

フリーソフトによるデータ解析・マイニング

R と基本統計量

データには、身長、体重、成績などのように量的計測できるデータと性別、血液型などのように性質を表すデータがある。前者を**量的データ**と呼び、後者を**質的データ**、あるいは**カテゴリカルデータ**と呼ぶ。

統計量(statistic)とは、統計データから計算、要約した数量のことである。基本統計量とは、通常広く使用されている合計、比率、平均、中央値、最頻値、分散、四分位数などを指す。

1. 計と比率

1.1 合計

量的データを分析する際には、データを加え合わせる作業が頻繁に行われる。

n 個のデータ x_1, x_2, \dots, x_n があつた時、これらのデータの値をすべて加えて合計を求める演算を

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_i + \dots + x_n$$

で表す。 $\sum_{i=1}^n x_i$ のような書き方もある。

例えば、ある販売部門の 10 人の第一四半期の売上が表 1 の通りであるとする。

表1 10人の年売上(単位は百万)

	A	B	C	D	E	F	G	H	I	J	合計
変数	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	$\sum_{i=1}^{10} x_i$
	⇕	⇕	⇕	⇕	⇕	⇕	⇕	⇕	⇕	⇕	⇕
売上	12	6	13	7	8	11	12	9	10	7	95

10 人 A、B、C、 \dots 、I、J の売上を x にし、それぞれ識別できるようにするため、 x に A、B、C、 \dots 、I、J の順に添字 1、2、3、 \dots 、9、10 をつけたものが $x_1, x_2, x_3, \dots, x_9, x_{10}$ とする。A の売上 x_1 は 12、B の売上 x_2 は 6 のように対応関係をつけると、10 人の売上の合計

は次のようになる。

$$\begin{aligned}\sum_{i=1}^{10} x_i &= x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} \\ &= 12 + 6 + 13 + 7 + 8 + 11 + 12 + 9 + 10 + 7 \\ &= 95\end{aligned}$$

R では **sum** 関数を用いてデータの合計を求めることができる。

```
>uriage<-c(12,6,13,7,8,11,12,9,10,7)
> sum(uriage)
[1] 95
```

1.2 比率

表 1 の A 氏の売上は 12 となっている。この値 12 が持っている意味は相対的なもので、単純に 12 の値が大きいか、小さいかは議論できない。この 12 が相対的に大きいか、小さいかは全体のなかで占める比率、あるいは割合を求めて議論をしなければならない。

ある値 x_j が全体のなかで占める比率は、 x_j を総数 $\sum_{i=1}^n x_i$ で割って求める。これを 記号で表すと次のようになる。

比率:	$\frac{\text{個別}}{\text{全体}} = \frac{x_j}{\sum_{i=1}^n x_i}$
	$\frac{\text{個別}}{\text{全体}} \times 100\% = \frac{x_j}{\sum_{i=1}^n x_i} \times 100\%$
百分率:	

漢字がわからない人は、 $\frac{\text{個別}}{\text{全体}}$ は理解できないが、どの国の人でも若干の数学教養をもっていれば、式

$$\frac{x_j}{\sum_{i=1}^n x_i}$$

の意味は簡単に理解できる。このような数式表記は一種の国際言語であるとも言える。上記の

Aの人のことを例にすると、A氏の売上の比率は(下記の記号は近似を意味する)

$$\frac{x_1}{\sum_{i=1}^n x_i} = \frac{12}{95} \quad 0.1263$$

で、これをパーセンテージ(percentage、百分率、百分比)で表すと次のとおりになる。

$$\frac{x_j}{\sum_{i=1}^n x_i} \times 100\% = \frac{12}{25} \times 100\% \quad 0.1263 \times 100\% = 12.63\%$$

Rでは、次のように比率の定義とおりに演算を行うことで、簡単に比率と割合を求めることができる。入力した売り上げのデータ `uriage` を用いて説明する。

```
> uriage/sum(uriage)
```

```
[1] 0.12631579 0.06315789 0.13684211 0.07368421 0.08421053 0.11578947  
[7] 0.12631579 0.09473684 0.10526316 0.07368421
```

```
> 100*(uriage/sum(uriage))
```

```
[1] 12.631579 6.315789 13.684211 7.368421 8.421053 11.578947 12.631579  
[8] 9.473684 10.526316 7.368421
```

パッケージ `sca` の中には比率データをパーセンテージに変換し、%記号を付けたデータを返す関数 `percent` がある。引数 `d` で小数点右の何桁まで出力するかを指定することができる。デフォルトは `d=0` になっているので、`d` を指定しない場合、小数点以下は返さない。

```
> library(sca)
```

```
> percent(uriage/sum(uriage))
```

```
[1] "13%" "6%" "14%" "7%" "8%" "12%" "13%" "9%" "11%" "7%"
```

```
> noquote(percent(uriage/sum(uriage)))
```

```
[1] 13% 6% 14% 7% 8% 12% 13% 9% 11% 7%
```

```
> noquote(percent(uriage/sum(uriage),d=1))
```

```
[1] 12.6% 6.3% 13.7% 7.4% 8.4% 11.6% 12.6% 9.5% 10.5% 7.4%
```

2. 中心を表す統計量

統計量にはそのデータの中心の位置を表す代表値として平均、中央値、最頻値が多く用いられている。

2.1 平均

ここでいうデータの平均(average、または mean)は、データの合計をデータの個数で割った算術平均である。表1のデータを用いて説明することにする。

	A	B	C	D	E	F	G	H	I	J	合計
変数	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	$\sum_{i=1}^{10} x_i$
	⇕	⇕	⇕	⇕	⇕	⇕	⇕	⇕	⇕	⇕	⇕
売上げ	12	6	13	7	8	11	12	9	10	7	95

このデータから一人当たりの平均売上を考えて見ましょう。一人あたりの平均売上は合計値95を人数10で割ることで求める。

$$\frac{12 + 6 + 13 + 7 + 8 + 11 + 12 + 9 + 10 + 7}{10} = \frac{95}{10} = 9.5$$

値9.5がこの10個のデータの平均値である。この具体的な計算過程を定義すると次のようになる。

n 個のデータ x_1, x_2, \dots, x_n があつた時、これらのデータの平均値 \bar{x} は次の式で求めます。

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j = \frac{1}{n} (x_1 + x_2 + \dots + x_j + \dots + x_n)$$

一般的に平均値は x の上に横線を引いた \bar{x} (エックスバーと呼ぶ) で表す。 \bar{x} を求める式は一見煩雑に見えるが、合計をデータの数で割る単純な算術計算である。

R では、算術平均を求める関数 `mean` が用意されている。データ `uriage` を用いてその使用例を示す。

```
> mean(uriage)
```

```
[1] 9.5
```

2.2 最頻値 (mode)

データのなかでもっとも頻繁に現れている値を**最頻値**、あるいは**モード(mode)**と呼ぶ。例えば、データ

1 3 5 4 3

では、3が2回現れ、それ以外は1回であるので、最も頻繁に現れている値は3となる。つまり、このデータの最頻値(モード)は3である。最頻値は出現頻度が最も高いデータであるので、データの中に必ず最頻値が存在するとは限らない。

Rでは最頻値を求める関数が用意されていないので、他の関数を用いて間接に求めることができる。例えば、ベクトルデータについては関数 `table` を用いると、各要素の出現頻度が集計される。返された結果からわかるように、3が2回現れていると集計されている。

```
> table(c(1,5,4,3,3))
```

```
1 3 4 5
```

```
1 2 1 1
```

2.3 中央値 (median)

データを大きさの順に並べた場合、中央に位置する値を**中央値**、あるいは**メディアン(median)**と呼ぶ。例えば、データ

2 3 5 3 4

があったとしよう。5個のデータを大きさの順に並べると

2 3 3 4 5

ここが中央

となり、中央に位置する値は矢印上の3であるので、このデータセットの中央値は3である。これはデータの数が奇数の場合であるが、データの数が次のように偶数の場合は、中央の両値を足して2で割った値を中央値とする。この例では、 $(3+4)/2=3.5$ が中央値となる。

Rでは最頻値(メジアン)を求める関数 `median` である。

2 3 3 4 5 7

ここが中央

```
> median(c(2, 3, 3, 4, 5, 7))
```

```
[1] 3.5
```

3 バラツキの特性値

データがどのような値を中心としているかはデータの特徴を知る上で重要な情報であるが、データがどの程度散らばっているかという情報も、データの特徴を把握する上で重要である。

3.1 範囲(range)

データの中で最も大きい値を最大(maximum)値と呼び、最も小さい値を最小(minimum)値と呼ぶ。データの範囲は、データの最小値から最大値までの区間を指す。

Rでの最大値、最小値、範囲を求める関数はそれぞれ関数 `max`、`min`、`range` である。次に `uriage` データセットを用いた例を示す。

```
> max(uriage)
```

```
[1] 13
```

```
> min(uriage)
```

```
[1] 6
```

```
> range(uriage)
```

```
[1] 6 13
```

3.2 分散と標準偏差

分散と標準偏差はデータの散らばり(バラツキ)の状況を表す統計量である。分散と標準偏差の定義に関する説明の前に、まず次の表2のA、B両氏の携帯電話料金のデータをみよう。両氏の6ヶ月間の電話料金の総額および平均は同じである。何が違うであろうか？

表2 A、B両氏の電話料金

	1月	2月	3月	4月	5月	6月	合計	平均
A氏	7206	6358	9809	8915	7850	8965	49103	8183.833
B氏	5550	4880	15914	4856	13013	4890	49103	8183.833

両氏のデータをそれぞれ散布図で表すと図1のとなる。横軸が月で、縦軸が料金で、図のなかの横線は平均値である。図からわかるようにA氏の月別の料金は平均値の周辺に集中しているが、B氏の月別料金はA氏より散らばっている。

このようにデータが平均値からどの程度散らばっているかでデータの特徴を説明することができる。データが平均値からどの程度散らばっているかを示す量として**分散**(variance)と**標準偏差**(standard deviation)と呼ばれている統計量がある。

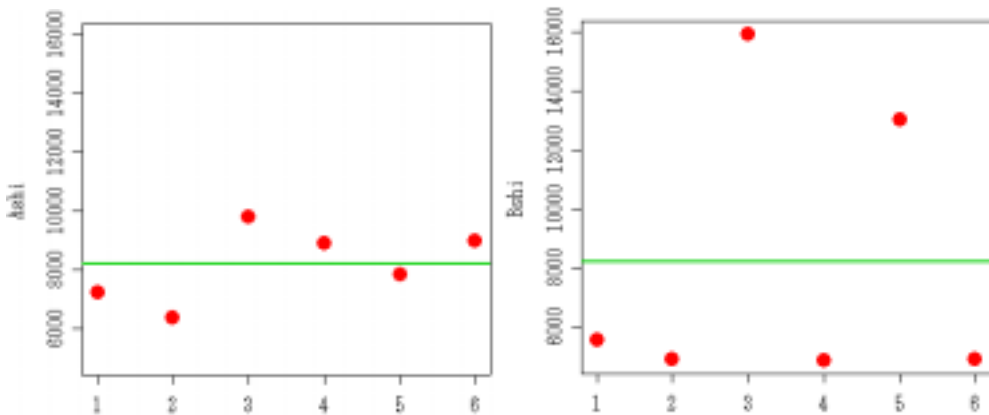


図1 電話料金の散布図

n 個のデータ x_1, x_2, \dots, x_n があつた時、これらのデータの分散 S^2 は次の式で定義されている。

$$S^2 = \frac{1}{n-1} \{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_j - \bar{x})^2 + \dots + (x_n - \bar{x})^2\}$$

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

また、標準偏差 S は分散 S^2 の正の平方根である。

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2}$$

R では分散を求める関数 `var`、標準偏差を求める関数 `sd` がある。次に A、B 両氏の電話料金の分散、標準偏差を求めるコマンドを次示す。それぞれの分散は 1637453、24570507 で B 氏の分散が A 氏よりはるかに大きいことがわかる。この結果を散布図と見比べると、バラツキが大きいデータの分散値が大きいことがわかる。この例のように、用いたデータの桁数と比べて分散の値の桁数がかなり大きい。データの桁数が多いと扱うのに不便であるので、場合によっては分散値の正の平方根を用いる。分散値の正の平方根を標準偏差 (standard deviation) と呼ぶ。

```
>Ashi<-c(7206,6358,9809,8915,7850,8965)
>Bshi<-c(5550,4880,15914,4856,13013,4890)
> var(Ashi)
[1] 1637453
> var(Bshi)
```

```
[1] 24570507
> sd(Ashi)
[1] 1279.630
> sd(Bshi)
[1] 4956.865
```

注：図1の散布図は次のコマンドで作成することができる。

```
> plot(1:6,Ash,pch=21,bg=2,col=2,cex=2,ylim=range(Bshi))
> abline(mean(Bshi),0,lw=2,col=3)
> plot(1:6,Bshi,pch=21,bg=2,col=2,cex=2,ylim=range(Bshi))
> abline(mean(Bshi),0,lw=2,col=3)
```

3.3 四分位数（しぶんいすう）

データを大きさの順にならべて、データを4等分したとき、各等分の境の値を四分位数と呼ぶ。例えば、データ

3 5 6 8 9 11 12 15 16

があるとする。データの数が奇数であるので、中央値は9となる。データ9を含むその左辺、右辺のデータの中央値はそれぞれ6、12である。その値6、9、12をそれぞれ**第1四分位数**(25%点)、**第2四分位数**(50%点)、**第3四分位数**(75%点)と呼ぶ。

3 5 6 8 9 11 12 15 16

0%点 25%点 50%点 75% 100%点

また下記の式

$$\frac{\text{第3四分位数} - \text{第1四分位数}}{2}$$

で得られた値を**四分位偏差**と呼ぶ。よって、上記のデータの四分位偏差は

$$\frac{12 - 6}{2} = 3$$

となる。この四分位偏差もデータのバラツキを示す一つの指標である。

Rで四分位数を求める関数は **quantile** である。四分位偏差は関数 **IQR** を2で割ることで、求めることが可能である。

```
> quantile(c(3,5,6,8,9,11,12,15,16))
```

```
0% 25% 50% 75% 100%
 3   6   9  12  16
```

```
> IQR(c(3,5,6,8,9,11,12,15,16))
```

```
[1] 6
```

