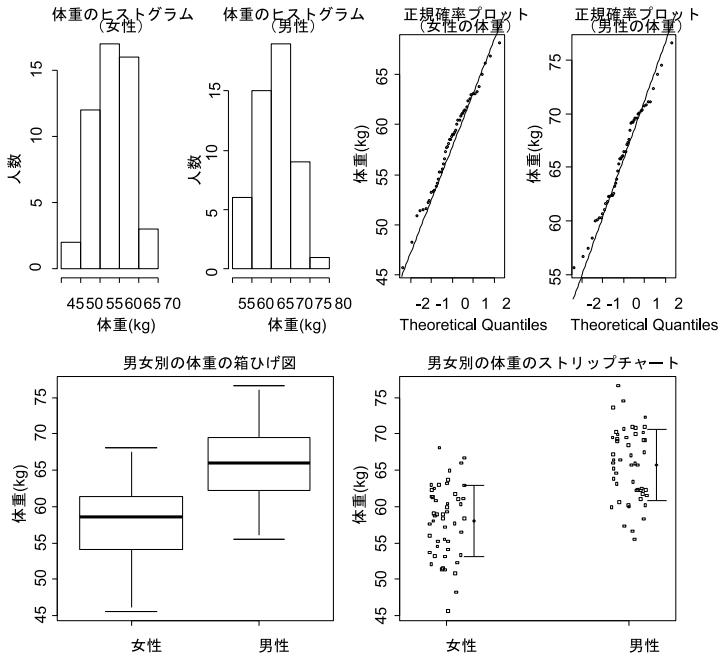
医学情報処理演習第 3 回「記述統計量」 2005 年 10 月 24 日 中澤　港

前回の課題の回答例

体重は比尺度をもつ数値型の量的変数なので，分布を示すにはヒストグラムまたは正規確率プロットを用いるのが普通である。目的によっては，箱ひげ図やストリップチャートも使えないことはない（男女別に分布をみるというよりも，男女間で分布を比較する目的なら，ヒストグラムや正規確率プロットより見やすい場合もある）。下枠内のコードを実行すれば，その下のグラフ群ができあがる。

it02-ans.R

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/p01.txt")
attach(dat)
mWT <- tapply(WT,SEX,mean); sWT <- tapply(WT,SEX,sd); IS <- c(1,2)+0.15
layout(rbind(c(1,2,3,4),c(5,5,6,6)))
hist(WT[SEX=="F"],main="体重のヒストグラム\n（女性）",xlab="体重 (kg)")
hist(WT[SEX=="M"],main="体重のヒストグラム\n（男性）",xlab="体重 (kg)")
qqnorm(WT[SEX=="F"],main="正規確率プロット\n（女性の体重）",ylab="体重 (kg)")
qqline(WT[SEX=="F"],lty=2)
qqnorm(WT[SEX=="M"],main="正規確率プロット\n（男性の体重）",ylab="体重 (kg)")
qqline(WT[SEX=="M"],lty=2)
levels(SEX) <- c("女性","男性")
boxplot(WT~SEX,main="男女別の体重の箱ひげ図",ylab="体重 (kg)")
stripchart(WT~SEX,method="jitter",vert=T,ylab="体重 (kg)",main="男女別の体重のストリップチャート")
points(IS,mWT,pch=18); arrows(IS,mWT-sWT,IS,mWT+sWT,code=3,angle=90,length=.1)
detach(dat)
```



データを記述する２つの方法

前回はデータを図示して直感的に全体像を把握する方法を示した。今回は，生データが含んでいる多くの情報を少ない数値に集約して示す方法を説明する。つまり，分布の特徴をいくつかの数値で代表させようというわけである。このような値を，代表値と呼ぶ。たんに代表値という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。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には，分布の位置を示す値と，分布の広がりを示す値がある。例えば，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，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。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。

一般に，統計処理の対象になっているデータは，仮想的な母集団（言い換えると，その研究結果を適用可能と考えられる範囲でもある）からの標本（サンプル）であり，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。母集団での位置や広がりを示す値は母数（parameter）と呼ばれ，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。例えば不偏分散は尺度母数の一つである。

中心傾向（Central Tendency）

平均（mean）

平均[*1]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（ミューと発音する）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[*2]。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[*3]。

[*1] ここで平均と呼んでいるのは，とくに断りがなければ，算術平均 (arithmetic mean) のことである。

[*2] 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。

[*3] 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。

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。もちろん，通常は，データを何かの変数に付値しておいて，関数は変数に対して適用するので，以下のように入力することになる。なお，数値型あるいは整数型ベクトルの変数 `X` について，`mean(X)` は `sum(X)/length(X)` と同値である。

```
X <- c(5,8,10,11,12)
mean(X)
```

中心傾向として有名なものには，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。

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。

文章通りに R の式を打つと，下記のようになる。

```
X <- c(2,4,6,8,30)
barX <- mean(X)
barX
x <- X-barX
x
sum(x)
```

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，

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

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。

例題３

ある大学の３つの学部で行われた TOEIC の試験の平均点が $\{440, 470, 610\}$ であったとする。これら３つの学部それぞれの人数が，順に $\{200$ 人，500 人，300 人$\}$ であったなら，この大学の TOEIC の総平均は何点か？

文章通りに R で実行する式は以下の通り。

```
P <- c(440,470,610)
N <- c(200,500,300)
sum(N*P)/sum(N)
```

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[*4]。

少数の極端な値の例をあげてみよう。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を計算してみると，

```
P <- c(5,10,30)
N <- c(80,15,5)
sum(N*P)/sum(N)
```

を実行すると７本とわかる。ところが，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，平均論文数は，

```
P2 <- c(P,2330)
N2 <- c(N,1)
sum(N2*P2)/sum(N2)
```

より 30 本となる。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，従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。つまりこれは，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。

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

[*4] 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。

分割する値である。中央値を求めるには式は使わないが，決まった手続き（アルゴリズム）として，並べかえ (sorting) は必要である。極端な外れ値の影響を受けにくいので，歪んだ分布に対する最も重要な central tendency の指標といえる。先のA大学教員の論文数の例では，1 人だけ 2330 本の論文数をもつ人が入ろうが入るまいが，論文数の中央値（つまり，論文数の順位が中央の人の論文数）は 5 本で変わらない。これだけではわかりにくいと思うので，もっと単純な例で考えてみよう。

例題 4

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[*5]，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[*6]。

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 4 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 5

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 6

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 5 と同様に計算すると，平均は 18.4，中央値は 12 となる。例題 5 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 7

次の分布の中央値は何か？　{4, 6, 9, 10, 11, 12}

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。R では以下の 1 行を打てばよい。

[*5] 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。

[*6] これに対して平均は，値の大きさによって変わるので，value sensitive であるといえる。

```
median(c(4,6,9,10,11,12))
```

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。例えば，例題７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。

例題８

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

例題９

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８でも例題９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。

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８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９で真の値は，{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。

ここでもう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。

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[*7]。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。なお，R では区間は Factor 型になってしまうので，以下のように区間の中央の値を数値型変数として入れて計算する。

it03-1.R

```
CA <- 10:1*5-3
FRE <- c(1,2,3,6,8,17,26,11,2,0)
X <- c(rep(CA,FRE))
median(X)
```

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。

[*7] 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。

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。

最頻値（Mode）

残る最頻値は，きわめて単純で，もっとも度数が多い値をいう。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[*8]。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，平均が最も小さく，中央値が次で，最頻値が最も大きくなる。右すそを引いた分布では逆になる。

例えば，例題９のデータで最頻値を求めるためには，R では以下のようにすればよい。

```
X <- c(7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 10, 10)
names(sort(table(X),dec=T))[1]
```

ただし，最頻値が複数あるかもしれないので，2 行目は `sort(table(X),dec=T)` としておいて，頻度の高い順に並べ替えられた度数分布表を見る方が確実である。

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[*9]，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，名義尺度のカテゴリ変数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

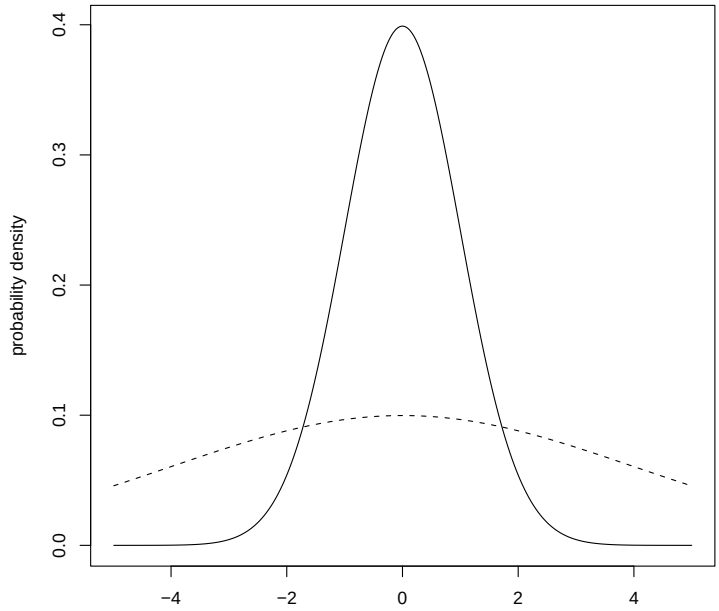
[*8] しかし隣り合う２つの値がともに最頻値である場合は二峰性だとはいわず，離れた２つの値が最頻値あるいはそれに近い場合，つまり度数分布やヒストグラムの山が２つある場合に，分布が二峰性だといい，２つの異なる分布が混ざっていると考えるのが普通である。

[*9] 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。

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。例えば，例題６のデータで幾何平均と調和平均を計算するには，以下のようにすればよい（各行の#以降はコメントなので打たなくていい)。

```
X <- c(2, 4, 7, 9, 12, 15, 17, 46, 54)
exp(mean(log(X))) # prod(X)^(1/length(X)) と同じ
1/(mean(1/X))
```

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。例えば，上図の２つの分布は[*10]，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。

[*10] この図を書くための R のプログラムは次の通り。

```
x <- seq(-5,5,length=1001)
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)
```

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。

範囲 (range)

範囲は，最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，１つの値として示すときは，最大値から最小値を引いた値になる。

例題 10

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 11

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影響を受けにくい値の方がよい。

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，ちょうど大きさの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。R では，`fivenum()` という関数によって，五数要約値が表示される。

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。R で数値型変数 X の 20 パーセンタイル値と 80 パーセンタイル値を計算するには `quantile(X,c(0.2,0.8))` のように `quantile()` 関数を使えばよいが，実はパーセンタイル値指定のデフォルトが c(0,0.25,0.5,0.75,1) なので，同順位のデータがなければ，`quantile(X)` は `fivenum(X)` と同じ結果になる。

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。結果の示し方としては，$Q3 - Q1$ という１つの値で示すのではなく，$[Q1, Q3]$ という形で示すことも多い。

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 12

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。

データはファイル `p03.txt` としてサーバに置いたので，R では，以下のようにすればよい[*11]。

it03-2.R

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/p03.txt")
X <- dat$WT
Q <- fivenum(X)
Q1 <- Q[2]
Q2 <- Q[3]
Q3 <- Q[4]
IQR <- Q3-Q1
SIQR <- IQR/2
cat("四分位範囲=",IQR," [",Q1,",",Q3,"], ","四分位偏差=",SIQR,"\n")
```

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[*12]。偏差の大きさは，分布のばらつきを反映している。

例えば，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最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。

[*11] 同順位があるため，`fivenum()` の代わりに `quantile()` を使うと若干異なる結果になる。
[*12] 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ぶ方がよい。

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。この例では，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。

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[*13]。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。即ち，不偏分散 V_{ub} は，

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

である。R で，数値型変数 X の不偏分散は，`var(X)` によって得られる。

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[*14]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。R で，数値型変数 X の不偏標準偏差は，`sd(X)` によって得られる。

[*13] 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である。

[*14] 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。

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[*15]，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ことに注意）を平均で割って 100 を掛けた値を変動係数という。即ち，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。

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。

課題

R にはさまざまなサンプルデータが含まれており，`try(data())` とすると一覧表示できるが，今回は，その中の `ChickWeight` を使ってみることにする。これは，含まれるタンパク質が異なる 4 種類の試験食を与えて飼育した 50 羽の鶏の体重を 0 日目から 20 日目まで測定したデータである（何日分か欠損がある）。

Chick という順序型の変数に鶏の個体番号が入っており，Diet という要因型の変数に食事の種類を示す数字が入っており，Time という数値型の変数に何日目の体重かという値が入っており，weight という数値型の変数に体重が入っている。以下のようにすると，`X` という変数に，餌の種類を問わず，20 日目の鶏の体重がすべて入るので，この `X` に適切な関数を適用して，20 日目の鶏の体重の平均，標準偏差（もちろん不偏標準偏差），中央値，四分位偏差を求めよ。

```
data(ChickWeight)
attach(ChickWeight)
X <- weight[Time==20]
```

結果は配布する紙に学籍番号，氏名と共に自筆して提出すること。結果の提出をもって出席確認とする。

[*15] 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。