

医学情報処理演習第3回「記述統計量」 2006年10月23日 中澤 港

前回の課題の回答例

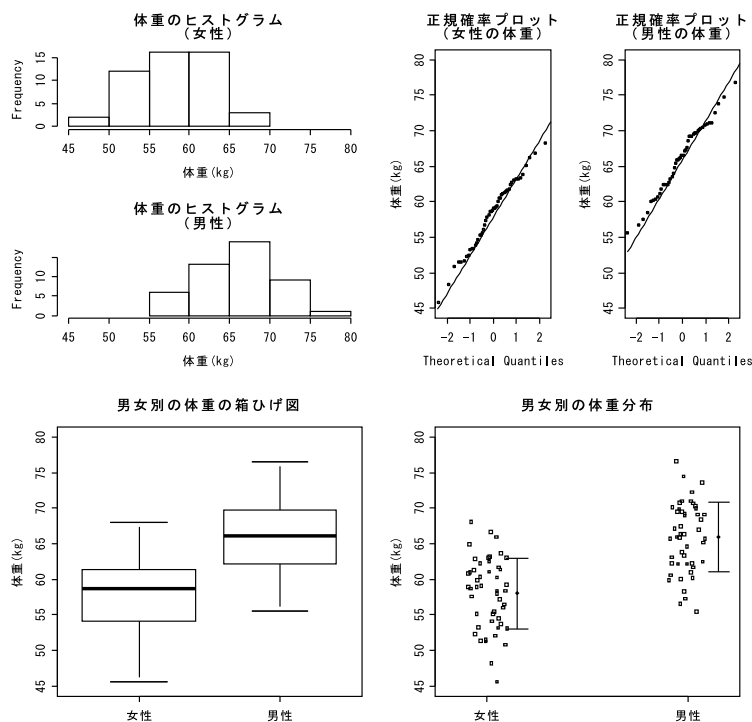
体重は比尺度をもつ数値型の量的変数なので、分布を示すにはヒストグラム、正規確率プロット、箱ひげ図、ストリップチャートなどを用いる。とくに、男女間で分布を比較する目的なら、ヒストグラムや正規確率プロットより箱ひげ図やストリップチャートの方が見やすい場合もある。下枠内のコードを実行すれば、その下のグラフ群ができてあがる。

it02-ans-2006.R

```

dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/p02.txt")
dat2 <- subset(dat,complete.cases(dat)); rm(dat); attach(dat2)
mwt <- tapply(wt,sex,mean); swt <- tapply(wt,sex,sd); IS <- c(1,2)+0.15
layout(matrix(c(1,2,5,5,1,2,5,5,3,3,6,6,4,4,6,6),4,4))
hist(wt[sex=="F"],main="体重のヒストグラム\n(女性)",xlab="体重 (kg)",xlim=c(45,80))
hist(wt[sex=="M"],main="体重のヒストグラム\n(男性)",xlab="体重 (kg)",xlim=c(45,80))
qqnorm(wt[sex=="F"],main="正規確率プロット\n(女性の体重)",ylab="体重 (kg)",ylim=c(45,80))
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)",ylim=c(45,80))
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)",ylim=c(45,80))
stripchart(wt~sex,method="jitter",vert=T,main="男女別の体重分布",
  ylab="体重 (kg)",ylim=c(45,80))
points(IS,mwt,pch=18); arrows(IS,mwt-swt,IS,mwt+swt,code=3,angle=90,length=.1)
detach(dat2)

```



データを記述する2つの方法

今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、 $\sum X = X_1 + X_2 + X_3 + \dots + X_N$ 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、もちろんサンプルサイズである^{*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が、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} と表するのが普通である。

例題

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、この例題の解を得るには、`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。しかし、定義から明らかなように、少数でも極端な外れ値があると、平均は強くその影響を受けてしまうという欠点がある。対策としては、ネイマンの外れ値の検定などを用いるか、あるいは数値以外の情報に基づいて、その極端な値がサンプルとして相応しくないと判定して解析から除外するか、それができない場合は、後述する中央値を使えばよい。

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、

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

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

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)
```

中央値 (median)

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。先に述べた通り、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。

例題

次のデータの中央値は何か？ { 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*⁴、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*⁵。

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、上の例題の解を得るには、median(c(1, 4, 6, 8, 40, 50, 58, 60, 62)) とすればよい。

例題

次のデータの平均と中央値は何か？ A={2, 4, 7, 9, 12, 15, 17} また、それに46と54が加わった B={2, 4, 7, 9, 12, 15, 17, 46, 54} の平均と中央値は何か？

Rで

```
x <- c(2,4,7,9,12,15,17)
mean(x)
median(x)
y <- c(x,46,54)
mean(y)
median(y)
```

とすると、Aの平均は約9.43、中央値は9であり、Bの平均は18.4、中央値は12となることがわかる。大きい方に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つ大きい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

次のデータの中央値は何か？ { 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

*⁴ 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。

*⁵ 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。

使うことになっている。R では以下の 1 行を打てばよい。

```
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、上の例題のデータが、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ません。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ません。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

データ C={7, 7, 7, 8, 8, 8, 9, 9, 10, 10} の中央値は何か？ また、D={7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 10, 10} の中央値は何か？

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、C も D 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、連続量データの数値はぴったりその値というわけではなく、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 C 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 D 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。以上のような考え方で真の中央値を求める関数 `truemedian()` を R で定義すると、次の枠内ようになる。

truemedian.R

```
truemedian <- function(X,h=1) {
  YY <- rep(0,length(X))
  XX <- table(X)
  q <- length(XX)
  k <- 0
  for (i in 1:q) {
    L <- as.numeric(names(XX)[i])-h/2
    for (j in 1:XX[[i]]) {
      k <- k+1
      YY[k] <- L+h*(2*j-1)/(2*XX[[i]])
    }
  }
  median(YY)
# 真の値を表示するには print(YY)
}
```

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。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見る(つまり総数を見る)。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*6。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、次のように区間の中央値を数値型変数として入れて計算すると、17という結果が得られることから15-19歳と判定できる。また、階級幅が5なので先に定義したtruemedian()関数を使えば、真の中央値として約19.3歳も得られる(注:言うまでもないが、先にtruemedian()関数の定義部分を実行しておかなくてはエラーになる)。

it03-1-2006.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)
truemedian(X,5)
```

*6 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、Cの中央値が下から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。

この考え方をすると、中央値が正確な分布の中央になっている（少なくともその近似になっている）という特性が強化される。

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。もっとも度数が多い値が、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*7。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、例題のデータ D 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。

```
D <- c(7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 10, 10)
names(sort(table(D),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(D),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。

使い分け

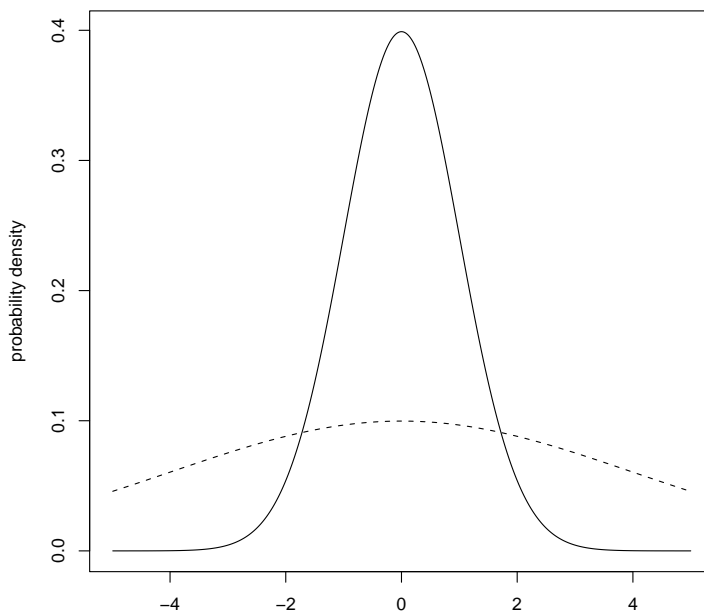
平均は、(1) 分布のすべての値を考慮した値である、(2) 同じ母集団からサンプリングを繰り返した場合に一定の値となる、(3) 多くの統計量や検定で使われている、という特長をもつ。標本調査値から母集団の情報を推論したい場合に、もっとも普通に使われる。しかし、(1) 極端な外れ値の影響を受けやすい、(2) 打ち切りのある分布では代表性を失う場合がある*8、という欠点があり、外れ値があったり打ち切りがあったりする分布では位置の指標として中央値の方が優れている。最頻値は、標本をとったときの偶然性の影響を受けやすいし、もっとも頻度が高い値以外の情報はまったく使われない。しかし、試験の点で何点の人が多か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、数値ベクトル X について幾何平均を得るには `exp(mean(log(X)))` または `prod(X)^(1/length(X))` とすれば良いし (但し後者の方が圧倒的に計算効率が悪い)、調和平均を得るには `1/(mean(1/X))` とすれば良い。

*7 隣り合う2つの値がともに最頻値である場合は二峰性とはいわず、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。

*8 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。

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、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。なお、上図を書くためのRのプログラムは次の通りである。

```
curve(dnorm(x,0,1),-5,5,lty=1,xlab="",ylab="probability density")
curve(dnorm(x,0,4),-5,5,lty=2,add=T)
```

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。値のとり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

次のデータの範囲はいくらか? { 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

次のデータの範囲はいくらか？ {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 Q_1 , Q_2 , Q_3 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、`quantile(X)` は `fivenum(X)` とほぼ同じ結果になる*9。

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、 $Q_3 - Q_1$ という 1 つの値で示すのではなく、 $[Q_1, Q_3]$ 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

*9 詳しく言うと、`fivenum()` 関数はごく単純に並べ替えて四分位数に該当する順位を求め、それが整数ならそのまま、小数なら前後の (つまり、上から数えた場合と下から数えた場合の) 平均を取った値を返す。`quantile()` 関数には順位の計算法が 9 種類含まれており、デフォルトは S と同じ Type 7 である。`fivenum()` と同じ結果を返すのは Type 2 らしいので、同じ結果が欲しければ `quantile(X, type=2)` とすればよい。Hyndman, R. J. and Fan, Y. (1996) Sample quantiles in statistical packages, *American Statistician*, 50: 361-365. は Type 8 を推奨している。

点が優れている。

例題

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てあるので、次のようにする。

it03-2-2006.R

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/p03.txt")
Q <- fivenum(dat$WT)
IQR <- Q[4]-Q[2]
SIQR <- IQR/2
cat("四分位範囲=", IQR, " [" , Q[2] , " , " , Q[4] , " ] , " , "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ため、分布のばらつきを正しく反映しない可能性があるため、すべてのデータを使ってばらつきを表す方法も考えたいわけである。そこで注目したいのが偏差、即ち生の値と平均との距離である^{*10}。偏差の大きさは、分布のばらつきを反映している。

例えば、 $E = \{11, 12, 13, 14, 15, 16, 17\}$ 、 $F = \{5, 8, 11, 14, 17, 20, 23\}$ は、どちらも平均は 14 だが、F が E よりもばらつきが大きいことは直感的にわかる。言い換えると、F の方が E よりも平均からの距離が大きい。これを 1 つの値として表したいわけである。ただ合計しただけでは、平均のところでは述べたように、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良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いうので、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。この例では、E の平均偏差は約 1.71、F 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。

```
meandev <- function(X) {
  mX <- mean(X)
  sum(abs(X-mX))/length(X) }
meandev(c(11, 12, 13, 14, 15, 16, 17))
meandev(c(5, 8, 11, 14, 17, 20, 23))
```

^{*10} 誤差と呼ばれることもあるが、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。

平均偏差はすべてのデータを使う利点があるが、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、絶対値にする代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^{*11}。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、 $\text{var}(X)$ 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、 $\text{Mean} \pm 2\text{SD}$ ^{*12}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、 $\text{sd}(X)$ によって得られる。

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

生データ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^{*13}、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・歪みが多い・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 \pm ばらつき、という形で示すことが多い。

^{*11} 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な上、桁落ちしにくいのでお薦めする。

^{*12} 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。

^{*13} 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。

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 `sleep` を使ってみる。患者を2群に分けて別々の催眠剤（変数 `group` で表される）を与えたときの睡眠時間の変化（変数 `extra`）が得られている。催眠剤1と催眠剤2それぞれについて、データの分布の位置とばらつきの指標値を計算せよ。

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。