データの分析【代表値～箱ひげ図】 p.176～186

青チャートⅠ：例題175～181

【態度目標】しゃべる、質問する、説明する、動く、協力する、貢献する

【内容目標】代表値を活用して情報を読み取れるようになろう

□データの整理

統計では，目的に応じてある集団について調査や実験を行い，それによって得られたデータを分析し，問題を解決したりする。ある高校で 1 年生の学力の傾向を調べて今後の指導に役立てたいとき，例えば学力テストを行って，生徒の得点，所属クラスといった情報を集めて分析するであろう。

この例での得点，所属クラスのように，ある集団を構成する人や物の特性を表すものを **変量** といい，実験や調査などで得られた変量の観測値や測定値，調査結果などの集まりを **データ** という。

学力テストの例では，得点のデータは数値として得られるが，所属クラスのデータは「A 組」「B 組」「C 組」のように，数値ではないものとして得られる。得点，温度，長さ，面積，重さのデータのように，数値として得られるデータを **量的データ** という。また，所属クラス，都道府県のデータのように，数値ではないもの（カテゴリーや状態で表されるもの）として得られるデータを **質的データ** という。

次のデータ 1 は，東京の 2018 年 4 月の日ごとの最高気温である。

データを構成する観測値や測定値の個数を，そのデータの **大きさ** という。例えば，上で示した最高気温のデータの大きさは 30 である。

データ 1

21.9	24.5	23.4	26.2	15.3	22.4	21.8	16.8	19.9	19.1
21.9	25.9	20.9	18.8	22.1	20.0	15.0	16.0	22.2	26.4
26.0	28.3	18.7	21.3	22.5	25.0	22.0	26.1	25.6	25.7

（気象庁ホームページより作成，単位は ℃）

□度数分布表とヒストグラム

データの散らばりの様子を **分 布** という。データの分布を見るための 1 つの方法として，**度数分布表** がある。

右の表は，データ 1 の度数分布表である。度数分布表において，区切られた各区間を **階 級**，各区間の幅を **階級の幅**（右の表では 3 ℃），各階級に入るデータの値の個数を **度 数** という。また，各階級の真ん中の値を **階級値** という。例えば，18 ℃ 以上 21 ℃ 未満の階級の階級値は 19.5 ℃ である。

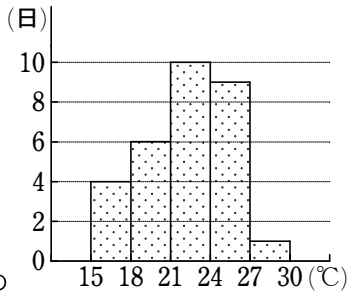
データ 1 の度数分布表

階級（℃）	度数（日）
15 以上 18 未満	4
18 ～ 21	6
21 ～ 24	10
24 ～ 27	9
27 ～ 30	1
計	30

※度数分布表の階級の幅は，データ全体の傾向がもっとも良く表せるように適切な大きさを選ぶことが大切である。

統計学ではスタージェスの公式というものもある
階級の数を n，データの大きさを N とすると
$n = 1 + \log_2 N$
（階級の数から正規分布になるための適切なデータ数を算出する感じ）

度数分布表に整理されたデータを柱状のグラフで表したものを **ヒストグラム** という。右の図は，データ 1 の度数分布表をヒストグラムで表したものである。ヒストグラムの各長方形の面積は，各階級の度数に比例する。ヒストグラムも，データの分布を見るための 1 つの方法である。



他にも ・相対度数…各階級の度数を，度数の合計で割った値。
ある階級の度数が全体に占める割合がわかる
・累積相対度数…相対度数を小さい階級からその階級の値まで合計して得られる値
・累積相対度数折れ線…各階級の累積相対度数を折れ線でつないだもの

データの分析【代表値～箱ひげ図】 p.176～186

青チャートⅠ：例題175～181

□データの代表値

データ全体の特徴を適当な 1 つの数値で表すことがある。その数値をデータの **代表値** という。よく用いられる代表値として，平均値，中央値，最頻値がある。

○ 平均値 (mean)

変量 x についてのデータが n 個の値 x_1，x_2，……，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，$\overline{x}$ で表す。

平均値 $\quad \overline{x}=\dfrac{1}{n}(x_1+x_2+\cdots\cdots+x_n)$

データの値が大きいときなどは **仮平均** を立てて，仮平均との誤差の平均を用いて求めることもある

例１） 東京の最高気温のデータの平均値

$$\frac{1}{30}(21.9+24.5+\cdots\cdots+25.7)=\frac{661.7}{30}\fallingdotseq 22.1\ (℃)$$ 終

平均値を用いることが多いが，極端にかけ離れた値(外れ値)がある場合やデータが非対称な分布のときは検討が必要

○中央値 (median)

次のデータは，A さんの行きの通学時間を，ある週の 5 日について調べた結果である。

42　38　40　44　96　(単位は分)

このデータの平均値は 52 分である。しかし，1 日だけ通学時間が極端に長かったために，この平均値は，他の 4 日の通学時間からは離れたものになっていて，このデータの代表値として適切とはいえない。 このような場合は，通学時間を大きさの順に並べたときの中央の位置にくる値，すなわち 42 分を代表値とすることが考えられる。

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，データの **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，中央に 2 つの値が並ぶが，その場合は 2 つの値の平均値を中央値とする。

例２） ある商品の価格について，A 町と B 町の店舗で調査した。

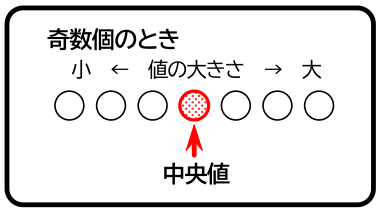
(1) **データの個数が奇数個のとき**

A 町では 5 店舗で調査して，次のデータが得られた。

260　270　280　280　300 (円)

小さい順から並べる (昇順に並べる)

このデータの中央値は　280 (円)



(2) **データの個数が偶数個のとき**

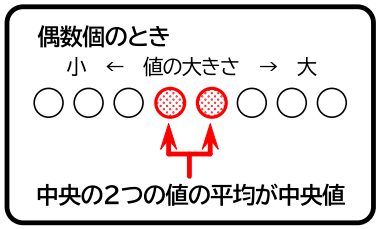
B 町では 6 店舗で調査して，次のデータが得られた。

100　260　270　280　280　280 (円)

小さい順から並べる (昇順に並べる)

このデータの中央値は

$$\frac{1}{2}(270+280)=275\ (円)$$ 終



中央値は
累積度数分布表があると
求めやすい

中央値は、平均値と異なり極端にかけ離れた値の影響をほとんど受けない。そのため多くの人の感覚に近い値が得られる。

平均値中央値

データの分析【代表値～箱ひげ図】 p.176～186

青チャートⅠ：例題175～181

○最頻値 (mode) (並数とも呼ばれる)

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。データが度数分布表に整理されているときは，度数が最も大きい階級の階級値を最頻値とする。

例３） 次の表は，ある店での 1 週間の靴のサイズ別の販売数である。

サイズ (cm)	24	24.5	25	25.5	26	26.5	27	計
販売数	2	7	13	15	22	10	4	73

最も大きい度数は 22
このときの階級値が最頻値
なので「最頻値は 26 cm」

最頻値は 26 cm である。 (終)

注意

① 最頻値が 2 つ以上ある場合があり，一意的に定まらないこともある。

② 階級の幅を変えると，階級値も度数も変化してしまうため，最頻値も変動してしまう。

生産計画や仕入れ計画を立てる際は，代表値として最頻値が適している。

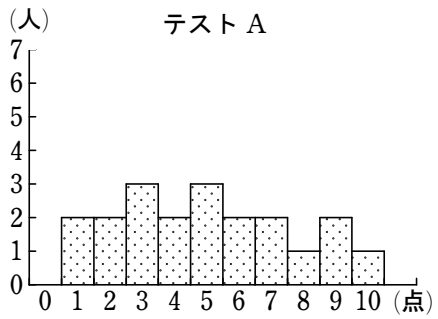
□データの散らばりと四分位範囲

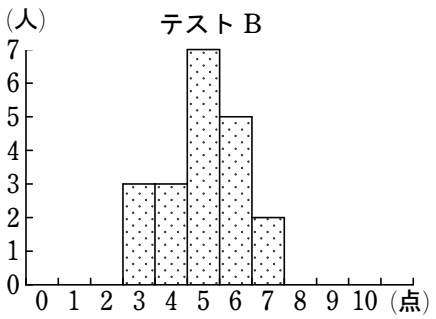
下のデータは，20 人の生徒に 10 点満点の 3 種類のテスト A，B，C を行い，その得点を小さい方から順に並べたものである。

テスト A	1	1	2	2	3	3	3	4	4	5	5	5	6	6	7	7	8	9	9	10
テスト B	3	3	3	4	4	4	5	5	5	5	5	5	5	6	6	6	6	6	7	7
テスト C	1	2	3	3	4	4	4	4	5	5	5	5	5	6	6	6	7	7	8	10

(**単位は点**)

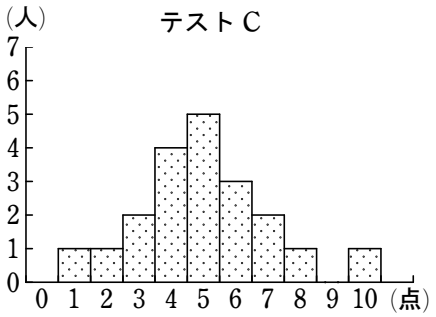
それぞれのデータをヒストグラムにすると，次のようになる。





どのデータも，中央値は 5 点，平均値は 5 点である。しかし，データの散らばりの様子にはかなりの違いが見られる。ここでは，データの散らばり具合いを比較する方法について学ぶ。散らばり具合いは，範囲，四分位範囲などの数値や，箱ひげ図などで表すことができる。

また，データの中に他の値から極端に離れた外れ値と呼ばれる値があった場合に，それをどのように分析するかについても考えよう。



データの分析【代表値～箱ひげ図】 p.176～186

青チャートⅠ：例題175～181

○範囲（range）

データの最大値と最小値の差を **範 囲** という。範囲は，データの散らばりの度合いを表す 1 つの量である。

注意 長所：簡単に求められる　　短所：外れ値など例外的な値が含まれていると影響を大きく受けやすい

例４） テスト A，B のデータの範囲は，次のようになる。

テスト A　$10-1=9$（点）　テスト B　$7-3=4$（点）

テスト A の方が範囲が大きいから，テスト A の方がデータの散らばりの度合いが大きいと考えられる。　終

□四分位範囲

○ 四分位数（quartile）

範囲は簡便な量であるが，散らばりの度合いを比較する量として適切でない場合も多い。例えば，前ページのテスト A とテスト C について，データの範囲はともに 9 点であり，範囲ではこれらの散らばりの違いを表せない。また，範囲はデータの最大値，最小値だけで決まるので，極端に離れた値があると，それによって範囲は大きく変わってしまう。

そこで，データを値の大きさの順に並べて 4 等分し，中央の 50 %のデータについて考えることがある。データを値の大きさの順に並べたとき，4 等分する位置にくる値を **四分位数** という。四分位数は，小さい方から **第 1 四分位数**，**第 2 四分位数**，**第 3 四分位数** といい，これらを順に Q_1，Q_2，Q_3 で表す。 第 2 四分位数は中央値である。

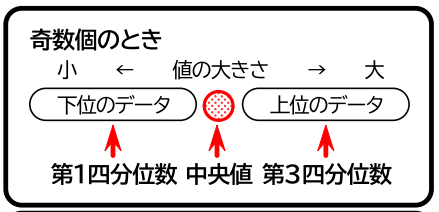
補足 第 1 四分位数は下側四分位，25 パーセンタイル値（25 percentile），
第 3 四分位数は上側四分位，75 パーセンタイル値（75 percentile）と呼ばれることもある

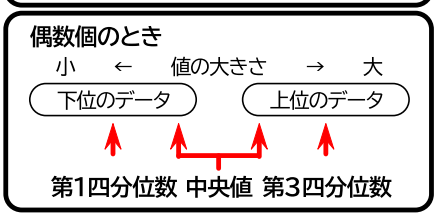
データを値の小さい方から順に左から並べたとき，左半分のデータを下位のデータ，右半分のデータを上位のデータと呼ぶことにする。ただし，データの大きさが奇数のとき，中央の位置にくる値は，下位のデータにも上位のデータにも含めないものとする。

第 1 四分位数 Q_1，第 3 四分位数 Q_3 を次で定める。

第 1 四分位数 Q_1 は　下位のデータの中央値
第 3 四分位数 Q_3 は　上位のデータの中央値

下位のデータ、上位のデータの個数の偶数・奇数で，抽出の仕方を変えよう





【注意】四分位数は，他にもいくつかの定め方がある（Excel等は求め方が違うので値が異なることもある）。

例５） (1) データ 2 3 5 7 11 13 17 19 23 について

第 2 四分位数すなわち中央値は　$Q_2=11$

第 1 四分位数は　$Q_1=\dfrac{3+5}{2}=4$

第 3 四分位数は　$Q_3=\dfrac{17+19}{2}=18$

(2) データ 2 3 5 7 11 13 17 19 23 29 について

第 2 四分位数すなわち中央値は　$Q_2=\dfrac{11+13}{2}=12$

第 1 四分位数は　$Q_1=5$

第 3 四分位数は　$Q_3=19$　　終

① データを値の大きさの順に並べ、中央値（第２四分位数）を求める。
② ①の中央値を境界としてデータの個数を２等分し、値が中央値以下の下位のデータと値が中央値以上の上　位のデータに分ける。ただし、データの大きさが奇数のとき①の中央値はどちらの組にも含めないものとする。
③ 下位のデータの中央値（第１四分位数），上位のデータの中央値（第３四分位数）を求める。

四分位数

データの分析【代表値～箱ひげ図】 p.176～186

青チャートⅠ：例題175～181

○**四分位範囲**（interquartile range, IQR）：中心付近のデータがどのくらい散らばっているかの目安

第 3 四分位数から第 1 四分位数を引いたもの，すなわち Q_3-Q_1 を **四分位範囲** という。四分位範囲は，データを値の大きさの順に並べたときの，中央の 50 %のデータの範囲にほぼ等しく（この中に中央値周辺に並ぶ約 50 %のデータが含まれる。），通常の範囲に比べて極端に離れた値の影響を受けにくい。

四分位範囲	Q_3-Q_1

【補足】四分位範囲を 2 で割った値を **四分位偏差**(quartile deviation) という。

四分位範囲は，データの散らばりの度合いを表す 1 つの量であり，データの値が中央値の周りに集中しているほど、四分位範囲は小さくなる傾向にある。逆に四分位範囲が大きいほど、中央値の周りの散らばりの度合いが大きいと考えられる。

例６）

テスト A	1	1	2	2	3	3	3	4	4	5	5	5	6	6	7	7	8	9	9	10
テスト B	3	3	3	4	4	4	5	5	5	5	5	5	5	6	6	6	6	6	7	7
テスト C	1	2	3	3	4	4	4	4	5	5	5	5	5	6	6	6	7	7	8	10

テスト A のデータについて，$Q_1=3$，$Q_3=7$ より

四分位範囲は　$Q_3-Q_1=7-3=4$（点）

テスト C のデータについて，$Q_1=4$，$Q_3=6$ より

四分位範囲は　$Q_3-Q_1=6-4=2$（点）

テスト A の方が四分位範囲が大きいから，
テスト A の方がデータの散らばりの度合いが
大きいと考えられる。　㊍

補足
幹葉図（みきはず）を用いる方法もある。
例えば一の位のみを書き並べたものは右のようになる

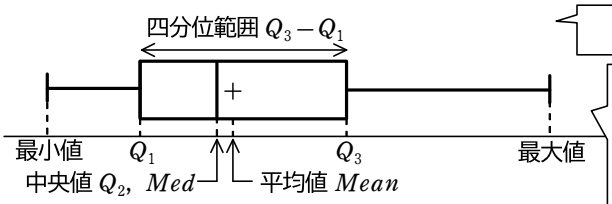
20	1　9
30	2　6　8
40	0　9
50	3　5
60	8
70	
80	0

□**箱ひげ図**（box plot, box and whisker plot）

データの分布を見るための図に **箱ひげ図** と呼ばれるものがある。
箱ひげ図は，データの最小値，第 1 四分位数，中央値，第 3 四分位数，最大値を，箱と線（髭）で表現する図である。箱の長さは四分位範囲を表す。なお，箱ひげ図に平均値を記入することもある。

ジョン・テューキーが1970年代に提唱したもの。ジョン・テューキーはノイマンと共同でコンピュータの設計に関わっており、bit や software という用語を考案したといわれる。

箱ひげ図



縦に書くこともある

補足 データの分布を「最小値、第1四分位数、中央値、第3四分位数，最大値」の5つの値に要約して表すことを**5数要約**（five−number summary）という。箱ひげ図は，これを簡潔に示したものである。

箱ひげ図は複数のデータの分布を比較するときに便利な図である。

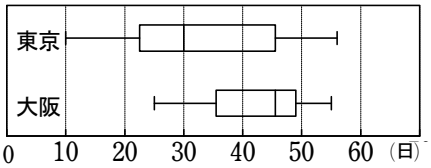
データの分析【代表値～箱ひげ図】 p.176～186

青チャートⅠ：例題175～181

例7） 次のデータは，東京と大阪について，2007 年から 2018 年までの最低気温が 25 ℃ 以上であった日の日数を，1 年ごとに集計した結果である。

東京　31　25　20　56　49　49　39　29　26　10　18　42
大阪　44　42　27　55　51　43　47　29　25　47　47　53
（気象庁ホームページより作成，単位は日）

	最小値	Q_1	中央値	Q_3	最大値
東京	10	22.5	30	45.5	56
大阪	25	35.5	45.5	49	55

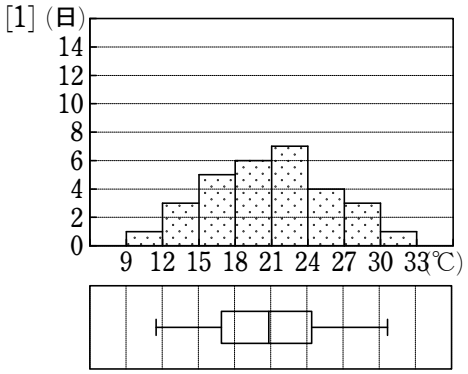


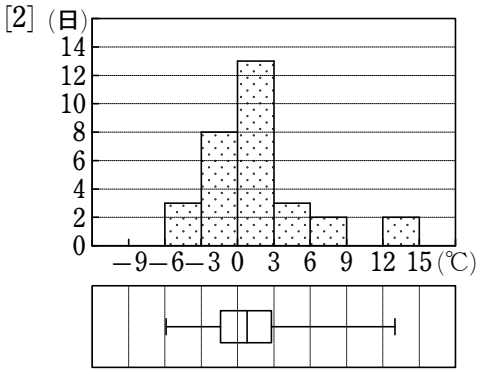
この 2 つのデータの箱ひげ図をかくと，右の図のようになる。

箱ひげ図から，データを値の大きさの順に並べたときの中央の 50 %のデータは，大阪の方が散らばりの度合いが小さく，大阪の方が値が大きい方に分布している傾向が読み取れる。 終

□ ヒストグラムと箱ひげ図

図 [1]，[2] は，札幌の 2018 年の日ごとの最高気温のデータをヒストグラムと箱ひげ図に表したもので，順に 6 月，12 月のものである。（データは 235 ページにある。）



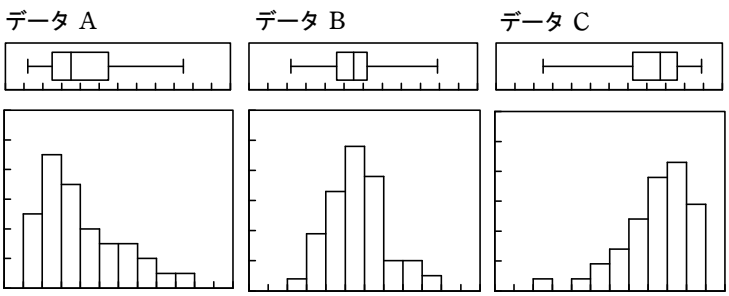


（気象庁ホームページより作成）

ヒストグラムの山の位置と，箱ひげ図の箱の位置がだいたい対応していることがわかる。

また，ヒストグラムのすそにあたる部分が，箱ひげ図のひげに対応している。ヒストグラムのすそが右に伸びていれば，箱ひげ図のひげも右に伸びる。箱ひげ図から，データの分布のおおまかな様子がわかる。

下の図は，あるデータ A，B，C のヒストグラムと箱ひげ図との対応である。



箱ひげ図では，ヒストグラムほどにはデータの分布が詳しく表現されないが，大まかな様子はわかる。

ただし，二山以上（多峰）になる様子など類推できないものもある。

データの分析【代表値～箱ひげ図】 p.176～186

青チャートⅠ：例題175～181

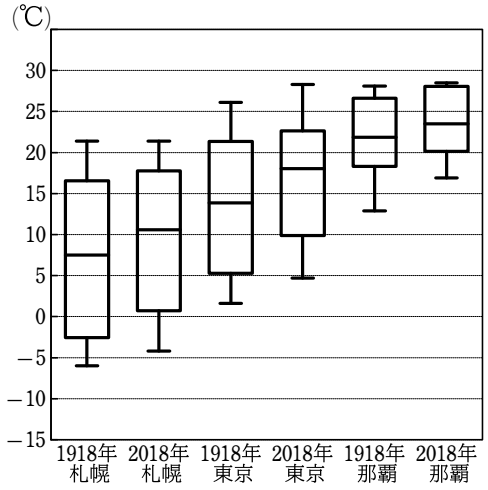
例） 箱ひげ図によるデータの分析

右の図は，1918 年と 2018 年の札幌，東京，那覇における，月ごとの平均気温のデータを箱ひげ図に表したものである。

この図から，たとえば，次のようなことが読み取れる。

1．那覇は 1 年を通して暖かく，寒暖の差が小さい。札幌は年間の寒暖の差が大きい。

2．どの都市も，1918 年より 2018 年の方が気温が高めの傾向にある。（特に Q_1 の上昇が著しい）



(気象庁ホームページより作成)

データの分析【代表値～箱ひげ図】 p.176～186

青チャートⅠ：例題175～181

□外れ値

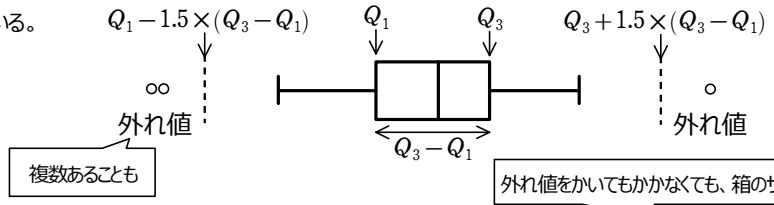
データの中に，他の値から極端に離れた値が含まれることがある。そのような値を **外れ値** という。外れ値の基準は複数あるが，たとえば，次のような値を外れ値とする。

(第 1 四分位数－1.5×四分位範囲) 以下の値

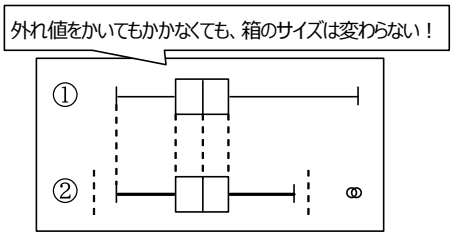
(第 3 四分位数＋1.5×四分位範囲) 以上の値

外れ値がある場合，次の図のような箱ひげ図が用いられることがある。

外れ値は ○ で示している。また，箱ひげ図の左右のひげは，データから外れ値を除いたときの最小値または最大値まで引いている。



【注意】四分位数は，外れ値を除かないすべてのデータの四分位数であり，その値にもとづいて箱をかく。例えば，右の図において，① はすべてを含めた箱ひげ図であるが，外れ値を ○ で示した箱ひげ図は ② のようになる。

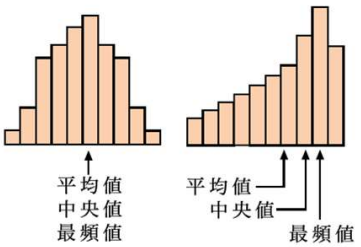


外れ値は，測定ミスや入力ミスによる異常な値であることもある。一方で，外れ値の背景を探ることにより，問題発見があったり，問題解決の手がかりが得られたりすることもある。たとえば，販売員の販売成績を調べたとき，並外れて成績がよい販売員がいたら，その販売員の工夫を探ることで，全体の販売成績を上げる対策を見いだせる可能性がある。

深める データの平均値，中央値，最頻値の中で，外れ値の影響を受けやすいものはどれか，説明してみよう。

たとえば，ある 5 人のテスト結果が 15点，13点，12点，16点，98点のとき，平均値は30.8点，中央値は15点である。98点は他の値より著しく高いため，98点の値を「はずれ値」として除いて考えると，平均値は14点と大きく変化するが，中央値は14点とあまり変化しない。

このことから，平均値の方が中央値よりも「外れ値」からの影響を受けやすいといえる。

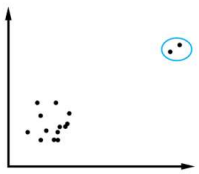


また，平均値，中央値，最頻値は，右の図のように，データが 1 つの山でほぼ左右対称に分布しているとき，互いに近い値となる。

一方，データが 1 つの山で左右対称に分布していないときは，平均値より中央値や最頻値の方が，データの代表的な位置を表すのに適切である。

さらに，相関係数も「外れ値」の影響を受けやすい。

右の散布図で表されたデータの相関係数は0.885であるが，円の中の 2 つの「はずれ値」を除いて相関係数を求めると0.023となる。



このため，相関係数を考えるときは相関係数だけでなく，散布図も調べることが重要である。