# 大学院・医学基礎技術演習・実験基本技術（医学統計学）テキスト（1）

中澤　港（生態情報学　准教授）

2007 年 5 月 9 日

本演習は，実験や調査によって得られる生データからデータファイルを作成し，統計処理ソフトウェア **R** で解析し，結果を読み，レポートをまとめるという一連の流れを身に付けられるように，コンピュータを使って演習を行う。

問い合わせ先：生態情報学　准教授　中澤　港（e-mail: nminato@med.gunma-u.ac.jp）

## 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。MS Windows では，まだ 32 bit 環境でしか動作しない。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珍しくないが，Vine Linux などでは容易にインストールできるように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。

演習室のコンピュータには，やや古いバージョンだが R 本体と，それをメニューで操作するためのパッケージ Rcmdr がインストールされている。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するので，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。日本では会津大学[*1]，筑波大学[*2]，東京大学[*3]のどれかを利用すべきだろう。

Windows　CRAN ミラーから R-2.5.0 のインストール用ファイル[*4]をダウンロードし，ダブルクリックして実行し，適当に問いあわせに答えるだけでインストールは完了する。起動して日本語メッセージが文字化けしていたら，(1) いったん上部メニューバーの「編集」の「GUI プリファレンス」を開いて，表示フォントを Courier から MS ゴシックに変更して「反映」と「保存」をクリックするか，(2) 日本語表示用の設定が書かれた環境設定ファイルをコピーすれば[*5]直る。グラフィック画面での日本語表示まで考えれば後者をお勧めする。

Macintosh　最新版である R-2.5.0 に対応している OS は，Mac OS X 10.4 (Tiger) 以降である。同じく CRAN ミラーから `R-2.5.0.dmg` をダウンロードしてダブルクリックし，できたフォルダ内の `R.mpkg` をダブルクリックしてインストールすればよい。本学社会情報学部・青木繁伸教授のサイトに詳細な解説記事[*6]があるので参照されたい。

Linux　Debian，RedHat/Fedora Core，Vine 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-2.5.0.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

[*1] `ftp://ftp.u-aizu.ac.jp/pub/lang/R/CRAN/`
[*2] `http://cran.md.tsukuba.ac.jp/`
[*3] `http://ftp.ecc.u-tokyo.ac.jp/CRAN/`
[*4] `R-2.5.0-win32.exe`
[*5] `http://www.okada.jp.org/RWiki/?%C6%FC%CB%DC%B8%EC%B2%BD%B7%C7%BC%A8%C8%C4`（RjpWiki の日本語化掲示板）を参考にされたい。
[*6] `http://aoki2.si.gunma-u.ac.jp/R/begin.html`

よっては多少のパッチを当てる必要がある。

## 1.1 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，使えるデバイス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[*7]。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。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，「上向き矢印キー」で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 2000/XP 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。

## 1.2 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。

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

**導入** install.packages() 例えば，CRAN から vcd 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("vcd",dep=TRUE)
```

とする。dep=TRUE は dependency（依存）が真という意味で，vcd が依存している，vcd 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。

[*7] 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テキストファイル.Renviron を置き，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。

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。

ただし，本演習では，こうしたコマンドベースの使い方をせず，Rcmdr ライブラリを使ったメニュー操作を基本にする。Rcmdr のメニューを起動するには，プロンプトに対して以下を打てばよい。なお，もし GUI プリファレンスが MDI になっていて使いにくいときは，SDI にして保存し，R を起動しなおせばよい。

```
library(Rcmdr)
```

# 2　データ入力・記述統計・図示

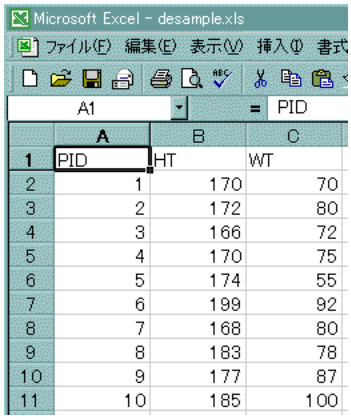
## 2.1　データ入力

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。

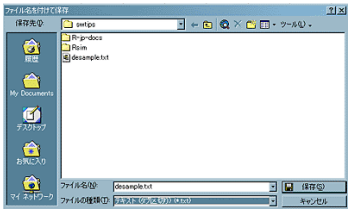
しかし実際にはもっとサイズの大きなデータについて，いろいろな分析を行う場合が多いので，データ入力と分析は別々に行うのが普通である。そのためには，同じ調査を繰り返しするとか，きわめて大きなデータであるとかでなければ，Microsoft Excel のような表計算ソフトで入力するのが手軽であろう。きわめて単純な例として，10 人の対象者についての身長と体重のデータが次の表のように得られているとする。

| 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。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。



あとは Rcmdr を使って，先ほど保存した `desample.txt` を読み込む。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，「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○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，「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，OK ボタンを押しただけでデータが読み込める。

なお，データ入力は，入力ミスを防ぐために，2 人以上の人が同じデータを入力し，それを比較するプログラムを実行して誤りをチェックする方法がよいとされる。しかし，現実には 2 人の入力者を確保するのが困難なため，1 人で 2 回入力して 2 人で入力する代わりにするか，あるいは 1 人で入力してプリントアウトした結果を元データと見比べてチェックするといった方法が使われることも多い。

## 2.2 欠損値の扱い

ここで注意しなければならないのは，欠損値の取扱いである。一般に，統計処理をする対象のデータは，母集団から標本抽出したサンプルについてのものである。サンプルデータを統計解析して，母集団についての情報を得るためには，そのサンプルが正しく母集団を代表していることが何より大切である。質問紙調査の場合でも，実験研究の場合でも，欠損値（質問紙なら無回答，非該当，わからない，等，実験研究なら検出限界以下，サンプル量不足，測定失敗等）をどのように扱うかによって，サンプルの代表性が歪められてしまうことがある。欠損が少なければあまり気にしなくていいが，たとえば，健診の際の食生活質問等で，「甘いものが好きですか」に対して無回答の人は，好きだけれどもそれが健康に悪いと判断されるだろうから答えたくない可能性があり，その人たちを分析から除くと，甘いもの好きの人の割合が，全体よりも少なめに偏った対象の分析になってしまう。なるべく欠損が少なくなるような努力をすべきだけれども，どうしても欠損のままに残ってしまった場合は，結果を解釈する際に注意する。

欠損値のコードは，通常，無回答 (NA) と非該当と不十分な回答が区別できる形でコーディングするが，ソフトウェアの上で欠損値を欠損値として認識させるためのコードは，分析に使うソフトウェアによって異なっているので（欠損値を表すコードの方を変更することも可能），それに合わせておくのも 1 つの方法である。デフォルトの欠損値記号は，R なら NA，SAS なら.（半角ピリオド）である。Excel ではブランク（何も入力しない）にしておくと欠損値として扱われる，入力段階で欠損値をブランクにしておくと，「入力し忘れたのか欠損値なのかが区別できない」という問題を生じるので，入力段階では決まった記号を入力しておいた方が良い。その上で，もし簡単な分析まで Excel でするなら，すべての入力が完了してから，検索置換機能を使って（Excel なら「編集」の「置換」。「完全に同一なセルだけを検索する」にチェックを入れておく），欠損値記号をブランクに変換すれば用は足りる。

次に問題になるのが，欠損値を含むデータをどう扱うかである。結果を解釈する上で一番紛れのない方法は，「1 つでも無回答項目があったケースは分析対象から外す」ということである[*8]（もちろん，非該当は欠損値ではあるが外してはならない）。その場合，統計ソフトに渡す前の段階で，そのケースのデータ全体（Excel 上の 1 行）を削除してしまうのが簡単である（もちろん，元データは別名で保存しておいて，コピー上で行削除）。質問紙調査の場合，たとえば 100 人を調査対象としてサンプリングして，調査できた人がそのうち 80 人で，無回答項目があった人が 5 人いたとすると，回収率 (recovery rate) は 80%（80/100）となり，有効回収率 (effective recovery rate) が 75%（75/100）となる。調査の信頼性を示す上で，これらの情報を明記することは重要である。目安としては有効回収率が 80% 程度は欲しい。

[*8] 最初からその方針ならば，1 つでも無回答項目があった人のデータは入力しないことに決めておく手もある。通常はそこまで思い切れないので，とりあえず入力は全部することが多い。

## 2.3 記述統計

記述統計はデータの特徴を把握する目的で用いる。しかし，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。Rcmdr からは，メニューバーの「統計量」の「要約」から「数値による要約」を選べばよい。

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 $\mu$（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[*9]。

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。

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[*10]，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

[*9] 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。

[*10] 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。

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[*11]。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。

**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，不偏標準偏差となる。もし分布が正規分布ならば，Mean±2SD[*12]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。

## 2.4 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。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ので，それをざっと示すことにする。

離散変数の場合は，以下のものが代表的である。

[*11] 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。

[*12] 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。

**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。Rcmdr では「グラフ」の「棒グラフ」を選ぶ。

**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 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 では描けない。

**円グラフ（ドーナツグラフ・パイチャート）**　円全体を 100 %として，各値の割合にしたがって中心から区切り線を引き，塗り分けた図である。ドーナツグラフでは２つの同心円にして，内側の円内を空白にする。R では `pie()` 関数を用いる。Rcmdr では「グラフ」の「円グラフ」を選ぶ。

連続変数の場合は，以下のものが代表的である。

**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。Rcmdr では「グラフ」の「ヒストグラム」を選ぶ。

**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。Rcmdr では「グラフ」の「QQ プロット」を選ぶ。

**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。Rcmdr では「グラフ」の「幹葉表示」を選ぶ。

**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。Rcmdr では「グラフ」の「箱ひげ図」を選ぶ。層別に描くこともできる。また，層別の平均と標準偏差を折れ線で結ぶことも「グラフ」の「平均のプロット」でできる。

**レーダーチャート**　複数の連続変数を中心点から放射状に数直線としてとり，データ点をつないで表される図である。それら複数の変数によって特徴付けられる性質のバランスをみるのに役立つ。１つのケースについて１つのレーダーチャートができるので，他のケースと比較するには，並べて描画するか，重ね描きする。R では `stars()` 関数を用いる。Rcmdr では描けない。

**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。Rcmdr では「グラフ」の「散布図」で描ける。

## 3　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

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[*13]。

２群の差がないという帰無仮説を検定する（つまり，差がないという帰無仮説の元で，現在得られている値以上に差がある値が偶然得られる確率＝有意確率＝が，偶然ではありえないくらい小さいかどうかを調べる）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
      i. ２群の間で分散に差がないという帰無仮説で $F$ 検定して仮説が棄却されない場合：$t$ 検定（R では `t.test(x,y,var.equal=T)`）
      ii. 仮説が棄却される場合：Welch の検定（R では `t.test(x,y)`）
   (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()`）。

これらの手法は，ほぼすべての初等統計の教科書に載っているが，簡単に説明しておく。

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。

## 3.1 母分散が既知で等しい $V$ である場合

$z_0 = |E(X) - E(Y)|/\sqrt{V/n_X + V/n_Y}$ が標準正規分布に従うことを使って検定する[*14]。

## 3.2 母分散が未知の場合

調査データを分析する場合は母分散が既知であることはほとんどなく，こちらが普通である。手順は以下の通り。

1. $F$ 検定（分散が等しいかどうか）：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。Rcmdr では「統計量」の「分散」から「分散の比の F 検定」を選ぶ。
2. 分散に差があるか差がないかによって，平均値が等しいかどうかの検定法は異なる（以下に詳述）。分散に差があるときは，その事実をもって別の母集団からとられた標本であると判断し，平均値が等しいかどうかを検定する意味はないとする考え方もあるが，一般には Welch の方法を使うか，ノンパラメトリックな方法を使って検定する。

## 3.3 分散に差がない場合

母分散 `S` を `S<-(DFX*SX+DFY*SY)/(DFX+DFY)` として推定し，

```
t0 <- abs(mean(X)-mean(Y))/sqrt(S/length(X)+S/length(Y))
```

が自由度 `DFX+DFY` の $t$ 分布に従うことから，帰無仮説「X と Y の平均値には差がない」を検定すると，`(1-pt(t0,DFX+DFY))*2` が有意確率となる。

---

[*13] 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 値関数の図示の方が遥かによい統計解析であると断言している。

[*14] 分布がひどく歪んでいる場合には，Mann-Whitney の U 検定（Wilcoxon の順位和検定と数学的に同値）を行う。後述するが，その場合は，代表値としても平均値と標準偏差でなく，中央値と四分位範囲または四分位偏差を表示するのが相応しい。

R では，`t.test(X,Y,var.equal=T)` とする。また，$F$ 検定のところで触れた量的変数と群分け変数という入力の仕方の場合は，`t.test(X~C,var.equal=T)` とする。ただしこれだと両側検定なので，片側検定したい場合は，`t.test(X,Y,var.equal=T,alternative="less")` などとする（`alternative="less"`は対立仮説が `X<Y` という意味なので，帰無仮説が `X>=Y` であることを意味する）。

## 3.4 分散が差がある場合（Welch の方法）

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

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を使うのが常道であるが[*15]，生データを図示する場合は `stripchart()` 関数を用いる。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))))
```

のように変換しておく必要がある。プロットするには次のように入力すればよい。

```
stripchart(X~C,method="jitter",vert=T)
MX <- tapply(X,C,mean); SX <- tapply(X,C,sd); IX <- c(1.1,2.1)
points(IX,MX,pch=18)
arrows(IX,MX-SX,IX,MX+SX,angle=90,code=3)
```

Rcmdr では，分散に差がある場合もない場合も「統計量」の「平均」の「独立サンプル t 検定」で検定できる。

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。
Rcmdr では「統計量」の「平均」の「対応のある t 検定」を選ぶ。

## 3.5 Wilcoxon の順位和検定

Wilcoxon の順位和検定は，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。Rcmdr では，「統計量」の「ノンパラメトリック検定」を選んで実行する。

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。以下，Wilcoxon の順位和検定の手順を箇条書きする。

[*15] 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。

1. 変数 $X$ のデータを $x_1, x_2, ..., x_m$ とし，変数 $Y$ のデータを $y_1, y_2, ..., y_n$ とする。
2. まず，これらをまぜこぜにして小さい方から順に番号をつける[*16]。例えば，$x_8[1], y_2[2], y_{17}[3], ..., x_4[N]$ のようになる（但し $N = m + n$）。
3. ここで問題にしたいのは，それぞれの変数の順位の合計がいくつになるかということである。ただし，順位の総合計は $(N + 1)N/2$ に決まっているので，片方の変数だけ考えれば残りは引き算でわかる。そこで，変数 $X$ だけ考えることにする。
4. $X$ に属する $x_i$（$i = 1, 2, ..., m$）の順位を $R_i$ と書くと，$X$ の順位の合計は

$$R_X = \sum_{i=1}^{m} R_i$$

となる。$R_X$ があまり大きすぎたり小さすぎたりすると，$X$ の分布と $Y$ の分布に差がないという帰無仮説 $H0$ が疑わしいと判断されるわけである。では，帰無仮説が成り立つ場合に，$R_X$ はどのくらいの値になるのだろうか？[*17]

5. もし $X$ と $Y$ に差がなければ，$X$ は $N$ 個のサンプルから偶然によって $m$ 個取り出したものであり，$Y$ がその残りである，と考えることができる。順位についてみると，$1, 2, 3, ..., N$ の順位から $m$ 個の数値を取り出すことになる。同順位がなければ，ありうる組み合わせは，${}_NC_m$ 通りある[*18]。
6. $X > Y$ の場合には，${}_NC_m$ 通りのうち，合計順位が $R_X$ と等しいかより大きい場合の数を $k$ とする（$X < Y$ の場合は，合計順位が $R_X$ と等しいかより小さい場合の数を $k$ とする）。
7. $k/{}_NC_m$ が有意水準 $\alpha$ より小さいときに $H0$ を疑う。$N$ が小さいときは有意になりにくいが，$N$ が大きすぎると計算が大変面倒である[*19]。そこで，正規近似を行う（つまり，期待値と分散を求めて，統計量から期待値を引いて分散の平方根で割った値が標準正規分布に近似的に従うという関係を用いて検定する）。
8. 帰無仮説 $H_0$ のもとでは，期待値は

$$E(R) = \sum_{i=1}^{m} E(R_i) = m(1 + 2 + ... + N)/N = m(N + 1)/2$$

（1 から $N$ までの値を等確率 $1/N$ でとるから）。分散はちょっと面倒で，

$$var(R) = E(R^2) - (E(R)^2)$$

から，

$$E(R^2) = E((\sum_{i=1}^{m} R_i)^2) = \sum_{i=1}^{m} E(R_i^2) + 2\sum_{i<j} E(R_i R_j)$$

となるので[*20]，

$$E(R_i^2) = (1^2 + 2^2 + ... + N^2)/N = (N + 1)(2N + 1)/6$$

---

[*16] 同順位がある場合の扱いは後述する。

[*17] 以下説明するように，順位和 $R$ をそのまま検定統計量として用いるのが Wilcoxon の順位和検定であり，$R_X, R_Y$ の代わりに，$U_X = mn + n(n + 1)/2 - R_Y$，$U_Y = mn + m(m + 1)/2 - R_X$ として，$U_X$ と $U_Y$ の小さいほうを $U$ として検定統計量として用いるのが，Mann-Whitney の U 検定である。また，$U_X - U_Y$ を検定統計量とするのが Kendall の S 検定である。有意確率を求めるために参照する表が異なる（つまり帰無仮説の下で検定統計量が従う分布の平均と分散は，これら 3 つですべて異なる）が，数学的には等価な検定である。R では，Wilcoxon の順位和統計量の分布関数が提供されているので，例えばここで得られた順位和を RS と書くことにすると，`2*(1-pwilcox(RS,m,n))` で両側検定の正確な有意確率が得られる。

[*18] R では `choose(N,m)` によって得られる。

[*19] もっとも，今ではコンピュータにやらせればよい。例えば R であれば，`wilcox.test(X,Y,exact=T)` とすれば，サンプル数の合計が 50 未満で同順位の値がなければ，総当りして正確な確率を計算してくれる。が，つい 15 年くらいまではコンピュータは誰もが使える道具ではなかったし，総当りをするには計算時間がかかりすぎた。今のコンピュータでもサンプルサイズが大きいと，総当りでは計算時間がかかりすぎて実用的でない。

[*20] 第 1 項が対角成分，第 2 項がそれ以外に相当する。$m = 2$ の場合を考えてやればわかるが，

$$E((\sum_{i=1}^{2} R_i)^2) = E((R_1 + R_2)^2) = E(R_1^2 + R_2^2 + 2R_1R_2) = \sum_{i=1}^{2} E(R_i^2) + 2\sum_{i<j} E(R_i R_j)$$

となる。

と

$$E(R_iR_j) = \frac{1}{N(N-1)}\{(\sum_{k=1}^{N} k)^2 - \sum_{k=1}^{N} k\}$$

$$= \frac{1}{N(N-1)}(\frac{N^2(N+1)^2}{4} - \frac{N(N+1)(2N+1)}{6})$$

$$= \frac{(N+1)(3N+2)}{12}$$

を代入して整理すると，結局，$var(R_X) = m(N+1)(N-m)/12 = mn(N+1)/12$ となる。

9. 標準化[*21]して連続修正[*22]し，$z_0 = \{|R_X - E(R_X)| - 1/2\}/\sqrt{var(R_X)}$ を求める。$m$ と $n$ が共に大きければこの値が標準正規分布に従うので，例えば $z_0 > 1.96$ ならば，両側検定で有意水準 5% で有意である。R で有意確率を求めるには，$z_0$ を `z0` と書けば，`2*(1-pnorm(z0,0,1))` とすればよい。
10. ただし，同順位があった場合は，ステップ 2 の「小さい方から順に番号をつける」ところで困ってしまう。例えば，変数 $X$ が $\{2,6,3,5\}$，変数 $Y$ が $\{4,7,3,1\}$ であるような場合には，$X$ にも $Y$ にも 3 という値が含まれる。こういう場合は，下表のように平均順位を両方に与えることで，とりあえず解決できる。

| 属する変数 | $Y$ | $X$ | $X$ | $Y$ | $Y$ | $X$ | $X$ | $Y$ |
|---|---|---|---|---|---|---|---|---|
| 値 | 1 | 2 | 3 | 3 | 4 | 5 | 6 | 7 |
| 順位 | 1 | 2 | 3.5 | 3.5 | 5 | 6 | 7 | 8 |

11. ただし，このやり方では，正規近似をする場合に分散が変わる[*23]。帰無仮説の下で，$E(R_X) = m(N+1)/2$ はステップ 8 と同じだが，分散が

$$var(R_X) = mn(N+1)/12 - mn/\{12N(N-1)\} \cdot \sum_{t=1}^{T}(d_t^3 - d_t)$$

となる。ここで $T$ は同順位が存在する値の総数であり，$d_t$ は $t$ 番目の同順位のところにいくつのデータが重なっているかを示す。上の例では，$T = 1$，$d_1 = 2$ となる。なお，あまりに同順位のものが多い場合は，この程度の補正では追いつかないので，値の大小があるクロス集計表として分析することも考慮すべきである（例えば Cochran-Armitage 検定などが考えられる）。

## 3.6 2 群の母比率の差の検定

たとえば，患者群 $n_1$ 名と対照群 $n_2$ 名の間で，ある特性をもつ者の人数がそれぞれ $r_1$ 名と $r_2$ 名だったとして，その特性の母比率に差がないという帰無仮説を考える。

2 群の母比率 $p_1, p_2$ が，各々の標本比率 $\hat{p_1} = r_1/n_1, \hat{p_2} = r_2/n_2$ として推定されるとき，それらの差を考える。差 $(\hat{p_1} - \hat{p_2})$ の平均値と分散は，$E(\hat{p_1} - \hat{p_2}) = p_1 - p_2, V(\hat{p_1} - \hat{p_2}) = p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$ となる。2 つの母比率に差が無いならば，$p_1 = p_2 = p$ とおけるはずなので，$V(\hat{p_1} - \hat{p_2}) = p(1-p)(1/n_1 + 1/n_2)$ となる。この $p$ の推定値として，$\hat{p} = (r_1 + r_2)/(n_1 + n_2)$ を使い，$\hat{q} = 1 - \hat{p}$ とおけば，$n_1p_1$ と $n_2p_2$ がともに 5 より大きければ，標準化して正規近似を使い，

$$Z = \frac{\hat{p_1} - \hat{p_2} - E(\hat{p_1} - \hat{p_2})}{\sqrt{V(\hat{p_1} - \hat{p_2})}} = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}} \sim N(0,1)$$

によって[*24]検定できる。

---

[*21] 何度も出てくるが，平均（期待値）を引いて分散の平方根で割る操作である。

[*22] これも何度も出てくるが，連続分布に近づけるために 1/2 を引く操作である。

[*23] 正確な確率を求めることができれば問題ないけれども，同順位がある場合には，R では正確な確率は求められない。

[*24] この $Z$ は離散値しかとれないため，連続分布である正規分布による近似の精度を上げるために，連続性の補正と呼ばれる操作を加え，かつ $p_1 > p_2$ の場合（つまり $Z > 0$ の場合）と $p_1 < p_2$ の場合 (つまり $Z < 0$ の場合）と両方考える必要があり，正規分布の対称性から絶対値をとって $Z > 0$ の場合だけ考え，有意確率を 2 倍する。即ち，

$$Z = \frac{|\hat{p_1} - \hat{p_2}| - (1/n_1 + 1/n_2)/2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}$$

として，この $Z$ の値が標準正規分布の 97.5% 点（R ならば `qnorm(0.975,0,1)`）より大きければ有意水準 5% で帰無仮説を棄却する。

数値計算をしてみるため，仮に，患者群 100 名と対照群 100 名で，喫煙者がそれぞれ 40 名，20 名だったとする。喫煙率に 2 群間で差がないという帰無仮説を検定するには，

```
p <- (40+20)/(100+100)
q <- 1-p
Z <- (abs(40/100-20/100)-(1/100+1/100)/2)/sqrt(p*q*(1/100+1/100))
2*(1-pnorm(Z))
```

より，有意確率が約 0.0034 となるので，有意水準 5% で帰無仮説は棄却される。つまり，喫煙率に 2 群間で差がないとはいえないことになる。

差の 95% 信頼区間を求めるには，サンプルサイズが大きければ正規分布を仮定できるので，原則どおりに差から分散の平方根の 1.96 倍を引いた値を下限，足した値を上限とすればよい。この例では，

```
dif <- 40/100-20/100
vardif <- 40/100*(1-40/100)/100+20/100*(1-20/100)/100
difL <- dif - qnorm(0.975)*sqrt(vardif)
difU <- dif + qnorm(0.975)*sqrt(vardif)
cat("喫煙率の差の点推定値=",dif,"  95% 信頼区間= [",difL,",",difU,"]\n")
```

より，[0.076,0.324] となる。しかし，通常は連続性の補正を行うので，下限からはさらに $(1/n_1 + 1/n_2)/2 = (1/100 + 1/100)/2 = 0.01$ を引き，上限には同じ値を加えて，95% 信頼区間は [0.066,0.334] となる。

R には，こうした比率の差を検定するための関数 `prop.test()` が用意されており，以下のように簡単に実行することができる。

```
smoker <- c(40,20)
pop <- c(100,100)
prop.test(smoker,pop)
```

母比率の推定と，その差があるかどうかの検定[*25]，差の 95% 信頼区間を一気に出力してくれる。Rcmdr では，「統計量」の「比率」を選ぶ。

3 群以上の母比率の差の検定

`prop.test()` 関数は，3 群以上の間でも，「どの群でも事象の生起確率に差がない」という帰無仮説を検定するのに使える。その帰無仮説が棄却されるときに，どの群間で差があるのかをみるには，検定の多重性（後述）が生じるので，平均値の差の場合と同様，第一種の過誤を調整する必要があり，ボンフェローニの方法やホルムの方法を用いることができる。R の関数は `pairwise.prop.test()` である。なお，3 群以上の間で事象の生起確率に一定の傾向がみられるかどうかを調べたい場合には，コクラン＝アーミテージの検定という手法がある。例えば，漁師 100 人，農民 80 人，事務職 30 人について便の検査をして，日本住血吸虫虫卵陽性者が 60 人，30 人，8 人だったとしたとき，職業的な貝との接触リスクに対して勝手に漁師を 4，農民を 2，事務職を 1 とスコアリングして，陽性割合の増加傾向が，このスコアと同じかどうかを調べることができる。この場合なら，R のコマンドは以下のようになる。Rcmdr には組み込まれていない。

```
total <- c(100,80,30)
epos <- c(60,30,8)
orisk <- c(4,2,1)
prop.trend.test(epos,total,orisk)
```

[*25] 連続性の補正済み，事象が生起しない場合についても考慮してカイ二乗適合度検定をしているのだが，この操作は次回説明する 2 つの変数の独立性のカイ二乗検定と数学的に等価である。

# 4　分散分析と多重比較

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[*26]。

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。

## 4.1　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）[*27]。

| 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 でデータを読み込んでから，`summary(aov(HEIGHT ~ VG))` とすれば（Rcmdr の場合は，メニューバーの Statistics から Means の One-Way ANOVA を選ぶ），例えば次のような結果が得られる。

---

[*26] なお，分散分析は本来，その効果をみるための実験計画をした上で実施するものだから，群ごとのサンプルサイズは揃っているべきだし，効果の有無を効率よく検出するのに適したサンプルサイズが設計されているべきだが，現実には実験計画されていないデータにも適用されている。適切なサンプルサイズは，母集団の均質性，サブグループ数，母集団のパラメータ推定に求めたい正確さ，注目している現象の出現頻度，予算などで変わってくる。詳しくは，永田靖 (2003) サンプルサイズの決め方，朝倉書店を参照されたいが，有意水準 5%，検出力 90% の場合なら，以下の式によって求めるのが基本となる。

- 2 つの集団の平均値の差を調べる場合：予測される標本平均が m1, m2，標本分散が d1, d2 なら，サンプルサイズは

$$\frac{(1.96+1.28)^2(d1+d2)}{(m1-m2)^2}$$

- 2 つの集団の罹患率の差を調べる場合：2 つの集団で予測される罹患率がそれぞれ r1，r2 なら，サンプルサイズは

$$\frac{(1.96+1.28)^2(r1+r2)}{(r1-r2)^2}$$

- 2 つの集団の比率の差を調べる場合：期待される比率を p1，p2 とすると，サンプルサイズは，

$$\frac{\{1.28\sqrt{p1(1-p1)+p2(1-p2)}+1.96\sqrt{(p1+p2)(1-(p1+p2)/2)}\}^2}{(p1-p2)^2}$$

[*27] `http://phi.med.gunma-u.ac.jp/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
```

このような結果の表を分散分析表という。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。

## 4.2 クラスカル＝ウォリス (Kruskal-Wallis) の検定

一元配置の分散分析は，各群が等しい母分散をもつ正規分布に従うことを仮定して行っているパラメトリックな分析法なので，データの分布がひどく歪んでいる場合は結果がおかしくなる場合がある[*28]。そこで，多群間の差を調べるためにもノンパラメトリックな方法がある。クラスカル＝ウォリス (Kruskal-Wallis) の検定と呼ばれる方法である。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。

[*28] 各群の母分散が等しいかどうかを調べる検定法として，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。同じ目的のノンパラメトリックな方法として，Fligner-Killeen の検定という方法もあり，fligner.test(Y~C) で実行できる。また，量的変数について，母集団で正規分布しているかどうかを調べる方法としては，既に説明したヒストグラムや正規確率プロットなどのグラフ表示による方法の他に，シャピロ＝ウィルク (Shapiro-Wilk) の検定と呼ばれる方法もある。詳しくは説明しないが，R では shapiro.test(Y) で実行できる。厳密に言えば，これらの検定で等分散性と分布の正規性が確認されない限り，一元配置分散分析の結果を解釈するには注意が必要なのだが，論文や本でもそこまで考慮されずに使われていることが多い。

- $H$ 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。

## 4.3 検定の多重性の調整

仮に，上述の南太平洋の島の 3 つの村での健診の例で，一元配置分散分析か Kruskal-Wallis の検定で有意差があったときに，具体的にどの村の間に有意差があるのかを調べるには，単純に考えると，$t$ 検定[*29]や順位和検定[*30]を繰り返せば良さそうである。この方法が使われている本や論文もないわけではない。しかし，3 つの村でこれをやると 3 つから 2 つを取り出す全ての組み合わせについて検定するので，3 回の比較をすることになり，個々の検定について有意水準を 5% にすると，全体としての第 1 種の過誤は明らかに 5% より大きくなる。もし村が 7 つあったら，7 つから 2 つを取り出す組み合わせは 21 通りあるので，1 つくらいは偶然によって有意差が出てしまう比較があっても全然おかしくない。したがって，先に述べた通り，$t$ 検定の繰り返しは第 1 種の過誤が大きくなってしまって不都合である。これに似た方法として無制約 LSD（最小有意差）法や Fisher の制約つき LSD 法（一元配置分散分析を行って有意だった場合にのみ LSD 法を行うという方法）があるが，これらも第 1 種の過誤を適切に調整できない（ただし制約つきの場合は 3 群なら大丈夫）ことがわかっているので，使ってはいけない。現在では，この問題は広く知られているので，$t$ 検定の繰り返しや LSD 法で分析しても論文は accept されない。

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，シェフェ (Scheffé) の方法，ダンカン (Duncan) 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。しかしこの中で，ダンカンの方法は，新多範囲検定などと呼ばれた時期もあったが，数学的に間違っていることがわかっているので，使ってはいけない。ボンフェローニの方法とシェフェの方法も検出力が悪いので，特別な場合を除いては使わない方が良い。せめてテューキーの HSD を使うべきである。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[*31]。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。

上記いくつかの方法が良く使われている原因は，用途が限定されているダネットとウィリアムズを除けば，たんにそれらが歴史的に古く考案され，昔の統計学の教科書にも説明されているからに過ぎない。現在では，かなり広い用途をもち，ノンパラメトリックな分析にも適応可能なホルム (Holm) の方法（ボンフェローニの方法を改良して開発された方法）が第一に考慮されるべきである。その上で，全ての群間の比較をしたい場合はペリ (Peritz) の方法，対照群との比較をしたいならダネットの逐次棄却型検定（これはステップダウン法と呼ばれる方法の 1 つであり，既に触れたダネットの方法とは別）も考慮すればよい。とはいえ，ソフトウェアによってはこれらの方法をサポートしていない場合もあると思われ，その場合はテューキーの HSD を使うべきである（もちろん場合によっては，ダネットかウィリアムズを使い分けねばならない）[*32]。

多重比較においては，帰無仮説が単純ではない。例えば，4 群間の差を調べるとしよう。一元配置分散分析での帰無仮説は，$\mu_1 = \mu_2 = \mu_3 = \mu_4$ である。これを包括的帰無仮説と呼び，$H_{\{1,2,3,4\}}$ と書くことにする。さて第 1 群から第 4 群までの母平均 $\mu_1 \sim \mu_4$ の間で等号関係が成り立つ場合をすべて書き上げてみると，$H_{\{1,2,3,4\}} : \mu_1 = \mu_2 = \mu_3 = \mu_4$，$H_{\{1,2,3\}} : \mu_1 = \mu_2 = \mu_3$，$H_{\{1,2,4\}} : \mu_1 = \mu_2 = \mu_4$，$H_{\{1,3,4\}} : \mu_1 = \mu_3 = \mu_4$，$H_{\{2,3,4\}} : \mu_2 = \mu_3 = \mu_4$，$H_{\{1,2\},\{3,4\}} : \mu_1 = \mu_2$ かつ $\mu_3 = \mu_4$，$H_{\{1,3\},\{2,4\}} : \mu_1 = \mu_3$ かつ $\mu_2 = \mu_4$，$H_{\{1,4\},\{2,3\}} : \mu_1 = \mu_4$ かつ $\mu_2 = \mu_3$，$H_{\{1,2\}} : \mu_1 = \mu_2$，$H_{\{1,3\}} : \mu_1 = \mu_3$，$H_{\{1,4\}} : \mu_1 = \mu_4$，$H_{\{2,3\}} : \mu_2 = \mu_3$，$H_{\{2,4\}} : \mu_2 = \mu_4$，$H_{\{3,4\}} : \mu_3 = \mu_4$ の 14 通りである。このうち，$H_{\{1,2,3,4\}}$ 以外のものを部分帰無仮説と呼ぶ。すべての 2 つの群の組み合わせについて差を調べるということは，$\{H_{\{1,2\}}, H_{\{1,3\}}, H_{\{1,4\}}, H_{\{2,3\}}, H_{\{2,4\}}, H_{\{3,4\}}\}$ が，考慮すべき部分帰無仮説の集合と

[*29] R では `t.test(height[vg=="X"],height[vg=="Y"])` など。

[*30] R では `wilcox.test(height[vg=="X"],height[vg=="Y"])` など。

[*31] 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。

[*32] もっとも，オープンソースで多くのコンピュータで無料で使える R がホルムの方法をデフォルトとしている現実を考えれば，そういう言い訳はもはや通用しない。

なる。一方，例えば第 1 群が対照群であって，他の群のそれぞれが第 1 群と差があるかどうかを調べたい場合は，$\{H_{\{1,2\}}, H_{\{1,3\}}, H_{\{1,4\}}$ が考慮すべき帰無仮説の集合となる。これらの集合をその多重比較における「帰無仮説族」と呼ぶ。

ここで多重比較の目的を「帰無仮説族」というコトバを使って言い換えてみる。個々の帰無仮説で有意水準を 5% にしてしまうと，帰無仮説族に含まれる帰無仮説のどれか 1 つが誤って棄却されてしまう確率が 5% より大きくなってしまう。それではまずいので，その確率が 5% 以下になるようにするために，何らかの調整を必要とするわけで，この調整をする方法が多重比較なのである。つまり，帰無仮説族の有意水準を定める（例えば 5% にする）ことが，多重比較の目的である[*33]。

R では，`pairwise.t.test(HEIGHT,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 $t$ 検定した結果を出してくれる[*34]。

また，`pairwise.wilcox.test(HEIGHT,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。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))` などとして，テューキーの HSD を行うことも可能である。また，CRAN (`http://cran.r-project.org/`) から `multcomp` パッケージをインストールすることによって，`simtest(HEIGHT ~ VG, type="Dunnett")` あるいは `simtest(HEIGHT ~ VG, type="Williams")` としてダネットやウィリアムズの方法を使うことも可能である（ただし，この例でこれらの方法を使うことは不適切である）。

Rcmdr の場合なら，「統計量」の「平均」から「一元配置分散分析」を選んで実行するときに，Pairwise comparisons of means に左にチェックを入れておけば，自動的に Tukey の HSD で検定の多重性を調整してくれる。

## 5 文献


- 大橋靖雄，浜田知久馬 (1995) 生存時間解析: SAS による生物統計. 東京大学出版会.
- 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.
- 永田　靖 (2003) サンプルサイズの決め方. 朝倉書店.


[*33] このことからわかるように，差のなさそうな群をわざと入れておいて帰無仮説族を棄却されにくくしたり，事後的に帰無仮説を追加したりすることは，統計を悪用していることになり，やってはいけない。

[*34] ただし，$t$ 検定とは言っても，`pool.sd=F` というオプションをつけない限りは，$t_0$ を計算するときに全体の誤差分散を使うので，ただの $t$ 検定の繰り返しとは違う。