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

中澤　港（公衆衛生学　准教授）

2011 年 5 月 25 日，6 月 1 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。

目次

問い合わせ先：群馬大学大学院医学系研究科公衆衛生学分野・准教授　中澤　港
e-mail: nminato@med.gunma-u.ac.jp

2010 年 5 月 26 日：第 0.5 版（途中まで）
2010 年 8 月 17 日：第 1.0 版（一応完成）
2010 年 11 月 8 日：第 1.1 版（生存時間解析の説明を追加し，何箇所か修正）
2010 年 11 月 10 日：第 1.1.1 版（何箇所か typo を修正）
2010 年 11 月 16 日：第 1.1.2 版（英語版と合わせた）
2010 年 11 月 22 日：第 1.1.3 版（ロジスティック回帰分析の例題中の変数名の誤訳を訂正）
2011 年 5 月 20 日：第 1.1.4 版（R のバージョン更新，multiple imputation について追記）
2011 年 6 月 1 日：第 1.1.4.1 版（subset の例示の修正）

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。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，R 本体[*1]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[*2]か兵庫教育大学[*3]のどちらかを利用すべきだろう。

1.1 R のインストール方法 [2011 年 5 月 20 日現在]

Windows　CRAN ミラーから R-2.13.0 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する。最近の Windows 版バイナリではインストーラに問題があって，インストールに使う言語として日本語を選ぶと途中からダイアログが文字化けするので，インストールには Japanese でなく English を選ぶことをお薦めする（インストールを英語でやっても，R を起動すると，ちゃんと日本語環境で使える）。English を選んでリターンキーを押すと，[Setup - R for Windows 2.13.0] というウィンドウが起動するので，[Next] というボタンをクリックし，ライセンス表示にも [Next] をクリックし，インストール先は通常デフォルトのまま `C:\Program Files\R\R-2.13.0` で [Next] をクリックする。次のインストールするコンポーネントを選ぶウィンドウでは，通常はデフォルトの User installation のままでも問題ないが，よほどディスク容量に余裕がないのでなければ，Full installation にすることをお薦めする。[Next] ボタンをクリックすると，スタートアップオプションをカスタマイズするか (Do you want to customize the startup options?) と尋ねるウィンドウが表示されるので，ここは**デフォルトの No (accept defaults) でなく，Yes (customized startup) の方をマークして** [Next] をクリックすることをお薦めする[*5]。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** [Next] をクリックする。次の Help Style はどちらを選んでも良いが，筆者は Plain text の方が好みである。[Next] をクリックすると，インターネット接続を標準設定にするか，Internet Explorer のプロキシ設定に合わせるかを聞いてくるので，以前の群馬大学のようにプロキシサーバを介さなくては学外のサイトが見えない状況の場合は Internet2 を選ぶ方が便利だが，通常はどちらでも問題ないはずである。[Next] をクリックするとスタートメニューフォルダ名を聞いてくるが，ここはデフォルトのまま R で問題ない。次のウィンドウではデスクトップアイコンを作り，R のバージョン番号をレジストリに記録し，.RData という拡張子をもつファイルを R に関連付けするというオプションがデフォルト指定されているが，通常はそのままで問題ないだろう[*6]。次に [Next] をクリックするとインストールが始まる。暫く待つとセットアップウィザードが完了したという意味のウィンドウが表示されるので，[Finish] をクリックする。以上の操作で R 本体のインストールは終わりである。デスクトップアイコンをダブルクリックするかクイック起動メニューのアイコンをクリックして R を起動した際に，もし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

[*1] 2011 年 5 月 20 日現在の最新版は 2.13.0 だが，演習室のコンピュータの R のバージョンはまだ 2.11.1 である。

[*2] `http://cran.md.tsukuba.ac.jp/`

[*3] `http://essrc.hyogo-u.ac.jp/cran/`

[*4] `R-2.13.0-win.exe`

[*5] 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 R Commander が非常に使いにくくなるからである。

[*6] ただし筆者はデスクトップのアイコンを増やしたくないので，Create a desktop icon のチェックは外し，代わりにその下の，Create a Quick Launch icon にチェックを入れている。

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[*7]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-2.13.0 に対応している OS は，Mac OS X 10.5 (Leopard) 以降である。同じく CRAN ミラーから `R-2.13.0.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）。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[*8]があるので参照されたい。

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-2.13.0.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[*9]。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[*10]。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[*11]。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 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。

[*7] `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。

[*8] `http://aoki2.si.gunma-u.ac.jp/R/begin.html`

[*9] 演習室のコンピュータでは，通常は，スタートメニューの中にしか起動アイコンがないので，それをデスクトップにコピーするとよい。

[*10] 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 `LANG=C LC_ALL=C` 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。

[*11] \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。

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。

```
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[*12]，

```
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 R Commander を使う

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 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。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。

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。

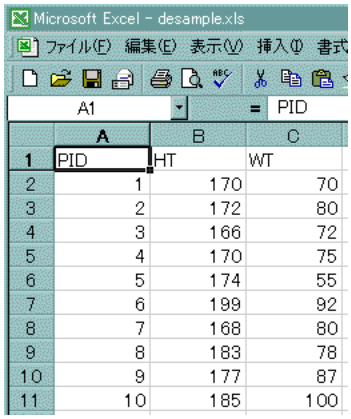
ごく小さな規模のデータについて単純な分析だけ行う場合，電卓で計算してもよいし，分析する手続きの中で直接数値を入れてしまってもよい。例えば，60 kg, 66 kg, 75 kg という 3 人の平均体重を R を使って求めるには，プロンプトに対して mean(c(60,66,75)) または (60+66+75)/3 と打てばいい。

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

[*12] 演習室のコンピュータは Windows Vista なので管理者権限がないとライブラリのインストールができないし，他にもいろいろな制約がある。

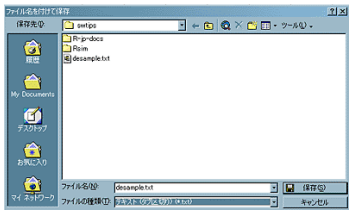
ば，Microsoft Excel のような表計算ソフトで入力するのが手軽であろう。きわめて単純な例として，10 人の対象者についての身長と体重のデータが次の表のように得られているとする。

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，「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，「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。
[おまけ情報] 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，「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。

2.2 入力ミスを防ぐためのデータ入力の原則

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。Excel のワークシートが 2 枚できたときに，それらを比較するには，1 つのブックの Sheet1 と Sheet2 にそれらを貼り付けておき，Sheet3 の一番左上のセル（A1）に，

```
=If(Sheet1!A1=Sheet2!A1,"","X")
```

と入力し，これをコピーして，Sheet3 上の全範囲（Sheet1 と Sheet2 に参照されているデータがある範囲）に貼り付けると，誤りがあるセルにのみ"X"という文字が表示される。元データを参照して Sheet1 と Sheet2 の不一致部分をすべて正しく直し終われば，Sheet3 が見かけ上空白になるはずである。

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。

2.3 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[*13]，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

[*13] 欠損値を表すコードの方を変更することも可能。例えば R では read.delim() などデータファイルを読み込む関数の中で，例えば na.string="-99"とオプション指定すれば，データファイル中の-99 を欠損値として変換しながら読み込んでくれる

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[*14]（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。

もう少し厳密に考えると，上述のごとくランダムでない欠損は補正のしようがないが，欠損がランダムな場合でも 2 通りの状況を分けて考える必要がある。即ち，MISSING COMPLETELY AT RANDOM (MCAR) の場合は単純に除去しても検出力が落ちるだけでバイアスはかからないが，MISSING AT RANDOM (MAR)，つまり欠測となった人とそうでない人の間でその変数の分布には差が無いが他の変数の分布に差がある場合には，単純に除去してしまうとバイアスがかかるのである。そのため，multiple imputation という欠損値を補う方法がいろいろ開発されている。R では，mitools[*15]と mice[*16]という 2 つのパッケージがあり，後者のメインテナは Dr. Stef van Buuren というオランダの方で，実は Multiple Imputation Online というサイト[*17]の Head をされている専門家なので，mice を使えばできると思われるが，やや複雑な話なので，本実習では扱わない[*18]。

2.4 記述統計

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。

[*14] 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。

[*15] `http://cran.r-project.org/web/packages/mitools/index.html`

[*16] `http://cran.r-project.org/web/packages/mice/index.html`

[*17] `http://www.multiple-imputation.com/`

[*18] 例えば，欠損のあるデータフレーム名を withmiss とすると，`library(mice)` の後で，`imp <- mice(withmiss)` とすると，元データや multiple imputation による欠損値推定の係数群が `imp` というオブジェクトに保存される。multiple imputation の方法は"sample"，"pmm"，"logreg"，"norm"，"lda"，"mean"，"polr"などから選べて，`mice()` 関数の `meth`=オプションで指定できる。欠損値が補完されたデータフレームを得るには，`est <- complete(imp, 2)` などとする。デフォルトでは 5 組の係数群が推定されるので，この例で指定した 2 により，そのうち 2 番目の係数群を使って推定されるデータフレームが得られる。あとは，このデータフレームを使って解析した複数の結果をまとめる必要がある。

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[*19]。

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[*20]，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。

[*19] 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。

[*20] 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。

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[*21]。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）[*22]。もし分布が正規分布ならば，Mean±2SD[*23]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。

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。さきほど読みこんだ身長・体重のデータで実行してみよう。

2.5 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。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。

例を示そう。Rcmdr で離散変数をグラフにする場合は，変数が要因型でなくてはいけないので，扱うデータを，MASS ライブラリに含まれている survey というデータに変えてみる。まず，R コマンダーのメニューの [ツール] の [パッケージのロード] を選んで表示されるウィンドウの中で，MASS を選ぶ。次に [データ] の [パッケージ内のデータ] の [アタッチされたパッケージからデータセットを読み込む] を選び，表示されるウィンドウの左の枠で MASS をダブルクリックし，次に右の枠で survey をダブルクリックし，[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。

[*21] 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。

[*22] 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。

[*23] 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。

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。

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 では上記 survey データをアクティブにした状態で，「グラフ」の「棒グラフ」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 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。survey データをアクティブにした状態で，変数として `Smoke` を選ぶと，喫煙習慣ごとの人数の割合に応じて円が分割された扇形に塗り分けられたグラフができる。**

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。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 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）。**

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。

> **Rcmdr では「グラフ」の「QQ プロット」を選ぶ。survey データでは，変数として `Age` を選ぶと，まったく正規分布でないので直線状でないし，`Pulse` を選ぶと，やや歪んでいるけれども概ね直線に乗るので正規分布に近いことがわかる。**

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。

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。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。

Rcmdr 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，変数として Pulse を選び，[層別のプロット] というボタンをクリックして表示されるウィンドウで，層別変数として Smoke を選んで [OK] ボタンをクリックしてから，戻ったウィンドウで再び [OK] をクリックすればいい。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「平均のプロット」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。

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() 関数を用いればできないことはないが，stars() で描かれるレーダーチャートは一般的なものと異なる。普通のレーダーチャートを描くには，plotrix ライブラリか fmsb ライブラリをインストールする必要がある。どちらも CRAN のミラーサイトからダウンロードしてインストールできる。install.packages("plotrix") と install.packages("fmsb") とすればよい。その上で，例えば後者の場合なら，library(fmsb) としてから example(radarchart) とすれば使い方がわかる。**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。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。

Rcmdr では「グラフ」の「散布図」で描ける。層別にマークを変えることもできる。survey データで年齢と身長の関係を，男女でマークを変えて見たいときは，x 変数として Age，y 変数として Height を選び，「平滑線」のチェックを外し，[層別のプロット] ボタンをクリックして層別変数として Sex を選び，「層別に線を描く」にチェックを入れ，戻った画面で [OK] をクリックすれば描画される。できあがった散布図の点をマウスでクリックし，値を確認したい時は，x 変数や y 変数を指定するウィンドウで，「点を確認する」にチェックを入れておく。確認したい点の上で左クリックするとレコード番号が表示され，右クリックするまで繰り返すことができる。

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。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[*24]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[*25]：Welch の検定（R では `t.test(x,y)`）[*26]
   （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()`）。

3.1　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。

[*24] 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。

[*25] `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。

[*26] 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，最近まで，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，本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た

Rcmdr では「統計量」の「分散」から「分散の比の F 検定」を選び，グループ (Group variable) として C を，目的変数 (Response variable) として X を選ぶ。ただし，グループ変数は要因型になっていないと候補として表示されないので，もし 0/1 で入力されていたら，予め「データ」の「アクティブデータセット内の変数の操作」で「数値変数を因子に変換」を用いて要因型にしておく（字面は 0/1 のままでも OK）。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，グループとして Sex を，目的変数として Height を選んで [OK] ボタンをクリックする。デフォルトでは両側検定されるが，仮説によっては片側検定をすることもあり，その場合は「対立仮説」の下のラジオボタンのチェックを変えればいい。男女それぞれの分散と，検定結果が「出力ウィンドウ」に表示される。

3.2 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。

Rcmdr では「統計量」の「平均値」の「独立サンプル t 検定」を選んで，グループ (Group variable) として C を，目的変数 (Response variable) として X を選んで，等分散と考えますか？　というラジオボタンは「No」にしておき，両側検定か片側検定かを選んでから [OK] ボタンをクリックする。ただし，グループ変数は要因型になっていないと候補として表示されないので，もし 0/1 で入力されていたら，予め「データ」の「アクティブデータセット内の変数の管理」で「数値変数を因子に変換」を用いて要因型にしておく（字面は 0/1 のままでも OK。具体的には下の例題を参照）。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，グループとして Sex を，目的変数として Height を選べばよい。Welch の方法による 2 標本 t 検定の結果が「出力ウィンドウ」に表示される。

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を使うのが常道であるが[*27]，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))))
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。プロットするには次のように入力すればよい（注：この手順は Rcmdr のメニューには入っていない）。

[*27] 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。

```
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)
```

3.3 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"))
```

Rcmdr では「統計量」の「平均」の「対応のある 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，出力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。

例題

Rcmdr のメニューで「データ」の「パッケージ内のデータ」の「アタッチされたパッケージからデータセットを読み込む」を選び，左側から datasets パッケージをダブルクリックし，右側から infert データフレームをダブルクリックすると，Trichopoulos *et al.* (1976) Induced abortion ans secondary infertility. *Br J Obst Gynaec*, 83: 645-650. で使われているデータを読み込むことができる。
アテネ大学の第一産婦人科を受診した続発性の不妊の 100 人の女性の 1 人ずつについて同じ病院から年齢，既往出生児数，教育歴をマッチングした健康な（不妊でない）女性 2 人ずつを対照として選ぶことを目指してサンプリングし，2 人の対照が見つかった不妊患者が 83 人だったので，この患者と対照全員を含むデータである（ただし 74 組目だけ対照が 1 人しかデータに含まれていないので，249 人でなく 248 人のデータとなっている。除かれたのはそれまでの自然流産と人工妊娠中絶が 2 回ずつあった人である）。
含まれている変数は以下の通りである。

education: **教育を受けた年数（3 水準の要因型）**
age: **年齢**
parity: **既往出生児数**
induced: **それまでの人工妊娠中絶回数（2 は 2 回以上）**
case: **不妊の女性が 1，対照が 0**
spontaneous: **それまでの自然流産回数（2 は 2 回以上）**
stratum: **マッチングした組の番号**
pooled.stratum: **プールした層番号**

(1) 不妊患者と対照の間で自然流産を経験した数に差がないという帰無仮説を検定せよ。(2) 各女性の自然流産の経験数と人工妊娠中絶の経験数に差がないという帰無仮説を検定せよ。有意水準はともに 5% とする。

本当は因果を逆に考えてロジスティック回帰またはポアソン回帰する方が筋がいいと思うが，ここでは敢えて平均値の差の検定をしてみる。2 群間で分布が異なるし対照群では正規分布から明らかに外れているが，それにも目をつぶって平均値の差の検定を行う。2 回以上というのを 2 回と扱っていいのかという点にも問題があるが，ここでは目をつぶる。R コンソールでは，この操作は単純である。必要な *t* 検定をするには，次のように打てば良い。

(1) `t.test(spontaneous ~ case, data=infert)`[*28]
(2) `t.test(infert$induced, infert$spontaneous, paired=TRUE)`

[*28] もし患者群と対照群の間で分散が等しいという帰無仮説を検定したいときは，`var.test(spontaneous ~ case, data=infert)` と打つ。

Rcmdr で群別に分布をみるには，群分け変数が要因型でなくてはならないので，まず「データ」の「アクティブデータセット内の変数の管理」の「数値変数を因子に変換」で `case` **を因子型に変えておく。変数として** `case` **を選び，因子水準は「水準名を指定」がチェックされた状態にして，新しい変数または複数の変数に対する接頭文字列のところが<変数と同じ>となっているのを** `group` **として（ここは，複数の数値型変数を一度に因子型に変換するときは接頭文字列を入力するが，1 つだけの場合は新しい変数名全体を打つ必要がある），因子型の変数名が** `group` **となるように指定する。すると水準名を指定するウィンドウが開くので，0 のところに** `control`**，1 のところに** `infertile` **と打つ。そうやって準備をしておいてから，「グラフ」の「箱ひげ図」で「変数（１つを選択）」として** `spontaneous` **を選び，「層別のプロット」ボタンをクリックして「層別変数（１つ選択）」として** `group` **を選ぶと，対照群と不妊群別々に箱ひげ図を描くことができる。値が 0, 1, 2 しかないので箱ひげ図よりも棒グラフあるいはヒストグラムの方がわかりやすいが，棒グラフやヒストグラムは Rcmdr では層別でプロットできないため，ここでは箱ひげ図を採用した。もちろん「平均のプロット」で標準偏差をエラーバーとする平均値を線で結んだプロットをさせてもよい。**

(1)「統計量」の「平均」の「独立サンプル t 検定」を選び，「等分散と考えますか？」で「No」にチェックが入っていることを確認し，グループを `group`**，目的変数を** `spontaneous` **にして [OK] をクリックすると Welch の方法による** t **検定が実行できる（なお，「統計量」の「分散」の「分散の比の F 検定」でグループを** `group`**，目的変数を** `spontaneous` **として両側検定を実行すると出力ウィンドウに表示される** `p-value` **が小さいので，2 群の分散にも統計的な有意差があることがわかる）。**

(2)「統計量」の「平均」の「対応のある t **検定」を指定し，第１の変数として** `spontaneous`**，第２の変数として** `induced` **を選んで [OK] ボタンをクリックすれば実行できる。**

3.4 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。

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[*29]。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？[*30]

[*29] 同順位がある場合の扱いは後述する。

[*30] 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

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[*31]。
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 α 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[*32]。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，

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

となるので[*33]，

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

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[*34]して連続修正[*35]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

定の正確な有意確率が得られる。

[*31] R では `choose(N,m)` によって得られる。

[*32] 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。

[*33] 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，

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

となる。

[*34] 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。

[*35] 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。

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[*36]。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）。

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。

Rcmdr では「統計量」の「ノンパラメトリック検定」で「2 標本ウィルコクソン検定」を選び，グループ変数として `Sex` を選び，応答変数として `Height` を選んで [OK] ボタンをクリックする。

3.5 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。

Rcmdr では，「統計量」，「ノンパラメトリック検定」，「対応のあるウィルコクソン検定」と選択し，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。

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，$n_1 p_1$ と $n_2 p_2$ がともに 5 より大きければ，標準化して正規近似を使い，

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

[*36] 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。

によって[*37]検定できる。

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，2 群間でその差がないという帰無仮説の検定[*38]，差の 95% 信頼区間を一気に出力してくれる。survey データフレームで「利き手が左である割合に男女で差が無い」という帰無仮説を検定するには，`prop.test(table(survey$Sex,survey$W.Hnd))` とすれば良い。

Rcmdr では，「統計量」の「比率」から「2 標本の比率の検定」を選ぶ。survey データフレームで，「利き手が左である割合に男女で差がない」という帰無仮説を検定するには，グループとして `Sex`，目的変数として `W.Hnd` を指定し，検定のタイプとして「連続修正を用いた正規近似」にチェックを入れて [OK] ボタンをクリックすればよい。

4　3 群以上の位置母数の差の検定

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

[*37] 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。

[*38] 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，カイ二乗適合度検定をしているので，後述する 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。

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[*39]。

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）[*40]。

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://phi.med.gunma-u.ac.jp/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。

[*39] 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，

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

[*40] 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
```

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。

ただし，一元配置分散分析は，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。

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[*41]。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。

Rcmdr では，メニューバーの「統計量」から「平均」の「１元配置分散分析」を選ぶ。バートレットの検定は，「統計量」の「分散」から選べる。シャピロ＝ウィルクの検定は，「統計量」の「要約」から「シャピロ・ウィルクの正規性の検定」を選ぶ。

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。等分散でない場合，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場合もあり，R では oneway.test() で実行できる。これは Rcmdr に組み込まれていないし，あまり一般的に使われてはいないが，本学社会情報学部の青木繁伸教授の数値実験によると，等分散性の検定をせず，常に oneway.test() を用いる方がよいことが既知である。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る。

4.2 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

[*41] 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。

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。

Rcmdr では，「統計量」，「ノンパラメトリック検定」，「クラスカルーウォリスの検定...」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。**現在のところ，Rcmdr のメニューには入っていない。**

4.3 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。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[*42]。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やホルムの方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やホルムの方法による検定の多重性の調整ができる。

例えば，上の例で，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，すべての 2 村落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[*43]。

[*42] 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。

[*43] `"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。

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。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, 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。

Rcmdr の場合は，「統計量」の「平均」から「１元配置分散分析」を選んで実行するときに，「2 組ずつの平均の比較（多重比較）」の左のボックス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た対比較を実行してくれるし，同時信頼区間のグラフも表示される。しかし，第一種の過誤を調整する方法は，まだサポートされていない。

4.4 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。**残念ながら，Rcmdr ではまだサポートされていない。**

```
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。

5 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。

２群ずつ比べて，どの群間で差があるのかをみようとすると，平均値の場合と同様に検定の多重性が生じるので，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になる。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。

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

Rcmdr では「統計量」「比率」メニューには１標本と２標本の場合しかないので指定できない。しかし，実は３群以上で「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ことは，後述するクロス集計表の考え方をすれば，「群分け変数と事象の有無を示す変数が独立」という帰無仮説の検定と同じことなので，「統計量」の「分割表」の「２元表」で行の変数，列の変数として，群分け変数と事象の有無を示す変数をそれぞれ指定すれば可能である。集計済みのデータの場合も，「統計量」「分割表」「２元表の入力と分析」を選び，この例なら行数を 2 のまま，列数を 3 にし，表に数値を入力して，[OK] ボタンをクリックすれば，検定ができる（ただし，`prop.test()` と異なり，Yates の連続性の修正を行うオプションは提供されていない点には注意が必要である）。なお，`pairwise.prop.test()` や `prop.trend.test()` は，Rcmdr ではまだサポートされていない。

生データから計算する場合について，MASS ライブラリの survey データセットを使って例示しよう。Clap という変数は，両手を叩き合わせたときにどちらが上に来るかを意味し，左，右，どちらでもない，という３つのカテゴリからなる。W.Hnd は字を書く手がどちらか，つまり利き手を意味する。両手を叩き合わせたときに上に来る手の３タイプ間で，左利きの割合に差が無いという帰無仮説を検定する。次いで，３タイプ中のすべての２タイプの組み合わせについて，左利きの割合に差が無いという帰無仮説を検定し，第一種の過誤をホルムの方法で調整してみる。R コンソールでは次の２行を打つだけで済む。

```
prop.test(table(survey$Clap, survey$W.Hnd))
pairwise.prop.test(table(survey$Clap, survey$W.Hnd), p.adjust.method="holm")
```

Rcmdr では，「統計量」「分割表」「２元表」を選び，行の変数として Clap を選び，列の変数として W.Hnd を選んで [OK] ボタンをクリックすると，カイ二乗検定の結果が表示される。検定の多重性の調整は Rcmdr ではサポートされていない。

6　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。

Rcmdr では，「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，“平滑線” の右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。

6.1 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。

6.2 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[*44]，擬似相関 (spurious correlation) であったり[*45]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)
```

Rcmdr では，「統計量」の「要約」の「相関の検定」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）。相関のタイプとして「ピアソンの積率相関」と「スピアマンの順位」と「ケンドールのタウ」が選べるようになっている。この例題ではピアソンの積率相関係数を求めるので，初期設定のまま「ピアソンの積率相関」にしておけばよい。検定についても「対立仮説」の下に「両側」「相関＜０」「相関＞０」の

[*44] 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。

[*45] 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。

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
0.600991
```

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[*46]，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。

男女別に相関係数の検定を実行するには，いろいろなやり方があるが，最も単純に考えれば，データセットそのものを男女別の部分集合に分け，それぞれについて分析すればよい。R コンソールでは次の４行を打つ（その前に，MASS ライブラリをメモリに読み込んでおかねばならないのは当然である）。

```
males <- subset(survey, Sex=="Male")
cor.test(males$Height, males$Wr.Hnd)
females <- subset(survey, Sex=="Female")
cor.test(females$Height, females$Wr.Hnd)
```

Rcmdr では，「データ」の「アクティブデータセット」の「アクティブデータセットの部分集合を抽出」を選び，表示されるウィンドウで，「すべての変数を含む」はチェックが入ったまま，「部分集合の表現」のボックスに Sex=="Male" **と入力し，「新しいデータセットの名前」に** Males **（既にある名前と重複しなければ何でもよい）と入力して [OK] ボタンをクリックすると，男性だけのデータフレーム** Males **ができてアクティブになる。ここで先ほどと同じ「統計量」「要約」「相関の検定」をすれば男性の身長と利き手の大きさについてピアソンの積率相関係数を求めて有意性の検定をすることができる。**
女性について同じことをするには，まず「データ」の「アクティブデータセット」の「アクティブデータセットの選択」で survey **を選び直し，「アクティブデータセットの部分集合を抽出」の「部分集合の表現」で** Sex=="Female"**，「新しいデータセットの名前」で** Females **として OK ボタンをクリックしてから「統計量」→「要約」→「相関の検定」を実行すればよい。**

[*46] 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。

順位相関係数の定義

なお，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。

[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。

6.3 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[*47]。

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。

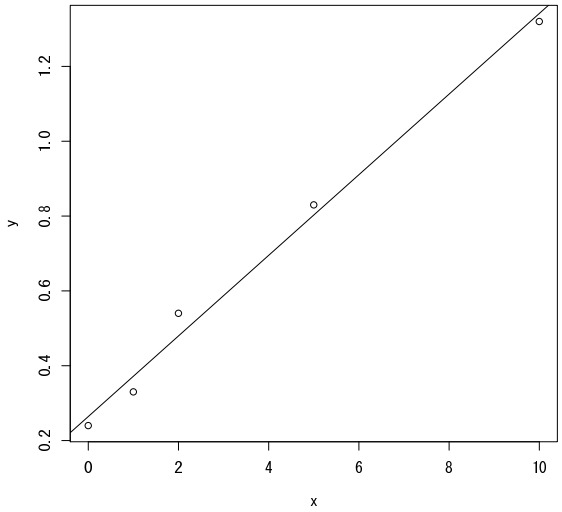
```
y <- c(0.24, 0.33, 0.54, 0.83, 1.32)
x <- c(0, 1, 2, 5, 10)
# 線形回帰モデルを当てはめる
res <- lm(y ~ x)
# 詳しい結果表示
summary(res)
# 散布図と回帰直線を表示する
plot(y ~ x)
abline(res)
# 吸光度 0.67 に対応する濃度を計算する
(0.67 - res$coef[1])/res$coef[2]
```

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% (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。

[*47] 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。



0.67 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。

Rcmdr 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。「データ」「新しいデータセット...」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，平滑線と書かれているチェックボックスのチェックを外してから [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，「統計量」「モデルへの適合」「線形回帰」と選び，目的変数として y**，説明変数として** x **を選ぶ。[OK] をクリックすると，アウトプットウィンドウに結果が表示される。**

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[*48]。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)
```

Rcmdr 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` **を選ぶ。その後 [OK] をクリックすると結果が得られる。**

6.4 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，回帰係数の推定値 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，

[*48] 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。

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 Rcmdr 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。

7 回帰モデルの応用

7.1 重回帰モデル

説明変数は 2 つ以上の変数を含むことができる。このような場合，モデルは「重回帰モデル」と呼ばれる。注意しなくてはならない点がいくつかあるが，基本的には線形モデルの右側に + でつないで説明変数群を与えるだけである。

例えば，これまで扱ってきた survey データで，利き手の大きさの分散を説明するために，身長のみならず，利き手でない方の手の大きさも使うことにしよう。R コンソールでは次のように打てばよい（もちろん，予め MASS ライブラリをロードしておかねばならない）。

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

Rcmdr では，「統計量」「モデルへの適合」「線形回帰」を選び，“目的変数” として `Wr.Hnd` をクリックし，“説明変数” として `Height` をクリックしてからキーボードの Ctrl キーを押しながら `NW.Hnd` もクリックする。その後，[OK] ボタンをクリックすると，結果がアウトプットウィンドウに示される。

重回帰モデルでは，個々の説明変数について推定される回帰係数は，他の説明変数の目的変数への影響を調整した上で，その変数独自の目的変数への影響を示す「偏回帰係数」である。しかし偏回帰係数の値は，各変数の絶対的な大きさに依存しているので，各説明変数の目的変数への影響の相対的な強さを示すものにはならない。そうした比較をしたければ，R コンソールで次のようにタイプして `stb` として得られる「標準化偏回帰係数」が利用できる。

```
sdd <- c(0,sd(res$model$Height),sd(res$model$NW.Hnd))
stb <- coef(res)*sdd/sd(res$model$Wr.Hnd)
stb
```

Rcmdr には，メニューアイテムとしては，この機能は提供されていないが，スクリプトウィンドウに表示されているコマンドを編集し，必要な範囲が選択された状態で [Submit] というボタンをクリックすれば，結果がアウトプットウィンドウに表示される。

7.2 当てはまりの良さの評価

データから得た回帰直線は，完璧にデータに乗ることはない。そこで，回帰直線の当てはまりのよさを評価する。a と b が決まったとして，$z_i = a + bx_i$ とおいたとき，$e_i = y_i - z_i$ を残差 (residual) と呼ぶ。残差は，y_i のばらつきのうち，回帰直線では説明できなかった残りに該当する。つまり，残差が大きいほど，回帰直線の当てはまりは悪いと考えられる。残差にはプラスもマイナスもあるので二乗和をとり，次の式で得られる「残差平方和」Q を定義することができる。

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - z_i)^2$$

$$= \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n - \frac{(n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}/n$$

残差平方和 Q は回帰直線の当てはまりの悪さを示す尺度であり，それを n で割った Q/n を残差分散という。残差分散（$\mathrm{var}(e)$ と書くことにする）と Y の分散 $\mathrm{var}(Y)$ とピアソンの相関係数 r の間には，

$$\mathrm{var}(e) = \mathrm{var}(Y)(1 - r^2)$$

という関係が常に成り立つので，

$$r^2 = 1 - \mathrm{var}(e)/\mathrm{var}(Y)$$

となる。このことから r^2 が 1 に近いほど回帰直線の当てはまりがよいことになる。その意味で，r^2 を「決定係数」と呼ぶ。また，決定係数は，Y のばらつきがどの程度 X のばらつきによって説明されるかを意味するので，X の「寄与率」と呼ぶこともある。

データによっては，何通りもの回帰直線の残差平方和が大差ないという状況がありうる。例えば，目的変数と説明変数が実はまったく無関係であった場合は，データの重心を通るどのような傾きの線を引いても残差平方和はほとんど同じになってしまう。言い換えれば，傾きや切片の推定値が不安定になる。

r^2 は説明変数が多ければ大きくなるので，通常は自由度で r^2 を調整した「自由度調整済み重相関係数の二乗」を決定係数と考える。この値は，R コンソールでも Rcmdr でも線形モデルの当てはめ結果の中で，`Adjusted R-Squared` として表示されている。

当てはまりの良さの別の尺度として，AIC（赤池の情報量基準：Akaike information criterion）も良く用いられる。とくに重回帰モデルでは，AIC も表示するのが普通である。R には `AIC()` という関数があり，線形回帰モデルの結果を付値したオブジェクトを，この関数に渡せば AIC が計算される（例えば `AIC(res)` のように使う）。ここでは AIC について詳しくは説明しないが，たくさんのオンライン資料や書籍で説明されている。例えば Wikipedia[*49]にも丁寧な説明がある。

7.3 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。

[*49] http://ja.wikipedia.org/wiki/%E8%B5%A4%E6%B1%A0%E6%83%85%E5%A0%B1%E9%87%8F%E8%A6%8F%E6%BA%96

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)
```

Rcmdr では，まず「データ」「パッケージ内のデータ」「アタッチされたパッケージからデータを読み込む...」を選んでから，datasets **パッケージ，**airquality **データフレームをそれぞれダブルクリックし，**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，平滑線の隣のチェックボックスのチェックを外して [OK] をクリックする。次に，「統計量」の「モデルへの適合」の「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 Rcmdr 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。

当てはまりを改善するには，説明変数を追加することが有効な場合がある。この例では，Wind あるいは Temp を説明変数に加えて重回帰モデルにすれば，当てはまりが改善する。R コンソールでは，次の 3 行を打てば重回帰モデルの当てはめができる。自由度調整済み重相関係数の二乗が約 60% にまで改善していることがわかる。

```
mres <- lm(Ozone ~ Solar.R + Wind + Temp, data=airquality)
summary(mres)
AIC(mres)
```

Rcmdr では，「統計量」「モデルへの適合」「線形モデル」を選び，左側のテキストボックスに `Ozone` **とタイプし，右側のテキストボックスに** `Solar.R + Wind + Temp` **とタイプして [OK] をクリックすると，同じ結果が得られる（AIC は自動的には出てこないが）。**

7.4 共分散分析 (ANACOVA/ANCOVA)

複数のグループがあって，どのグループに属するサンプルについても，同じ説明変数と目的変数が調べられているとき，それらの関係がグループによって異なるかどうか調べたい場合がある。共分散分析は，このような場合に用いられる。

典型的なモデルは，

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{C} + \beta_2 \mathrm{X} + \beta_{12} \mathrm{C} \cdot \mathrm{X} + \varepsilon$$

となる。ここで，**C** は 2 値変数，**X** と **Y** は量的な変数（連続変数）である。**C** の 2 群間で **Y** の平均値に差があるかどうかを比べたいのだが，**Y** が **X** と相関がある場合に（このとき **X** を共変量と呼ぶ），**X** と **Y** の回帰直線の傾き (slope) が **C** の 2 群間で差がないときに，**X** による影響を調整した **Y** の修正平均（adjusted mean; 調整平均ともいう）を，**C** の 2 群間で比べる。修正平均は **C** の各変数についての係数（2 群の場合，基準にする変数の係数はゼロ）に，共変量の平均に共変量の係数を掛けたものを加え，さらに切片を加えることによって計算できる。

ただし，2 本の回帰直線がともに充分な説明力をもっていて，かつ 2 本の回帰直線の間で傾きに差がない場合でないと，修正平均の比較には意味がない。そもそも回帰直線の説明力が低ければその変数は共変量として考慮する必要がないし，傾きが違っていれば群分け変数と独立変数の交互作用が従属変数に関して有意に影響しているということなので，2 群を層別して別々に解釈する方が良い。

いま，C で群分けされる 2 つの母集団における，(X, Y) の間の母回帰直線を，$y = \alpha_1 + \beta_1 x$，$y = \alpha_2 + \beta_2 x$ とすれば，共分散分析は次の手順で進める。

(1) **傾きに差がないという帰無仮説の検定**　$H_0 : \beta_1 = \beta_2$，$H_1 : \beta_1 \neq \beta_2$ を検定する。各群について，X と Y の平均と変動と共変動を出しておけば[*50]，仮説 H_1 のもとでの残差平方和

$$d_1 = SS_{Y1} - (SS_{XY1})^2/SS_{X1} + SS_{Y2} - (SS_{XY2})^2/SS_{X2}$$

と仮説 H_0 のもとでの残差平方和

$$d_2 = SS_{Y1} + SS_{Y2} - (SS_{XY1} + SS_{XY2})^2/(SS_{X1} + SS_{X2})$$

を計算して $F = (d_2 - d_1)/(d_1/(N-4))$ が H_0 のもとで第 1 自由度 1，第 2 自由度 $N-4$ の F 分布に従うことを使って傾きが等しいかどうかの検定ができる。

(2) **傾きに差がないとき，y 切片に差がない帰無仮説の検定**　$\beta_1 = \beta_2$ のもとで（即ち，共通の傾き β を，$\beta = (SS_{XY1} + SS_{XY2})/(SS_{X1} + SS_{X2})$ として推定し），$H_0' : \alpha_1 = \alpha_2$，$H_1' : \alpha_1 \neq \alpha_2$ を検定する。帰無仮説 H_0' のもとで全部のデータを使った残差平方和 $d_3 = SS_Y - (SS_{XY})^2/SS_X$ を計算して，$F = (d_3 - d_2)/(d_2/(N-3))$ が第 1 自由度 1，第 2 自由度 $N-3$ の F 分布に従うことを使って検定できる。H_0' が棄却された場合は各群の平均を共通の傾きに代入すれば各群の切片が求められるし，棄却されない場合は 2 群を一緒にして普通の単回帰分析をすることになる。

(3) **傾きに有意差があるとき，層別解析**　$\beta_1 = SS_{XY1}/SS_{X1}$，$\beta_2 = SS_{XY2}/SS_{X2}$ として別々に傾きを推定し，y 切片 α もそれぞれの式に各群の平均値を入れて計算する。

[*50] サンプルサイズ $N1$ の第 1 群に属する x_i, y_i について，$E_{X1} = \sum x_i/N1$，$SS_{X1} = \sum (x_i - E_{X1})^2$，$E_{Y1} = \sum y_i/N1$，$SS_{Y1} = \sum (y_i - E_{Y1})^2$，$E_{XY1} = \sum x_i y_i/N1$，$SS_{XY1} = \sum (x_i y_i - E_{XY1})^2$。第 2 群も同様。

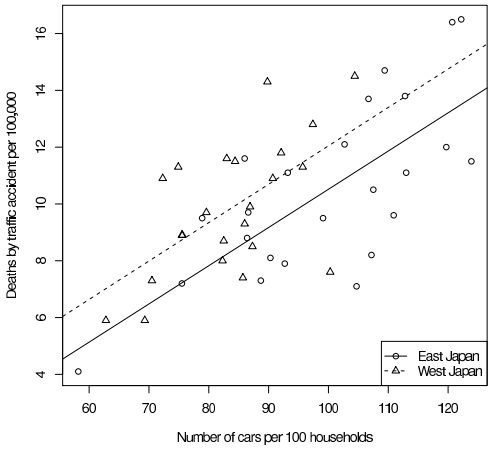
例題

http://phi.med.gunma-u.ac.jp/grad/sample3.dat は，都道府県別のタブ区切りテキストデータファイルである。変数としては，都道府県名 (PREF)，日本の東西（REGION），1990 年の 100 世帯当たり乗用車台数（CAR1990)，1989 年の人口 10 万人当たり交通事故死者数 (TA1989)，1985 年の国勢調査による人口集中地区居住割合 (DIDP1985) が含まれている（REGION の 1 は東日本，2 は西日本を意味する）。
このデータについて，東日本と西日本で，100 世帯当たり乗用車台数で調整した人口 10 万人当たり交通事故死者数に差があるか，共分散分析によって検討せよ[a]。

[a] （注）実は乗用車台数の影響を調整しなければ人口当たり交通事故死者数は東西で有意な差はない。

R コンソールに打ち込むコマンドは以下の通りである。

```
sample3 <- read.delim("http://phi.med.gunma-u.ac.jp/grad/sample3.dat")
plot(TA1989 ~ CAR1990, pch=as.integer(REGION), data=sample3,
 xlab="Number of cars per 100 households", ylab="Deaths by traffic accident per 100,000")
east <- subset(sample3,REGION=="East")
regeast <- lm(TA1989 ~ CAR1990, data=east)
summary(regeast)
west <- subset(sample3,REGION=="West")
regwest <- lm(TA1989 ~ CAR1990, data=west)
summary(regwest)
abline(regeast,lty=1)
abline(regwest,lty=2)
legend("bottomright",pch=1:2,lty=1:2,legend=c("East Japan","West Japan"))
summary(lm(TA1989 ~ REGION*CAR1990, data=sample3))
anacova <- lm(TA1989 ~ REGION+CAR1990, data=sample3)
summary(anacova)
cfs <- dummy.coef(anacova)
cfs[[1]] + cfs$CAR1990 * mean(sample3$CAR1990) + cfs$REGION
```



最後のモデルの REGION の係数がゼロと差が無いという帰無仮説の検定の p 値は，0.0319 である。このことから，CAR1990 の影響を調整した上でも TA1989 には，東日本と西日本の間に，有意水準 5% で統計学的に有意な差があると言える。最後の 2 行によって，東日本，西日本それぞれの，修正平均値は，次のように表示される。

```
    East     West
 9.44460 10.96650
```

Rcmdr では，まずデータセット名 sample3 としてインターネットからデータを読み込むため，「データ」「データのインポート」「テキストファイル，クリップボードまたは URL から読み込み...」を選び，“データセット名を入力” の右にあるテキストボックスに sample3 と打ち，“インターネット URL” の隣のラジオボタンをチェックし，“フィールド区切り” の「タブ」の隣のラジオボタンをチェックし，[OK] ボタンをクリックする。表示されるウィンドウで http://phi.med.gunma-u.ac.jp/grad/sample3.dat と打ち，[OK] をクリックすると，インターネットからデータが読み込まれ，sample3 という名前のデータフレームがアクティブになる。
次に REGION で層別した散布図を描く。「x 変数」を CAR1990，「y 変数」を TA1989 とし，“平滑線” の隣のチェックを外して [OK] し，「グループ別にプロット」をクリックして REGION を選んで [OK] し，元のウィンドウでも [OK] すると，散布図と 2 本の回帰直線が描かれ，2 本の回帰直線はほぼ平行に見える（丁寧にやるには，ここで東西日本別々に層別して，CAR1990 によって TA1989 が説明されるかをみるため，単回帰分析を行う。東日本のサブセットと西日本のサブセットを作って分析すればよい。CAR1990 の係数は東西どちらでも有意にゼロと異なる。したがって，その影響を調整することに意味はあると思われることが確認できる）。
次に，傾きに差があるかを解析する。「統計量」「モデルへの適合」「線形モデル」でモデル名 LinearModel.3 として左辺の目的変数として TA1989 を，右辺の説明変数群として REGION+CAR1990+REGION:CAR1990 を指定する。結果をみると，REGIONWest:CAR1990 の行に示されている交互作用効果の p 値は 0.990 である。この値は，2 本の回帰直線の傾きに統計学的な有意差がないことを意味する。
そこで今度は，乗用車所有台数で調整した交通事故死者数の修正平均に差があるかどうかをみるため，交互作用項を除いて回帰を行う。再び「統計量」「モデルへの適合」「線形モデル」で，モデル名を LinearModel.4 とし，左辺はそのまま TA1989 で，右辺の説明変数を CAR1990+REGION に変えて線形モデルの当てはめを実行する。この結果，REGIONWest の行の p 値は 0.0319 なので，REGION という変数は，有意水準 5% で，CAR1990 の影響を調整しても TA1989 に対して統計学的に有意な影響をもっていることが示された。ただし，修正平均は R コンソールと同じコマンドをスクリプトウィンドウに打って選択し，「Submit」ボタンをクリックしなくては計算できない。単純な平均値は東日本が 10.5，西日本が 9.87 であるが，乗用車保有台数の影響を調整した修正平均は，東日本が 9.44，西日本が 11.0 と逆転し，かつ有意水準 5% で統計学的な有意差があるといえた。

7.5 ロジスティック回帰分析

ロジスティック回帰分析は，従属変数（目的変数。ロジスティック回帰分析では反応変数，あるいは応答変数と呼ぶこともある）が 2 値変数であり，二項分布に従うので `lm()` ではなく，`glm()` を使う。

ロジスティック回帰分析の思想としては，例えば疾病の有無を，複数のカテゴリ変数によって表される要因の有無と年齢のような交絡因子によって説明するモデルをデータに当てはめようとする。量的な変数によって表される交絡を調整しながらオッズ比を計算できるのが利点であり，医学統計ではもっともよく使われる手法の一つである。

疾病の有無は 0/1 で表され，データとしては有病割合（総数のうち疾病有りの人数の割合）となるので，そのままではモデルの左辺は 0 から 1 の範囲しかとらないが，右辺は複数のカテゴリ変数と量的変数（多くは交絡因子）からなるので実数のすべての範囲をとる。そのため，左辺をロジット変換（自身を 1 から引いた値で割って自然対数をとる）する。

つまり，疾病の有病割合を P とすると，ロジスティック回帰モデルは次のように定式化できる。

$$\ln(P/(1-P)) = b_0 + b_1X_1 + ...b_kX_k$$

もし X_1 が要因の有無を示す 2 値変数で，$X_2, ...X_k$ が交絡であるなら，$X_1 = 0$ の場合を $X_1 = 1$ の場合から引けば，

$$b1 = \ln(P_1/(1-P_1)) - \ln(P_0/(1-P_0)) = \ln(P_1 * (1-P_0)/(P_0 * (1-P_1)))$$

となるので，b_1 が他の変数の影響を調整したオッズ比の対数になる。対数オッズ比が正規分布するとすれば，オッズ比の 95% 信頼区間が

$$\exp(b_1 \pm 1.96 \times \mathrm{SE}(b_1))$$

として得られる。

例題

`library(MASS)` の `data(birthwt)` は，Springfield の Baystate 医療センターの 189 の出生について，低体重出生とそのリスク因子の関連を調べたデータで，次の変数を含んでいる。低体重出生の有無を反応変数としたロジスティック回帰分析をせよ。

low	低体重出生の有無を示す 2 値変数（児の出生時体重 2.5 kg 未満が 1）
age	年齢
lwt	最終月経時体重（ポンド単位。略号 lb. で，1 lb. は 0.454 kg に当たる）
race	人種（1＝白人，2＝黒人，3＝その他）
smoke	喫煙の有無（1＝あり）
ptl	早期産経験数
ht	高血圧の既往（1＝あり）
ui	子宮神経過敏の有無（1＝あり）
ftv	妊娠の最初の 3 ヶ月の受診回数
bwt	児の出生時体重 (g)

データには多くの変数が含まれているが，本来，ロジスティック回帰分析では，反応変数に対する効果を見たい変数と交絡因子となっている変数はすべて説明変数としてモデルに投入するべきである（説明変数と反応変数の両方と有意な相関があれば交絡因子となっている可能性がある）。

ここでは，丁寧な考察を経て，独立変数が人種，喫煙の有無，高血圧既往の有無，子宮神経過敏の有無，最終月経時体重，早期産経験数となったとしよう。ロジスティック回帰分析の前に，数値型で入っているカテゴリ変数を要因型に変換しておく必要がある。R コンソールでは次のように打てばよい。

```
library(MASS)
data(birthwt)
birthwt$clow <- factor(birthwt$low, labels=c("NBW","LBW"))
birthwt$crace <- factor(birthwt$race, labels=c("white","black","others"))
birthwt$csmoke <- factor(birthwt$smoke, labels=c("nonsmoke","smoke"))
birthwt$cht <- factor(birthwt$ht, labels=c("normotensive","hypertensive"))
birthwt$cui <- factor(birthwt$ui, labels=c("uterine.OK","uterine.irrit"))
```

Rcmdr では，「データ」「アクティブデータセット内の変数の管理」「数値変数を因子に変換」を選び，まず変数として `low` を選び，新しい変数名を `clow` として [OK] ボタンをクリックする。数値 0 が水準 1 となり（NBW と名付ける），数値 1 が水準 2 となる（LBW と名付ける）。次に `race` を選び，新変数名を `crace` として [OK] ボタンをクリックし，出てくるウィンドウで第 1 水準に"white"，第 2 水準に"black"，第 3 水準に"others"とカテゴリ名を指定し，[OK] ボタンをクリックする。`smoke`，`ht`，`ui` についても同様にカテゴリ変数 `csmoke`，`cht`，`cui` に変換する。

ロジスティック回帰モデルをこのデータに当てはめるには，R コンソールでは次の 2 行を打てば良い。

```
res <- glm(clow ~ crace+csmoke+cht+cui+lwt+ptl, family=binomial(logit), data=birthwt)
summary(res)
```

もしモデルがどの程度データを説明しているのか評価したければ，線型重回帰モデルの自由度調整済み重相関係数の代わりに，Nagelkerke の R^2 を計算することができる。

```
require(fmsb)
NagelkerkeR2(res)
```

Rcmdr では，「統計量」「モデルへの適合」「一般化線型モデル」で，式の左辺に `clow`（**因子**）**をクリックして代入し（たんに** `clow` **と入る），右辺に** `crace+csmoke+cht+cui+lwt+ptl` **と打つ（またはクリックして選ぶ）。リンク関数族を** `binomial` **にして，リンク関数を** `logit` **にして [OK] する。Rcmdr では Nagelkerke の** R^2 **を求めるオプションはない。**

R コンソールでも Rcmdr でも，表示される結果は次の通りである。

```
Call:
glm(formula = clow ~ crace + csmoke + cht + cui + lwt + ptl,
    family = binomial(logit), data = birthwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9049  -0.8124  -0.5241   0.9483   2.1812

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.086550   0.951760  -0.091  0.92754
craceblack        1.325719   0.522243   2.539  0.01113 *
craceothers       0.897078   0.433881   2.068  0.03868 *
csmokesmoke       0.938727   0.398717   2.354  0.01855 *
chthypertensive   1.855042   0.695118   2.669  0.00762 **
cuiuterine.irrit  0.785698   0.456441   1.721  0.08519 .
lwt              -0.015905   0.006855  -2.320  0.02033 *
ptl               0.503215   0.341231   1.475  0.14029
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 201.99  on 181  degrees of freedom
AIC: 217.99

Number of Fisher Scoring iterations: 4
```

この出力からロジスティック回帰分析の結果は，下表のようにまとめられる。このように，量的な変数は表の下に共変量として調整したと書くのが普通である。また，係数は対数オッズ比でなく指数をとってオッズ比に直し，95% 信頼区間も表示する。この操作は残念ながら Rcmdr ではまだできないので，分析結果が付値されている変数（回帰分析のモデルを指定するウィンドウの中で「モデル名」として指定したもの）が `GLM.1` だったとすると，R コンソールで，`exp(coef(GLM.1))` とすればオッズ比の点推定量が得られるし，`exp(confint(GLM.1))` とすれば 95% 信頼区間が得られる。

表. Baystate 医療センターにおける低体重出生リスクのロジスティック回帰分析結果

独立変数 *	オッズ比	95% 信頼区間 下限	上限	p 値
人種（白人）				
黒人	3.765	1.355	10.68	0.011
他の有色人種	2.452	1.062	5.878	0.039
喫煙あり（なし）	2.557	1.185	5.710	0.019
高血圧既往あり（なし）	6.392	1.693	27.3	0.008
子宮神経過敏あり（なし）	2.194	0.888	5.388	0.085

AIC: 217.99，D_{null}: 234.67（自由度 188），D: 201.99（自由度 181）

* カッコ内はリファレンスカテゴリ。これらの変数の他，最終月経時体重と早期産経験数を共変量としてロジスティック回帰モデルに含んでいる。

8　2 つのカテゴリ変数による分割表について独立性の仮説を検定する

カテゴリ変数間の関係を調べるにはどうしたらよいだろうか？　もちろん，カテゴリ変数についても関連の強さをみる指標はあって，ファイ係数（記号は ρ を用いるのが普通）と呼ばれる指標は，要因の有無，発症の有無を 1,0 で表した場合のピアソンの積率相関係数と同じ計算式で得られる。θ_1, θ_2 を発症者中の要因あり割合，非発症者中の要因あり割合として，$\rho = \sqrt{(\pi_1 - \pi_2)(\theta_1 - \theta_2)}$ である。また，疫学研究では，人数あるいは人年の比を取ることによって，要因があった群が，要因がなかった群に比べて，どれくらい発症しやすいかを調べることが多い（オッズ比やリスク比やハザード比を求め，その 95% 信頼区間が 1 を含まないかどうかで，要因の有無が発症の有無に有意に影響しているかどうかを判定することが慣例的に行われる）。

しかし，2 つのカテゴリ変数の関係を考えるとき，一般に，もっともよく行われるのは，それらが独立であるという帰無仮説を立てて検定することである。

カテゴリ変数のもつ統計的な情報は，カテゴリごとの度数だけである。そこで，2 つのカテゴリ変数の間に関係について検討したいときには，まず**それらの組み合わせの度数**を調べた表を作成する（R では `table()` という関数が使える）。これを**クロス集計表**と呼ぶ。とくに，2 つのカテゴリ変数が，ともに 2 値変数のとき，そのクロス集計は 2 × 2 クロス集計表 [2 by 2 cross tabulation]（2 × 2 分割表 [2 by 2 contingency table]）と呼ばれ，その統計的性質が良く調べられている。

8.1　独立性のカイ二乗検定

独立性の検定としては，2 つのカテゴリ変数の間に**関連がない**と仮定した場合に推定される期待度数を求めて，それに観測度数が適合するかを検定するカイ二乗検定が最も有名である[*51]。

	A	$\bar{A}$
B	a 人	b 人
$\bar{B}$	c 人	d 人

2 つのカテゴリ変数 A と B が，それぞれ「あり」「なし」の 2 つのカテゴリ値しかとらないとき，これら 2 つのカテゴリ変数の組み合わせは「A も B もあり（$A \cap B$）」「A なし B あり（$\bar{A} \cap B$）」「A あり B なし（$A \cap \bar{B}$）」「A も B もなし（$\bar{A} \cap \bar{B}$）」の 4 通りしかない。それぞれの度数を数えあげた結果が，上記の表として得られたときに，母集団の確率構造が，

[*51] もちろん，ある種の関連が仮定できれば，その仮定の元に推定される期待度数と観測度数との適合を調べてもいいが，一般に，2 つのカテゴリ変数の間にどれくらいの関連がありそうかという仮定はできないことが多い。そこで，関連がない場合の期待度数を推定し，それが観測値に適合しなければ関連がないとはいえない，と推論するのである。

	A	$\bar{A}$
B	π_{11}	π_{12}
$\bar{B}$	π_{21}	π_{22}

であるとわかっていれば，$N = a + b + c + d$ として，期待される度数は，

	A	$\bar{A}$
B	$N\pi_{11}$	$N\pi_{12}$
$\bar{B}$	$N\pi_{21}$	$N\pi_{22}$

であるから，

$$\chi^2 = \frac{(a - N\pi_{11})^2}{N\pi_{11}} + \frac{(b - N\pi_{12})^2}{N\pi_{12}} + \frac{(c - N\pi_{21})^2}{N\pi_{21}} + \frac{(d - N\pi_{22})^2}{N\pi_{22}}$$

として，自由度３のカイ二乗検定をすればよいことになる。しかし，普通，π は未知である。そこで，$Pr(\bar{A}) = 1 - Pr(A)$ かつ，この２つのカテゴリ変数が独立ならば $Pr(A \cap B) = Pr(A)Pr(B)$ と考えれば良い[*52]ことを使って，$Pr(A)$ と $Pr(B)$ を母数として推定する[*53]。$Pr(A)$ の点推定量は，B を無視して A の割合と考えれば $(a + c)/N$ であることは自明である。同様に，$Pr(B)$ の点推定量は，$(a + b)/N$ となる。したがって，$\pi_{11} = Pr(A \cap B) = Pr(A)Pr(B) = (a + c)(a + b)/(N^2)$ となる。

同様に考えれば，母集団の各セルの確率は下式で得られる。

$$\pi_{12} = (b + d)(a + b)/(N^2)$$

$$\pi_{21} = (a + c)(c + d)/(N^2)$$

$$\pi_{22} = (b + d)(c + d)/(N^2)$$

これらの値を使えば，

$$\chi^2 = \frac{\{a - (a + c)(a + b)/N\}^2}{\{(a + c)(a + b)/N\}} + \frac{\{b - (b + d)(a + b)/N\}^2}{\{(b + d)(a + b)/N\}} + \frac{\{c - (a + c)(c + d)/N\}^2}{\{(a + c)(c + d)/N\}} + \frac{\{d - (b + d)(c + d)/N\}^2}{\{(b + d)(c + d)/N\}}$$

$$= \frac{(ad - bc)^2 \{(b + d)(c + d) + (a + c)(c + d) + (b + d)(a + b) + (a + c)(a + b)\}}{(a + c)(b + d)(a + b)(c + d)N}$$

分子の中括弧の中は N^2 なので，結局，

$$\chi^2 = \frac{N(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}$$

ただし通常は，イェーツの連続性の補正を行う。カイ二乗分布は連続分布なので，各度数に 0.5 を足したり引いたりしてやると，より近似が良くなるという発想である。この場合，

$$\chi_c^2 = \frac{N(|ad - bc| - N/2)^2}{(a + c)(b + d)(a + b)(c + d)}$$

が自由度１のカイ二乗分布に従うと考えて検定する。**なお，$|ad - bc|$ が $N/2$ より小さいときは補正の意味がないので，$\chi^2 = 0$ とするのが普通である。$|ad - bc| < N/2$ のとき，R の `chisq.test()` では Yates の元論文の主旨に従うということで補正されてしまうけれども，`prop.test()` では補正されない。**

実際の検定はクロス集計表が既に得られているとき，例えば a=12, b=8, c=9, d=10 などとわかっていれば，R コンソールでは，次のように入力すれば行列の定義とカイ二乗検定ができる。

[*52] この帰無仮説は，合計に比例する割合で人数配分が行われていることに相当するので，B あり群と B なし群のそれぞれについて，A ありの割合に差がないという，比率の差の検定の帰無仮説と数学的に等価である。

[*53] $Pr(X)$ はカテゴリ X の出現確率を示す記号である。また，２つの母数をデータから推定するので，得られるカイ二乗統計量が従う分布の自由度は３より２少なくなり，自由度１のカイ二乗分布となる。

```
x <- matrix(c(12,9,8,10), 2, 2)
# x <- matrix(c(12,8,9,10), 2, 2, byrow=TRUE) is also possible.
chisq.test(x)
```

Rcmdr では，「統計量」「分割表」「２元表の入力と分析」で表示される表の各セルに直接数字を入力し，必要な統計量のチェックボックスにチェックを入れて OK ボタンをクリックするだけである。

例題

肺ガンの患者 100 人に対して，１人ずつ性・年齢が同じ健康な人を対照として 100 人選び（この操作をペアマッチサンプリングという），それぞれについて過去の喫煙の有無を尋ねた結果，患者群では過去に喫煙を経験した人が 80 人，対照群では過去に喫煙を経験した人が 55 人だった。肺ガンと喫煙は無関係といえるか？　独立性のカイ二乗検定をせよ。

帰無仮説は，肺ガンと喫煙が無関係（独立）ということである。クロス集計表を作ってみると，

	肺ガン患者群	健康な対照群	合計
過去の喫煙経験あり	80	55	135
過去の喫煙経験なし	20	45	65
合計	100	100	200

となる。肺ガンと喫煙が無関係だという帰無仮説の下で期待される各カテゴリの人数は，

	肺ガンあり	肺ガンなし
喫煙あり	$135 \times 100/200 = 67.5$	$135 \times 100/200 = 67.5$
喫煙なし	$65 \times 100/200 = 32.5$	$65 \times 100/200 = 32.5$

となる。従って，連続性の補正を行なったカイ二乗統計量は，

$$\chi_c^2 = (80-68)^2/67.5 + (55-67)^2/67.5 + (20-32)^2/32.5 + (45-33)^2/32.5 = 13.128...$$

となり，自由度１のカイ二乗分布で検定すると `1-pchisq(13.128,1)` より有意確率は 0.00029... となり，有意水準 5% で帰無仮説は棄却される。つまり，肺ガンの有無と過去の喫煙の有無には 5% 水準で統計学的に有意な関連があるといえる。

R コンソールでは次の 1 行を打つだけで上の結果を得ることができる。

```
chisq.test(matrix(c(80,20,55,45),2,2))
```

Rcmdr では，「統計量」「分割表」「２元表の入力と分析」で，対応するセルに直接人数を入力して [OK] をクリックすればよい。

例題

MASS パッケージの survey データフレームで，性別 (Sex) が利き手 (W.Hnd) と独立であるという帰無仮説を検定せよ。

R コンソールでは次の２行を打てばできる。

```
require(MASS)
chisq.test(xtabs(~ Sex+W.Hnd, data=survey))
```

得られる p 値は 0.6274 であり，性別と利き手の間に統計学的に有意な関連があるとはいえないことを意味する。

Rcmdr では survey **データフレームをアクティブにした後で，「統計量」「分割表」「２元表...」を選び，行の変数として** Sex **を，列の変数として** W.Hnd **を選び，[OK] をクリックする。アウトプットウィンドウに，** X-squared = 0.5435, df = 1, p-value = 0.461 **と結果が表示される。これは Yates の補正なしの値である。Rcmdr では，**chisq.test() **関数は必ず** correct=FALSE **オプション付きで実行される。**
しかし，別の方法で Yates の補正ありの独立性のカイ二乗検定と同じ結果を得ることはできる。「統計量」の「比率」から「２標本の比率の検定」を選ぶ。既に書いた通り，グループとして Sex**，目的変数として** W.Hnd **を指定し，検定のタイプとして「連続修正を用いた正規近似」にチェックを入れて [OK] ボタンをクリックすれば，p 値として 0.6274 という Yates の補正をしたときの値が得られる。**

8.2 フィッシャーの正確確率

期待度数が低い組み合わせがあるときには，カイ二乗検定での正規近似が非常に悪い近似になる。そういう場合，カテゴリを併合して変数を作り直し，組み合わせの種類を減らして，各組み合わせの頻度を上げる方法もあるが，もっといい方法が考案されている。それがフィッシャーの正確確率（検定）である。

ある２次元クロス集計表が与えられたとして，周辺度数を固定して（各々の変数については母比率が決まっていると仮定して）すべての組み合わせを考え，それらが起こる確率（超幾何分布に従う）を１つずつ計算し，得られている集計表が得られる確率よりも低い確率になるような表が得られる確率をすべて足し合わせてしまえば，２つのカテゴリ変数の間に関連がないという帰無仮説の下でそういう表が偶然得られる確率がどれほど低いのかを，直接計算することができる。こうして計算される確率を，フィッシャーの正確確率という。これなら，近似ではないので，期待度数が低い組み合わせがあっても問題ない。

もう少し丁寧に言うと，サイズ N の有限母集団があって，そのうち変数 A の値が 1 である個体数が m_1，1 でない個体数が m_2 あるときに，変数 B の値が 1 である個体数が n_1 個（1 でない個体数が $n_2 = N - n_1$ 個）あるという状況を考え，この n_1 個のうち変数 A の値が 1 である個体数がちょうど a である確率を求めることになる。これは，m_1 個から a 個を取り出す組み合わせの数と m_2 個から $n_1 - a$ 個を取り出す組み合わせの数を掛けて，N 個から n_1 個を取り出す組み合わせの数で割った値になる。これと同じ周辺度数をもつ２×２分割表のうち，確率がこれと同じかこれよりも小さい表の確率をすべて足し合わせたものが，「変数 A と変数 B が独立」という帰無仮説が成り立つ確率になる。

有限母集団からの非復元抽出になるので，平均 $E(a)$ と分散 $V(a)$ は，

$$E(a) = n_1 m_1 / N$$

$$V(a) = \{(N - n_1)/(N - 1)\} n_1 (m_1/N)(m_2/N) = (m_1 m_2 n_1 n_2)/\{N^2(N - 1)\}$$

となる。実際には組み合わせ計算が多いので，手計算で実行することはまずありえず，統計ソフトにやらせることになる。また，個々の２×２分割表の確率は離散値をとるので，同じ確率の表がありうる場合に，それを足し算に含めるのかどうかは難しい点である。

しかし，フィッシャーの正確確率は，近似を使わないので，クロス集計表を使って２つのカテゴリ変数間の独立性の検定をするときは，コンピュータが使えるならば，サンプルサイズがよほど大きくない限り，常にカイ二乗検定ではなく，フィッシャーの正確な確率を求めるべきである。R コンソールで実行するのは簡単で，カイ二乗検定で chisq.test() と書かれていたところを，fisher.test() で置き換えればいい。

例題

MASS パッケージの survey データフレームについて，性別 (Sex) と喫煙習慣 (Smoke) が独立であるとしたときに，実際得られている組み合わせあるいはそれより起こりにくい組み合わせが偶然得られる確率を，フィッシャーの正確確率によって計算せよ。

R コンソールでは次の２行を打てばよい。

```
require(MASS)
fisher.test(xtabs(~Sex+Smoke, data=survey))
```

Rcmdr では，survey データフレームをアクティブにしておき，「統計量」「分割表」「２元表の分析」として行の変数として Sex，列の変数として Smoke を選んでから，「フィッシャーの正確検定」にチェックを入れて [OK] ボタンをクリックする。

どちらも `p-value = 0.3105` という同じ結果を示す。したがって，性別と喫煙習慣が無関係である可能性は有意水準 5% で否定できない。

9 繰り返し測定または複数の評価者による分割表

順序変数またはカテゴリ変数について，各個人の同じ変数で２時点での値があるか，あるいは複数の評価者による評価値がある場合，その結果は２次元クロス集計表としてまとめることができ，この表は検査＝再検査信頼性，あるいは評価者間信頼性を調べるのに使うことができる。しかし，この目的ではカイ二乗検定もフィッシャーの正確確率も不適切である。なぜなら，各個人の２時点の値や，複数の評価者により同じ人を評価した値は，明らかに独立ではないからである。知りたいのは，偶然の一致とは考えられないほど良く一致しているかどうかといったことになる。

２つの測定値の一致を知りたい場合は，カッパ統計量（κ）を使うことができる。この場合，帰無仮説は，２つの測定値の一致が偶然の一致と同じということであり，対立仮説は，偶然の一致よりも有意に大きい一致になる。

逆に介入効果を知りたい場合など，２時点での測定値があっても，偶然よりも違いがあることを明らかにしたい場合もある。この場合は帰無仮説はカッパ統計量と同じだが，対立仮説が**２つの測定値が偶然の一致よりも違っている**ことになる。この場合はマクネマー (McNemar) の検定が使える。ただし，変数が単なるカテゴリではなく順序変数の場合で，カテゴリ数が３以上なら，ウィルコクソンの符号付き順位検定を使うこともできるし，その方が適切である場合が多い。

9.1 カッパ統計量

検査再検査信頼性を評価するために次の表が得られたとしよう。

	再検査	
	陽性	陰性
検査結果 陽性	a	b
検査結果 陰性	c	d

もし２回の検査結果が完璧に一致していたら $b = c = 0$ となるが，通常は $b \neq 0$ かつ／または $c \neq 0$ である。ここで，２回の検査結果の一致確率は，$P_o = (a + d)/(a + b + c + d)$ と定義できる。完全な一致の場合，$b = c = 0$ から $P_o = (a+d)/(a+d) = 1$ となる。逆に完全な不一致の場合, $a = d = 0$ から $P_o = 0$ となる。一致の程度が偶然と同じならば, 期待される一致確率 P_e は, 次の式で計算できる。 $P_e = \{(a+c)(a+b)/(a+b+c+d)+(b+d)(c+d)/(a+b+c+d)\}/(a+b+c+d)$

ここで，κ を $\kappa = (P_o - P_e)/(1 - P_e)$ と定義すると，完全な一致のとき $\kappa = 1$，偶然と同程度の一致のとき $\kappa = 0$，偶然の一致より悪い一致のとき $\kappa < 0$ となる。κ の分散 $V(\kappa)$ は $V(\kappa) = P_e/\{(a + b + c + d) \times (1 - P_e)\}$ となるので，$\kappa/\sqrt{V(\kappa)}$ として標準化した統計量は標準正規分布に従う。そこで，$\kappa = 0$ という帰無仮説を検定したり，κ の 95% 信頼区間を計算することが可能になる。

追加パッケージ `vcd` には，`Kappa()` という関数があって，κ の点推定量を計算できるし，`confint()` 関数を適用すれば 95% 信頼区間も計算できる。また，筆者も κ を計算する関数 `Kappa.test` を開発し公開している。この関数は `fmsb` パッケージに含まれている。残念なことに，現在のところ，Rcmdr では κ を計算する項目は提供されていない。

数値計算をしてみよう。次の表を考える。

	再検査	
	陽性	陰性
検査陽性	12	4
検査陰性	2	10

R コンソールでは，fmsb パッケージをインストールしてあれば，次の 2 行を打つだけで，次の枠内に示す結果がすべて得られる。

```
require(fmsb)
Kappa.test(matrix(c(12,2,4,10),2,2))
```

```
$Result
        Estimate Cohen's kappa statistics and test the null
        hypothesis that the extent of agreement is same as random
        (kappa=0)
data:  matrix(c(12, 2, 4, 10), 2, 2)
Z = 3.0237, p-value = 0.001248
95 percent confidence interval:
 0.2674605 0.8753967
sample estimates:
[1] 0.5714286

$Judgement
[1] "Moderate agreement"
```

ここで “Judgement”（一致度の判定）は，Landis JR, Koch GG (1977) Biometrics, 33: 159-174 の基準によっている。もし κ が負ならば"No agreement"（不一致），0-0.2 なら"Slignt agreement"（微かな一致），0.2-0.4 なら"Fair agreement"（多少の一致），0.4-0.6 なら"Moderate agreement"（中程度の一致），0.6-0.8 なら"Substantial agreement"（かなりの一致），0.8-1.0 だと"Almost perfect agreement"（ほぼ完全な一致）と表示される。大雑把なガイドラインに過ぎないが，実用的な基準である。

9.2 マクネマーの検定

マクネマーの検定は，元々は２×２クロス集計表について開発された。次の表を考えてみよう。

	介入後	
	あり	なし
介入前 あり	a	b
介入前 なし	c	d

マクネマーの検定では，次のように定義する χ_0^2 を計算する。この χ_0^2 統計量は，帰無仮説（２回の測定結果の一致の程度が偶然と差が無い）の下で自由度１のカイ二乗分布に従う。

$$\chi_0^2 = \frac{(b-c)^2}{(b+c)}$$

連続性の補正をする場合は次の式になる（ただし b と c が等しいときは $\chi_0^2 = 0$ とする）。

$$\chi_0^2 = \frac{(|b-c|-1)^2}{(b+c)}$$

拡張マクネマー検定は，M × M クロス集計に適用できるようにしたものである。セル [i,j] に入る人数を n_{ij}（i, j = 1, 2, ..., M）とすると，次の式で χ_0^2 を計算することができ，この χ_0^2 統計量は帰無仮説（[i,j] に入る確率と [j,i] に入る確率が同じ）の下で自由度 M(M-1)/2 のカイ二乗分布に従う。

$$\chi_0^2 = \frac{\sum_{i<j}\left(n_{ij}-n_{ji}\right)^2}{(n_{ij}+n_{ji})}$$

R コンソールでは，既に関数が提供されていて，対応する 2 変数間の分割表を TABLE と書くことにすると，マクネマーの検定は，mcnemar.test(TABLE) とするだけでできる。なお，残念ながら Rcmdr のメニューにはまだ入っていない。

10 生存時間解析

生存時間解析は Rcmdr 本体には入っていない。しかし，プラグインが 2 種類発表されている。1 つは John Fox 教授自身が開発した RcmdrPlugin.survival であり，もう 1 つは Dr. Daniel C. Leucuta というルーマニアの研究者が開発した RcmdrPlugin.SurvivalT である[*54]。いずれも，survival ライブラリの機能の基本的なものに対して，グラフィカルなユーザーインターフェースを提供するものである。ここでは，R コンソールで次のように打ち，前者をインストールしたものとする。

```
install.package(RcmdrPlugin.survival)
```

この入門的な授業の枠内で生存時間解析を十分説明することは難しいが，R の survival パッケージでは，生存時間解析を実行するための多くの関数が提供されており，かなり高度な解析まで実行できる。

Rcmdr の [ツール] から [プラグインのロード] を選んで RcmdrPlugin.survival を選んで [OK] をクリックすると，Rcmdr を再起動するかどうか聞いてくるので，[はい] をクリックすると，次のような生存時間解析関係のメニューが Rcmdr に追加された状態になる（以下，この状態を “Rcmdr+RcmdrPlugin.survival” と表記する）：「データ」の「Survival data」，「統計量」の「Survival analysis」，「モデルへの適合」のいくつかの項目，「モデル」の「グラフ」のいくつかの項目，「モデル」の「数値による診断」の「Test proportional hazards」である。

10.1 生存時間解析とは

実験においては，化学物質などへの 1 回の曝露の影響を時間を追ってみていくことが良く行われる。時間ごとに何らかの量の変化を追うほかに，エンドポイントを死亡とした場合，死ぬまでの時間を分析することで毒性の強さを評価することができる。このような期間データを扱う方法としては，一般に生存時間解析 (Survival Analysis または Event History Analysis) と呼ばれるものがある。なかでもよく知られているものが Kaplan-Meier の積・極限推定量である（現在では一般に，カプラン＝マイヤ推定量と呼ばれている）。カプラン＝マイヤ推定量は，イベントが起こった各時点での，イベントが起こる可能性がある人口（リスク集合）あたりのイベント発生数を 1 から引いたものを掛け合わせて得られる，ノンパラメトリックな最尤推定量である。また，複数の期間データ列があったときに，それらの差を検定したい場合は，ログランク検定や一般化ウィルコクソン検定が使われる。

それらのノンパラメトリックな方法とは別に，イベントが起こるまでの時間が何らかのパラメトリックな分布に当てはまるかどうかを調べる方法もある。当てはめる分布としては指数分布やワイブル分布がある。中間的なものとして，イベントが起こるまでの期間に対して，何らかの別の要因群が与える効果を調べたいときに，それらが基準となる個体のハザードに対して $\exp(\sum \beta z_i)$ という比例定数の形で掛かるとする比例ハザード性だけを仮定し，分布の形は仮定しないコックス回帰（比例ハザードモデルとも呼ばれる）は，セミパラメトリックな方法といえる。別の要因群の効果は，パラメトリックなモデルに対数線形モデルの独立変数項として入れてしまう加速モデルによって調べることもできる。

R では生存時間解析をするための関数は survival パッケージで提供されており，R コンソールで library(survival) または require(survival) とタイプして survival パッケージをメモリにロードした後では，Surv() で生存時間クラスをもつオブジェクトの生成，survfit() でカプラン＝マイヤ法，survdiff() でログランク検定，coxph() でコックス回帰，survreg() で加速モデルの当てはめを実行できる。なお，生存時間解析について，より詳しく知りたい方は，大橋，浜田 (1995) などを参照されたい。

[*54] このプラグインについての説明が，Leucuta DC, Achimas-Cadariu A (2008) Statistical graphical user interface plug-in for survival analysis in R statistical and graphics language and environment. *Applied Medical Informatics*, 23(3-4): 57-62. という論文として発表されている。

10.2 カプラン＝マイヤ法

まず，カプラン＝マイヤ推定量についての一般論を示す。イベントが起こる可能性がある状態になってから，イベントが起こった時点を $t_1, t_2, ...$ とし，t_1 時点でのイベント発生数を d_1，t_2 時点でのイベント発生数を d_2，以下同様であるとする。また，時点 $t_1, t_2, ...$ の直前でのリスク集合の大きさを $n_1, n_2, ...$ で示す。リスク集合の大きさとは，その直前でまだイベントが起きていない個体数である。観察途中で死亡や転居などによって打ち切りが生じるために，リスク集合の大きさはイベント発生によってだけではなく，打ち切りによっても減少する。従って n_i は，時点 t_i より前にイベント発生または打ち切りを起こした個体数を n_1 から除いた残りの数となる。なお，イベント発生と打ち切りが同時点で起きている場合は，打ち切りをイベント発生直後に起きたと見なして処理するのが慣例である。このとき，カプラン＝マイヤ推定量 $\hat{S}(t)$ は，

$$\hat{S}(t) = (1 - d_1/n_1)(1 - d_2/n_2)... = \prod_{i<t}(1 - d_i/n_i)$$

として得られる。その標準誤差はグリーンウッドの公式により，

$$var(\hat{S}) = \hat{S}^2 \times \sum_{i<t} \frac{d_i}{n_i(n_i - d_i)}$$

で得られる。なお，カプラン＝マイヤ推定量を計算するときは，階段状のプロットを同時に行うのが普通である。

R コンソールでは，`library(survival)` としてパッケージを呼び出し，`dat <- Surv(`**生存時間，打ち切りフラグ**`)` 関数で生存時間データを作り（打ち切りフラグは 1 でイベント発生，0 が打ち切り。ただし区間打ち切りの場合は 2 とか 3 も使う），`res <- survfit(dat~1)` でカプラン＝マイヤ法によるメディアン生存時間が得られ（群分け変数 `C` によって群ごとにカプラン＝マイヤ推定をしたい場合は，`res <- survfit(dat~C)` とすればよい），`plot(res)` とすれば階段状の生存曲線が描かれる。イベント発生時点ごとの値を見るには，`summary(res)` とすればよい[*55]。

例題

`survival` ライブラリに含まれているデータ `aml` は，急性骨髄性白血病 (acute myelogenous leukemia) 患者が化学療法によって寛解した後，ランダムに 2 群に分けられ，1 群は維持化学療法を受け（維持群），もう 1 群は維持化学療法を受けずに（非維持群），経過観察を続けて，維持化学療法が再発までの時間を延ばすかどうかを調べたデータである[a]。以下の 3 つの変数が含まれている。

time　生存時間あるいは観察打ち切りまでの時間（週）
status　打ち切り情報（0 が観察打ち切り，1 がイベント発生）
x　維持化学療法が行われたかどうか（Maintained が維持群，Nonmaintained が非維持群）

薬物維持化学療法の維持群と非維持群で別々に，再発までの時間の中央値をカプラン＝マイヤ推定し，生存曲線をプロットせよ。

[a] 出典: Miller RG: Survival Analysis. John Wiley and Sons, 1981. 元々は，Embury SH, Elias L, Heller PH, Hood CE, Greenberg PL, Schrier SL: Remission maintenance therapy in acute myelogenous leukaemia. *Western Journal of Medicine*, 126, 267-272, 1977. のデータ。研究デザインは Gehan と似ている。

R コンソールでは以下を入力する。

```
library(survival)
print(res <- survfit(Surv(time,status)~x, data=aml))
plot(res,xlab="(Weeks)",lty=1:2,main="Periods until remission of acute myelogenous leukaemia")
legend("right",lty=1:2,legend=levels(aml$x))
```

2 行目でカプラン＝マイヤ法での計算がなされ，下枠内が表示される。

[*55] 参考までに書いておくと，生データがイベント発生の日付を示している場合，間隔を計算するには `difftime()` 関数や `ISOdate()` 関数を使う。例えば 1964 年 8 月 21 日生まれの人の今日の年齢は `integer(difftime(ISOdate(2007,6,13),ISOdate(1964,8,21)))/365.24` とすれば得られるし，さらに 12 を掛ければ月単位になる。

```
                records n.max n.start events median 0.95LCL 0.95UCL
x=Maintained         11    11      11      7     31      18      NA
x=Nonmaintained      12    12      12     11     23       8      NA
```

この表は，維持群が 11 人，非維持群が 12 人，そのうち再発が観察された人がそれぞれ 7 人と 11 人いて，維持群の再発までの時間の中央値が 31 週，非維持群の再発までの時間の中央値が 23 週で，95% 信頼区間の下限はそれぞれ 18 週と 8 週，上限はどちらも無限大であると読む。

3 行目で 2 群別々の再発していない人の割合の変化が生存曲線として描かれ，4 行目で凡例が右端に描かれる。

Rcmdr+RcmdrPlugin.survival では，まず survival パッケージ内の aml をアクティブにする。「データ」の「アタッチされたパッケージからデータを読み込む」では，`aml (leukemia)`**ではなくて，**`leukemia`**の方を選ばないとうまくいかなかったので注意されたい。**
カプラン＝マイヤ推定は，「統計量」＞「Survival analysis」＞「Estimate survival function」と進み，“Time or start/end times” として [time] を選び，“Event indicator” として [status] を選ぶ。次に Strata だが，aml データで Maintained 群と Non-maintained 群別々にカプラン＝マイヤ推定したいときは，ここに表示されている [x] をクリックする。全データを一括して推定したい場合は何もしない（間違って一度クリックしてしまうと，たぶん解除できないので，その場合は諦めてウィンドウを閉じ，やり直すのが無難である）。
このメニューは生存曲線も描いてくれる。“Confidence Intervals” として “Log”，“Log-log”，“plain”，“none” を指定できる。デフォルトは “Log” である（つまり，`survfit()`**関数で，**`conf.type=`**オプションを指定しない場合は，**`summary()`**で出力される各イベント発生時点の信頼区間は，**`conf.type="log"`**として計算される）。大橋・浜田 (1995) の p.68 の出力を見ると，SAS バージョン 6 の計算は，**`conf.type="plain"`**とした場合と一致していたし，Statistics in Medicine に 1997 年に掲載されていたチュートリアル論文での計算（**`http://phi.med.gunma-u.ac.jp/swtips/survival.html`**を参照のこと）は，**`conf.type="log-log"`**とした場合と一致した。“Plot confidence intervals” は，常に生存曲線に信頼区間をつけて描画する [Yes] と常に信頼区間は描画しない [No] の他に，[Default] が指定できる。[Default] は，Strata が指定されている場合は [No]，Strata が無い場合は [Yes] が指定された場合と同じ動作をする（なお，描画しない場合でも，アウトプットウィンドウには，summary() の結果として，各イベント発生時点での生存確率の信頼区間の値は表示される）。“Confidence level” は信頼水準であり，デフォルトは.95，つまり 95 %である。ここを.99 とすれば 99 %信頼区間が求められる。“Mark censoring times” にはデフォルトでチェックが入っており，生存曲線のグラフの打ち切り発生時点で短いティックマークがプロットされる。このチェックを外すと打ち切りレコードは描画されない。“Method” は [Kaplan-Meier] の他にも 2 つ選べ，“Variance Method” も [Greenwood] でない方法も選べるが，ここはデフォルトのままでいいと思う。“Quantiles to estimate” のボックスに [.25,.5,.75] と入っているが，アウトプットウィンドウで.5 のところに表示される値が生存時間の中央値として推定されるカプラン＝マイヤ推定量である。日本語環境で使う場合に重要なのは，一番下の"部分集合の表現"のボックスに表示されている [＜全ての有効なケース＞] という文字列を消してから [OK] ボタンをクリックすることである。おそらく，このプラグインパッケージのバグと思われるが，文字列を消さないで [OK] をクリックすると，[＜全ての有効なケース＞] という文字列そのものがオプションとして渡されてしまいエラーを生じる。**

10.3 ログランク検定

次に，ログランク検定を簡単な例で説明する。

8 匹のラットを 4 匹ずつ 2 群に分け，第 1 群には毒物 A を投与し，第 2 群には毒物 B を投与して，生存期間を追跡したときに，第 1 群のラットが 4,6,8,9 日目に死亡し，第 2 群のラットが 5,7,12,14 日目に死亡したとする。この場合，観察期間内にすべてのラットが死亡し，正確な生存時間がわかっているので，観察打ち切りがないデータとなっていて計算しやすい。

ログランク検定の思想は，大雑把に言えば，死亡イベントが起こったすべての時点で，群と生存／死亡個体数の 2 × 2 クロス集計表を作り，それをコクラン＝マンテル＝ヘンツェル流のやり方で併合するということである。上記の例で

は，死亡イベントが起こった時点 1〜8 において各群の期待死亡数を計算し，各群の実際の死亡数との差をとって，それに時点の重みを掛けたものを，各時点における各群のスコアとして，群ごとのスコアの合計を求める。２群しかないので，各時点において群１と群２のスコアの絶対値は同じで符号が反対になる。２群の生存時間に差がないという帰無仮説を検定するためには，群１のスコアの２乗を分散で割った値をカイ二乗統計量とし，帰無仮説の下でこれが自由度１のカイ二乗分布に従うことを使って検定する。なお，重みについては，ログランク検定ではすべて１である。一般化ウィルコクソン検定では，重みを，2 群を合わせたリスク集合の大きさとする（そうした場合，もし打ち切りがなければ，検定結果は，ウィルコクソンの順位和検定の結果と一致する）。つまり，ログランク検定でも一般化ウィルコクソン検定でも，実は期間の情報はまったく使われず，死亡順位の情報だけが使われている。

記号で書けば次の通りである。第 i 時点の第 j 群の期待死亡数 e_{ij} は，時点 i における死亡数の合計を d_i，時点 i における j 群のリスク集合の大きさを n_{ij}，時点 i における全体のリスク集合の大きさを n_i とすると，

$$e_{ij} = d_i \cdot n_{ij}/n_i$$

と表される。上の例では，$e_{11} = 1 \cdot n_{11}/n_1 = 4/8 = 0.5$ となる。時点 i における第 j 群の死亡数を d_{ij}，時点の重みを w_i と表せば，時点 i における群 j のスコア u_{ij} は，

$$u_{ij} = w_i(d_{ij} - e_{ij})$$

となり，ログランク検定の場合（以下，重みは省略してログランク検定の場合のみ示す）の群 1 の合計スコアは

$$u_1 = \sum_i d_{i1} - e_{i1}$$

となる。上の例では，

$$u_1 = (1 - 4/8) + (0 - 3/7) + (1 - 3/6) + (0 - 2/5) + (1 - 2/4) + (1 - 1/3) + (0 - 0/2) + (0 - 0/1)$$

である。これを計算すると約 1.338 となる。分散は，分散共分散行列の対角成分を考えればいいので，

$$V = V_{jj} = \sum_i \frac{(n_i - n_{ij})n_{ij}d_i(n_i - d_i)}{{n_i}^2(n_i - 1)}$$

となる。この例の数値を当てはめると，

$$V = \frac{(8-4)\times 4}{8^2} + \frac{(7-3)\times 3}{7^2} + \frac{(6-3)\times 3}{6^2} + \frac{(5-2)\times 2}{5^2} + \frac{(4-2)\times 2}{4^2} + \frac{(3-1)\times 1}{3^2}$$

となり，計算すると，約 1.568 となる。したがって，$\chi^2 = 1.338^2/1.568 = 1.14$ となり，この値は自由度１のカイ二乗分布の 95% 点である 3.84 よりずっと小さいので，有意水準 5% で帰無仮説は棄却されない。つまりこれだけのデータでは，差があるとはいえないことになる（もちろん，サンプルサイズを大きくすれば違う結果になる可能性もある）。

R では，`Surv(time,event)` と `group`（注：ここで time は生存時間，event は 1 がイベント観察，0 が観察打ち切りを示すフラグ，group がグループを示す）を，`survdiff()` 関数に与えることによってログランク検定が実行できる。打ち切りレコードがない場合は，event は省略できる。なお，生存時間解析の関数はすべて `survival` パッケージに入っているので，まず `library(survival)` とすることは必須である。

この例では，R コンソールでは以下を入力する。

```
library(survival)
time <- c(4,6,8,9,5,7,12,14)
event <- c(1,1,1,1,1,1,1,1)
group <- c(1,1,1,1,2,2,2,2)
dat <- Surv(time,event)
survfit(dat~group)
survdiff(dat~group)
```

得られる結果の一番下を見ると，$\chi^2 = 1.2$，自由度 1，$p = 0.268$ となっているので，有意水準 5% で，2 群には差がないことがわかる。

カプラン＝マイヤ法の例題で使った aml データで，維持群と非維持群の生存時間に差が無いという帰無仮説をログランク検定するには，R コンソールでは以下のように打つ。

```
library(survival)
survdiff(Surv(time,status)~x, data=aml)
```

次の枠内の結果が得られる。p 値が 0.0653 なので，有意水準 5% で統計学的に有意な差があるとはいえない。

```
Call:
survdiff(formula = Surv(time, status) ~ x, data = aml)

                N Observed Expected (O-E)^2/E (O-E)^2/V
x=Maintained    11        7    10.69      1.27      3.40
x=Nonmaintained 12       11     7.31      1.86      3.40

 Chisq= 3.4  on 1 degrees of freedom, p= 0.0653
```

Rcmdr+RcmdrPlugin.survival でログランク検定をしたい場合は，`leukemia`**（つまり** `aml`**）データをアクティブにしてから，統計量＞ Survival analysis ＞ Compare Survival Functions でできる。"Time or start/end times" を [time]，"Event indicator"を [status]，"Strata"を [x] にして（色が反転した選択された状態にして），"rho"を 0 のまま，"部分集合の表現"ボックスの中を削除して [OK] ボタンをクリックすると，ログランク検定の結果がアウトプットウィンドウに表示される（"rho"を 1 にすると，Peto-Peto 流の一般化ウィルコクソン検定の結果が得られる）。**

10.4 コックス回帰

コックス回帰も簡単に説明しておく。カプラン＝マイヤ推定やログランク検定は，まったく母数の分布を仮定しない方法だった。コックス回帰は，「比例ハザード性」を仮定するため，比例ハザードモデルと呼ばれる。コックス回帰の基本的な考え方は，イベント発生に影響する共変量ベクトル $z_i = (z_{i1}, z_{i2}, ..., z_{ip})$ をもつ個体 i の，時点 t における瞬間イベント発生率 $h(z_i, t)$（これをハザード関数と呼ぶ）として，

$$h(z_i, t) = h_0(t) \cdot \exp(\beta_1 z_{i1} + \beta_2 z_{i2} + ... + \beta_p z_{ip})$$

を想定するものである。$h_0(t)$ は基準ハザード関数と呼ばれ，すべての共変量のイベント発生への影響がゼロである「基準人」の，時点 t における瞬間イベント発生率を意味する。$\beta_1, \beta_2, ..., \beta_p$ が推定すべき未知パラメータであり，共変量が $\exp(\beta_x z_{ix})$ という比例定数の形でイベント発生に影響するので，このことを「比例ハザード性」と呼ぶのである。なお，Cox が立てたオリジナルのモデルでは，z_i が時間とともに変わる，時間依存性共変量の場合も考慮されていたが，現在，通常行われるコックス回帰では，共変量の影響は時間に依存しないもの（時間が経過しても増えたり減ったりせず一定）として扱うことが多い。そのため，個体間のハザード比は時点によらず一定になるという特徴をもつ。つまり，個体 1 と個体 2 で時点 t のハザードの比をとると，基準ハザード関数 $h_0(t)$ が分母分子からキャンセルされるので，ハザード比は常に，

$$\frac{\exp(\beta_1 z_{11} + \beta_2 z_{12} + ... + \beta_p z_{1p})}{\exp(\beta_1 z_{21} + \beta_2 z_{22} + ... + \beta_p z_{2p})}$$

となる。このため，比例ハザード性を仮定できれば，基準ハザード関数の形について（つまり，生存時間分布について）特定のパラメトリックモデルを仮定する必要がなくなる。この意味で，比例ハザードモデルはセミパラメトリックであるといわれる。

生存関数とハザード関数の関係について整理しておく。T をイベント発生までの時間を表す非負の確率変数とすると，生存関数 $S(t)$ は，$T \geq t$ となる確率である。$S(0) = 1$ となることは定義より自明である。ハザード関数 $h(t)$ は，ある瞬間 t にイベントが発生する確率なので，

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)} \\ &= -\frac{dS(t)}{dt} \frac{1}{S(t)} = -\frac{d(\log(S(t))}{dt} \end{aligned}$$

である。累積ハザード関数 $H(t)$ は，

$$H(t) = \int_0^t h(u)du = -\log S(t)$$

となり，式変形すると，

$$S(t) = \exp(-H(t))$$

とも書ける。共変量ベクトルが z である個体の累積ハザード関数を $H(z,t)$，生存関数を $S(z,t)$ と書けば，前者については，比例ハザード性が成立していれば，

$$H(z,t) = \int_0^t h(z,u)du = \int_0^t h_0(u)\exp(\beta z)du = \exp(\beta z)H_0(t)$$

が成り立ち，それを代入すると，後者について

$$S(z,t) = \exp(-H(z,t)) = \exp\{-\exp(\beta z)H_0(t)\}$$

となる。両辺の対数を取って符号を逆にして再び両辺の対数を取ると，

$$\log(-\log S(z,t)) = \beta z + \log H_0(t)$$

となる。この式から，共変量で層別して，横軸に生存時間をとり，縦軸に生存関数の対数の符号を逆にしてもう一度対数をとった値をとって散布図を描くと，層間で βz だけ平行移動したグラフが描かれることがわかる。これを**二重対数プロット**と呼ぶ。逆に考えれば，二重対数プロットを描いてみて，層ごとの散布図が平行になっていなければ，「比例ハザード性」の仮定が満たされないので，コックス回帰は不適切といえる。

パラメータ β の推定には，部分尤度という考え方が用いられる。時点 t において個体 i にイベントが発生する確率を，時点 t においてイベントが 1 件起こる確率と，時点 t でイベントが起きたという条件付きでそれが個体 i である確率の積に分解すると，前者は生存時間分布についてパラメトリックなモデルを仮定しないと不明だが，後者はその時点でのリスク集合内の個体のハザードの総和を分母，個体 i のハザードを分子として推定できる。すべてのイベント発生について，後者の確率だけをかけあわせた結果を L とおくと，L は，全体の尤度から時点に関する尤度を除いたものになり，その意味で部分尤度あるいは偏尤度と呼ばれる。サンプルサイズを大きくすると真の値に収束し，分布が正規分布で近似でき，分散もその推定量としては最小になるという意味での，「良い」推定量として，パラメータ β を推定するには，この部分尤度 L を最大にするようなパラメータを得ればよいことを Cox が予想したので（後にマルチンゲール理論によって証明された），比例ハザードモデルをコックス回帰という[*56]。R でのコックス回帰の基本は，`coxph(Surv(time,cens)~grp+covar,data=dat)` という形になる。

例題

`aml` データで維持の有無が生存時間に与える影響をコックス回帰せよ。

[*56] なお，同時に発生したイベントが 2 つ以上ある場合は，その扱い方によって，Exact 法とか，Breslow 法，Efron 法，離散法などがあるが，可能な場合は Exact 法を常に使うべきである。また，離散法は，離散ロジスティックモデルに対応する推定法となっていて，生存時間が連続量でなく，離散的にしか得られていない場合に適切である。Breslow 法を使う統計ソフトが多いが，R の `coxph()` 関数のデフォルトは Efron 法である。Breslow 法よりも Efron 法の方が Exact 法に近い結果となる。

```
require(survival)
summary(res <- coxph(Surv(time,status)~x, data=aml))
KM <- survfit(Surv(time,status)~x, data=aml)
par(family="sans",las=1,mfrow=c(1,3))
plot(KM, lty=1:2, main="aml データのカプラン＝マイヤプロット")
legend("topright", lty=1:2, legend=c("維持","非維持"))
plot(survfit(res),
 main="aml データで維持の有無を共変量とした\n 基準生存曲線と 95 %信頼区間")
plot(KM, fun=function(y) {log(-log(y))}, lty=1:2, main="aml データの二重対数プロット")
```

2 行目で得られる結果は，下枠内の通りである。

```
Call:
coxph(formula = Surv(time, status) ~ x, data = aml)

  n= 23
                coef exp(coef) se(coef)    z     p
xNonmaintained 0.916       2.5    0.512 1.79 0.074

               exp(coef) exp(-coef) lower .95 upper .95
xNonmaintained       2.5        0.4     0.916      6.81

Rsquare= 0.137   (max possible= 0.976 )
Likelihood ratio test= 3.38  on 1 df,   p=0.0658
Wald test            = 3.2  on 1 df,   p=0.0737
Score (logrank) test = 3.42  on 1 df,   p=0.0645
```

p 値が 0.074 なので，有意水準 5% で「維持化学療法の有無が生存時間に与えた効果がない」という帰無仮説は棄却されない[*57]。exp(coef) の値 2.5 が，2 群間のハザード比の推定値になるので，維持群に比べて非維持群では 2.5 倍死亡ハザードが高いと考えられるが，95% 信頼区間が 1 を挟んでおり，有意水準 5% では統計学的に有意な違いではない。

3 行目以降により，左に 2 群別々に推定したカプラン＝マイヤプロットが描かれ，中央に維持療法の有無を共変量としてコックス回帰したベースラインの生存曲線が描かれ，右に二重対数プロットが描かれる。コックス回帰をした場合のカプラン＝マイヤプロットは，中央のグラフのように，比例ハザード性を前提として，群の違いを 1 つのパラメータに集約させ，生存関数の推定には 2 つの群の情報を両方用い，共変量の影響も調整して推定したベースラインの生存曲線を 95% 信頼区間つきで描かせるのが普通である。

どうしても共変量の影響を考えてコックス回帰したベースラインの生存曲線を 2 群別々に描きたい場合は，coxph() 関数の中で，subset=(x=="Maintained") のように指定することによって，群ごとにパラメータ推定をさせることができるが，その場合は独立変数に群分け変数を入れてはいけない。2 つ目のグラフを重ねてプロットするときは par(new=T) をしてから色や線種を変えてプロットすればいいが，信頼区間まで重ね描きすると見にくいのでお薦めしない。

コックス回帰で共変量の影響をコントロールできることの意味をもう少し説明しておく。例えば，がん患者の生存時間を分析するとき，進行度のステージ別の影響は無視できないけれども，これを調整するには，大別して 3 つの戦略がありうる。

1. ステージごとに別々に分析する。
2. 他の共変量の影響はステージを通じて共通として，ステージを層別因子として分析する
3. ステージも共変量としてモデルに取り込む

[*57] なお，最下行に Score (logrank) test とあるが，これは Rao の Score 検定の結果であり，survdiff() により実行されるログランク検定の結果ではない。

3 番目の仮定ができれば，ステージも共変量としてイベント発生への影響を定量的に評価できるメリットがあるが，そのためには，ステージが違ってもベースラインハザード関数が同じでなければならず，やや非現実的である。また，ステージをどのように共変量としてコード化するかによって結果が変わってくる（通常はダミー変数化することが多い）。2 番目の仮定は，ステージによってベースラインハザード関数が異なることを意味する。R の `coxph()` 関数で，層によって異なるベースラインハザードを想定したい場合は，`strata()` を使ってモデルを指定する。例えば，がん患者の生存時間データで，生存時間の変数が `time`，打ち切りフラグが `status`，治療方法を示す群分け変数が `x` であるときに，がんの進行度を表す変数 `stage` があったとすると，進行度によってベースラインハザード関数が異なることを想定して，治療方法によって生存時間に差が出るかどうかコックス回帰で調べたければ，`coxph(Surv(time,status)~x+strata(stage), data=aml)` とすればよい（但し aml データには stage は含まれていないので注意）。

なお，コックス回帰はモデルの当てはめなので，残差分析や尤度比検定，重相関係数の 2 乗などを用いて，よりよいモデル選択をすることができる。ただし，ベースラインハザード関数の型に特定の仮定を置かないと AIC は計算できない。

Rcmdr+RcmdrPlugin.survival でのコックス回帰は，統計量＞モデルへの適合＞ Cox regression model を選ぶ。まず “Time or start/end times” は [time]，“Event indicator” は [status]，“Strata” と “Clusters” は選択しない（“Strata” を指定すると，上述の通り，その層ごとにベースラインハザードが異なると仮定して推定してしまう）。次に，“Method for Ties” はデフォルトが [Efron] になっているが，[Breslow] や [Exact] も選べる。“Robust Standard Errors” は [Default]，[Yes]，[No] から選べる。“変数” としてハザード比を求めたいグループ変数や共変量を + でつないで指定するが，leukemia データでは “x [因子]” しか候補がないので，それをクリックすると，~の右側のボックスには [x] と入る。“部分集合の表現” の下のボックスの中を削除してから [OK] をクリックすれば，コックス回帰の結果が得られる。

11 レポート課題

http://phi.med.gunma-u.ac.jp/grad/worldfactbook2011.txt は，米国 CIA の web サイトで zip 形式に圧縮して公開されている「The world factbook 2011」[*58]から作ったタブ区切りテキスト形式のデータである。

このデータに含まれている変数は，次の通りである。

COUNTRY　国名
IMR　乳児死亡率（0 歳での死亡数/出生 1000）
LIFEEXP　ゼロ歳平均余命（いわゆる平均寿命）
TFR　合計出生率
NDAIDS　HIV/AIDS による年間死亡数
APHIVAIDS　HIV/AIDS の成人有病割合 (%)
GDPPCUSD　米ドル換算購買力平価ベース 1 人当たり GDP（国内総生産）
PUNEMP　失業者割合 (%)
GINI　国ごとの家族の所得の不平等度を示す Gini の集中係数（完全な平等のとき 0，完全な不平等で 100）

近年，社会疫学という研究分野において，健康が社会のありようによって影響を受けることが指摘されており，ゼロ歳平均余命や乳児死亡率といった健康指標が，所得の不平等度や国内総生産といった社会経済因子から受ける影響を調べることも行われている。このデータをその視点で統計解析せよ。

図示や記述統計を必ず実行してデータの性状を把握し，**検討すべき作業仮説を立て**，その仮説を検討するのに**適した分析方法**を選択し，**結果を明記**した上で**健康指標と社会経済因子の関係について考察**を展開すること。

提出期限は 6 月末日までとする。Microsoft Word，OpenOffice.org Writer，または pdf 形式で A4 サイズ 3 枚以内にまとめて，中澤准教授 (nminato@med.gunma-u.ac.jp) までメールの添付ファイルとして提出すること。

12 文献

- 青木繁伸 (2009) R による統計解析. オーム社
- 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.
- 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.
- 永田　靖 (2003) サンプルサイズの決め方. 朝倉書店.
- http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/Getting-Started-with-the-Rcmdr.pdf（作成者である John Fox 自身による R Commander の入門テキスト）
- http://www.ec.kansai-u.ac.jp/user/arakit/documents/Getting-Started-with-the-Rcmdr-ja.pdf（日本語版メニュー作成者である荒木孝治さんによる邦訳）

[*58] https://www.cia.gov/library/publications/the-world-factbook/index.html