

統計解析ソフト「R」の導入

竹田裕一

統計解析ソフト「R」はオープンソースフリーウェアで、誰でも自由にダウンロードして使用することができます。ここでは、基本的な使い方の説明をいたします。

1. データの入力方法

データの入力方法は直接入力とファイルやクリップボードからの読み込みの2種類があります。例えば、5つのデータをxに代入する場合、

```
> x <- c(52.3,24.7,32.6,43.1,50.0)
```

と入力します。データの数が多い場合などはcsvファイルをExcel等で作成し、

```
> x <- read.csv("ファイル名.csv",header=T,row.name=1)
```

等のようにして読み込むことが可能です。

2. 基本的な統計量の関数

「R」では基本的な統計量を求める関数が用意されており、簡単に求めることが可能です。

標本平均 : `mean(x)` 標本分散 : `var(x)` 標本標準偏差 : `sd(x)`

中央値 : `median(x)` 共分散 : `var(x,y)` 相関係数 : `cor(x,y)`

例えば、52.3,24.7,32.6,43.1,50.0 の標本平均、標本分散、標本標準偏差を求めるのであれば、次のようになります。

```
> x <- c(52.3,24.7,32.6,43.1,50.0)
```

```
> mean(x)
```

```
[1] 40.54
```

```
> var(x)
```

```
[1] 137.073
```

```
> sd(x)
```

```
[1] 11.70782
```

3. 確率計算

「R」では様々な確率分布や確率点を求める関数が用意されており、その確率分布に従う乱数も簡単に作ることができます。

- ・ $n=10, p=0.6$ の 2 項分布に従う確率変数が 6 になる確率
> dbinom(6,10,p=0.6)
[1] 0.2508227
- ・ 標準正規分布に従う確率変数が 1.96 以下になる確率
> pnorm(1.96)
[1] 0.9750021
- ・ 標準正規分布に従う確率変数の上側 5%点
> qnorm(1-0.05)
[1] 1.644854
- ・ 自由度 5 の t 分布に従う確率変数が 2.52 以下になる確率
> pt(2.52,5)
[1] 0.9734127
- ・ 自由度 9 の t 分布に従う確率変数の下側 1%点
> qt(0.01,9)
[1] -2.821438
- ・ 平均 157, 分散 5^2 の正規分布に従う乱数 5 個
> rnorm(5,157,5)
[1] 150.0091 161.3488 150.5507 154.5813 144.7027

4. 区間推定

4 章の区間推定は、統計量の計算と確率点の関数を用いて簡単に求めることができます。例えば P.54 の例題の場合

```
> x <- c(1.4,1.5,1.3,1.5,1.4,1.7, …… ,1.7,1.5)
> mean(x)-qt(1-0.05/2,19)*sd(x)/sqrt(20)
[1] 1.371418
> mean(x)+qt(1-0.05/2,19)*sd(x)/sqrt(20)
[1] 1.508582
```

となるので、 $1.37 < \mu < 1.51$ と求めることができます。また、P.58 のような二標本の平均の区間推定の場合、共通な母分散（標準偏差）の推定を行うことによって

```
> x <- c(16.4,15.5,15.8,16.1,15.1,16.9,15.6,14.8,16.5,16.3)
> y <- c(14.9,13.8,14.1,14.6,14.7,13.1,14.9,13.4,14.5,14.0)
> s <- sqrt(((10-1)*var(x)+(10-1)*var(y))/((10-1)+(10-1)))
[1] 0.6437736
> (mean(x)-mean(y))-qt(1-0.05/2,18)*s*sqrt(1/10+1/10)
[1] 1.095135
> (mean(x)-mean(y))+qt(1-0.05/2,18)*s*sqrt(1/10+1/10)
[1] 2.304865
```

$1.10 < \mu < 2.30$ と求めることができます。

5. 平均に関する仮説検定

5, 6章の平均に関する仮説検定は、区間推定と同様に統計量の計算と確率点の関数を用いて簡単に求めることができます。統計量は

```
> x <- c(51.0,45.9,48.8,54.0,53.5,48.0,44.5,46.0,50.3,48.0)
> (mean(x)-50.2)/(sd(x)/sqrt(10))
[1] -1.187918
```

自由度9のt分布の両側5%点は

```
> qt(1-0.05/2,9)
[1] 2.262157
```

のように求められ、 $|-1.19| < 2.262$ であるから、帰無仮説を棄却することができないことが分かります。また、母分散が未知の場合はt.testの関数を用いて統計量等の計算を行うことが可能です。同じ問題をt.testを使うと

```
> t.test(x,mu=50.2,alternative="two.sided")
```

One Sample t-test

data: x

t = -1.1879, df = 9, p-value = 0.2653

alternative hypothesis: true mean is not equal to 50.2

(以下省略)

のような結果が得られ、p-value の値と有意水準の値を比較することによって帰無仮説を棄却するかどうか判断します。今回の場合、有意水準 5%との比較ですので、帰無仮説を棄却することができません。

8. 相関係数

10 章の相関係数は基本的な統計量の関数で簡単に求めることが可能で、散布図についても `plot(x,y)` で簡単に描くことができます。相関係数に関する検定については、関数 `cor.test` で行うことが可能で、P.123 の例の場合、

```
> x <- c(149,157,166,156,158,156,152,148,155,160,165,152,163,163)
> y <- c(153,167,160,152,163,168,160,162,154,157,162,152,157,155)
> cor.test(x, y, method="pearson")
Pearson's product-moment correlation
```

data: x and y

t = 0.4845, df = 12, p-value = 0.6367

alternative hypothesis: true correlation is not equal to 0

(以下省略)

のように同じ結果が得られます。

9. 回帰直線

11 章の回帰直線については関数 `lm` を使い、関数 `abline` を使うと、散布図に回帰直線を引くことができます。例えば、P.129 の例の場合

```
> x <- c(37,45,48,55,58,67,75)
> y <- c(112,132,120,154,136,162,164)
> lm(y~x)
Call:
lm(formula = y ~ x)
```

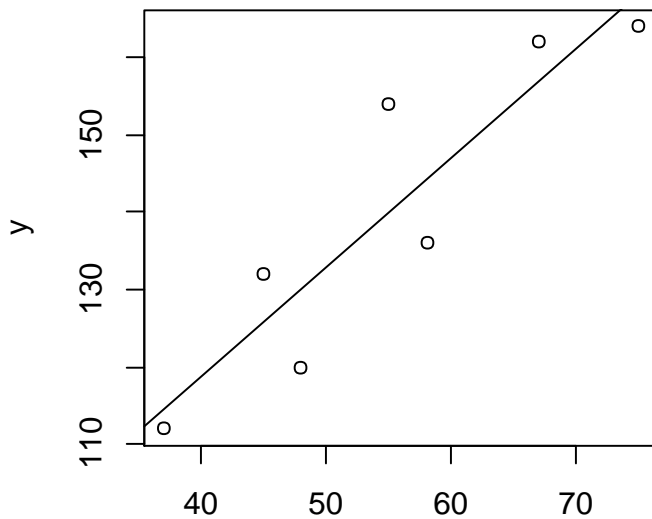
Coefficients:

(Intercept)	x
61.949	1.419

のように回帰直線が $y=62+1.4x$ であることが分かる（精密計算なため、本文とは結果が少し異なります）。また、散布図と回帰直線は

```
> plot(x,y)
```

```
>
```



```
abline(lm(y~x))
```

x

によって、次のようなグラフが得られます。