

外国語教育研究における R を用いた統計処理入門

川口 勇作 (愛知学院大学 教養部)

ykawa@dpc.agu.ac.jp

FLEAT VII ワークショップ @早稲田大学

2019/8/6

概要

本ワークショップでは、統計解析環境 R を用いた、外国語教育研究における統計処理の手法を習得することを目指します。はじめに、統計分析の際に必要な R の基本的な操作方法や動作についての導入を行います。その後、記述統計量(平均、標準偏差など)の算出、*t*検定、分散分析、相関分析、回帰分析といった、外国語教育研究においてよく用いられる分析方法の実習を行います。

R の基本知識・操作

- 命令を入力して Enter を押す
- #から始まる行はコメントアウトされる
- 使用する値は、矢印(<- / ->)やイコール(=)を用いて、変数に入れる

```
> hako <- c(1, 2, 3, 4, 5)
```
- 行列から要素を取り出すには、**変数名[行数, 列数]** というように指定する
- データフレームから要素を取り出すには、**変数名\$列名** というように指定する

データ読み込み

1. まず、作業ディレクトリを確認

```
> getwd()
```
2. 読み込みたいファイルを作業ディレクトリに移動
3. そこから

```
> read.csv("ファイル名.csv", header = T) #ファイル名は拡張子(.csv)まで含める
```
4. 見出し行の有無の指定(指定がない場合は見出し行ありとして認識)
ありの場合: header = T なしの場合: header = F

必要となるパッケージのインストール・読み込み

- パッケージのインストール

```
> install.packages("パッケージ名") #パッケージ名を" "で括る
```
- インストールしたパッケージを読み込む(=使える状態にする)方法

```
> library(読み込みたいパッケージ名) #パッケージ名前後の" "はあってもなくても OK
```

記述統計量の算出

- **記述統計**を算出するための関数一覧 ※すべて引数はベクトル
 - データの個数を求める: `length()`
 - 平均値を求める: `mean()`
 - 中央値を求める: `median()`
 - 最頻値を求める: `mode()`
 - 最大値を求める: `max()`
 - 最小値を求める: `min()`
 - 不偏分散を求める: `var()`
 - 不偏分散にもとづく標準偏差を求める: `sd()`
- **記述統計量**を一気に算出するための関数: **summary(変数名)**
※引数はデータフレーム・行列でも可
 - 平均値・第1四分位・中央値・第3四分位・最小値が算出される
- **記述統計量**を一気に算出するための関数: **describe(変数名)** [`psych` パッケージ]
※引数はデータフレーム・行列でも可
 - 標本サイズ・平均・標準偏差・中央値・刈り込み平均・中央絶対偏差・最小値・最大値・データ範囲・歪度・尖度・標準誤差が算出される

```
> library(psych)
> describe(変数名)
```
- **ヒストグラム**(度数分布図)を作成する: **hist(変数名)** ※引数はベクトル

```
> hist(変数名)
```
- **箱ひげ図**(box plot)を作成する/蜂群図(bee swarm plot)を重ね描きする

```
> 変数名 <- read.csv("ファイル名.csv", header = T)
> boxplot(変数名)
> library(beeswarm)
> beeswarm(変数名, add = T)
```

t 検定

- **t 検定**を実行する ※引数はベクトル

```
> t.test(変数名 1, 変数名 2, paired = T/F)
```

 - `paired` オプションは、対応ありの場合 T、対応なしの場合 F を選ぶ
- **効果量(Cohen's d)**を算出する

```
> library(effsize)
> cohen.d(変数名 1, 変数名 2)
```

一元配置分散分析

- **ANOVA 君**を読み込む(事前に ANOVA 君を作業ディレクトリに保存しておく)
> source("anovakun_483.txt")
> source("anovakun_483.txt", encoding = 'CP932') # Mac/Linux の方はこちら
- **一元配置分散分析**を実行する ※引数はベクトル
> anovakun(変数名, "要因計画", 要因の水準数, holm = T, eta = T)
 - holm オプションを T にすると、多重比較の方法が Holm 法になる
 - eta オプションを T にすると、効果量(イータ二乗)を算出する

二元配置分散分析

- **ANOVA 君**を読み込む(事前に ANOVA 君を作業ディレクトリに保存しておく)
> source("anovakun_483.txt")
> source("anovakun_483.txt", encoding = 'CP932') # Mac/Linux の方はこちら
- **二元配置分散分析**を実行する ※引数はベクトル
> anovakun(変数名, "要因計画", 要因 1 の水準数, 要因2の水準数, holm = T, peta = T)
 - holm オプションを T にすると、多重比較の方法が Holm 法になる
 - peta オプションを T にすると、効果量(偏イータ二乗)を算出する

相関分析

- **散布図**を作成する ※引数は行列
> plot(変数名)
- **相関係数(行列)・ p 値・信頼区間**を算出する ※cor.test 関数の引数はベクトル
> cor(変数名, method = "相関係数の種類") # 相関係数(行列)
> cor.test(変数名 1, 変数名 2, method = "相関係数の種類") # 相関係数 + p 値・信頼区間
 - method では、"pearson", "spearman", "kendall" のいずれかを指定(デフォルトは "pearson")
 - method が "spearman", "kendall" の場合は、信頼区間は算出されない

メモ

回帰分析

- 単回帰分析を実行する ※引数はベクトル
 - > 結果を入れる変数名 <- lm(従属変数 ~ 独立変数)
 - > summary(結果を入れた変数名)
- 重回帰分析を実行する
 - > 結果を入れる変数名 <- lm(従属変数 ~ 独立変数 1 + 独立変数2 + ... + 独立変数 n)
 - > summary(結果を入れた変数名)
- 多重共線性の指標 VIFを算出(2 未満なら問題ないといわれる)
 - > library(car)
 - > vif(結果を入れた変数)

メモ

連絡先

川口 勇作 (愛知学院大学 教養部)

Mail: ykawa@dpc.agu.ac.jp

サイト: <https://y-kawaguchi.com>

今回のワークショップの情報ページ: <https://y-kawaguchi.com/fleat7ws>