第 4 章

データを 1 つの値にまとめる（記述統計量）

4.1 データを記述する 2 つの方法

データを 1 つの値にまとめるとは，分布の特徴を 1 つの値で代表させる，ということである。このような値を，代表値と呼ぶことにしよう[*1]。代表値は，記述統計量 (descriptive statistics) の 1 つである。

分布の特徴を代表させる値としては，誰でも思いつくだろうが，2 つが考えられる。分布の位置と，分布の広がりである。たとえば，正規分布だったら，$N(\mu, \sigma^2)$ という形で表されるように，平均 μ（ミューと発音する），分散 σ^2 という 2 つの値によって分布が決まるわけだが，この場合，μ が分布の位置を決める情報で，σ^2 が分布の広がりを決める情報である。

一般に，調査データは，仮想的な母集団からの標本（サンプル）[*2]と考えられ，データから計算される代表値は，母集団での分布の位置や広がりを推定するために使われる。その意味で，これらの代表値は母数（parameter）と呼ばれる。分布の位置を決める母数を位置母数 (location parameter)，分布の広がりを決める母数を尺度母数 (scale parameter) と呼ぶ。

分布の位置を示す代表値は central tendency（中心傾向）と呼ばれ，分布の広がりを示す代表値は variability（ばらつき）と呼ばれる。

本章で用いる例題は，Grimm (1993) の第 3 章と第 4 章から引用してアレンジした

*1 たんに代表値と言った場合は分布の位置を指すことが多いが，ここではもう少し広い意味で用いる。

*2 サンプリング理論については，統計学というよりは調査法や実験計画法の範疇になるので，それらの成書を参照されたい。

ものが多い。代表値のような基礎的なことについてきちんと説明された教科書は意外に少ない中で，Grimm の本は丁寧に書かれていて，名著といってよい。英文も平易なのでお薦めする。

4.2　中心傾向（Central Tendency）

4.2.1　平均（mean）

平均は，分布の位置を示す指標として，もっとも頻繁に用いられる。実験的仮説検証のためにデザインされた式の中でも，頻繁に用いられる。記述的な指標の1つとして，平均は，いくつかの利点と欠点をもっている。日常生活の中でも平均をとるという操作は普通に行われるから説明不要かもしれないが，数式で書くと以下の通りである。

母集団の平均 μ）は，

$$\mu = \frac{\sum X}{N}$$

である。X はその分布における個々の値であり，N は値の総数である。$\sum$（シグマと発音する）は，一群の値の和を求める記号である。すなわち，$\sum X = X_1 + X_2 + X_3 + ... + X_N$ である。

標本についての平均を求める式も，母集団についての式と同一である。ただし，数式で使う記号が若干異なっている[*3]。標本平均 $\bar{X}$（エックスバーと発音する）は，

$$\bar{X} = \frac{\sum X}{n}$$

である。n は，もちろん標本サイズである[*4]。

例題1

値が $\{5, 8, 10, 11, 12\}$ である母集団の平均はいくらか？

値が5つしかない母集団というものは想像しにくいかもしれないが，$\mu = (5 + 8 + 10 + 11 + 12)/5 = 9.2$ であることは，小学生でもわかるだろう。R で平均を計算するには，`mean()` という関数を使う。たとえば，例題1の解を得るには，`mean(c(5,8,10,11,12))` とすればよい。

[*3] 一般に母集団についての統計量を示す記号にはギリシャ文字を使うことになっている。

[*4] 記号について注記しておくと，集合論では $\bar{X}$ は集合 X の補集合の意味で使われるが，代数では確率変数 X の標本平均が $\bar{X}$ で表されるということである。同じような記号が別の意味で使われるので混乱しないように注意されたい。補集合は X^C という表記がなされる場合も多いようである。標本平均は $\bar{X}$ と表すのが普通である。

さて，本章で取り上げる中心傾向には，平均の他に，あと2つ，中央値 (median) と最頻値 (mode) がある。どれも分布の中心の位置がどの辺りかを説明するものだが，中心性 (centrality) へのアプローチが異なっている。

平均は，中心性を示すために，どんなやり方をとっているのだろうか？ たまたまその値が平均と同じであったという希な値を除けば，各々の値は，平均からある距離をもって存在する。言い換えると，各々の値は，平均からある程度の量，ばらついている。ある値が平均から離れている程度は，単純に $X - \bar{X}$ である。この，平均からの距離を，偏差（あるいは誤差）といい，x という記号で書く。つまり，$x = X - \bar{X}$ である。次の例を見ればわかるように，偏差は正の値も負の値もとるが，その合計は0になるという特徴をもつ。どんな形をしたどんな平均のどんなに標本サイズが大きいデータだろうと，偏差の和は常に0である。式で書くと，$\sum x = \sum(X - \bar{X}) = 0$ ということである。言い方を変えると，偏差の和が0になるように，平均によって調整が行われたと見ることもできる。平均は，この意味で，分布の中心であるといえる。

例題 2

標本 A が $\{2, 4, 6, 8, 10\}$ という5つのデータからなり，標本 B が $\{2, 4, 6, 8, 30\}$ という5つのデータからなるとき，A の標本平均は6であるから，それぞれの値の偏差は $\{-4, -2, 0, 2, 4\}$ となり，その合計は0である。B についても確かめよ。

標本平均は (2+4+6+8+30)/5=10 で，それぞれの値の偏差は $\{-8, -6, -4, -2, 20\}$ となるので，確かにその合計は0となる。分布 B は分布 A よりも平均が大きい。

4.2.2 重み付き平均 (weighted mean)

重み付き平均は，各々の値にある重みをかけて合計したものを，重みの合計で割った値である。式で書くと，

$$\bar{X} = \frac{n_1(\bar{X}_1) + n_2(\bar{X}_2) + \ldots + n_n(\bar{X}_n)}{n_1 + n_2 + \ldots + n_n}$$

ここでは標本サイズが異なる複数の平均の総平均 (grand mean) を計算する場合について説明する。

Y大学の3つの学部の学生が TOEIC (Test of English for International Communication：国際コミュニケーションのための英語テスト) を受験したところ，学部ごとの得点の平均がA学部 440 点，B学部 470 点，C学部 410 点だったとしよう。これらの値からY大学全体の TOEIC の平均点を求めたいときは，どうしたらよいだろうか？ 単純にこれらを足して3で割った 440 点としていいのだろうか？

大学全体としてのTOEICの平均は，3つの学部のどこに所属する学生であるかにかかわりなく，全員の得点を足して，その人数で割って得るべきものである。そうだとすれば，単純に3つの値を足して3で割るのでは具合が悪い。各学部の人数は異なるので，人数の多い学部の得点の方が，総平均には余計に影響するだろうからだ。こういう場合は，各学部の人数をそれぞれの平均点に掛けて（つまり各学部の得点総和に戻して）足し合わせ，それを人数の和（つまり大学全体の人数）で割れば良いことが直感的にわかるだろう。これはまた，各学部の人数の大学全体の人数に占める割合を重みとして各学部の得点に掛け，それを足し合わせることと同値でもある。

例題3

TOEICの平均点が $\{440, 470, 410\}$ であった3つの学部それぞれの人数が $\{200$ 人，100人，300人$\}$ であったなら，この大学のTOEICの総平均は何点か？

$$\frac{200}{(200+100+300)}\times 440+\frac{100}{(200+100+300)}\times 470+\frac{300}{(200+100+300)}\times 410=430$$

より，430点となる。Rで実行するときは，

```
p <- c(440,470,410)
n <- c(200,100,300)
sum(p*n)/sum(n)
```

とするとよい[*5]。

[*5] 最後の行は `sum(p*n/sum(n))` でもよい。

練習問題

3つの年齢群ごとの平均血圧が下の表のように記録されているとき，すべての年齢群をプールした，血圧の総平均を求めよ[a]。

	年齢		
	20-39	40-59	60+
収縮期血圧 (mmHg)	118	128	145
拡張期血圧 (mmHg)	70	78	82
人数	13	12	16

[a] ただし，血圧の意味合いは年齢によって変わってくるからこそ，ふつう敢えて年齢群別に平均を出すわけだから，年齢群をプールした血圧の平均を出すことには，あまり意味はない。ここは単なる計算練習だと思って欲しい。また，ここで述べたような意味での重み付き平均を計算する必要があるのは，集計済みの二次資料から指標値を再計算するような場合なので，生データがあれば生データから計算すれば済むことである。

収縮期血圧の総平均は，$(118\times13+128\times12+145\times16)/(13+12+16)=131$ より，131 mmHg となり，拡張期血圧の平均は，$(70\times13+78\times12+82\times16)/(13+12+16)=77$ より，77 mmHg となる。これも R で実行するときは，例題3と同様に，

```
SBP <- c(118,128,145)
DBP <- c(70,78,82)
n <- c(13,12,16)
gSBP <- sum(SBP*n)/sum(n)
gDBP <- sum(DBP*n)/sum(n)
cat("SBP 総平均=",gSBP,", DBP 総平均=",gDBP,"\n")
```

とするとわかりやすい。

4.2.3 度数分布の平均

度数分布の平均も，重み付き平均に似た概念である。離散変数の平均の場合に，度数分布を出して，各値にその度数を掛けたものの和を度数の総和で割ることで得られる。これは，言い換えると，度数で重み付けした平均である。

$$\mu=\frac{\sum Xf}{\sum f}$$

という式になる。

平均は，例題2を見ればわかるように，少数の極端な値の影響を受けやすいという

欠点をもつ。1つだけ極端な値があったからといって，あまりに値がそちらに引っ張られてしまっては，分布の位置を代表する値としては具合が良くない[*6]。

例題 4

A大学の学長選挙で，B氏が，A大学の研究水準を上げるという公約を掲げて当選したとしよう。4年後の次の選挙のときに，B氏は自分が公約を果たしたと宣伝したいわけだが，彼の定義によると，大学の研究水準が上がるとは，教員の論文数の平均が増えるということである。ところで，B氏が当選した当時の教員数は100人いて，そのうち発表論文数が5本の人が80人，10本の人が15人，30本の人が5人いたとしよう。この時点での平均論文数は $(5\times 80+10\times 15+30\times 5)/100=7$ なので7本である。その後4年間誰も1本も論文を書かなかったとしても，2年目にたまたま2330本の論文をもつ教員が1人着任したら，何が起こるかを考えてみよう。

平均論文数は $(5\times 80+10\times 15+30\times 5+2330)/101=30$ から，30本となってしまう。そこで，B氏は，大威張りで，任期中に平均論文数は4倍以上に増えたと報告することができる。元々A大学にいた教員の論文数はまったく変わらず，したがってたいした研究環境を提供できていないと思われるにもかかわらず，である。B氏が公約を果たしたと宣伝しても嘘ではないことになるが，何か妙である。

例題4は，極端に高い値が，平均を高く押し上げてしまったという例である。分布の位置の指標としては，極端な外れ値に対してこんなに敏感であっては具合が良くない。こういう極端な値が含まれている歪んだ分布の場合には，平均という指標は誤解を生んでしまうので，相応しくないことになる。

4.2.4　中央値（median）

そこで登場するのが**中央値**である。中央値は，全体の半分がその値より小さく，半分がその値より大きい，という意味で，分布の中央である。言い換えると，中央値は，頻度あるいは値の数に基づいて分布を2つに等分割する値である。中央値を求めるには式は使わない（決まった手続き＝アルゴリズムとして，並べかえ (sorting) は必要）。極端な外れ値の影響を受けにくい（言い換えると，外れ値に対して頑健である）。歪んだ分布に対するもっとも重要な central tendency の指標が中央値である。

---

*6 その1つが，実は測定ミスであったり，異質な対象だったりして，外れ値である場合もあり，その場合は平均の計算に入れないこともある。あまり機械的にやるのは良くないが，ネイマンの外れ値の検定を使うのも一案である。

例題 5

次の分布の中央値は何か？ {1, 4, 6, 8, 40, 50, 58, 60, 62}

この場合，小さい方から数えても大きいほうから数えても 5 番目の値である 40 が中央値であることは自明である。次に小さい値である 50 との距離や次に大きい値である 8 との距離は中央値を考える際には無関係である。中央値を求めるには，値を小さい順に並べかえて[*7]，ちょうど真中に位置する値を探せばよい。この意味で，中央値は値の順序だけに感受性をもつ（= rank sensitive である）といえる[*8]。

R で中央値を計算するには，`median()` という関数を使う。たとえば，例題 5 の解を得るには，`median(c(1, 4, 6, 8, 40, 50, 58, 60, 62))` とすればよい。

例題 6

次の標本分布の平均と中央値は何か？ {2, 4, 7, 9, 12, 15, 17}

R で

```
x <- c(2,4,7,9,12,15,17)
mean(x)
median(x)
```

とすると，平均は約 9.43，中央値は 9 であるとわかる。

例題 7

次の標本分布の平均と中央値は何か？ {2, 4, 7, 9, 12, 15, 17, 46, 54}

例題 6 と同様に計算すると，平均は 18.4，中央値は 12 となる。例題 6 に比べると，右側に 2 つの極端な値を加えただけだが，平均はほぼ倍増してしまう。それに対して，中央値は 1 つ右側の値に移るだけであり，中央値の方が極端な値が入ることに対して頑健といえる。

ところで，値の数が奇数だったら，このように順番が真中というのは簡単に決められるが，値が偶数個だったらどうするのだろうか？

例題 8

次の分布の中央値は何か？ {4, 6, 9, 10, 11, 12}

[*7] 値の数が少ない場合には，手作業で並べかえを行えばよいが，大量のデータを手作業で並べかえるのは大変である。コンピュータのプログラムに値を並べかえさせるアルゴリズムには，単純ソート，バブルソート，シェルソート，クイックソートなどがある。

[*8] 平均は値の大きさによって変わるので，value sensitive であるといえる。

中央値が9と10の間にくることは明らかである。そこで，普通は9と10を平均した9.5を中央値として使うことになっている。もっとも，本来整数値しかとらないような値について，中央値や平均として小数値を提示することに意味があるかどうかは問題である。たとえば，例題8の分布が，ある地方の水泳プールで6日間観察したときの，1日当たりの飛び込みの回数を示すものだとしよう。中央値が9.5ということになると，9.5回の飛び込みというのは何を表すのか？　半分だけ飛び込むということはありえない。つまり実体はない，単なる指標値だということになる。同様に平均についても，世帯当たりの平均子ども数が2.4人とかいうとき，0.4人の子どもは実体としてはありえない。しかし，分布の位置を示す指標としては有用なので，便宜的に使っているのである。

例題9

次の分布の中央値は何か？　｛7, 7, 7, 8, 8, 8, 9, 9, 10, 10｝

例題10

次の分布の中央値は何か？　｛7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 10, 10｝

このように同順位の値（tieという）がある場合は，事態はやや複雑である。順番で言えば，例題9でも例題10でも中央値は8と8の間に来るはずだから，8と思うであろう。実際，SAS，SPSSなどの有名ソフトを初めとして，Microsoft ExcelやRに至るまで，ほぼすべての統計ソフトは，8という答えを出してくるし，一般にはそれで問題ない。

さて，もう1歩進めて，度数分布表から中央値を計算する場合を考えてみよう。ちょっと複雑だが，理解するのは難しくない。下表は，年齢階級ごとの人数の分布であり，これから年齢の中央値を求める方法を考えることにする。

同順位の値の注意

ただし，厳密に考えると，簡単に 8 と言えない。Grimm (1993) が指摘するように，分布の値を示す数値は，間隔の中点と考えるべきだからである。普通はそこまで厳密に考える必要はないが，参考までに説明しておこう。

要点は，『それぞれの値を，表示単位によって規定される区間の中点と考え，同順位の値があるときは，それが区間内に均等に散らばると考える』ということである。これは直感的に考えても合理的であろう。

たとえば，1 1 1 2 2 2 3 3 3 という，表示単位 1 のデータがあるとき，真の値がそれぞれ等間隔に散らばっているならば，0.67 1.00 1.33 1.67 2.00 2.33 2.67 3.00 3.33 と考えるのが自然である。これなら，それぞれの値が 1/3 間隔になっているし，中点 1 で示される値 0.67 1.00 1.33 の平均は 1 となるので，どこにも矛盾がない。

この例から帰納的に考えて，その区間の下限の値を L とし，階級幅を h とし，同順位の個数を fm 個とし，1 つ下の区間までに F 個のサンプルがあるとすれば，F+1 番目，F+2 番目，...，F+fm 番目 の値はそれぞれ，L+1/(2fm)*h，L+3/(2fm)*h，...，L+(2fm-1)/(2fm)*h となる。つまり，**F+x 番目の値は，L+(2x-1)/(2fm)*h** となる。この式から例題 9 の 3 つの 8 の真の値がいくつになるか計算すると，

4 番	5 番	6 番
7.67	8.00	8.33

となって，5 番と 6 番の間は 8.17 となる。

同じく例題 10 で真の値を計算すると，|6.67 7.00 7.33 7.60 7.80 8.00 8.20 8.40 8.75 9.25 9.75 10.25| となるので，中央値は 8.00 と 8.20 の間で 8.10 となる。|1 1 2 2 3 3| という表示単位 1 のデータでは，真の値は |0.75 1.25 1.75 2.25 2.75 3.25| と推定されるので，中央値は 1.75 と 2.25 の平均で 2 となる。

年齢階級	度数	累積度数
45-49	1	76
40-44	2	75
35-39	3	73
30-34	6	70
25-29	8	64
20-24	17	56
15-19	**26**	**39**
10-14	11	13
5-9	2	2
0-4	0	0

まず，累積度数の最大の数をみる（つまり総数をみる）。この例では 76 である。中

央値の順位は $(76+1)/2 = 38.5$ 位となる[*9]。38.5 番目の値を含む年齢階級を探すと，15–19 である。そこで，単純に統計ソフトが出してくる中央値は 15-19 歳となる[*10]。

5 歳の階級幅の中のどこに中央値があるのかというところまで推定しようとなると，もう少し厳密に考えねばならなくなる。つまり，Grimm 流に 15-19 歳の 26 人の値が均等に散らばっていると考えると，{14.5+5/52, 14.5+15/52, 14.5+25/52, ..., 14.5+245/52, 14.5+255/52} となるから，38.5 位の値は，最後の 2 つの平均をとって，$14.5 + (245 + 255)/104 \approx 19.3$ から約 19.3 歳となる。

このやり方は，中央値が正確な分布の中央（少なくともその近似）になっているという特性を強めるものである。式で書けば，中央値は，

$$L + \left[\frac{N/2 - F}{f_m} \cdot h\right]$$

となる。ここで，L は中央順位を含む階級の正確な下限，F は中央順位を含む階級より下の値の総度数，f_m は中央順位を含む階級の度数，h は階級幅である。

この式は以下のように導かれる。

1. サンプル数 N が奇数のとき，$(N+1)/2$ 番目が中央値なので，$F+x = (N+1)/2$ を x について解いて $L + (2x-1)/(2f_m) * h$ に代入すれば，

   $$L + (N + 1 - 2F - 1)/(2f_m) * h = L + (N/2 - F)/f_m * h$$

   となる。
2. N が偶数のとき，中央値は $N/2$ 番目と $N/2+1$ 番目の間なので，$F+x = N/2$ と $F + x = N/2 + 1$ を x について解いて $L + (2x-1)/(2f_m) * h$ に代入した

   $$L + (2(N/2 - F) - 1)h/(2f_m)$$

   と

   $$L + (2(N/2 + 1 - F) - 1)h/(2f_m)$$

   の平均となって，やはり

   $$L + (N/2 - F)/f_m * h$$

   で良いことになる。

[*9] Grimm (1993) には 76 を 2 で割って 38 番目の値が中央値であると書かれているが，論理的整合性を欠く。もし総数を 2 で割った順位の値が中央値だとすると，例題 8 の答えが下から 3 番目で 9 ということになってしまう。総数に 1 を加えて 2 で割る方が論理的整合性が高い。

[*10] 繰り返すが，普通はこの解で問題ない。

4.2.5 最頻値（Mode）

残る最頻値は，きわめて単純である。もっとも度数が多い値を探すだけである。もっとも数が多い値が，もっとも典型的だと考えるわけである。データを見ると，最頻値が2つある場合があり，この場合は分布が二峰性 (bimodal) だという[*11]。すべての値の出現頻度が等しい場合は，最頻値は存在しない。

分布の形によって，平均，中央値，最頻値の関係は変わってくる。歪んでいない分布ならば，ばらつきの程度によらず，これら3つの値は一致する。二峰性だと最頻値は2つに分かれるが，平均と中央値はその間に入るのが普通である。左すそを引いた分布では，平均がもっとも小さく，中央値が次で，最頻値がもっとも大きくなる。右すそを引いた分布では逆になる。

平均は，(1) 分布のすべての値を考慮した値である，(2) 同じ母集団からサンプリングを繰り返した場合に一定の値となる，(3) 多くの統計量や検定で使われている，という特長をもつ。標本調査値から母集団の因果関係を推論したい場合に，もっとも普通に使われる。しかし，(1) 極端な外れ値の影響を受けやすい，(2) 打ち切りのある分布では代表性を失う場合がある[*12]，という欠点があり，外れ値があったり打ち切りがあったりする分布では位置の指標として中央値の方が優れている。最頻値は，標本をとったときの偶然性の影響を受けやすいし，もっとも頻度が高い値以外の情報はまったく使われない。しかし，試験の点で何点の人が多かったかを見たい場合は最頻値が役に立つし，名義尺度については最頻値しか使えない。

ここで上げた3つの他に，幾何平均 (geometric mean) や調和平均 (harmonic mean) も，分布の位置の指標として使われることがある。幾何平均はデータの積の累乗根（対数をとって平均を出して元に戻すことと同値），調和平均はデータの逆数の平均の逆数であり，どちらもゼロを含むデータには使えない。大きな外れ値の影響を受けにくいという利点があり，幾何平均は，とくにデータの分布が対数正規分布に近い場合によく用いられる。

[*11] しかし隣り合う2つの値がともに最頻値である場合は二峰性だとはいわず，離れた2つの値が最頻値あるいはそれに近い場合，つまり度数分布やヒストグラムの山が2つある場合に，分布が二峰性だといい，2つの異なる分布が混ざっていると考えるのが普通である。

[*12] 氷水で痛みがとれるまでにかかる時間とか，年収とか。無限に観察を続けるわけにはいかないし，年収は下限がゼロで上限はビル・ゲイツのそれのように極端に高い値があるから右すそを長く引いた分布になる。平均年収を出している統計表を見るときは注意が必要である。年収の平均的な水準は中央値で表示されるべきである。

4.3 ばらつき（Variability）

分布を特徴付けるには，分布の位置だけではなく，分布の広がり具合の情報も必要である。たとえば，図 4.1 の 2 つの分布は，次に示す R で描いた

```
x <- c(1:1000)/100-5
z1 <- dnorm(x,0,1)
z2 <- dnorm(x,0,4)
plot(x,z1,type='l',lty=1,ylab='probability density',xlab='')
points(x,z2,type='l',lty=2)
```

ものだが，どちらも平均 0 の正規分布なので中央値も最頻値も共通だが，実線で書かれた幅が狭い方が標準偏差 1，破線で書かれた幅が広い方が標準偏差 4 と，標準偏差が大きく異なるために，まったく違った外見になっている。標準偏差は，もっとも良く使われる分布の広がり具合の指標である。

広がり具合を示す指標は，ばらつき (variability) と総称される。ばらつきの指標には，範囲，四分位範囲，四分位偏差，平均偏差，分散（及び不偏分散），標準偏差（及び不偏標準偏差）がある。

4.3.1 範囲 (range)

範囲は，もっとも単純なばらつきの尺度である。値のとる全範囲そのものである。つまり，最大値から最小値を引いた値になる。

例題 11

次の分布の範囲はいくらか？ $\{17, 23, 42, 44, 50\}$

いうまでもなく，$50 - 17 = 33$ である。ばらつきの尺度として範囲を使うには，若干の問題が生じる場合がある。極端な外れ値の影響をダイレクトに受けてしまうのである。次の例を考えてみよう。

例題 12

次の分布の範囲はいくらか？ $\{2, 4, 5, 7, 34\}$

答えは $34 - 2 = 32$ なのだが，2, 4, 5, 7 というきわめて近い値 4 つと，かけ離れて大きい 34 という値からなるのに，32 という範囲は，全体のばらつきが大きいかのような誤った印象を与えてしまう。ばらつきの指標としては，分布の端の極端な値の

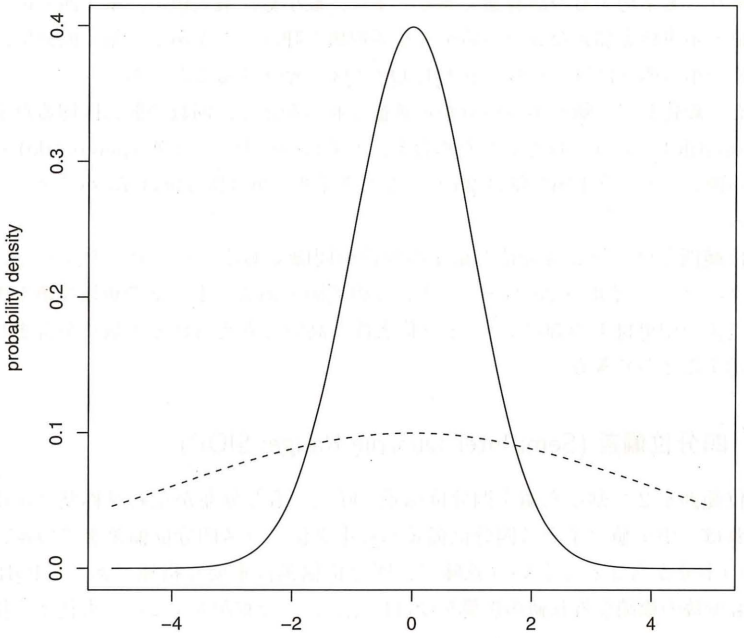

図 4.1 平均が同じで標準偏差が異なる正規分布

影響を受けにくい値の方がよい。

4.3.2 四分位範囲 (Inter-Quartile Range; IQR)

そこで登場するのが四分位範囲である。その前に，分位数について説明しよう。値を小さい方から順番に並べかえて，4 つの等しい数の群に分けたときの 1/4, 2/4, 3/4 にあたる値を，四分位数 (quartile) という。1/4 の点が第 1 四分位，3/4 の点が第 3 四分位である (つまり全体の 25 %の値が第 1 四分位より小さく，全体の 75 %の値が第 3 四分位より小さい)。2/4 の点というのは，ちょうど順番が真中ということだから，第 2 四分位は中央値に等しい。

ちょっと考えればわかるように，ちょうど4等分などできない場合がもちろんあって，上から数えた場合と下から数えた場合で四分位数がずれる可能性があるが，その場合はそれらを平均するのが普通である。また，最小値，最大値に，第1四分位，第3四分位と中央値を加えた5つの値を五数要約値と呼ぶことがある。第1四分位，第2四分位，第3四分位は，それぞれ Q1, Q2, Q3 と略記することがある。

これを一般化して，値を小さい方から順番に並べかえて，同数の群に区切る点を分位数 (quantile) という。百等分した場合を，とくにパーセンタイル (percentile) という。言い換えると，第1四分位は25パーセンタイル，第3四分位は75パーセンタイルである。

四分位範囲とは，第3四分位と第1四分位の間隔である。パーセンタイルでいえば，75パーセンタイルと25パーセンタイルの間隔である。上と下の極端な値を排除して，全体の中央付近の50％（つまり代表性が高いと考えられる半数）が含まれる範囲を示すことができる。

4.3.3 四分位偏差 (Semi Inter-Quartile Range; SIQR)

四分位範囲を2で割った値を四分位偏差と呼ぶ。もし分布が左右対称型の正規分布であれば，中央値マイナス四分位偏差から中央値プラス四分位偏差までの幅に全データの半分が含まれるという意味で，四分位偏差は重要な指標である。IQR も SIQR も少数の極端な外れ値の影響を受けにくいし，分布が歪んでいても使える指標である。

例題 13

パプアニューギニアのある村で成人男性28人の体重を量ったところ，{50.5, 58.0, 47.5, 53.0, 54.5, 61.0, 56.5, 65.5, 56.0, 53.0, 54.0, 56.0, 51.0, 59.0, 44.0, 53.0, 62.5, 55.0, 64.5, 55.0, 67.0, 70.5, 46.5, 63.0, 51.0, 44.5, 57.5, 64.0}（単位は kg）という結果が得られた。このデータから，四分位範囲と四分位偏差を求めよ。

R の `fivenum()` 関数を使うと，Q1=52.00，Q2=55.50，Q3=61.75 とわかる。これより，四分位範囲は Q3−Q1=9.75，四分位偏差はそれを2で割って4.975である。

4.3.4 平均偏差 (mean deviation)

偏差の絶対値の平均を平均偏差と呼ぶ。四分位範囲や四分位偏差は，全データのうちの限られた情報しか使わないので，分布のばらつきを正しく反映しない可能性があ

る。そこで，すべてのデータを使ってばらつきを表す方法を考えよう。すべての生の値は，平均からある距離をもって分布している。この距離は既に述べたように偏差あるいは誤差と呼ばれる[*13]。偏差の大きさは，分布のばらつきを反映している。

例題 14

分布 A が {11, 12, 13, 14, 15, 16, 17}，分布 B が {5, 8, 11, 14, 17, 20, 23} だとする。どちらも平均は 14 である。しかし，分布 B は分布 A よりもばらつきが大きい。言い換えると，分布 B の方が分布 A よりも平均からの距離が大きい。しかし，それをどうやって 1 つの値として表すことができるだろうか？

ただ合計しただけでは，平均のところで述べたように，偏差の総和は必ずゼロになってしまう。これはマイナス側の偏差がプラス側の偏差と打ち消しあってしまうためなので，偏差の**絶対値**の総和を出してやればいいというのがもっとも単純な発想である。それだけだと標本サイズが大きいほど大きくなってしまうので，値 1 つあたりの偏差の絶対値を出してやるために標本サイズで割ることが考えられる。これが平均偏差の考え方である。

すなわち，平均偏差 MD は，

$$MD = \frac{\sum |X - \bar{X}|}{n}$$

で定義される。$\bar{X}$ は平均，n は標本数である。例題 14 の場合，分布 A の平均偏差は約 1.71，分布 B の平均偏差は約 5.14 である。これらの値は，次の R プログラムによって計算される。

```
A <- c(11, 12, 13, 14, 15, 16, 17)
B <- c(5, 8, 11, 14, 17, 20, 23)
mA <- mean(A)
mB <- mean(B)
sum(abs(A-mA))/NROW(A)
sum(abs(B-mB))/NROW(B)
```

平均偏差はすべてのデータを使い，かつ少数の外れ値の影響は受けにくいという利点があるが，絶対値を使うために他の統計量との数学的な関係がなく，標本データから母集団統計量を推定するのに使えないという欠点がある。

[*13] 誤差の方が意味が広いので，この意味で使う場合は偏差と呼んだ方がいいと思う。

4.3.5 分散 (variance)

マイナス側の偏差とプラス側の偏差を同等に扱うためには，絶対値にするかわりに二乗しても良い。つまり，偏差の二乗和の平均をとるわけである。これが分散という値になる。分散 V は，

$$V = \frac{\sum (X - \bar{X})^2}{n}$$

で定義される[*14]。標本サイズ n で割る代わりに自由度 $n-1$ で割って，不偏分散 (unbiased variance) という値にすると，標本データから母集団の分散を推定するのに使える。すなわち，不偏分散 V_{ub} は，

$$V_{ub} = \frac{\sum (X - \bar{X})^2}{n-1}$$

である。

4.3.6 標準偏差 (standard deviation)

分散の平方根をとったものが標準偏差である。平均と次元を揃えるという意味をもつ。不偏分散の平方根をとったものは，不偏標準偏差となる。もし分布が正規分布ならば，Mean±2SD[*15]の範囲にデータの 95% が含まれるという意味で，標準偏差は便利な指標である。

練習問題

例題 14 の 2 つの分布について，不偏分散と不偏標準偏差を計算せよ。

以下の R プログラムにより，A の不偏分散が 4.67，A の不偏標準偏差が 2.16，B の不偏分散が 42，B の不偏標準偏差が 6.48 だとわかる。

[*14] 電卓などで計算するときは，これを式変形して得られる $V = \sum X^2/n - \bar{X}^2$（2 乗の平均から平均の 2 乗を引く）という形の方が簡単だが，2 つの大きい数の引き算を行うことになって計算上の桁落ちの可能性が増えるので，コンピュータプログラム上は定義式通りに計算すべきである。

[*15] 普通このように 2SD と書かれるが，正規分布の 97.5 パーセント点は 1.959964... なので，この 2 は，だいたい 2 くらいという意味である。

```
A <- c(11, 12, 13, 14, 15, 16, 17)
B <- c(5, 8, 11, 14, 17, 20, 23)
cat("A の不偏分散=",var(A),
" / A の不偏標準偏差=",sd(A),"\n")
cat("B の不偏分散=",var(B),
" / B の不偏標準偏差=",sd(B),"\n")
```

Rのbaseには不偏分散と不偏標準偏差を求める関数しかないです。

4.3.7 標準誤差 (standard error) と変動係数 (coefficient of variation)

生データの分布のばらつきの指標ではないが，関連するのでここで示しておく。不偏標準偏差を標本サイズの平方根 $\sqrt{n}$ で割った値は，平均の推定幅を示す値となり[*16]，標準誤差 (standard error; SE) として知られている。SD と SE を混用している論文も散見されるが，意味がまったく違う。また，標準偏差（不偏標準偏差ではない）を平均で割って 100 を掛けた値を変動係数という。すなわち，平均に対して，全測定値が何 % ばらついているかを示す，相対的なばらつきの指標である。これは測定誤差を示すときなどに使われる値であり，母集団統計量である。

4.4 まとめ

データの分布は，位置とばらつきを示す 2 つの値で代表させるのが普通である。分布に外れ値が多い・歪みが大きい・尺度水準が低いなどの理由で，分布を仮定できない場合は，中央値と四分位偏差を用い，そうでない場合は平均と（不偏）標準偏差を用いて，位置±ばらつき，という形で示すことが多い。

参考までに，次のページに Microsoft Excel と R による代表値の求め方を一覧形式でまとめておくので，必要に応じて参照されたい。

[*16] 平均の分散は生データの分散の $1/n$ になることと，n が大きいとき，元の分布によらず平均は正規分布に近づく（中心極限定理）ため。

求める代表値など	Excelの関数または手順 （範囲A1:Y1にデータがあるとして）	Rの関数または手順 （x <- c(...) などのやり方で変数xにデータを入れたとして）
最頻値	離散データなら=MODE(A1:Y1)で良いが，連続量なら，ツール>分析ツール>ヒストグラムでヒストグラムを書いて最大度数のデータ区間を探し，その区間の中点を最頻値とする。	hist(x) でヒストグラムを書いて最大度数のデータ区間を探し，その区間の中点を最頻値とする。 hist(x,c(min(x),5,8,max(x))) などとすれば，xの最小値から5まで，5から8まで，8からxの最大値までという3つの区間で度数を計算させることができる。本来はhist(x,5) とすれば5つの区間という形の指定ができるはずなのだが，区間の数によってうまくいったりいかなかったりした。 なお，hist(x,plot=F) とすれば，グラフを書く代わりに数値を表示させられる。
中央値	=MEDIAN(A1:Y1)	median(x) ただし，xの中にNA（欠損値）を含む場合は， median(x,na.rm=T) または median(x[!is.na(x)]) とする。以下同様。
平均	=AVERAGE(A1:Y1) 調和平均は=HARMEAN(A1:Y1)，幾何平均は=GEOMEAN(A1:Y1)で求められる。	mean(x) 調和平均は 1/mean(1/x)，幾何平均は exp(mean(log(x))) で求められる。
範囲	=MAX(A1:Y1)−MIN(A1:Y1)	max(x)-min(x) か range(x)
四分位範囲	=QUARTILE(A1:Y1,3) − QUARTILE(A1:Y1,1)	IQR(x) または，y<-quantile(x);y[4]-y[2] または，fivenum(x)[4]-fivenum(x)[2] でも良い。
四分位偏差	=(QUARTILE(A1:Y1,3) − QUARTILE(A1:Y1,1))/2	IQR(x)/2 または，y<-quantile(x);(y[4]-y[2])/2　または，(fivenum(x)[4]-fivenum(x)[2])/2 でも良い。
平均偏差	=AVEDEV(A1:Y1)	組み込み関数にはないが，sum(abs(x-mean(x)))/NROW(x) で得られる。
不偏分散	=VAR(A1:Y1) （不偏でない分散は=VARP(A1:Y1)で得られる）	var(x) 不偏でない分散は組み込み関数にはないが，sum((x-mean(x))^2)/NROW(x) で得られる。

不偏標準偏差	=STDEV(A1:Y1)（不偏でない標準偏差は=STDEVP(A1:Y1) で得られる）	sd(x) 不偏でない標準偏差は，sqrt(sum((x-mean(x))^2)/NROW(x)) で得られる。(*)
タブ区切りデータファイルの読み込み	そのままドラッグ&ドロップ	1行目に変数名が入っているなら，x<-read.delim("d:/sample.dat",header=T) などとする (**)。それぞれの変数は，たとえばx$ageのようにして参照できる。1行目が変数名でなくすぐにデータである場合は，x<-read.delim("d:/sample.dat",header=F) とする。この場合，変数名は x$V1, x$V2, ... として参照できる。いちいち x$とつけるのが面倒なら，attach(x) とすれば V1 とか V2 だけで参照できる。
カンマ区切りデータファイルの読み込み	そのままドラッグ&ドロップ	1行目に変数名が入っているなら，x<-read.csv("d:/sample.dat",header=T) とする (**)。1行目が変数名でなくすぐにデータである場合は，x<-read.csv("d:/sample.dat",header=F) とする。
データの編集	表にそのまま打ち込む	de(x$V1,x$V5) などとすれば表形式で指定した変数の値を編集できる。表の上でマウスを右クリックすると操作メニューがでる。
データの書き出し	ファイルから保存を選ぶ	コンマ区切りでデータフレーム x をマイドキュメントの sample.dat に書き出すには，write.table(x,"d:/sample.dat",sep=",") とする。タブ区切りなら sep="\t"とすればよい。

(*) もちろん，不偏でない分散を出すときに，Vx<-sum((x-mean(x))^2)/NROW(x) などとして値を保存しておいて，sqrt(Vx) とするのがエレガントである。

(**) \を/に置き換えたファイル名をフルパスで書く。ただし，2バイトコードが入ったディレクトリ名やファイル名は，文字化けするので使いにくい（半角英数字のファイル名に書き換えておくべきである)。また，Windows2000 の場合，マイドキュメントフォルダのフルパスは，普通，C:/Documents and Settings/nakazawa/My Documents/のようになるが，長いパスを打つのは面倒なので，D:ドライブのルートディレクトリなどにデータを置くと，指定が容易である。