「R・エクセル・基礎数学」について

OpenIntro Statistics(4th Edition) では統計計算ソフト R や基礎数学が利用されている。ここで R, エクセル, 基礎数学について久保川・国友「統計学」付録にほぼ沿って簡単に解説する。なおこのメモは不完全であるから適宜、更新するものとする。(2021.3.13)

1　R 入門

[1] R について　R または R 言語は S 言語などを発展させた統計計算のソフトウエアである。元々は統計家 Ross Ihaka, Robert Gentleman などが中心になり開発が進められたが，1997 年ごろからかなりの数の統計家が参加しより充実した形で開発がオープンリソースで続けられている。R の特徴として次のような点が挙げられる。

- R はフリーなソフトなので，R を提供しているサイトからダウンロードし，容易にインストールすることができる。
- R はグラフィクスが充実している。
- 様々なパッケージが提供されており，最新の統計手法が利用可能である。その意味では，常に進化し続けている。
- 単なるデータ解析にとどまらず，シミュレーション実験など本格的な数値計算を行うこともできる。

R はまず R 本体を起動させるパッケージをインストールするように設計されているが，その他に利用可能な統計ソフトウエアは多数にのぼる。市販ではなくオープンリソースであるのでインストールや R 内部の計算についての責任は利用者であることが重要な点である。日本でも計算統計に関心のある統計家を中心に R の日本語化などの取り組みもあるが，統計数理研究所で開発された統計計算プログラムなども R に移植され広範な統計計算において R を利用することができるようになっている。

[2] R をダウンロードしよう　例えば Google で「R」と検索すると Web サイト”The R Project for Statistical Computing”を見つけことができる。”Download CRAN”から国別のミラーサイト，日本の統計数理研究所のサイト” http://cran.ism.ac.jp ”にアクセスし，自分のパソコンの OS の種類に応じてダウンロードするファイルを選択する。次に，” base ” のリンクをクリックする。OS が Windows の場合，一番上のリンク” Download R4.0.4 for Windows ”をクリックする。バージョンアップすると ” R4.0.4 ” の後ろの数字も変わる。現在

(2021.3.13) では ” R4.0.4 ” が最新版である。インストーラーを起動し指示通りにインストールする。なお、インストール時の選択のときアイコンに印をつけておくと画面上にアイコンが現れる。アイコンをクリックするだけで R が立ち上がる。

[3] R を利用しよう

(1) 四則演算．例えば 2+3 を求めるときには，> の後ろに 2+3 を入力しリターンキーを押すと答えが次の行に現れる。四則演算は次のようになる。

```
> 2+3
[1] 5
> 2-3
[1] -1
> 2*3
[1] 6
> 2/3
[1] 0.6666667
```

(2) データの入力．3 個のデータ 4, 7, 11 を A に入力するには, A <- c(4, 7, 11) と書く。実際 A の中身を print(A) で出力すると次のようになる。

```
> A <- c(4, 7, 11)
> print(A)
[1] 4 7 11
```

3×2 の行列として B に入力するには，

```
> B <- matrix(c(1, 2, 3, 4, 5, 6),3,2)
> print(B)
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

とすればよい。また 2×3 の行列として C に入力するには，

```
> C <- matrix(c(1,2,3,4,5,6),2,3)
> print(C)
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

とすればよい。また data1.txt という名前のテキストファイルに入っているデータを D として読み込むときには

```
> D <- read.table(”data1.txt”, header=T)
```

などとし，data2.csv という名前のエクセルファイルに入っているデータを E として読み込むときには

```
> E <- read.csv("data2.csv", header=T)
```

などとすればよい。 print(D), print(E) として D, E の中身を出力してみると，正しく読み込まれているかを確かめることができる。

(3) 基本統計量の計算．10 人の数学の得点が 52, 65,42,55,48,62,95,74,58,52 で与えられているとき，その最小値，第 1 四分位点，メディアン，平均，第 3 四分位点，最大値，分散, 標準偏差の値を求めてみよう。最小値，第 1 四分位点，メディアン，平均，第 3 四分位点，最大値を一括して与えるには summary(·) というコマンドを使う。

```
> MS <- c(52, 65,42,55,48,62,95,74,58,52)
> summary(MS)
 Min.   1st Qu.  Median    Mean   3rd Qu.    Max.
42.00    52.00    56.50    60.30    64.25    95.00
> var(MS)        % 分散
[1] 230.4556
> sd(MS)         % 標準偏差
[1] 15.18076
```

(4) ヒスとグラム等のグラフ表示．幹葉表示を描くには stem(·) を使う。

```
> stem(MS, scale=2)
```

グラフを描画するには，まず x11() と入力し，作図デバイスを立ち上げる。続いて，ヒストグラム，箱ひげ図を表示してみる。

```
> x11()
> hist(MS)
> boxplot(MS,range=0)
```

(5) 相関係数． 6 人の生徒の (数学, 理科) の得点が，(52, 43), (80, 75), (45, 44), (70, 65), (53, 58), (58, 55) で与えられるときに，数学と理科の得点の相関係数を求めてみよう。

```
> sugaku <- c(52,80,45,70,53,58)
> rika <- c(43,75,44,65,58,55)
> cor(sugaku,rika)        % 相関係数
```

数学と理科の得点データを x-y 平面にプロットしてみる。グラフィックスの出力デバイスが立ち上がっていないときには，x11() と入力して画面を立ち上げておく。

```
> plot(sugaku, rika, xlab="MATHEMATICS", ylab="SCIENCE",
        xlim=c(35,85), ylim=c(35,85))
```

と入力すると，得点データが x-y 平面にプロットされる。

(6) 回帰分析．上のデータを用いて理科の得点を数学の得点で説明するという回帰分析を行ってみる。次のうに入力すると，x-y 平面上に回帰直線とともにデータがプロットされる。また回帰係数の推定値など回帰分析の結果を出力するには summary(·) というコマンドを使う。

```
> reg <- lm(rika∼sugaku)        % 理科 (y), 数学 (x) を指定する
> abline(reg)        % 回帰直線を引く
> summary(reg)        % 回帰分析の結果を表示する
> x1 <- c(1,2,3,4,5)
> y <- c(11,23,36,42,55)
> reg <- lm(y∼x1)        % 回帰
>plot(x1,resid(reg))     % 横軸 x1 として 残差をプロットする。
```

重回帰分析やロジスティック回帰分析は次のように入力する。

```
> x2 <- c(1,4,5,9,15)        % 回帰直線を引く
> reg1 <- lm(y∼x1+x2) % 回帰分析
> z <- c(0,0,0,1,1)        %2 値データ
> reg2 <- glm(z∼1+x1+x2,binomial(logit)) % ロジスティック回帰
```

(7) 偏相関係数の計算．16 個の 3 変数のデータ (0, 1, 2), . . ., (3, 8, 8) が与えられてとき，偏相関を計算してみよう。16 行 3 列の行列の形でデータを次のように入力する。

```
> pdata <- matrix(c(0,0,0,0,1,1,1,1,2,2,2,2,3,3,3,3,
1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,
2,4,0,2,4,6,2,4,6,8,4,6,8,10,6,8),16, 3)
```

この行列の 2 列目と 3 列目の相関係数の計算と，(1 列目，2 列目), (2 列目，3 列目), (3 列目，1 列目) の各ペアーのデータをプロットするには，次のようにする。

```
> cor(pdata[,2], pdata[,3])
> pairs(pdata)
```

偏相関を計算する簡単な方法は，psych というパッケージを利用することである。このパッケージが組み込まれていないときには，メニュー「パッケージ」の中の「パッケージの読み込み」をクリックし，psych をクリックすると組み込まれる。library(psych) と入力して，利用可能な状態にする。以下は，1 列目の影響を取り除いた 2 列目と 3 列目の偏相関，3 列目の影響を取り除いた 1 列目と 2 列目の偏相関，2 列目の影響を取り除いた 1 列目と 3 列目の偏相関を計算している。

```
> library(psych)
```

> partial.r(pdata, c(2,3),c(1))

> partial.r(pdata, c(1,2),c(3))

> partial.r(pdata, c(1,3),c(2))

(8) 確率と分位点の値. 正規分布の分布関数 $\Phi(a) = \int_{-\infty}^{a} (2\pi)^{-1/2} \exp\{-x^2/2\} dx$ を求めたいときには，pnorm(a) を用いる。また $\Phi(c_a) = a$ となる分位点 c_a の値を求めるには qnorm(a) を用いる。例えば次のようになる。

> pnorm(0)

[1] 0.5

> pnorm(-1)

[1] 0.15

> pnorm(-2)

[1] 0.02

> qnorm(0.5)

[1] 0

> qnorm(0.95)

[1] 1.64

> qnorm(0.975)

[1] 1.96

> qnorm(0.99)

[1] 2.33

> qnorm(0.995)

[1] 2.57

(9) ローレンツ曲線とジニ係数. 10 個のデータ 100, 100, 95, 90, 90, 90, 80, 50, 30, 20 のローレンツ曲線を描くには，例えば次のように入力する。ここで lenght(data) は data の要素の個数，sort(data) は data を小さい順に並べ直すこと，cumsum(y1) は y1 の累積和，seq(0, 1, length=n+1) は区間 [0, 1] を n 等分したときの区切りの点を並べたものを意味する。具体的には, cumsum(1,2,3)=(1, 3, 6) となる。

> data <- c(100, 100, 95, 90, 90, 90, 80, 50, 30, 20)

> n <- length(data)

> y1 <- sort(data)

> y2 <- cumsum(y1)

> y <- c(0, y2/max(y2))

> x <- seq(0, 1, length=n+1)

> plot(x, y, type="l", ann=F,xlim=c(0, 1), ylim=c(0, 1))

> plot(segments(0,0,1,1), lty=”dotted”, add=TRUE)

集中度を表すローレンツ曲線は data を大きい順に並べ直すので sort(data, decreasing=T) を用いる。またジニ係数は次のようにして求められる。

> 2*sum(x-y)/n

(10) スクリプトの利用．R の入力画面に直接入力するのではなく，メニュー「ファイル」にある「新しいスクリプト」を作成し，そこに R の命令を記述するとよい。実行したい命令文だけをマウスで範囲選択し，マウスの右ボタンで実行すると，R の入力画面の方に結果が出力される。スクリプトは自動保存されないので，気がついたときに保存しておいた方がよい。最後に，R を終了するには

> q()

と入力する。

[4] 参考文献 基本的な使い方が上で説明したが，他にも様々なコマンドが用意されており，R でできることは実に幅広い。インストールの仕方や R の基本的な使い方，様々な統計手法や確率の計算，グラフの使い方などについては，R の解説書を参考にしてほしい。例えば，小暮厚之 (2009)，金明哲 (2007) があげられる。これから R を利用しようとして実際にコードを打つ際には以下のサイトの資料が参考になる。http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html

2　エクセル入門

[1] エクセルについて エクセルはマイクロソフトから販売している表計算ソフトで広く利用されている。エクセルの表画面の上に直接データを入力し，様々な関数や分析ツールを適用することにより分析結果を求める。分析ツールの中には回帰やヒストグラムなど利用しやすいものが含まれているので，エクセルのアドイン・プログラムである [分析ツール] をインストールして利用することが望ましい。以下の説明は [エクセル 2010] に関する内容である。

[2] 分析ツールを読み込む まず，アドイン・プログラムである [分析ツール] を次のようにして読み込む。

(1) [ファイル] タブをクリックし，左メニューの下にある [オプション] をクリックする。

(2) [アドイン] をクリックし，[管理] ボックスの一覧の [Excel アドイン] をクリックする。その右隣の [設定] ボタンをクリックする。[有効なアドイン] ボックスが現れるので，[分析ツール] チェックボックスをオンにし，[OK] をクリックする。

(3) 分析ツールを読み込むと，[データ] タブの [分析] で [データ分析] を利用できるようになる。

[3] エクセルを使ってみる

(1) データの入力．列 A の 1 行にはデータの名前を記入する。具体的なデータの数値は 2 行目の A2 から順次書き入れる。例えば，10 個のデータが A2 から A11 に入力されているとする。

(2) 平均，分散．C1 のセルに 10 個のデータの合計を表示させるには，セル C1 をクリックし，そこに「=SUM(A2:A11)」と入力しリターンキーを押すと合計した値が表示される。その他様々な特性値を求めることができる。平均を求めたいときには，上の SUM を AVERAGE に代えればよい。メディアン，モード，分散，標準偏差は MEDIAN, MODE, VARP, STDEVP を用いる。

(3) データの変換．A2 から A11 に入力されているデータを自然対数で変換したいときには，セル D2 をクリックして「=」を入力し，[数式] タブの [関数ライブラリ] で [関数/三角] をクリックし，[LN] をクリックする。A2 と入力し [OK] をクリックすると，自然対数で変換した数値がセル D2 に出力される。次に，セル D2 をクリックし，マウスの右ボタンを押して [コピー] をクリックする。セル D3 からセル D11 をドラッグし，マウスの右ボタンをクリックして，[貼り付けのオプション] の中の [fx] をクリックすると，A3 から A11 の数値を自然対数で変換した数値が D3 から D11 に出力される。他のデータの変換も同様に行うことができる。

(4) 基本統計量．(2) のようにして平均や分散を求めることができるが，[分析ツール] を用いると，様々な基本統計量を一度の計算してくれる。[データ] タブの [分析] で [データ分析] をクリックし，[分析ツール] ボックスの中の [基本統計量] をクリックし，[OK] ボタンをクリックする。データが A2 から A11 に入力されているときには，[基本統計量] の中の [入力範囲] をクリックしてから，A2 から A11 をドラッグすると，データが指定される。下の [出力先] をクリックして，出力したい場所のセルを指定する。その下の [統計情報] ボックスにチェックを入れて，[OK] ボタンをクリックすると，平均，分散などの基本統計量の値が出力される。

(5) ヒストグラム．[データ] タブの [分析] で [データ分析] をクリックし，[分析ツール] ボックスの中の [ヒストグラム] をクリックし，[OK] ボタンをクリックする。データが A2 から A11 に入力されているときには，[ヒスとグラム] の中の [入力範囲] をクリックしてから，A2 から A11 をドラッグすると，データが指定される。下の [出力先] をクリックして，出力したい場所のセルを指定する。その下の [グラフ作成] ボックスにチェックを入れて，[OK] ボタンをクリックすると，ヒストグラムが表示される。

(6) 確率の計算．例えば正規分布の確率は, =NORMSDIST(a1) とすると a1 に対応する分布関数の値, =NORMSINV(b1) とすると b1 に対応する分位点を

求めることができる。二項分布 Bin(n,p) の場合には数値 x に対応する確率は =BINOMDIST(x,n,p,TRUE) と入力すればよい。

(7) 回帰分析. [データ] タブの [分析] で [データ分析] をクリックし，[分析ツール] ボックスの中の [回帰分析] をクリックし，[OK] ボタンをクリックする。Y のデータが A2 から A11 に，X のデータが B2 から B11 に入力されているときには，[回帰分析] の中の [入力 Y 範囲] をクリックしてから，A2 から A11 をドラッグし，また [入力 X 範囲] をクリックしてから，B2 から B11 をドラッグする。下の [一覧の出力先] をクリックして，出力したい場所のセルを指定する。[OK] ボタンをクリックすると，回帰分析の結果が出力される。

[4] 参考文献 分析ツールの中には他にも様々な統計手法が組み込まれている。詳しくは，例えば，森棟公夫，他 (2008) の 3 章を参照してほしい。

3 数学の基本的な知識

(1) 記号の定義と基本事項. 自然対数の底 e を

$$e = \lim_{x\to\infty}(1+x^{-1})^x = \lim_{x\to 0}(1+x)^{1/x}$$

の極限で定義する。実数 λ に対して $\lim_{x\to\infty}(1+\lambda x^{-1})^x = e^{\lambda}$ が成り立つ。例えばエクセルで $n = 10, 100, 1000, \cdots, = (1+1/n)^n$ として確かめてみよう。

実数 a に対して $(a)_k = a(a-1)\cdots(t-k+1)$ とし，2 項係数 ${}_nC_k$ を

$$\left(\begin{array}{c} a \\ k \end{array}\right) = \frac{(a)_k}{k!}$$

と書くこともある。n が正の整数のときには $(n)_k = n!/(n-k)!$, $(n)_n = n!$ となる。 $0! = 1$ と定義する。また

$$\left(\begin{array}{c} n+1 \\ k \end{array}\right) = \left(\begin{array}{c} n \\ k \end{array}\right)\left(\begin{array}{c} n \\ k-1 \end{array}\right), \quad \left(\begin{array}{c} -n \\ k \end{array}\right) = (-1)^k\left(\begin{array}{c} n+k-1 \\ k \end{array}\right)$$

が成り立つ。

スターリングの公式：n が大きいときに近似式

$$n! \approx \sqrt{2\pi}e^{-n}n^{n+1/2}$$

が成り立つ。

数学的帰納法により次の等式が示される。

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \qquad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6},$$

$$\sum_{k=1}^{n} k^3 = \Big[\frac{n(n+1)}{2}\Big]^2, \quad \sum_{k=1}^{n} k^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$$

次の等式は容易に確かめられる。

$$\sum_{k=0}^{n} ar^k = a\frac{1-r^{n+1}}{1-r}$$

(2) 微積分.

微分：関数 $f(x)$ の $x = x_0$ における微分係数を

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

で定義する。微分係数が存在するとき $f(x)$ は $x = x_0$ で微分可能であるという。区間 I の任意の点で微分可能であるとき $f(x)$ は I で微分可能であるという。$f'(x) = (d/dx)f(x)$ と書いて導関数という。$f'(x)$ が連続であるとき，$f(x)$ は連続微分可能であるという。例えば, $(e^x)' = e^x$, $(a^x)' = (\log a)a^x$, $a > 0$, $(logx)' = 1/x$, $x > 0$, $(\sin x)' = cosx$, $(\cos x)' = -\sin x$, $(\tan x)' = 1/\cos^2 x$ となる。

ロピタルの定理： $\lim_{x \to a} f(x) = \lim_{x \to a} g(x) = 0$ となる連続微分可能な関数 $f(x)$, $g(x)$ に対して，$\lim_{x \to a} f'(x)/g'(x)$ が存在するとき次の等式が成り立つ。

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)}$$

統計学では、例えば連続型の確率分布関数の導出や期待値・分散の計算に際して関数の積分 (integral) 演算を行う必要が出てくる。f が区間 I 上で連続とし、f の任意の原始関数を F とする。すなわち、

$$F'(x) = f(x), \quad x \in I$$

とすれば、

$$\int_a^b f(x)dx = F(b) - F(a), \quad a, b \in I$$

が成り立つ。

正規分布の密度関数: 正規分布 $N\left(\mu, \sigma^2\right)$ の密度関数は

$$n\left(x|\,\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (-\infty < x < +\infty)$$

とすると、正規分布にしたがう確率変数 X は任意の実数 y に対して

$$P(X \leq y) = \Phi(y) = \int_{-\infty}^{y} f(x)dx$$

となり、$\int_{-\infty}^{+\infty} n\left(x|\,\mu, \sigma^2\right)\, dx = 1$ となることが知られている (参考文献を参照)。

4 参考文献

- 金明哲 (2007) R によるデータサイエンス. 森北出版
- 小暮厚之 (2009) R による統計データ分析入門. 朝倉書店
- 久保川達也・国友直人 (2016) 統計学. 東京大学出版会
- 国友直人 (2015) 応用をめざす数理統計学. 朝倉書店
- 森棟公夫，照井伸彦，中川満，西埜晴久，黒住英司 (2008) 統計学. 有斐閣
- 竹村彰通 (1997) 統計. 共立出版