R

R.data

サンプルデータ

■ R をインストールした時点で、サンプルデータもインストールされている。

・ data() で一覧表示される。

■一覧

・https://www.trifields.jp/r-sample-data-491
・https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html
・https://www.openintro.org/data/

■その他

・https://r-dir.com/reference/datasets.html

■そのデータの説明は、データ名の前に ? を付けて「実行」すると、Help に表示される

■言語データのサンプル

・例

・与格交代する動詞の例

・ ?verbs とすると、Help に説明が出る

■身長と体重と性別のサンプルデータ

https://helloacm.com/the-machine-learning-case-study-how-to-predict-weight-over-heightgender-using-linear-regression/

・どんなデータか、その概要を知る
  ・ str() でデータの概要が表示される

```
> str(kimatsu)
'data.frame':   18 obs. of  2 variables:
 $ kokugo : num  83 45 73 50 22 67 77 89 66 90 ...
 $ suugaku: num  90 55 90 43 33 55 48 98 56 75 ...
```

・グラフにして様子を見る
  ・ plot()

■ランダムサンプリング sample(数値の範囲 , サンプル数)

```
 sample(1:100, 50)
  [1] 68 39  1 34 87 43 14 82 59 51 85 21 54 74  7 73 79 37 83 97 44 84 33 35 70 96 42 38 20 28 72 80
40 69 25 99 91 75
 [39]  6 24 32 94  2 45 18 22 92 90 98 64
```

・これで、ランダムな数字を必要なだけ出しておいて、
・それを「要素番号」として指定することで、
・データフレーム中のデータをランダムに選び出せる。
・ランダムに選んだ残りは - をつけて指定すればよい

■ rnorm(**個数** , **平均** , **標準偏差**)

・ダミーのデータを作り出す。（何か試してみたいときに）
・平均と標準偏差を省略して個数だけ指定した場合、平均０、標準偏差１の正規分布データからのランダム抽出
　・平均５０で、標準偏差１０のデータを４０人分
　　・やってみてヒストグラム作るとわかりますが、４０人程度では、ランダムだと、「え、これが」と思われるような分布になります。４００でもまだまだ、４０００になるとまあまあ、４００００だとかなり安定した正規分布になりますね。
　　・同じグラフでも、boxplot だと、４０でもきれいにつりあいの取れたグラフになります。

R **で扱うデータの種類**

・図でわかりやすい　↓

https://cell-innovation.nig.ac.jp/surfers/vector_difference.html

■**ベクトル**

・プログラミングで言う「配列」
・「変数名」をつける
・要素が入っている
・変数名 <- c(要素をカンマで区切って並べる)
　・例：国語と数学の得点
　・kokugo <- c(83, 45, 73, 50, 22)
　・suugaku <- c(90, 55, 90, 43, 33)
・要素は前から順番に位置が決まっている
・変数名をタイプすると内容が表示される

```
> kokugo
[1] 83 45 73 50 22
```

■**行列** matrix

■**リスト**

■**データフレーム**

・プログラミングで言う「多次元配列」
・行が項目
・列が測定値
・複数のベクトルを合わせてデータフレームを作ることもできる
・変数名 <- data.frame(ベクトル名をカンマで区切って並べる)
　・例：期末試験の得点（国語と数学）
　・kimatsu <- data.frame(kokugo, suugaku)
・変数名をタイプすると内容が表示される

```
>kimatsu
```

```
  kokugo suugaku
1     83       90
2     45       55
3     73       90
4     50       43
5     22       33
```

・個々のベクトルを指定するときは　$ を使う
　・国語の得点　kimatsu$kokugo
　・数学の得点　kimatsu$suugaku

・逆に、あるデータフレームの特定の列を取り出して、別のデータフレームにする

```
icnaleMLS.MLT <- data.frame(icnale$MLS, icnale$MLT)
```

・何行何列あるか調べる

19282 行・10 列あることがわかる。

・カラム名を確認

・データフレームかどうか、確認

・データフレーム内の複数の列で並べ替え

```
data[order(data$pid, data$week),]
```

・pid のカラムで並べ替えてから、week のカラムで並べ替える
　・事前
　・事後

データの型

■ factor （因子）

・カテゴリー変数
・順序の有り無しに分けられる
　・順序なし factor()
　・順序あり ordered()

・Reference

https://sites.google.com/site/leihcrev/r/ordered-and-unordered-factors

■ factor の並び順を指定する

```
factor( データ , levels = c(" はじめ ", " つぎに ", " さいごに "))
```

```
f.data$students <- factor(f.data$students, levels=c("year2", "year3", "natives"))
```

データの型の変換

・as.vector()
・as.numeric()
・as.matrix()
・as.data.frame()
・unlist()　　list を vector に
・as.factor()

・すべての列を変換

■ factor 型の変換に注意

・ 数値をいったん factor 型にすると、factor のレベルを表す数字（1 から始まる）が付く
・ それを再度数値に変換すると、元の数字ではなく、factor のレベルの数字が数値になるので注意
・ いったん文字列にしてから戻す

```
as.integer(as.character(y.df$SL))
as.numeric(as.character(y.df$SL))
```

・ Reference https://a-habakiri.hateblo.jp/entry/2016/12/18/213416

データの型の確認

・dim()
・class()
・mode()
・typeof()

・summary()
・str()

■間違いやすい点のまとめ

https://cell-innovation.nig.ac.jp/surfers/R_point.html

事前にデータを作っておく

・表計算ソフトなどで、テキストファイル（タブ区切りや CSV）で保存しておく。
・一件一行
・一番上の行は、変数名
・欠損値は NA と記入しておく
・ファイルの読み込み
　・データフレーム名 <- read.table(choose.files())
　・これでウインドウが開くのでどのファイルを読み込むか指定する

■エクセルから「コピペ」することもできる

1. エクセル上でデータを一覧にしておく
    1. 一番上の行は見出しとして行名を入れておく。これが R でも使われるので、簡潔な記号にしておくのがよい。
2. 取り入れたいデータ部分をマウスでドラッグして選び、Ctrl+C とかでコピーする。
3. ウィンドウを R に切り替えて、
4.R のコンソールで以下の命令を打って、実行（エンターキーを押す）。

```
データフレーム名 <- read.delim("clipboard")
もしくは、
データフレーム名 <- read.table("clipboard", header=T)
```

・データフレーム名は、R の中でのデータの名前
・見出しの有り無しのデフォルトが、read.delim と read.table で違うので注意
・見出しの一部を変更するには、要素番号で見出しを指定して上書きする。

```
> names(KSL)[3] <- "SL"
```

・KSL というデータの３番目の見出しを SL に変更する例

■テキストファイル読み込み： readLines(), scan()

・一文一行ずつ読み込んで全体はベクトル。一文一行が一要素。

・readLines()

```
x <- readLines(choose.files())
```

・「incomplete final line」（）という警告が出る場合、warn=F とする。

```
x <- readLines(" ファイル名 ", warn=F)
```

・scan()

```
x <- scan(choose.files(), what="char", sep="\n")
```

■要素の比較　%in%

```
x %in% y
```

■欠損値を計算から外すオプション

```
na.rm=T
```

データの読み込み（ファイルから読み込む）

■ csv ファイル

・read.csv()
・tidyverse::read_csv()
　・こちらの方がトラブル少ない
　・デフォルトで、ヘッダーあり

■タブ区切りファイル

```
read.delim(" ファイル名 ")
```

・ヘッダーはデフォルトで TRUE になっている。

■読み込む際に、データの型をカラムごとに指定する： colClasses=c()

データの一部（サブセット）を取り出す：条件に当てはまるものだけを選ぶ

```
subset( データフレーム名 , 条件 )
```

・or 条件は |
・and 条件は &
・文字のマッチは ==" 文字 "
・否定は !=

■例

```
文長が 4 以上：  SL >= 4
文長が 4 以上 14 以下： SL >=4 & SL <= 14
```

■該当するデータの特定の列だけを出力

```
subset( データフレーム名 , 条件 , 出力列 )
```

・出力列の例

```
select = c(MHD, MDD)
```

```
KSL.sub <- subset(KSL, SL >=4 & SL <= 14, select = c(SL, MHD, MDD))
```

・文長が４以上１４以下に該当するデータのうち、SL と MHD と MDD のカラムだけを取り出す。

・単純に要素番号の指定だけでもできる

```
> jp2gram2 <- jp2gram[,3:4]
```

■特定の行を削除するには != で該当しないものを残せばよい

```
NP.dat.jp2b <- NP.dat.jp %>% subset(Group != "Am")
```

・しかし、これだと、Group に Am というレベルが残ったままになる。
・使ってないレベルを削除するのは droplevels()

```
NP.dat.jp2c <- droplevels(NP.dat.jp2)
```

データの選択

・Reference

https://kazutan.github.io/JSSP2018_spring/data_handling.html

・TOEIC スコアがないものを除外する

■条件を複数重ねる

・これで、TOEIC も TOEFL も両方スコアのあるデータだけになった。

データの一部（特定のカラム）を取り出す

・いくつものカラムが並ぶデータの特定のカラムだけを使いたい場合
  ・連続していれば、データフレーム [,2:4] とすれば、2 列目から 4 列目だけを取り出せる
  ・非連続している場合は、data.frame() で、カラム名を指定して取り出す

```
data.frame(x$ID, x$age, x$score)
```

データフレーム内のカテゴリー変数のレベル名の並び替え

・デフォルトだとアルファベット順になる。
・特定の順番に並べ替える

データの一部を除く [- 行番号 , - 列番号]

■削除する行・列番号にマイナスをつけて処理する。

・ １列目を除く データフレーム名 [, -1]
・ ２列目を除く データフレーム名 [, -2]
・ １行目を除く データフレーム名 [-1 ,]
・ ２行目を除く データフレーム名 [-2 ,]

・一列目にファイル名が入っていて、ファイル名以外の部分で、相関を見たいとき

```
chart.Correlation( データフレーム名 [ , -1])
```

■複数の行・列の指定は c() を使って範囲を指定して削除

・ ２列目から 4 列目を除く データフレーム名 [, -c(2:4)]

データの欠損値を埋める

```
データ [is.na( データ )] <- 値
```

```
t2b.dat[is.na(t2b.dat)] <- "jp"
```

・NA だったところに jp と入る

行に ID 番号をつける

```
cbind(ID=1:nrow( データ ), データ )
```

データに行番号をつける。（ランクの順位を数値として入れておくなど）

■行番号を rank というカラムに追加する

■行の names 属性を独立したカラムとして追加する

・これで、該当の項目を grep で検索できる
・therefore は 175 位

ファイルへの書き出し

```
write( 変数 , " ファイル名 ", ncol= 列数 )
```

・は、write.table() で書き出す。

```
write.table( データフレーム名 , " ファイル名 ")
```

・フィールド間が半角スペース

```
write.csv( データフレーム名 , " ファイル名 .csv")
```

・フィールド間がカンマ

・csv ファイルをエクセルで読み込む場合
・エクセルは、デフォルトは日本語文字コードは S-JIS
・ファイルに BOM という情報が付いていないと S-JIS と思って読み込む。ファイルが UTF-8 だと文字化ける。
・BOM をつけて csv ファイル出力するには、tidyverse パッケージの中の dplyr ライブラリーを使って write_excel_csv() 命令で保存する。

データの操作

■データフレームの縦結合： dplyr::bind_rows

・2 つ以上も OK

■データフレームの横結合： dplyr::bind_cols

・2 つ以上も OK

■変数を縦につなげる

```
rbind( 変数１ , 変数２ )
```

```
新しいデータフレーム <- rbind( データフレーム１ , データフレーム２ )
```

・カラム名が同じでないとエラーになる。「名前が以前の名前と一致しません」
・カラム名の確認
・二つのカラム名の比較： を使う

```
setdiff(colnames( データ１ ), colnames( データ２ ))
```

・カラム名の確認

```
colnames( データ )[4]
```

・カラム名の変更

```
colnames( データ )[4] <- " 新しい名前 "
```

・一度につなげられるのは二つ。
・三つつなげる場合は、二段階になる。

■変数を横につなげる

```
cbind()
```

■列を追加する

・空（NA）の値をもった列をつけ足す。

```
データフレーム $ 追加する列名 <- NA
```

データを縦横に一連の同じ処理をする

■行単位で処理（二つ目の引数 1）：たとえば、平均点を出す

```
apply(x, 1, 関数 )
```

■列単位で処理（二つ目の引数 2）：たとえば、総合点を出す

```
apply(x, 2, 関数)
```

■一部の列だけを集計して合計の列を付け足す

・２列目から９列目を足して、total の列を右端に付け足す

・数値データの部分だけ合計する（NA は除く）

```
mutate("Total" = rowSums(across(where(is.numeric)), na.rm=TRUE))
```

データのフォーマットを Wide format から Long format にする

・縦横で「集計」してあるものを、一行一件の Long format に変換する

■ gather()

■ pivot_longer()

個々のオブジェクトをファイルに保存

■保存

```
save( オブジェクト , file=" ファイル名 .Rdata")
```

■オブジェクトは、複数あってもよい

```
save( オブジェクト , オブジェクト , オブジェクト , file=" ファイル名 .Rdata")
```

■読み込み

```
load(" ファイル名 .Rdata")
```

読み込んだ結果、もともとのオブジェクトがもともとのオブジェクト名で復元される。

読み込むときに別名にしたいとき

■保存

```
saveRDS( オブジェクト , file=" ファイル名 .rds")
```

■復元

```
readRDS(file=" ファイル名 .rds")
```

```
別名 <- readRDS(file=" ファイル名 .rds")
```

data.frame から取り出したものを matrix に変換

・二行目の２から１７番目の要素だけ取り出す

```
x <- fragJ11[2,2:17]
```

・マトリックスに変換

```
y <- as.matrix(x)
```

・見出しに NULL を入れて見出しをなくす

```
dimnames(y) <- NULL
```

・data.matrix() もある。

行列を入れ替える： t 関数

Box Plot Diagram to Identify Outliers

boxplot と外れ値の説明　↓　わかりやすい
https://www.whatissixsigma.net/box-plot-diagram-to-identify-outliers/

References

■ R for Data Science

https://r4ds.had.co.nz/

https://note.com/mitti1210/n/n386d216fcb0a