R 練習問題

作成者：五十嵐未来

1　基礎編

問題 1 - 1

$5 + 9$ を計算し、結果をコンソールに表示せよ。

問題 1 - 2

15^9 を計算し、結果をコンソールに表示せよ。

問題 1 - 3

$100/7$ を計算し、結果をコンソールに表示せよ。

問題 1 - 4

$a = 9, b = 7, c = 4$ のとき、$(a + b)^c$ を計算し、結果をコンソールに表示せよ。

問題 1 - 5

`"Hello world!"`という文字列が格納されたオブジェクトを用意し、それを次のようにコンソールに表示せよ。

```
[1]  "Hello world!"
```

問題 1 - 6

`"Hello"`と`"world!"`が格納されたオブジェクトをそれぞれ用意し、それらを次のように一行でコンソールに表示せよ（paste 関数を用いて二つのオブジェクトを結合すること）。

```
[1] "Hello world!"
```

問題 1 - 7

"Hello"と"world!"が格納されたオブジェクトをそれぞれ用意し、それらを-でつないでコンソールに表示せよ。

```
[1] "Hello-world!"
```

問題 1 - 8

1 から 10 までの数字が格納されたベクトルを用意し、それらをカンマ区切りでコンソール上に表示せよ。

```
[1] "1,2,3,4,5,6,7,8,9,10"
```

問題 1 - 9

$5, 3$ が代入されているオブジェクト a, b を用意し、それらの和を次のような形式でコンソールに表示せよ（[1] "a+b=8"の形ではないことに注意）。

```
a+b=8
```

問題 1 - 10

ベクトル a を $a = (38, 14, 25, 62, 8, 71, 64, 29, 92)$ とし、a の平均と標準偏差を求め、有効数字 3 桁に丸めた次のような形式でコンソール上に表示せよ。

```
Mean of a is 44.78
SD of a is 28.66
```

2　関数

問題 2 - 1

ベクトルを入力するとその総和を出力する関数を作成せよ。

問題 2 - 2

第 1 引数に入力ベクトル、第 2 引数に平均又は標準偏差のどちらを実行するかを表す文字を指定することで、第 2 引数に応じてベクトルの平均値又は標準偏差を出力する関数を作成せよ。

問題 2 - 3

行列を入力すると、各行の最大値とその位置をコンソール上に表示する関数を作成せよ。
例えば、$A = \begin{pmatrix} 3 & 5 & 6 \\ 7 & 2 & 4 \\ 8 & 8 & 4 \end{pmatrix}$ を作成した `each_row_max` 関数に入力すれば以下のように出力される。

```
> each_row_max(A)
row: 1 max: 6 loc: 3
row: 2 max: 7 loc: 1
row: 3 max: 8 loc: 1, 2
```

問題 2 - 4

平均 0、標準偏差 1 の正規乱数を一つ発生させ、ベクトルに昇順で格納し続ける関数を作成せよ。ただし、発生した正規乱数が絶対値で 2 を超えていた場合は繰り返しをやめ、「Error: generated value is over 2 or -2」という文をエラーとして表示せよ。

3　ベクトル

問題 3 - 1

1 から 100 までの交差が 5 の数列を作成せよ。

問題 3 - 2

paste 関数と rep 関数を駆使して下の行列を作成せよ。

```
     [,1]           [,2]           [,3]           [,4]           [,5]
[1,] "user1/itemA"  "user1/itemB"  "user1/itemC"  "user1/itemD"  "user1/itemE"
[2,] "user2/itemA"  "user2/itemB"  "user2/itemC"  "user2/itemD"  "user2/itemE"
[3,] "user3/itemA"  "user3/itemB"  "user3/itemC"  "user3/itemD"  "user3/itemE"
[4,] "user4/itemA"  "user4/itemB"  "user4/itemC"  "user4/itemD"  "user4/itemE"
[5,] "user5/itemA"  "user5/itemB"  "user5/itemC"  "user5/itemD"  "user5/itemE"
```

問題 3 - 3

標準正規乱数が 100 個格納されたベクトルを用意し、正である要素のみについて和を計算せよ。

問題 3 - 4

1 から 100 までの値を取り得る整数乱数を 100 個用意し、5 の倍数のみを取り出して表示せよ

問題 3 - 5

1 から 100 までの値を取り得る整数乱数を 100 個用意し、「2 の倍数かつ 3 の倍数」または「5 の倍数」を NA に置き換えよ。

4　行列・配列・リスト

問題 4 - 1

以下の手順で行列を作成せよ。

1. 行数・列数が 10 の空行列 A を用意する
2. 取り得る値が 1 から 10、確率は全て等確率、試行回数が 1 の多項分布から独立に 2 回乱数を発生させる
3. 行列 A の要素のうち、2 で得た乱数に対応する行番号・列番号の要素に 1 を加算する
4. 2 と 3 を 10000 回繰り返す

問題 4 - 2

1 から 100 までの整数を次のように保存した行列を作成し、偶数列のみを降順に並び替えよ。

問題 4 - 3

各要素が独立した正規分布に従う、10×10 行列に対して特異値分解を行い、それにより復元した行列と元々の行列が一致することを確認せよ。

問題 4 - 4

各次元数が (3,3,2) で、各要素が独立した正規分布に従う 3 次元配列を作成し、3 次元目の 1 つ目の要素の転置行列と、2 つ目の要素の逆行列をリスト形式で格納した（要素名はそれぞれ `trans` と `inverse` とする）オブジェクトを返す関数を作成せよ。

問題 4 - 5

長さが 3 の初期化したリスト（各要素が `NULL`）を作成し、名前をそれぞれ `original`、`mean`、`max` とする。標準正規乱数を 10 個発生させ、その乱数のベクトル・平均値・最大値を各要素に格納する。このリストに対して次の 3 つの操作を行え。

1. 乱数ベクトルをベクトルで取り出す
2. 乱数ベクトルをリストで取り出し、リスト形式になっていることを確認する
3. 平均値と最大値を同時に取り出す

5　apply 系関数

問題 5 - 1

各次元数が (4,3,2) で、各要素が独立した正規分布に従う 3 次元配列を作成し、2 次元目の方向に平均を取った 4×2 行列を得よ（apply 関数を使用すること）。

問題 5 - 2

a-z と A-Z の計 52 種類の文字列から、1 から 10 個重複を許したサンプリング抽出（個数は整数乱数で決定する）を 10 回繰り返し、各サンプルをリスト形式で格納する。このリストの各要素数をベクトルで表せ（sapply 関数を使用すること）。

問題 5 - 3

上で作成したリストの各要素に含まれる小文字（a-z）を取り出せ（lapply 関数を使用すること）。

問題 5 - 4

標準正規乱数 100 個のベクトルと、同じ次元の 0 ベクトルを用意し、各要素を比較して大きい方を格納したベクトルを作成せよ（mapply 関数を使用すること）。

6　データフレーム

問題 6 - 1

下表の要素が格納されたデータフレームを作成せよ。

	product	price	sales
1	apple	150	23
2	apple	152	24
3	peach	156	28
4	peach	156	21
5	grape	155	27

問題 6 - 2

R に組み込まれたデータセットである `iris` データ（アヤメに関する観測データ）を読み込み、以下の操作を実行せよ。

1. 各 `Species` についてそれぞれ平均値を算出する

2. `Sepal.Length` が 5.0 以上かつ `Sepal.Width` が 3.0 以上の行を抽出し、各 `Species` の数を数える
3. 架空の変数 `Leaves`（葉の数）が、`iris` の各サンプルに対して観測されたと仮定して適当なベクトルを作成し、データフレームに加える
4. アヤメの新種 `newayame` が発見されたと仮定し、各変数について適当な観測データを作成しデータフレームに加える

問題 6 - 3

あるユーザーのツイートデータ（`https://igarashim.github.io/data/tweet.tsv` からダウンロード可能）をデータフレームとして読み込み、以下の操作を実行せよ。ただし、csv ファイルに含まれている変数は `time, tw_id, fav, RT, text` であり、それぞれツイート投稿時刻、Tweet ID、いいね数、RT 数、本文である。そのため、読み込み時にそれぞれの変数を文字列、文字列、整数、整数、文字列として読み込むことが望ましい。また、tsv ファイルであることから分かるように Tab 区切りのテキストファイル形式であり、文字コードには UTF-8 が使用されている。

1. いいね、または RT が 1 回以上されているツイートのみを抽出して新しいデータフレームを作成し、読み込んだ時と同様の形式の tsv ファイルとして任意の場所に出力する
2. `time` 変数を文字列型から日付型へと変更し、いいねが最も多かったツイートと、RT が最も多かったツイートの投稿時刻の差を計算する
3. `tw_id` 変数を文字列から数値型へと変更し、いいねが最も多かったツイートと、RT が最も多かったツイートの Tweet ID をコンソール上に表示する（ただし、`9.79924e+17` のような指数表現ではなく、`979923999143682048` のように数値で表示する）

7　文字列操作

問題 7 - 1

a-z の 26 種類のアルファベットを全て結合した文字列を作成し（アルファベットの間にはスペースなどの区切り文字を入れない）、その文字列の長さを数えよ。

問題 7 - 2

a-z と A-Z の計 52 種類のアルファベットを全てカンマ区切りで結合した文字列を作成せよ。ただし z と A の間だけはカンマではなく改行文字を挿入すること。

```
a,b,c,...,x,y,z
A,B,C,...,X,Y,Z
```

問題 7 - 3

問題 6 - 3 で読み込んだデータフレームの text 変数に対して、文字列ベクトルから、スペースで単語に区切ったリストへと変更せよ。

```
[[1]]
[1] "StarCraft" "is" ... "a..." "https://t.co/brr8imkDDk"

[[2]]
[1] ...
```

問題 7 - 4

問題 6 - 3 で読み込んだデータフレームの全ツイートの text 変数について、is という文字が使われている回数を数えよ。is が使われてさえいればよいため、disappoint や finished などもカウントに含めてよい。ただし IS のように大文字は含めない。

問題 7 - 5

問題 6 - 3 で読み込んだデータフレームの全ツイートの text 変数について、ユーザー名（@+ 数字またはアルファベットの文字列）を消去せよ。

問題 7 - 6

問題 6 - 3 で読み込んだデータフレームの全ツイートの text 変数に対して、全ての文字を小文字に直し、ユーザー名・URL 文字列（http://または https://から始まる文字列）・記号・数字を削除したうえで、文字列ベクトルから、スペースで単語に区切ったリストへと変更せよ。

```
[[1]]
[1] "starcraft" "is" ... "a"

[[2]]
[1] ...
```

8　図の作成

以下では、あるデータセットに対して、パッケージを用いない作図と ggplot2 というパッケージを用いた作図の 2 種類を行う。ここで用いられるデータセットは全て https://igarashim.github.io/data/data.RData（オブジェクト名：df）からダウンロード可能であるため、適宜保存し使用せよ。

問題 8 - 1

R 組み込みのデータセットである `AirPassengers` は、クラシックスボックス&ジェンキンス航空会社の、1949 年から 1960 年の国際線旅客数に関する月次データである。これを加工したデータセット `Problem8-1` を用いて、以下の折れ線グラフを作成せよ。ただし、横軸が日付（変数名：`Year`）、縦軸が乗組員である（`Passengers`）。

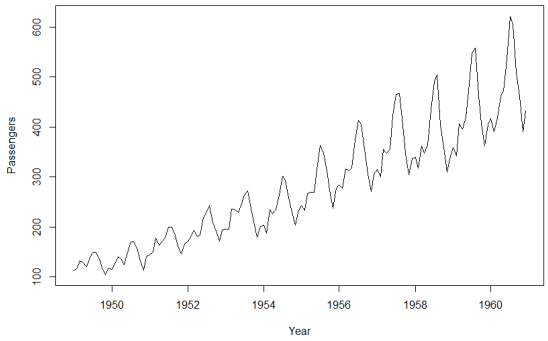

図 1　R による作図

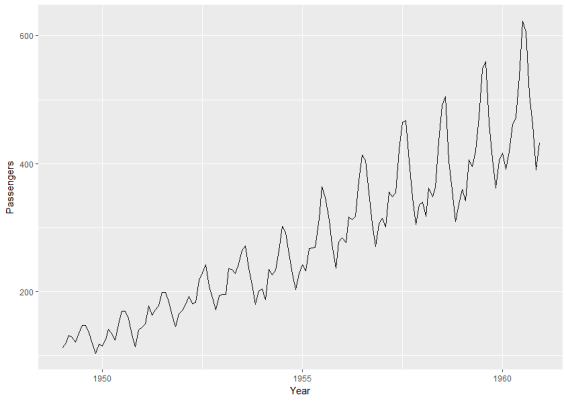

図 2　ggplot による作図

問題 8 - 2

乱数によって発生させたデータ（`x,y`）からなるデータセット `Problem8-2` を用いて、以下の散布図を作成せよ。ただし、白抜き（黒）丸がデータの散布図、直線が `y` を `x` に回帰した際の 95% 信頼区間付き回帰直線を表す。

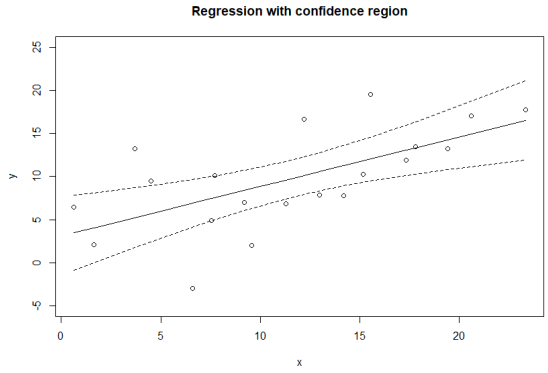

図 3　R による作図

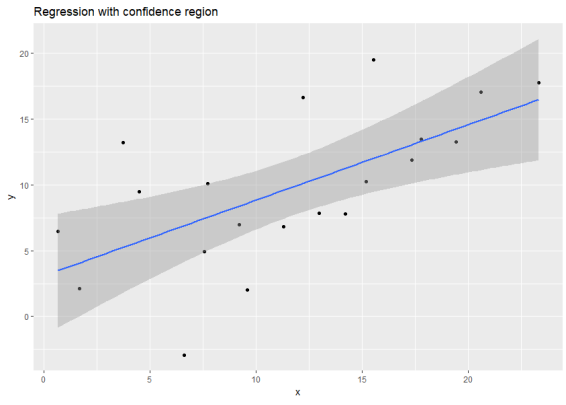

図 4　ggplot による作図

問題 8 - 3

R 組み込みのデータセットである `Titanic` は、タイタニック号の乗客の生存と、性別・年齢・乗船クラスをまとめたデータである。オリジナルデータセットでは、生存変数が二値であるが、これから生存率を計算する加工を施したデータセット `Problem8-3` を用いて、以下の折れ線グラフを作成せよ。ただし、横軸が乗船クラス（`Class`）、縦軸が生存率（`Survived`）であり、色は性別（`Sex`）を、データ点の形状は年齢（`Age`）をそれぞれ表している。

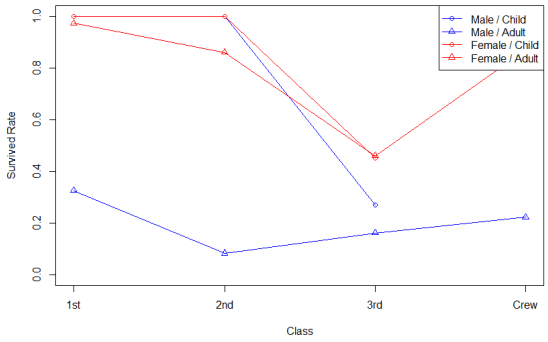

図 5　R による作図

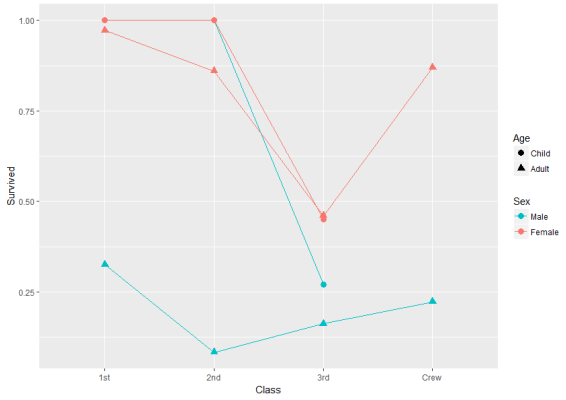

図 6　ggplot による作図

問題 8 - 4

R 組み込みのデータセットである `iris` について、アヤメの種類ごとに各変数の平均値を計算したデータセット `Problem8-4` を用いて、以下の棒グラフを作成せよ。ただし、横軸はアヤメの種類（`Species`）、縦軸はがく片の幅の平均値（`Sepal.Width`）である。

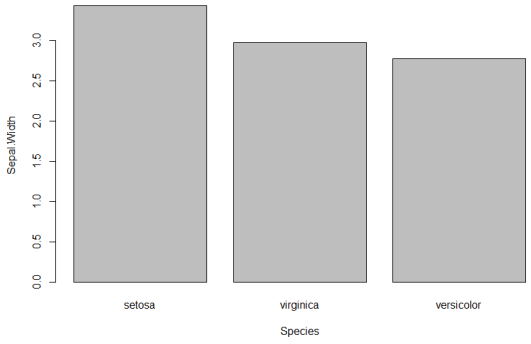

図 7　R による作図

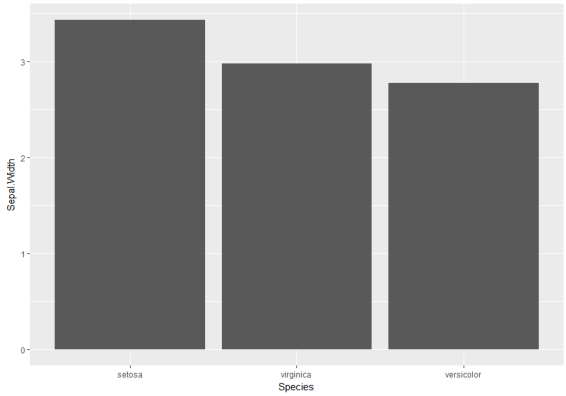

図 8　ggplot による作図

問題 8 - 5

問題 8 - 4 と同じデータセットを用いて、以下の棒グラフを作成せよ。ただし、横軸はアヤメの種類（`Species`）、縦軸は各変数の平均値（`Sepal.Length`,`Sepal.Width`,`Petal.Length`,`Petal.Width`）である。

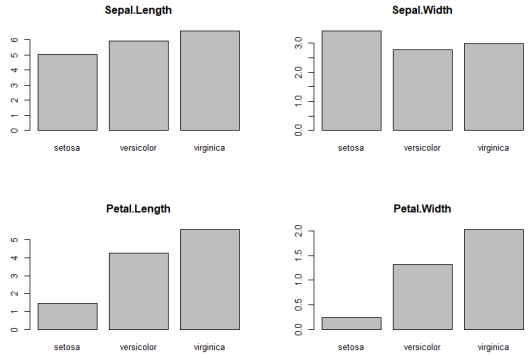

図 9　R による作図

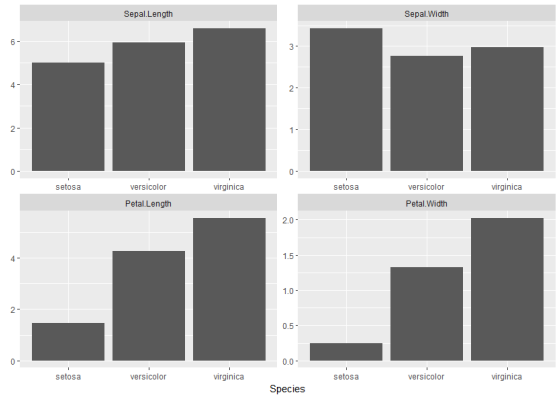

図 10　ggplot による作図

問題 8 - 6

乱数によって発生させたデータ（`A`,`B`,`C`）からなるデータセット `Problem8-6` を用いて、以下のヒストグラムを作成せよ。ただし、色の透過は 50% であり、それぞれに使われている色は、`A` が#F8766D、`B` が#00BFC4、`C` が#7CAE00 である。

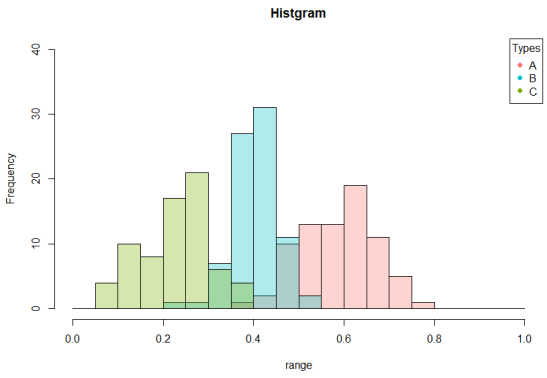

図 11　R による作図

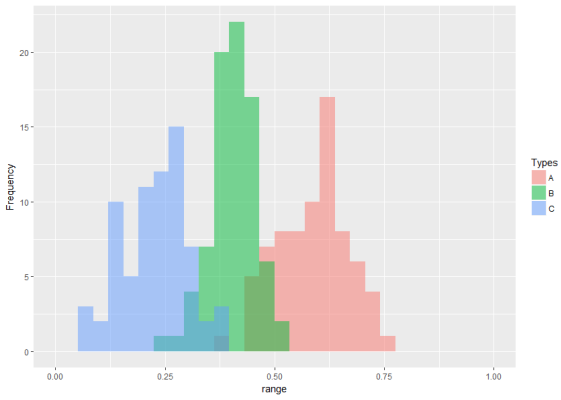

図 12　ggplot による作図

9　応用編

問題 9 - 1

フリーダウンロードが可能なデータセットを豊富に扱っている UCI Machine Learning Repository（`https://archive.ics.uci.edu/ml/index.php`）へアクセスして適当なデータセットをダウンロードし、以下の 3 種類の方法で重回帰分析を行え。

1. `lm` 関数を用いて重回帰（QR 分解を利用した最小二乗法が使われている）を行い、結果を要約する
2. 最尤法による重回帰を行い、結果を要約する
   （方針：データとモデルを入力すると、対数尤度を返す関数を作成し、`optim` 関数を用いて最大化計算を行う）
3. ギブスサンプリングによって重回帰係数を推定する関数を自作し、上記二つの方法による推定値と比較する

問題 9 - 2

映画.com の映画レビューページ（例「未来のミライ」`https://eiga.com/movie/88326/review/`）から評点（0.0 から 5.0 まで 0.5 刻みの 11 段階）とレビュー本文を収集せよ。ただし、一定の Web ページへの集中アクセスを防ぐため、`Sys.sleep` 関数等を用いた処理を必ず入れること。
（方針：`rvest` パッケージを用いて、html を xml 形式のテキストとして取得し、html のタグなどを頼りに欲しい情報を取得するのが一般的である）

問題 9 - 3

問題 9 - 2 で取得したレビューデータに対して、Latent Dirichlet Allocation（LDA）を適用し、レビューごとのトピック分布を説明変数に、評点（4.0 以上を高評点とした二値データ）を目的変数とした分類を行え。なお、分類手法には何を用いても良いとする（ロジスティック回帰・サポートベクターマシン・深層学習などさまざまな分類手法が適用可能である）。