

Rによる基本的な統計解析の方法

遠山 2019.1.11

はじめに

ここで書いてある事項は R で統計解析をする際に、参考になるように作ったものです。

私的に使いやすい方法を紹介しています。**この資料を複写・配布することはお控えください。**

また、何かおかしい点が見つかったら、遠山 (asako.toyama@gmail.com) までご連絡ください。

この手引きでは、基本的に、**ツールとしての使用方法**しか書いていません。統計でやっていることの意味については、教科書や授業等で勉強してください。

また、R のコードは WEB で役に立つサイトが沢山ありますので、参考にして下さい。殆ど WEB 上の情報だけでも R は独学可能だと思います。

使用するデータファイル

3章・4章・5章：**sample1.csv, sample1.xlsx**

7章：**sample2.xlsx**

8章：**factor_sample.xlsx**

【クイックスタート】

既に RStudio がインストールされていて、新しく解析を始めるときは、まず、**2.4 節(P7,8)**からはじめてください。その後、自分の必要な分析のページを参照してください。

目次

| | |
|--|----|
| Rによる基本的な統計解析の方法 | i |
| 1. Rをそのまま使う場合（※授業ではR Studioを使うので、1.1のみ行ってください） | 1 |
| 1.1. Rのインストール..... | 1 |
| 1.2. Rを開く..... | 2 |
| 1.3. Rエディタの保存..... | 3 |
| 1.4. ワーキングディレクトリの設定..... | 4 |
| 2. RStudioを使う場合 | 5 |
| 2.1. RStudioについて | 5 |
| 2.2. RStudioのインストール..... | 5 |
| 2.3. RStudioの画面の見方..... | 6 |
| 2.4. 実験毎に“プロジェクト”を作ると便利です。..... | 7 |
| 3. データの読み込み..... | 9 |
| 3.1. csvファイルを使う場合..... | 9 |
| 3.2. クリップボードから直接データを読み込む場合 ★便利..... | 9 |
| 4. 記述統計量の算出とデータの図示..... | 10 |
| 4.1. 読み込んだファイルのデータの全体像を確認する..... | 10 |
| 4.2. 特定の列の情報を確認する..... | 10 |
| 4.3. ヒストグラムを描く..... | 11 |
| ■ RTのヒストグラム（度数分布を示すグラフ）を描く..... | 11 |
| ■ もっと綺麗なヒストグラムを描きたい時..... | 11 |
| ■ 画像の保存の仕方（pngとして保存する場合）..... | 11 |
| 4.4. 列同士の関係性を確認する..... | 12 |
| ■ 2つの質的変数の連関をみる..... | 12 |
| ■ 質的変数の各水準ごとに量的変数を集計する..... | 12 |
| ■ 条件毎の値の特徴（四分位数）を図示する。..... | 12 |
| ■ 量的変数間の関係を図示する..... | 13 |
| ■ 量的変数間の相関係数を確認する..... | 14 |
| 4.5. 読み込んだファイルを分割する..... | 14 |
| 5. 相関..... | 16 |
| ■ 使用場面..... | 16 |
| ■ 作図..... | 16 |
| ■ 統計..... | 16 |

| | |
|--|----|
| ■ Advance..... | 17 |
| 6. t 検定..... | 19 |
| 6.1. 対応のある標本間での比較（被験者内検定）..... | 19 |
| 6.2. 対応のない標本間での比較（被験者間検定）..... | 19 |
| 6.3. 片側検定に関して..... | 20 |
| 7. カイ二乗検定など..... | 21 |
| ■ 使用場面..... | 21 |
| ■ 統計..... | 21 |
| 8. 分散分析..... | 23 |
| 8.1. ANOVA 君を使う環境設定..... | 23 |
| 8.2. ANOVA 君用にエクセルを整理し、分析をかける。..... | 23 |
| ■ 被験者内 1 要因 3 水準 の場合..... | 24 |
| ■ 被験者間要因（2 水準）と被験者内要因（3 水準）の混合計画の場合..... | 25 |
| ■ 被験者内 2 要因（2 水準と 3 水準）の場合..... | 26 |
| 8.3. 分散分析に関して補足..... | 27 |
| ■ 球面性検定..... | 27 |
| ■ 自由度の調整..... | 27 |
| ■ 多重比較..... | 27 |
| ■ 効果量を出力する..... | 27 |
| 8.4. 結果の見方..... | 28 |
| ■ 被験者内 1 要因 3 水準 の場合..... | 28 |
| ■ 被験者内 2 要因（2 水準と 3 水準）の場合..... | 29 |
| 9. おまけ..... | 30 |
| 9.1. 地味によく使う機能..... | 30 |
| 9.2. 直接コピー & ペーストでデータを読み込む方法..... | 30 |
| 9.3. 欠損値について..... | 30 |
| 9.4. もっと詳しく学びたい人は..... | 31 |
| ■ 参考 URL..... | 31 |
| ■ 参考図書..... | 31 |

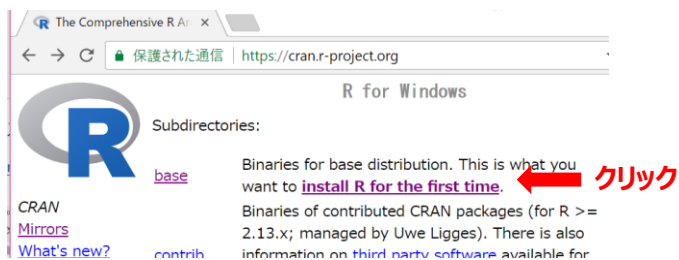
1. R をそのまま使う場合（※R Studio をメインで使う場合は、1.1 のみ行ってください）

1.1. R のインストール

- ① <https://cran.r-project.org/> へ。
- ② 自分の PC に合わせてクリック。（以降、Windows を選んだ場合の例を示します。）



③

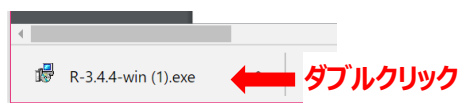


- ④ 下記をクリックすると、ダウンロードが始まります。

ここで、R.3.4.4 は、R のバージョンを表しています。サイトにアクセスした時の最新バージョンが表示されているので、それをダウンロードしてください。



- ⑤ ダウンロードが完了したら、exe ファイルをダブルクリックして、インストールを開始します。どんどん「次へ」を押して、インストールを完了させましょう。



1.2. R を開く

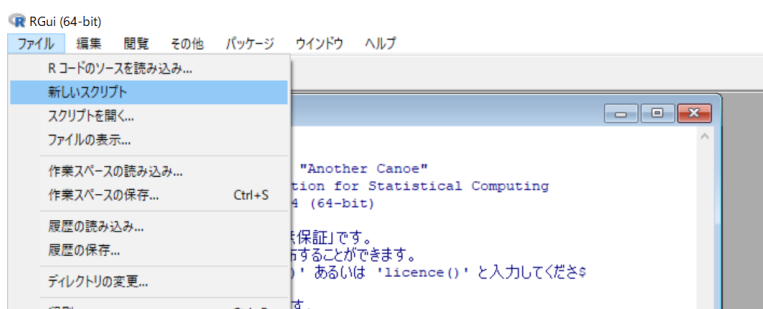
コードを書く画面は **R console** と **R エディタ** の2種類があります。

R を開くと、**R consol** という画面がでできます。

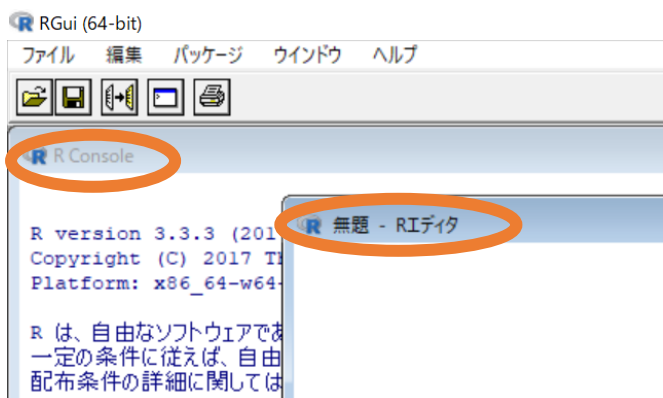
R consol 上にコードを書いて Enter キーを押すと、そのコードが直ちに実行されます。
ただし、R consol 上に書いたコードは保存できません。

自分の書いたコードを保存するためには、**R エディタ** を使います。

下記のように、**ファイル > 新しいスクリプト** をクリックし、R エディタの画面を開きます。



下記のような感じで、**R Console** と **R エディタ** を開いた状態にしましょう。



【コードの実行方法】

R エディタにコードを書いた後、1行ずつ、実行する方法

→ 実行したいコードの行にカーソルを合わせ、「Ctrl + R」を押すと、R Console 上で実行されます。

R エディタにコードを書いた後、全てのコード、または一部を実行する方法

→ 実行したいコードを選択した状態で、「Ctrl + R」を押すと、R Console 上で実行されます。

なお、コードの冒頭に **#** をつけると、コメントアウトできるので、実行時に読み込まれません。

1.3. R エディタ の 保存

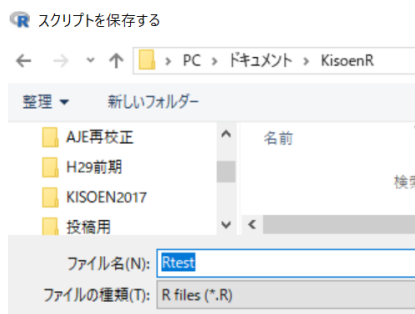
① 保存したい R エディタを開いた状態で、**ファイル> 保存** を選択



②「新しいフォルダー」をクリック。新しいフォルダーの名前をつける。（下記の例だと『KisoenR』）



② 上で作ったフォルダをダブルクリック。ファイル名を付けて（下記の例だと『Rtest』）, 「保存」をクリック。



これで、保存が完了です。

※ 次からは、このファイルを直接ダブルクリックするか、R を開いた状態で、**ファイル> スクリプトを開く** で開くことができます。

ファイル名や、ファイルの中で、日本語を使うことは避けましょう！なるべく**英数字を使いましょう！！**

1.4. ワーキングディレクトリの設定

現在のワーキングディレクトリ（作業をしている場所）を確認する：`getwd()`

ワーキングディレクトリを設定する：`setwd("C:/Users/asako/OneDrive/ドキュメント/KisoenR")`

コードの意味：ワーキングディレクトリを "" 内のフォルダに移動。

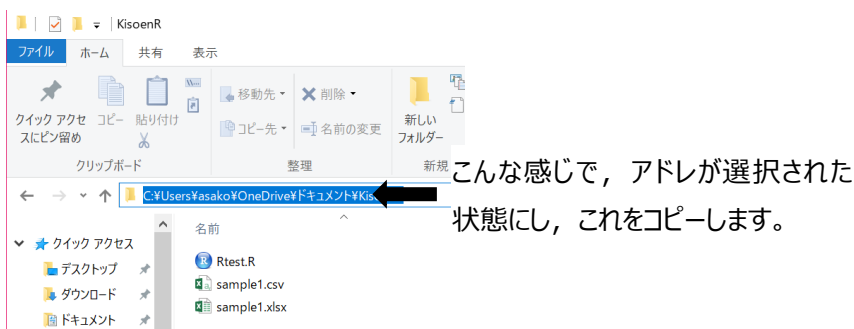
""の中は各自で決めたフォルダのアドレスを入れます。

【どのフォルダをワーキングディレクトリにするか？】

基本的に、データを保存しているフォルダを使うとよいでしょう。R エディタも同じフォルダに保存しましょう。

【フォルダのアドレスは？】

特定のフォルダのアドレスを確認するためには、まずエクスプローラでそのフォルダを開きます。上部にアドレスが表示されているので、そこを一度クリックすると、青色になります。これが、アドレスが選択されている状態です。この状態で「Ctrl + C」でアドレスをコピーできます。



これを、`setwd("")`の、"" 内に貼り付け（「Ctrl + V」）れば良いのですが、その後、必ず、**スラッシュを / に変更してください！** そうしないと、エラーがでます。

* csv ファイルや、R エディタのファイルは、全て同じフォルダ内に保存しておくといいですね。

* フォルダ名は半角の英数字を使いましょう。

設定が終わったら、再度 `getwd()` と入力して、ちゃんと設定されているか、確認してください。

2. RStudio を使う場合

2.1. RStudio について



私は **RStudio** というソフトウェアを使って、R を使用しています。

関数と文字で色分けしてくれたり、使いやすい画面で非常に便利です。

一旦インストールが完了したら、R を開く必要はありません。RStudio を直接開いて、作業します。

(実行などのショートカットキーは R と異なります。)

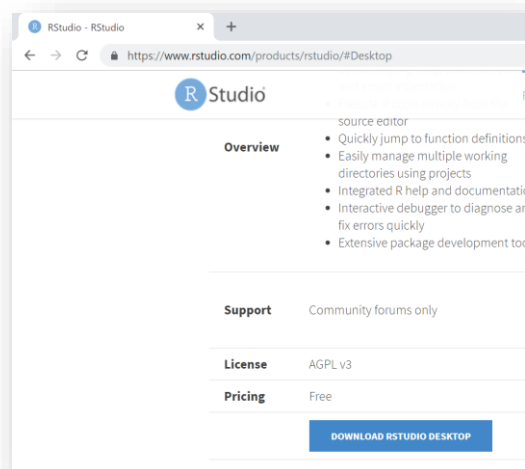
2.2. RStudio のインストール

大学のパソコン室のパソコンには予めインストールされているかもしれないので、まずは、それを確認してみてください。

インストールされていない場合は、RStudio のページから、**Desktop 版**をダウンロードしましょう。(Free の方です。)
google 検索などで、RStudio と検索すると、一番上に出てくると思います。

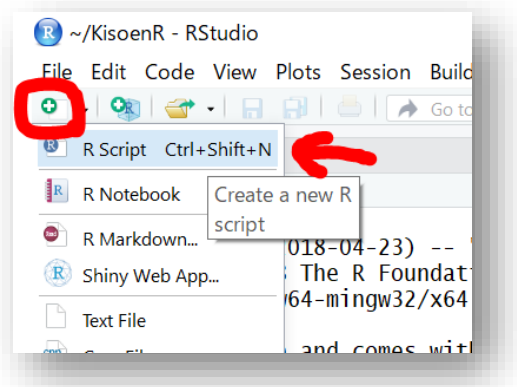
(google 検索では、検索が“日本語のページを検索”になっている場合があるので、“すべての言語”に設定し直すと、一番上に出てくるはずです。)

そこから、ダウンロードのページへ進み、ダウンロードとインストールを完了してください。

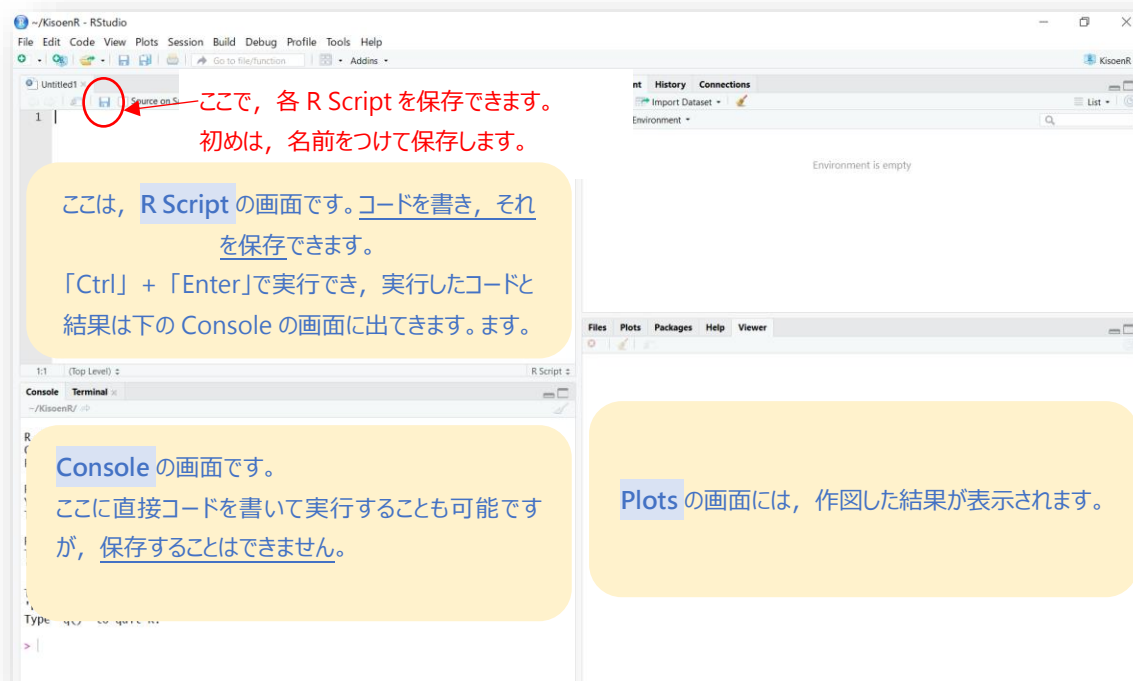


2.3. RStudio の画面の見方

画面の見方を説明する前に、まずは R Script というものを新規作成してください。
左上端の十字のマークをクリックし、[R Script] というものをクリックすると作成できます。



各画面について、皆さんが特に使う部分だけざっくり紹介します。この内容はしっかりおさえましょう。



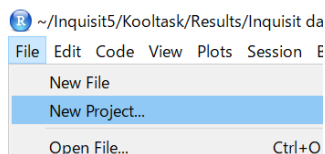
2.4. 実験毎に“プロジェクト”を作ると便利です。

R プロジェクトとは、一連の解析を1つにまとめておくもので、ワーキングディレクトリや履歴も一緒に保存してくれるので、大変便利です。Rstudioに慣れてない方はその良さは分かりにくいかもしれませんが、とりあえずこれは必ず作成するようにしましょう。(R プロジェクトを作成しておくと、1.4 節で説明したような作業が不要になります。)

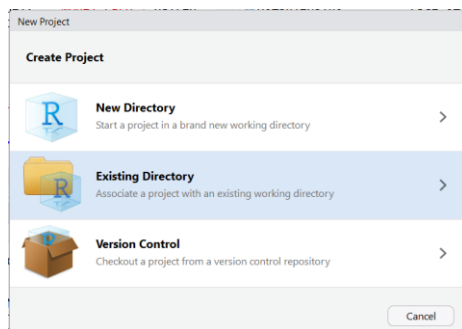
以下に、演習での実験データの整理方法について、私のお薦めするやり方を書いていきます。
(慣れてきたら自分流のルールを作ってみてください。)

- ① **実験の結果を入れるフォルダを作成**します。実験の結果はすべてここに保存します。
- ② 上で作ったフォルダに、RStudioで**プロジェクトを作成**します。手順は下記の通りです。

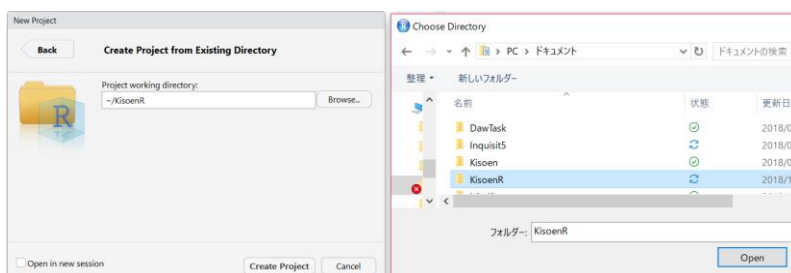
1. RStudioを開いたら、File > New project... をクリック



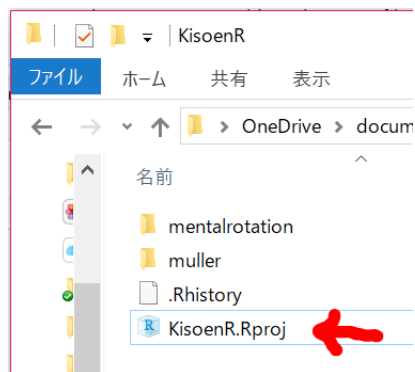
2. Existing Directory をクリック。



3. [Browser...] から、①で作成したフォルダを選択し、ダブルクリック。その後、ディレクトリが指定したものに変わっていることを確認して、[Create Project] をクリック。



4. フォルダの中に、拡張子が“.Rproj”というファイル（プロジェクトファイル）ができていることを確認しましょう。



これでプロジェクトが1つ作れました。

※ 次回以降、Rstudioを開く際は、このプロジェクトファイルを直接ダブルクリックして開きましょう。

【参考】フォルダの中はこんな感じになります。

ここがフォルダのアドレスです。クリックすると青くなってコピーすることができます。
Rコード `getwd()`: 今このフォルダで作業しているか確認
Rコード `setwd("アドレス")`: 特定のフォルダに設定しなおす

anova君のソースコード
→ これがあると、分散分析 (ANOVA) をするのに便利なRの関数が使えます。作業フォルダの中に入っている必要があります。

Rのスクリプト (拡張子は.R)
→ 書いたコードが保存されています。これを直接開くよりも、プロジェクトファイルを開いて、その中から開くようにしましょう。

Rプロジェクトファイル (拡張子は .Rproj)
→ 毎回、プロジェクトファイルをクリックしてスタートするようにしましょう。自動的にRプロジェクトが入っているフォルダが作業フォルダとして認識されます。このプロジェクトの中でRのスクリプトを動かしたりします。

データ (拡張子は.xlsx や .csv)
→ データはExcelで作成すると便利です。ただし、Rで読み込む形にするには csv ファイルを別途保存します。

.Rhistory や .Rata は自身で直接操作することはありませんが、Rstudioを動かす上での情報が保存されています。むやみに消さないように。

その他、Rstudioの中で、図 (拡張子 .png など) を保存すると、デフォルトで、この作業フォルダの中に保存されます。

3. データの読み込み

3.1. csv ファイルを使う場合

Rに読み込むデータは、基本的に、エクセルで作成し、**csvファイル**として保存しましょう。

【補足：エクセルファイルの保存形式について】

Rで読み込むデータは、基本的に、エクセルで作成し csv の拡張子で保存したものを使います。

エクセルを保存する時は、拡張子が .xlsx がデフォルトです。この拡張子だと、色や数式の情報や、グラフなどがそのまま保存されます。また複数のワークシートの情報が保存できます。

一方、拡張子を .csv で保存したエクセルファイルは、色や式、グラフの情報は保存されず、ワークシートも1つしか保存されないので注意してください。

csv ファイルを読み込む：`d <- read.csv("sample1.csv", header = T)`

コードの意味：sample1.csv というファイル名のファイルを読み込み、d と名前を付ける。

1 行目は、ヘッダーである (True の T) 。

"" の中に、読み込む csv ファイルの名前。

ヘッダーがなく、1 行目からデータが入っている場合は header=F とします。

ちなみに、header= は書かなくても、大体自動的に判断してくれます。

注意点：""を忘れずに！ 拡張子 (.csv) も忘れずに！

3.2. クリップボードから直接データを読み込む場合 ★便利

この方法は結構便利です。是非使ってみてください。

- ① エクセルファイルや CSV ファイルで、読み込みたいセルを選択。
- ② 「Ctrl + C」でクリップボードにコピー。
- ③ Rに `d <- read.table("clipboard", header=T)` と打ち込んで実行。

これだけで、コピーした内容が d という変数名で保存されます。(d の部分は好きな変数名をどうぞ)

4. 記述統計量の算出 と データの図示

4.1. 読み込んだファイルのデータの全体像を確認する

読み込んだファイル全体を表示するコード : `d` ←ちゃんと読み込めているか、まず確認。
 データの上 6 行のみ提示する : `head(d)` ←データ量が多い場合は一部だけ表示して確認。
 データの要約を確認 : `summary(d)`
 行の数を確認 : `nrow(d)`
 列の数を確認 : `ncol(d)`
 データ値の形式を確認 (数値なのか、文字なのか) : `str(d)`

4.2. 特定の列の情報を確認する

ここでは、特に RT についての記述統計量を出す例を示します。

RT の平均 : `mean(d$RT)`

ここで、**\$マーク**は「~の」くらいの意味です。つまり、`d$RT` は、「`d` というデータの中の RT のデータ」の列です。

RT の中央値 : `median(d$RT)`

得られたデータを母集団から取ってきた標本と考え、母集団のばらつきを考えると、

RT の不偏分散 : `var(d$RT)`

RT の不偏標準偏差 : `sd(d$RT)`

得られたデータを母集団として、そのばらつきを考えると、

RT の分散 : `var(d$RT)*(length(d$RT)-1)/length(d$RT)`

RT の標準偏差 : `sqrt(var(d$RT)*(length(d$RT)-1)/length(d$RT))` #標本分散の平方根

下記のように、標本分散を求める関数を作ると、便利かもしれません。

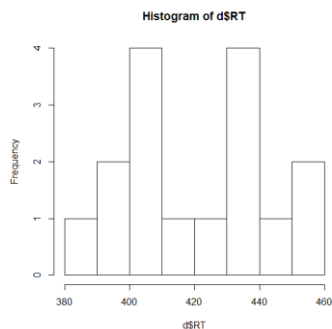
```
variance <- function(x) { var(x)*(length(x)-1)/length(x) }
```

一度上のコードを実行すれば、RT の標本分散は `variance(d$RT)`、標本標準偏差は `sqrt(variance(d$RT))` となります。

4.3. ヒストグラムを描く

■ RTのヒストグラム（度数分布を示すグラフ）を描く

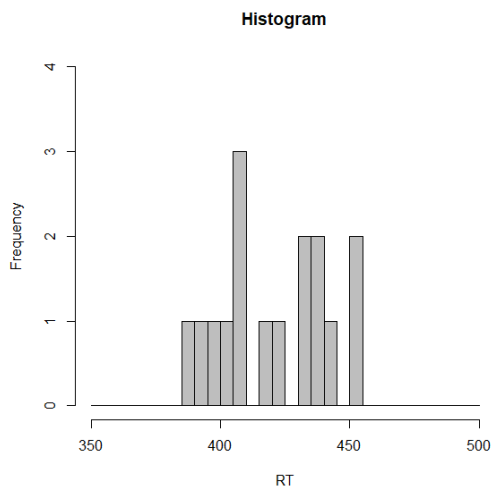
```
hist(d$RT)
```



■ もっと綺麗なヒストグラムを描きたい時

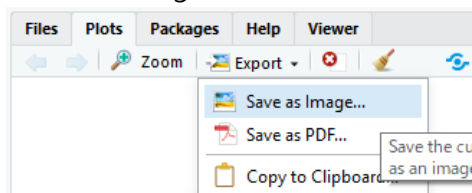
```
hist(d$RT, breaks=seq(350,500,5), main="Histogram", xlab="RT", ylim=c(0,4), col="gray")
```

上の図と、下記の図で、何が変わったか確認して、コードの意味を理解しましょう。



■ 画像の保存の仕方（pngとして保存する場合）

図が出てきた個所の上に「Export」から「Save a Image...」を選んで保存しましょう。



保存先は、次の画面の、「Directory...」から選べます。

4.4. 列同士の関係性を確認する

■ 2つの質的変数の連関をみる

例) 性別によって、不安の高い人・低い人がどれくらいいるか知りたい。

※sample1.csv の Gender も AnxietyHL も質的変数です。

クロス集計表の作成 : `table(d$Gender, d$AnxietyHL)`

このコードをうつと、下記のようにクロス集計表が作成されます。各セルには度数が入っています。

例えば、女性で不安が高いに分類される人は5人。

```
> table(d$Gender, d$AnxietyHL)
```

```
      High Low
f      5   3
m      4   4
```

■ 質的変数の各水準ごとに量的変数を集計する

例) 条件毎に、RT の平均値を出したい。

※sample1.csv の Condition(条件) は、c1 と c2 があります。

`tapply()`関数 が便利です。`tapply(量的変数, 質的変数, 関数)` を入力します。

性別毎の RT の平均値の計算 : `tapply(dRT, dCondition, mean)`

下記のような感じで、条件毎の平均値が確認できます。mean のところは適宜、適当な関数を入れて下さい。

```
> tapply(d$RT, d$Condition, mean)
      c1      c2
417.5 425.0
```

■ 条件毎の値の特徴（四分位数）を図示する。

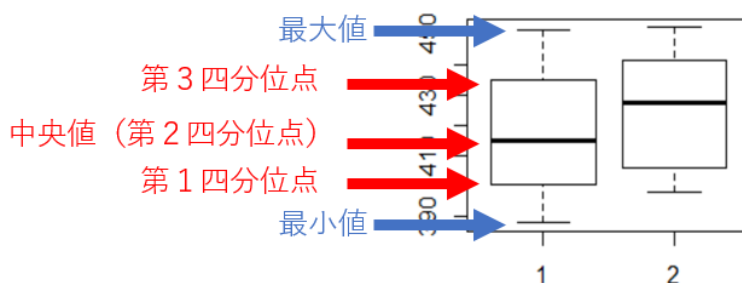
箱ひげ図を表示する：

```
boxplot( d$RT[d$Condition == "c1"], d$RT[d$Condition == "c2"] )
```

※ コード中の [] の中で、どの条件の RT のデータを使うかを限定しています。

参考までに、主な比較演算子の記号は下記の通りです。

等号 ==
否定 !=
以上 >=
以下 <=
より大きい >
より小さい <



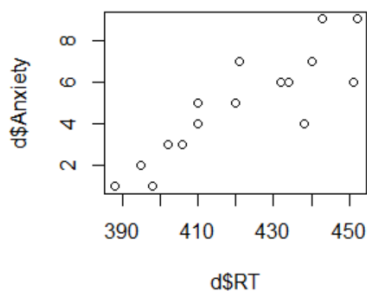
上記のような図が箱ひげ図です。最大値と最小値の外に○があるときは、外れ値です。

外れ値はデフォルトでは、(第1四分位数-1.5*(第3四分位数-第1四分位数))より小さい、または、(第3四分位数+1.5*(第3四分位数-第1四分位数))より大きい値、と定義されています。

■ 量的変数間の関係を図示する

散布図を描く：`plot(dRT, dAnxiety)`

出てきた図を見ると、不安の高い人ほど、反応時間が遅いという関係がありそうです。



■ 量的変数間の相関係数を確認する

相関係数を出す : `cor(dRT, dAnxiety)`

下記のように、反応時間と不安の間には、 $r = 0.86$ という高い相関がありました。

```
> cor(d$RT, d$Anxiety)
[1] 0.8625582
```

4.5. 読み込んだファイルを分割する

女性のデータのみ限定する : `d_f <- subset(d, Gender == "f")`

コードの意味 : d のデータのうち、Gender が f である行を取り出し、d_f と名前をつける。

== 以外の比較演算子の書き方については、P13 を参照してください。

新しいデータの名前は、上の例だと d_f としましたが、好きなように付けてください。

※ データの中の文字は、" " で囲む必要があります！（上記の例だと "f"）

試しに、d_f と打ち込んでみると、下記のように female のみのデータがでできます。確認してください。

```
> d_f <- subset(d, Gender == "f")
> d_f
  Sub Gender Condition  RT Anxiety
1  1     f         c1  420      5
2  2     f         c1  451      6
3  3     f         c1  432      6
4  4     f         c1  438      4
5  5     f         c2  398      1
6  6     f         c2  402      3
7  7     f         c2  421      7
8  8     f         c2  410      5
> |
```

この subset() をうまく使うことで、男女別に平均値を出したり、が可能になりますね。
ただ、データを分割せずとも、table() や tapply() などを活用しても良いですね。

この章で使ったサンプルデータのエクセルバージョンは sample1.xlsx でした。このデータの AnxietyHL の列では、Anxiety が 5 以上の人を “High”, それ以外を “Low” としました。

| | A | B | C | D | E | F | |
|----|-----|--------|-----------|-----------|---------|-----|---|
| 1 | Sub | Gender | Condition | AnxietyHL | Anxiety | RT | M |
| 2 | 1 | f | c1 | High | 5 | 420 | |
| 3 | 2 | f | c1 | High | 6 | 451 | |
| 4 | 3 | f | c1 | High | 6 | 432 | |
| 5 | 4 | f | c1 | Low | 4 | 438 | |
| 6 | 5 | f | c2 | Low | 1 | 398 | |
| 7 | 6 | f | c2 | Low | 3 | 402 | |
| 8 | 7 | f | c2 | High | 7 | 421 | |
| 9 | 8 | f | c2 | High | 5 | 410 | |
| 10 | 9 | m | c1 | Low | 1 | 388 | |
| 11 | 10 | m | c1 | Low | 2 | 395 | |
| 12 | 11 | m | c1 | Low | 4 | 410 | |
| 13 | 12 | m | c1 | Low | 3 | 406 | |
| 14 | 13 | m | c2 | High | 7 | 440 | |
| 15 | 14 | m | c2 | High | 9 | 443 | |
| 16 | 15 | m | c2 | High | 6 | 434 | |
| 17 | 16 | m | c2 | High | 9 | 452 | |

R で読み込んだ sample1.csv ファイルの方には色はついていませんが, sample1.xlsx と値は同じです。

5. 相関

■ 使用場面

2つの変数間の関係を知りたいとき。

例) 英語の成績が良い人は、数学の成績も良いのか？

■ 作図

相関はあるかどうかの検定結果は、外れ値の影響を受けやすいです。

統計の結果だけに頼らず、まず必ず散布図は確認しましょう。

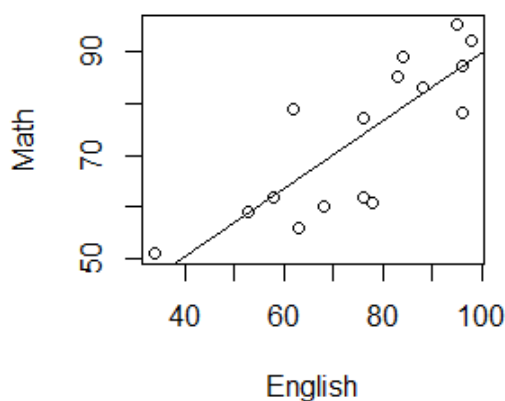
散布図を確認する：`plot(d$MathTest~ d$EnglishTest, ylab="Math", xlab="English")`

※ `ylab` や `xlab` はそれぞれ、Y 軸のラベルと X 軸のラベルですが、特段必要なければ、指定する必要はありません。

回帰直線を追加する：`abline(lm(d$ MathTest ~ d$ EnglishTest))`

※ `plot()` の式は、"`~`" の代わりに "`,`" でもいいのですが（: x 軸と y 軸にくるものが逆になります）、回帰直線のコードに合わせて "`~`" を使う方が、間違いを防げます。

※ `abline(a,b)` で、 $y = a + bx$ の回帰直線を描く関数です。この括弧の中に、線形回帰を実行する関数である `lm()` を入れると、そこで求められた回帰式の切片と傾きを使って直線を引いてくれます。



■ 統計

代表的な相関の出し方は、下記の2種類あります。一般に、2つの変数のうちどちらか一方でも正規分布とみなせなければスピアマンの順位相関係数を使う、ということになります。（実際はピアソンでやっているものが多い。）

ピアソンの積率相関：`cor.test(d$MathTest, d$EnglishTest)`

スピアマンの順位相関 : `cor.test(d$MathTest, d$EnglishTest, method = "spearman")`

結果の見方

```
> cor.test(d$MathTest, d$EnglishTest)

Pearson's product-moment correlation

data: d$MathTest and d$EnglishTest
t = 5.2702, df = 14, p-value = 0.0001184 p 値
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5365737 0.9337030
sample estimates:
 cor
0.8153976 相関係数 (r)
```

報告の書き方の例

AとBに関して、ピアソンの積率相関を求めたところ、有意な正の相関が認められた ($r = .36, p < .05$)。

AとBに関して有意な負の相関が認められた ($r = -.36, p = .02$)。

AとBに有意な相関はなかった ($r = .23, n.s.$)。

AとBに有意な相関はなかった ($r = .23, p = .15$)。

■ Advance

`cor()` を使うと、複数の変数間の相関を一気にみることも可能です。

(`cor()` に入れるデータは数値データのみにする必要があります。)

ただし有意差の検定には、やはり `cor.test()` で1対比較してください。

また、"PerformanceAnalytics" というパッケージの `chart.Correlation()` 関数で下記のような図が簡単に書けます。見るべき相関が沢山あるときは便利です。

1. 必要なパッケージのインストール : `install.packages("PerformanceAnalytics")`
※この作業が必要なのは、1番最初だけです。
2. 関数を使うためにパッケージを読み込む : `library("PerformanceAnalytics")`
3. 関数の実行 : `chart.Correlation(d)`

`chart.Correlation()` 関数を使うときも、データは数値データのみにおきましょう。

(この関数自体は、設定を変えれば、質的データも扱うことは可能です。詳しく知りたい人はネットで調べてください。)

使用例)

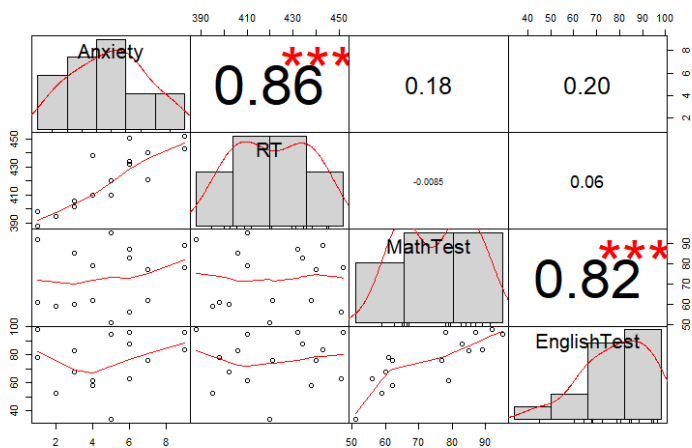
- ① Sample1.exle を開き、下記のように、E～H 列までを選択して Ctrl + C でコピーします。

| | A | B | C | D | E | F | G | H |
|----|-----|--------|-----------|-----------|---------|-----|----------|------------|
| 1 | Sub | Gender | Condition | AnxietyHL | Anxiety | RT | MathTest | EnglishTes |
| 2 | | 1 f | c1 | High | 5 | 420 | 51 | 34 |
| 3 | | 2 f | c1 | High | 6 | 451 | 56 | 63 |
| 4 | | 3 f | c1 | High | 6 | 432 | 87 | 96 |
| 5 | | 4 f | c1 | Low | 4 | 438 | 62 | 58 |
| 6 | | 5 f | c2 | Low | 1 | 398 | 61 | 78 |
| 7 | | 6 f | c2 | Low | 3 | 402 | 60 | 68 |
| 8 | | 7 f | c2 | High | 7 | 421 | 62 | 76 |
| 9 | | 8 f | c2 | High | 5 | 410 | 95 | 95 |
| 10 | | 9 m | c1 | Low | 1 | 388 | 92 | 98 |
| 11 | | 10 m | c1 | Low | 2 | 395 | 59 | 53 |
| 12 | | 11 m | c1 | Low | 4 | 410 | 79 | 62 |
| 13 | | 12 m | c1 | Low | 3 | 406 | 85 | 83 |
| 14 | | 13 m | c2 | High | 7 | 440 | 77 | 76 |
| 15 | | 14 m | c2 | High | 9 | 443 | 89 | 84 |
| 16 | | 15 m | c2 | High | 6 | 434 | 83 | 88 |
| 17 | | 16 m | c2 | High | 9 | 452 | 78 | 96 |

- ② 下記のコードを実行します。

```
library("PerformanceAnalytics") #ライブラリの読み込み
d<-read.table("clipboard", header=T) #先ほどコピーした部分を読み込み，dとする。
chart.Correlation(d) #作図する。
```

- ③ 結果を確認します。散布図も検定結果も見ることができ、便利ですね。



6. t 検定

使用例

t 検定は、2つの平均を比べたい時に使います。まずは下記の2つの場合のどちらになるかで検定方法が違います。6.1と6.2では、両側のt検定について扱います。片側のt検定については、6.3で触れます。

6.1. 対応のある標本間での比較（被験者内検定）

```
t(d$A1, d$A2, paired = T)
```

paired = Tで、対応のあるt検定であることを表します。

※ 等分散性については気にしなくてよい（験者内検定の場合は、各ペアの差の分散を使うため）

6.2. 対応のない標本間での比較（被験者間検定）

まずは、下記のコードで等分散性を検定（通常のt検定は、2群の分散が等しい場合に使用できる。）

```
var.test(d$A1, d$A2) #この検定は「2群の分散は等しい」という帰無仮説です。
```

等分散性を検定の結果、p値が.05より大きい場合

→ 2群の分散が等しいという帰無仮説は棄却できない。

→ 通常のt検定を行う

```
t.test($A1, d$A2, paired = F, var.equal = T)
```

paired = Fで、対応のないt検定であることを指定します。

var.equal = Tで、等分散性を仮定することを指定します。

等分散性を検定の結果、p値が.05より小さい場合

→ 2群の分散が等しいという帰無仮説は棄却される。（つまり、等分散性を仮定できない）

→ ウェルチのt検定を行う

```
t.test($A1, d$A2, paired = F, var.equal = F)
```

paired = Fで、対応のないt検定であることを指定します。

var.equal = Fで、等分散性を仮定できないことを指定します。

6.3. 片側検定に関して

これまでの検定は、両側の t 検定に関するものでした。

片側検定をしたい場合は、p 値を 2 で割ればいだけです。t 値の結果や、df はそのままです。

- 一応、下記のように書けば、数学の方が英語より点数が高い、という帰無仮説の片側の t 検定

```
t.test(data$math, data$english, paired=T, alternative="greater")
```

- また、下記のように書くと、数学の方が英語より点数が低い、という帰無仮説の片側 t 検定

```
t.test(data$math, data$english, paired=T, alternative="less")
```

報告の書き方の例

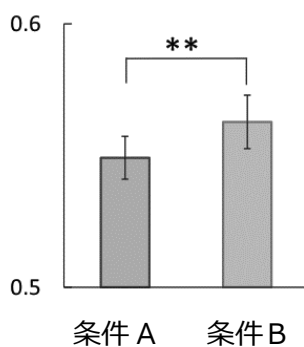
t 検定により A と B の反応時間について検討した結果、**Aの方が有意に**反応時間が長かった ($t(104) = 2.21, p < .05$)。

※ t 検定の t は小文字です。(大文字だと、別の意味になってしまうので注意しましょう。)

図を描くのは、エクセルを使った方が簡単です。

※ 誤差バーが何を表すのかも、図の説明で書いておきましょう。

※ 図中にアスタリスク (*) などで、有意かどうかを視覚的に分かるようにしておくが見やすいですが、これは必須ではありません。



** : $p < .001$

7. カイ二乗検定など

カイ二乗検定は、2つの質的変数の連関に関する検定です。この検定は近似ですが、一般に、クロス集計表のセルの期待値に 10 未満のものがある場合は、**フィッシャーの直接確率検定**（直接 p 値を計算する方法）を用いる方が良いとされています。（どちらがよいかは議論のあるところのようです。）

■ 使用場面

「群（Group）」（実験群（E）/コントロール群（C））で、「反応（Response）」（Yes /No）に連関があるか？」を検討したい。

帰無仮説 H0: 2 要因の比率に差がない。

対立仮説 H1: 2 要因の比率に差がある。

■ 統計

① データを読み込みこむ（今回は、sample2.xlsx から直接コピペで取り込みます。）

```
d2<-read.table("clipboard",header=T)
```

| | A | B | C |
|----|-----|-------|----------|
| 1 | Sub | Group | Response |
| 2 | | 1 E | Yes |
| 3 | | 2 E | Yes |
| 4 | | 3 E | Yes |
| 5 | | 4 E | Yes |
| 6 | | 5 E | Yes |
| 7 | | 6 E | Yes |
| 8 | | 7 E | Yes |
| 9 | | 8 E | Yes |
| 10 | | 9 E | Yes |
| 11 | | 10 E | Yes |
| 12 | | 11 E | No |
| 13 | | 12 E | No |
| 14 | | 13 C | Yes |
| 15 | | 14 C | Yes |
| 16 | | 15 C | Yes |

② クロス集計表を作成する

```
> table(d2$Group,d2$Response)
```

```

      No Yes
C      7   6
E      2  10

```

③ 検定をする（実際は下記のうち、適切な 1 つを行います。）


```
chisq.test(d2$Group, d2$Response)          #ピアソンのカイ2乗検定
chisq.test(d2$Group, d2$Response, correct=F) #イエーツの連続修正を行ったカイ2乗検定
fisher.test(d2$Group, d2$Response)         #フィッシャーの直接確率検定
```

上記の関数の中には、2つの要因（d2\$Group, d2\$Response）を入れていますが、ここに直接集計表（table(d2\$Group, d2\$Response)）を入れても同じ結果になります。

【補足：fisher.test()について】

フィッシャーの直接確率検定で片側検定を行う場合は、"alternative = " を使います。

```
fisher.test(d2$Group, d2$Response, alternative = "less")
fisher.test(d2$Group, d2$Response, alternative = "greater")
```

alternative は、2*2 のクロス集計表を検定するときのみ使えて、"two.sided", "greater", "less" のいずれかを指定できます。


なお、fisher.test() では、オッズ比（odds ratio）も計算されていますが、ここで計算されるオッズ比は、条件付きの最尤推定量です。また、fisher.test() では、p 値とオッズ比の信頼区間の計算方法が違っているので、同じ結果にならないことがあります。exact2x2 パッケージの fisher.exact() を使うと良いかもしれません。

※ ころへんの話は、奥村晴彦 先生のサイトがわかりやすいので参考にしてみてください。

<https://oku.edu.mie-u.ac.jp/~okumura/stat/fishertest.html>

8. 分散分析

8.1. ANOVA 君を使う環境設定

ここでは、井関龍太先生の ANOVA 君という R 用の分散分析関数を利用する手法を紹介します。まずは、anovakun_482.txt ファイルをコピーして、カレントディレクトリのフォルダに入れます。（RStudio でプロジェクトファイル  を入れているフォルダです。）これで準備完了です。

最新の txt ファイルは、井関龍太先生のホームページ上にありますので必要に応じてダウンロードして下さい。Google で「ANOVA 君」と検索すると、検索結果のトップが下記の URL になっていると思います。そのページ内を探して下さい。

（この txt ファイルの番号は、更新されているので、482 ではないかもしれません。）

<http://riseki.php.xdomain.jp/index.php?ANOVA%E5%90%9B>

8.2. ANOVA 君用にエクセルを整理し、分析をかける。

ANOVA 君を使用して分析をする際は、データを特定の順番で並べる必要があります。ここではその作成例と、コードを実行するまでの概要です。（この節は **factor_sample.xlsx** を参照してください。）

以下では、よく使いそうな下記の3つの分析を紹介します。

- ・ 被験者内 1 要因 3 水準 の場合
- ・ 被験者内 2 要因（2 水準と 3 水準）の場合
- ・ 被験者間要因（2 水準）と被験者内要因（3 水準）の混合計画の場合

この3種類がわかれば、要因数や水準数が違って多くの分析は対応できると思います。

（そのほかの分や詳細については、井関先生のホームページを参照して下さい。）

■ 被験者内 1 要因 3 水準 の場合

| | A | B | C |
|---|------|------|------|
| 1 | con1 | con2 | con3 |
| 2 | 5 | 6 | 1 |
| 3 | 4 | 4 | 2 |
| 4 | 6 | 4 | 3 |
| 5 | 3 | 2 | 1 |
| 6 | 5 | 1 | 3 |
| 7 | 6 | 5 | 2 |
| 8 | 2 | 5 | 1 |
| 9 | 5 | 7 | 2 |

※ 1 行目に 3 つの水準の名前を入れます。(これは自分の整理のために、分析では使いません。)

※ 2 行目以降の各列は一人ひとりの被験者のデータ情報です。(e.g., 各条件の平均値)

分析

① 下記のコードで ANOVA 君の関数を使えるように読み込みます。

```
source("anovakun_482.txt")
```

② エクセルの該当箇所を Ctrl + C でコピーします。

③ 下記のコードで、R に読み込みます。

```
d <- read.table("clipboard", header=T)
```

※ コピーしていたデータを読み込み、d という名前をつけます。

※ 1 行目も読み込んだ場合は、header = T とすることで、「1 行目はヘッダー」と宣言。

※ 1 行目を読み込まないのであれば、header = F とします。

④ ちゃんとデータが読み込めているか、d とコンソールに打って確認します。

```
d
```

⑤ 分散分析を実行します。

```
anovakun(d, "sA", condition=c("15","30","60"), gg=T)
```

※ anovakun() という関数で、分散分析を実施します。

※ d は、先程作ったデータです。

※ "sA" とあるのは実験計画を指定しています。s のあとのアルファベットは、被験者内要因があることを示します。要因の数だけ A,B,C... とアルファベットを増やします。今回は 1 要因 なので、A だけです。

※ condition = c("15","30","60") というのは、A の要因に condition という名前をつけ、その中の水準名を 15, 30, 60 とする、と指定しています。

※ gg=T とすると、自由度の補正で、Greenhouse-Geisser の方法を使って ϵ を算出し、自由度の調整をします。

■ 被験者間要因（2水準）と被験者内要因（3水準）の混合計画の場合

| | A | B | C | D |
|---|-----|-------|-------|-------|
| 1 | sub | con15 | con30 | con60 |
| 2 | a1 | 5 | 6 | 1 |
| 3 | a1 | 4 | 4 | 2 |
| 4 | a1 | 6 | 4 | 3 |
| 5 | a1 | 3 | 2 | 1 |
| 6 | a2 | 5 | 1 | 3 |
| 7 | a2 | 6 | 5 | 2 |
| 8 | a2 | 2 | 5 | 1 |
| 9 | a2 | 5 | 7 | 2 |

※ A 列に被験者間要因の水準を指定します。a1,a2,a3…としましょう。

※ 1 行目は、A は被験者間要因名、B, C, D には、被験者内要因の水準名を入れます。（自分の整理のために、分析では使いません。）

※ 2 行目以降の各列は一人ひとりの被験者のデータ情報です。（e.g., 各条件の平均値）

分析

① 下記のコードで ANOVA 君の関数を使えるように読み込みます。

```
source("anovakun_482.txt")
```

② エクセルの該当箇所を Ctrl + C でコピーします。

③ 下記のコードで、R に読み込みます。

```
d <- read.table("clipboard", header=T)
```

※ コピーしていたデータを読み込み、d という名前をつけます。

※ 1 行目も読み込んだ場合は、header = T とすることで、「1 行目はヘッダー」と宣言。

※ 1 行目を読み込まないのであれば、header = F とします。

④ ちゃんとデータが読み込んでいるか、d とコンソールに打って確認します。

```
d
```

⑤ 分散分析を実行します。

```
anovakun(d, "AsB", group=c("group1","group2"), condition=c("15","30","60"),gg=T)
```

※anovakun()という関数で、分散分析を実施します。d は、先程作ったデータです。

※ "AsB" とあるのは実験計画を指定しています。s の前には被験者間要因、s の後には被験者内要因の数だけ A,B,C…とアルファベットを入れます。今回はどちらも 1 要因 なので s の前に A、s の後に B とします。

※ その後、アルファベットで指定した要因の順番に、要因名と各水準の名前を指定します。例えば、condition = c("15","30","60") というのは、B の要因に condition という名前をつけ、その中の水準名を 15, 30, 60 とする、と指定しています。

※ gg=T とすると、自由度の補正で、Greenhouse-Geisser の方法を使って ϵ を算出し、自由度の調整をします。

■ 被験者内2要因（2水準と3水準）の場合

| | A | B | C | D | E | F |
|----|-------|-------|-------|-------|-------|-------|
| 1 | U15 | U30 | U60 | D15 | D30 | D60 |
| 2 | 3.5 | 9.25 | 8.25 | 30.5 | 26.5 | 22.5 |
| 3 | 12.75 | 17 | 12 | 29.75 | 32.25 | 24.25 |
| 4 | 2.5 | 5 | -5.25 | 22.25 | 17.75 | 9.25 |
| 5 | 27.75 | 28.25 | 29.75 | 17 | 20.25 | 21.5 |
| 6 | 29.5 | 32 | 26.75 | 39.75 | 35.75 | 30 |
| 7 | 16.75 | 22 | 7 | 28.5 | 22.25 | 23.25 |
| 8 | 33.5 | 32.25 | 32.25 | 27 | 26.75 | 21 |
| 9 | 28 | 22.75 | 20 | 36.5 | 32.75 | 26 |
| 10 | 15 | 9.25 | 8.5 | 28.5 | 26 | 26 |
| 11 | 24.25 | 23 | 17.75 | 32.5 | 28 | 24 |
| 12 | 9.5 | 7 | 7 | 9.5 | 6.5 | 6.25 |

※ 被験者内要因を入れ子構造で入れる必要があります。上の例では、要因 A（Up か Down か）がまず大きな区分けとして、A~C 列と D~F 列にわかれています。その下に要因 B（15 度、30 度、60 度）がそれぞれきています。

※ 1 列名は、上記の点が分かるように名前を入れます。（これは自分の整理のために、分析では使いません。）

※ 2 行目以降の各列は一人ひとりの被験者のデータ情報です。（e.g., 各条件の平均値）

分析

① 下記のコードで ANOVA 君の関数を使えるように読み込みます。

```
source("anovakun_482.txt")
```

② エクセルの該当箇所を Ctrl + C でコピーします。

③ 下記のコードで、R に読み込みます。

```
d <- read.table("clipboard", header=T)
```

※ コピーしていたデータを読み込み、d という名前をつけます。

※ 1 行目も読み込んだ場合は、header = T とすることで、「1 行目はヘッダー」と宣言。

※ 1 行目を読み込まないのであれば、header = F とします。

④ ちゃんとデータが読み込んでいるか、d とコンソールに打って確認します。

```
d
```

⑤ 分散分析を実行します。

```
anovakun(d, "sAB", group=c("Up","Down"), condition=c("15","30","60"))
```

※ anovakun() という関数で、分散分析を実施します。d は、先程作ったデータです。

※ "sAB" とあるのは実験計画を指定しています。s の後には被験者内要因の数だけ A,B,C... とアルファベットを入れます。今回は 2 要因あるので、sAB となります。

※ その後、アルファベットで指定した要因の順番に、要因名と各水準の名前を指定します。例えば、condition = c("15","30","60") というのは、B の要因に condition という名前をつけ、その中の水準名を 15, 30, 60 とする、と指定しています。

※ gg=T とすると、自由度の補正で、Greenhouse-Geisser の方法を使って ϵ を算出し、自由度の調整をします。

8.3. 分散分析に関して補足

下記、詳しくは井関先生のホームページで、「ANOVA 君の使い方」のページを参考にしてください。

■ 球面性検定

- ANOVA 君では、デフォルトの球面性検定は **Mendoza の多標本球面性検定** というものです。
- ただし、**Mauchly の球面性検定** にも変更できます。その際は、「**mau = T**」をコードに追加します。

■ 自由度の調整

- 球面性の仮定が満たされないことが確認された場合、 ϵ (イプシロン) による自由度の補正を行いますが、ANOVA 君では、デフォルトでは、**自由度の調整を行いません**。
- **Greenhouse-Geisser** の方法で ϵ を算出し、**自由度の調整をする場合は、「gg = T」** をコードに追加します。
- 他の、 ϵ の算出方法としては、「下限値」「Huynh-Feldt-Lecoutre」「Chi-Muller」等があります。

コードは下記のような感じになります。

```
anovakun(d, "sAB",
         letter = c("normal", "reverse"),
         rotation = c("0", "60", "120", "180", "240", "300"),
         mau = T, # Mauchlyの球面性検定を指定する場合
         gg = T)
```

■ 多重比較

- ANOVA 君では、デフォルトの多重比較の方法は、Shaffer の方法 (Modified Sequentially Rejective Bonferroni Procedure) です。
- これも変更することができます。例えば、「holm = T」とすることで、Holm の方法を採用します。

■ 効果量を出力する

例えば、anovakun() の中に、**peta=T** を追加してあげるだけで、**偏イータ二乗を結果で表示**します。

8.4. 結果の見方

■ 被験者内1要因3水準の場合

以下では、下記のようなデータ（被験者内1要因3水準，被験者数は11名）を分析にかけた場合の，結果の見方を紹介します。

| | A | B | C |
|---|----|----|----|
| 1 | 15 | 30 | 60 |
| 2 | 17 | 18 | 15 |
| 3 | 21 | 25 | 18 |
| 4 | 12 | 11 | 2 |
| 5 | 22 | 24 | 26 |
| 6 | 25 | 21 | 28 |

[sA-Type Design] 指定したデザイン

This output was generated by anovakun 4.7.2 under R version 3.3.3.
It was executed on Fri May 11 02:04:21 2018.

<< DESCRIPTIVE STATISTICS >> 記述統計の結果

```
-----
angle  n    Mean   S.D.
-----
 15  11  22.9091  7.9430
 30  11  22.0909  8.0679
 60  11  18.0909  8.2153
-----
```

<< SPHERICITY INDICES >> 球面性の仮定の検定（"有意でない"場合は、仮定が満たされている。）

== Mendoza's Multisample Sphericity Test and Epsilons ==

```
-----
Effect  Lambda approx.Chi df    p      LB    GG    HF    CM
-----
angle  0.2227    2.7035  2  0.2588 ns  0.5000 0.7940 0.9194 0.8197
-----
```

LB = lower.bound, GG = Greenhouse-Geisser
HF = Huynh-Feldt-Lecoutre, CM = Chi-Muller

<< ANOVA TABLE >> 分散分析表 angle 要因、有意差あり($p < .001$)

```
-----
Source      SS  df    MS  F-ratio  p-value
-----
s  1845.6364  10  184.5636
-----
angle  146.2424  2   73.1212  13.1642  0.0002 ***
s x angle  111.0909  20   5.5545
-----
```

```
-----
Total 2102.9697 32 65.7178
      +p < .10, *p < .05, **p < .01, ***p < .001
```

<< POST ANALYSES >>

< MULTIPLE COMPARISON for "angle" > ④ 多重比較の結果

```
== Shaffer's Modified Sequentially Rejective Bonferroni Procedure ==
== The factor < angle > is analysed as dependent means. ==
== Alpha level is 0.05. ==
```

```
-----
angle  n    Mean   S.D.
-----
  15  11  22.9091  7.9430
  30  11  22.0909  8.0679
  60  11  18.0909  8.2153
-----
```

```
-----
Pair   Diff  t-value  df    p  adj.p          ↓ angle 要因の水準ペア毎の検定結果
-----
15-60  4.8182  3.9996  10  0.0025  0.0076  15 > 60 *
30-60  4.0000  3.9641  10  0.0027  0.0076  30 > 60 *
15-30  0.8182  1.0930  10  0.3000  0.3000  15 = 30
-----
```

output is over -----///

■ 被験者内2要因（2水準と3水準）の場合

この資料の最後にスライド資料があります。これは授業中にメンタルローテーションのデータに関して、ANOVA 君を用いて、**2要因の被験者内分散分析**を実施した結果を例にとり、その結果の見方について解説しますので、そちらを参照してください。

そのときのコードは下記の通りです。

```
# 全ての被験者内効果について、Greenhouse-Geisserのεによる調整を行う場合は、「gg=T」を追加。
anovakun(d, "sAB",
         letter = c("normal", "reverse"),
         rotation = c("0", "60", "120", "180", "240", "300"),
         mau = T, # Mauchlyの球面性検定を指定する場合
         gg = T)
```


9. おまけ

9.1. 地味によく使う機能

- R consol にカーソルを合わせた状態で、↑キーを押すと、前に書いたコードが使えます。
- R では、関数の前に ? をつけると、詳しい説明が見ることができます。英語です。

9.2. 直接コピー & ペーストでデータを読み込む方法

本文中でも紹介しましたが、この方法を知っていると、csv ファイルをいちいち作らずとも、エクセルから直接データをコピーして使えるので便利です。

①エクセルの該当部分をコピー (Ctrl + C)

②コードをうつ `k <- read.table("clipboard", header=T)`

コードの意味：クリップボードにコピーしたデータを読み込み、k と名前を付ける。

1 行目は、ヘッダーである (True の T) 。

必ず、ちゃんとデータが読み込まれているか、確認してみてくださいね。

9.3. 欠損値について

データに欠損値が含まれる場合は、該当箇所には NA と入れておきます。

k というデータから、NA のある行を削除したデータを作成して k1 と命名する場合は、下記のようにします。

```
k1 <- na.omit(k)
```

また、欠損値を含んだまま相関をかけると、エラーになってしまいます。その場合、cor () の中に、下記のように、use="complete.obs" を追記してあげることで、欠損値のないペアだけを使って相関を出してくれます。

```
cor(k$con1, k$con2, use="complete.obs")
```

9.4. もっと詳しく学びたい人は...

今は、オンラインで色々な情報を得ることができます。統計に関しても、沢山勉強になるサイトがあるので、是非参考にしてみましょう。

■ 参考 URL

奥村晴彦先生のページ

<https://oku.edu.mie-u.ac.jp/~okumura/stat/>

井関龍太先生のページ

<http://riseki.php.xdomain.jp/index.php?ANOVA%E5%90%9B>

統計科学研究所の資料

http://www.statistics.co.jp/reference/software_R/free_software-R.htm

■ 参考図書

村井潤一郎（2013） はじめての R: ごく初歩の操作から統計解析の導入まで 北大路書房