神奈川県高等学校教科研究会
数学部会 講演会 資料
2011 年 1 月 12 日

数学Ⅰ『データの分析』について

～四分位数、箱ひげ図など～

竹田　裕一[1]

新聞・雑誌やインターネット上に様々な指標や調査結果（株価指数や世論調査など）があるが、その結果を正しく理解するためには統計の知識が必要となる。このため新学習指導要領では、小学校・中学校・高等学校に資料の活用（統計）が必修として入ることになった。

今回の講演会では数学Ⅰの「資料の活用：データの分析」の内容のうち、従来の数学B・Cに含まれない「四分位数」,「四分位偏差」,「箱ひげ図」を中心に解説を行う。

1　中学で新しく教える範囲

現在の学習指導要領において数学Bおよび数学Cで学習する内容の一部が中学の数学で学習することになった。数学Ⅰで新しく教える「四分位数」などは、この内容を踏まえているので復習を兼ねて定義などを簡単に解説しておく。

1.1　分布を代表する統計量と散らばり具合を表す統計量（数学B ⇒ 中学１年）

分布を代表する統計量として３つの代表値「**平均 (mean)**」,「**中央値 (median)**」,「**最頻値 (mode)**」がある。これらの違いを簡単に解説すると

・平均値 (mean)

数値的な中心を表す代表値で、すべてのデータの値 $(x_1, x_2, \cdots, x_n)$ を足しその個数 (n) で割った値である。

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

・中央値 (median)

確率的な中心を表す代表値で、大きさの順に並べたとき中央にくる値である。
偶数個の場合は、中央に並ぶ値の平均値を中央値とする。

・最頻値 (mode)

頻度に関する代表値で、頻度が最大（最もよく現れる）となる値である。
通常は度数分布表に整理したときに、最大度数をもつ階級値を使う。

また平均値からの散らばり具合を表す統計量として「**分散 (variance)**」がある。一般に分散の値が小さいほど平均値の近くにデータがあることを示している。各データの値 x_i の値と平均値 $\bar{x}$ の差を**x_iの偏差**とよび、**分散** (s^2) と**標準偏差** (s) の定義は次のとおり。

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

[1]神奈川工科大学　基礎・教養教育センター
E-mail： y-takeda@ctr.kanagawa-it.ac.jp

1.2　母集団と標本（数学Ｃ ⇒ 中学３年）

調査対象の性質・特徴などを調査するときに、調査対象すべて（**母集団**）を調べることが可能であればすべてのデータを調査する（**全数調査**）ことによって正確な結果を得ることができる。しかしながら、全数調査が物理的な問題（無限個ある場合や破壊検査など）や時間・費用の問題で実施できないことが多い。その場合、母集団の一部分（**標本**）を調査する（**標本調査**）ことによって、性質・特徴を推測することになる。正しい推測をするためには、母集団を正しく定義することと、母集団から偏りなく標本を取り出すことが重要となる。

例．選挙

選挙は１票でも多い人が当選するので必ず全数調査が行われている。しかしながら、選挙速報の「当選確実」は出口調査などの標本調査から推測された結果である。推測であるために、当選確実を間違えることや、票差が僅差であるとき、当選確実が出ないことがある。

調査事項：誰が当選するか
母集団：投票した人すべて
標　本：ある投票所で投票した人の一部
（いわゆる出口調査）

選挙の場合、母集団を有権者すべてと考える場合もあるが、当選者を決めるのは投票数であるから投票した人のみを母集団と考えればよい。

前節で解説した代表値は、母集団の真の値（**母平均**や**母分散**）ではなく標本を使って推定した値（**標本平均**や**標本分散**）である。推定値の計算方法はいろいろな種類があり、それぞれ異なる基準に従って作られたものである。母平均 μ の推定量である標本平均 $\bar{x}$ は前節で定義した式が使われることが多いが、母分散 σ^2 の推定量である標本分散 s^2 は多くの本で２つの推定量が示されている。

不偏推定量（数値的に近い）

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

最尤推定量（確率的にあてはまりが良い）

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

一般に母分散 σ^2 の推定量は数値的に近い不偏推定量が使われることが多い。Excel で分散の値を求める場合、不偏推定量を VAR、最尤推定量を VARP で求めることができる（正確には、関数 VARP は与えられたデータが母集団すべてである場合の母分散の値を求めている）。

2　高等学校で新しく教える範囲

新学習指導要領において、数学 I において学習することになった「データの分析」の内容は、現在の数学Ｂの「統計とコンピュータ」をほぼそのまま移行した形になっている。現在の内容からの変更点は、いくつかの用語の変更（資料 ⇒ データ，相関図 ⇒ 散布図）と、代表値の内容が中学に移ったことから、新しい内容として「**四分位数**」,「**四分位範囲**」,「**箱ひげ図**」が加えられた点にある。「四分位数」や「四分位範囲」は分散や標準偏差と同様に分布の散らばり具合を表す統計量であり、「箱ひげ図」は「四分位数」を使って、分布の散らばり具合をグラフ化し直感的にわかりやすくしたものである。次節以降で「四分位数」,「四分位範囲」,「箱ひげ図」の解説と、平均や分散との違いについて述べる。

2.1　四分位数と四分位範囲

四分位数はデータを大きさの順に並べたとき、４つに分割した所にある３つの数であり、値の小さなほうから**第１四分位数**，**第２四分位数**，**第３四分位数**とよぶ。第２四分位数についてはデータを大きさの順に並べて中央に来る値であるから中央値と一致している。また、データの散らばり具合を第１四分位数，第２四分位数，第３四分位数と最大値，最小値の５点で表す方法があり、**５数要約**とよぶ。ちなみに、四分位数と同じ考え方で**三分位数**や**五分位数**なども定義可能である。

2.1.1　四分位数の求め方

四分位数は次のように求める。

１）データを大きさの順に並べ、中央値を求め、第２四分位数とする。

２）中央値のデータを含まない中央値以下のデータのみで中央値を求め、第１四分位数とする。

３）中央値のデータを含まない中央値以上のデータのみで中央値を求め、第３四分位数とする。

データの個数が偶数の場合は、中央値のときと同様に中央に並ぶ値の平均値を中央値とする。

例１．次のような１５個のデータの場合

順位	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
データ	11	11	12	16	16	16	18	18	19	20	20	20	22	23	24

１）第２四分位数 (中央値) を求める

８番目が中央の値なので第２四分位数は 18

２）第１四分位数を求める

１～７番目のデータの中央値を求めればよい。

４番目が中央の値なので第１四分位数は 16

３）第３四分位数を求める

９～１５番目のデータの中央値を求めればよい。

１２番目が中央の値なので第３四分位数は 20

最小値	第１四分位数	第２四分位数	第３四分位数	最大値
11	16	18	20	24

例２．次のような８個のデータの場合

順位	1	2	3	4	5	6	7	8
データ	10	11	11	13	14	18	20	21

１）第２四分位数 (中央値) を求める

４番目と５番目が中央の値なので第２四分位数は $\frac{13+14}{2} = 13.5$

２）第１四分位数を求める

１～４番目のデータの中央値を求めればよい。

２番目と３番目が中央の値なので第１四分位数は $\frac{11+11}{2} = 11$

３）第３四分位数を求める

５～８番目のデータの中央値を求めればよい。

６番目と７番目が中央の値なので第３四分位数は $\frac{18+20}{2} = 19$

最小値	第１四分位数	第２四分位数	第３四分位数	最大値
10	11	13.5	19	21

2.1.2　四分位範囲の求め方

四分位範囲はデータの中心付近の５０％が含まれる区間の大きさを示す。

四分位範囲＝第３四分位数－第１四分位数

また、四分位範囲の半分の大きさ**四分位偏差**とよぶ。

四分位偏差＝四分位範囲÷２

この他、データの１００％が含まれる区間の大きさを示す**範囲**がある。

範　囲　＝最大値－最小値

実際に前節の例の場合のそれぞれの値を求めると

例１．の場合

範囲	四分位範囲	四分位偏差
$24-11=13$	$20-16=4$	$4\div 2=2$

例２．の場合

範囲	四分位範囲	四分位偏差
$21-10=11$	$19-11=8$	$8\div 2=4$

ここで定義した四分位数や四分位範囲はあくまでも真の値を推定した値であり、1.2 節でも述べたように推定値の計算方法はいろいろな種類があり、それぞれ値が異なる。そのため、ソフトによってはここで定義した値と異なる値が出てくることがある。実際 Excel で四分位数を求める関数 QUARTILE を使って四分位数を求めると、ほとんど同じ値が得られるが、例２の第３四分位数のみ 19 ではなく 18.5 が得られる。

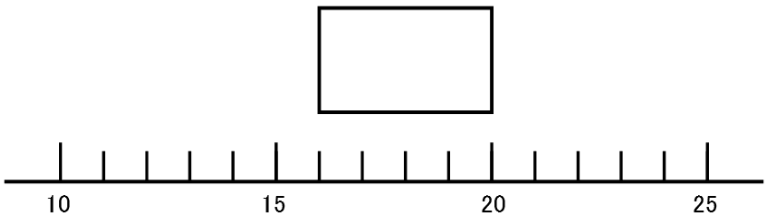
2.2　箱ひげ図のかき方

箱ひげ図は５数要約を元に次のようにしてかく。

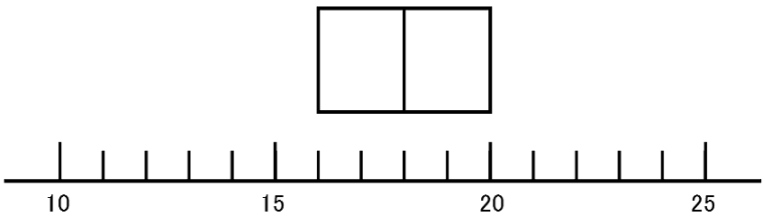
例１の場合

１）第１四分位数と第３四分位数を両端とした箱（長方形）をかく。

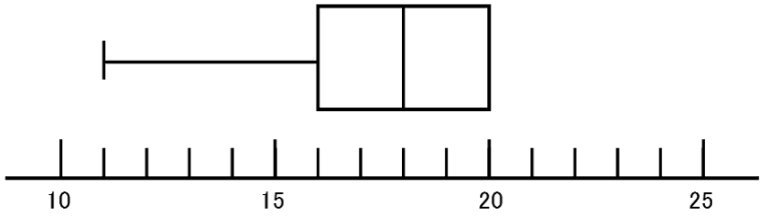
（長方形の幅が四分位範囲に対応する）



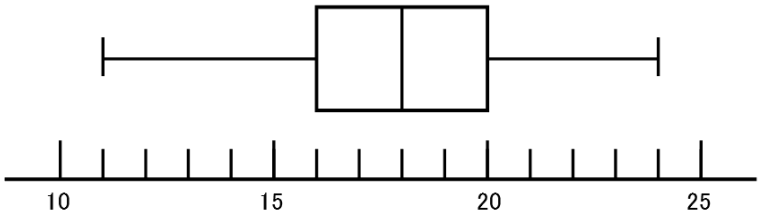
２）第２四分位数（中央値）に線を引く。



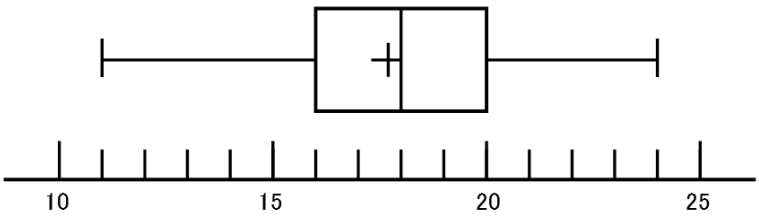
3）最小値に線を引き、箱の左側までひげ（線）をかく。



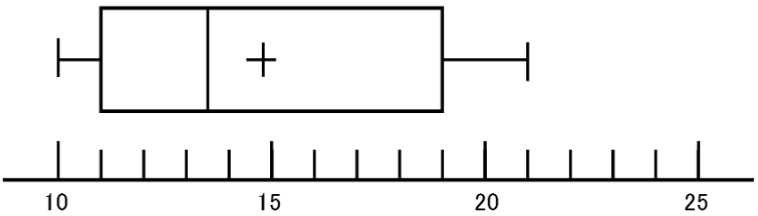
4）最大値に線を引き、箱の右側までひげ（線）をかく。



一般に箱ひげ図は4）のような形になるが、場合によっては平均値 (17.73) に印をつけることがある。



例２の場合の箱ひげ図をかくと次のようになる。



上記の箱ひげ図は横方向であるが、株価のローソク足のような縦方向の箱ひげ図をかくこともある。

箱ひげ図を見ると、「データが多く集まっている所」や「分布が左右対称」などがわかる。実際に例１はほぼ左右対称になっていて、中央値を含む５０％のデータは中央値の近くにまとまっていることがわかる。例２は前半５０％のデータは 10 から 13.5 の狭い範囲にあるが、後半５０％のデータが 13.5 から 21 と範囲の幅が倍以上になっていることから、明らかに左右対称ではない（右側のすそが長い分布）ことがわかる。このように箱ひげ図を見るとある程度ヒストグラムの形も予測できる。

演習．次のデータの５数要約を求め、箱ひげ図を完成させよ。

番号	1	2	3	4	5	6	7	8	9	10	11	12	13
データ	21	20	22	10	32	18	11	13	15	20	24	16	25

(平均値 $\bar{x} = 19$)

3　まとめ

新聞やインターネットに載っている情報の多くは、分布の真ん中を示す値として平均値のみが書かれていることが多く、中央値が書かれていることが少ない。これは母集団分布が左右対称であれば「平均値＝中央値」が成立するために平均値を載せておけば十分だからである。特に母集団分布が正規分布の場合、母平均 μ と母分散 σ^2 がわかれば分布の形が決まってしまうため、平均と分散さえ推定すれば分布の形も推定できる（$[\mu-\sigma, \mu+\sigma]$ の間に６８％，$[\mu-2\sigma, \mu+2\sigma]$ の間に９５％のデータが入る）。

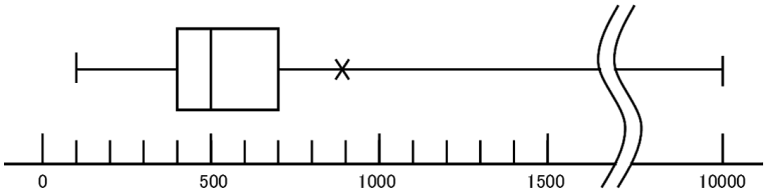
母集団分布が左右対称で無い場合、平均値は直感的な真ん中の値を示していないことがある。例えば所得の例ように一部の人間が大きな収入がある場合、平均値は感覚よりも大きな値となってしまう。この場合平均値よりも中央値の方が感覚と一致している。

例．２０人の年収が次のようなとき、平均値・中央値はいくらになるか。

１００万：	１名	２００万：	１名	３００万：	２名	４００万：	３名
５００万：	４名	６００万：	３名	７００万：	３名	８００万：	１名
１０００万：	１名	１億：	１名				

答）平均値：９９０万円　　中央値：５００万円

上記のデータでは、９割の人が平均値以下の年収となってしまい真ん中の値とはいえない。これは明らかに年収１億円の人がいるために、平均値が上がってしまったためである。このデータに対して箱ひげ図をかくと次のようになる（実際に箱ひげ図をかくときは１億円を**外れ値**として扱うこともある）。



（第１四分位数：４００万円，第３四分位数：７００万円）

箱ひげ図を見ると平均値が箱の中に入っておらず、明らかに右すそが長い分布であることがわかる。実際の年収の分布は厚生労働省が毎年公表しており、「平成２０年 国民生活基礎調査の概況」（http://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa08/index.html）をみると平均値５５６万円に対して中央値４４８万円と１００万円以上の差がある。この他に五分位数が公表されており、

第１：２１０万円　　第２：３６１万円　　第３：５４９万円　　第４：８１４万円

である。詳細なデータが無いためと最大値が大きすぎると考えられるので箱ひげ図をかくことができないが、実際のデータも右すそが長い分布であることがわかる（実際にヒストグラムが載っており、右すそが長い）。そのため平均値のみが一人歩きすると、自分の年収が平均収入よりも低いと嘆く人が多いことになる（平均年収より１００万円以上少ない人が５０％以上いる）。

この例のように母集団分布の左右対称が崩れた場合、平均・分散を使うより中央値や四分位数を使ったほうが適当である。さらに左右対称であるかどうかを判断するためにはヒストグラムや箱ひげ図などのグラフを見ると傾向がわかる。ヒストグラムは階級の幅を変えると、分布の形も変わってしまうことがあるが、箱ひげ図は一意に決まる５数要約を使ってかくので、必ず同じ形となる利点がある。

現在のデータ分析はコンピュータにデータを入力すればいろいろな指標（統計量）やグラフを簡単に得ることができるが、それ故にあきらかに間違った結果や、ミスリードを誘うような結果も散見される。小学校から大学までうまく連携して正しい統計の知識を教えることによって、目先の数字にだまされずに正しい判断ができるような人を育てて行きたいと思っています。