R と EZR を用いた実験データの解析

中澤　港

2013 年 11 月 5 日

このテキストは，『エビデンスベーストヘルスケア特講』用のテキスト[*1]から，公衆衛生実習で実施する実験テータの統計解析に必要な最小限の部分を抜粋し，補足説明を加筆したものである。詳しくは元テキストを参照されたい。

目次

[*1] http://minato.sip21c.org/ebhc-text.pdf

問い合わせ先：神戸大学大学院保健学研究科国際保健学領域・教授　中澤　港
e-mail: minato-nakazawa@umin.net

1 R の基本

R は MS Windows，Mac OS，Linux など，さまざまな OS で動作する。中間栄治さんが早い段階で開発に参加してくださったおかげで，テキスト画面でもグラフィック画面でも日本語の表示が可能だし，岡田昌史さんや間瀬茂さんを中心に組織されたユーザグループの協力によって，インターフェースの多くの部分で日本語に翻訳されたメッセージが利用可能である[*2]。Windows 版や Mac OS 版は，通常，実行形式になっているものをダウンロードしてインストールする。Linux では tar で圧縮されたソースコードをダウンロードして，自分でコンパイルすることも難しくないが，ubuntu などではコンパイル済みのバイナリを提供してくれている人もいるので，それを使う方が容易にインストールできるかもしれない。

R はフリーソフトなので，自分のコンピュータにインストールすることも自由にできる。R 関連のソフトウェアは CRAN (The Comprehensive R Archive Network) からダウンロードすることができる。CRAN のミラーサイトが各国に存在し，ダウンロードは国内のミラーサイトからすることが推奨されているので，日本では筑波大学[*3]，兵庫教育大学[*4]，統計数理研究所[*5]のどれかを利用すべきだろう。

1.1 R のインストール方法 [2013 年 11 月 5 日現在]

Windows　CRAN ミラーから R-3.0.2 のインストール用ファイル[*6]をダウンロードし，ダブルクリックして実行する。インストール途中で，スタートアップオプションをカスタマイズするかどうか尋ねるダイアログが表示されるので，ここは**いいえ（デフォルト）でなく，はい（カスタマイズする）の方をマークして**「次へ」をクリックすることをお薦めする[*7]。次に表示されるウィンドウで **SDI (separate windows) にチェックを入れて**「次へ」をクリックするのが重要である。他のオプションはどうでもいい。

Macintosh　Mac OS X のバージョンに注意。同じく CRAN ミラーから `R-3.0.2.pkg` をダウンロードしてダブルクリックしてインストールすればよい（ただし，それだと Tcl/Tk がインストールされないので，別途 tools フォルダの中もインストールする必要があるらしい）。群馬大学社会情報学部・青木繁伸教授のサイトに詳細な解説記事[*8]があるので参照されたい。

[*2] この翻訳作業は，R の大きなバージョンアップの際には毎回必要になるので，`http://www.okada.jp.org/RWiki/`の**日本語化掲示板**でボランティアが募集される。
[*3] `http://cran.md.tsukuba.ac.jp/`
[*4] `http://essrc.hyogo-u.ac.jp/cran/`
[*5] http://cran.ism.ac.jp/
[*6] `R-3.0.2-win.exe`
[*7] スタートアップオプションがデフォルトでは，R を起動した後のすべてのウィンドウが，1 つの大きなウィンドウの中に表示される MDI モードになってしまうのだが，それだと Rcmdr/EZR が非常に使いにくくなるからである。
[*8] `http://aoki2.si.gunma-u.ac.jp/R/begin.html`

Linux　Debian，RedHat，ubuntu など，メジャーなディストリビューションについては有志がコンパイルしたバイナリが CRAN にアップロードされているので，それを利用すればインストールは容易であろう。マイナーな環境の場合や，高速な数値演算ライブラリを使うなど自分のマシンに最適化したビルドをしたい場合は，CRAN からソース `R-3.0.2.tar.gz` をダウンロードして展開して自力でコンパイルする。最新の環境であれば，`./configure` と `make` してから，スーパーユーザになって `make install` で済むことが多いが，場合によっては多少のパッチを当てる必要がある。

1.2　R の使い方の基本

以下の解説は Windows 版による。基本的に Linux 版でも Mac OS X 版でも大差ないが，使えるグラフィックデバイスやフォントなどが多少異なるので，適宜読み替えられたい。なお，以下の本文中，\記号は¥の半角と同じものを意味する。

Windows では，インストールが完了すると，デスクトップまたはクイック起動メニューに R のアイコンができている。Rgui を起動するには，デスクトップの R のアイコンをダブルクリックするだけでいい[*9]。ウィンドウが開き，作業ディレクトリの.Rprofile が実行され，保存された作業環境.RData が読まれて，

```
>
```

と表示されて入力待ちになる。この記号>をプロンプトと呼ぶ。R への対話的なコマンド入力は，基本的にプロンプトに対して行う。閉じ括弧を付け忘れたり命令や関数の途中で改行してしまった場合はプロンプトが継続行を意味する + となることに注意されたい。なお，Windows では，どうしても継続行状態から抜けられなくなってしまった場合，ESC キーを押すとプロンプトに戻ることができる。

入力した命令や関数は，「ファイル」メニューの「履歴の保存」で保存でき，後で「ファイル」の Source で呼び出せば再現できる。プロンプトに対して source("プログラムファイル名") としても同じことになる（但し，Windows ではファイルパス中，ディレクトリ（フォルダ）の区切りは/または\\で表すことに注意[*10]。できるだけ 1 つの作業ディレクトリを決めて作業することにする方が簡単である。演習室のコンピュータでは，通常，マイドキュメントが作業ディレクトリになっているはずである）。

また，キーボードの↑を押せば既に入力したコマンドを呼び戻すことができる。

なお，R をインストールしたディレクトリの `bin` にパスを通しておけば，Windows 7/8 のコマンドプロンプトで `R` と打っても，R を起動することができる。この場合は，コマンドプロンプトが R コンソールの代わりにシェルとして動作する。

1.3　Rgui プロンプトへの基本操作

終了　`q()`

付値　<- 例えば，1，4，6 という 3 つの数値からなるベクトルを X という変数に保存するには次のようにする。

```
X <- c(1, 4, 6)
```

定義　`function()` 例えば，平均と標準偏差を計算する関数 `meansd()` の定義は次の通り。

[*9] 前もって起動アイコンを右クリックしてプロパティを選択し，「作業フォルダ (S)」に作業ディレクトリを指定しておくとよい。環境変数 R_USER も同じ作業ディレクトリに指定するとよい（ただし，システムの環境変数または作業ディレクトリに置いたテキストファイル.Renviron に，`R_USER="c:/work"`などと書いておくと，それが優先される）。また，企業ユーザなどで proxy を通さないと外部のネットワークと接続できない場合は，Windows のインターネットの設定できちんと proxy を設定した上で，起動アイコンのプロパティで，「起動コマンドのリンク先」末尾に`--internet2` と付しておく。また，日本語環境なのに R だけは英語メニューで使いたいという場合は，ここに `LANGUAGE="en"` と付しておけばいいし，R のウィンドウが大きな 1 つのウィンドウの中に開く MDI ではなく，別々のウィンドウで開く SDI にしたければ，ここに`--sdi` と付しておけばいい。

[*10] \という文字（バックスラッシュ）は，日本語キーボードでは¥ である。

```
meansd <- function(X) { list(mean(X), sd(X)) }
```

導入 `install.packages()` 例えば，CRAN から Rcmdr ライブラリをダウンロードしてインストールするには，

```
install.packages("Rcmdr", dep=TRUE)
```

とする。最初のダウンロード利用時には，ライブラリをどのミラーサーバからダウンロードするかを聞いてくるので，通常は国内のミラーサーバを指定すればよいだろう。筆者は筑波大学のサーバを利用することが多い。`dep=TRUE` は dependency（依存）が真という意味で，Rcmdr が依存している，Rcmdr 以外のライブラリも自動的にダウンロードしてインストールしてくれる。なお，`TRUE` は `T` でも有効だが，誤って `T` を変数として別の値を付値してしまっていると，意図しない動作をしてしまい，原因を見つけにくいバグの元になるので，できるだけ `TRUE` とフルスペル書いておくことが推奨されている。

ヘルプ `?` 例えば，*t* 検定の関数 `t.test` の解説をみるには，`?t.test` とする。

関数定義は何行にも渡って行うことができ，最終行の値が戻り値となる。関数内の変数は局所化されているので，関数内で変数に付値しても，関数外には影響しない。関数内で変数の値を本当に変えてしまいたいときは，通常の付値でなくて，`<<-`（永続付値）を用いる。

1.4 EZR を使う

このようなコマンドベースの使い方に習熟するには一定の時間が必要である。世界各地で R ユーザが開発した追加機能パッケージが多数公開されているが，なかでもカナダ・マクマスタ大学の John Fox 教授が開発した Rcmdr（R Commander）は，メニュー形式で R を操作できるパッケージとして有名である。Rcmdr のメニューはカスタマイズすることができるし，プラグインという仕組みで機能追加もできるので，自治医科大学の神田善伸教授が医学統計向けにフルカスタマイズし機能追加したものが EZR である[*11]。

EZR をインストールするには，Rcmdr をインストールした後で，

```
install.packages("RcmdrPlugin.EZR", dep=TRUE)
```

と打てば良い。

Rcmdr のメニューを起動するには，プロンプトに対して `library(Rcmdr)` と打てばよい。暫く待てば R Commander の GUI メニューが起動する。なお，いったん R Commander を終了してしまうと，もう一度 `library(Rcmdr)` と打っても Rcmdr は起動しない。そうではなくて，`Commander()` と打つのが正しい。ただし，`detach(package:Rcmdr)` と打って Rcmdr をアンロードしてからなら，もう一度 `library(Rcmdr)` と打つことで R Commander の GUI メニューを呼び出すことができる。

そこから EZR を呼び出すには，メニューの「ツール」から，「Rcmdr プラグインのロード」を選び，プラグインとして RcmdrPlugin.EZR を選んで OK ボタンをクリックする。少し待つと，「R コマンダーを再起動しないとプラグインを利用できません。再起動しますか？」と尋ねるダイアログが表示されるので，「はい (Y)」をクリックすると EZR が起動する。

なお，本来はオリジナルの Rcmdr メニューが「標準メニュー」に残るはずなのだが，プラグイン化と両立できなかったそうで，「標準メニュー」は階層構造が崩れてしまっているのが残念なところである。ここは今後のバージョンアップに期待したい[*12]。

[*11] `http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmed.html`

[*12] 2013 年 8 月 23 日に，Rcmdr パッケージがメジャーバージョンアップして 2.0.0 となったことがアナウンスされた。RcmdrPlugin.EZR がいつ対応するかは不明であるが，それまでは，**自治医大のサイトで公開されている，EZR を組み込み済みの R（若干古いバージョン）をダウンロードして利用すると良い。Windows 環境では，この場合，起動アイコンをクリックするだけで，R だけではなく，自動的に EZR まで起動してくれる。**

コラム: 3 つ以上の値を測る 3 つの理由

triplicate の理由　本実習の二酸化窒素測定では，プレートリーダーには同じサンプルを 3 箇所のウェルに入れて，3 つずつの値を測定した。ウェルの中に空気が入ってしまった場合など，2 つしか測っていなくて 2 つの値がかけ離れていると，どちらが正しいかわからないが，3 つあれば，2 つ以上のウェルに空気が混入してしまう可能性は低いので，おそらく近い 2 つの値の方が正しいだろうと判定できる。duplicate でなく triplicate するのはそのためである。もちろん，triplicate でも 96 穴のプレートならばいっぺんに全部測れるくらいしかサンプルがなく，コストが変わらないというのも大きな理由である。

サンプルサイズが 3 以上でなくてはいけない理由　滴定や二酸化炭素濃度測定を 1 つのサンプルについて 3 回以上した理由は，2 つ以上のサンプル間で平均値を比べる際のバラツキの信頼性を担保するためである。2 つしか値がないと，個々の測定値が平均値から相対的にどれくらい外れているかを評価できない。偏差の絶対値が標準偏差と一致してしまうため，実測値が何であろうと，偏差値を計算すると大きい方が 60，小さい方が 40 になってしまう。

検量線を書くための標準試料が 3 つ以上の濃度でなくてはいけない理由　検量線は，予め濃度または絶対量がわかっている標準試料を横軸に取り，縦軸に吸光度などをとって，両者の関係を直線として作図し（統計学的には後述する回帰直線），測定したい未知試料の吸光度から濃度または絶対量を逆算するために用いる。2 種類の標準試料しかないと，直線は必ず 2 つの測定点を通ってしまうので，標準試料調製に失敗していても気づくことすらできない。液体の塩分濃度の簡易測定器であるカーディソルト（堀場製作所）などでも，純水でゼロ点調整後のスロープの調整には 0.1% と 5% の標準液で校正する，3 点校正を行う。標準液の保存状態が悪くて濃度が狂ってしまっていると校正がうまくできないが，そのことに気づくことができるのが 3 点校正のメリットである（その場合は未開封の標準液を用意して校正をやり直す）。

2　実験計画法

実験的研究は，どんなものであれ，注意深くデザインされねばならない。……というわけで，実験計画法について説明する。

実験計画法は，R.A. Fisher がロザムステッドで行った農学研究に始まるが，保健医療分野では，この種のデザインは，毒性試験や臨床試験で用量反応関係を分析するために必須である。もちろん，ヒトを対象にした研究は，実施前に倫理審査を通らねばならず，倫理審査に提出する書類には，適切なサンプルサイズの設定を含む，適切な研究デザインが記述されねばならない。実験計画法には，目的に応じて様々なデザインがある。本稿では以下の 3 つについて説明する。

- 単一群，事前-事後デザイン
- 平行群間比較試験（完全無作為化法）
- クロスオーバー法

2.1　単一群，事前-事後デザイン

このデザインを使うと，研究者は，個々の対象者に対して同じ精度で測定された，何かの処理の前後で測定値に変化がないかどうかを評価することができる。通常使われる検定手法は，対応のある t 検定や，ウィルコクソンの符号順位検定になる（事前や事後の測定点が 2 時点以上ある場合は，反復測定分散分析やフリードマンの検定になる）。なお，対応のある t 検定は，個人ごとに算出した変化量の平均値がゼロという帰無仮説を検定する一標本 t 検定と，数学的には同値である。以下のような研究が典型的な例である。

- 慢性関節リウマチ (RA) 患者の手術前後の血清コルチゾールレベルを比較することで，手術の効果を判定
- うつ病患者の音楽療法の前後で，質問紙によるうつ得点を比較することで，音楽療法の効果を判定
- 珈琲を飲む前後で単純計算にかかる時間や正答率を比較することで，珈琲を飲むことが計算能力や集中力に影響するかを判定

2.2 平行群間比較試験（完全無作為化法）

これは非常に単純である。研究参加に同意した対象者各人に対して，完全にランダムに（行き当たりばったりに，ではなく），いくつかの処理（曝露）の１つを割り付け，処理間での比較をするというものである。

無作為化 (randomization) の方法にはいくつかある。Fleiss JL (1986)“The design and analysis of clinical experiments”は，乱数表 (random number table) の代わりに乱数順列表 (random permutation table) を使うことを推奨している。

しかし，今ではコンピュータソフトが使えるので，紙の表を使わなくても，簡単にランダム割り付けができる。

結果の解析は 2 群間での平均値の比較なら Welch の方法による t 検定，3 群以上の間での平均値の比較なら一元配置分散分析 (One-way ANOVA) になる。

2.3 クロスオーバー法 (cross-over design)

クロスオーバー法では，それぞれの対象者が 2 種類の処理を受ける。このとき，適当な間隔（ウォッシュアウト期間と呼ぶ。前の処理の影響のキャリーオーバーを避けるために設ける）をおくことと，処理の順番が違う２つのグループを設定することが必要である。

Hilman BC et al. “Intracutaneous immune serum globulin therapy in allergic children.”, JAMA. 1969; 207(5): 902-906. を例として説明しよう。

この研究では，同意が得られた 574 人から，まず研究目的に対して不適格な 43 人を除外し，531 人をランダムに 2 群に分けた。グループ 1 が 266 人，グループ 2 が 265 人となった。グループ 1 に処理 A を行い，34 人が脱落した。同時にグループ 2 には処理 B を行い，脱落が 15 人であった。その後，2 ヶ月のウォッシュアウト期間をおき，グループ 2 の 250 人に処理 A を，グループ 1 の 232 人に処理 B を行ったところ，それぞれ 45 人，29 人が脱落したので，2 回の処理を完了したのは合計 408 人となった。

本実習の産業衛生の実験では，クロスオーバー法で作業環境がストレス増加に与える影響を評価した。このデザインは少々複雑である。処理の順番が結果に影響しなければクロスオーバーにしなくても良いのだが，今回の実験デザインでは，作業前後での唾液アルブミン濃度増加が，作業環境の照度によって影響を受けないかどうか，照度条件が，先に明条件，後に暗条件か，その逆かという順番による影響がないかどうか，それらの交互作用がないかどうかをすべて調べる必要がある。仮に下表のようなデータが得られたとする（分析用のデータは，このような形式で表計算ソフトに入力する[*13]。変数 PID は被験者の識別番号を意味する。変数 Ord.LD は照度条件の順番を意味し，LD が先に明条件の群，DL が先に暗条件の群である。変数 LD は照度条件を意味し，L が明条件，D が暗条件である。変数 T0.Alb と T1.Alb は作業前後の唾液アルブミン値を意味する）。

[*13] `http://minato.sip21c.org/pubhealthpractice/crossovertest.txt` として web から入手できる。

PID	Ord.LD	LD	T0.Alb	T1.Alb
1	LD	L	15	21
2	LD	L	17	19
3	LD	L	19	25
4	LD	L	21	23
5	LD	L	23	27
6	DL	D	15	34
7	DL	D	17	27
8	DL	D	19	39
9	DL	D	21	28
10	DL	D	23	41
1	LD	D	15	35
2	LD	D	17	28
3	LD	D	19	40
4	LD	D	21	29
5	LD	D	23	42
6	DL	L	15	22
7	DL	L	17	20
8	DL	L	19	26
9	DL	L	21	24
10	DL	L	23	28

分析には（他のやり方もあるが，ここでは最も簡単な方法のみ示す),「ファイル」の「データのインポート」メニューからこのデータを読み込ませた後,「統計解析」の「連続変数の解析」から「対応のある 2 群以上の間の平均値の比較（反復測定分散分析)」を選んで,「反復測定したデータを示す変数（2 つ以上選択)」の枠から T0.Alb と T1.Alb を選び,「群別する変数を選択 (0〜複数選択可)」の下の枠から LD と Ord.LD を選んでから OK をクリックする。結果は下の枠内のように得られる（２つのグラフが自動的に描画される)。

```
Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

                                  SS num Df Error SS den Df        F    Pr(>F)
(Intercept)                   22944.1      1    402.8     16 911.3843 1.554e-15 ***
Factor1.LD                      291.6      1    402.8     16  11.5829  0.003634 **
Factor2.Ord.LD                    0.0      1    402.8     16   0.0000  1.000000
Factor1.LD:Factor2.Ord.LD         2.5      1    402.8     16   0.0993  0.756738
Time                            980.1      1    154.8     16 101.3023 2.510e-08 ***
Factor1.LD:Time                 291.6      1    154.8     16  30.1395 4.942e-05 ***
Factor2.Ord.LD:Time               0.0      1    154.8     16   0.0000  1.000000
Factor1.LD:Factor2.Ord.LD:Time    2.5      1    154.8     16   0.2584  0.618160
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

右端の Pr(>F) という列をみると，Factor1.LD の行が 5% より小さいので明条件と暗条件ではアルブミン値が有意に異なり，Time の行も 5% より小さいので作業前後でアルブミン値が有意に異なり，Factor1.LD:Time の行も 5% より小さいので作業条件と時間の交互作用効果が有意である（即ち，**明条件か暗条件かで，作業前後でのアルブミン値の変化の仕方が有意に異なる**）ことがわかる。この結果では，Factor2.Ord.LD や，それを含む行は Pr(>F) が大きい値なので，クロスオーバーデザインにはしたけれども，明暗条件の順序は結果に影響しなかったといえる。

なお，群別がない場合は，球面性検定の結果が表示され，その結果が統計学的に有意だった場合に用いられる 2 種類の補正の結果も表示されるが，このように群別変数を指定して交互作用効果を見る場合は計算されない。

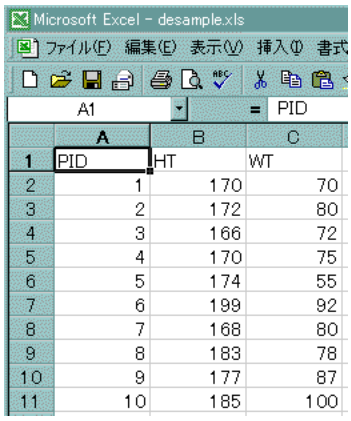
3　データ入力

研究によって得られたデータをコンピュータを使って統計的に分析するためには，まず，コンピュータにデータを入力する必要がある。データの規模や利用するソフトウェアによって，どういう入力方法が適当か（正しく入力でき，かつ効率が良いか）は異なってくる。

ごく小さな規模のデータについて単純な分析だけ行う場合，電卓で計算してもよいし，分析する手続きの中で直接数値を入れてしまってもよい。例えば，60 kg, 66 kg, 75 kg という 3 人の平均体重を R を使って求めるには，プロンプトに対して `mean(c(60,66,75))` または `(60+66+75)/3` と打てばいい。

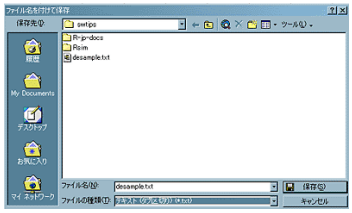
しかし実際にはもっとサイズの大きなデータについて，いろいろな分析を行う場合が多いので，データ入力と分析は別々に行うのが普通である。そのためには，同じ調査を繰り返しするとか，きわめて大きなデータであるとかでなければ，Microsoft Excel のような表計算ソフトで入力するのが手軽であろう。きわめて単純な例として，10 人の対象者についての身長と体重のデータが次の表のように得られているとする。

対象者 ID	身長 (cm)	体重 (kg)
1	170	70
2	172	80
3	166	72
4	170	75
5	174	55
6	199	92
7	168	80
8	183	78
9	177	87
10	185	100



まずこれを Microsoft Excel などの表計算ソフトに入力する。一番上の行には変数名を入れる。日本語対応 R なら漢字やカタカナ，ひらがなも使えるが，半角英数字（半角ピリオドも使える）にしておくのが無難である。入力が終わったら，一旦，そのソフトの標準の形式で保存しておく（Excel ならば*.xls 形式，OpenOffice.org の calc ならば*.ods 形式）。入力完了した状態は，右の画面のようになる。

次に，この表をタブ区切りテキスト形式で保存する。Microsoft Excel の場合，メニューバーの「ファイル (F)」から「名前を付けて保存」を選び，現れるウィンドウの一番下の「ファイルの種類 (T)」のプルダウンメニューから「テキスト（タブ区切り）(*.txt)」を選ぶと，自動的にその上の行のファイル名の拡張子も xls から txt に変わるので，「保存 (S)」ボタンを押せば OK である（下のスクリーンショットを参照）。複数のシートを含むブックの保存をサポートした形式でないとかいう警告が表示されるが無視して「はい」を選んでよい。その直後に Excel を終了しようとすると，何も変更していないのに「保存しますか」と聞く警告ウィンドウが現れるが，既に保存してあるので「いいえ」と答えてよい（「はい」を選んでも同じ内容が上書きされるだけであり問題はない）。この例では，desample.txt ができる。



R コンソールを使って，このデータを `Dataset` という名前のデータフレームに読み込むのは簡単で，次の 1 行を入

力するだけでいい（ただしテキストファイルが保存されているディレクトリが作業ディレクトリになっていなくてはいけない）。

```
Dataset <- read.delim("desample.txt")
```

Rcmdr でのタブ区切りテキストデータの読み込みは，メニューバーの「データ」から「データのインポート」の「テキストファイルまたはクリップボードから」を開いて[a]，「データセット名を入力：」の欄に適当な参照名をつけ（変数名として使える文字列なら何でもよいのだが，デフォルトでは **Dataset** となっている），「フィールドの区切り記号」を「空白」から「タブ」に変えて（「タブ」の右にある〇をクリックすればよい），**OK** ボタンをクリックしてからデータファイルを選べばよい。
なお，データをファイル保存せず，**Excel** 上で範囲を選択して「コピー」した直後であれば，「データのインポート」の「テキストファイル又はクリップボードから」を開いてデータセット名を付けた後，「クリップボードからデータを読み込む」の右のチェックボックスにチェックを入れておけば，（データファイルを選ばずに）**OK** ボタンを押しただけでデータが読み込める。
Windows 版では，**R-2.9.0** と **Rcmdr1.4-9** 以降なら，**RODBC** ライブラリの機能によって **Excel** ファイルを直接読み込むこともできる。「データ」の「データのインポート」の「**from Excel, Access, or dBase dataset**」を開いて[b]，「データセット名を入力：」の欄に適当な参照名をつけ，**Excel** ファイルを開くとシートを選ぶウィンドウが出てくるので，データが入っているシートを選べば自動的に読み込める。

[a] EZR では「ファイル」「データのインポート」の「ファイルまたはクリップボード，URL からテキストデータを読み込む」を開くが，後は同じである。
[b] EZR では，「ファイル」「データのインポート」の「Excel，Access，dBase のデータをインポート」を開く。

4　データの図示

データの大局的性質を把握するには，図示をするのが便利である。人間の視覚的認識能力は，パターン認識に関してはコンピュータより遥かに優れていると言われているから，それを生かさない手はない。また，入力ミスをチェックする上でも有効である。つまり，データ入力が終わったら，**何よりも先に図示をすべき**といえる。

変数が表す尺度の種類によって，さまざまな図示の方法がある。離散変数の場合は，度数分布図，積み上げ棒グラフ，帯グラフ，円グラフが代表的である。

survey というデータを使って例示する。**EZR** では **MASS** パッケージが最初からロードされているので，「ファイル」「パッケージに含まれるデータを読み込む」から，パッケージとして MASS をダブルクリックし，データセットとして survey をダブルクリックしてから「**OK**」ボタンをクリックする。

“survey” というデータは，アデレード大学の学生 237 人の調査結果であり，含まれている変数は，`Sex`（性別：要因型），`Wr.Hnd`（字を書く利き手の親指と小指の間隔，cm 単位：数値型），`NW.Hnd`（利き手でない方の親指と小指の間隔，cm 単位：数値型），`W.Hnd`（利き手：要因型），`Fold`（腕を組んだときにどちらが上になるか？：右が上，左が上，どちらでもない，の 3 水準からなる要因型），`Pulse`（心拍数／分：整数型），`Clap`（両手を叩き合わせた時，どちらが上にくるか？：右，左，どちらでもない，の 3 水準からなる要因型），`Exer`（運動頻度：頻繁に，時々，しない，の 3 水準からなる要因型），`Smoke`（喫煙習慣：ヘビースモーカー，定期的に吸う，時々吸う，決して吸わない，の 4 水準からなる要因型），`Height`（身長：cm 単位の数値型），`M.I`（身長の回答がインペリアル（フィート／インチ）でなされたか，メトリック（cm ／ m）でなされたかを示す要因型），`Age`（年齢：年単位の数値型）である。

このデータフレームを使って，いくつかのグラフを描いてみよう。

4.1 カテゴリ変数の場合

度数分布図 値ごとの頻度を縦棒として，異なる値ごとに，この縦棒を横に並べた図である。離散変数の名前を X とすれば，R では `barplot(table(X))` で描画される。

> **EZR では「グラフ」「棒グラフ（頻度）」を選び，変数として Smoke を選ぶと，喫煙習慣ごとの人数がプロットされる。**

積み上げ棒グラフ 値ごとの頻度の縦棒を積み上げた図である。R では

```
fx <- table(X)
barplot(matrix(fx,NROW(fx)),beside=F)
```

で描画される。**Rcmdr や EZR では描けない。**

帯グラフ 横棒を全体を 100% として各値の割合にしたがって区切って塗り分けた図である。R では

```
px <- table(X)/NROW(X)
barplot(matrix(pc,NROW(pc)),horiz=T,beside=F)
```

で描画される。**これも Rcmdr や EZR では描けない。**

4.2 連続変数の場合

ヒストグラム 変数値を適当に区切って度数分布を求め，分布の様子を見るものである。R では `hist()` 関数を用いる。デフォルトでは「適当な」区切り方として “Sturges” というアルゴリズムが使われるが，明示的に区切りを与えることもできる。また，デフォルトでは区間が「〜を超えて〜以下」であり，日本で普通に用いられる「〜以上〜未満」ではないことにも注意されたい。「〜以上〜未満」にしたいときは，`right=FALSE` というオプションを付ければ良い。R コンソールで年齢 (`Age`) のヒストグラムを描かせるには，`hist(survey$Age)` だが，「10 歳以上 20 歳未満」から 10 歳ごとの区切りでヒストグラムを描くように指定するには，`hist(survey$Age, breaks=1:8*10, right=FALSE)` とする。

> **Rcmdr や EZR では「グラフ」の「ヒストグラム」を選ぶ。survey データでは，変数として Age を選べば，年齢のヒストグラムが描ける（アデレード大学の学生のデータのはずだが，70 歳以上の人や 16.75 歳など，大学生らしくない年齢の人も含まれている）。Rcmdr や EZR では「〜以上〜未満」にはできない（裏技的には，予め「〜以上〜未満」のカテゴリデータに変換しておき，「棒グラフ（頻度）」で描画することはできる）。**

正規確率プロット 連続変数が正規分布しているかどうかを見るものである（正規分布に当てはまっていれば点が直線上に並ぶ）。R では `qqnorm()` 関数を用いる。例えば，survey データフレームの心拍数 (`Pulse`) について正規確率プロットを描くには，`qqnorm(survey$Pulse)` とする。

> **EZR では QQ プロットはできないが，「統計解析」「連続変数の解析」「正規性の検定 (Kolmogorov-Smirnov 検定)」で検定と同時にヒストグラムに正規分布の曲線を重ね描きしてくれるので，正規分布と見なせるかどうかは見当がつく。**

幹葉表示 (stem and leaf plot) 大体の概数（整数区切りとか 5 の倍数とか 10 の倍数にすることが多い）を縦に並べて幹とし，それぞれの概数に相当する値の細かい部分を葉として横に並べて作成する図。R では `stem()` 関数を用いる。同じデータで心拍数の幹葉表示をするには，`stem(survey$Pulse)` とする。

EZR では「グラフ」「幹葉表示」を選び，「変数（１つ選択）」の枠内から Pulse を選んで「OK」ボタンをクリックすれば Output ウィンドウにテキスト出力される。さまざまなオプションが指定可能である。

箱ヒゲ図 (box and whisker plot)　縦軸に変数値をとって，第１四分位を下に，第３四分位を上にした箱を書き，中央値の位置にも線を引いて，さらに第１四分位と第３四分位の差（四分位範囲）を 1.5 倍した線分をヒゲとして第１四分位の下と第３四分位の上に伸ばし，ヒゲの先より外れた値を外れ値として○をプロットした図である。カテゴリによって層別した箱ヒゲ図を横に並べて描くと，大体の分布の様子と外れ値の様子が同時に比較できるので便利である。R では `boxplot()` 関数を用いる。例えば，survey データで喫煙状況 (Smoke) 別に心拍数 (Pulse) の箱ヒゲ図を描くには，`boxplot(survey$Pulse ~ survey$Smoke)` とする。

EZR では「グラフ」の「箱ひげ図」を選ぶ。survey データで喫煙状況別に心拍数の箱ひげ図を描かせるには，「群別する変数（0〜1 つ選択）で Smoke を選び，上下のひげの位置として「第 1 四分位数-1.5x 四分位範囲、第 3 四分位数 +1.5x 四分位範囲」の左のラジオボタンをチェックして「OK」をクリックするだけでよい。似た用途のグラフとして，層別の平均とエラーバーを表示して折れ線で結ぶことも「グラフ」の「棒グラフ（平均値）」または「折れ線グラフ（平均値）」でできる。survey データで喫煙習慣ごとに心拍数の平均値とエラーバーを表示して折れ線で結びたいなら，「因子」として Smoke，「目的変数」として Pulse を選べばよい。エラーバーとしては標準誤差（デフォルト），標準偏差，信頼区間から選択できる。

散布図 (scatter plot)　２つの連続変数の関係を２次元の平面上の点として示した図である。R では `plot()` 関数を用いる。異なる群ごとに別々のプロットをしたい場合は `plot()` の `pch` オプションで塗り分けたり，`points()` 関数を使って重ね打ちしたりできる。点ごとに異なる情報を示したい場合は `symbols()` 関数を用いることができるし，複数の連続変数間の関係を調べるために，重ね描きしたい場合は `matplot()` 関数と `matpoints()` 関数を，別々のグラフとして並べて同時に示したい場合は `pairs()` 関数を用いることができる。データ点に文字列を付記したい場合は `text()` 関数が使えるし，マウスで選んだデータ点にだけ文字列を付記したい場合は `identify()` 関数が使える。`survey` データで，R コンソールを使って，横軸に年齢 (Age)，縦軸に身長 (Height) をとってプロットしたいときは，`plot(Height ~ Age, data=survey)` と打てば良い。

EZR では「グラフ」の「散布図」で描ける（「散布図行列」メニューで `pairs()` の機能も実装されている）。層別にマークを変えることもできる。

5　記述統計・分布の正規性・外れ値

記述統計は，(1) データの特徴を把握する目的，また (2) データ入力ミスの可能性をチェックする目的で計算する。あまりにも妙な最大値や最小値，大きすぎる標準偏差などが得られた場合は，入力ミスを疑って，元データに立ち返ってみるべきである。

記述統計量には，大雑把にいって，分布の位置を示す「中心傾向」と分布の広がりを示す「ばらつき」があり，中心傾向としては平均値，中央値，最頻値がよく用いられ，ばらつきとしては分散，標準偏差，四分位範囲，四分位偏差がよく用いられる。

中心傾向の代表的なものは以下の３つである。

平均値（mean）　分布の位置を示す指標として，もっとも頻繁に用いられる。実験的仮説検証のためにデザインされた式の中でも，頻繁に用いられる。記述的な指標の１つとして，平均値は，いくつかの利点と欠点をもっている。日常生活の中でも平均をとるという操作は普通に行われるから説明不要かもしれないが，数式で書くと以下の通りである。

母集団の平均値 μ（ミューと発音する）は，

$$\mu = \frac{\sum X}{N}$$

である。X はその分布における個々の値であり，N は値の総数である。$\sum$（シグマと発音する）は，一群の値の和を求める記号である。すなわち，$\sum X = X_1 + X_2 + X_3 + ... + X_N$ である。

標本についての平均値を求める式も，母集団についての式と同一である。ただし，数式で使う記号が若干異なっている。標本平均 $\bar{X}$（エックスバーと発音する）は，

$$\bar{X} = \frac{\sum X}{n}$$

である。n は，もちろん標本サイズである[*14]。

ちなみに，重み付き平均は，各々の値にある重みをかけて合計したものを，重みの合計で割った値である。式で書くと，

$$\bar{X} = \frac{n_1(\bar{X}_1) + n_2(\bar{X}_2) + ... + n_n(\bar{X}_n)}{n_1 + n_2 + ... + n_n}$$

中央値（median） 中央値は，全体の半分がその値より小さく，半分がその値より大きい，という意味で，分布の中央である。言い換えると，中央値は，頻度あるいは値の数に基づいて分布を２つに等分割する値である。中央値を求めるには式は使わない（決まった手続き＝アルゴリズムとして，並べ替え (sorting) は必要）。極端な外れ値の影響を受けにくい（言い換えると，外れ値に対して頑健である）。歪んだ分布に対する最も重要な central tendency の指標が中央値である。R で中央値を計算するには，`median()` という関数を使う。なお，データが偶数個の場合は，普通は中央にもっとも近い２つの値を平均した値を中央値として使うことになっている。

最頻値（Mode） 最頻値はもっとも度数が多い値である。すべての値の出現頻度が等しい場合は，最頻値は存在しない。R では `table(X)[which.max(table(X))]` で得られる（ただし，複数の最頻値がある場合は，これだと最も小さい値しか表示されないので要注意）。

平均値は，(1) 分布のすべての値を考慮した値である，(2) 同じ母集団からサンプリングを繰り返した場合に一定の値となる，(3) 多くの統計量や検定で使われている，という特長をもつ。標本調査値から母集団の因果関係を推論したい場合に，もっとも普通に使われる。しかし，(1) 極端な外れ値の影響を受けやすい，(2) 打ち切りのある分布では代表性を失う場合がある[*15]，という欠点があり，外れ値があったり打ち切りがあったりする分布では位置の指標として中央値の方が優れている。最頻値は，標本をとったときの偶然性の影響を受けやすいし，もっとも頻度が高い値以外の情報はまったく使われない。しかし，試験の点で何点の人が多かったかを見たい場合は最頻値が役に立つし，名義尺度については最頻値しか使えない。

ここで上げた 3 つの他に，幾何平均 (geometric mean) や調和平均 (harmonic mean) も，分布の位置の指標として使われることがある。幾何平均はデータの積の累乗根（対数をとって平均値を出して元に戻したもの），調和平均はデータの逆数の平均値の逆数であり，どちらもゼロを含むデータには使えない。大きな外れ値の影響を受けにくいという利点があり，幾何平均は，とくにデータの分布が対数正規分布に近い場合によく用いられる。

一方，分布のばらつき（Variability）の指標として代表的なものは，以下の 4 つである。

四分位範囲 (Inter-Quartile Range; IQR) 四分位範囲について説明する前に，分位数について説明する。値を小さい方から順番に並べ替えて，４つの等しい数の群に分けたときの 1/4, 2/4, 3/4 にあたる値を，四分位数 (quartile) という。1/4 の点が第１四分位，3/4 の点が第３四分位である（つまり全体の 25% の値が第１四分位より小さく，全体の 75% の値が第３四分位より小さい）。2/4 の点というのは，ちょうど順番が真中ということだから，第２四分位は中央値に等しい。ちょっと考えればわかるように，ちょうど４等分などできない場合がもちろんあって，上から数えた場合と下から数えた場合で四分位数がずれる可能性があるが，その場合はそれらを平均するのが普通である。また，最小値，最大値に，第１四分位，第３四分位と中央値を加えた５つの値を五数要約値と呼ぶことがある（R では `fivenum()` 関数で五数要約値を求めることができる）。第１四分位，第２四分位，第３四分位

[*14] 記号について注記しておくと，集合論では $\bar{X}$ は集合 X の補集合の意味で使われるが，代数では確率変数 X の標本平均が $\bar{X}$ で表されるということである。同じような記号が別の意味で使われるので混乱しないように注意されたい。補集合は X^C という表記がなされる場合も多いようである。標本平均は $\bar{X}$ と表すのが普通である。

[*15] 氷水で痛みがとれるまでにかかる時間とか，年収とか。無限に観察を続けるわけにはいかないし，年収は下限がゼロで上限はビル・ゲイツのそれのように極端に高い値があるから右すそを長く引いた分布になる。平均年収を出している統計表を見るときは注意が必要である。年収の平均的な水準は中央値で表示されるべきである。

は，それぞれ Q1, Q2, Q3 と略記することがある。四分位範囲とは，第 3 四分位と第 1 四分位の間隔である。上と下の極端な値を排除して，全体の中央付近の 50%（つまり代表性が高いと考えられる半数）が含まれる範囲を示すことができる。

四分位偏差 (Semi Inter-Quartile Range; SIQR)　四分位範囲を 2 で割った値を四分位偏差と呼ぶ。もし分布が左右対称型の正規分布であれば，中央値マイナス四分位偏差から中央値プラス四分位偏差までの幅に全データの半分が含まれるという意味で，四分位偏差は重要な指標である。IQR も SIQR も少数の極端な外れ値の影響を受けにくいし，分布が歪んでいても使える指標である。

分散 (variance)　データの個々の値と平均値との差を偏差というが，マイナス側の偏差とプラス側の偏差を同等に扱うために，偏差を二乗して，その平均をとると，分散という値になる。分散 V は，

$$V = \frac{\sum (X - \mu)^2}{N}$$

で定義される[*16]。標本数 n で割る代わりに自由度 $n - 1$ で割って，不偏分散 (unbiased variance) という値にすると，標本データから母集団の分散を推定するのに使える。即ち，不偏分散 V_{ub} は，

$$V_{ub} = \frac{\sum (X - \bar{X})^2}{n - 1}$$

である（R では `var()` で得られる）。

標準偏差 (standard deviation)　分散の平方根をとったものが標準偏差である。**平均値と次元を揃える意味**をもつ。不偏分散の平方根をとったものは，不偏標準偏差と呼ばれる（R では `sd()` で得られる）[*17]。もし分布が正規分布ならば，Mean±2SD[*18]の範囲にデータの 95% が含まれるという意味で，標準偏差は便利な指標である。なお，名前は似ているが，「標準誤差」はデータのばらつきでなくて，推定値のばらつきを示す値なので混同しないように注意されたい。例えば，平均値の標準誤差は，サンプルの不偏標準偏差をサンプルサイズの平方根で割れば得られるが，意味は，「もし標本抽出を何度も繰り返して行ったとしたら，得られる標本平均のばらつきは，一定の確率で標準誤差の範囲におさまる」ということである。

上記の記述統計量を計算するには，EZR では，「統計解析」の「連続変数の解析」の「連続変数の要約」を選べばよい。

分布の正規性の検定は，EZR では，「統計解析」の「連続変数の解析」の「正規性の検定」を選び，変数として `wbc` を選んで OK ボタンをクリックするだけでできる。自動的にコルモゴロフ＝スミルノフ検定とシャピロ＝ウィルク検定の結果が表示される。この例では結果が異なるが，これらの検定は分布の正規性へのアプローチが異なるので，結果が一致しないこともある。多くの検定手法がデータの分布の正規性を仮定しているが，この検定で正規性が棄却されたからといって機械的に変数変換やノンパラメトリック検定でなくてはいけないとは限らない。独立 2 標本の平均値の差がないという仮説を検定するための t 検定は，かなり頑健な手法なので，正規分布に従っているといえなくても，そのまま実行してもいい場合も多い。

外れ値の検定は，EZR では，「統計解析」の「連続変数の解析」の「外れ値の検定」を選んで，変数として `wbc` を指定して OK ボタンをクリックするだけでいい。外れ値を NA で置き換えた新しい変数を作成することも右のオプション指定から容易にできる。しかし，既に書いた通り，この結果だけで機械的に外れ値を除外することはお薦めできない。また，この例では “No outliers were identified.” と表示されるので，統計学的に外れ値があるとはいえない。

6　独立２標本の差の検定

医学統計でよく使われるのは，伝統的に仮説検定である。仮説検定は，意味合いからすれば，元のデータに含まれる情報量を，仮説が棄却されるかどうかという 2 値情報にまで集約してしまうことになる。これは情報量を減らしすぎで

[*16] 実際に計算するときは２乗の平均から平均の２乗を引くとよい。

[*17] 不偏分散は母分散の不偏推定量だが，不偏標準偏差は不偏分散の平方根なので分散の平方根と区別する意味で不偏標準偏差と呼ばれるだけであって，一般に母標準偏差の不偏推定量ではない。

[*18] 普通このように 2SD と書かれるが，正規分布の 97.5 パーセント点は 1.959964... なので，この 2 は，だいたい 2 くらいという意味である。

あって，点推定量と信頼区間を示す方がずっと合理的なのだが，伝統的な好みの問題なので，この演習でも検定を中心に説明する。もっとも，Rothman とか Greenland といった最先端の疫学者は，仮説検定よりも区間推定，区間推定よりも p 値関数の図示[*19]の方が遥かによい統計解析であると断言している。

典型的な例として，独立にサンプリングされた２群の平均値の差がないという帰無仮説の検定を考えよう。通常，研究者は，予め，検定の有意水準を決めておかねばならない。検定の有意水準とは，間違って帰無仮説が棄却されてしまう確率が，その値より大きくないよう定められるものである。ここで２つの考え方がある。フィッシャー流の考え方では，**p 値**（有意確率）は，観察されたデータあるいはもっと極端なデータについて帰無仮説が成り立つ条件付き確率である。もし得られた p 値が小さかったら，帰無仮説が誤っているか，普通でないことが起こったと解釈される。ネイマン＝ピアソン流の考え方では，帰無仮説と対立仮説の両方を定義しなくてはならず，研究者は繰り返しサンプリングを行ったときに得られる，この手続きの性質を調べる。即ち，本当は帰無仮説が正しくて棄却されるべきではないのに誤って棄却するという決断をしてしまう確率（これは「偽陽性」あるいは第一種の過誤と呼ばれる）と，本当は誤っている帰無仮説を誤って採択してしまう確率（第二種の過誤と呼ばれる）の両方を調べる。これら２つの考え方は混同してはならず，厳密に区別すべきである。

通常，有意水準は 0.05 とか 0.01 にする。上述の通り，検定の前に決めておくべきである。得られた有意確率がこの値より小さいとき，統計的な有意性があると考えて帰無仮説を棄却する。

独立２群間の統計的仮説検定の方法は，以下のようにまとめられる。

1. 量的変数の場合
   (a) 正規分布に近い場合[*20]：Welch の検定（R では `t.test(x,y)`）[*21]
   (b) 正規分布とかけ離れている場合：Wilcoxon の順位和検定（R では `wilcox.test(x,y)`）
2. カテゴリ変数の場合：母比率の差の検定（R では `prop.test()`）。

6.1 等分散性についての F 検定

まず，標本調査によって得られた独立した２つの量的変数 X と Y（サンプル数が各々 n_X と n_Y とする）について，平均値に差があるかどうかを検定することを考える。

２つの量的変数 `X` と `Y` の不偏分散 `SX<-var(X)` と `SY<-var(Y)` の大きい方を小さい方で（以下の説明では `SX>SY` だったとする）割った `F0<-SX/SY` が第 1 自由度 `DFX<-length(X)-1`，第 2 自由度 `DFY<-length(Y)-1` の F 分布に従うことを使って検定する。有意確率は `1-pf(F0,DFX,DFY)` で得られる。しかし，`F0` を手計算しなくても，`var.test(X,Y)` で等分散かどうかの検定が実行できる。また，１つの量的変数 `X` と１つの群分け変数 `C` があって，`C` の２群間で `X` の分散が等しいかどうか検定するというスタイルでデータを入力してある場合は，`var.test(X~C)` とすればよい。

EZR では，「統計解析」の「連続変数の解析」から「2 群の等分散性の検定（F 検定）」を選び，目的変数（１つ選択）の枠から X を，グループ（1 つ選択）の枠から C を選び（EZR では要因型にしなくても選べる），OK ボタンをクリックする。survey データで，「男女間で身長の分散に差がない」という帰無仮説を検定するには，目的変数として Height を，グループとして Sex を選び，OK ボタンをクリックすると，男女それぞれの分散と検定結果が Output ウィンドウに表示される。

[*19] リスク比あるいはオッズ比が 1 と差がないという帰無仮説については，fmsb ライブラリに `pvalueplot()` 関数として実装済みである。

[*20] `shapiro.test()` で Shapiro-Wilk の検定ができるが，その結果を機械的に適用して判断すべきではない。

[*21] それに先立って２群の間で分散に差がないという帰無仮説で F 検定し，あまりに分散が違いすぎる場合は，平均値の差の検定をするまでもなく，２群が異なる母集団からのサンプルと考えられるので，平均値の差の検定には意味がないとする考え方もある。また，かつては，まず F 検定して 2 群間で分散に差がないときは通常の t 検定，差があれば Welch の検定，と使い分けるべきという考え方が主流だったが，群馬大学社会情報学部青木繁伸教授や三重大学奥村晴彦教授のシミュレーション結果により，F 検定の結果によらず，平均値の差の検定をしたいときは常に Welch の検定をすればよいことがわかっている。

6.2 Welch の方法による t 検定

$t_0 = |E(X) - E(Y)|/\sqrt{S_X/n_X + S_Y/n_Y}$ が自由度 ϕ の t 分布に従うことを使って検定する。但し，ϕ は下式による。

$$\phi = \frac{(S_X/n_X + S_Y/n_Y)^2}{\{(S_X/n_X)^2/(n_X-1) + (S_Y/n_Y)^2/(n_Y-1)\}}$$

R では，`t.test(X,Y,var.equal=F)` だが，`var.equal` の指定を省略した時は等分散でないと仮定して Welch の検定がなされるので省略して `t.test(X,Y)` でいい。量的変数 `X` と群分け変数 `C` という入力の仕方の場合は，`t.test(X~C)` とする。survey データで「男女間で平均身長に差がない」という帰無仮説を検定したいときは，`t.test(Height ~ Sex, data=survey)` とする。

EZR の場合は「統計解析」「連続変数の解析」から「2 群間の平均値の比較（t 検定）」を選び，目的変数として Height を，比較する群として Sex を選び，「等分散と考えますか？」の下のラジオボタンを「No (Welch test)」の方をチェックして，「OK」ボタンをクリックすると，結果が Output ウィンドウに表示される。男女それぞれの平均，不偏標準偏差と検定結果の p 値が示され，エラーバーが上下に付いた平均値を黒丸でプロットし，それを直線で結んだグラフも自動的に描かれる。

なお，既に平均値と不偏標準偏差が計算されている場合の図示は，エラーバー付きの棒グラフがよく使われるが[*22]，棒グラフを描く時は基線をゼロにしなくてはいけないことに注意されたい。生データがあれば，`stripchart()` 関数を用いて，生データのストリップチャートを描き，その脇に平均値とエラーバーを付け足す方がよい。そのためには，量的変数と群別変数という形にしなくてはいけないので，たとえば，２つの量的変数 `V <- rnorm(100,10,2)` と `W <- rnorm(60,12,3)` があったら，予め

```
X <- c(V, W)
C <- as.factor(c(rep("V", length(V)), rep("W", length(W))))
x <- data.frame(X, C)
```

または

```
x <- stack(list(V=V, W=W))
names(x) <- c("X", "C")
```

のように変換しておく必要がある[*23]。プロットするには次のように入力すればよい[*24]。

```
stripchart(X~C, data=x, method="jitter", vert=TRUE)
Mx <- tapply(x$X, x$C, mean)
Sx <- tapply(x$X, x$C, sd)
Ix <- c(1.1, 2.1)
points(Ix, Mx, pch=18, cex=2)
arrows(Ix, Mx-Sx, Ix, Mx+Sx, angle=90, code=3)
```

[*22] R では，`barplot()` 関数で棒グラフを描画してから，`arrows()` 関数でエラーバーを付ける。

[*23] この操作は，EZR でも「アクティブデータセット」の「変数の操作」から「複数の変数を縦に積み重ねたデータセットを作成する」を選べば簡単に実行できる。

[*24] EZR では「グラフ」「ドットチャート」でプロットする変数として X，群分け変数として C を選ぶことで，生データについては jitter ではないが似たグラフを描くことができる。また，平均値とエラーバーを線で結んだグラフは「グラフ」「折れ線グラフ（平均値）」で描くことができる。ただし，両者を重ね合わせることは，2013 年 8 月時点の EZR のメニューからはできない。

6.3 対応のある２標本の平均値の差の検定

各対象について２つずつの値があるときは，それらを独立２標本とみなすよりも，対応のある２標本とみなす方が切れ味がよい。全体の平均に差があるかないかだけをみるのではなく，個人ごとの違いを見るほうが情報量が失われないのは当然である。

対応のある２標本の差の検定は，paired-*t* 検定と呼ばれ，意味合いとしてはペア間の値の差を計算して値の差の母平均が 0 であるかどうかを調べることになる。R で対応のある変数 `X` と `Y` の paired-*t* 検定をするには，`t.test(X,Y,paired=T)` で実行できるし，それは `t.test(X-Y,mu=0)` と等価である。

survey データで「親指と小指の間隔が利き手とそうでない手の間で差がない」という帰無仮説を検定するには，R コンソールでは，`t.test(survey$Wr.Hnd, survey$NW.Hnd, paired=TRUE)` と打てばよい。グラフは通常，同じ人のデータは線で結ぶので，例えば次のように打てば，差が 1 cm 以内の人は黒，利き手が 1 cm 以上非利き手より大きい人は赤，利き手が 1 cm 以上非利き手より小さい人は緑で，人数分の線分が描かれる。

```
Diff.Hnd <- survey$Wr.Hnd - survey$NW.Hnd
C.Hnd <- ifelse(abs(Diff.Hnd)<1, 1, ifelse(Diff.Hnd>0, 2, 3))
matplot(rbind(survey$Wr.Hnd, survey$NW.Hnd), type="l", lty=1, col=C.Hnd, xaxt="n")
axis(1, 1:2, c("Wr.Hnd", "NW.Hnd"))
```

EZR では「統計解析」「連続変数の解析」から「対応のある 2 群間の平均値の検定（paired t 検定）」を選ぶ。第１の変数として `Wr.Hnd` を，第２の変数として `NW.Hnd` を選び，[OK] ボタンをクリックすると，Output ウィンドウに結果が得られる。有意水準 5% で帰無仮説は棄却され，利き手の方がそうでない手よりも親指と小指の間隔が有意に広いといえる。

6.4 Wilcoxon の順位和検定

Wilcoxon の順位和検定は，データが外れ値を含んでいる場合や正規分布から大きく外れた分布である場合に用いられる，連続分布であるという以外にデータの分布についての仮定を置かない（ノンパラメトリックな）検定の代表的な方法である。パラメトリックな検定でいえば，*t* 検定を使うような状況，つまり，独立２標本の分布の位置に差がないかどうかを調べるために用いられる。Mann-Whitney の U 検定と（これら２つほど有名ではないが，Kendall の S 検定とも）数学的に等価である。データがもつ情報の中で，単調変換に対して頑健なのは順位なので，これを使って検定しようという発想であるが，本稿では詳細な原理の説明は省略する。

例として，survey データで，身長 (`Height`) の分布の位置が男女間で差がないという帰無仮説を検定してみよう。R コンソールでは簡単で，`wilcox.test(Height ~ Sex, data=survey)` で良い。

EZR では「統計解析」の「ノンパラメトリック検定」から「2 群間の比較（Mann-Whitney U 検定）」を選び，目的変数として `Height` を，比較する群として `Sex` を選んで「OK」ボタンをクリックする。検定結果が Output ウィンドウに表示されるだけでなく，箱ひげ図も同時に描かれる。

6.5 Wilcoxon の符号付き順位検定

Wilcoxon の符号付き順位検定は，対応のある *t* 検定のノンパラメトリック版である。ここでは説明しないが，多くの統計学の教科書に載っている。

実例だけ出しておく。`survey` データには，利き手の大きさ（親指と小指の先端の距離）を意味する `Wr.Hnd` という変数と，利き手でない方の大きさを意味する `NW.Hnd` という変数が含まれているので，これらの分布の位置に差が無いという帰無仮説を有意水準 5% で検定してみよう。

同じ人について利き手と利き手でない方の手の両方のデータがあるので対応のある検定が可能になる。R コンソール

では，wilcox.test(survey$Wr.Hnd, survey$NW.Hnd, paired=TRUE) とすればよい。

EZR では，「統計解析」「ノンパラメトリック検定」から「対応のある 2 群間の比較（Wilcoxon の符号付き順位和検定）」を選び，第 1 の変数として左側のリストから Wr.Hnd を選び，第 2 の変数として右側のリストから NW.Hnd を選んで [OK] ボタンをクリックするだけである。順位和検定のときと同じく検定方法のオプションを指定できるが，通常はデフォルトで問題ない。

7　3 群以上の比較

3 群以上を比較するために，**単純に 2 群間の差の検定を繰り返すことは誤り**である。なぜなら，n 群から 2 群を抽出するやりかたは ${}_nC_2$ 通りあって，1 回あたりの第 1 種の過誤（本当は差がないのに，誤って差があると判定してしまう確率）を 5% 未満にしたとしても，3 群以上の比較全体として「少なくとも 1 組の差のある群がある」というと，全体としての第 1 種の過誤が 5% よりずっと大きくなってしまうからである。

この問題を解消するには，**多群間の比較という捉え方をやめて，群分け変数が注目している量の変数に与える効果があるかどうかという捉え方にする**のが一つの方法であり，具体的には一元配置分散分析やクラスカル＝ウォリス (Kruskal-Wallis) の検定がこれに当たる。

そうでなければ，有意水準 5% の 2 群間の検定を繰り返すことによって全体として第 1 種の過誤が大きくなってしまうことが問題なので，**第 1 種の過誤を調整することによって全体としての検定の有意水準を 5% に抑える**方法もある。このやり方は「多重比較法」と呼ばれる。

7.1　一元配置分散分析

一元配置分散分析では，データのばらつき（変動）を，群間の違いという意味のはっきりしているばらつき（群間変動）と，各データが群ごとの平均からどれくらいばらついているか（誤差）をすべての群について合計したもの（誤差変動）に分解して，前者が後者よりもどれくらい大きいかを検討することによって，群分け変数がデータの変数に与える効果があるかどうかを調べる。

例えば，南太平洋の 3 つの村 X, Y, Z で健診をやって，成人男性の身長や体重を測ったとしよう。このとき，データは例えば次のようになる（架空のものである）[*25]。

ID 番号	村落 (VG)	身長 (cm)(HEIGHT)
1	X	161.5
2	X	167.0
（中略）		
22	Z	166.0
（中略）		
37	Y	155.5

村落によって身長に差があるかどうかを検定したいならば，HEIGHT という量的変数に対して，VG という群分け変数の効果があるかどうかを一元配置分散分析することになる。R コンソールでは以下のように入力する。

```
> sp <- read.delim("http://minato.sip21c.org/grad/sample2.dat")
> summary(aov(HEIGHT ~ VG, data=sp))
```

すると，次の枠内に示す「分散分析表」が得られる。

[*25] http://minato.sip21c.org/grad/sample2.dat として公開してあり，R から read.delim() 関数で読み込み可能な筈である。

```
            Df  Sum Sq Mean Sq F value   Pr(>F)
VG           2  422.72  211.36  5.7777 0.006918 **
Residuals   34 1243.80   36.58
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

右端の*の数は有意性を示す目安だが，確率そのものに注目してみるほうがよい。Sum Sq のカラムは偏差平方和を意味する。VG の Sum Sq の値 422.72 は，村ごとの平均値から総平均を引いて二乗した値を村ごとの人数で重み付けした和であり，群間変動または級間変動と呼ばれ，VG 間でのばらつきの程度を意味する。Residuals の Sum Sq の値 1243.80 は各個人の身長からその個人が属する村の平均身長を引いて二乗したものの総和であり，誤差変動と呼ばれ，村によらない（それ以外の要因がないとすれば偶然の）ばらつきの程度を意味する。Mean Sq は平均平方和と呼ばれ，偏差平方和を自由度 (Df) で割ったものである。平均平方和は分散なので，VG の Mean Sq の値 211.36 は群間分散または級間分散と呼ばれることがあり，Residuals の Mean Sq の値 36.58 は誤差分散と呼ばれることがある。F value は分散比と呼ばれ，群間分散の誤差分散に対する比である。この場合の分散比は第 1 自由度 2，第 2 自由度 34 の F 分布に従うことがわかっているので，それを使った検定の結果，分散比がこの実現値よりも偶然大きくなる確率（Pr(>F) に得られる）が得られる。この例では 0.006918 なので，VG の効果は 5% 水準で有意であり，帰無仮説は棄却される。つまり，身長は村落によって有意に異なることになる。

EZR で sample2.dat **を** sp **というデータフレームに読み込むには，[ファイル] の [データのインポート] から [テキストファイル，クリップボードまたは URL から] と進んで，[データフレーム名を入力:] のところに** sp **と打ち，[インターネット URL] の右側のラジオボタンをチェックし，フィールド区切りを [タブ] として [OK] をクリックして表示されるダイアログに** http://minato.sip21c.org/grad/sample2.dat **と入力して [OK] する。**
ANOVA を実行するには，[統計解析] の [連続変数の解析] で [三群以上の間の平均値の比較（一元配置分散分析 one-way ANOVA)] を選び，「目的変数」として HEIGHT **を，「比較する群」として** VG **を選び，[OK] をクリックすればよい。エラーバー付きの棒グラフが自動的に描かれ，アウトプットウィンドウには分散分析表に続いて，村ごとの平均値と標準偏差の一覧表が表示される。右端の p 値は一元配置分散分析における** VG **の効果の検定結果を再掲したものになっている。**

古典的な統計解析では，各群の母分散が等しいことを確認しないと一元配置分散分析の前提となる仮定が満たされない。母分散が等しいという帰無仮説を検定するには，バートレット (Bartlett) の検定と呼ばれる方法がある。R では，量的変数を Y，群分け変数を C とすると，bartlett.test(Y~C) で実行できる[*26]。この結果得られる p 値は 0.5785 なので，母分散が等しいという帰無仮説は有意水準 5% で棄却されない。これを確認できると，安心して一元配置分散分析が実行できる。

EZR では，メニューバーの「統計解析」から「連続変数の解析」の「三群以上の等分散性の検定 (Bartlett 検定)」を選び，「目的変数」として HEIGHT**，「グループ」として** VG **を選んで [OK] する。**

しかし，このような 2 段階の検定は，検定の多重性の問題を起こす可能性がある。群馬大学の青木繁伸教授や三重大学の奥村晴彦教授の数値実験によると，等分散であるかどうかにかかわらず，2 群の平均値の差の Welch の方法を多群に拡張した方法を用いるのが最適である。R では oneway.test() で実行できる。上記，村落の身長への効果をみる例では，oneway.test(HEIGHT ~ VG, data=sp) と打てば，Welch の拡張による一元配置分散分析ができて，以下の結果が得られる。

[*26] もちろん，これらがデータフレーム dat に含まれる変数ならば，bartlett.test(Y~C, data=dat) とする。

```
> oneway.test(HEIGHT ~ VG, data=sp)

One-way analysis of means (not assuming equal variances)

data:  HEIGHT and VG
F = 7.5163, num df = 2.00, denom df = 18.77, p-value = 0.004002
```

残念ながら **EZR** ではメニューにないので，スクリプトウィンドウで aov の部分を `oneway.test` に書き直して「実行」するしかない。

7.2 クラスカル＝ウォリス (Kruskal-Wallis) の検定と Fligner-Killeen の検定

多群間の差を調べるためのノンパラメトリックな方法としては，クラスカル＝ウォリス (Kruskal-Wallis) の検定が有名である。R では，量的変数を `Y`，群分け変数を `C` とすると，`kruskal.test(Y~C)` で実行できる。以下，Kruskal-Wallis の検定の仕組みを箇条書きで説明する。

- 「少なくともどれか 1 組の群間で大小の差がある」という対立仮説に対する「すべての群の間で大小の差がない」という帰無仮説を検定する。
- まず 2 群の比較の場合の順位和検定と同じく，すべてのデータを込みにして小さい方から順に順位をつける（同順位がある場合は平均順位を与える）。
- 次に，各群ごとに順位を足し合わせて，順位和 $R_i (i = 1, 2, ..., k;\ k$ は群の数) を求める。
- 各群のオブザーベーションの数をそれぞれ n_i とし，全オブザーベーション数を N としたとき，各群について統計量 B_i を $B_i = n_i\{R_i/n_i - (N+1)/2\}^2$ として計算し，

$$B = \sum_{i=1}^{k} B_i$$

として B を求め，$H = 12 \cdot B/\{N(N+1)\}$ として H を求める。同順位を含むときは，すべての同順位の値について，その個数に個数の 2 乗から 1 を引いた値を掛けたものを計算し，その総和を A として，

$$H' = \frac{H}{1 - \frac{A}{N(N^2-1)}}$$

により H を補正した値 H' を求める。

- H または H' から表を使って（データ数が少なければ並べかえ検定によって）有意確率を求めるのが普通だが，$k \geq 4$ で各群のオブザーベーション数が最低でも 4 以上か，または $k = 3$ で各群のオブザーベーション数が最低でも 5 以上なら，H や H' が自由度 $k - 1$ のカイ二乗分布に従うものとして検定できる。

上の例で村落の身長への効果をみるには，R コンソールでは，`kruskal.test(HEIGHT ~ VG, data=sp)` と打てば結果が表示される。

EZR では，「統計解析」，「ノンパラメトリック検定」，「**3** 群以上の間の比較 **(Kruskal-Wallis** の検定**)**」と選び，「グループ」として `VG` を，「目的変数」として `HEIGHT` を選び，**[OK]** をクリックするだけである。

Fligner-Killeen の検定は，グループごとのばらつきに差が無いという帰無仮説を検定するためのノンパラメトリックな方法である。Bartlett の検定のノンパラメトリック版といえる。上の例で，身長のばらつきに村落による差が無いという帰無仮説を検定するには，R コンソールでは，`fligner.test(HEIGHT ~ VG, data=sp)` とすればよい。

EZR のメニューには入っていないので，必要な場合はスクリプトウィンドウにコマンドを打ち，選択した上で「実行」ボタンをクリックする。

7.3 検定の多重性の調整を伴う対比較

多重比較の方法にはいろいろあるが，良く使われているものとして，ボンフェローニ (Bonferroni) の方法，ホルム (Holm) の方法，シェフェ (Scheffé) の方法，テューキー (Tukey) の HSD，ダネット (Dunnett) の方法，ウィリアムズ (Williams) の方法がある。最近は，FDR(False Discovery Rate) 法もかなり使われるようになった。ボンフェローニの方法とシェフェの方法は検出力が悪いので，特別な場合を除いては使わない方が良い。テューキーの HSD またはホルムの方法が薦められる。なお，ダネットの方法は対照群が存在する場合に対照群と他の群との比較に使われるので，適用場面が限定されている[*27]。ウィリアムズの方法は対照群があって他の群にも一定の傾向が仮定される場合には最高の検出力を発揮するが，ダネットの方法よりもさらに限られた場合にしか使えない。

テューキーの HSD は平均値の差の比較にしか使えないが，ボンフェロー二の方法，ホルムの方法，FDR 法は位置母数のノンパラメトリックな比較にも，割合の差の検定にも使える。R コンソールでは，`pairwise.t.test()`，`pairwise.wilcox.test()`，`pairwise.prop.test()` という関数で，ボンフェローニの方法，ホルムの方法，FDR 法による検定の多重性の調整ができる。fmsb ライブラリを使えば，`pairwise.fisher.test()` により，Fisher の直接確率法で対比較をした場合の検定の多重性の調整も可能である。

なお，Bonferroni のような多重比較法で p 値を調整して表示するのは表示上の都合であって，本当は帰無仮説族レベルでの有意水準を変えているのだし，p.adjust.method="fdr"でも，p 値も有意水準も調整せず，帰無仮説の下で偶然 p 値が有意水準未満になって棄却されてしまう確率（誤検出率）を計算し，帰無仮説ごとに有意水準に誤検出率を掛けて p 値との大小を比較して検定するということになっているが，これは弱い意味で帰無仮説族レベルでの有意水準の調整を意味する，と原論文に書かれているので，統計ソフトが p 値を調整した値を出してくるのはやはり表示上の都合で，本当は有意水準を調整している（参照：Benjamini Y, Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Royal Stat. Soc. B, 57: 289-300, 1995.）。

Bonferroni，Holm，FDR という 3 つの多重比較の考え方はシンプルでわかりやすいので，ここで簡単にまとめておく。k 個の帰無仮説について検定して得られた p 値が p(1)<p(2)<...<p(k) だとすると，有意水準 α で帰無仮説族の検定をするために，Bonferroni は p(1) から順番に α/k と比較し，p(i) ≧ α/k になったところ以降判定保留，Holm は p(i) ≧ α/i となったところ以降判定保留とする。有意水準 α で fdr をするには，まず p(k) を α と比較し，次に p(k-1) を $\alpha \times$ (k-1)/k と比較し，と p 値が大きい方から比較していき，p(i)< $\alpha \times$ i/k となったところ以降，i 個の帰無仮説を棄却する。R の pairwise.*.test() では，Bonferroni ならすべての p 値が k 倍されて表示，Holm では小さい方から i 番目の p 値が i 倍されて表示，fdr では小さい方から i 番目の p 値が k/i 倍されて表示されることによって，表示された p 値を共通の α との大小で有意性判定ができるわけだが，これは表示上の都合である。（**残念ながら，FDR 法はまだ Rcmdr や EZR のメニューには含まれていない**）

実例を示そう。先に示した 3 村落の身長データについて，どの村落とどの村落の間で身長に差があるのかを調べたい場合，R コンソールでは，

```
pairwise.t.test(sp$HEIGHT, sp$VG, p.adjust.method="bonferroni")
```

とすれば，2 村落ずつのすべての組み合わせについてボンフェローニの方法で有意水準を調整した p 値が表示される[*28]。

また，`pairwise.wilcox.test(sp$HEIGHT,sp$VG,p.adjust.method="bonferroni")` とすれば，ボンフェローニの方法で有意水準を調整した，すべての村落間での身長の差を順位和検定した結果を出してくれる。これらの関数で，`p.adjust.method` を指定しなければホルムの方法になるが，明示したければ，`p.adjust.method="holm"`とすればよい。FDR 法を使うには，`p.adjust.method="fdr"`とすればよい。R でもボンフェローニが可能なのは，一番単純な方法であるという理由と，ホルムの方法に必要な計算がボンフェローニの計算を含むからだと思われる。なお，R を使って分析するのだけれども，データがきれいな正規分布をしていて，かつ古典的な方法の論文しか accept しない

[*27] ただし，対照群が他の群との比較のすべての場合において差があるといいたい場合は，多重比較をするのではなくて，*t* 検定を繰り返して使うのが正しいので，注意が必要である。もちろんそういう場合は多くはない。

[*28] `"bonferroni"`は`"bon"`でも良い。また，`pairwise.*`系の関数では `data=`というオプションが使えないので，データフレーム内の変数を使いたい場合は，予めデータフレームを `attach()` しておくか，またはここで示したように，変数指定の際に一々，“データフレーム名 $ ” を付ける必要がある。

雑誌に対してどうしても投稿したい，という場合は，TukeyHSD(aov(HEIGHT ~ VG, data=sp)) などとして，テューキーの HSD を行うことも可能である。

EZR では，一元配置分散分析メニューのオプションとして実行できる。「統計解析」「連続変数の解析」「3 群以上の平均値の比較（一元配置分散分析 one-way ANOVA)」を選んで，「目的変数」として HEIGHT を，「比較する群」として VG を選んでから，下の方の「↓ 2 組ずつの比較 (post-hoc 検定) は比較する群が 1 つの場合のみ実施される」から欲しい多重比較法の左側のボックスにチェックを入れてから「OK」ボタンをクリックする。

いずれのやり方をしても，TukeyHSD の場合だと，2 群ずつの対比較の結果として，差の推定値と 95% 同時信頼区間に加え，Tukey の方法で検定の多重性を調整した p 値が下記のように表示され，検定の有意水準が 5% だったとすると，Z と Y の差だけが有意であることがわかる。

```
> TukeyHSD(AnovaModel.3, "factor(VG)")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = HEIGHT ~ factor(VG), data = sp, na.action = na.omit)

$`factor(VG)`
         diff        lwr      upr     p adj
Y-X -2.538889 -8.3843982  3.30662 0.5423397
Z-X  5.850000 -0.9598123 12.65981 0.1038094
Z-Y  8.388889  2.3382119 14.43957 0.0048525
```

7.4 Dunnett の多重比較法

Dunnett の多重比較は，コントロールと複数の実験群の比較というデザインで用いられる。以下，簡単な例で示す。例えば，5 人ずつ 3 群にランダムに分けた高血圧患者がいて，他の条件（食事療法，運動療法など）には差をつけずに，プラセボを 1 ヶ月服用した群の収縮期血圧 (mmHg 単位) の低下が 5, 8, 3, 10, 15 で，代表的な薬を 1 ヶ月服用した群の低下は 20, 12, 30, 16, 24 で，新薬を 1 ヶ月服用した群の低下が 31, 25, 17, 40, 23 だったとしよう。このとき，プラセボ群を対照として，代表的な薬での治療及び新薬での治療に有意な血圧降下作用の差が出るかどうかを見たい（悪くなるかもしれないので両側検定で）という場合に，Dunnett の多重比較を使う。R でこのデータを bpdown というデータフレームに入力して Dunnett の多重比較をするためには，次のコードを実行する。

```
bpdown <- data.frame(
 medicine=factor(c(rep(1,5),rep(2,5),rep(3,5)), labels=c("プラセボ","代表薬","新薬")),
 sbpchange=c(5, 8, 3, 10, 15, 20, 12, 30, 16, 24, 31, 25, 17, 40, 23))
summary(res1 <- aov(sbpchange ~ medicine, data=bpdown))
library(multcomp)
res2 <- glht(res1, linfct = mcp(medicine = "Dunnett"))
confint(res2, level=0.95)
summary(res2)
```

つまり，基本的には，multcomp ライブラリを読み込んでから，分散分析の結果を glht() 関数に渡し，linfct オプションで，Dunnett の多重比較をするという指定を与えるだけである。multcomp ライブラリのバージョン 0.993 まで使えた simtest() 関数は，0.994 から使えなくなったので注意されたい。

EZR では，まず「ファイル」「データのインポート」「ファイルまたはクリップボード，**URL** からテキストデータを読み込む」として，「データセット名を入力」の右側のボックスに bpdown と入力し，「データファイルの場所」として「インターネットの **URL**」の右側のラジオボタンをクリックし，「フィールドの区切り記号」を「タブ」にして「**OK**」ボタンをクリックする。表示される **URL** 入力ウィンドウに `http://minato.sip21c.org/bpdown.txt` と打って「**OK**」ボタンをクリックすれば，上記データを読み込むことができる。
そこで「統計解析」の「連続変数の解析」から「**3** 群以上の平均値の比較（一元配置分散分析 **one-way ANOVA)**」を選んで，「目的変数」として sbpchange，「比較する群」として medicine を選び，「**2** 組ずつの比較 **(Dunnett** の多重比較**)**」の左のチェックボックスをチェックしてから「**OK**」ボタンをクリックすればいい。
なお，このデータで処理名を示す変数 medicine の値として o.placebo，1.usual，2.newdrug のように先頭に数字付けた理由は，それがないと水準がアルファベット順になってしまい，**Dunnett** の解析において新薬群がコントロールとして扱われてしまうからである。
ノンパラメトリック検定の場合は，「統計解析」の「ノンパラメトリック検定」から「**3** 群以上の間の比較 **(Kruskal-Wallis** 検定**)**」と選び，「目的変数」を sbpchange，「グループ」を medicine にし，「**2** 組ずつの比較 **(post-hoc** 検定、**Steel** の多重比較**)**」の左のチェックボックスをチェックして「**OK**」ボタンをクリックすれば，**Steel** の多重比較が実行できる。

8　２つの量的な変数間の関係

２つの量的な変数間の関係を調べるための，良く知られた方法が２つある。相関と回帰である。いずれにせよ，まず散布図を描くことは必須である。

MASS ライブラリの survey データフレームで，身長と利き手の大きさ（親指の先端と小指の先端の距離）の関係を調べるには，R コンソールでは，`require(MASS)` として MASS ライブラリをメモリに読み込んだ後であれば，`plot(Wr.Hnd ~ Height, data=survey)` とするだけである。もし男女別にプロットしたければ，`pch=as.integer(Sex)` というオプションを指定すれば良い。

EZR では，「ファイル」「パッケージに含まれるデータを読み込む」から左の枠の MASS でダブルクリックし，右の枠に現れるデータフーレムの下の方へスクロールして survey でダブルクリックしてから **OK** ボタンをクリックした後に，「グラフ」「散布図」と選び，**x** 変数として Height を，**y** 変数として Wr.Hnd を選び，**“**最小 **2** 乗直線**”** の左側のチェックボックスのチェックを外し，**[OK]** をクリックする。男女別にプロット記号を変えたい場合は，「層別のプロット」というボタンをクリックし，層別変数として Sex を選んで **[OK]** をクリックし，元のウィンドウに戻ったら再び **[OK]** をクリックすればよい。

8.1　相関と回帰の違い

大雑把に言えば，相関が変数間の関連の強さを表すのに対して，回帰はある変数の値のばらつきがどの程度他の変数の値のばらつきによって説明されるかを示す。回帰の際に，説明される変数を（従属変数または）目的変数，説明するための変数を（独立変数または）説明変数と呼ぶ。２つの変数間の関係を予測に使うためには，回帰を用いる。

8.2　相関分析

一般に，2 個以上の変量が「かなりの程度の規則正しさをもって，増減をともにする関係」のことを相関関係 (correlation) という。相関には正の相関 (positive correlation) と負の相関 (negative correlation) があり，一方が増えれば他方も増える場合を正の相関，一方が増えると他方は減る場合を負の相関と呼ぶ。例えば，身長と体重の関係は正の相関である。

散布図で相関関係があるように見えても，見かけの相関関係 (apparent correlation) であったり[*29]，擬似相関 (spurious correlation) であったり[*30]することがあるので，注意が必要である。

相関関係は増減をともにすればいいので，直線的な関係である必要はなく，二次式でも指数関数でもシグモイドでもよいが，通常，直線的な関係をいうことが多い（指標はピアソンの積率相関係数）。曲線的な関係の場合，直線的になるように変換したり，ノンパラメトリックな相関の指標（順位相関係数）を計算する。順位相関係数としてはスピアマンの順位相関係数が有名である。

ピアソンの積率相関係数 (Pearson's Product Moment Correlation Coefficient) は，r という記号で表し，２つの変数 X と Y の共分散を X の分散と Y の分散の積の平方根で割った値であり，範囲は $[-1, 1]$ である。最も強い負の相関があるとき $r = -1$，最も強い正の相関があるとき $r = 1$，まったく相関がないとき（2 つの変数が独立なとき），$r = 0$ となることが期待される。X の平均を $\bar{X}$，Y の平均を $\bar{Y}$ と書けば，次の式で定義される。

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

相関係数の有意性の検定においては，母相関係数がゼロ（＝相関が無い）という帰無仮説の下で，実際に得られている相関係数よりも絶対値が大きな相関係数が偶然得られる確率（これを「有意確率」という。通常，記号 p で表すので，「p 値」とも呼ばれる）の値を調べる。偶然ではありえないほど珍しいことが起こったと考えて，帰無仮説が間違っていたと判断するのは有意確率がいくつ以下のときか，という水準を有意水準といい，検定の際には予め有意水準を（例えば 5% と）決めておく必要がある。例えば $p = 0.034$ であれば，有意水準 5% で有意な相関があるという意思決定を行なうことができる。p 値は，検定統計量

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

が自由度 $n - 2$ の t 分布に従うことを利用して求められる。

散布図を描いた survey データフレームの身長と利き手の大きさの間でピアソンの相関係数を計算し，その有意性を検定するには，R コンソールでは次の 1 行を打てばよい（スピアマンの順位相関について実行したい時は，methos=spearman を付ける）。

```
cor.test(survey$Height, survey$Wr.Hnd)
```

EZR では，「統計解析」の「連続変数の解析」から「相関係数の検定（Pearson の積率相関係数）」を選び，変数として Height と Wr.Hnd を選ぶ（Ctrl キーを押しながら変数名をクリックすれば複数選べる）。検定については「対立仮説」の下に「両側」「相関＜０」「相関＞０」の３つから選べるようになっているが，通常は「両側」でよい。OK をクリックすると，Rcmdr の出力ウィンドウに次の内容が表示される。

```
Pearson's product-moment correlation

data:  survey$Height and survey$Wr.Hnd
t = 10.7923, df = 206, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5063486 0.6813271
sample estimates:
      cor
0.6009909
```

[*29] 例）同業の労働者集団の血圧と所得。どちらも一般に加齢に伴って増加する。

[*30] 例）ある年に日本で植えた木の幹の太さと同じ年に英国で生まれた少年の身長を 15 年分，毎年 1 回測ったデータには相関があるようにみえるが，直接的な関係はなく，どちらも時間経過に伴って大きくなるために相関があるように見えているだけである。

これより，身長と利き手の大きさの関係について求めたピアソンの積率相関係数は，$r = 0.60$（95% 信頼区間が [0.50, 0.69]）であり[*31]，`p-value < 2.2e-16`（有意確率が 2.2×10^{-16} より小さいという意味）より，「相関が無い」可能性はほとんどゼロなので，有意な相関があるといえる。なお，相関の強さは相関係数の絶対値の大きさによって判定し，伝統的に 0.7 より大きければ「強い相関」，0.4〜0.7 で「中程度の相関」，0.2〜0.4 で「弱い相関」とみなすのが目安なので，この結果は中程度の相関を示すといえる。

順位相関係数の定義

スピアマンの順位相関係数 ρ は[a]，値を順位で置き換えた（同順位には平均順位を与えた）ピアソンの積率相関係数と同じである。X_i の順位を R_i，Y_i の順位を Q_i とかけば，

$$\rho = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2$$

となる。スピアマンの順位相関係数がゼロと差がないことを帰無仮説とする両側検定は，サンプル数が 10 以上ならばピアソンの場合と同様に，

$$T = \frac{\rho \sqrt{n-2}}{\sqrt{1-\rho^2}}$$

が自由度 $n-2$ の t 分布に従うことを利用して行うことができる。ケンドールの順位相関係数 τ は，

$$\tau = \frac{(A-B)}{n(n-1)/2}$$

によって得られる。ここで A は順位の大小関係が一致する組の数，B は不一致数である。
R コンソールで順位相関係数を計算するには，`cor.test()` 関数の中で，`method="spearman"`または `method="kendall"` と指定すれば良い。EZR では，「統計解析」「ノンパラメトリック検定」「相関係数の検定（Spearman の順位相関係数）」から，解析方法のところで Spearman か Kendall の横のラジオボタンを選んで OK ボタンをクリックすれば計算できる。

[a] ピアソンの相関係数の母相関係数を ρ と書き，スピアマンの順位相関係数を r_s と書く流儀もある。

8.3 回帰モデルの当てはめ

回帰は，従属変数のばらつきを独立変数のばらつきで説明するというモデルの当てはめである。十分な説明ができるモデルであれば，そのモデルに独立変数の値を代入することによって，対応する従属変数の値が予測あるいは推定できるし，従属変数の値を代入すると，対応する独立変数の値が逆算できる。こうした回帰モデルの実用例の最たるものが**検量線**である。検量線とは，実験において予め濃度がわかっている標準物質を測ったときの吸光度のばらつきが，その濃度によってほぼ完全に（通常 98% 以上）説明されるときに（そういう場合は，散布図を描くと，点々がだいたい直線上に乗るように見える），その関係を利用して，サンプルを測ったときの吸光度からサンプルの濃度を逆算するための回帰直線である（曲線の場合もあるが，通常は何らかの変換をほどこし，線形回帰にして利用する）。

検量線の計算には，(A) 試薬ブランクでゼロ点調整をした場合の原点を通る回帰直線を用いる場合と，(B) 純水でゼロ点調整をした場合の切片のある回帰直線を用いる場合がある。例えば，濃度の決まった標準希釈系列 (0, 1, 2, 5, 10 $\mu g/\ell$) について，純水でゼロ点調整をしたときの吸光度が，(0.24, 0.33, 0.54, 0.83, 1.32) だったとしよう。吸光度の変数を y，濃度を x と書けば，回帰モデルは $y = bx + a$ とおける。係数 a と b（a は切片，b は回帰係数と呼ばれる）は，次の偏差平方和を最小にするように，最小二乗法で推定される。

$$f(a,b) = \sum_{i=1}^{5} (y_i - bx_i - a)^2$$

この式を解くには，$f(a,b)$ を a ないし b で偏微分したものがゼロに等しいときを考えればいいので，次の 2 つの式

[*31] 95% 信頼区間の桁を丸めて示す場合，真の区間を含むようにするために，四捨五入ではなく，下限は切り捨て，上限は切り上げにするのが普通である。

が得られる。

$$b = \frac{\sum_{i=1}^{5} x_i y_i/5 - \sum_{i=1}^{5} x_i/5 \cdot \sum_{i=1}^{5} y_i/5}{\sum_{i=1}^{5} {x_i}^2/5 - \left(\sum_{i=1}^{5} x_i/5\right)^2}$$

$$a = \sum_{i=1}^{5} y_i/5 - b \cdot \sum_{i=1}^{5} x_i/5$$

これらの a と b の値と，未知の濃度のサンプルについて測定された吸光度（例えば 0.67 としよう）から，そのサンプルの濃度を求めることができる。注意すべきは，サンプルについて測定された吸光度が，標準希釈系列の吸光度の範囲内になければならないことである。回帰モデルが標準希釈系列の範囲外でも直線性を保っている保証は何もないのである[*32]。

R コンソールでは，`lm()`（linear model の略で線形モデルの意味）を使って，次のようにデータに当てはめた回帰モデルを得ることができる。

```
y <- c(0.24, 0.33, 0.54, 0.83, 1.32)
x <- c(0, 1, 2, 5, 10)
# 線形回帰モデルを当てはめる
res <- lm(y ~ x)
# 詳しい結果表示
summary(res)
# 散布図と回帰直線を表示する
plot(y ~ x)
abline(res)
# 吸光度 0.67 に対応する濃度を計算する
(0.67 - res$coef[1])/res$coef[2]
```

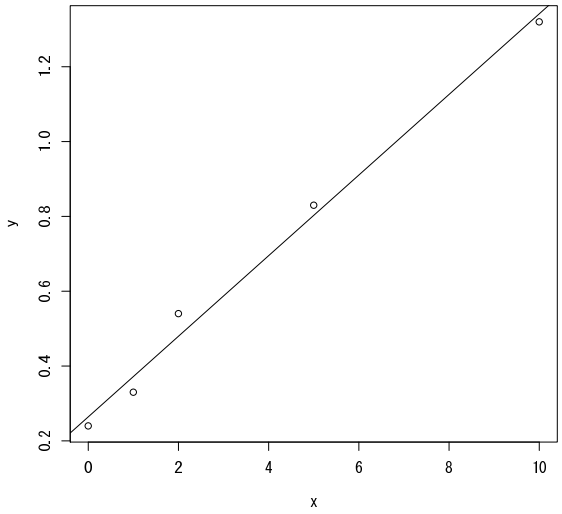
結果は次のように得られる。

```
Call:
lm(formula = y ~ x)

Residuals:
       1        2        3        4        5
-0.02417 -0.04190  0.06037  0.02718 -0.02147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.26417    0.03090   8.549 0.003363 **
x            0.10773    0.00606  17.776 0.000388 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04894 on 3 degrees of freedom
Multiple R-squared: 0.9906,     Adjusted R-squared: 0.9875
F-statistic:   316 on 1 and 3 DF,  p-value: 0.0003882
```

推定された切片は $a = 0.26417$，回帰係数は $b = 0.10773$ である。また，このモデルはデータの分散の 98.75%

[*32] 回帰の外挿は薦められない。サンプルを希釈したり濃縮したりして吸光度を再測定し，標準希釈系列の範囲におさめることをお薦めする。

(0.9875) を説明していることが，Adjusted R-squared からわかる。また，p-value は，吸光度の分散がモデルによって説明される程度が誤差分散によって説明される程度と差が無いという帰無仮説の検定の有意確率である。



0.67 という吸光度に相当する濃度は，3.767084 となる。したがって，この溶液の濃度は，3.8 $\mu g/\ell$ だったと結論することができる。

EZR では，データはデータセットとして入力しなくてはいけない。「ファイル」「新しいデータセットを作成する」を選び，データセット名を入力：と書かれたテキストボックスに workingcurve と打って **[OK]** ボタンをクリックする。データエディタウィンドウが表示されたら，**[var1]** をクリックして，変数エディタの変数名というテキストボックスに y と打ち，型として **"numeric"** の方のラジオボタンをクリックしてから，キーボードの **Enter** キーを押す。次いで，同様にして **[var2]** を **[x]** に変える。それから，それぞれのセルに吸光度と濃度のデータを入力し，データエディタウィンドウを閉じる（通常は「ファイル」「閉じる」を選ぶ）。
散布図と回帰直線を描くには，「グラフ」「散布図」を選んで，**x** 変数として x を，**y** 変数として y を選び，**[OK]** ボタンをクリックする。
線形回帰モデルを当てはめるには，「統計解析」「連続変数の解析」「線形回帰（単回帰，重回帰）」と選び，目的変数として y，説明変数として x を選び，「モデル解析用に解析結果をアクティブモデルとして残す」の左のチェックボックスにチェックを入れて **[OK]** をクリックする。アウトプットウィンドウに結果が表示される（後述する多重共線性をチェックするために **VIF** を計算しようとしてエラーが表示されるが気にしなくて良い）。

検量線以外の状況でも，同じやり方で線形回帰モデルを当てはめることができる。survey データフレームに戻ってみよう[*33]。もし利き手の幅の分散を身長によって説明したいなら，線形回帰モデルを当てはめるには，R コンソールでは次のようにタイプすればいい。

```
res <- lm(Wr.Hnd ~ Height, data=survey)
summary(res)
```

EZR では，ロゴのすぐ右の **"データセット:"** の右側をクリックして survey を指定し，survey データセットをアクティブにしてから，「統計量」「モデルへの適合」「線形回帰」と選び，目的変数として Wr.Hnd，説明変数として Height を選び，「モデル解析用に解析結果をアクティブモデルとして残す」の左のチェックボックスにチェックを入れてから **[OK]** をクリックすると結果が得られる。

[*33] もちろん，survey データセットを使う前には，MASS パッケージをロードしておく必要がある。

8.4 推定された係数の安定性を検定する

回帰直線のパラメータ（回帰係数 b と切片 a）の推定値の安定性を評価するためには，t 値が使われる。いま，Y と X の関係が $Y = a_0 + b_0X + e$ というモデルで表されるとして，誤差項 e が平均 0，分散 σ^2 の正規分布に従うものとすれば，切片の推定値 a も，平均 a_0，分散 $(\sigma^2/n)(1 + M^2/V)$（ただし M と V は x の平均と分散）の正規分布に従い，残差平方和 Q を誤差分散 σ^2 で割った Q/σ^2 が自由度 $(n - 2)$ のカイ二乗分布に従うことから，

$$t_0(a_0) = \frac{\sqrt{n(n-2)}(a - a_0)}{\sqrt{(1 + M^2/V)Q}}$$

が自由度 $(n - 2)$ の t 分布に従うことになる。

しかしこの値は a_0 がわからないと計算できない。a_0 が 0 に近ければこの式で $a_0 = 0$ と置いた値（つまり $t_0(0)$。これを切片に関する t 値と呼ぶ）を観測データから計算した値が $t_0(a_0)$ とほぼ一致し，自由度 $(n - 2)$ の t 分布に従うはずなので，その絶対値は 95% の確率で t 分布の 97.5% 点（サンプルサイズが大きければ約 2 である）よりも小さくなる。つまり，データから計算された t 値がそれより大きければ，切片は 0 でない可能性が高いことになるし，t 分布の分布関数を使えば，「切片が 0 である」という帰無仮説に対する有意確率が計算できる。

回帰係数についても同様に，

$$t_0(b) = \frac{\sqrt{n(n-2)V}b}{\sqrt{Q}}$$

が自由度 $(n - 2)$ の t 分布に従うことを利用して，「回帰係数が 0」であるという帰無仮説に対する有意確率が計算できる。有意確率が充分小さければ，切片や回帰係数がゼロでない何かの値をとるといえるので，これらの推定値は安定していることになる。

R コンソールでも EZR でも，線形回帰をした結果の中の，`Pr(>|t|)` というカラムに，これらの有意確率が示されている。

9 回帰モデルを当てはめる際の留意点

身長と体重のように，どちらも誤差を含んでいる可能性がある測定値である場合には，一方を説明変数，他方を目的変数とすることは妥当でないかもしれない（一般には，身長によって体重が決まるなど方向性が仮定できれば，身長を説明変数にしてもよいことになっている）。また，最小二乗推定の説明から自明なように，回帰式の両辺を入れ替えた回帰直線は一致しない。従って，どちらを目的変数とみなし，どちらを説明変数とみなすか，因果関係の方向性に基づいて（先行研究や臨床的知見を参照し）きちんと決めるべきである。

回帰を使って予測をするとき，外挿には注意が必要である。とくに検量線は外挿してはいけない。実際に測った濃度より濃かったり薄かったりするサンプルに対して，同じ関係が成り立つという保証はどこにもないからである（吸光度を y とする場合は，濃度が高くなると分子の重なりが増えるので飽和 (saturate) してしまい，吸光度の相対的な上がり方が小さくなっていき，直線から外れていく）。サンプルを希釈したり濃縮したりして，検量線の範囲内で定量しなくてはいけない。

例題

組み込みデータ `airquality` は，1973 年 5 月 1 日から 9 月 30 日まで 154 日間のニューヨーク市の大気環境データである。含まれている変数は，`Ozone`（ppb 単位でのオゾン濃度），`Solar.R`（セントラルパークでの 8:00 から 12:00 までの 4000 から 7700 オングストロームの周波数帯の太陽放射の強さを Langley 単位で表した値），`Wind`（LaGuardia 空港での 7:00 から 10:00 までの平均風速，マイル／時），`Temp`（華氏での日最高気温），`Month`（月），`Day`（日）である。日照の強さを説明変数，オゾン濃度を目的変数として回帰分析せよ。

R コンソールでは，次の 4 行を打てば良い。

```
plot(Ozone ~ Solar.R, data=airquality)
res <- lm(Ozone ~ Solar.R, data=airquality)
abline(res)
```

```
summary(res)
```

EZR では，まず「ファイル」「パッケージに含まれるデータを読み込む」から左の枠の datasets をダブルクリックし，右の枠に現れるデータフーレムの下の方へスクロールして airquality をダブルクリックしてから **OK** ボタンをクリックして airquality データフレームをアクティブにする。次いで「グラフ」「散布図」を選び，**x** 変数を Solar.R，**y** 変数を Ozone として **[OK]** をクリックする。次に，「統計解析」「連続変数の解析」「線形回帰」を選ぶ。目的変数として Ozone を，説明変数として Solar.R を選んで **OK** ボタンをクリックする。

R コンソールでも EZR でも得られる結果は同じで，次の枠内の通りである。

```
Call:
lm(formula = Ozone ~ Solar.R, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-48.292 -21.361  -8.864  16.373 119.136

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.59873    6.74790   2.756 0.006856 **
Solar.R      0.12717    0.03278   3.880 0.000179 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 31.33 on 109 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-Squared: 0.1213,Adjusted R-squared: 0.1133
F-statistic: 15.05 on 1 and 109 DF,  p-value: 0.0001793
```

得られた回帰式は Ozone = 18.599 + 0.127 · Solar.R であり，最下行をみると F 検定の結果の p 値が 0.0001793 ときわめて小さいので，モデルの当てはまりは有意である。しかし，その上の行の `Adjusted R-squared` の値が 0.11 ということは，このモデルではオゾン濃度のばらつきの 10% 余りしか説明されないことになり，あまりいい回帰モデルではない。

10 文献

- 新谷歩 (2011) 今日から使える医療統計学講座【Lesson 3】サンプルサイズとパワー計算. 週刊医学界新聞，2937 号 (http://www.igaku-shoin.co.jp/paperDetail.do?id=PA02937_06)
- 青木繁伸 (2009) R による統計解析. オーム社
- 古川俊之 [監修]，丹後俊郎 [著] (1983) 医学への統計学. 朝倉書店.
- 中澤　港 (2003) R による統計解析の基礎. ピアソン・エデュケーション.
- 中澤　港 (2007) R による保健医療データ解析演習. ピアソン・エデュケーション.
- 神田善伸 (2012) EZR でやさしく学ぶ統計学〜EBM の実践から臨床研究まで〜，中外医学社[*34].

[*34] http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmed.html