラムを確率と見なす考え方について理解し、案を理由とともに記述している。	5において、ヒストグラムを確率と見なす考え方について理解し、案を記述している。	5において、案を記述していない。		

引用文献

- (1) 文部科学省（2017）『中学校学習指導要領解説数学編』，p.89.
- (2) 東京都立富士高等学校・東京都立富士高等学校附属中学校（2021）『令和3年度スーパーサイエンスハイスクール研究開発実施計画書【開発型・実践型】』



データ分析講座 I ②

ループリックを見返しながら、自己の到達度を確認し、改善を図りましょう。

育成したい資質・能力	グランドデザインの観点	評価の観点	評価の対象	高度に達成されている	達成されている	一部に課題あり	自己評価	教員による評価
				A	B	C		
理数的解決力	分析解析	思考・判断・表現	7の記述	7において、自力または解説を参考にして、値をすべて求めている。	7において、自力または解説を参考にして、値を求めようとしているが、途中で終わっているなど、不足がある。	7において、値を1つも求めている。		
理数的解決力	分析解析	思考・判断・表現	8の記述	8において、箱ひげ図から数値を読み取り、箱ひげ図を読み取ったことから分かることを記述している。	8において、箱ひげ図から数値を読み取っているが、箱ひげ図を読み取ったことから分かることを記述していない。	8において、箱ひげ図から数値を読み取っていない。		
理数的解決力	分析解析	思考・判断・表現	9の記述	9において、箱ひげ図がヒストグラムより優れている点について、理由とともに記述している。	9において、箱ひげ図がヒストグラムより優れている点について記述しているが、理由がない。	9において、記述していない。		

7 五数要約

データの分布の特徴を表す数値として代表値を学習しましたが、代表値の比較だけでは、わからないことがあります。

データ1 : 4 6 4 7 4 8 4 9 5 0 5 1 5 2 5 3 5 4

データ2 : 1 0 2 0 3 0 4 0 5 0 6 0 7 0 8 0 9 0

データ1とデータ2の平均値と中央値は、それぞれ50になっています。同じ値ですが、散らばりの度合いが違います。度数分布表やヒストグラムでも散らばりの度合いを調べることができますが、数値で調べることはできないのでしょうか。

データの散らばりの度合いを調べる簡単な方法に、**範囲**（レンジ）があります。範囲は、最大値から最小値を引いた値になります。

次のデータの範囲をそれぞれ求めましょう。また、そこから分かることを書きましょう。

①データ1の範囲

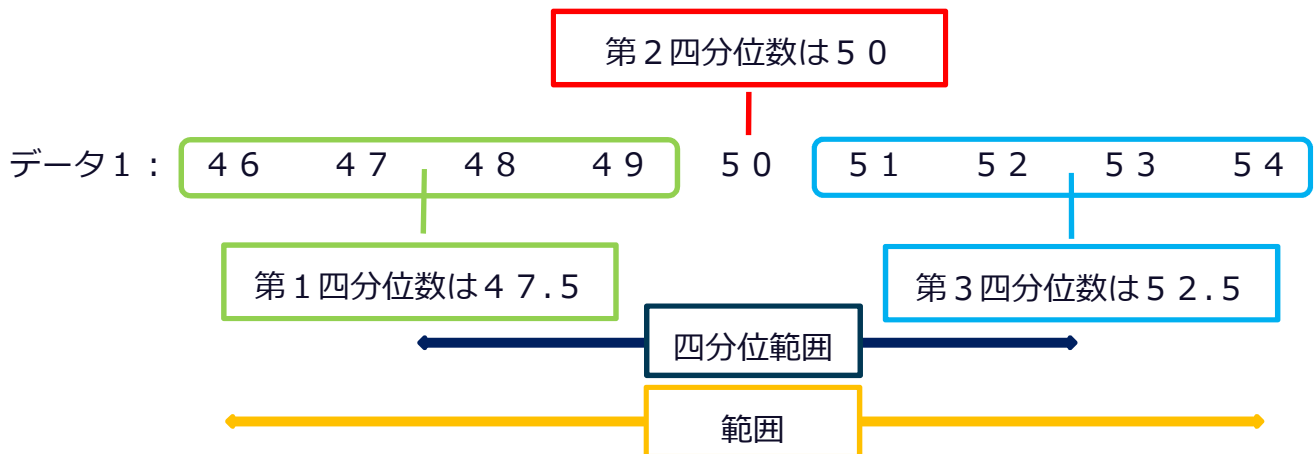
②データ2の範囲

③分かること



散らばりの度合いを表す数値として、**四分位数**があります。データを値の大きさ順に並べたとき、4等分する位置にくる値です。小さい方から、**第1四分位数**、**第2四分位数**（中央値）、**第3四分位数**といいます。この3つの数に、**最小値**と**最大値**を合わせた5つの数でデータの散らばりの度合いを表すことを、**五数要約**といいます。

四分位数について考察していきましょう。まず、第2四分位数は、中央値になります。中央値を境にして、データを分けたとき、それぞれのデータの中央値が、第1四分位数と第3四分位数になります。



また、第3四分位数から第1四分位数を引いた値を**四分位範囲**といい、中央の50%のデータの範囲にほぼ等しくなります。四分位範囲を2で割った値を**四分位偏差**といいます。

データ1の四分位範囲は5で、四分位偏差は2.5です。

データ2について、五数要約、四分位範囲、四分位偏差をそれぞれ求めましょう。

10	20	30	40	50	60	70	80	90
①最小値			②第1四分位数			③第2四分位数		
④第3四分位数			⑤最大値					
⑥四分位範囲				⑦四分位偏差				

先ほどは、データの個数が奇数の場合を考えましたが、次は、データの個数が偶数の場合を考えます。第2四分位数は、中央値になります。中央値を境にして、データを分けます。それぞれのデータの中央値が、第1四分位数と第3四分位数になります。





8 箱ひげ図

次のデータは、ある中学 1 学年 100 人分のハンドボール投げの記録です。単位は〔m〕です。

16	12	27	18	18	23	22	24	15	13
26	12	24	24	15	10	18	15	18	18
18	18	15	16	21	11	12	20	26	27
16	20	25	21	18	18	23	16	18	24
16	18	14	18	14	14	18	15	14	18
23	23	23	14	14	21	21	27	25	23
20	22	27	18	18	14	18	18	27	24
15	25	15	24	23	21	25	25	15	16
24	11	25	23	13	13	20	15	20	26
18	20	25	22	23	23	21	22	16	22

上のデータを用いて箱ひげ図を作成すると、図 3 のようになります。箱の部分とひげの部分が、それぞれの値を表しています。箱の長さは四分位範囲を表しています。

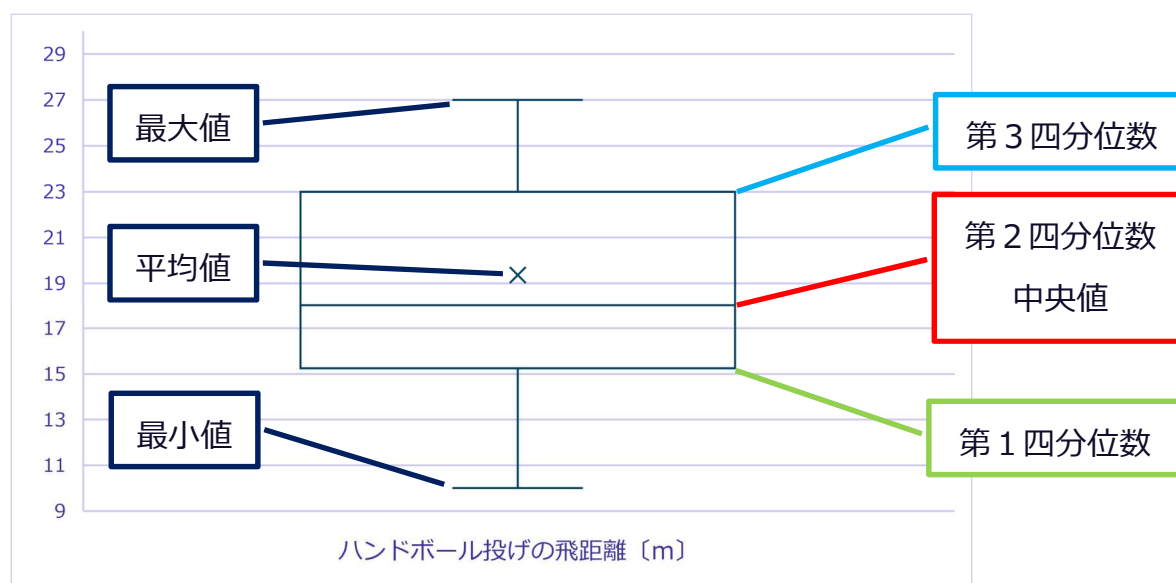


図 3 箱ひげ図

箱ひげ図は、ヒストグラムのように分布の詳細を表現することはできません。しかし、数値だけでは判断しづらかった代表値を、具体的に示すことができます。

図 3 の箱ひげ図から読み取れる数値や、分かることを書きましょう。



9 ヒストグラムと箱ひげ図

図4は、先ほどのデータをヒストグラムで表したものです。図5は、図3の箱ひげ図を横にした図です。

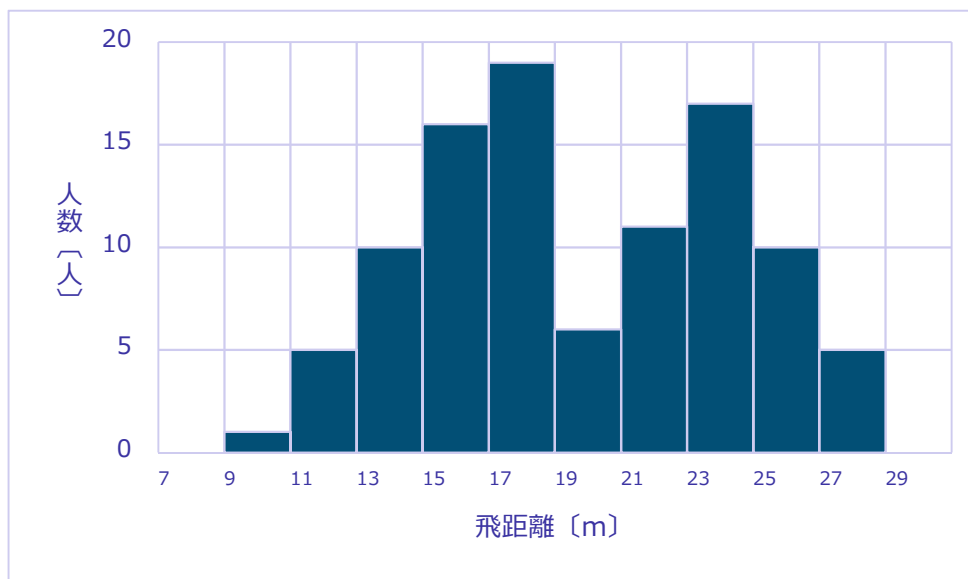


図4 ヒストグラム

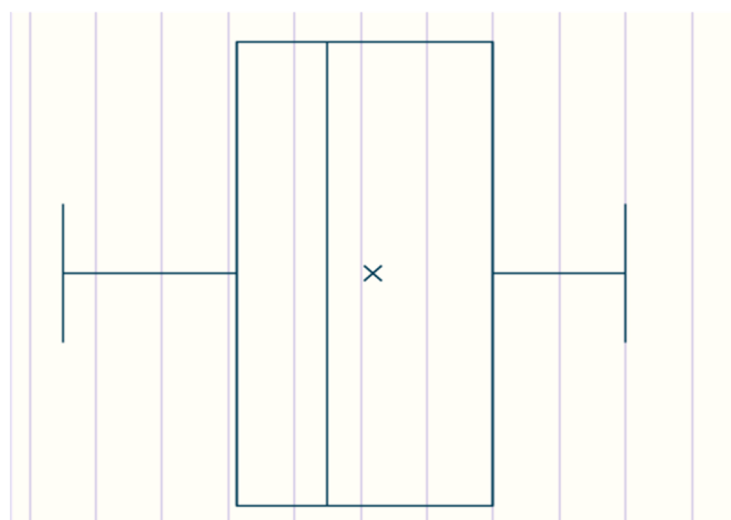


図5 箱ひげ図を横にした図

図4と図5を比較して分かるように、それぞれに長所と短所があります。何の情報を優先的に読み取りたいか、もしくは比較したいかによって使い分けましょう。ヒストグラムは分布の詳細を表現することができます。箱ひげ図はどうでしょうか。

箱ひげ図がヒストグラムより優れている点を、理由とともに書きましょう。



1 0 箱ひげ図による比較

箱ひげ図は、データ間の比較に向いています。図6は、4種類のデータを箱ひげ図で表したものです。数字の羅列と比べると、散らばりの度合いが視覚化されていて、分かりやすくなっています。

データA :	1	2	3	4	5	6	7	8	9	10
データB :	1	2	3	3	4	4	4	6	7	10
データC :	3	4	4	4	5	5	6	6	7	7
データD :	1	1	2	3	7	7	7	8	9	10

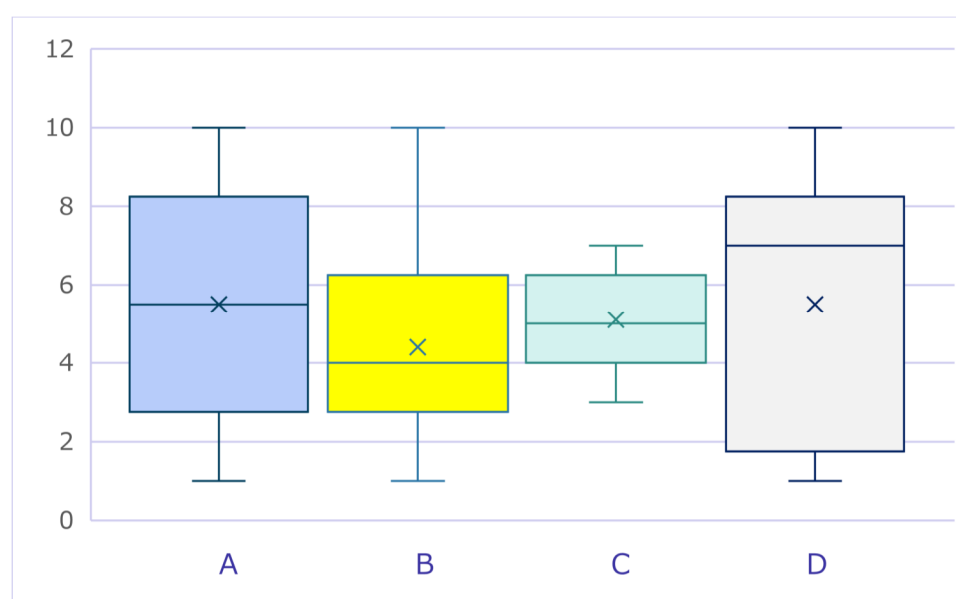


図6 箱ひげ図の比較

1 1 箱ひげ図と外れ値

箱ひげ図のひげの長さは、四分位範囲の1.5倍までとすることがあります。その場合は、ひげの長さを超えるものは、外れ値として扱います。

外れ値の判断は難しく、外れ値を判断するときには、慎重にならなければなりません。

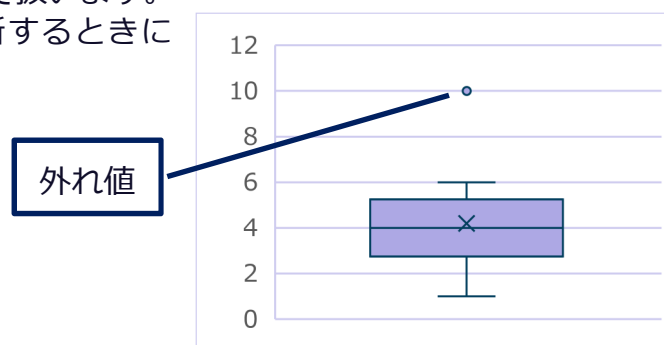


図7 四分位範囲の1.5倍が上限



12 ルーブリックによる自己評価

育成したい資質・能力	学習活動	評価の観点	評価の対象	高度に達成されている	達成されている	一部に課題あり	自己評価	教員による評価
				A	B	C		
理数的解決力	分析解析	思考・判断・表現	7の記述	7において、自力または解説を参考にして、値をすべて求めている。	7において、自力または解説を参考にして、値を求めようとしているが、途中で終わっているなど、不足がある。	7において、値を1つも求めている。		
理数的解決力	分析解析	思考・判断・表現	8の記述	8において、箱ひげ図から数値を読み取り、箱ひげ図を読み取ったことから分かることを記述している。	8において、箱ひげ図から数値を読み取っているが、箱ひげ図を読み取ったことから分かることを記述していない。	8において、箱ひげ図から数値を読み取っていない。		
理数的解決力	分析解析	思考・判断・表現	9の記述	9において、箱ひげ図がヒストグラムより優れている点について、理由とともに記述している。	9において、箱ひげ図がヒストグラムより優れている点について記述しているが、理由がない。	9において、記述していない。		

引用文献

- (1) 文部科学省（2017）『中学校学習指導要領解説数学編』，p.89.
- (2) 東京都立富士高等学校・東京都立富士高等学校附属中学校（2021）『令和3年度スーパーサイエンスハイスクール研究開発実施計画書【開発型・実践型】』



データ分析講座Ⅰ③

ループリックを見返しながら、自己の到達度を確認し、改善を図りましょう。

育成したい資質・能力	グランドデザインの観点	評価の観点	評価の対象	高度に達成されている	達成されている	一部に課題あり	自己評価	教員による評価
				A	B	C		
理数的解決力	分析解析	思考・判断・表現	14の記述	14において、自力または解説を参考にして、偏差、偏差の絶対値、偏差の2乗の合計と平均値など、すべて求めている。	14において、自力または解説を参考にして、偏差、偏差の絶対値、偏差の2乗の合計と平均値などを求めようとしているが、途中で終わっているなど、不足がある。	14において、偏差、偏差の絶対値、偏差の2乗の合計と平均値などを1つも求めている。		
理数的解決力	分析解析	思考・判断・表現	15と16の記述	15と16において、自力または解説を参考にして、分散と標準偏差をともに求めている。	15と16において、自力または解説を参考にして、分散を求めているが、標準偏差を求めている。	15と16において、分散や標準偏差をどちらも求めている。		

13 散布度

データの分布の散らばりを表す数値として、散布度があります。散布度には、前回の講座で学習した四分位偏差や、これから学習する分散や標準偏差があります。四分位偏差は、中央値を基準とした散らばりの度合いを表しますが、分散と標準偏差は、どちらも平均値からの散らばりの度合いを表します。

表3 1日あたりの平均売上〔万円〕

店	A	B	C	D	E	合計	平均値
夏	21	23	25	26	30	125	25
冬	10	15	25	35	40	125	25

夏と冬のどちらが平均値からの散らばりの度合いが大きいでしょうか。大きいと判断した

理由を書きましょう。



1.4 偏差

各データの値と平均値の差を、**偏差**といいます。各偏差の大きさを比較しながら、散らばりの度合いを判断してみましょう。

$$(\text{偏差}) = (\text{値}) - (\text{平均値})$$

ここで、平均値を記号 $\bar{x}$ で表すことにします。各偏差は表4のようになります。

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

表4 変数 x の平均値 $\bar{x}$ と偏差

	1	2	3	...	n
x	x_1	x_2	x_3		x_n
偏差	$x_1 - \bar{x}$	$x_2 - \bar{x}$	$x_3 - \bar{x}$		$x_n - \bar{x}$

各偏差の大きさを1つの数値で表すことができると、データの比較がしやすくなります。そこで、偏差の平均値を考えてみましょう。表3の例を使用してください。

①表5を完成させましょう。

表5 偏差の合計と平均値

店	A	B	C	D	E	合計	平均値
夏	21	23	25	26	30	125	25
偏差							
冬	10	15	25	35	40	125	25
偏差							

一般的に、偏差の合計と平均値はともに0となることが分かっています。そのため、これでは比較することはできません。次に、偏差の絶対値の平均値や、偏差の2乗の平均値を考えてみましょう。



②表 6 を完成させましょう。この場合の平均値を、**平均偏差**といいます。

表 6 偏差の絶対値の合計と平均値

店	A	B	C	D	E	合計	平均値
夏	2 1	2 3	2 5	2 6	3 0	1 2 5	2 5
偏差の絶対値							
冬	1 0	1 5	2 5	3 5	4 0	1 2 5	2 5
偏差の絶対値							

③表 7 を完成させましょう。この場合の平均値を、**分散**といいます。

表 7 偏差の 2 乗の合計と平均値

店	A	B	C	D	E	合計	平均値
夏	2 1	2 3	2 5	2 6	3 0	1 2 5	2 5
偏差の 2 乗							
冬	1 0	1 5	2 5	3 5	4 0	1 2 5	2 5
偏差の 2 乗							

平均偏差と分散のどちらも、平均値からの散らばりの度合いを表すことができます。しかしながら、表 6 の平均偏差が使われることはほとんどありません。表 7 の分散は便利な性質をもっていることが知られており、データ分析講座 II で学習する共分散や相関係数を求めるときにも必要です。



1.5 分散

分散は、1変量について、各データの平均値を基準とした散らばりの度合いを表す値です。偏差の2乗の平均値を、分散といいます。**分散 s^2** を式で表すと次のようになります。

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

分散の値が大きいほど、平均値を基準とした散らばりの度合いが大きいと解釈できます。平均値との差が大きいデータが多いほど、分散の値が大きくなると考えられるからです。

表8 数学の得点

学生	1	2	3	4	5	6	7	8	9	10
数学	40	75	65	75	20	60	50	70	30	55

表8のデータを用いて、数学の得点の分散を求めましょう。

<自分で解くためのスペース>

<解説を書くためのスペース>



1.6 標準偏差

標準偏差は、1 変量について、各データの平均値を基準とした散らばりの度合いを表す値です。偏差の 2 乗の平均値を分散といいます。その分散の値の正の平方根が標準偏差です。標準偏差 s_x を式で表すと次のようになります。

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

2 乗（平方）すると a になる数を、 a の平方根といいます。例えば、2 は 2 乗すると 4 になり、 -2 も 2 乗すると 4 になります。つまり、4 の平方根は 2 と -2 の二つです。また、2 の正の平方根は、記号 $\sqrt{\quad}$ を用いて、 $\sqrt{2}$ と表します。読み方は「ルート 2」で、実際の値は、 $\sqrt{2} = 1.41421356 \cdots$ で、循環しない無限小数（無理数）であることが知られています。

分散と標準偏差の単位を確認します。分散は計算過程で 2 乗されるので、データの単位が 2 乗されると考えられます。例えば、元のデータの単位が「点」であるとき、分散の単位は「点²」となります。そこで、分散の平方根となる標準偏差を考えることによって、標準偏差の単位は「点」となり、データの単位と同じものになります。分散では単位を省略しますが、標準偏差では単位を明記します。

表 8 のデータを用いて、数学の得点の標準偏差を求めましょう。

<自分で解くためのスペース>

<解説を書くためのスペース>

1.7 単位について

分散と標準偏差は、どちらも平均を基準としたばらつき具合を表しています。単位がデータと同じものになる標準偏差が、より分かりやすい指標になります。分散や標準偏差は、単位が a 倍になるとそれぞれ a^2 倍、 a 倍となります。



1 8 視覚的な説明

分散の視覚的な説明をします。あくまでも補足説明です。青い点はデータを表しています。緑の直線は平均値を表しています。

分散は、偏差の2乗の平均値です。偏差の2乗は、赤い正方形の面積として捉えることができるので、分散の大きさは、赤い正方形の面積の和として捉えることができます（正確にはその平均値）。平均値からの散らばりの度合いが小さいほど、それぞれの赤い正方形が小さくなり、面積の和も小さくなることが分かります。

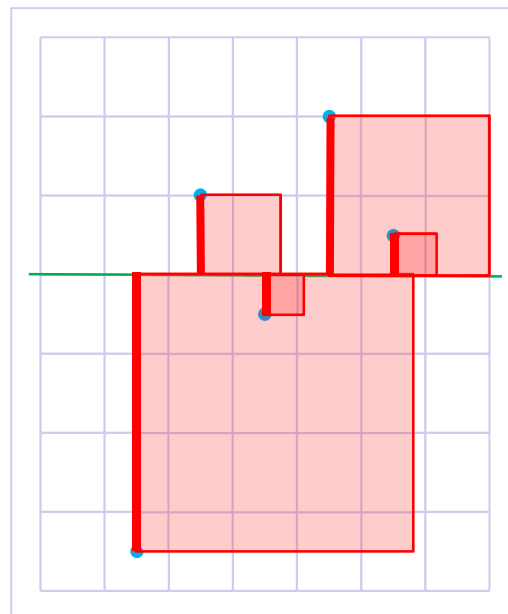


図8 分散を面積で表現

1 9 ルーブリックによる自己評価

育成したい資質・能力	グランドデザインの観点	評価の観点	評価の対象	高度に達成されている	達成されている	一部に課題あり	自己評価	教員による評価
				A	B	C		
理数的解決力	分析解析	思考・判断・表現	14の記述	14において、自力または解説を参考にして、偏差、偏差の絶対値、偏差の2乗の合計と平均値など、すべて求めている。	14において、自力または解説を参考にして、偏差、偏差の絶対値、偏差の2乗の合計と平均値などを求めようとしているが、途中で終わっているなど、不足がある。	14において、偏差、偏差の絶対値、偏差の2乗の合計と平均値などを1つも求めている。		
理数的解決力	分析解析	思考・判断・表現	15と16の記述	15と16において、自力または解説を参考にして、分散と標準偏差をともに求めている。	15と16において、自力または解説を参考にして、分散を求めているが、標準偏差を求めている。	15と16において、分散や標準偏差をどちらも求めている。		

引用文献

- (1) 南風原朝和（2021）『心理統計学の基礎ー統合的理解のために』，有斐閣アルマ，p.31.
- (2) 東京都立富士高等学校・東京都立富士高等学校附属中学校（2021）「令和3年度スーパーサイエンスハイスクール研究開発実施計画書【開発型・実践型】」