

Rによる保健医療データ解析演習

中澤 港

rev 1.00 / 7-Nov-2007

目次

第 1 章	R の導入とデータ入力	9
1.1	統計処理ソフトの選択	9
1.2	データ入力	15
1.3	欠損値について	17
1.4	R での読み込みと基本操作	19
1.5	課題	21
第 2 章	基本的な図示	23
2.1	尺度と変数	23
2.2	名義尺度 (nominal scale)	24
2.3	順序尺度 (ordinal scale)	25
2.4	間隔尺度 (interval scale)	28
2.5	比尺度 (ratio scale)	28
2.6	データの図示の目的	29
2.7	名義尺度や順序尺度をもつ変数の図示	31
2.8	連続変数の場合	38
2.9	塗り分け地図	44
2.10	課題	47
第 3 章	記述統計量	49
3.1	データを記述する 2 つの方法	49
3.2	中心傾向 (Central Tendency)	50
3.3	ばらつき (Variability)	59
3.4	まとめ	64
3.5	課題	65

第 4 章	標本統計量と母数推定	67
4.1	標本統計量と母数	67
4.2	標本抽出	67
4.3	中心極限定理	73
4.4	信頼区間	74
4.5	自由度	76
4.6	課題	76
第 5 章	データの分布と検定	77
5.1	はじめに	77
5.2	ベルヌーイ試行と 2 項分布	77
5.3	2 項分布のシミュレーション	78
5.4	2 項分布の理論分布	79
5.5	正規分布	80
5.6	χ^2 分布	81
5.7	t 分布	81
5.8	F 分布	82
5.9	検定の考え方と第 1 種, 第 2 種の過誤	82
5.10	両側検定と片側検定	83
5.11	分布の正規性の検定	84
5.12	課題	87
第 6 章	2 群の平均値の差の検定	89
6.1	母平均と標本平均の差の検定	89
6.2	独立 2 標本の平均値の差の検定	91
6.3	対応のある 2 標本の平均値の差の検定	95
6.4	課題	97
第 7 章	一元配置分散分析と多重比較	99
7.1	多群の平均値を比較する 2 つの思想	99
7.2	一元配置分散分析	100
7.3	検定の多重性を調整する「多重比較」	103
7.4	課題	110
第 8 章	相関と回帰	111

8.1	相関と回帰の違い	111
8.2	相関	111
8.3	回帰	117
8.4	課題	128
第 9 章	計数データと比率の解析	129
9.1	母比率を推定する方法	129
9.2	推定値の確からしさ	130
9.3	母比率の信頼区間	131
9.4	カテゴリ 2 つの場合の母比率の検定	133
9.5	カテゴリが 3 つ以上ある場合の母比率の検定	134
9.6	サイコロの正しさの検定	138
9.7	2 群間の比率の差	138
9.8	3 群以上の比率の差	140
9.9	課題	142
第 10 章	クロス集計	143
10.1	複数のカテゴリ変数を分析するために	143
10.2	2 つのカテゴリ変数の独立性の検定	144
10.3	研究デザインと疫学指標	153
10.4	スクリーニングにおける ROC 分析	168
10.5	交絡を考える	173
10.6	課題	178
第 11 章	量的データのノンパラメトリックな分析	179
11.1	2 群の分布の位置の差に関するノンパラメトリックな検定	179
11.2	多群間の分布の位置の差の検定	189
11.3	課題	192
第 12 章	一般化線型モデル	193
12.1	一般化線型モデルとは?	193
12.2	モデルの記述法	194
12.3	変数の種類と数の違いによる線型モデルの分類	196
12.4	重回帰分析についての留意点	197
12.5	多重共線性 (multicollinearity)	198

12.6	モデルの評価	201
12.7	変数選択	206
12.8	採択されたモデルを使った予測	207
12.9	共分散分析	209
12.10	ロジスティック回帰分析	215
12.11	課題	220
第 13 章	生存時間解析	221
13.1	生存時間解析概論	221
13.2	カプラン = マイヤ法 : <code>survfit()</code> 関数	223
13.3	ログランク検定 : <code>survdif()</code> 関数	231
13.4	コックス回帰—比例ハザードモデル : <code>coxph()</code> 関数	233
13.5	課題	243
第 14 章	文献	245
14.1	R に関する日本語の文献	245
14.2	R に関する英語の文献	245
14.3	疫学・統計学についての文献	246
14.4	R に関するウェブサイト	246
第 15 章	【課題解答例】	247
	コマンド索引	274
	事項索引	278

はじめに ~ 本書の狙い

2003年に「Rによる統計解析の基礎」を出版してから、次々とRの解説書が出版され、わが国におけるRの普及には目覚ましいものがあった。群馬大学社会情報学部の青木繁伸教授による「Rによる統計処理」^{*1}や、筑波大学の岡田昌史講師による「RjpWiki」^{*2}のようなインターネット上のメディアの充実に加え、中間英治さんによるコーディング、RjpWikiを通して集まったボランティアの翻訳作業と、R開発コアチームの協力によって、メッセージの日本語化や2バイト文字対応にもかなりの進展があった。この数年で、きわめて使いやすいツールになったといえるだろう。

筆者自身の経験でも、2006年には統計数理研究所や計量生物学会のチュートリアルセミナーなど、Rについて解説する機会を何度ももったことから考えると、この優れたデータ解析環境は、かなり広まってきたと思う。ただ、これまで、医学データ解析にフォーカスした日本語の本は、Peter Dalgaardの“Introductory Statistics with R”を岡田昌史さんたちが訳された「Rによる医療統計学」(2007年、丸善)しかないと言っていい状況であった。本書は、「Rによる統計解析の基礎」をベースにしているが、群馬大学医学部で2年生を対象に行った「医学情報処理演習」のために、保健医療データを使った例題を自力で解く形が中心になるように、抜本的に書き換えたものである。

「Rによる統計解析の基礎」は統計学の考え方や理屈の説明に主眼があり、実際の計算手段としてRを使うやり方をいくつか例示したものだだったが、本書はむしろ、考え方や理屈は最小限にとどめ、保健医療分野のデータを相手にしたときのRの使い方と結果の読み方、まとめ方に主眼をおいた。実際にRを使う際に役に立つようにするため、細かくコマンド索引をつけた^{*3}。

*1 <http://aoki2.si.gunma-u.ac.jp/R/>

*2 <http://www.okada.jp.org/RWiki/>

*3 巻末参照。なお、コマンド索引中、()内はそのコマンド定義が含まれる、標準でロードされていないライブラリ名を示す。

本書の作成には R はもちろんだが、その他にもさまざまなフリーソフトウェアを用いた。R で作成した図を加工して EPS 形式で出力するために、OpenOffice.org^{*4}を利用した。加工しない場合は、R から直接 pdf 形式で出力し、Ghostscript^{*5}を使って EPS 形式に変換した。組版には p^LA^TE_X2e といくつかのスタイルファイル (jsbook, tascmac, fancybox, dvipdfm, makeidx) を利用した。p^LA^TE_X2e 関係の情報は、三重大学・奥村晴彦教授^{*6}が管理運営されている TeX Wiki^{*7}から入手した。ここに記して御礼申し上げる。

末筆ながら演習を通じてコメントをくださった同僚諸氏ならびに学生諸君、web で公開した草稿に対してコメントくださった方々に感謝申し上げます。もちろん、本書の内容に間違いがあれば、それは著者個人の責任である。

2007 年 11 月 7 日 中澤 港

*4 日本ユーザ会 (<http://ja.openoffice.org/>) のサイトを見れば一通り使えるようになると思う。

*5 <http://www.ring.gr.jp/pub/text/TeX/ptex-win32/gs/> から日本語対応版がダウンロードできる。

*6 jsbook を開発・公開されている方でもある。

*7 <http://oku.edu.mie-u.ac.jp/~okumura/texwiki/>

第 1 章

R の導入とデータ入力

1.1 統計処理ソフトの選択

保健医療分野で扱うデータは、実験、臨床、調査などで直接得たり、官庁統計などの二次資料から得たりするので、信頼性も量もさまざまである。しかし、いずれにせよ、その統計的な処理はほとんどコンピュータを使って行われるので、まずは、適切なソフトウェアを選択する必要がある。しかし、ソフトウェアは無数にあって、どれを使うのが適当なのかわからないという人も多いと思われる。

1つの方針としては、所属する研究室や身の回りで使っている人が多いソフトウェアを使うというのは合理的である。データを共用することもできるし、わからなくなったときに、すぐに誰かに尋ねることができる。この方法の欠点は、研究室を移ったときにそのメリットが失われることと、何か新しいことにチャレンジする場合に（いうまでもないが、論文を書くためには、研究のどこかが新しくなくてはいけない）、自分でやり方を探索せねばならず、上述のメリットがあまりないことである^{*1}。

では、とくにそういう制約がないとしたら、どういうソフトウェアがいいだろうか。下表は、国際的によく使われているソフトウェアを比較したものである。Rをお勧めする理由は明白であろう。

^{*1} なお、Mac OS で動作する StatView というソフトは、簡便な操作性もあって臨床系で良く使われてきたが、SAS 社に買収された後、開発・販売が停止されるに至り、今後サポートが期待できないので、これから統計解析を始める人は使わない方が無難であろう。

比較項目	SAS	SPSS	JMP	Excel	EpiInfo	R
メニュー操作						*
プログラム実行						
統計手法				×		
作図能力						
信頼性						
価格	×				**	**
動作 OS					×	
解説書						

* R Commander (Rcmdr) という無料のライブラリがカナダ・McMaster 大学の John Fox 教授により開発されており、それをインストールすれば、かなりの動作がメニューから操作できる。メニュー自体はテキストファイルとして別に用意されていたので、当初有志によって日本語訳が作られたが、現在では、関西大学の荒木孝治教授^{*2}の貢献により国際化版になっているので、日本語によるメニュー操作が可能である。

** 完全に無料。

1.1.1 フリーソフト利用上の心得

R のようなオープンソースのフリーソフトウェアの開発はボランティアベースで行われている。利用者もただ利用するだけではなく、開発に参加することで、R ユーザ共同体全体の利益に貢献することが望ましい。ソースコードが書けるとか、メッセージの翻訳などで貢献できれば最高だが、利用者としてのフィードバックをするだけでも十分に役に立つ。こうしたオープンソースのフリーソフトの信頼性は、多くのユーザが世界中で使って、ちゃんと使えているという事実によって担保されるからである。

フィードバックをする際には、相応のマナーがある。R の場合、本体の開発のコアチームは r-devel というメーリングリストで連絡を取りながら開発しているが、一般ユーザからの質問やバグレポートは r-help^{*3} というメーリングリストになされるのが普通である（高度にテクニカルな内容であれば、直接 r-devel に投稿してもよい）。その際、以下の点に注意すべきである。

1. R 本体及び使用ライブラリは最新版にする。
2. ?によって関数ヘルプ（英語）を出して熟読する。
3. `help.search("keyword")` や `RSiteSearch("keyword")` を使ってヘルプ

^{*2} <http://www.ec.kansai-u.ac.jp/user/arakit/R.html> を参照されたい。

^{*3} <https://www.stat.math.ethz.ch/mailman/listinfo/r-help>

- (英語)を見ても解決しないか確かめる。
4. 既に解決しているかもしれないので、最新の NEWS^{*4} (英語)を見ておく。
 5. FAQ^{*5} (英語)を見ておく。
 6. 公式入門マニュアル “An Introduction to R”^{*6} (英語)の関連箇所を読んでおく。
 7. MASS など、本の付録的なライブラリの場合は、その本を読んでおく。
 8. 内容を再現確認できるだけのコードやデータや使用ライブラリや使用 OS, 言語口ケールを明記して、質問文やバグレポートを書く。
 9. 以上の条件をすべて満たした上で r-help に簡潔な英語で投稿する。html メールは不可なので、必ずテキスト形式で投稿すること。pdf 以外のバイナリは添付しないこと。巨大なデータがないと再現確認ができない事例では、web 上にデータファイルを置いて URL のみを投稿すること (もし公開できないデータなら、同じ問題を起こし、かつ公開可能なサンプルデータを自力で作って公開すること)。できれば所属を含む短い署名をつけること。メールがスレッドで管理されているので、新しい話題を投稿するときは、返信でなくて、メールを新規作成すること。

この手順をきちんと踏んでいれば、おそらく投稿後 1 時間もしないうちに、世界のどこかから返事が届くだろう。コアチームの一員で通称 R 教授こと Prof. Brian Ripley の実に正しいけれども辛口で簡潔すぎるコメントが届いて凹むこともあるかもしれない。しかし、たいていの場合、何らかの形で問題が解決する。ごく稀に、統計的な思想上の理由から、開発コアチームが仕様として問題を受け付けられないこともあるが、そういう場合は代替的なライブラリや関数が見つかることが多い^{*7}。

英語が苦手な方は、RjpWiki^{*8}の Q & A (初級者コース)などに書き込んでみることもできると思うが、その場合でも、R 本体やライブラリを最新版にすることと、過去の書き込みをチェックして同一事例がないかどうかをチェックすること、書き込み

*4 CRAN ミラーの `src/base/NEWS`

*5 CRAN ミラーの `faqs.html`

*6 CRAN ミラーの `doc/manuals/R-intro.pdf`

*7 例えば、カイ二乗検定の関数 `chisq.test()` における Yates の連続性の補正は、`ad-bc` の絶対値が $N/2$ より小さくてもなされてしまう。これは Yates の元論文がそうになっているから、というコアチームの意思による。しかし、`prop.test()` における Yates の補正は、多くの統計ソフトが採用している基準である「`ad-bc` の絶対値が $N/2$ より小さい場合はカイ二乗値をゼロとする」になっている。

*8 <http://www.okada.jp.org/RWiki/>

を読んだ人が手元でその問題を再現できるだけの情報を過不足なく提供することは必須である。ボランティアベースで開発されている以上、皆の時間を無駄に使わせないための配慮は礼儀であろう。それを踏まえてこそコミュニティ皆が幸せになれるというものである。

1.1.2 Rの動作環境とインストール

RはMS Windows, Mac OS, Linuxなど、さまざまなOSで動作する。MS Windowsでは、まだ32 bit環境でしか動作しない。Linuxではtarで圧縮されたソースコードをダウンロードして、自分でコンパイルすることも珍しくないが、Vine Linuxなどでは容易にインストールできるようにコンパイル済みのバイナリを提供してくれている人もいる。R関連のソフトウェアはCRAN (The Comprehensive R Archive Network) からダウンロードすることができる。CRANのミラーサイトが各国に存在するので、ダウンロードは国内のミラーサイトからすることが推奨されている。日本では会津大学*9, 筑波大学*10, 東京大学*11のどれかを利用すべきだろう。

Windows CRANミラーからR-2.6.0のインストール用ファイル*12をダウンロードし、ダブルクリックして実行し、適当に問いあわせに答えるだけでインストールは完了する。起動して日本語メッセージが文字化けしていたら、(1) いったん上部メニューバーの「編集」の「GUIプリファレンス」を開いて、表示フォントをCourierからMSゴシックに変更して「反映」と「保存」をクリックするか、(2) 日本語表示用の設定が書かれた環境設定ファイルをコピーすれば*13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-2.6.0に対応しているOSは、Mac OS X 10.4 (Tiger)以降である。同じくCRANミラーからR-2.6.0.dmgをダウンロードしてダブルクリックし、できたフォルダ内のR.mpkgをダブルクリックしてインストールすればよい。群馬大学社会情報学部・青木繁伸教授のサイトに詳細な解説記事*14があるので参照されたい。

*9 <ftp://ftp.u-aizu.ac.jp/pub/lang/R/CRAN/>

*10 <http://cran.md.tsukuba.ac.jp/>

*11 <http://ftp.ecc.u-tokyo.ac.jp/CRAN/>

*12 R-2.6.0-win32.exe

*13 <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。

*14 <http://aoki2.si.gunma-u.ac.jp/R/begin.html>

Linux Debian, RedHat/Fedora Core, Vine など、メジャーなディストリビューションについては有志がコンパイルしたバイナリが CRAN にアップロードされているので、それを利用すればインストールは容易であろう。マイナーな環境の場合や、高速な数値演算ライブラリを使うなど自分のマシンに最適化したビルドをしたい場合は、CRAN からソース R-2.6.0.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。

1.1.3 R の使い方の基本

以下の解説は Windows 版による。基本的に Linux 版でも Mac OS X 版でも大差ないが、使えるデバイスなどが多少異なるので、適宜読み替えられたい。なお、以下の本文中、\記号は¥の半角と同じものを意味する。

Windows では、インストールが完了すると、デスクトップに R のアイコンができています。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*15。ウィンドウが開き、作業ディレクトリの .Rprofile に書かれた内容が実行され、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て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う場合、**[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

*15 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うい。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、システム的环境変数または作業ディレクトリにテキストファイル .Renviron を置き、その中に R_USER="c:/work"などと書いておくと、そちらが優先される。Windows のファイルパス記述におけるディレクトリ (フォルダ) 区切り記号は通常\だが、R の中では\とするか、Linux と同じ / を用いる。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」末尾にスペースを空けて --internet2 と付しておく。

「ファイル」の「Rコードのソースを読み込み」(英語版では“Source”となっている)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(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)。また、「上向き矢印キー」で既に入力したコマンドを呼び戻すことができる。

なお、Rをインストールしたディレクトリの bin にパスを通しておけば、Windows 2000/XP のコマンドプロンプトで R と打っても、R を起動することができる。この場合は、コマンドプロンプトが R コンソールの代わりにシェルとして動作する。

1.1.4 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)) }
```

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(中括弧の中という意味)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。

導入 install.packages() 例えば、CRAN から vcd をダウンロードしてインストールするには、

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

とする。dep=TRUE は dependency (依存) が真という意味で、vcd が依存している、vcd 以外のライブラリも自動的にダウンロードしてインストールしてくれる。なお、TRUE は T でも有効だが、誤って T を変数として別の値を付値してしまっていると、意図しない動作をしてしまい、原因を見つけにくいバグの元になるので、できるだけ TRUE とフルスペル書いておくことが推奨されて

いる。ただし本書では、紙幅の都合上、大抵のところでは、TRUE/FALSE の代わりに T/F で済ませている。

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。

なお、世界中の研究者が GIS を含む空間統計解析やゲノム解析などに至るまでさまざまな追加ライブラリを公開しているため、R 本体に含まれていなくても、CRAN 内で検索すればたいがい解析法は見つかる。もしなければ、自分で新しい拡張関数やライブラリを作って公開することもできる。本書でもいくつかの汎用関数定義をしている^{*16}。

ちなみに、R バイナリに組み込まれていて、R 本体を起動するだけで自動的にロードされるライブラリは、`base`, `datasets`, `grDevices`, `graphics`, `grid`, `methods`, `splines`, `stats`, `stats4`, `tcltk`, `tools`, `utils` であり、推奨ライブラリで、将来全バイナリに組み込まれる予定なのは、`KernSmooth`, `MASS`, `boot`, `class`, `cluster`, `foreign`, `lattice`, `mgcv`, `nlme`, `nnet`, `rpart`, `spatial`, `survival` である。これらは、Windows 版バイナリには入っていてインストール済みだがロードはされていないので、使うときは `library(survival)`、あるいは `require(survival)` のようにしてロードせねばならない。`search()` でロード済みライブラリ一覧、`.packages(all.avail=T)` でインストール済み一覧が表示される。ロード済みのライブラリをアンロードするには `detach(package:survival)` などとする。

1.2 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。

さらに、入力以前に、エディティングとコーディングがきちんとできていないと、いくら正しく入力しても意味がない。エディティングとは、生データをエラーチェックなど精査して、回答そのものをチェックし、コーディングできるようにする過程であり、コーディングとは、調査や実験によって得た生のデータを、どういう形の変数

^{*16} `source("http://phi.med.gunma-u.ac.jp/msb/msb-funcs.R")` を実行すれば使えるようになる。なお、標準的な日本語版の Windows 環境ではそのまま大丈夫なはずだが、MacOS や Linux 上で日本語文字コードに問題が生じた場合は、直接 `source()` で読み込まず、ブラウザで開いて文字コードを SJIS にして文字化けを解消してコピーし、スクリプトエディタ等にペーストしてから実行すればよい。

として保存するかを決めた一定の規則を定めることである。厳密にやる場合は、コード表（英語では coding sheet といい、調査票上の回答カテゴリーや分析機器からプリントアウトされた生の数値と、入力すべきデータとの項目ごとの対応表のこと）をつくり、データ形式も記載することが多い。コード表に基づいて調査票やプリントアウト上にコードを振ってから、初めてデータ入力ができることになる。

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 人の平均体重を求めるには、Microsoft Excel では、1 つのセルの中に=AVERAGE(60,66,75) とか=(60+66+75)/3 と打てばいいし、R ならば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。分析には R を使うとした場合、同じ調査を繰り返すとか、きわめて大きなデータであるとかでなければ、Microsoft Excel のような表計算ソフト*17で入力するのが手軽であろう*18。

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

	A	B	C
1	PID	HT	WT
2	1	170	70
3	2	172	80
4	3	166	72
5	4	170	75
6	5	174	55
7	6	199	92
8	7	168	80
9	8	183	78
10	9	177	87
11	10	185	100

*17 Excel 以外には、例えばフリーソフトとして公開されている OpenOffice.org の calc は、Excel 形式のファイルを読み書きでき、概ね同じような感覚で使える表計算ソフトである。

*18 大きなデータや繰り返し調査などの場合は、html でフォームを書き、httpd サーバソフトを動作させ、cgi を使ってデータ入力するとか、ACCESS などのデータベースソフトを使って入力フォームを設計して入力する方が間違いがないし効率も良い。前者については <http://phi.med.gunma-u.ac.jp/swtips/webdb.html> も参照されたいが、本当に大規模なデータ入力をしなくてはいけない場合は、専門家に相談することをお薦めする。

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。バージョン 2 以降の R はマルチバイト文字にも対応しているので、漢字やカタカナ、ひらがなも使えるが、変数名は半角英数字（半角ピリオドも使える）にしておくのが無難である。アルファベットの大文字と小文字は区別されることに注意されたい。必要なら別にラベルやコメントをつけることができる。この例の場合、対象者 ID を PID、身長を HT、体重を WT とするといいい。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。Windows ではファイルの種類が拡張子によって決まっていて、Microsoft Excel 標準形式のファイルには xls という拡張子がつくので、例えば desample.xls というファイル名になる（前掲の表右を参照）。

次に、R で使用するために、この表をタブ区切りテキスト形式で保存する^{*19}。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^{*20}、「保存 (S)」ボタンを押せば OK である。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 Microsoft 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（もっとも、「はい」を選んでも同じ内容が上書きされるだけなので問題はない）。この例では、desample.txt ができる。

1.3 欠損値について

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

^{*19} 実は、<http://treetron.googlepages.com/> で公開されていて CRAN にも登録されている xlsReadWrite ライブラリに含まれている read.xls() 関数を使えば、タブ区切りテキスト形式になくても Microsoft Excel 97 以降 2003 までの *.xls ファイルを直接データフレームに読み込むことができる。

^{*20} Windows では初期設定のままでは登録されたファイル形式の拡張子は隠されるが、間違いなくファイルを読み書きするために拡張子を常に表示する設定にしておくことをお勧めする。XP の場合なら、予めエクスプローラのフォルダオプションの「表示」タブの「登録されている拡張子は表示しない」のチェックを外しておく。

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。欠損値が少なければあまり気にしないでいいが、例えば、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と思って「はい」とは答えにくく、さりとして嘘はつきたくないので無回答にしたという可能性があり、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(No Answer)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(もちろん、多くのソフトで、欠損値を表すコードの方を変更することも可能)、それに合わせておくのも1つの方法である。デフォルトの欠損値記号は、RならNA、SASなら.(半角ピリオド)である。Microsoft Excelでは空白(何も入力しない)にしておくのと欠損値として扱われる。ただし、入力段階で欠損値を空白にしておくと、「入力し忘れたのか欠損値なのか」が区別できない」という問題を生じるので、入力段階では決まった記号を入力しておいた方がよい。その上で、もし簡単な分析までMicrosoft 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(Microsoft 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)、欠損値記号を空白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(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(表計算ソフト上の1行)を削除してしまうのが簡単である(通常、元データは別名で保存しておいて、コピー上で行削除する)。質問紙調査の場合、例えば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80%程度は欲しい。

1.4 Rでの読み込みと基本操作

あとはRで読み込めばいい。この例のように、複数の変数を含む変数名付きのデータを読み込むときは、データフレームという構造（複数の型の異なる変数をまとめ、1つの塊として扱うための特別なリスト構造）に付値するのが普通である。保存済みのタブ区切りテキスト形式データが作業ディレクトリにあって `desample.txt` というファイル名だとすれば、Rのプロンプトに対して、

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

と打てば、`dat` というデータフレームにデータが付値される^{*21}。確認のためにデータを表示させたい場合は、ただ `dat` と打てばいいし、データ構造を見たければ、`str(dat)` とすればよい^{*22}。数値型（`numeric`、ただしこの場合はすべて整数値なので整数型 `integer` になっている）の変数として `PID`、`HT`、`WT` の3つが読み込まれたことがわかる。

Rの変数の型には、この他に、カテゴリ変数に対応する要因型（`factor`）、順序を表す順序型（`ordered`）、真（`TRUE`）か偽（`FALSE`）を示す論理型（`logical`）などがあり、型によって適用可能な分析方法が違ってくる。数値型変数を無理やり順序扱いか要因扱いはすることは可能だが、その逆は一般に不適切である。要因型変数に対して数値型変数にしか使えないような分析法を使いたいときは、ダミー変数化という方法を使うことができる。

また、`summary(dat)` とすれば、この `dat` というデータフレームに含まれるすべての変数について、型に応じて適切な要約をしてくれる。数値型の変数については、欠損値の個数も `NA's` として表示される^{*23}。

読み込まれた変数に対して分析したいとき、例えばこの例の身長の平均値と標準偏差を求めたいければ、分析したい変数名の前にデータフレーム名と`$`をつけて、

^{*21} 外部データファイルは、作業ディレクトリにおいてあれば、この例のようにファイル名だけ指定すればいいが、フルパスで指定すればコンピュータ内のどこにあってもいいし、ネットワークドライブでもいいし、URLで指定すればインターネット上のものでも読み込める。

^{*22} `str` は `structure`（構造）の略であり、もちろん `structure(dat)` と打つても有効である。

^{*23} なお、1つでも欠損値がある人はデータから取り除いて処理をしたい場合は、`dat2 <- subset(dat, complete.cases(dat))` とすることにより、元のデータフレーム `dat` に対して、欠損値を含まないサブセット `dat2` ができるので、以後 `dat2` について分析すればよい。

```
cat("mean=", mean(dat$HT), "sd=", sd(dat$HT), "\n")
```

とすればよい*24。いちいち `dat$`と打つのが面倒ならば、`attach(dat)` とすれば、それ以降のセッション中、`detach(dat)` するまで、`dat$`を入力しなくても良くなる。例えば、このデータで身長と体重の相関係数を出して検定したいときは次のようにする。

```
attach(dat)
cor.test(HT, WT)
detach(dat)
```

この例のようにサイズが小さなデータで、何度も使うわけではなく、その場で計算させれば終わりという場合、わざわざファイルを作らなくても、Microsoft Excel の画面上で必要なデータ範囲を選択し、コピーしておいてから、R のコンソールに移って `dat <- read.delim("clipboard")` とすることで読み込み可能である。このやり方は Microsoft Excel に限らず、web 上の表 (table タグで作られているもの) をブラウザで選択・コピーする場合にも概ね通用する。

なお、R コマンドー (Rcmdr) を使って、次のようにメニュー形式でデータファイルを読み込むこともできる。他のライブラリを使えるようにするときも基本的に同じだが、R コマンドーを呼び出すには次の枠内を打つ。

```
library(Rcmdr)
```

この状態から、先ほど保存した `desample.txt` を読み込むためには、メニューバーの Data (日本語版では「データ」) から Import Data (日本語版では「データのインポート」) の From Text File (日本語版では「テキストファイルまたはクリップボードから」) を開いて、Enter name for data set: (日本語版では「データセット名を入力:」) の欄に適当な参照名をつけ (先の例では `dat` となっていたもの。変数名として使える文字列なら何でもよいのだが、Rcmdr のデフォルトでは Dataset となって

*24 変数名指定は、そのデータフレーム名の中でユニークになるところまでいい。つまり、この例だと、`dat$HT` は `dat$H` で事足りる。しかし、間違わないようにするため、省略しない方がいいと思う。参考までに書いておくと、`$`を使う代わりに、`dat[["HT"]]` としてもいいし、`dat[, "HT"]` としても同じことである。この参照方法の場合、"HT"は変数の中からそれにマッチする順番を探して返すので、変数の出現順序がわかっていれば数字でもいい。このデータフレームでは HT という変数が 2 番目に出てくるので、`dat[[2]]` や `dat[,2]` も同じ意味になる。

いる), Field Separator (日本語版では「フィールドの区切り記号」) を White space (日本語版では「空白」) から Tabs (日本語版では「タブ」) に変えて (Tabs の右にある をクリックすればよい), OK ボタンをクリックすればよい。後は Rcmdr のメニューから選んでいくだけで、いろいろな分析ができる。

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

1.5 課題

<http://phi.med.gunma-u.ac.jp/msb/data/p01.xls> を Microsoft Excel^{*25} に読み込み、タブ区切りテキストファイル経由またはクリップボード経由のどちらでもいいので R に読み込んで、このデータフレームに含まれるすべての変数について、変数名、型、有効なサンプルサイズを求めよ。

*25 OpenOffice.org の calc でもよい。

第 2 章

基本的な図示

2.1 尺度と変数

本章の習得目標は、入力済みのデータについて、適切な図示を行うことであるが、適切な図示の方法は、データの性質によって変わってくる。そのため、図示の方法に先立って、尺度と変数についてざっとさらしておく。

尺度とは、研究対象として取り上げる操作的概念を数値として扱うときのモノサシの目盛り（の種類）、言い換えると、「データに何らかの値を対応させる基準」である。尺度は、名義尺度、順序尺度、間隔尺度、比尺度（比例尺度ともいう）の 4 つに分類される。

変数とは、モノサシで測定された値につける名前である。変数は、それが表す尺度の水準によって分類されるが、一般には、定性的変数（カテゴリ変数）は名義尺度をもち、定量的変数は順序尺度、間隔尺度、または比尺度をもち、定量的変数には、整数値しかとらない離散変数と、実数値をとりうる連続変数がある。順序尺度は離散変数である。

前章で説明したように、R の変数の型には、integer (int) で表される整数型、numeric (num) で表される数値型、factor で表される要因型、character で表される文字列型などがある。整数型の変数は離散変数であり、数値型の変数は連続変数である*1。要因型や文字列型の変数はカテゴリ変数である*2。同じ関数でも、変数の型によって動作が異なる場合が多いので、変数の型（及びその変換）については注意が必

*1 as.numeric() を使えば、離散変数を数値型扱いすることは可能である。しかし、整数でない実数に対して as.integer() を用いて整数型にすると、小数点以下が切り捨てられて値が変わってしまう。

*2 as.ordered() を使って順序型にすることもできる。

要である。

順序尺度は離散変数、間隔尺度は離散の場合も連続の場合もあるが連続変数であることが多く、比尺度は連続変数である。定性的変数と離散変数の中には、「1か0,あるいは1か2,のように、2種類の値しかとらない」「2分変数 (dichotomous variable) 」*3や、「AかBかC,のように3種類の値しかとらない」「3分変数 (trichotomous variable) 」がある。変数がとり得る値の範囲を、その変数の定義域と呼ぶ。

変数は、被験者や研究対象のちがいによって、複数の異なったカテゴリあるいは数値に分かれるのでなければ意味がない。例えば、その研究のすべての対象者が男性であれば、性別という変数を作ることは無意味である*4。

対応する尺度の種類によって、変数は、図示の仕方も違おうし、代表値も違おうし、適用できる統計解析手法も違ってくる。ここでは簡単にまとめるが、より詳しく知りたい方は、池田央『調査と測定』(新曜社)等の専門書を参照されたい。

2.2 名義尺度 (nominal scale)

- 値の差も値の順序も意味をもたず、質的データの分類基準を与える。
- 例えば、性別とか職業とか居住地とか病名は、名義尺度をもつカテゴリ変数である。
- 変数の型としては、文字列型が要因型になる。
- 性別というカテゴリ変数は、例えば、男性なら“M”, 女性なら“F” という具合に文字列値をとることもできるが、一般には男性なら1, 女性なら2というように、数値を対応させる。これは、前回触れたとおり、コーディング (coding) と呼ばれる手続きである。
- 関心のある事象が、例えば血液中のヘモグロビン濃度のように、性別ばかりでなく、授乳や妊娠によって影響を受ける場合は、調査対象者を、男性なら1, 授乳も妊娠もしていない女性は2, 授乳中の女性は3, 妊娠中の女性は4, という具合に、生殖状態 (性別及び授乳, 妊娠) という名義尺度をあらわす変数にコード化する場合もある。
- 名義尺度を表す値にはそれを他の値と識別する意味しかない。統計解析では、カテゴリごとの度数を求めたり、クロス集計表を作って解析するほかには、グループ分けや層別化に用いられるのが普通である。なお、3つ以上のカテゴリ

*3 2 値変数と呼ぶこともある。

*4 ただし、後に別のデータと併合することを考えて、敢えて作っておく場合もある。

をもつ変数を、より複雑な統計解析に使う場合は、ダミー変数として値ごとの有無を示す複数の2分変数群に変換することもある。例えば、人種という変数の定義域が{白人, 黒人, ヒスパニック, アジア系}であれば、この変数の尺度は名義尺度である。白人を1, 黒人を2, ヒスパニックを3, アジア系を4と数値を割り振っても、名義尺度であるには違いない。しかし、人種という変数を無くして、代わりに、白人か、黒人か、ヒスパニックかという、それぞれが0/1で表される3つのダミー変数を導入することによって、同じ情報を表現することができる。例えば、白人ならば、白人という変数の値が1となり、黒人という変数とヒスパニックという変数の値はともに0となる。ダミー変数は、平均値をとると、「1に当てはまるケースの割合」と一致するため、本来なら量的な変数にしか使えないような多くの統計手法の対象になりうる。

2.3 順序尺度 (ordinal scale)

- 値の差には意味がないが、値の順序には意味があるような尺度。
- 変数の型は、順序型 (Ord.factor) になる。Rでは、読んだだけで順序型と自動判定されることはないので、順序型変数を用いたい場合は、数値型または整数型としてデータ入力しておき、`as.ordered()` を使って型変換する。
- 例えば、尿検査での潜血の程度について +++ , ++ , + , ± , - で表される尺度は、+ の数を数値として、例えば 3, 2, 1, 0.5, 0 とコーディングしても、3と2の差と2と1の差が等しいわけではなく、3は2よりも潜血が高濃度に検出され、2は1よりも高濃度だという順序にしか意味がないから、順序尺度である。3, 2, 1, 0.5, 0 とコーディングしておき、`as.ordered(o.blood)` のようにして型変換すべきである。
- 順序尺度を表す値は、順序の情報だけに意味があるので、変数の定義域が 3, 2, 1 であろうと、15, 3.14159265358979, 1 であろうと同じ意味をもつ。しかし、意味が同じなら単純な方がいいので、1から連続した整数値を割り当てて、順位そのものを定義域にするのが通例である（上の例のように元データに近い形にすることもある）。同順位がある場合の扱いも何通りか提案されている。
- 注意しなければならないのは、本来は順序尺度であっても、もっともらしい仮定を導入して得点化し、間隔尺度であるとみなす場合も多い、ということである。例えば「まったくその通り」「まあそう思う」「どちらともいえない」「たぶん違うと思う」「絶対に違う」の 5, 4, 3, 2, 1 などは本来は順序尺度だが、等間

隔な得点として扱われる場合が多い。質問紙調査などで、いくつかの質問から得られるこのような得点の合計によって何らかの傾向を表す合成得点を得ることが頻繁に行われるが、得点を合計する、という操作は各質問への回答がすべて等間隔であり、変数ごとの重みも等しいという仮定を置いているわけである（たとえ調査者が意識していなくても、尺度構成をしていることになる）。合成得点は、通常、間隔尺度扱われる。合成得点が示す尺度の信頼性を調べるためにクロンバック (Cronbach) の α 係数^{*5} という統計量がよく使われるが、 α 係数の計算に分散が使われていることから、それが間隔尺度扱われていることがわかる。

2.3.1 Cronbach の α 係数

質問紙調査においては、多くの概念は直接聞き取ることができないので、複数の質問を組み合わせることによって対象者の差異をより細かく把握しようと試みることがある。回答が同じように変動していれば、それらの質問によって同じ上位概念を聞き取れている信頼性が高いと考え、それを指標化したものが Cronbach の α 係数である。

例えば、自然への親近感を聞き取りたい場合に、

- (1) あなたは自然が好きですか？ 嫌いですか？
(好き、どちらかといえば好き、どちらかといえば嫌い、嫌い)

だけでは対象者は4群にしか分かれな（順序尺度として数値化すると、好きを4点、嫌いを1点として1点から4点の4段階）。しかし、

- (2) 休日に海や山で過ごすのと映画館や遊園地で遊ぶのとどちらが好きですか？
(海や山、どちらかといえば海や山、どちらかといえば映画館や遊園地、映画館や遊園地)

を加えて、これも「海や山」を4点、「映画館や遊園地」を1点とする順序尺度として扱うことにすれば、(1)と(2)の回答の合計点を計算すると、2点から8点まで

^{*5} 次節で説明する。統計学では α という記号は有意水準、回帰分析の切片など、いろいろな意味で使われることがあるので、この統計量を指すときは「クロンバックの」をつけるべきである。

の7群に回答者が類別される可能性があり、より細かい把握が可能になる。さらに、

- (3) 無人のジャングルで野生生物の観察をする仕事に魅力を感じますか？
それとも感じませんか？
(感じる、どちらかといえば感じる、どちらかといえば感じない、感じない)

の4点を加えると、3点から12点までの10段階になる。この合計得点を「自然への親近感」を表す尺度として考えてみると、3つの項目は同じ概念を構成する項目(下位概念)として聞き取られているので、互いに回答が同じ傾向になることが期待される。つまり(1)で好きと答えた人なら、(2)では海や山と答える人が多いだろうし、(3)では感じないと答えるよりも感じると答える人が多いと考える。同じ概念を構成する質問に対して同じ傾向の回答が得られれば、その合計得点によって示される尺度は、信頼性が高いと考えられる。

上記3つの質問に対して一貫した答えが得られたかどうかを調べる方法の1つに折半法がある。例えば質問(1)と(3)の合計点の変数 x_{13} と質問(2)の点の変数 x_2 という具合に、同じ概念を構成する全質問を2つにわけて、 x_{13} と x_2 の相関係数を $r_{x_{13}x_2}$ とすれば、これらの質問の信頼性係数 $\alpha_{x_{13}x_2}$ は、

$$\alpha_{x_{13}x_2} = \frac{2r_{x_{13}x_2}}{1 + r_{x_{13}x_2}}$$

となるというのがスピアマン・ブラウンの公式である。

折半法では通常、奇数番目の項目と偶数番目の項目に二分するが、(1)の点と(2)と(3)の合計点という分け方もあるわけで、下位概念が3つ以上ある質問だったら、これらの回答に一貫して同じ傾向があるかどうかをスピアマン・ブラウンの公式で出そうと思うと、 α の値はいくつもの(n 個の下位概念からなるなら、 n 項目を2つに分ける組み合わせの数だけ)できる。この例では、 $\alpha_{x_1x_{23}}$ 、 $\alpha_{x_{12}x_3}$ も計算する必要がある。

それをまとめてしまおうというのが Cronbach の α 係数で、仮に(1)(2)(3)の合計得点が「自然への親近感」を表す変数 x_t だとして、(1)(2)(3)の得点をそれぞれ変数 x_1 、 x_2 、 x_3 とすれば、Cronbach の α 係数は、

$$\alpha = \frac{3}{3-1} \left(1 - \frac{s_{x_1}^2 + s_{x_2}^2 + s_{x_3}^2}{s_{x_t}^2} \right)$$

となる(s_{x_1} は x_1 の不偏標準偏差である。以下同様)。 α 係数が0.8以上なら十分な、0.7でもまあまあの、内的一貫性(信頼性)がその項目群にはあるとみなされる。

X, Y, Z が同じ概念の下位尺度となるスコアの変数だとして, α 係数を計算するための R のコードを次の枠内に示す。1 行の中に複数の文を入れる場合は, このように ; (セミコロン) で区切れればよい。

```
T <- X+Y+Z
VX <- var(X); VY <- var(Y); VZ <- var(Z); VT <- var(T)
alpha <- (3/2)*(1-(VX+VY+VZ)/VT)
print(alpha)
```

2.4 間隔尺度 (interval scale)

- 値の差に意味があるが, ゼロに意味がない尺度。^{*6}
- 変数の型は数値型か整数型である。
- 例えば, 体温は間隔尺度である。体温が摂氏 39 度であることは, 摂氏 36 度に比べて「平熱より 3 度高い」という意味をもつが, $39/36$ を計算して 1.083 倍といっても意味がない。
- 間隔尺度をもつ変数に対しては, 平均や相関など, かなり多くの統計手法が適用できるが, 意味をもたない統計量もある。例えば, 標準偏差を平均値で割った値を % 表示したものを変動係数というが, 身長という変数でも, 普通に cm 単位や m 単位やフィート単位で表した比尺度なら変動係数に意味があるが, 100 cm を基準とした cm 単位や, 170 cm を基準とした 2 cm 単位のように間隔尺度にしてしまった場合の変動係数には意味がない。変動係数は, 分布の位置に対する分布のばらつきの相対的な大きさを意味するので, 分布の位置がゼロに対して固定されていないと意味がなくなってしまうのである。

2.5 比尺度 (ratio scale)

- 値の差に意味があり, かつゼロに意味がある尺度。^{*7}
- 変数の型は数値型または整数型であるが, 数値型としておくべきである。
- 例えば, cm 単位で表した身長とか, kg 単位で表した体重といったものは, 比

^{*6} より正確に言えば, 値の比に意味がない尺度ということになる。ただし, 値の差の比には意味がある。

^{*7} より正確に言えば, 値の比にも意味がある尺度ということになる。

尺度である。予算額といったものも、0 円に意味がある以上、比尺度である。ただし、予算額には 0 円やマイナスが普通にありえるし、何 % 成長とか何 % 削減という扱いより絶対値の増減が問題にされる場合が多いので、間隔尺度とすべきという見方もある。

2.6 データの図示の目的

データの図示の目的は大別して 2 つある。1 つは見せるためであり、もう 1 つは考えるためである。もちろん、両者の機能を併せもつグラフも存在するが、重視すべきポイントが変わってくるので、一般には、この 2 つは別のグラフになる。

見せるためのグラフでも、プレゼンテーションやポスターに使うグラフと、投稿論文に載せるグラフは、一般に別物である。前者は、1 つのグラフに 1 つのことだけを語らせる必要があり、とにかくわかりやすさが最大のポイントであるのに対して、後者は複数の内容を語らせることも可能である。これは、見る人が 1 枚のグラフを見るために使える時間からくる制約である。例えば、日本の都道府県別 TFR（合計出生率）の年次変化のグラフを示すのに、プレゼンテーションならば左図のようにした方が見やすいが、論文に載せる場合は右図のようにする方が良い。いずれの場合も統計ソフトだけで仕上げるのは（不可能ではないが）面倒だし、管理上も不都合なので、プレゼンテーションソフト*⁸か描画ソフト*⁹に貼り付けて仕上げるのが普通である。

*⁸ PowerPoint のほかには、フリーソフトの OpenOffice.org に含まれている Impress というものがある。有名であり、概ね PowerPoint と互換である。OpenOffice.org はフリーソフトでありながらきわめて高機能なオフィスソフトである。<http://ja.openoffice.org/>を参照されたい。

*⁹ OpenOffice.org に含まれている Draw も使える。他には、高価なソフトではあるが、もし使える環境にあれば Adobe の Illustrator がよい。

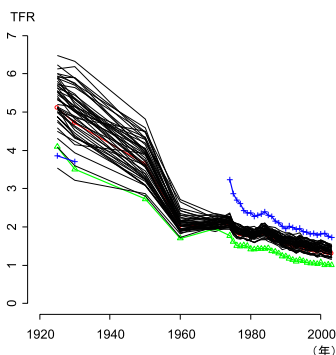
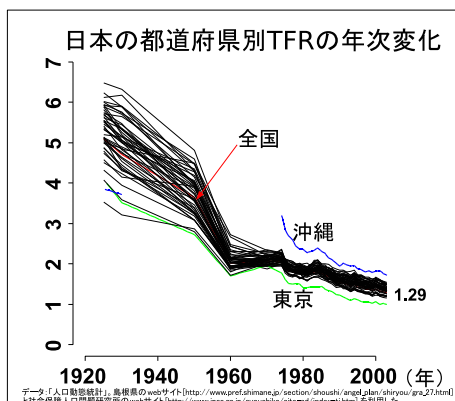


図2.日本の都道府県別合計出生率(TFR)の年次変化
(△:東京, +:沖縄, マークなし:他道府県, ○:全国).
データは「人口動態統計」に基づく。

見せるためのグラフについて詳しく知りたい方は、山本義郎 (2005) 『レポート・プレゼンに強くなるグラフの表現術』講談社現代新書 (ISBN4-06-149773-1)を一読されることをお勧めする。本書では考えるためのグラフに絞って説明する。考えるためのグラフに必要なのは、データの性質に忠実に作るということである。データの大局的性質を把握するために、ともかくたくさんのグラフを作って多角的に眺めてみよう。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、それを生かさなない手はない。統計解析は、いろいろな仮定をおいて理論構築されているので、ただソフトウェアの計算結果の数値だけを妄信してしまうのは危険である。図示されたものをみれば、直感的なチェックができるので、仮定を満たしていない統計手法を使ってしまう危険が避けられる場合が多い。つまり、

統計解析前に図示は必須

であると心得よう。R で図示をした場合、最大の利点は、その図をベクトルグラフィックスとして加工したり再利用できることである。図を作った後で、pdf 形式あるいは jpg 形式、png 形式、tiff 形式などで画像として保存しておくことも可能だが、Windows 環境ならばメタファイル形式にしておく再加工が容易である (Macintosh や Linux 環境なら postscript 形式がよいと思われる)。しかし、たくさんの図を作ったときは、ある程度まとめて管理できた方が便利だし、コメントもつけておく方が、

再利用するときに役に立つと思われる。そのためにも、前述の通り、作った図は、メタファイルとしてプレゼンテーションソフトや描画ソフトに貼り付けておくことをお勧めする。なお、描画ソフト内でも日本語を扱うことを考えると、Windows 環境でメタファイルを使う際は、描画前に次の枠内のように Windows 内蔵の TrueType フォントを指定しておくといよい。動作が軽くなるし、描画ソフトに読み込んだり貼り付けたりした後、加工するために切り離れたときに日本語が化けない。以下の例では MS ゴシックを JP1 ファミリ、MS 明朝を JP2 ファミリと名付け、`par()` で描画に使われるフォントファミリを JP1 とすることで描画に使われるフォントが MS ゴシックになる。なお、`par()` はグラフィックデバイス指定後 (Windows の場合は画面表示なら指定しなくてもいい)、描画コマンドを実行する前に実行せねばならない。

```
windowsFonts(JP1=windowsFont("MS Gothic"),JP2=windowsFont("MS Mincho"))
par(family="JP1")
```

では、具体的な図示の方法に入ろう。変数が表す尺度の種類によって、さまざまな図示の方法があるので、それをざっと示すことにする。

2.7 名義尺度や順序尺度をもつ変数の図示

要因型または順序型の変数についての作図は、カテゴリごとの度数を情報として使うことになる。順序型の場合はカテゴリの順番も情報となる。そのため、作図関数に渡す値は一般にデータそのものではなく、その集計結果になる^{*10}。もちろん、既に表の形になっている場合は、そのまま作図関数に渡すことができる。なお、以下の例の多くは、R のプロンプトで `source()` 関数を使って実行可能である^{*11}。

^{*10} `table()` 関数を使って度数分布を求め、その結果を作図関数に与えるのが普通である。

^{*11} 例えば枠の上辺に `c02-1.R` と書かれていれば、

```
source("http://phi.med.gunma-u.ac.jp/msb/c02-1.R")
```

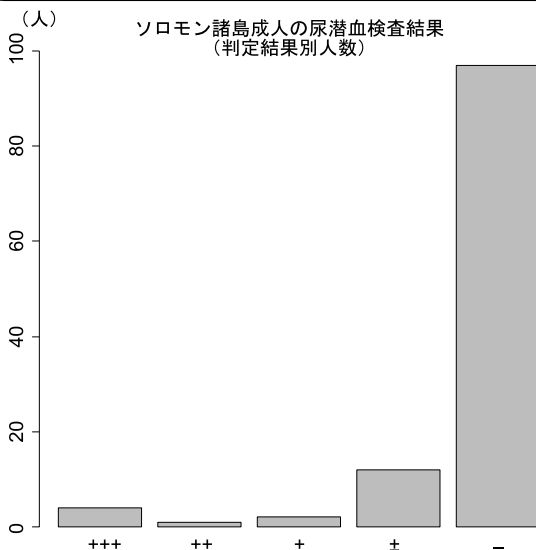
などとする。または、R Console の「ファイル」で「スクリプトを開く」を選び、ファイル名を入力するところに URL を打つと R Editor にコードを読み込めるので、その後、「編集」の「全て実行」を選んで実行させてもよい。ただし、日本語文字コードが CP932 になっているので、Linux や MacOS では文字化けが起こる可能性がある。`source()` 関数の引数に `encoding="CP932"` を入れば、この文字化けは回避できるはずである。

2.7.1 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を度数分布図という。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。例えば、ソロモン諸島の M 村の成人について尿検査をして、潜血の結果が、+++ が 4 人、++ が 1 人、+ が 2 人、± が 12 人、- が 97 人だったとしよう。これを度数分布図として棒グラフを作成するには、どうしたらいいだろうか。この例では、`table(X)` に当たる部分が既に与えられているので、次の枠内のように、まずカテゴリ別度数を `c()` で与え、`names()` を使ってカテゴリに名前を付けてから、`barplot()` 関数で棒グラフを描画すればよい（以下の例では OpenOffice.org の Draw を使って加工済みのグラフを載せておく）。

```
c02-1.R
```

```
ob <- c(4,1,2,12,97)
names(ob) <- c("+++","++","+","±","-")
barplot(ob,ylim=c(0,100),
        main="ソロモン諸島成人の尿潜血検査結果\n(判定結果別人数)")
```



なお、合計で割って縦軸を割合にした方が見やすい場合もある。

2.7.2 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。上のデータで積み上げ棒グラフを描くには以下のようにする。最初の2行は変わらない。残りは、`barplot()`の引数を変えることと、カテゴリ名をグラフィック画面上の適切な位置に書き込むためのコードである。`as.matrix()`は、ベクトルを行列に型変換する関数である。ここでは5行1列の行列になるので、`as.matrix(ob)`は、`matrix(ob,5)`と同じことを表すが、`as.matrix()`の方が名前も保存される点が優れている。`length()`は、ベクトルの長さを意味するので、`length(ob)`は5である。`for (in) { }`は決まった回数 of 繰り返し実行を指示する制御文である。ここでは、`i`の中身を1から5まで変えながら`{}`内を繰り返す。`sum()`は合計を意味する。`text()`は、グラフィック画面の指定座標に文字列を表示する関数である。`paste`は文字列変数の内容を評価して表示する場合につける。`names`は、名前付き変数に名前をつけたり名前を参照する際に用いる。^{*12}

c02-2.R

```
ob <- c(4,1,2,12,97)
names(ob) <- c("+++", "++", "+", "±", "-")
ii <- barplot(as.matrix(ob), beside=F, ylim=c(0,120),
             main="ソロモン諸島成人の尿潜血検査結果")
oc <- ob
for (i in 1:length(ob)) { oc[i] <- sum(ob[1:i])-ob[i]/2 }
text(ii, oc, paste(names(ob)))
```

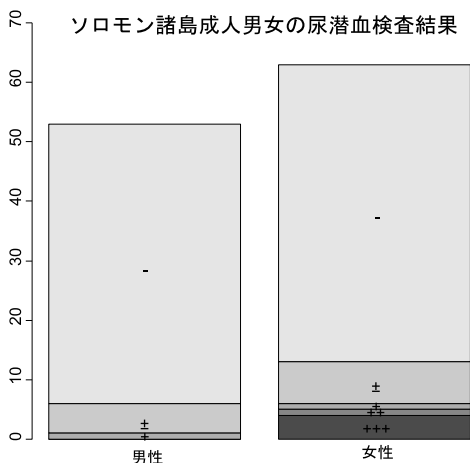
積み上げ棒グラフは単独で用いるよりも、複数の積み上げ棒グラフを並べて比較するのに向いている。例えば、上の結果を男女別に見ると、男性では+++と++が0人、+が1人、±が5人、-が47人、女性では+++が4人、++が1人、+が1人、±が7人、-が50人だったとき、男女別々に積み上げ棒グラフを描いて並べると、内訳を男女で比較することができる。実行するためのRのコードは次の通りである。1行目で男性、2行目で女性のカテゴリ別度数を付値する。3行目の`cbind`は、ベクトルをカラムとして結合する関数である。結果として`obx`は5行2列の行列となる。次の行の`rownames()`関数によって行の名前、つまり潜血のカテゴリ名を付値し、さらに次の行の`colnames()`関数によって列名を与えている。次の行の

^{*12} 5-6行目は、実は`oc <- cumsum(ob)-ob/2`でも済む。`cumsum()`は累和を求める関数である。

barplot() 関数では、積み上げ棒グラフを描くために beside=F というオプションを与えている。このオプションがないと、積み上げでない度数分布図が2つ並べて描かれる。なお、barplot() 関数の返り値は、棒の X 座標である。残りの5行は、まず各カテゴリの四角形の中央となる Y 座標を保存するための配列変数 oc を用意し、男女別々に Y 座標を計算し、カテゴリ名を text() 関数でグラフィック画面上に表示している。このとき、表示する文字列は paste() 関数で括っておく方が安全である。

c02-3.R

```
obm <- c(0,0,1,5,47)
obf <- c(4,1,1,7,50)
obx <- cbind(obm,obf)
rownames(obx) <- c("+++", "++", "+", "±", "-")
colnames(obx) <- c("男性", "女性")
ii <- barplot(obx, beside=F, ylim=c(0,70),
              main="ソロモン諸島成人男女の尿潜血検査結果")
oc <- obx
for (i in 1:length(obx[,1])) { oc[i,1] <- sum(obx[1:i,1])-obx[i,1]/2 }
for (i in 1:length(obx[,2])) { oc[i,2] <- sum(obx[1:i,2])-obx[i,2]/2 }
text(ii[1], oc[,1], paste(rownames(obx)))
text(ii[2], oc[,2], paste(rownames(obx)))
```

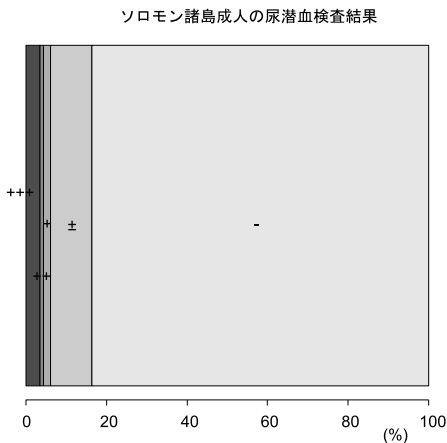


2.7.3 帯グラフ

横棒を全体を 100% として各カテゴリの割合にしたがって区切って塗り分けた図である。内訳を見るのに向いている。複数並べて構成比を比べたいときに効果を発揮する。ソロモン諸島成人の尿潜血検査結果について帯グラフを描くための R のコードを c02-4.R に示す。1 行目はカテゴリ別人数の付値, 2 行目は人数を構成割合 (%) に変える計算 (sum()) で求めた合計人数で割った後, 100 を掛けて % 表示にしている), 3 行目で names() 関数を使ってカテゴリ名を付けて準備が完了する。4 行目が (継続行で 5 行目も) 実際に帯グラフを書くところである。度数分布図や積み上げ棒グラフと同じ barplot() 関数を使うのだが, オプションで horiz=T を付けると棒が横に寝てくれる。割合を塗り分けるためには積み上げと同じ形にしなくてはいけないので, beside=F オプションも付ける。横軸が 0 から 100 までと固定しているので, xlim=c(0,100) も必須である。最後の 2 行は, 帯の部分ごとにカテゴリ名を表示するためのコードである。

c02-4.R

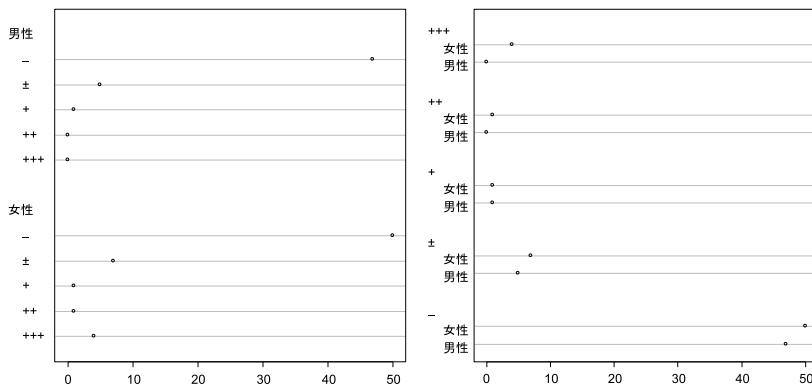
```
ob <- c(4,1,2,12,97)
obp <- ob/sum(ob)*100
names(obp) <- c("+++", "++", "+", "±", "-")
ii <- barplot(as.matrix(obp), horiz=T, beside=F, xlim=c(0,100),
  xlab="%", main="ソロモン諸島成人の尿潜血検査結果")
oc <- cumsum(obp)-obp/2
text(oc, ii, paste(names(obp)))
```



2.7.4 ドットチャート

棒グラフの棒を描く代わりに上端に点を打ったグラフである。複数のドットチャートを並列することもできる。基本的に `barplot()` の代わりに `dotchart()` を使えばよい。ソロモン諸島成人男女の尿潜血の例では、次の枠内のコードを用いればよい。

```
c02-5.R
obm <- c(0,0,1,5,47)
obf <- c(4,1,1,7,50)
obx <- cbind(obm,obf)
rownames(obx) <- c("+++","++","+", "±", "-")
colnames(obx) <- c("男性","女性")
layout(t(1:2))
dotchart(obx)
dotchart(t(obx))
```



2.7.5 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、各カテゴリ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。構成比を見るには、帯グラフよりも直感的にわかりやすい場合も多い。ただし、認知心理学者の Cleveland は、その著書【Cleveland WS (1985) *The elements of graphing data*. Wadsworth, Monterey, CA, USA.】の p.264 で、「円グラフで示すことができるデータは、常にドットチャートでも示すことができる。このことは、共通した軸上の位置の判定が、正確度の低い角度の判定の代わりに使えることを意味する。」と、実験研究の結果から述べているし、R の help ファイルによると、構成比を示すためにも円グラフよりもドットチャートや帯グラフあるいは積み上げ棒グラフを使うことが薦められているので、円グラフはむしろ「見せるためのグラフ」として使う際に価値が高いといえよう。

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^{*13}。ソロモン諸島成人の尿潜血検査結果について円グラフを描かせる R のコードを `c02-6.R` に示す。

*13 R-1.5 以前は `piechart()` 関数だったが置き換えられた

```
c02-6.R  
ob <- c(4,1,2,12,97)  
names(ob) <- c("+++", "++", "+", "±", "-")  
pie(ob)
```

2.8 連続変数の場合

2.8.1 ヒストグラム

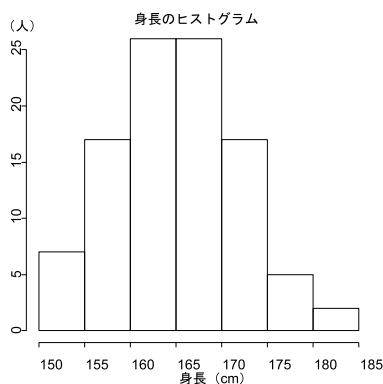
変数値を適当に区切って度数分布を求め、分布の様子を見るものである。Microsoft Excel ではツールのアドインの分析ツールに含まれているヒストグラム作成機能で区切りも与えてやらないと作成できず、非常に面倒だが、R では `hist()` 関数にデータベクトルを与えるだけである、棒グラフとの違いは、横軸（人口ピラミッド^{*14}のように90度回転して縦軸になることもある）が、連続していることである（区切りに隙間があってはいけない）。縦軸はデフォルトでは頻度だが、`freq=F` オプションが `prob=T` オプションをつければ全体を1とした割合にできる。

基本的に、区切りにはアプリオリな意味はないので、分布の形を見やすくするとか（区切りを明示的に指定しない限り、Sturgesの式により勝手に見やすい区切りが選ばれる）、10進法で切りのいい数字にするとかでよい。対数軸にする場合も同様である。第1章で使用した身長と体重のデータをタブ区切りテキスト形式で保存して変数名を大文字に変えたファイル^{*15}を用いて、身長のヒストグラムを描かせるコードは次の枠内の通りである。なお、MASSライブラリに `truehist()` という関数が含まれており、それも使える。`truehist()` 関数では、明示的に区切りを指定しない場合はScottの式を用いて区切り数が選ばれることと、縦軸が基本的に頻度でなく全体を1とした割合になっていること、デフォルトでヒストグラムに色が付いていることが大きな違いのようである。

```
c02-7.R  
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p01.txt")  
hist(dat$HT, main="身長のヒストグラム")
```

^{*14} 詳しくは <http://phi.med.gunma-u.ac.jp/demography/makepyramid.html> を参照。

^{*15} <http://phi.med.gunma-u.ac.jp/msb/data/p01.txt> として公開してある。

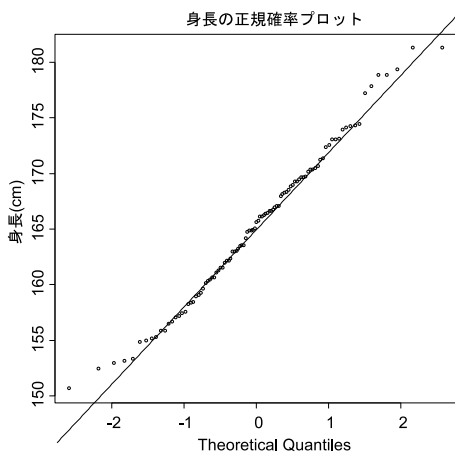


2.8.2 正規確率プロット

連続変数が正規分布しているかどうかを見るグラフである。正規分布に当てはまっていれば点が直線上に並び、ずれていればどのようにずれているかを読み取ることができる。ヒストグラムに比べると、正規確率プロットから分布の様子を把握するには熟練を要するが、区切りの恣意性によって分布の様子が違って見える危険がないので、ヒストグラムと両方実施すべきである。ヒストグラムで示したのと同じデータについて正規確率プロットを描かせるには、次の枠内を打てばよい。qqnorm() で正規確率プロットが描かれ、qqline() で、もしデータが正規分布していればこの直線上に点がプロットされるはず、という直線が描かれる。ここで lty=2 は線種の指定で、破線を描かせている。実線ならば lty=1 とする。

c02-8.R

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p01.txt")
qqnorm(dat$HT,main="身長の正規確率プロット",ylab="身長 (cm)")
qqline(dat$HT,lty=2)
```



2.8.3 幹葉表示 (stem and leaf plot)

だいたいの概数 (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を幹葉表示 (かんようひょうじ、あるいは、みきはひょうじと読む) という。英語で “stem and leaf plot” という方が通りがよい。R では `stem()` 関数を用いる。ただしテキスト出力画面に出力されるため、グラフィックとして扱うには少々工夫が必要である。言ってみればヒストグラムを90度回転して数字で作るようなグラフだが、各階級の内訳がわかる利点がある。身長の例では、次の枠内のように打てば、テキスト画面に幹葉表示が得られるし、`gstem()` 関数を次のように定義して `stem()` の代わりに使えばグラフィック画面にも出力できる。要は、`capture.output()` 関数を使って幹葉表示のテキスト画面への出力を変数 `stem.out` に取り込み、`length()` 関数で幹の長さを計算してから `plot(..., type="n", axes=F, xlab="", ylab="")` で何も表示せずにグラフィック画面の座標系だけ定義し、最後に `text()` 関数で座標を指定して幹と葉をグラフィック画面に書き込むようにする。

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p01.txt")
stem(dat$HT)
```



```
gstem.R  
  
gstem <- function (X,D=1) {  
  .stem.out <- capture.output(stem(X,D))  
  .stem.len <- length(.stem.out)  
  plot(c(1,2),c(1,.stem.len),type="n",axes=F,xlab="",ylab="")  
  text(rep(1,.stem.len),.stem.len:1,.stem.out,pos=4)  
}
```

2.8.4 箱ヒゲ図 (box and whisker plot)

次章で詳述するが、データを小さい方から順番に並べて、ちょうど真中にくる値を中央値 (median) といい、小さい方から $1/4$ の位置の値を第 1 四分位 (first quartile)、大きいほうから $1/4$ の位置の値を第 3 四分位 (third quartile) という。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。身長データだと次の枠内を打てばよい。

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p01.txt")  
boxplot(dat$HT)
```

2.8.5 ストリップチャート (stripchart)

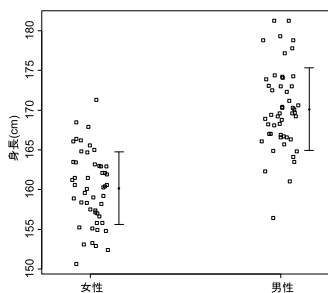
2 群間で平均値を比較する場合などに、群ごとに大まかに縦軸での位置を決め、横軸には各データ点の正確な値をプロットしたストリップチャートを描くと見やすい。群の数によって縦軸と横軸は入れ換えた方が見やすいこともある。R では `stripchart()` 関数を用いる。縦軸と横軸を入れ換えるには、`vert=T` オプションをつける。横に平均値と標準偏差を表すエラーバーを付加することも多い。

身長データを男女間で比較するためにストリップチャートとエラーバーを描く (縦軸横軸は上記説明とは逆にした) コードの例を `c02-9.R` に示す。1 行目で web 上のデータファイルを `dat` というデータフレームに読み込み、2 行目で `subset()` の条件として `complete.cases()` を指定することで欠損値を除いたサブセットを作っ

てから^{*16}, 3行目に `dat` を `attach` して変数名だけで参照可能にし, 4行目では `tapply(HT,SEX,mean)` によって身長を表す `HT` の `SEX` 別の平均値を計算し, 長さ2のベクトルとして `mHT` という変数に付値する。5行目は同様に標準偏差を性別に計算して `sHT` という変数に付値する。6行目はストリップチャートの横に描くエラーバーの横軸上の座標 (チャートの中心から右側に2つのチャートの間隔の15% だけずらした位置) を計算して `IS` という変数に付値するコードである。7行目で実際にストリップチャートを描かせ, 8行目の `points()` で平均値をプロットし, 9行目の `arrows()` で平均値の上下にエラーバーを描かせている。 `arrows()` は元々矢印を描く関数だが, `code=3`, `angle=90` というオプションをつけることでエラーバーの上下に横線が引かれる。その横線の長さは `length=.1` という指定によって決まる。

c02-9.R

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p01.txt")
dat <- subset(dat,complete.cases(dat))
attach(dat)
mHT <- tapply(HT,SEX,mean)
sHT <- tapply(HT,SEX,sd)
IS <- c(1,2)+0.15
stripchart(HT~SEX,method="jitter",vert=T,ylab="身長 (cm)")
points(IS,mHT,pch=18)
arrows(IS,mHT-sHT,IS,mHT+sHT,code=3,angle=90,length=.1)
detach(dat)
```



^{*16} ここで欠損値を除くのは, `sd()` が `na.rm=T` オプションを付けずに欠損値を含むベクトルを与える
とエラーになるので, それを回避するためである。

2.8.6 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を散布図という。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、複数の変数について、2つずつの関係を同時散布図として描画したい場合は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。もっとも基本的な使い方として、身長と体重の関係を男女別にマークを変えてプロットするなら、次の枠内のようにする。

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p01.txt")
plot(dat$HT,dat$WT,pch=paste(dat$SEX),xlab="身長 (cm)",ylab="体重 (kg)")
```

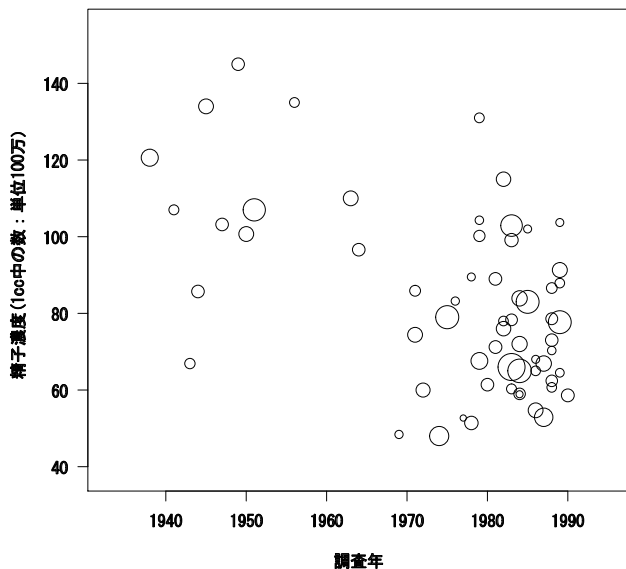
symbols()関数の使用例を以下に示す。これは、1992年になされたCarlsenらの「全世界で男性の精子数が減少し続けている」という衝撃的な報告の元となったメタアナリシスで使われたプロットである。調査年を横軸に、精子濃度を縦軸にとり、半径がサンプルサイズの自然対数に比例する円をプロットしたものである^{*17}。

semen.R

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/semen.txt")
attach(dat)
SIZE <- log(NUMBER)/5
symbols(YEAR,CONC, circle=SIZE, inches=F, xlab="調査年",
        ylab="精子濃度 (1cc 中の数 : 単位 100 万)",
        main="精子濃度経年変化のメタアナリシス")
detach(dat)
```

^{*17} なお、半径をサンプルサイズの平方根に比例させれば、面積がサンプルサイズに比例するが、その状態では円同士があまりひどく重なり合わないようには比例定数をきわめて小さくする必要があり、サンプルサイズが小さいデータが点としてしか表示されなくなってしまうので、図示のテクニックとして自然対数に比例させた。比例定数を1/5にしたのは試行錯誤による。原論文でもサンプルサイズの対数に比例した円になっているのは、おそらく同じ理由からであろう。

精子濃度経年変化のメタアナリシス



2.8.7 レーダーチャート (radar chart)

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をレーダーチャートと呼ぶ。英語では radar chart または spiderweb chart ともいう。それら複数の変数によって特徴付けられる性質のバランスをみるのに役立つ。1つのケースについて1つのレーダーチャートができるので、他のケースと比較するには、並べて描画するか、重ね描きする。Rでは `stars()` 関数を用いることで並べ描きも重ね描きもできるが、各々の軸について値が相対化されてしまうと、単独のチャートが描けないという制限がある（データとしては常に行列かデータフレームを与えねばならない）。詳細は省略するが、`example(stars)` によって例示を見ればだいたいの使い方がわかるだろう。

2.9 塗り分け地図

以上説明した基本的なグラフ以外にも、遺伝子データなどを使った系統関係を示す樹状図（デンドログラム）や、生存関数など、外部ライブラリを使えば、目的に応じ

て多様なグラフが作成できる。ここではその一例として、`maptools` ライブラリを用いた^{*18}塗り分け地図の作り方を簡単に説明する。塗り分け地図とは、地図上の小区域ごとの指標値を何段階かに区分し、どの区分に属するかによって個々の小区域を塗り分けた地図である^{*19}。必要な情報は区域に関連付けられた指標値と、地図情報データである。地図情報データはデジタイザなどを使って独自に作成することも不可能ではないが、通常、既に公開されているシェイプファイルを web サイトからダウンロードする。シェイプファイルとは、ESRI 社が提供している GIS 用の地図情報データの形式であり、拡張子が `shp` となっている。全世界について、郡レベルくらいまでは無料でダウンロードできる。ESRI 社が無償公開している ArcExplorer などを使ってマウスで場所を選ぶ方法もあるが、場所が決まっていれば、DIVA-GIS のサイト^{*20}から Country で国名を選び、Theme を Administrative Boundaries と指定して zip 圧縮されたファイルをダウンロードする方が手取り早い。

ここでは、群馬県の各市町村を、65 歳以上人口比率という指標値にしたがって 5 段階に塗り分ける例を示す。日本の場合、DIVA-GIS で得られる最小の地域区分は都道府県なので、市町村単位のシェイプファイルを得るため、ESRI Japan 社の全国市町村界データ^{*21}から `japan_ver61.zip` をダウンロードする。RjpWiki の中にある都道府県別のシェイプファイルデータ^{*22}を利用する手もある。

群馬県の市町村別 65 歳以上人口比率は、群馬県の年齢別人口（平成 18 年 10 月 1 日現在）^{*23}から、第 3 表を Microsoft Excel 形式で^{*24}ダウンロードできる。これを地図情報とつなげるには市町村名を市町村コードとリンクさせる必要があるが、市町村コードの表も web 上から得られる^{*25}。この 2 つのファイルから、変数として JCODE（市町村コード）と AP2006（市町村別 65 歳以上人口比率）をもつタブ区切りテキストファイル^{*26}を作成した。

^{*18} CRAN から `install.packages(maptools,dep=T)` として事前にインストールする必要がある。

^{*19} GIS (Geographic Information Systems) の専門用語ではコロプレス図と呼ばれる。

^{*20} <http://www.cipotato.org/DIVA/data/DataServer.htm>

^{*21} http://www.esri.jp/gis_data/japanshp/japanshp.html。頻繁に更新されて、バージョン番号が付されている。

^{*22} <http://www.okada.jp.org/RWiki/?ShapeFile%A5%E9%A5%A4%A5%D6%A5%E9%A5%EA>

^{*23} <http://toukei.pref.gunma.jp/NBJ2006.htm>

^{*24} http://toukei.pref.gunma.jp/sokuhou_temp/nbj2006_005.xls

^{*25} http://www.lasdec.nippon-net.ne.jp/com/addr/kaku_ken/gunnmaken.htm の「団体コード」は、最初の 2 桁の 10 が群馬県を意味し、次の 3 桁が市町村を表し、最後の 1 桁がチェックディジットである。

^{*26} <http://phi.med.gunma-u.ac.jp/msb/data/agedprop.txt>。ただし、平成 18 年 10 月 1 日から榛名町（市町村コード 10321）が高崎市に合併されたのに、シェイプファイルは旧榛名町のまま

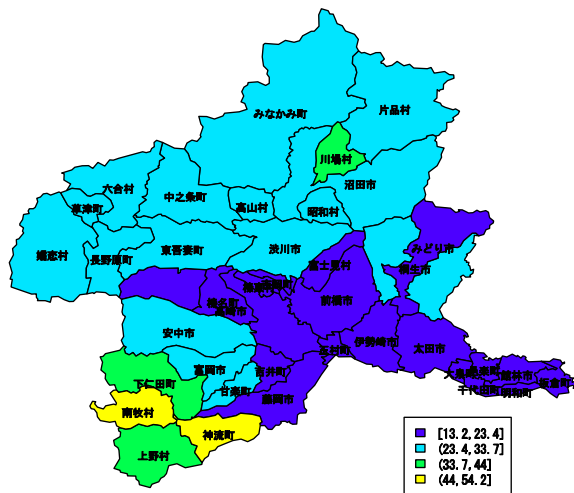
あとは、作業ディレクトリに `japan_ver61.zip` を展開してできるすべてのファイルと `agedprop.txt` をおいて以下のコードを実行すれば塗り分け地図ができあがる。市町村名は重心に描かれる。桐生市が変な位置に描かれているのは、飛び地があるためである。描画色は `topo.colors()` を用いて指定したが、`cm.colors()` だとシアンからマゼンタまで、`heat.colors()` だと赤から黄色まで、`rainbow()` だと虹のグラデーションを指定できるし、`c("red","blue","green","yellow")` のように指定することもできる。

```
classmap.R
```

```
library(maptools)
jpn <- read.shape("./japan_ver61.shp")
jpndata <- jpn$att.data
jpnpoly <- Map2poly(jpn,region.id=attr(jpn,"region.id"))
jpnxy <- get.Pcent(jpn)
gunmadata <- subset(jpndata,jpndata$PREF=="群馬県")
gunmapoly <- subset(jpnpoly,jpndata$PREF=="群馬県")
x <- subset(jpnxy[,1],jpndata$PREF=="群馬県")
y <- subset(jpnxy[,2],jpndata$PREF=="群馬県")
aged <- read.delim("./agedprop.txt")
gunmadata <- merge(aged,gunmadata,sort=F,by="JCODE")
DD <- gunmadata$AP2006
classes <- cut(DD,seq(min(DD),max(DD),length=5),include.lowest=T)
cols <- topo.colors(4)
plot(gunmapoly,col=cols[ordered(classes)],xlab="",ylab="",axes=F)
legend(max(x)-0.3*(max(x)-min(x)),min(y)+0.1*(max(y)-min(y)),
  legend=names(table(classes)),cex=0.6,fill=cols)
title("群馬県市町村の65歳以上高齢者割合4区分(2006年)")
text(x,y,gunmadata$CITY1,cex=0.5,pos=1,offset=0)
```

入っているため、そのままではうまく結合できない。タブ区切りテキストファイル内に榛名町を追加し(65歳以上人口比率としては高崎市のデータを用いて)、データを結合した。平成の大型市町村合併は時期をずらして起きているため、このようなデータ間の不整合が起こることがある。

群馬県市町村の65歳以上高齢者割合4区分 (2006年)



2.10 課題

<http://phi.med.gunma-u.ac.jp/msb/data/p01b.txt> から、体重 (wt) の分布を男女別に図示せよ。

第3章

記述統計量

3.1 データを記述する2つの方法

前章では、データを図示して直感的に全体像を把握する方法を示した。本章では、生データが含んでいる多くの情報を少ない数値に集約して示す方法を説明する。つまり、分布の特徴をいくつかの数値で代表させようというわけである。このような値を、代表値と呼ぶ。たんに代表値という場合は分布の位置を指すことが多いが、ここではもっと広い意味で用いる。代表値は、記述統計量 (descriptive statistics) の1つである。

分布の特徴を代表させる値には、分布の位置を示す値と、分布の広がりを示す値がある。例えば、正規分布だったら、 $N(\mu, \sigma^2)$ という形で表されるように、平均値 μ 、分散 σ^2 という2つの値によって分布が決まるわけだが、この場合、 μ が分布の位置を決める情報で、 σ^2 が分布の広がりを決める情報である。分布の位置を示す代表値は central tendency (中心傾向) と呼ばれ、分布の広がりを示す代表値は variability (ばらつき) と呼ばれる。

一般に、統計処理の対象になっているデータは、仮想的な母集団 (言い換えると、その研究結果を適用可能と考えられる範囲ともいえる) からの標本 (サンプル) であり、データから計算される代表値は、母集団での分布の位置や広がりを推定するために使われる。母集団での位置や広がりを示す値は母数 (parameter) と呼ばれ、分布の位置を決める母数を位置母数 (location parameter)、分布の広がりを決める母数を尺度母数 (scale parameter) と呼ぶ。例えば不偏分散は尺度母数の一つである母分散の推定量となる。

3.2 中心傾向 (Central Tendency)

3.2.1 平均値 (mean)

平均値^{*1}は、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。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、数式で書くこと以下の通りである。

母集団の平均 μ (ミュウと発音する) 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、 $\sum X = X_1 + X_2 + X_3 + \dots + X_N$ である。

標本についての平均を求める式も、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^{*2}。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、もちろんサンプルサイズである^{*3}。

例題

パプアニューギニア高地辺縁部のある集落には成人男性が7人しかいなかった。その全員の身長を測ったら、{145 cm, 148 cm, 152 cm, 153 cm, 155 cm, 157 cm, 161 cm} であった。この集落の成人男性の平均身長は何センチメートルか？

$\mu = (145 + 148 + 152 + 153 + 155 + 157 + 161)/7 = 140 + (5 + 8 + 12 + 13 + 15 + 17 + 21)/7 = 140 + 13 = 153$ より、153 cm であることは、小学生でもわかる

^{*1} ここで平均値と呼んでいるのは、とくに断りがなければ、算術平均 (arithmetic mean) のことである。

^{*2} 一般に母集団についての統計量を示す記号にはギリシャ文字を使うことになっている。

^{*3} 記号について注記しておくが、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} と表するのが普通である。

だろう。R で平均値を計算するには、`mean()` という関数を使う。例えば、この例題の解を得るには、

```
mean(c(145, 148, 152, 153, 155, 157, 161))
```

とすればよい。もちろん、通常は、データを何かの変数に付値しておいて、関数は変数に対して適用するので、以下のように入力することになる。なお、数値型あるいは整数型の変数 X について、`mean(X)` は `sum(X)/length(X)` と同値である。`length()` はベクトルの長さ、即ちデータ数を返し、`sum()` は合計を返す。

```
X <- c(145, 148, 152, 153, 155, 157, 161)
mean(X)
```

中心傾向として有名なものには、平均値の他に、あと2つ、中央値 (median) と最頻値 (mode) がある。どれも分布の中心がどの辺りに位置するかを説明するものだが、中心性 (centrality) へのアプローチが異なっている。

平均値の中心性へのアプローチを説明しよう。「偶然にも値が平均値と同じであった」という稀な値を除けば、各々の値は、平均値からある距離をもって存在する。言い換えると、各々の値は、平均値からある程度の量、ばらついている。ある値が平均値から離れている程度は、単純に $X - \bar{X}$ である。この、平均値からの距離を偏差といい、 x という記号で書く。つまり、 $x = X - \bar{X}$ である。次の例を見ればわかるように、偏差は正の値も負の値もとるが、その合計は0になるという特徴をもつ。どんな形をしたどんな平均値のどんなに標本サイズが大きいデータだろうと、偏差の和は常に0である。式で書くと、 $\sum x = \sum (X - \bar{X}) = 0$ ということである。言い方を変えると、偏差の和が0になるように、平均値によって調整が行われたと見ることもできる。平均値が分布の中心であるといえるのは、この意味においてである。しかし、定義から明らかなように、少数でも極端な外れ値があると、平均値は強くその影響を受けてしまうという欠点がある。対策としては、外れ値の検定^{*4}を用いることもできるが、むしろ数値以外の情報に基づいて、その極端な値がサンプルとして相応しくないと判定して解析から除外するか、それができない場合は、後述する中央値を使うのが望ましい。外れ値だからといって機械的に除くのはバイアスの原因となりうる。

^{*4} outliers ライブラリを利用すれば、 χ^2 、コクラン、ディクソン、グラブスなどの方法が使える。それぞれ、`chisq.out.test()`、`cochran.test()`、`dixon.test()`、`grubbs.test()` として実装されている。

3.2.2 重み付き平均 (weighted mean)

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。ここでは標本サイズが異なる複数の平均の総平均 (grand mean) を計算する場合について説明する。サイズがそれぞれ n_1, n_2, \dots, n_k であるような k 個の平均値 $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$ の総平均値は、次の式で得られる。

$$\bar{X} = \frac{n_1(\bar{X}_1) + n_2(\bar{X}_2) + \dots + n_k(\bar{X}_k)}{n_1 + n_2 + \dots + n_k}$$

例題

ある大学の 3 つの学部で行われた TOEIC の試験の平均点が { 440, 470, 610 } であったとする。これら 3 つの学部それぞれの人数が、順に { 200 人, 500 人, 300 人 } であったなら、この大学の TOEIC の総平均は何点か？

文章通りに実行する R のコードは次の通りである。3 行目のように、ベクトル同士を * でつなげば、要素ごとの掛け算がなされる。

```
P <- c(440, 470, 610)
N <- c(200, 500, 300)
sum(N*P)/sum(N)
```

3.2.3 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が、決まった手続き (アルゴリズム) として、並べ換え (sorting) は必要である。先に述べた通り、極端な外れ値の影響を受けにくいので、歪んだ分布に対するもっとも重要な central tendency の指標といえる。

例題

次のデータの中央値は何か？ { 1, 4, 6, 8, 40, 50, 58, 60, 62 }

この場合、小さい方から数えても大きいほうから数えても 5 番目の値である 40 が

中央値であることは自明である。小さいほうから数えて 40 の次に位置する 50 との距離や、大きいほうから数えて 40 の次に位置する 8 との距離は、中央値を考える際には無関係である。中央値を求めるには、値を小さい順に並べ換えて*5、ちょうど真中に位置する値を探せばよい。この意味で、中央値は値の順序だけに感受性をもつ (= rank sensitive である) といえる*6。

R で中央値を計算するには、英語で中央値を意味する `median()` という関数を使う。例えば、上の例題の解を得るには、`median(c(1,4,6,8,40,50,58,60,62))` とすればよい。

例題

次のデータの平均値と中央値は何か？ $A = \{2, 4, 7, 9, 12, 15, 17\}$ また、それに 46 と 54 が加わった $B = \{2, 4, 7, 9, 12, 15, 17, 46, 54\}$ の平均値と中央値は何か？

R で

```
x <- c(2,4,7,9,12,15,17)
mean(x)
median(x)
y <- c(x,46,54)
mean(y)
median(y)
```

とすると、A の平均値は約 9.43、中央値は 9 であり、B の平均値は 18.4、中央値は 12 となることがわかる。大きい方に 2 つの極端な値を加えたただだが、平均値はほぼ倍増してしまう。それに対して、中央値は 1 つ大きい値に移るだけであり、中央値の方が極端な値が入ることに対して頑健 (ロバスト) といえる。

このようにデータが奇数個だったら、順番が真中というのは簡単に決められる。では、データが偶数個だったらどうするのだろうか？

例題

次のデータの中央値は何か？ $\{4, 6, 9, 10, 11, 12\}$

*5 値の数が少ない場合には、手作業で並べ換えればよいが、大量のデータを手作業で並べ換えるのは現実的でない。コンピュータのプログラムに値を並べ換えさせるアルゴリズムには、単純ソート、バブルソート、シェルソート、クイックソートなどがある。

*6 これに対して平均値は、値の大きさによって変わるので value sensitive であるといえる。

中央値が9と10の間にくることは明らかである。そこで、普通は9と10を平均した9.5を中央値として使うことになっている。Rでは以下の1行を打てばよい。

```
median(c(4,6,9,10,11,12))
```

もっとも、本来整数値しかとらないような値について、中央値や平均値として小数値を提示することに意味があるかどうかは問題である。例えば、上の例題のデータが、ある地方の水泳プールで6日間観察したときの、1日当たりの飛び込みの回数を示すものだとしよう。中央値が9.5ということになると、9.5回の飛び込みというのは何を表すのか？ 半分だけ飛び込むということにはありえない。つまり実体はない、単なる指標値だということになる。同様に平均値についても、世帯当たりの平均子供数が2.4人とかいうとき、0.4人の子供は実体としてはありえない。しかし、分布の位置を示す指標としては有用なので、便宜的に使っているのである。

例題

データ $C = \{7, 7, 7, 8, 8, 8, 9, 9, 10, 10\}$ の中央値は何か？ また、 $D = \{7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 10, 10\}$ の中央値は何か？

このように同順位の値 (tie という) がある場合は、事態はやや複雑である。順番で言えば、CもDも中央値は8と8の間に来るはずだから、8と思うであろう。実際、SAS、SPSSなどの有名ソフトを初めとして、Microsoft ExcelやRに至るまで、ほぼすべての統計ソフトは、8という答えを出してくるし、一般にはそれで問題ない。

ただし、厳密に考えると、簡単に8と言えない。Grimm (1993) が指摘するように、連続量データの数値はぴったりその値というわけではなく、間隔の midpoint と考えるべきだからである。つまり、測定単位1のとき、8というデータが意味するのは7.5以上8.5未満である。普通はそこまで厳密に考える必要はないが、参考までに説明しておく。

要点は、『それぞれの値を、表示単位によって規定される区間の midpoint と考え、同順位の値があるときは、それが区間内に均等に散らばると考える』ということである。これは直感的に考えても合理的であろう。

例えば、1 1 1 2 2 2 3 3 3 という、表示単位1のデータがあるとき、真の値がそれぞれ等間隔に散らばっているならば、0.67 1.00 1.33 1.67 2.00 2.33 2.67 3.00 3.33 と考えるのが自然である。これなら、それぞれの値が1/3間隔になっているし、midpoint 1で示される値0.67 1.00 1.33の平均値は1となるので、どこにも矛盾がない。

この例から帰納的に考えて、その区間の下限の値をLとし、階級幅をhとし、同

順位の個数を fm 個とし, 1つ下の区間までに F 個のサンプルがあるとすれば, $F+1$ 番目, $F+2$ 番目, ..., $F+fm$ 番目の値はそれぞれ, $L+1/(2fm)*h$, $L+3/(2fm)*h$, ..., $L+(2fm-1)/(2fm)*h$ となる。つまり, $F+x$ 番目の値は, $L+(2x-1)/(2fm)*h$ となる。

この式から C の 3 つの 8 の真の値がいくつになるか計算すると,

4 番 5 番 6 番

7.67 8.00 8.33

となって, 5 番と 6 番の間は 8.17 となる。

同じく D で真の値は, {6.67 7.00 7.33 7.60 7.80 8.00 8.20 8.40 8.75 9.25 9.75 10.25} となるので, 中央値は 8.00 と 8.20 の間で 8.10 となる。{1 1 2 2 3 3} という表示単位 1 のデータでは, 真の値は {0.75 1.25 1.75 2.25 2.75 3.25} と推定されるので, 中央値は 1.75 と 2.25 の平均値で 2 となる。以上のような考え方で真の中央値を求める関数 `truemedian()` を R で定義すると, 次の枠内のようになる。

```
truemedian <- function(X,h=1) {
  YY <- rep(0,length(X))
  XX <- table(X)
  q <- length(XX)
  k <- 0
  for (i in 1:q) {
    L <- as.numeric(names(XX)[i])-h/2
    for (j in 1:XX[[i]]) {
      k <- k+1
      YY[k] <- L+h*(2*j-1)/(2*XX[[i]])
    }
  }
  median(YY)
  # 真の値を表示するには print(YY)
}
```

ここでもう 1 歩進めて, 度数分布表から中央値を計算する場合を考えてみよう。次の表は年齢階級ごとの人数の分布であり, これから年齢の中央値を求める方法を考えることにする。

年齢階級	度数	累積度数
45-49	1	76
40-44	2	75
35-39	3	73
30-34	6	70
25-29	8	64
20-24	17	56
15-19	26	39
10-14	11	13
5-9	2	2
0-4	0	0

まず、累積度数の最大の数をみる（つまり総数をみる）。この例では76である。中央値の順位は $(76 + 1)/2 = 38.5$ 位となる*⁷。38.5 番目の値を含む年齢階級を探すと、15-19 である。そこで、単純に統計ソフトが出してくる中央値は 15-19 歳となる。なお、R では区間は Factor 型になってしまうので、c03-1.R に示すように区間の中央の値を数値型変数として入れて（: という連続する整数を示す演算子は最高の優先順位をもつので、10:1*5-3 は c(47,42,37,32,27,22,17,12,7,2) と同値である）median() を計算すると 17 という結果が得られることから、15-19 歳と判定できる。また、階級幅が 5 なので先に示したように truemedian() 関数を定義した上で*⁸c03-1.R を実行すれば、真の中央値としての約 19.3 歳も得られる。

```
c03-1.R
CA <- 10:1*5-3
FRE <- c(1,2,3,6,8,17,26,11,2,0)
X <- c(rep(CA,FRE))
median(X)
truemedian(X,5)
```

この考え方をすると、中央値が正確な分布の中央になっている（少なくともその近似になっている）という特性が強化される。

*⁷ Grimm (1993) には 76 を 2 で割って 38 番目の値が中央値であると書かれているが、論理的整合性を欠く。もし総数を 2 で割った順位の値が中央値だとすると、C の中央値が下から 3 番目で 9 ということになってしまう。総数に 1 を加えて 2 で割る方が論理的整合性が高い。

*⁸ source("http://phi.med.gunma-u.ac.jp/msb/msb-funcs.R") とすればよい。

3.2.4 最頻値 (Mode)

残る最頻値は、きわめて単純で、もっとも度数が多い値をいう。もっとも数が多い値が、もっとも典型的だと考えるわけである。データを見ると、最頻値が2つある場合があり、この場合は分布が二峰性 (bimodal) だという*⁹。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。

分布の形によって、平均、中央値、最頻値の関係は変わってくる。歪んでいない分布ならば、ばらつきの程度によらず、これら3つの値は一致する。二峰性だと最頻値は2つに分かれるが、平均と中央値はその間に入るのが普通である。左すそを引いた分布では、平均がもっとも小さく、中央値が次で、最頻値がもっとも大きくなる。右すそを引いた分布では逆になる。

例えば、例題のデータ D で最頻値を求めるためには、R では以下のようにすればよい。table() により度数分布が得られ、sort(,dec=T) で度数の降順 (つまり頻度が高い順) に並べ替えられ、names() でカテゴリ名つまり値が頻度順に得られ、[1] をつけることで1番目、つまりもっとも頻度の高い値が表示されるわけである。

```
D <- c(7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 10, 10)
names(sort(table(D),dec=T))[1]
```

ただし、最頻値が複数あるかもしれないので、2行目は sort(table(D),dec=T) としておいて、頻度の高い順に並べ替えられた度数分布表を見る方が確実である。

3.2.5 使い分け

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。標本調査データから母集団の情報を推論したい場合に、もっとも普通に使われる。しかし、(1) 極端な外れ値の影響を受けやすい、(2) 打ち切りのある分布では代表性を失う場合がある*¹⁰、という欠点があり、外れ値があったり打ち切

*⁹ 隣り合う2つの値がともに最頻値である場合は二峰性とはいわず、離れた2つの値が最頻値あるいはそれに近い場合、つまり度数分布やヒストグラムの山が2つある場合に、分布が二峰性だといい、2つの異なる分布が混ざっていると考えるのが普通である。

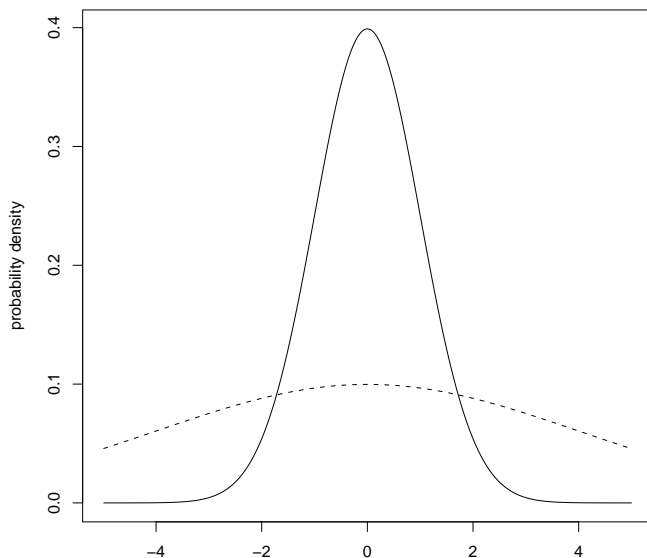
*¹⁰ 氷水で痛みがとれるまでにかかる時間とか年収とかいったものが良い例である。無限に観察を続けるわけにはいかないし、年収は下限がゼロで上限はビル・ゲイツのそのように極端に高い値があるから右すそを長く引いた分布になる。平均年収を出している統計表を見るときは注意が必要であ

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、名義尺度のカテゴリ変数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 (対数をとって平均を出して元に戻すことと同値)、調和平均はデータの逆数の平均の逆数であり、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。例えば、数値ベクトル X について幾何平均を得るには $\exp(\text{mean}(\log(X)))$ または $\text{prod}(X)^{1/\text{length}(X)}$ とすれば良いし (ただし後者の方が圧倒的に計算効率が悪い)、調和平均を得るには $1/(\text{mean}(1/X))$ とすれば良い。

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。

3.3 ばらつき (Variability)



分布を特徴付けるには、分布の位置だけではなく、分布の広がり具合の情報も必要である。例えば、上図の2つの分布は、どちらも平均0の正規分布なので中央値も最頻値も共通だが、実線で書かれた幅が狭い方が標準偏差1、破線で書かれた幅が広い方が標準偏差4と、標準偏差が大きく異なるために、まったく違った見かけになっている。標準偏差は、もっとも良く使われる分布の広がり具合の指標である。なお、上図を書くためのRのプログラムは次の通りである。curve()はグラフィック画面に関数を描画するのに便利である。引数として、横軸の変数 x を含む関数と横軸の描画範囲をコンマで区切って与えればよい。線種の指定は `lty=1` などとする。1が実線、2が破線である。`add=T` をつけると以前の描画を消さずに重ね描きできる。

```
curve(dnorm(x,0,1),-5,5,lty=1,xlab="",ylab="probability density")
curve(dnorm(x,0,4),-5,5,lty=2,add=T)
```

広がり具合を示す指標は、ばらつき (variability) と総称される。ばらつきの指標には、範囲、四分位範囲、四分位偏差、平均偏差、分散 (及び不偏分散)、標準偏差 (及び不偏標準偏差) がある。

3.3.1 範囲 (range)

範囲は、もっとも単純なばらつきの尺度である。値のとり全範囲そのもの、つまり最小値から最大値までの区間となる。1つの値として示すときは、最大値から最小値を引いた値になる。例えば、{17, 23, 42, 44, 50} というデータの範囲は、いうまでもなく、 $50 - 17 = 33$ である。しかし、ばらつきの尺度として範囲を使うには、若干の問題が生じる場合がある。極端な外れ値の影響をダイレクトに受けてしまうのである。次の例を考えてみよう。

例題

次のデータの範囲はいくらか？ {2, 4, 5, 7, 34}

答えは $34 - 2 = 32$ なのだが、2, 4, 5, 7 というきわめて近い値4つと、かけ離れた大きい34という値からなるのに、32という範囲は、全体のばらつきが大きいのかのような誤った印象を与えてしまう。ばらつきの指標としては、分布の端の極端な値の影響を受けにくい方がよいと考えるのが自然である。

3.3.2 四分位範囲 (Inter-Quartile Range; IQR)

そこで登場するのが四分位範囲である。その前に、分位数について説明しよう。値を小さい方から順番に並べ換えて、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、ちょうど大きさの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、もちろん、ちょうど4等分などできない場合がある。その場合、上から数えたときと下から数えたときで四分位数がずれるが、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。第 1 四分位, 第 2 四分位, 第 3 四分位は, それぞれ Q_1 , Q_2 , Q_3 と略記することがある。R では, `fivenum()` という関数によって, 五数要約値が得られる^{*11}。

これを一般化して, 値を小さい方から順番に並べ換えて, 同数の群に区切る点を分位数 (quantile) という。百等分した場合を, とくにパーセンタイル (percentile) という。言い換えると, 第 1 四分位は 25 パーセンタイル, 第 3 四分位は 75 パーセンタイルである。R で数値型変数 X の 20 パーセンタイル値と 80 パーセンタイル値を計算するには `quantile(X, c(0.2, 0.8))` のように `quantile()` 関数を使えばよいが, 実はパーセンタイル値指定のデフォルトが `c(0, 0.25, 0.5, 0.75, 1)` なので, `quantile(X)` は `fivenum(X)` とほぼ同じ結果になる^{*12}。

四分位範囲とは, 第 3 四分位と第 1 四分位の間隔である。パーセンタイルでいえば, 75 パーセンタイルと 25 パーセンタイルの間隔である。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。結果の示し方としては, $Q_3 - Q_1$ という 1 つの値で示すだけでなく, $[Q_1, Q_3]$ という区間の形で示すことも多い。

3.3.3 四分位偏差 (Semi Inter-Quartile Range; SIQR)

四分位範囲の幅を 2 で割った値を四分位偏差と呼ぶ。もしデータが正規分布していれば, 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。

四分位範囲も四分位偏差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点が優れている。

^{*11} 長さ 5 の数値ベクトルなので, `[]` を使えば個々の要素を参照できる。例えば X というデータの第 1 四分位は `fivenum(X)[2]` となる。

^{*12} 詳しく言うと, `fivenum()` 関数はごく単純に並べ替えて四分位数に該当する順位を求め, それが整数ならそのまま, 小数なら前後の (つまり, 上から数えたときと下から数えたときの) 平均を取った値を返す。一方, `quantile()` 関数には順位の計算法が 9 種類含まれており, デフォルトは `S` と同じ `Type 7` である。`fivenum()` と同じ結果を返すのは `Type 2` らしいので, 同じ結果が欲しければ `quantile(X, type=2)` とすればよい。Hyndman, R. J. and Fan, Y. (1996) Sample quantiles in statistical packages, *American Statistician*, 50: 361-365. は `Type 8` を推奨している。

例題

パプアニューギニアのある村で成人男性 28 人の体重を量ったところ、{50.5, 58.0, 47.5, 53.0, 54.5, 61.0, 56.5, 65.5, 56.0, 53.0, 54.0, 56.0, 51.0, 59.0, 44.0, 53.0, 62.5, 55.0, 64.5, 55.0, 67.0, 70.5, 46.5, 63.0, 51.0, 44.5, 57.5, 64.0}(単位は kg) という結果が得られた。このデータから、四分位範囲と四分位偏差を求めよ。

データは `http://phi.med.gunma-u.ac.jp/msb/data/p02.txt` としてサーバに置いてあるので、次のようにする。

c03-2.R

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p02.txt")
Q <- fivenum(dat$WT)
IQR <- Q[4]-Q[2]
SIQR <- IQR/2
cat("四分位範囲=", IQR, " [", Q[2], ", ", Q[4], "], ", "四分位偏差=", SIQR, "\n")
```

3.3.4 平均偏差 (mean deviation)

偏差の絶対値の平均を平均偏差と呼ぶ。四分位範囲や四分位偏差は、全データのうちの限られた情報しか使わないため、分布のばらつきを正しく反映しない可能性がある。すべてのデータを使ってばらつきを表す方法も考えたいわけである。そこで注目したいのが偏差、即ち個々の生の値と平均値との距離である^{*13}。偏差の大きさは、分布のばらつきを反映している。

例えば、 $E = \{11, 12, 13, 14, 15, 16, 17\}$, $F = \{5, 8, 11, 14, 17, 20, 23\}$ は、どちらも平均は 14 だが、 F が E よりもばらつきが大きいことは直感的にわかる。言い換えると、 F の方が E よりも平均値からの距離が大きい。これを 1 つの値として表したいわけである。ただ合計しただけでは、平均値のところで述べたように、偏差の総和は必ずゼロになってしまう。これはマイナス側の偏差がプラス側の偏差と打ち消しあってしまうためなので、偏差の絶対値の総和を出してやればよいというのがもっとも単純な発想である。それだけだと標本サイズが大きいほど大きくなってしまっているので、値 1 つあたりの偏差の絶対値を出してやるために標本サイズで割ることが考えられる。これが平均偏差の考え方である。

*13 誤差と呼ばれることもあるが、誤差の方が意味が広いので、この意味で使う場合は偏差と呼ぶ方がよい。

すなわち、平均偏差 MD は、

$$MD = \frac{\sum |X - \bar{X}|}{n}$$

で定義される。 \bar{X} は平均、 n は標本数である。この例では、E の平均偏差は約 1.71、F の平均偏差は約 5.14 である。これらの値は、次の R プログラムによって計算される。abs() は絶対値をとる関数である。

```
meandev <- function(X) {
  mX <- mean(X)
  sum(abs(X-mX))/length(X) }
meandev(c(11, 12, 13, 14, 15, 16, 17))
meandev(c(5, 8, 11, 14, 17, 20, 23))
```

平均偏差はすべてのデータを使う利点があるが、絶対値を使うために他の統計量との数学的な関係がなく、標本データから母集団統計量を推定するのに使えないという欠点がある。

3.3.5 分散 (variance)

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は、絶対値にする代わりに 2 乗しても良い。つまり、偏差の二乗和 (平方和) の平均をとるわけである。これが分散という値になる。分散 V は、

$$V = \frac{\sum (X - \bar{X})^2}{n}$$

で定義される*14。標本サイズ 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、数値型変数 X の不偏分散は、var(X) によって得られる。

*14 電卓などで計算するときは、これを式変形して得られる $V = \sum X^2/n - \bar{X}^2$ (2 乗の平均値から平均値の 2 乗を引く) という形の方が簡単な上、桁落ちしにくいのでお勧めする。

3.3.6 標準偏差 (standard deviation)

分散の平方根をとったものが標準偏差である。平均値と次元を揃える意味をもつ。不偏分散の平方根をとったものは、不偏標準偏差となる。もしデータが正規分布に従っていれば、 $\text{Mean} \pm 2\text{SD}^{*15}$ 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。R で、数値型変数 X の不偏標準偏差は、 $\text{sd}(X)$ によって得られる。

3.3.7 標準誤差 (standard error) と変動係数 (coefficient of variation)

生データのばらつきの指標ではないが、関連するのでここで示しておく。不偏標準偏差を標本サイズの平方根 \sqrt{n} で割った値は、平均値の推定幅を示す値となり^{*16}、標準誤差 (standard error; SE) として知られている。SD と SE を混用している論文も散見されるが、意味がまったく違う。また、標準偏差 (不偏標準偏差ではないことに注意) を平均で割って 100 を掛けた値を変動係数 (coefficient of variation; C.V.) という。即ち、平均値に対して、全データが何 % ばらついているかを示す、相対的なばらつきの指標である。これは測定データが母集団からのサンプルであることを前提としておらず、測定誤差 (あるいは測定精度というべきかもしれない) を示すときなどに使われる値である。

3.4 まとめ

データの分布は、位置とばらつきを示す 2 つの値で代表させるのが普通である。分布に外れ値が多い・歪みが大きい・尺度水準が低いなどの理由で、分布を仮定できない場合は、中央値と四分位偏差を用い、そうでない場合は平均値と (不偏) 標準偏差を用いて、位置 \pm ばらつき、という形で示すことが多い。

^{*15} 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。

^{*16} 平均の分散は生データの分散の $1/n$ になることと、 n が大きいとき、元の分布によらず平均値の分布は正規分布に近づく (中心極限定理) ため。

3.5 課題

Rにはさまざまなサンプルデータが含まれており、`try(data())`とすると一覧表示できるが、今回は、その中の `sleep` を使ってみる。患者を 2 群に分けて別々の催眠剤（変数 `group` で表される）を与えたときの睡眠時間の変化（変数 `extra`）が得られている。催眠剤 1 と催眠剤 2 それぞれについて、データの分布の位置とばらつきの指標値を計算せよ。

第 4 章

標本統計量と母数推定

4.1 標本統計量と母数

通常、統計解析が相手にするのは標本データの場合が多く、得られる統計量も標本統計量であるが、本当に知りたいのは母集団の統計量（母数）であるため、標本統計量から母数を推定しなければならない。本章ではその仕組みを考える。まず標本抽出をシミュレートしてみる。

4.2 標本抽出

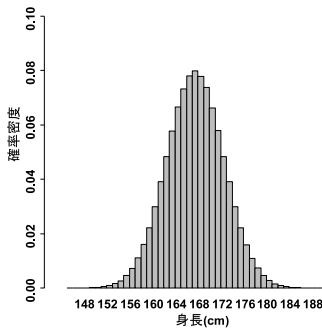
100 万人の成人男性の身長からなる母集団が、表 4-1 のようになっていたとしよう。表 4-1 のように計算すれば、 X の平均値は $\mu = \sum x \cdot p(x) = 167.9998$ 、 X の分散は $\sigma^2 = 25.09521$ である*¹。

ここから 10 個の標本を抽出することを考える。なるべく母集団全体の情報を偏りなく代表させるためには、身長が書かれた 100 万枚のカードを袋に入れてよくかきまぜ、10 枚取り出すことが考えられる。無作為に抽出される最初のカードが x である確率は、 $p(x)$ である。この 1 回目の抽出を X_1 であらわす。2 枚目のカードについても同じである（厳密に言えば、1 枚とっては元に戻して再びよくかきまぜて抽出する

*¹ もっとも、表 4-1 は R の正規乱数を用い、`X <- rnorm(1000000,168,5)` として、母平均 168、母標準偏差 5 の正規乱数を 100 万個発生させた架空のデータなので、この結果は当然である。実行するたびに異なるデータになってしまうのを避けるためには、乱数発生関数 `rnorm()` の前に乱数発生アルゴリズムとその初期値を `RNGkind("Mersenne-Twister",normal.kind="Inversion"); set.seed(1)` などとして固定すればよい。

表 4-1. 成人男子の身長からなる母集団と母平均 μ , 母分散 σ^2 の計算

身長 (cm) (x)	度数	相対度数 ($p(x)$)	$x \cdot p(x)$	$(x - \mu)^2 \cdot p(x)$
144	2	0.000002	0.000288	1.151979e-03
145	3	0.000003	0.000435	1.586970e-03
146	6	0.000006	0.000876	2.903942e-03
147	17	0.000017	0.002499	7.490844e-03
148	19	0.000019	0.002812	7.599834e-03
149	56	0.000056	0.008344	2.021553e-02
150	125	0.000125	0.018750	4.049901e-02
151	219	0.000219	0.033069	6.328937e-02
152	463	0.000463	0.070376	1.185248e-01
153	915	0.000915	0.139995	2.058690e-01
154	1609	0.001609	0.247786	3.153541e-01
155	2649	0.002649	0.410595	4.476659e-01
156	4550	0.004550	0.709800	6.551761e-01
157	7214	0.007214	1.132598	8.728592e-01
158	11005	0.011005	1.738790	1.100452e+00
159	16081	0.016081	2.556879	1.302498e+00
160	22098	0.022098	3.535680	1.414195e+00
161	29903	0.029903	4.814383	1.465155e+00
162	39048	0.039048	6.325776	1.405625e+00
163	48312	0.048312	7.874856	1.207694e+00
164	57703	0.057703	9.463292	9.231469e-01
165	66639	0.066639	10.995435	5.996634e-01
166	73332	0.073332	12.173112	2.932638e-01
167	78051	0.078051	13.034517	7.801682e-02
168	79829	0.079829	13.411272	3.828679e-02
169	77866	0.077866	13.159354	7.790011e-02
170	73767	0.073767	12.540390	2.951326e-02
171	66321	0.066321	11.340891	5.969761e-03
172	57993	0.057993	9.974796	9.279896e-04
173	48410	0.048410	8.374930	1.210356e-04
174	39081	0.039081	6.800094	1.407019e-05
175	29967	0.029967	5.244225	1.468475e-06
176	22055	0.022055	3.881680	1.411597e-07
177	15810	0.015810	2.798370	1.280672e-08
178	10875	0.010875	1.935750	1.087548e-09
179	7309	0.007309	1.308311	8.844242e-11
180	4596	0.004596	0.827280	6.618482e-12
181	2726	0.002726	0.493406	4.607095e-13
182	1519	0.001519	0.276458	2.977333e-14
183	939	0.000939	0.171837	2.112812e-15
184	462	0.000462	0.085008	1.182752e-16
185	224	0.000224	0.041440	6.473767e-18
186	128	0.000128	0.023808	4.147301e-19
187	50	0.000050	0.009350	1.805042e-20
188	31	0.000031	0.005828	1.240027e-21
189	14	0.000014	0.002646	6.174129e-23
190	5	0.000005	0.000950	2.420048e-24
191	4	0.000004	0.000764	2.116040e-25
合計	1000000	1.00	$\mu = 167.9998$	$\sigma^2 = 25.09521$ $\sigma = 5.009512$



「復元抽出」でないと、1枚目に何が出たかに依存して相対度数が変化するため、条件付き確率はまったく同じにはならないが、100万くらいの大きな母集団からサイズ10の標本を取り出すのであれば、「非復元抽出」(1度サンプルを取り出したら元に戻さない)であっても、 $p(x)$ はほとんど変化しないし、1枚目に何がでるかという場合を総当りして合計すれば、2枚目に何が出るかという確率も厳密に $p(x)$ となる。このように、無作為に選び出されるカードは、すべて母集団における相対度数によって与えられる確率で、母集団の値のどれかをとりうる、確率変数である。

サイズ10の標本の標本平均 \bar{X} は、 $\bar{X} = \frac{1}{10}(X_1 + X_2 + \dots + X_{10})$ なので、その期待値は、

$$E(\bar{X}) = \frac{1}{10}(E(X_1) + E(X_2) + \dots + E(X_{10})) = \frac{1}{10}(\mu + \mu + \dots + \mu) = \frac{1}{10} \cdot 10 \cdot \mu = \mu$$

となって母平均に一致する。標本平均の分散は

$$\begin{aligned} V(\bar{X}) &= V\left\{\frac{1}{10}(X_1 + X_2 + \dots + X_{10})\right\} \\ &= \left(\frac{1}{10}\right)^2 \{V(X_1) + V(X_2) + \dots + V(X_{10})\} \\ &= \left(\frac{1}{10}\right)^2 (\sigma^2 + \sigma^2 + \dots + \sigma^2) = \left(\frac{1}{10}\right)^2 \cdot 10 \cdot \sigma^2 \\ &= \frac{\sigma^2}{10} \end{aligned}$$

したがって、標本平均の標準偏差は、

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{10}}$$

となる。これは、標本から推定される母集団の平均値の信頼性を示す値なので、平均の標準誤差と呼ばれる。標本の標準偏差が生データのばらつきを示す値であってサンプルサイズに依存しないのに対し、標準誤差は平均値の信頼性を示す値なので、一般にサンプルサイズを大きくすれば小さくできる値であるため、標準誤差を示す時はサンプルサイズも明記しなくてはならない。

一方、母集団の分散の推定値としての標本の不偏分散は、

$$V(X) = \frac{1}{9} \sum_{i=1}^{10} (X_i - \bar{X})^2$$

である。なぜ $\frac{1}{10}$ でなくて $\frac{1}{9}$ かというと、 \bar{X} は 10 個の X_i からの計算値なので、偏差の情報は見かけ上 10 個あるようにみえるが、9 個分しかないからである（10 個の値の平均と 9 個の値が決まれば 10 個目の値はその計算値として示せてしまう）。厳密な証明はやや高度なので省略するが^{*2}、以下、数値シミュレーションで試してみよう。

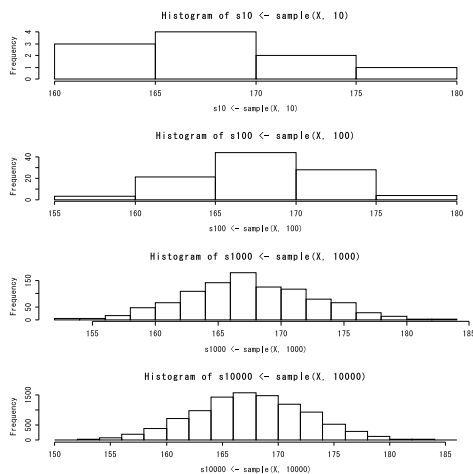
R で標本抽出をする関数は、`sample()` である。`replace=オプション`で復元、非復元を選択できる（デフォルトは非復元）。`sample()` 関数は乱数を使って標本抽出しているので、乱数を初期化して揃えない限り、実行するたびに結果は変わる。`c04-1.R` に示すコードでは、まず HT に身長の階級値、NUM に各階級の人数を付値し、`rep(HT, NUM)` により各 HT が NUM 人ずつ繰り返していることを X に付値して母集団を定義した後、`RNGkind("Mersenne-Twister", normal.kind="Inversion")` で使う乱数の種類をメルセンヌツイスターと指定し（正規乱数の生成法は逆関数法）、`set.seed(1)` を使って乱数の初期値を指定しているので、常に同じ結果が得られる。

次の `layout(c(1,2,3,4))` によってグラフィック画面が縦方向に 4 等分され、続く 4 行の `hist()` で、上から順に 4 つの（母集団 X から 10 人、100 人、1000 人、10000 人を無作為抽出した標本データの）ヒストグラムが描かれる。ヒストグラムを描く行では同時に標本を `s10` などの変数に付値しておく。その変数を使って、最後の 2 行で標本データの平均値と標準偏差を計算し表示している。

c04-1.R

```
HT <- 144:191
NUM <- c(2,3,6,17,19,56,125,219,463,915,1609,2649,4550,7214,11005,
        16081,22098,29903,39048,48312,57703,66639,73332,78051,79829,77866,
        73767,66321,57993,48410,39081,29967,22055,15810,10875,7309,4596,
        2726,1519,939,462,224,128,50,31,14,5,4)
X <- rep(HT, NUM)
RNGkind("Mersenne-Twister", normal.kind="Inversion")
set.seed(1)
layout(c(1,2,3,4))
hist(s10 <- sample(X,10))
hist(s100 <- sample(X,100))
hist(s1000 <- sample(X,1000))
hist(s10000 <- sample(X,10000))
print(c(mean(s10), mean(s100), mean(s1000), mean(s10000)))
print(c(sd(s10), sd(s100), sd(s1000), sd(s10000)))
```

^{*2} 参考書としては、竹村彰通『現代数理統計学』創文社をお薦めする。



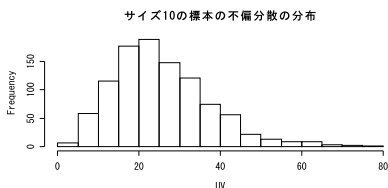
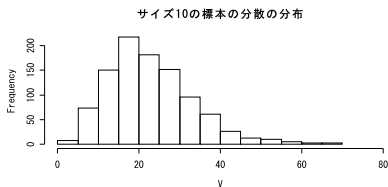
サンプルサイズが大きくなるにつれて、標本分布は母集団の分布に近づく（平均値からみても、不偏標準偏差からみても）。もっとも、ある程度大きなサンプルになると、それほどの差は感じられない。ここで標本の分散でなくて不偏分散の方が母集団の分散に近いことをサンプルサイズが 10 の場合について確かめるには次のようにする。なお、これまで何度か出てきたが、`for () { }` は、`()` 内の条件が満たされている間、`{ }` 内の命令を繰り返せという意味の制御文である。R に限らず、多くのプログラム言語にこうした「繰り返し制御文」は存在し、一般に `for` ループと呼ばれる。母集団のデータを定義して乱数を初期化するところまでは先のコードと同じである。次に 1000 回の無作為抽出について標本の分散と不偏分散をとっておくためにそれぞれ長さ 1000 の配列変数 V と UV を定義する（値はゼロ）。そうしておいてから、`for(i in 1:1000)` によって i が 1 から 1000 まで 1 つずつ増えながら繰り返される `for` ループの中で、母集団からサイズ 10 のサンプルを無作為抽出して s_{10} に付値し、その平均値を計算して m_{10} に付値し、分散と不偏分散を計算してそれぞれ $V[i]$ と $UV[i]$ に付値する。`for` ループを抜けたら作図に入る。まず `layout(c(1,2))` でグラフィック画面を上下 2 分割してから、次の 2 行の `hist()` で分散、不偏分散それぞれのヒストグラムを描き、最後に 1000 個ずつの分散と不偏分散の平均値を計算して表示する。

```

c04-2.R
HT <- 144:191
NUM <- c(2,3,6,17,19,56,125,219,463,915,1609,2649,4550,7214,11005,
16081,22098,29903,39048,48312,57703,66639,73332,78051,79829,77866,
73767,66321,57993,48410,39081,29967,22055,15810,10875,7309,4596,
2726,1519,939,462,224,128,50,31,14,5,4)
X <- rep(HT,NUM)
RNGkind("Mersenne-Twister",normal.kind="Inversion")
set.seed(1)
V <- rep(0,1000)
UV <- rep(0,1000)
for (i in 1:1000) {
  s10 <- sample(X,10)
  m10 <- mean(s10)
  V[i] <- sum((s10-m10)^2)/10
  UV[i] <- sum((s10-m10)^2)/9
}
layout(c(1,2))
hist(V,main="サイズ 10 の標本の分散の分布",xlim=c(0,80))
hist(UV,main="サイズ 10 の標本の不偏分散の分布",xlim=c(0,80))
print(c(mean(V),mean(UV)))

```

図をみると、標本の分散も不偏分散も右裾を引いた分布になっていて、分布の位置がずれていることがわかる。標本の分散の期待値よりも不偏分散の平均（つまり期待値）の方が母集団の分散に近いことが確認できる。



4.3 中心極限定理

さて、このような標本抽出のプロセスでは、実行するたびに標本抽出されるサンプルは異なるが、その平均値が一定の法則に従うことが知られている。つまり、ほとんどの任意の母集団から抽出された無作為標本の平均 \bar{X} の分布は、標本の大きさ n が増大するにつれて、平均 μ 、標準偏差 σ/\sqrt{n} の正規分布に近づく*3これを中心極限定理という。

厳密な証明は省略し、R で試してみることにする。次のプログラムは 0 から 100 まで一様分布する乱数 5000 個と平均 100、標準偏差 10 の正規分布に従う乱数 5000 個をあわせた母集団（元の分布）から、標本の大きさ（サンプルサイズ）を増やしなから標本抽出するものである。

c04-3.R

```
X <- c(runif(5000,0,100),rnorm(5000,100,10))
tsd <- function(X) { sqrt(var(X)*(length(X)-1)/length(X)) }
layout(matrix(c(1,3,2,4),2,2))
hist(X,xlim=c(0,140),freq=F,main="元の分布")
Z5 <- rep(0,1000)
for (i in 1:1000) { Z5[i] <- mean(sample(X,5)) }
hist(Z5,xlim=c(0,140),freq=F)
Y2 <- dnorm(0:140,mean(X),tsd(X)/sqrt(5))
lines(0:140,Y2,col="red")
Z30 <- rep(0,1000)
for (i in 1:1000) { Z30[i] <- mean(sample(X,30)) }
hist(Z30,xlim=c(0,140),freq=F)
Y3 <- dnorm(0:140,mean(X),tsd(X)/sqrt(30))
lines(0:140,Y3,col="red")
Z200 <- rep(0,1000)
for (i in 1:1000) { Z200[i] <- mean(sample(X,200)) }
hist(Z200,xlim=c(0,140),freq=F)
Y4 <- dnorm(0:140,mean(X),tsd(X)/sqrt(200))
lines(0:140,Y4,col="red")
```

layout(matrix(c(1,3,2,4),2,2)) はグラフィック画面を 2 行 2 列に分割し、1 番目の図を左上、3 番目の図を左下、2 番目の図を右上、4 番目の図を右下に描く

*3 μ は母平均、 σ は母集団の標準偏差である。

ように指定する命令である。それぞれの図は `hist()` でヒストグラムを描いた後に、`lines()` に `dnorm()` を与えて正規分布の曲線を赤で (`col="red"`) 重ね描きさせている。

サンプルサイズが大きくなるほど、正規分布に近づくと同時に、標本平均の標準偏差（つまり標準誤差）が小さくなっていき、ヒストグラムの幅が狭くなって、理論分布（赤い線）に近づくのが一目瞭然である。

前章で示したように、生データのばらつきを示す指標として標準偏差が適切なのは、生データの分布が正規分布に近い場合に限られるが、中心極限定理から考えると、平均値の信頼性を示す値としての標準誤差は、データの分布によらず、ある程度サンプルサイズが大きければ常に適切な指標となる。

4.4 信頼区間

標本平均の期待値は母平均に一致するので、標本平均は母平均のよい点推定量になっている。しかし、実際に得られた標本平均は、母平均に完全には一致しないのが普通であろう。そこで、それがどれくらい確からしい推定かを考えることに意味が出てくる。一つの手としては、標本から計算されるある区間の中に、正しく μ を含む割合が 95% であるような区間を推定する（つまり、95% の信頼度をもって区間推定をする）ことが考えられる。

標本サイズが大きい場合は、 \bar{X} の正規分布において、ちょうど 95% の確率を囲むような最短の範囲を選択すればよい。そのために、正規分布の両側の裾 2.5% 分を除く中央部分をとることができる。この部分は、母平均 μ を中心として、大きい方と小さい方に、 \bar{X} の標準偏差（つまり標準誤差）の 1.96 倍（1.96 は標準正規分布の 97.5% 点である）だけ動かした範囲を含む。つまり、

$$\Pr\left(\mu - 1.96 \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

である。この式を変形すると、

$$\Pr\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

となる。 σ は未知なので、通常、標本の不偏標準偏差 $\text{sd}(X)$ を用いる。つまり、 μ の 95% 信頼区間は、 $\bar{X} - 1.96\text{sd}(X)/\sqrt{n}$ から $\bar{X} + 1.96\text{sd}(X)/\sqrt{n}$ までとなる。

標本サイズが小さい場合は、未知の σ に対して $\text{sd}(X)$ を代入することが無視でき

ない誤差の原因となる。そこで、 $\text{sd}(X)$ を使って定義される t という統計量、

$$t = \frac{\bar{X} - \mu}{\text{sd}(X)/\sqrt{n}}$$

が自由度 $n - 1$ の t 分布に従うことから、 t 分布の 2.5% 点から 97.5% 点までを 95% 信頼区間とすることで、この誤差を回避することができる。自由度が無限大になれば t 分布は正規分布に一致するので、結局、常にこちらで計算すればよいことになる。

すなわち、標本サイズ n 、標本平均 \bar{X} 、標本の不偏標準偏差 $\text{sd}(X)$ のとき、母平均の 95% 信頼区間は、

$$\bar{X} - t_{0.025}\text{sd}(X)/\sqrt{n}$$

から

$$\bar{X} + t_{0.025}\text{sd}(X)/\sqrt{n}$$

までになる。R で自由度 $n - 1$ の t 分布の 97.5% 点を与える関数は $\text{qt}(0.975, n-1)$ なので、R のプログラム上で、例えば `c04-1.R` で定義した X からのサイズ 100 のサンプル `s100` から母平均の 95% 信頼区間を推定するには、次のようにすればよい。

c04-4.R

```
HT <- 144:191
NUM <- c(2,3,6,17,19,56,125,219,463,915,1609,2649,4550,7214,11005,
16081,22098,29903,39048,48312,57703,66639,73332,78051,79829,77866,
73767,66321,57993,48410,39081,29967,22055,15810,10875,7309,4596,
2726,1519,939,462,224,128,50,31,14,5,4)
X <- rep(HT,NUM)
RNGkind("Mersenne-Twister",normal.kind="Inversion")
set.seed(1)
s100 <- sample(X,100)
barX <- mean(s100)
sdX <- sd(s100)
t975 <- qt(0.975,length(s100)-1)
rootn <- sqrt(length(s100))
print(barX - t975*sdX/rootn)
print(barX + t975*sdX/rootn)
```

この結果から、95% 信頼区間を表示するときは、`[167.47, 169.19]` のように記載するのが普通である。ただし、もちろん、R にはもっと楽にこの計算をしてくれる関数

が用意されていて、標本データが付値されている数値型変数 X について、`t.test(X)` とすれば、母平均がゼロという帰無仮説の検定結果（検定については次章で説明する）とともに、母平均の推定値と母平均の 95% 信頼区間が表示される。`t.test(s100)` が表示する母平均の推定値と 95% 信頼区間は、上のコードで得られる結果と一致する。

4.5 自由度

不偏分散の計算のところで出てきたが、自由度について直感的に理解するためには、次のように考えることもできる。 n 個の数値からなる標本には、はじめ n の自由度がある。しかし、自由度の 1 つは、 \bar{X} を計算するときに使われてしまい、`sd(X)` を計算するための偏差 $X_i - \bar{X}$ に対しては $n - 1$ の自由度しか残らない。逆に見れば、サイズ n の標本について、はじめの $n - 1$ 個の偏差は自由だが、最後の偏差は、全偏差の合計がゼロにならねばならない（そのように平均という量はとられている）ために決定済みであり、結局自由に決められる個数は $n - 1$ となる。

一般には、データの数から、推定した母数の数を引いた値が、その統計量や分布の自由度になると考えればよい。

4.6 課題

50 から 99 までの値が 1000 ずつ、合計 50000 個の値からなる母集団があるとする。そこから 5 個のサンプルを 100 回抽出した時と 25 個のサンプルを 100 回抽出したときの標本平均の分布を図示し、2 つの分布を比較して考察せよ。

第5章

データの分布と検定のご概念

5.1 はじめに

量的なデータの場台はヒストグラムや正規確率プロットによって分布の様子をみることができるし、カテゴリデータの場台は各カテゴリの度数分布図（あるいは割合）をみることによって、分布の様子をみることができる。本章ではいくつかの理論分布を紹介し、データの分布が理論分布に適合しているかどうかを調べる検定法を紹介し、併せて検定の考え方そのものにも説明を加える。仮説検定という考え方は、伝統的な統計解析の中では、かなり重要な部分を占めてきたので、ここできちんと整理しておく。

5.2 ベルヌーイ試行と2項分布

まずはカテゴリデータの分布から説明する。1回の実験で事象 S が事象 F のどちらかが起こり、しかもそれらが起こる可能性が、 $\Pr(S) = p, \Pr(F) = 1 - p = q$ で何回実験しても変わらないとき、これをベルヌーイ試行という。ベルヌーイ試行では、事象 F は事象 S の余事象になっている。

例えば、不透明な袋に黒い玉と白い玉が500個ずつ入っていて、そこから中を見ないで1つの玉を取り出して色を記録して（事象 S は「玉の色が黒」、事象 F は「玉の色が白」）袋に戻す実験はベルヌーイ試行である*1。

*1 注：袋に戻さないと1回実験するごとに事象の生起確率が変わっていくのでベルヌーイ試行にならない。なお、サンプリングとみれば、これは復元抽出である。

ベルヌーイ試行を n 回行って、 S がちょうど k 回起こる確率は、

$$\Pr(X = k) = {}_n C_k p^k q^{n-k}$$

である。 ${}_n C_k$ は言うまでもなく n 個のものから k 個を取り出す組み合わせの数である。2 項係数と呼ばれる。このような確率変数 X は、「2 項分布に従う」といい、 $X \sim B(n, p)$ と表す。 $E(X) = np$, $V(X) = npq$ である。

5.3 2 項分布のシミュレーション

正二十面体サイコロ（各面には 1 から 20 までの数字が割り振られているものとする*2）を n 回（ $n = 4, 10, 20, 50$ ）投げたときの、1 から 4 までの目が出る回数を 1 試行と考えれば、これはベルヌーイ試行である。1 回投げたときに 1 から 4 までの目が出る確率は理論的には 0.2（= 母比率は 0.2）と考えられるので、試行 1000 セットの度数分布を描く R のプログラムは次のようになる。最初の関数定義 `times <- function(n) {}` は、正二十面体サイコロを n 回振ったときに目が 4 以下の回数をカウントする関数を定義している。中で使っている `ifelse(condition, res1, res2)` は、`condition` が真なら `res1` を、偽なら `res2` を返す関数である。ベクトルに対しても使えるのが便利である。

c05-1.R

```
times <- function(n) {
  dice <- as.integer(runif(n,1,21))
  hit <- sum(ifelse(dice<5,1,0))
  return(hit)}

a <- c(4,10,20,50)
layout(matrix(1:4,nr=2))
for (i in 1:4) {
  y <- 1:1000
  for (k in 1:1000) { y[k] <- times(a[i]) }
  barplot(table(y),main=paste("n=",a[i]))
}
```

*2 ただし、通常売られている正二十面体サイコロは、0 から 9 が 2 個ずつ刻印されているようである。

5.4 2項分布の理論分布

この例で、各 n についての理論的な確率分布は

$$\Pr(X = k) = {}_n C_k 0.2^k 0.8^{n-k}$$

である。R では `choose(n,k)` が n 個の中から k 個を選び出す組み合わせの種類数を返す関数なので、図を描くための R のプログラムは下記の通りとなる。

```
c05-2.R
layout(matrix(1:4,nr=2))
a <- c(4,10,20,50)
for (i in 1:4) {
  n <- a[i]
  k <- 0
  chk <- 1:(n+1)
  names(chk) <- 0:n
  while (k <= n) {
    chk[k+1] <- choose(n,k)*(0.2^k)*(0.8^(n-k))
    k <- k+1
  }
  barplot(chk,main=paste("n=",n))
}
```

ただし、前にも触れた `dnorm()` や `dt()` など、R には様々な確率分布についての関数があり、`choose(n,k)*(0.2^k)*(0.8^(n-k))` は `dbinom(k,n,0.2)` と同値である。このように、確率変数が取りうる各値に対して、その値をとる確率を与える関数を確率密度関数 (probability density function) という。値が小さいほうからそれを全部足した値を与える関数 (つまり、その確率変数の標本空間の下限から各値までの確率密度関数の定積分) を分布関数 (あるいは確率母関数 (probability generating function)、累積確率密度関数) と呼ぶ。

2項分布の確率変数の定義域は整数値なので、飛び飛びの値となる。その意味で、このような分布を離散分布という。離散分布には、2項分布の他には、ポアソン分布などがある*3。それに対して、正規分布や t 分布など、確率変数の定義域が実数であ

*3 ポアソン分布は、独立した事象の生起件数の分布であり、例えば、意図的な出産抑制がまったくない人類集団での完結出生児数の分布はポアソン分布に従うことが期待される。

る分布を、連続分布という。

5.5 正規分布

n が非常に大きい場合は、2 項分布 $B(n, p)$ の確率 $\Pr(X = np + d)$ という値が、

$$\frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{d^2}{2npq}\right)$$

で近似できる。一般にこの極限 (n を無限大に限りなく近づけた場合) である、

$$\Pr(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

という形をもつ確率分布を正規分布と呼び、 $N(\mu, \sigma^2)$ と書く。

$z = (x - \mu)/\sigma$ と置けば、

$$\Pr(Z = z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

となる。これを標準正規分布と呼び、 $N(0, 1)$ と書く。

既に示したように、R で標準正規分布の確率密度関数を $[-5, 5]$ の範囲でプロットするには、`curve(dnorm(x), -5, 5)` と打てば良い。`curve()` 関数は、連続分布をプロットするときに、定義域を x として、始点と終点をコマで区切って与えれば曲線を描画してくれる、便利な関数である。重ね描きする場合は、`add=T` を引数リストに加える。例えば、いま描いた標準正規分布の確率密度関数のグラフの上に、同じ範囲で平均値 1、標準偏差 2 の正規分布の確率密度関数を赤い破線で重ね描きするには、`curve(dnorm(x, 1, 2), add=T, col="red", lty=2)` とすればよい。

標準正規分布の 97.5% 点 (その点より小さい値をとる確率が 0.975 になるような点。その点を与える関数を分位点関数と呼ぶ) を得るには、`qnorm(0.975)` とすればよいし、-1.96 より小さな値をとる確率を得るには、`pnorm(-1.96)` とすればよい。

R では一般に、分布名が `fable` だとすると (注: 念のため書いておくと `fable` などという名前の分布は存在しないが)、確率密度関数が `dfable()`、確率母関数が `pfable()`、分位点関数が `qfable()` で得られる。また、その分布に従う n 個の乱数を得るには、`rfable(n)` とする。

5.6 χ^2 分布

X_1, X_2, \dots, X_v が互いに独立に標準正規分布 $N(0, 1)$ に従うとき,

$$V = \sum_{i=1}^v X_i^2$$

の分布を自由度 v の χ^2 分布 (カイ二乗分布) という。この分布の確率密度関数は,

$$f(x|v) = \frac{1}{2\Gamma(v/2)} \left(\frac{v}{2}\right)^{v/2-1} \exp\left(-\frac{v}{2}\right)$$

である。

なお, 言うまでもないが, Γ はガンマ関数で, 正の実数 α に対して,

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} \exp(-x) dx$$

であり, 正の整数 α に対しては $\Gamma(\alpha) = (\alpha - 1)!$ である。

χ^2 分布においては期待値 $E(x) = v$ であり, 分散 $V(x) = 2v$ である。自由度 1 の χ^2 分布を $[0, 10]$ の範囲でプロットするには, `curve(dchisq(x, 1), 0, 10)` とすればよい。他の自由度のものを重ね描きするには, 例えば自由度 2 の χ^2 分布を赤破線で重ね描きしたければ, `curve(dchisq(x, 2), 0, 10, add=T, col="red", lty=2)` とすればよい。

自由度 1 の χ^2 分布の 95% 点を得るには, `qchisq(0.95, 1)` とすればよいし, 3.84 より小さな値をとる確率を得るには, `pchisq(3.84, 1)` とすればよい。

5.7 t 分布

標準正規分布に従う確率変数 U と, 自由度 v の χ^2 分布 $\chi^2(v)$ に従う確率変数 V があり, それらが独立のとき,

$$T = U/\sqrt{V/v}$$

に従う分布のことをステューデントの t 分布という。この確率密度関数は

$$f(t) = \frac{\Gamma((v+1)/2)}{\sqrt{v}\Gamma(1/2)\Gamma(v/2)} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}$$

である。これは、ステューデントというペンネームで論文を書いていたギネス社の技師ゴセット (Gosset WS) が初めて導いた分布である。

自由度 20 の t 分布の確率密度関数を $[-5, 5]$ の範囲でプロットするには、`curve(dt(x, 20), -5, 5)` とすればよい。これが標準正規分布より裾が長い分布であることを見るために標準正規分布を赤い点線で重ね描きするには、続けて、`curve(dnorm(x), -5, 5, add=T, col="red", lty=2)` とすればよい。

また、自由度 20 の t 分布の 97.5% 点を得るには、`qt(0.975, 20)` とすればよいし、2 より小さな値をとる確率を得るには、`pt(2, 20)` とすればよい。

5.8 F 分布

V_1 と V_2 が独立で、自由度がそれぞれ ν_1, ν_2 の χ^2 分布に従う統計量であるとする。このとき、

$$F = \frac{V_1/\nu_1}{V_2/\nu_2}$$

が従う分布を自由度 (ν_1, ν_2) の F 分布という。 F 分布の確率密度関数は、

$$f(F) = \frac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{F^{(\nu_1/2)-1}}{\left(1 + \frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}$$

で与えられる。 $B(\alpha, \beta)$ はベータ関数で、ガンマ関数を用いれば

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

と書ける。自由度 (ν_1, ν_2) の F 分布を $F(\nu_1, \nu_2)$ と書き、その上側 100% 点を $F_\alpha(\nu_1, \nu_2)$ と書く。R で第 1 自由度 (ν_1) 9, 第 2 自由度 (ν_2) 14 の F 分布の確率密度関数を $[0, 10]$ の範囲でプロットするには、`curve(df(x, 9, 19), 0, 10)` とすればよいし、同じ F 分布の 95% 点 (上側 5% 点) を得るには `qf(0.95, 9, 19)` とすればよいし、同じ F 分布に従う統計量が 5 より小さな値をとる確率を得るには `pf(5, 9, 19)` とすればよい。

5.9 検定の考え方と第 1 種、第 2 種の過誤

検定とは、帰無仮説 (一般には、差がない、という仮説) の下で得られた統計量を、既知の確率分布をもつ量と見た場合に、その値よりも外れた値が得られる確率 (これ

を「有意確率」と呼ぶ)がどれほど小さいかを調べ、有意水準^{*4}より小さければ、統計学的に意味があることと捉え(統計学的に有意である、という)、帰無仮説がおかしいと判断して棄却する(つまり、「差がある」と判断する。一般に「統計学的な有意差 (statistically significant difference) があった」と表現する)という意思決定を行うものである。なお、帰無仮説を棄却できなかった場合は、サンプルサイズが小さいなどの理由で検出力が不足して棄却できないだけかもしれないので、本当に差がないのかどうかはわからない。つまり、「統計学的な有意差はなかった」とは、「差があるとはいえなかった」ということであり、積極的な意思決定には至っていないわけである。

この意思決定が間違っていて、本当は帰無仮説が正しいのに、間違っただけで帰無仮説を棄却してしまう確率の上限値は、有意水準と等しいので、その意味で、有意水準を第1種の過誤と(α エラーとも)呼ぶ(逆に、本当は帰無仮説が正しくないのに、その差を検出できず、有意でないと判断してしまう確率を、第2種の過誤と(β エラーとも)呼び、1から第2種の過誤を引いた値が検出力になる)。

なお、有意確率の小ささは、あくまで、データからみた帰無仮説のありえなさを示すだけであって、差の大きさを意味するのではない。ここを勘違いしたレポートなどが時折見られるので注意されたい。

5.10 両側検定と片側検定

2つの量的変数 X と Y の平均値の差の検定をする場合(平均値の差の検定については次章で詳しく触れる)、それぞれの母平均を μ_X 、 μ_Y と書けば、その推定量は $\mu_X = \text{mean}(X) = \sum X/n$ と $\mu_Y = \text{mean}(Y) = \sum Y/n$ となる。

両側検定では、帰無仮説 $H_0: \mu_X = \mu_Y$ に対して対立仮説(帰無仮説が棄却された場合に採択される仮説) $H_1: \mu_X \neq \mu_Y$ である。 H_1 を書き直すと、「 $\mu_X > \mu_Y$ または $\mu_X < \mu_Y$ 」ということである。つまり、 t_0 を「平均値の差を標準誤差で割った値」として求めると、 t_0 が負になる場合も正になる場合もあるので、有意水準 5% で検定して有意になる場合というのは、 t_0 が負で t 分布の下側 2.5% 点より小さい場合と、 t_0 が正で t 分布の上側 2.5% 点(つまり 97.5% 点)より大きい場合の両方を含む。 t 分布は原点について対称なので、結局両側検定の場合は、上述のように差の絶対値を分子にして、 t_0 の t 分布の上側確率^{*5}を 2 倍すれば有意確率が得られることになる。

^{*4} 分析者が決める一定の確率。当該研究分野の伝統に従うのが普通である。先行研究があればそれに従う。他に基準がなければ 5% か 1% にすることが多い。

^{*5} t 分布の確率密度関数を t_0 から無限大まで積分した値、即ち、 t 分布の分布関数の t_0 のところの値

片側検定は、データ以外の情報から X と Y の間に大小関係が仮定できる場合に言い、例えば、 X の方が Y より小さくなっているかどうかを検定したい場合なら、帰無仮説 $H_0: \mu_X \geq \mu_Y$ に対して対立仮説 $H_1: \mu_X < \mu_Y$ となる。この場合は、 t_0 が正になる場合だけ考えればよい。有意水準 5% で検定して有意になるのは、 t_0 が t 分布の上側 5% 点（つまり 95% 点）より大きい場合である。なお、R で平均値の差の検定を行うための関数は、平均値の信頼区間のところでも出てきた `t.test()` だが、詳しくは次章で説明する。

5.11 分布の正規性の検定

高度な統計解析をするときには、データが正規分布する母集団からのサンプルであるという仮定を置くことが多いが、それを実際に確認することは難しいので、一般には、分布の正規性の検定を行うことが多い。考案者の名前から Shapiro-Wilk の検定と呼ばれるものが代表的である。

5.11.1 Shapiro-Wilk の検定

Shapiro-Wilk の検定の原理をざっと説明すると、 $Z_i = (X_i - \mu)/\sigma$ とおけば、 Z_i が帰無仮説「 X が正規分布にしたがう」の下で $N(0, 1)$ からの標本の順序統計量となり、 $c(i) = E[Z(i)]$ 、 $d_{ij} = Cov(Z(i), Z(j))$ が母数に無関係な定数となるので、「 $X(1) < X(2) < \dots < X(n)$ の $c(1), c(2), \dots, c(n)$ への回帰が線型である」を帰無仮説として、そのモデルの下で σ の最良線型不偏推定量 $\hat{\sigma} = \sum_{i=1}^n a_i X(i)$ と $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$ を用いて、 $W = (k\hat{\sigma}^2)/S^2$ を検定統計量として検定するものである。 k は $\sum_{i=1}^n (ka_i)^2 = 1$ より求められる。

R には `shapiro.test()` という関数が実装されているので、数値型変数 X の分布が正規分布から有意に外れていないかを検定するには、単純に

```
shapiro.test(X)
```

とすればよい。変数 X のデータ数（ベクトルの要素数、R のコードでは `length(X)`）は、3 から 5000 の間でなければならない。2 以下では分布を考える意味がなく、また、検定統計量 W の分布がモンテカルロシミュレーションによって得られたものであるため、あまりに大きなサンプルサイズについては値が与えられていない。

を 1 から引いた値。R では `1-pt(t0, 自由度)`。

例題

http://phi.med.gunma-u.ac.jp/msb/data/p02.txt にあるパプアニューギニア成人男性の体重データは正規分布に従っているといえるか、Shapiro-Wilk の検定をせよ。

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p02.txt")
shapiro.test(dat$WT)
```

とすると、 $W = 0.9799$, $p\text{-value} = 0.8473$ と表示される。R で検定を行なう関数では、大抵の場合、有意確率は $p\text{-value}$ として表示されるが、ここでも $p\text{-value}$ が有意確率を意味している。0.8473 は 0.05 よりずっと大きいので、この成人男性の体重データが正規分布に従っているという帰無仮説の下でこのようなデータが偶然得られることは十分考えられる。したがって、正規分布に従っているという帰無仮説は棄却されない。

5.11.2 Geary の検定

正規性の検定にはたくさんの方法が提案されているが、ここではもう 1 つだけ紹介しておこう。提案者の名前から、ギアリー (Geary) の検定と呼ばれるものである。現在のところ、R にはデフォルトでは入っていないが^{*6}、比較的簡便で使いやすい検定である。以下、原理をざっと説明する^{*7}。

^{*6} moments パッケージに含まれている `geary()` 関数で Geary の尖度統計量 G を計算することができる。また、同じパッケージの `bonett.test()` 関数を使えば、「尖度が $\sqrt{2/\pi}$ に等しくない」を対立仮説とする Bonett-Seier 検定を実行できる。ここで紹介している Geary の検定と似たような有意確率が得られる。

^{*7} この説明は柴田義貞 (1981) 『正規分布 特性と応用』(東京大学出版会) に依拠している。先に用語を説明しておく、 $f(x)$ が 1 次元分布 F の確率密度関数として、正の実数 α について

$$\nu_{\alpha'} = \int_{-\infty}^{\infty} |x|^{\alpha} f(x) dx$$

が有限確定となるとき、この値を分布 F の原点のまわりの α 位の絶対モーメントと呼ぶ。また、

$$\mu_{r'} = \int_{-\infty}^{\infty} x^r f(x) dx$$

を原点のまわりの r 次のモーメントと呼ぶ。 $\mu_{1'}$ は平均値である。さらに、

$$\nu_{\alpha} = \int_{-\infty}^{\infty} |x - \mu_{1'}|^{\alpha} f(x) dx$$

左右対称な分布について、裾の長さを、平均値のまわりの1位の絶対モーメント ν_1 (つまり平均偏差) を平均値のまわりの2次のモーメントの平方根 $\sqrt{\mu_2}$ (つまり標準偏差) で割ったもので測ることにすると、その一致推定量 G は、

$$G = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{\sqrt{n \sum_{i=1}^n (X_i - \bar{X})^2}}$$

となる。この G が Geary の尖度統計量と呼ばれる。 G を用いて帰無仮説 H_0 : 「データ X が正規分布からの標本」を検定することができる。対立仮説の下での分布が正規分布よりも裾が長い対称分布 (例えば t 分布のような) であれば $G < g_0$ のとき帰無仮説を棄却する。 g のパーセント点については、 u_α を標準正規分布の $100\alpha\%$ 点として、 n が大きければ近似的に

$$g(\alpha; n) \simeq \sqrt{\frac{2}{\pi}} + u_\alpha \sqrt{1 - \frac{3}{\pi} \frac{1}{\sqrt{n}}}$$

で得られることがわかっているので、R のプログラムを次のように定義すれば*8、Geary の正規性の検定を行う関数 `geary.test()` ができる。

```
geary.test <- function(X) {
  m.X <- mean(X)
  l.X <- length(X)
  G <- sum(abs(X-m.X))/sqrt(l.X*sum((X-m.X)^2))
  p <- (1-pnorm((G-sqrt(2/pi))/sqrt(1-3/pi)*sqrt(l.X)))*2
  cat("Geary's test for normality:\n G=",G," / p=",p,"\n")
}
```

なお、作図の説明で触れた `hist(X)` で全体の様子を見たり、`qqnorm(X)` をしても、分布の正規性をチェックするにはいい方法である。`qqnorm(X)` で描かれるグラフは、 X が正規分布に従っていれば直線に乗るはずであり、外れているときにどのように外れているかが見える。

を平均値のまわりの α 位の絶対モーメントと呼び、

$$\mu_r = \int_{-\infty}^{\infty} (x - \mu_1)^r f(x) dx$$

を平均値のまわりの r 次のモーメントと呼ぶ。

*8 この定義は <http://phi.med.gunma-u.ac.jp/msb/msb-funcs.R> に含めてある。

5.12 課題

MASS ライブラリに含まれている低体重出生についてのデータフレーム `birthwt` 内に含まれている出生体重を示す変数 (`bwt`) が正規分布に従っていると言えるかどうか、作図により検討した上で検定せよ。なお、MASS ライブラリ内のデータフレームを使うには、最初に `library(MASS)` とすればよい。答えだけでなく手順も書くこと。

第 6 章

2 群の平均値の差の検定

6.1 母平均と標本平均の差の検定

まずは標本平均と母平均の差の検定を扱ってみる。なお、ここでは検定だけを説明するが、R の出力には 95% 信頼区間も表示されるので、統計的な結果としては、こちらの方が実は情報量が多い*1。ただ、疫学の専門誌以外では仮説検定が求められることが多い。おそらく、保健医療の現場ではやるかやらないかの意思決定が求められるため、仮説検定ですぱっと割り切ってしまう方が役に立つからだと思われる。けれども、意思決定を仮説検定に過度に委ねてしまうのは危険であり、データそのものを丁寧にみるが大変重要であることを忘れてはならない。

サイズ n の標本 X について、標本平均 $E(X) = \sum_{i=1}^n X_i/n$ と既知の母平均 μ_X の差の検定は、母分散 V_X が既知のとき、

$$z_0 = \frac{|E(X) - \mu_X|}{\sqrt{V_X/n}}$$

が標準正規分布に従うことを使って検定できる*2。 V_X が未知のときは、標本の不偏

*1 「平均値の差がない」という帰無仮説を有意水準 5% で検定するよりも、平均値の差の 95% 信頼区間を推定する方が情報量が多い。平均値の差の 95% 信頼区間が 0 をまたいでいれば、「差がない」という帰無仮説が棄却されることがわかるので、信頼区間を表示すれば仮説検定の結果も同時に分かる。けれども、区間推定をしたということは、区間そのものの方が、有意差の有無を判別するという意思決定よりも重要だとみなしているということなので、その結果について検定的な解釈をすべきではない。

*2 つまり、 $E(X)$ が、平均値 μ_X 、標準偏差 $\sqrt{V_X/n}$ の正規分布に従うということ。これは中心極限定理そのものである。

分散

$$S_X = \sum_{i=1}^n (X_i - E(X))^2 / (n - 1) = \text{var}(X)$$

を使って、

$$t_0 = \frac{|E(X) - \mu_X|}{\sqrt{S_X/n}}$$

が自由度 $n - 1$ の t 分布に従うことを使って検定できる（ただし、ランダムサンプルで、母集団の分布が正規分布であることを暗黙のうちに仮定している）。つまり、 t_0 が自由度 $n - 1$ の t 分布の 2.5% 点より小さいか 97.5% 点より大きかったら、有意水準 5% で有意差があるとみなす。前章でも述べたように、この場合、帰無仮説が「差がない」であり、対立仮説は「大きい小さい」なので、このような両側検定になる。実用上、両側検定の場合は、 t 分布はゼロに対して左右対称なので、有意確率は、 t_0 に対する確率母関数の値を 1 から引いた上側確率を 2 倍すれば得られる*3。

第4章で示した未知の母平均の信頼区間の推定は、この裏返しである。つまり、母平均の 95% 信頼区間の下限は、不偏分散を標本数 n で割ったものの平方根に自由度 $n - 1$ の t 分布の 97.5% 点を掛けた値を標本平均から引いた値になり、上限は、同じ値を標本平均に足した値になる。

R では、既に示したように、`t.test()` 関数がこれらを両方やってくれる。例えば、量的変数 X が母平均 120 の母集団からのランダムサンプルであるという帰無仮説を検定するには、`t.test(X,mu=120)` とする。`X <- rnorm(100,120,10)` の場合と、`X <- rnorm(100,110,10)` の場合について結果を比べてみるとよい。

例題

平成 10 年の国民栄養調査によれば、50-59 歳男性の平均 BMI (Body Mass Index の略語で、キログラム単位の体重をメートル単位の身長²で割った値) は 23.6 であった。同じ年に A 社の職員健診を受診した 50-59 歳男性 248 人の平均 BMI が 24.6 で、その不偏分散が 8.6 であったとき、A 社の 50-59 歳男性の BMI の平均値は全国平均と差があるといえるかどうか検定せよ。

母分散が未知なので、標本の不偏分散で代用すれば、 $t_0 = |24.6 - 23.6| / \sqrt{8.6/248} = 5.37$ より、自由度 247 の t 分布で 5.37 よりも大きい値をとる確率はほぼ 0 なので、

*3 データ以外の情報によって予め X が母平均より小さくなることはないわかっているときは、小さい側を考えなくてよくなるので、有意確率は t_0 に対する t 分布の確率母関数の値を 1 から引いた上側確率そのものとなるし、95% 信頼区間も 95% 点を考えればよい。このような場合を片側検定とよぶ。

両側検定のために 2 倍しても有意差があるといえる。R のプロンプトに対して以下のように入力すると、有意確率が得られる。

```
t0 <- (24.6-23.6)/sqrt(8.6/248)
2*(1-pt(t0,247))
```

6.2 独立 2 標本の平均値の差の検定

次に、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。

6.2.1 母分散が既知で等しい V である場合 (稀)

この場合は、言い換えると、これらの独立 2 標本が同じ母集団からのサンプルであるというのが帰無仮説になる。 $z_0 = |E(X) - E(Y)| / \sqrt{V/n_X + V/n_Y}$ が標準正規分布に従うことを使って検定する。

6.2.2 母分散が未知の場合 (通常はこちら)

1. F 検定 (2 群の分散に差が無いという帰無仮説の検定) : 2 つの量的変数 X と Y の不偏分散 $SX <- \text{var}(X)$ と $SY <- \text{var}(Y)$ 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 $DFX <- \text{length}(X) - 1$, 第 2 自由度 $DFY <- \text{length}(Y) - 1$ の F 分布に従うことを使って検定する (一般に、互いに独立な分散の比は F 分布に従うと考えてよい)。有意確率は $1 - \text{pf}(F0, DFX, DFY)$ で得られる。しかし、 $F0$ を手計算しなくても、 $\text{var.test}(X, Y)$ で分散に差がないかどうかの検定が実行できる*4。また、1 つの量的変数 Z と 1 つの群分け変数 C があって、 C の 2 群間で Z 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、

*4 「R による統計解析の基礎」では第 3 刷まで、『この場合は、R が勝手に入れ替えてくれるので、 X の不偏分散の方が Y の不偏分散より大きいかどうか気にしなくてもよい。』と書いていたが、実は、古川・丹後「医学への統計学」(朝倉書店)で 2 つの方法の 1 つとして触られている、「帰無仮説: $SX = SY$, 対立仮説: $SX \neq SY$ 」で大小を区別せず F 比を算出して両側検定するのがデフォルトになっているので注意されたい。

`var.test(Z~C)` とすればよい^{*5}。

- 分散に差があるか差がないかによって、平均値が等しいかどうかの検定法は異なる。分散に差があるときは、その事実をもって別の母集団からとられた標本であると判断し、平均値が等しいかどうかを検定する意味はないとする考え方もあるが、一般には Welch の方法を使うか、ノンパラメトリックな方法^{*6}を使って検定する。

6.2.3 分散に差がない場合の検定法

まず母分散 S を $S \leftarrow (DFX * SX + DFY * SY) / (DFX + DFY)$ として推定する (2 つの分散の自由度で重み付けした平均をとる)。

`t0 <- abs(mean(X) - mean(Y)) / sqrt(S / length(X) + S / length(Y))` が自由度 $DFX + DFY$ の t 分布に従うことから、帰無仮説「 X と Y の平均値には差がない」を検定すると、 $(1 - pt(t0, DFX + DFY)) * 2$ が有意確率となる。両側検定なので上側確率を出して 2 倍する。

R では、`t.test(X, Y, var.equal=T)` とする。また、先に触れた量的変数と群分け変数という入力の方の場合は、`t.test(X~C, var.equal=T)` とする。ただしこれだと両側検定なので、片側検定したい場合は、

```
t.test(X, Y, var.equal=T, alternative="less")
```

^{*5} ちなみに、ここで説明した X , Y 型の変数と Z , C 型の変数は相互に変換することが可能である。 X , Y が与えられているときは、

```
Z <- c(X, Y)
C <- factor(c(rep(1, length(X)), rep(2, length(Y))), labels=c("X", "Y"))
```

として Z , C が得られるし、逆に Z , C が与えられていれば、

```
X <- Z[C=="X"]
Y <- Z[C=="Y"]
```

のようにして X , Y を得られる。

^{*6} 詳しくは第 11 章で述べるが、例えば Mann-Whitney の U 検定 (Wilcoxon の順位和検定と数学的に同値) が良く用いられる。その場合は、代表値としても平均値と標準偏差でなく、中央値と四分位範囲または四分位偏差を表示するのが相応しい。ただし、ノンパラメトリックな方法は、本来は分散が異なる場合にはあまり適切でなく、分布が歪んでいたり外れ値がある場合に有効である。

などとする (alternative="less"は対立仮説が $X < Y$ という意味なので、帰無仮説が $X \geq Y$ であることを意味する)。

6.2.4 分散に差がある場合の検定法 (Welch の方法)

分散が異なる場合は、

$$t_0 = \frac{|E(X) - E(Y)|}{\sqrt{S_X/n_X + S_Y/n_Y}}$$

が自由度 ϕ の t 分布に従うことを使って検定する。ただし 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分散が等しくないと仮定して Welch の方法で検定されるので省略して `t.test(X,Y)` でいい。量的変数と群分け変数という入力の仕方の場合には、`t.test(X~C)` とする。

実は、`var.equal` には分散が等しいという命題が真か偽かを判定する論理変数を指定すれば良いので、`t.test(X,Y,var.equal=(var.test(X,Y)$p.value>=0.05))` とすれば、自動的にこれら 2 つの場合に応じた分析が行われる。しかし、ただ機械的にそう実行するのではなく、`stripchart()` や `boxplot()` などで 2 群をプロットし、生データのばらつきと位置の様子を確認した上で、`var.test()` の結果もみて、それに応じて `t.test()` を実行する方がよい^{*7}。

^{*7} なお、Welch の方法を用いるかどうかの事前検定としての分散比の F 検定の有意水準は 5% でなく 20% 程度にせよという意見や、常に Welch の方法を用いるべきという意見、逆に常に普通の t 検定でよいという意見を主張する統計学者もいる。

例題

2001年に、厚生科学研究で「少子化の見通しに関する専門家調査」が行われた。この調査は、「人口学、経済学、家族社会学、公衆衛生学を中心とした専門家を対象として少子化研究会のメンバーが対象候補者を抽出し、回答者の偏りや不足等について検討を加えた上で、748名を対象として調査を実施した」もので、回収率は44%であった。この調査では、2025年の合計出生率（TFR^a）がいくつになるかという予測値があるが、出生率がそのうち回復するとみるか、低下し続けるとみるかという質問項目もあり、この答えの違いによって、2025年のTFRの予測値には違いがあると考えられる。回復するとみる人たちの2025年のTFRの予測値は、1.40 1.40 1.56 1.50 1.40 ...（後略）となっており（サンプルサイズ58、平均1.487、不偏分散0.0275）、低下し続けるとみる人たちの予測値は、1.38 1.30 1.15 1.31 1.37 ...（後略）となっていた（サンプルサイズ221、平均1.356、不偏分散0.0211）。2群の平均値に有意な差があるといえるか、有意水準5%で検定せよ。

^a 女性の年齢別出生率の合計である。分母が総人口である birth rate を普通出生率というのに対して、女性人口を分母とする fertility rate を特殊出生率と呼んだという歴史的経緯から、合計特殊出生率ともいう。英語の total fertility rate の頭文字をとって TFR という呼び名も有名である。

Rで計算するには、まず次の枠内により F 検定を行なう。

```
F0 <- 0.0275/0.0211
1-pf(F0,57,220)
```

すると、0.0915... という結果が得られるので分散に有意水準5%で有意差はないといえる。したがって、Welchの方法でなく通常の t 検定を行う。

```
S <- ((58-1)*0.0275+(221-1)*0.0211)/(58+221-2)
t0 <- abs(1.487-1.356)/sqrt(S/58+S/221)
2*(1-pt(t0,58+221-2))
```

結果として $8.97506e-09$ が得られ（これはコンピュータの浮動小数点表示で、 8.97506×10^{-9} という意味）、5%より遙かに小さいので、出生率の見通しの異なる専門家集団間で、2025年合計出生率の予測値の平均値には有意差があったといえる。

なお、このように、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ことが多い。誤解を生む場合があるので、必ずしもいい図示ではないのだが、伝統的に良く使われている。作図のコードを次の枠内に示す。max(X+SX) は、平均値に不偏標準偏差を足した値について、回復派と低下派

の大きい方を意味する。barplot の中で `ylim=c(0,max(X+SX)*1.5)` としているのは、その値の 1.5 倍が入るように Y 軸の上限をとることを意味する。ただし、1.5 倍にとくに意味はなく、Y 軸の最上端よりもエラーバーの上端が上になってしまうのを防ぐための処置である。

```
X <- c(1.487,1.356)
names(X) <- c("回復派","低下派")
SX <- c(sqrt(0.0275),sqrt(0.0211))
IX <- barplot(X,ylim=c(0,max(X+SX)*1.5),main="専門家の 2025 年 TFR 予想")
arrows(IX,X,IX,X+SX,angle=90)
```

一方、生データがあるときの図示には、`stripchart()` か `boxplot()` 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。例えば、平均値 10、標準偏差 2 の正規乱数 100 個からなる変数 V と、平均値 12、標準偏差 3 の正規乱数 60 個からなる変数 W を比較して図示するためのコードは次の枠内の通り。なお、最後の行の `t.test()` 関数は、これら 2 つの変数の平均に有意差があるかどうかを検定するためのコードである。

c06-1.R

```
RNGkind("Mersenne-Twister")
set.seed(1)
V <- rnorm(100,10,2)
W <- rnorm(60,12,3)
X <- c(V,W)
C <- as.factor(c(rep("V",100),rep("W",60)))
stripchart(X~C,method="jitter",vert=T,ylim=c(0,20))
MX <- tapply(X,C,mean)
SX <- tapply(X,C,sd)
IX <- c(1.1,2.1)
points(IX,MX,pch=18)
arrows(IX,MX-SX,IX,MX+SX,angle=90,code=3)
t.test(V,W,var.equal=(var.test(V,W)$p.value>=0.05))
```

6.3 対応のある 2 標本の平均値の差の検定

先の例題と同じ専門家調査の結果で、2005 年の予測値と 2025 年の予測値に差があるかないかという問題を考えよう。この場合は同じ人について両方の値があるので、

全体の平均に差があるかないかだけを見るのではなく、個人ごとの違いを見るほうが情報量が失われない。このような場合は、独立2標本の平均値の差の検定をするよりも、対応のある2標本として分析する方が切れ味がよい(差の検出力が高い)^{*8}。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、ペア間で値の差を計算し、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)` または `t.test(X-Y,mu=0)` で実行できる(どちらでも等価である)。

2025年の予測値は、1.38 1.50 1.30 ... (後略)であり(回答数は311, 平均値は1.385, 不偏分散は0.0252), 2005年の予測値は、1.30 1.35 1.34 ... (後略)であった(回答数は311, 平均値は1.334, 不偏分散は0.00259)。これを普通に *t* 検定するなら、明らかに分散が異なるので、Welchの方法による検定で $t_0 = 5.37$, 自由度が373.1より両側検定の有意確率は 1.37×10^{-7} となる。一方、対応のある *t* 検定をすると、2025年と2005年の予測値の差が-0.08 -0.15 0.04 ... (後略)となり、サンプル数311, 平均-0.0508, 不偏分散0.0192より, $t_0 = 6.46$ となる。これを自由度310の *t* 分布で上側確率を求めて2倍すれば, $p = 3.942 \times 10^{-10}$ となり, こちらの方が有意確率は小さくなる。「差がない」という帰無仮説のありえなさが, 対応のある *t* 検定の方がよりはっきりするということである。いずれにせよ5%よりずっと小さいので, 2025年の予測値と2005年の予測値は5%水準で有意差があったといえる。

例題

10人の健康な日本人成人男性ボランティアを募り、同じ日の9:00と21:00に採血をして血清鉄濃度(mg/L)を測定した結果が下表のように得られたとする(注:架空のデータである)。9:00と21:00の血清鉄濃度に有意差があるといえるか? 有意水準5%で検定せよ。

時刻 \ 人	1	2	3	4	5	6	7	8	9	10
9:00	0.98	0.87	1.12	1.34	0.88	0.91	1.04	1.21	1.17	1.09
21:00	1.03	0.78	1.04	1.52	0.97	0.84	1.32	1.12	1.09	1.32

対応がある場合の図示は、次のコード `e06-2.R` に示すように、1組づつ線で結ぶことが多い。`plot()` 関数の中で `type="l"` はシンボルは打たずに線のみ引くということを意味する。この関数では1人目だけの線を引き、2人目以降は最後に `lines()` を

^{*8} 分布が歪んでいる場合や、分布が仮定できない場合の対応のある2標本の分布の位置の差があるかどうかを検定するには、ウィルコクソンの符号順位検定を用いる。Rでは `wilcox.test(変数1, 変数2, paired=T)` で実行できる。詳細は第11章で説明する。

for ループで繰り返して描画する。plot() 中の xaxt="n" は X 軸の値のラベルを抑制することを意味する(そうしないと X 軸の値ラベルが 1 と 2 になってしまう)。axis() を使って, "9:00" と "21:00" というラベルを表示しているが, axis() の最初の引数 1 は X 軸を意味する。Y 軸の場合はこの値を 2 にする。

```
c06-2.R
```

```
BX <- c(0.98,0.87,1.12,1.34,0.88,0.91,1.04,1.21,1.17,1.09)
AX <- c(1.03,0.78,1.04,1.52,0.97,0.84,1.32,1.12,1.09,1.32)
t.test(BX,AX,paired=T)
plot(c(1,2),c(BX[1],AX[1]),type="l",ylim=c(0,2),xaxt="n",xlab="",
      ylab="血清鉄濃度 (mg/L)",col=1)
axis(1,1:2,c("9:00","21:00"))
for (j in 2:length(BX)) { lines(c(1,2),c(BX[j],AX[j]),col=j) }
```

3 行目の t.test() により次の枠内の結果が得られ, p-value = 0.3852 が 0.05 よりずっと大きいので, 有意水準 5% で統計学的な有意差はなかったといえる。

```
data: BX and AX
t = -0.9128, df = 9, p-value = 0.3852
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.14609201  0.06209201
sample estimates:
mean of the differences
 -0.042
```

6.4 課題

20 匹の 8 週齢の ICR マウスをランダムに 10 匹ずつ 2 群にわけて, 片方には普通餌を自由に食べさせ, もう片方には高脂肪餌を自由に食べさせ, 飲水, 運動などもとくに制限せずに 1 週間飼育したとする。この 1 週間の前後でのマウスの体重 (g) が, 次の表^{*9}のように得られたとき, 高脂肪餌の摂取は普通餌摂取に比べてマウスの体重を有意に増加させる効果があると言えるかどうか検定せよ。群別の体重変化を図示し

^{*9} <http://phi.med.gunma-u.ac.jp/msb/data/p06.txt> としてダウンロードできる。変数名は, 普通餌開始時が NDS, 普通餌終了時が NDE, 高脂肪餌開始時が HFDS, 高脂肪餌終了時が HFDE となっている。

た上で、帰無仮説を明示し、手順も書くこと。

普通餌		高脂肪餌	
開始時	終了時	開始時	終了時
30.3	31.6	29.5	31.2
28.7	29.4	31.1	34.1
30.2	31.1	30.1	31.7
30.5	31.4	31.3	32.8
30.7	31.4	31.8	34.2
30.4	31.2	30.5	32.3
29.4	30.9	29.9	31.7
29.4	31.0	28.4	30.8
30.0	31.7	29.3	30.3
29.0	29.6	30.4	32.6

第 7 章

一元配置分散分析と多重比較

7.1 多群の平均値を比較する 2 つの思想

前章では 2 群の平均値を比較したが、本章では 3 群以上の（多群の）平均値を比較する方法を説明する。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（既に述べた通り、本当は差がないのに誤って差があると判定してしまう確率）を 5% 未満にしたとしても、3 群以上の比較全体として「少なくとも 1 組の差のある群がある」というと、全体としての第 1 種の過誤が 5% よりずっと大きくなってしまふからである。

この問題を解消するには、(1) 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。

別のアプローチとして、(2) 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は大きくなってしまふ第 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」と呼ばれる。

これら 2 つのアプローチは別々に行うというよりも、段階を踏んで行ふものと考えるのが一般的である*¹。一元配置分散分析やクラスカル=ウォリスの検定によって群

*¹ ただし、永田、吉田 (1997) が指摘するように、段階を踏んで実行すると、ここにまた検定の多重性の問題が生じるので、両方はやるべきではない、という考え方にも一理ある【典拠：永田靖、吉田道弘『統計的多重比較法の基礎』、サイエンティスト社、1997 年】。つまり、厳密に考えれば、群分け変数が量的変数に与える効果があるかどうかを調べたいのか、どの群とどの群の間で量的変数に差があるのかを調べたいのかによって、これら 2 つのアプローチを使い分けるべきかもしれない。

間に何らかの差があると結論されてから、初めて、どの群とどの群の差があるのかを調べるために多重比較を使うというわけである。その意味で、多重比較は *post hoc* な解析と呼ばれることがある。仮に多重比較で有意な結果が出たとしても、一元配置分散分析の結果が有意でなければ、偶然のばらつきの効果が群間の差よりも大きいということなので、特定の群間の差に意味があると結論することはできない。

7.2 一元配置分散分析

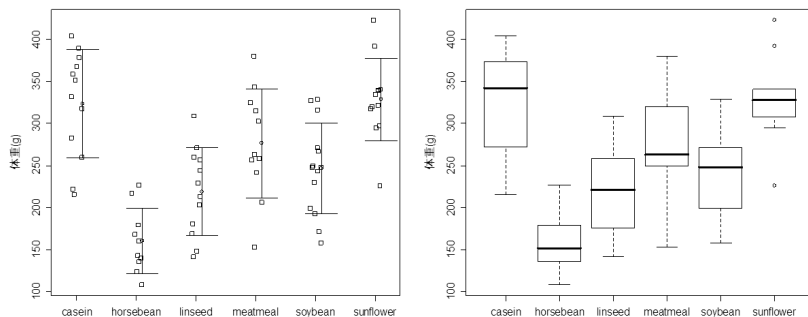
一元配置分散分析の思想は、データのばらつき（変動）を、群間の違いという意味のはっきりしているばらつき（群間変動、あるいは級間変動と呼ばれる）と、各データが群ごとの平均からどれくらいばらついているか（誤差）のすべての群についての合計（誤差変動）とに分解し、前者が後者よりもどれくらい大きいかを検討することによって、「群分け変数がデータの変数に与える効果が誤差に比べて有意に大きいかどうか」を調べるということである。帰無仮説は、「群分け変数がデータの変数に与える効果が誤差の効果に比べて大きくない」ということになる。言い換えると「すべての群の母平均値が等しい」が帰無仮説である。

では、具体的に、R に含まれているデータ `chickwts` で説明しよう。これは、既に一部使ったが、71 羽の鶏を孵化直後にランダムに 6 群に分けて、それぞれ異なる餌（カゼイン [casein]、ソラマメ [horsebean]、アマニの種 [linseed]、肉の配合餌 [meatmeal]、大豆 [soybean]、ヒマワリの種 [sunflower]）を与え、6 週間後に何グラムになったかを示すデータである（R Console で `?chickwts` と入力してヘルプをみると、出典は、Anonymous (1948) *Biometrika*, 35: 214. である）。すべての値を次の表に示す。

餌	その餌を食べて 6 週間育った鶏の体重 (g)
casein	368, 390, 379, 260, 404, 318, 352, 359, 216, 222, 283, 332
horsebean	179, 160, 136, 227, 217, 168, 108, 124, 143, 140
linseed	309, 229, 181, 141, 260, 203, 148, 169, 213, 257, 244, 271
meatmeal	325, 257, 303, 315, 380, 153, 263, 242, 206, 344, 258
soybean	243, 230, 248, 327, 329, 250, 193, 271, 316, 267, 199, 171, 158, 248
sunflower	423, 340, 392, 339, 341, 226, 320, 295, 334, 322, 297, 318

`chickwts` はデータフレームであり、体重を示す数値型変数 `weight` と、餌の種類を示す因子型変数 `feed` という形でデータが入っている。変数が 2 つで、オブザーベーションが 71 個という形になっていることは、`str(chickwts)` とすれば確認で

きる。餌の種類によって鶏の体重に差が出るかをみるためには、まずグラフ表示を試みると、何となく差がありそうに見える。



そこで、群間で体重に差がないという帰無仮説を検定するためには、weight という量的変数に対して、feed という群分け変数の効果を見る形で一元配置分散分析することになる。R Consoleに入力するコマンドは、`summary(aov(weight~feed))` または `anova(lm(weight~feed))` である。どちらでも同じ結果が次の枠内の通りに得られる。後者は、一元配置分散分析が線型モデルの一種であることを利用した書き方だが、ここでは前者を用いる。

```

              Df Sum Sq Mean Sq F value    Pr(>F)
feed           5 231129   46226  15.365 5.936e-10 ***
Residuals     65 195556     3009
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

このような結果の表を分散分析表という。右端の*の数は、最下行の Signif. codes に示されているように有意確率を記号化し(***は有意確率が0.001未満であることを意味する)、有意性を示す目安にしているのだが、有意確率そのもの ($\Pr(>F)$ の下の数字) に注目してみるほうがよい。

Sum Sq は、平方和 (sum of squares) の略である。feed の Sum Sq の値 231129 は、餌の種類が異なる群ごとの平均値から総平均を引いて二乗した値を、餌の種類が異なる群ごとの鶏の個体数で重み付けした和である。群間変動または級間変動と呼ばれ、feed 間でのばらつきの程度を意味する。Residuals の Sum Sq の値 195556 は各鶏の体重から、その鶏が属する餌群の鶏の平均体重を引いて二乗したものの総和で

あり、誤差変動と呼ばれ、餌群によらない（それ以外の要因がないとすれば偶然的）ばらつきの程度を意味する。

Mean Sq は平均平方和 (mean square) の略であり、平方和を自由度 (Df) で割ったものである。平均平方和は不偏分散なので、feed の Mean Sq の値 46226 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 3009 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 5、第 2 自由度 65 の F 分布に従う。一般に、一元配置分散分析の場合は、対立仮説の下では $F > 1$ であることが期待されるため右片側検定すればよい。したがって、分散比がこの実現値よりも偶然大きくなる確率は、 $1 - \text{pf}(15.365, 5, 65)$ で得られる。 $\text{Pr}(>F)$ の下の数字は、まさにその値を示すものである。この例では $5.936e-10$ と*²、ほとんどゼロといえるくらい小さいので、feed 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、鶏の体重は、生後 6 週間に与えた餌の種類によって差があることになる。

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いるため、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、パートレット (Bartlett) の検定と呼ばれる方法がある。R では `bartlett.test`(量的変数~群分け変数) で実行できる。帰無仮説「各群の母分散が等しい」が棄却された場合は、第 11 章で説明するクラスカル=ウォリスの検定のようなノンパラメトリックな方法を使うのが 1 つの方法であるが、むしろ、Welch の方法を一元配置分散分析に拡張した検定法*³を用いるのが最良*⁴である。

この例では、`bartlett.test(weight~feed)` と入力して得られる結果の p-value をみると、0.66 であり、5% よりずっと大きいので帰無仮説は棄却されず、一元配置分散分析を実行しても問題ないことになる。作図と検定を実行する R のコードは次の枠内の通りである*⁵。

*² 注： 5.936×10^{-10} の意味。 $1 - \text{pf}(15.365, 5, 65)$ の結果とは 10^{-13} の位で 1 違うのは F value の丸め誤差による。

*³ `oneway.test()` 関数の `var.equal=FALSE` オプションとして実装されている。以後、本書では「Welch の拡張による一元配置分散分析」と呼ぶ。なお、`oneway.test()` 関数は、`var.equal=TRUE` オプションで実行すれば通常の一元配置分散分析を実行する。

*⁴ <http://aoki2.si.gunma-u.ac.jp/lecture/BF/sankouzu.html> に、群馬大学の青木繁伸教授がシミュレーションによって検討した結果が示されている。

*⁵ ただし、きれいな作図のため、次のようにグラフィックデバイスとして Windows 拡張メタファイルを指定し、図の大きさとフォントも指定した。

```
c07-1.R
attach(chickwts)
mw <- tapply(weight,feed,mean)
sw <- tapply(weight,feed,sd)
ix <- 1:length(table(feed))+0.1
layout(t(1:2))
stripchart(weight~feed,vert=T,method="jitter",ylab="体重 (g)")
points(ix,mw)
arrows(ix,mw-sw,ix,mw+sw,angle=90,code=3)
boxplot(weight~feed,ylab="体重 (g)")
print(res.bt <- bartlett.test(weight~feed))
ifelse(res.bt$p.value<0.05,
cat("不等分散！ Bartlett の検定で p=",res.bt$p.value,"\n",
"Welch の方法による一元配置分散分析の結果で p=",
oneway.test(weight~feed,var.equal=F)$p.value),
summary(aov(weight~feed)))
detach(chickwts)
```

7.3 検定の多重性を調整する「多重比較」

この鶏の体重の例では、一元配置分散分析の結果、餌群の効果が有意だったので、次に調べたいことは、具体的にどの餌とどの餌の間で差がでてくるかであろう。

単純に考えると、2種類の餌ずつ、

layout(t(1:2))の前にwin.metafile("./c07-1.emf",width=12,height=6,pointsize=12)と、par(family="sans")を実行し、作図終了後、つまりboxplot(...)の次の行にdev.off()を実行した。できた拡張メタファイルをIrfanViewで開いて、AdobeのポストスクリプトプリンタドライバであるAcrobat Distiller Jに、用紙サイズを幅12cm、高さ6cm、ポストスクリプトオプションでEPSと指定して出力し、直接Encapsulated Postscriptファイルを作成してpL^AT_EX2_εに取り込んだ。

```

c07-2.R
attach(chickwts)
kf <- names(table(feed))
k <- length(kf)
for (i in 1:(k-1)) { for (j in (i+1):k) {
  cat("** Compare ",kf[i]," and ",kf[j]," **\n")
  print(RV <- var.test(weight[feed==kf[i]],weight[feed==kf[j]]))
  ifelse(RV$p.value<0.05,VRES<-FALSE,VRES<-TRUE)
  print(t.test(weight[feed==kf[i]],weight[feed==kf[j]],var.equal=VRES))
}}
detach(chickwts)

```

と t 検定を繰り返せば良さそうであり、この方法が使われている本や論文もないわけではない。しかし、6種類の餌についてこれをやると6つから2つを取り出す全ての組み合わせについて検定するため、15回の比較をすることになり、個々の検定について有意水準を5%にすると、全体としての第1種の過誤は明らかに5%より大きくなる。したがって、先に述べた通り、 t 検定の繰り返しは不都合である。これに似た方法として無制約LSD(最小有意差)法やFisherの制約つきLSD法(一元配置分散分析を行って有意だった場合にのみLSD法を行うという方法)があるが、これらも第1種の過誤を適切に調整できない(ただし制約つきの場合は3群なら大丈夫)ことがわかっているので、使ってはいけない。現在では、この問題は広く知られているので、 t 検定の繰り返しやLSD法で分析しても論文は受理されない。

多重比較の方法にはいろいろあるが^{*6}、ボンフェローニ(Bonferroni)の方法、シェフェ(Scheffé)の方法、ダンカン(Duncan)の方法、テューキー(Tukey)のHSD、ダネット(Dunnett)の方法、ウィリアムズ(Williams)の方法がよく使われている。このうち、ダンカンの方法は、数学的に間違っていることがわかっているので使ってはいけない。ボンフェローニの方法とシェフェの方法も検出力が低いので、特別な場合を除いては使わない方がよい。データが正規分布に近ければ、テューキーのHSDを使うべきである。ダネットの方法は対照群が存在する場合に対照群と他の群との比較に使われるので、適用場面が限定されている^{*7}。ウィリアムズの方法は対照群があつ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發揮する。

^{*6} 以下、一般論は、中澤 港(2003)「Rによる統計解析の基礎」第10章の内容と基本的に同じである。

^{*7} 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。

上記いくつかの方法がよく使われている理由は、限定された用途で高い検出力をもつダネットとウィリアムズを除けば、たんにそれらが歴史的に古く考案され、昔の統計学の教科書にも説明され、多くのソフトウェアに実装されているからに過ぎない。現在では、かなり広い用途をもち、ノンパラメトリックな分析にも適応可能なホルム (Holm) の方法 (ボンフェローニの方法を改良して開発された方法) が第一に考慮されるべきである。その上で、全ての群間の比較をしたい場合はペリ (Peritz) の方法、対照群との比較をしたいならダネットの逐次棄却型検定 (これはステップダウン法と呼ばれる方法の1つであり、既に触れたダネットの方法とは別) も考慮すればよい。

多重比較においては、帰無仮説が単純ではない。例えば、3群間の差を調べるとしよう。一元配置分散分析での帰無仮説は、 $\mu_1 = \mu_2 = \mu_3$ である。これを包括的帰無仮説と呼び、 $H_{\{1,2,3\}}$ と書くことにする。さて第1群から第3群までの母平均 $\mu_1 \sim \mu_3$ の間で等号関係が成り立つ場合をすべて書き上げると、 $H_{\{1,2,3\}} : \mu_1 = \mu_2 = \mu_3$ 、 $H_{\{1,2\}} : \mu_1 = \mu_2$ 、 $H_{\{1,3\}} : \mu_1 = \mu_3$ 、 $H_{\{2,3\}} : \mu_2 = \mu_3$ の4通りである。このうち、 $H_{\{1,2,3\}}$ 以外のものを部分帰無仮説と呼ぶ。

すべての2つの群の組み合わせについて差を調べるということは、 $\{H_{\{1,2\}}, H_{\{1,3\}}, H_{\{2,3\}}\}$ が、考慮すべき部分帰無仮説の集合となる。第1群が対照群で他の群のそれぞれが第1群と差があるかどうかを調べたい場合は、考慮すべき部分帰無仮説の集合は $\{H_{\{1,2\}}, H_{\{1,3\}}\}$ となって、「すべての2つの群の組み合わせについて調べる」場合とは異なる。これらの集合をその多重比較における「帰無仮説族」と呼ぶ。

ここで多重比較の目的を「帰無仮説族」というコトバを使って言い換えてみる。個々の帰無仮説で有意水準を5%にしてしまうと、帰無仮説族に含まれる帰無仮説のどれか1つが誤って棄却されてしまう確率が5%より大きくなってしまう。それではまずいので、その確率が5%以下になるようにするために、何らかの調整を必要とするわけで、この調整をする方法が多重比較なのである。つまり、帰無仮説族の有意水準を定める (例えば5%にする) ことが、多重比較の目的である。このことからわかるように、差のなさそうな群をわざと入れておいて帰無仮説族を棄却されにくくしたり、事後的に帰無仮説を追加したりすることは、統計を悪用していることになり、やってはいけない。

計算方法については、ボンフェローニとホルム、テューキーのHSDだけを簡単に紹介する。より詳しく知りたい場合には、永田、吉田 (1997) を参照されたい。

7.3.1 ボンフェローニの方法

ボンフェローニの方法とは、ボンフェローニの不等式に基づく多重比較法である。きわめて単純な考え方に基づいているために、適用可能な範囲が広いが、検出力が落ちてしまいがちなので、ベストな方法ではない。ボンフェローニの不等式とは、 k 個の事象 E_i ($i = 1, 2, \dots, k$) に対して成り立つ、

$$Pr(\cup_{i=1}^k E_i) \leq \sum_{i=1}^k Pr(E_i)$$

をいう。左辺は k 個の事象 E_i のうち少なくとも1つが成り立つ確率を示し、右辺は各事象 E_i が成り立つ確率を加え合わせたものなので、この式が成り立つことは自明であろう（個々の事象がすべて独立な場合にのみ等号が成立する）。

次に、この不等式を多重比較にどうやって応用するかを示す。まず、帰無仮説族を $\{H_{01}, H_{02}, \dots, H_{0k}\}$ とする。 E_i を「正しい帰無仮説 H_{0i} が誤って棄却される事象」と考える。この表現をボンフェローニの不等式にあてはめれば、

$$\begin{aligned} & Pr(\text{正しい帰無仮説のうちの少なくとも1つが誤って棄却される}) \\ & \leq \sum_{i=1}^k Pr(\text{正しい帰無仮説 } H_{0i} \text{ が誤って棄却される}) \end{aligned}$$

右辺が α 以下になるためには、もっとも単純に考えれば、足しあわされる各項が α/k に等しいかより小さければよい。つまり、ボンフェローニの方法とは、有意水準 α で帰無仮説族を検定するために、個々の帰無仮説の有意水準を α/k にするものである*8。手順としてまとめると、以下の通りである。

1. 帰無仮説族を明示し、そこに含まれる帰無仮説の個数 k を求める。
2. 帰無仮説族についての有意水準 α を定める。 $\alpha = 0.05$ または $\alpha = 0.01$ と定めることが多い。
3. 帰無仮説族に含まれているそれぞれの帰無仮説に対して検定統計量 T_i ($i = 1, 2, \dots, k$) を選定する。
4. データを取り、検定統計量 T_i を計算する。

*8 ここで注意しなければいけないことは、検定すべき帰無仮説族に含まれる個々の帰無仮説は、データをとるまえに定められていなければならないことである。データをとった後で有意になりそうな帰無仮説を k 個とってきて帰無仮説族を構成するのは、帰無仮説族に対しての第1種の過誤をコントロールできないので不適切である。

5. 各検定統計量 T_i について有意水準 α/k に対応する棄却限界値（通常は分布関数の $(1 - \alpha/k) \times 100\%$ 点）を c_i とするとき、 $T_i \geq c_i$ ならば H_{0i} を棄却し、 $T_i < c_i$ なら H_{0i} を保留する（採択ではない）。

なお R の `pairwise.t.test()` 関数など（後で説明するが `pairwise.prop.test()` 関数 `pairwise.wilcox.test()` 関数でも同様）の `p.adjust.method="bonferroni"` では、各々の帰無仮説の有意水準を α/k とする代わりに、各々の帰無仮説に対して得られる有意確率が k 倍されて（ただし 1 を超えるときは 1 として）表示されるので、各々の比較に対して表示される有意確率と帰無仮説族について設定したい有意水準との大小によって仮説の棄却 / 保留を判断してよい。

7.3.2 ホルムの方法

ボンフェローニの方法では、すべての H_{0i} について有意水準を α/k としたのが良くなかったので、ホルムの方法は、そこを改良したものである。以下、ホルムの方法の手順をまとめる。

1. 帰無仮説族を明示し、そこに含まれる帰無仮説の個数 k を求める。
2. 帰無仮説族についての有意水準 α を定める。 $\alpha = 0.05$ または $\alpha = 0.01$ と定めることが多い。ここまではボンフェローニの方法と同じである。
3. $\alpha_1 = \alpha/k$, $\alpha_2 = \alpha/(k-1)$, ..., $\alpha_k = \alpha$ を計算する。
4. 帰無仮説族に含まれているそれぞれの帰無仮説に対して検定統計量 T_i ($i = 1, 2, \dots, k$) を選定する。
5. データを取り、検定統計量 T_i を計算する。
6. 各検定統計量 T_i について有意確率 P_i を求め、小さい順に並べ換える。
7. P_i の小さいほうから順に α_i と P_i の大小を比べる。
8. $P_i > \alpha_i$ ならばそれよりも有意確率が大きい場合の帰無仮説をすべて保留して終了する。 $P_i \leq \alpha_i$ なら H_{0i} を棄却して、次に小さい P_i について比較する。 $i = k$ となるまで繰り返す。

ホルムの方法についても、R の多重比較の `p.adjust.method="holm"` オプションでは^{*9}、手順 7. で P_i と α_i の大小を比べる代わりに $P'_i = P_i \times (k - i + 1)$ が表示されるので、値そのものを有意水準と比較すればよい。

*9 デフォルトがホルムの方法なので、`p.adjust.method` を指定しなければホルムの方法になる。

7.3.3 テューキーのHSD

テューキーのHSDでは、母集団の分布は正規分布とし、すべての群を通して母分散は等しいと仮定する。

データが第1群から第 a 群まであって、各々が n_i 個($i = 1, 2, \dots, a$)のデータからなるものとする。第 i 群の j 番目のデータを x_{ij} と書くことにすると、第 i 群の平均値 \bar{x}_i と分散 V_i は、

$$\begin{aligned}\bar{x}_i &= \sum_{j=1}^{n_i} x_{ij} / n_i \\ V_i &= \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 / (n_i - 1)\end{aligned}$$

となり、誤差自由度 P_E と誤差分散 V_E は、

$$\begin{aligned}P_E &= N - a = n_1 + n_2 + \dots + n_a - a \\ V_E &= \sum_{i=1}^a \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 / P_E = \sum_{i=1}^a (n_i - 1) V_i / P_E\end{aligned}$$

で得られる。

簡単にいえば、テューキーのHSDは、すべての群間の比較について誤差分散を使った t_0 統計量を計算し、 t 分布ではなくて、ステューデント化された範囲の分布(Studentized range distribution)と呼ばれる分布の $(1 - \alpha) \times 100\%$ 点を $\sqrt{2}$ で割った値との大小で有意水準 α の検定をする方法である。以下手順としてまとめる。

1. 帰無仮説族を明示する。テューキーのHSDの場合は、通常、

$$\{H_{\{1,2\}}, H_{\{1,3\}}, \dots, H_{\{1,a\}}, H_{\{2,3\}}, \dots, H_{\{a-1,a\}}\}$$

2. 有意水準 α を定める。 $\alpha = 0.05$ または $\alpha = 0.01$ と定めることが多い。
3. データを取り、すべての群について \bar{x}_i, V_i を計算し、 P_E, V_E を計算する。
4. すべての2群間の組み合わせについて、検定統計量 t_{ij} を

$$t_{ij} = (\bar{x}_i - \bar{x}_j) / \sqrt{V_E(1/n_i + 1/n_j)}$$

により計算する($i, j = 1, 2, \dots, a; i < j$)

5. $|t_{ij}| \geq q(a, P_E; \alpha) / \sqrt{2}$ なら $H_{\{i,j\}}$ を棄却し、 i 群と j 群の平均値には差があると判断する(比較の形からわかるように、これは両側検定である)。 $|t_{ij}| < q(a, P_E; \alpha) / \sqrt{2}$ なら $H_{\{i,j\}}$ を保留する。ここで $q(a, P_E; \alpha)$ は、群数 a 、自由度 P_E のステューデント化された範囲の分布の $(1 - \alpha) \times 100\%$ 点である。つまり、 $\alpha = 0.05$ ならば、 $q(a, P_E, 0.05)$ は、群数 a 、自由度 P_E の

チューデント化された範囲の分布の 95% 点である。R では、この値を与える分位点関数は、群数 a 、自由度 df として、`qtukey(0.95, a, df)` だが、すべての群間比較を手計算するのは面倒なので、`TukeyHSD()` 関数を使って自動的に実行させるのが普通である。

上の鶏の体重の例について、実際に R で多重比較をしてみよう。

```
attach(chickwts)
pairwise.t.test(weight, feed, p.adjust.method="bonferroni")
detach(chickwts)
```

とすれば、ボンフェローニの方法で有意水準を調整した、すべての餌群間の体重の差を t 検定した結果の有意確率を、下三角行列の形で出してくれる*¹⁰。もし `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。データが正規分布に従っていれば、`TukeyHSD(aov(weight~feed))` としてテューキーの HSD を行ってもよい。`TukeyHSD()` 関数の引数は、通常、`aov()` の結果の分散分析表である*¹¹。実行結果は、2 群ずつの組み合わせのそれぞれについて、テューキーの HSD で調整した、差の 95% 信頼区間という形で得られる。`lwr` と `upr` の数値の間にゼロが含まれない組み合わせには有意水準 5% で有意差がある。

なお、CRAN から `multcomp` パッケージをインストールしておけば他の多重比較も可能である*¹²。

*¹⁰ ただし、 t 検定とは言っても、`pool.sd=F` というオプションをつけない限りは、 t_0 を計算するときに全体の誤差分散を使うので、ただの t 検定の繰り返しとは違う。

*¹¹ ここでは一元配置の場合しか説明しなかったが、すべての組み合わせのサンプルサイズが等しければ、二元配置分散分析でも同様に実行できる。

*¹² 管理者権限があれば、`install.packages("multcomp", dep=T)` でインストールできる。`library(multcomp)` と打ってから、`simtest(weight~feed, type="Dunnett")` とするとダネットの方法での多重比較が実行できる。群分け変数の最初のカテゴリが対照群であるとみなされるので、このデータでは `casein` を給餌されて育った鶏が対照群となる。もっともこれは、このデータに相応しい解析ではない。

多群の平均値を比較したいときの手順をまとめる。

- 群ごとの分布の正規性をチェックする。正規分布と大きくずれていたらノンパラメトリックな分析法を考える（その場合は、分布の位置も中央値で比較することになる）。そうでなければ次へ。
- Bartlett の検定で「群間で分散に差がない」帰無仮説を検定する。帰無仮説が棄却されたら差がないとはいえないのでノンパラメトリックな方法を考えるか Welch の拡張による一元配置分散分析を適用する。棄却されなければ次へ。
- 一元配置分散分析で群分け変数がデータに有意な効果を与えているか検討する。有意でなければ群間で平均値に差がないと判断される。有意なら次へ。
- どの群とどの群に差があるか、検定の多重性を調整しながら検定する。通常はテューキーの HSD でよい。

7.4 課題

<http://phi.med.gunma-u.ac.jp/msb/data/p07.txt> は、バプアニューギニアのある地方の 4 つの村に居住する成人男性を対象に、約 20 年前に行われた血液検査の結果得られた、ヘモグロビン濃度のデータを一部加工した（架空のデータを付け加えたり削除したりした）ものである。VIL という変数が村の番号を示し（その地方にある 13 の村のうち、異なる生態学的条件を代表する典型的な 4 つの村を選んだ）、HB がヘモグロビン濃度 (g/dL) を示す。村によってヘモグロビン濃度に差があるかどうか検討せよ。

第 8 章

相関と回帰

8.1 相関と回帰の違い

相関と回帰は、どちらも 2 つの変数の関係を扱うので混同されやすいが思想は異なる。相関は、変数間の関連の強さを表すものである。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ものである。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。

相関では関連の向きを考えないが、回帰は必ず向きがある。作図においても、2 つの変数の関係をみるには、まず散布図を作成するのは共通だが、回帰をみる場合は、必ず独立変数を横軸 (X 軸) にとる。さらに、相関をみる場合は集中楕円 (棄却楕円とも確率楕円ともいう) を重ね、回帰をみる場合は回帰直線とその信頼区間や予測区間を重ねてプロットする。

8.2 相関

関係とか関連とかいっても、その中身は多様である。例えば、 $pV = nRT$ のような物理法則は、測定誤差を別にすれば 100% 成り立つ関係である。身長と体重の間には、そのような例外のない関係は成り立たない (つまり、関係にばらつきがある)。しかし、無関係ではないことは直感的にも理解できるし、「身長の高い人は体重も概して重い傾向がある」ことは間違いない。一般に、2 個以上の変数が「かなりの程度の規則性をもって、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。

8.2.1 見かけの相関・擬似相関

相関関係があっても、それが見かけ上のものである（それらの変量とともに、別の変量と真の相関関係をもっている）場合がある。具体例としては、血圧と所得の間に正の相関があるという命題は、データをとってみれば、多くの場合に成り立つであろう。しかしこれは、おそらくどちらも年齢や摂取エネルギー量との間に真の相関関係が存在するのであって、それらの影響を制御したら（例えば同年齢で同じような食生活をしている人だけについてみるという層別化あるいは限定をしたら）、血圧と所得の間の正の相関は消えてしまうだろう。この場合、見かけの相関があることは、たまたまそのデータで成り立っているだけであって、科学的仮説としての意味に乏しい。

時系列データや地域相関のデータでは、擬似相関 (spurious correlation) が見られる場合もある。例えば、日本の砂糖輸入量と溺死・溺水者数の年次別データをプロットしてみると、負の相関関係があるように見えるのだが、両者の間には真の関係はない。ある年に日本で植えた木の幹の太さと、同じ年にイギリスで生れた少年の身長を 15 年分、毎年 1 回測ったデータをプロットすると、おそらくは正の相関関係があるように見えるのだが、両者の間には関係がないのは明らかである（どちらも年次と真の相関があるとはいえるだろうが）。複数の種類の異なるデータをまとめて見ることで見かけの相関が生じてしまう場合もあるので、注意が必要である。

8.2.2 直線的な相関・直線に乗らない相関

先に太字で示した定義の通り、相関関係は「増減をとにもする」関係であればいいので、その関係が直線的であろうとなかろうと問題ない。二次曲線、三次曲線、シグモイド状、あるいは階段関数状などの相関関係もありうる。しかし、一般には、直線的な関係があるという限定的な意味で使われる場合が多い。なぜなら、相関を表すための代表的な指標である相関係数^{*1} r が、直線的な関係の強さを示すための指標だからである。より厳密に言えば、 r が「直線的な関係の強さを示す指標」であるためには、その 2 つの変数が二次元正規分布に従っていることを前提とする。

直線に乗らない相関関係を捉えるには、2 つのアプローチがある。1 つは直線的な関係になるように対数変換などの変換をほどこすことで、もう 1 つはノンパラメト

^{*1} 普通、ただ相関係数といえは、ピアソンの積率相関係数 (Pearson's Product Moment Correlation Coefficient) を指し、通常、 r という記号で表す。

リックな相関係数（分布の形によらない、例えば順位の情報だけを使った相関係数）を使うことである。ノンパラメトリックな相関係数にはスピアマン (Spearman) の順位相関係数 ρ や、ケンドール (Kendall) の順位相関係数 τ がある。

ピアソンの積率相関係数とは、 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る（ X と Y の共変動を X の変動と Y の変動の積の平方根で割った値ともいえる）。式で書けば、相関係数の推定値 r は、 X の平均値を \bar{X} 、 Y の平均値を \bar{Y} と書けば、

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

となる。母相関係数がゼロかどうかという両側検定のためには、それがゼロであるという帰無仮説の下で、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検定すればよい*2。

R で変数 X と Y の相関係数を計算して有意確率を得るには次の枠内の 4 行を打てばよい*3。1 行目は X と Y の共分散 $\text{cov}(X, Y)$ を X と Y それぞれの不偏分散 ($\text{var}(X)$ と $\text{var}(Y)$) の積の平方根 ($\text{sqrt}()$ 関数) で割って、相関係数を計算しているが、実は `print(r <- cor(X, Y))` で置き換え可能である。2 行目は `length()` 関数を使ってデータ数を計算して n という変数に付値し、3 行目で検定統計量 t_0 を計算し、4 行目で t 分布の分布関数を使って有意確率を計算している。

```
print(r <- cov(X, Y)/sqrt(var(X)*var(Y)))
n <- length(X)
t0 <- r*sqrt(n-2)/sqrt(1-r^2)
print(2*(1-pt(abs(t0), n-2)))
```

*2 既に説明したとおり、検定は、帰無仮説を立てて、それが正しいときに、現在得られているデータ以上に外れたデータが偶然得られる確率がどれほどかを計算し（有意確率）、その確率が 0.05 とか 0.01 といった有意水準より小さいときに、偶然ではありえないほど小さいと判断し、帰無仮説棄却するという意思決定を行うためのプロセスである。有意確率を計算するためには、通常、帰無仮説が正しいとしたときに既知の確率分布に従うはずの量（検定統計量）を計算し、その既知の確率分布の分布関数のその値に対応する値を 1 から引けば（片側検定のとき）有意確率となる。両側検定の場合はその確率を 2 倍する。この原理は、たいていの検定に共通している。

*3 もっとも、`cor.test(X, Y)` とすれば、信頼区間の計算も含めて全部やってくれる。

相関係数の信頼区間は、サンプルサイズがある程度大きければ（通常は 20 以上）、正規近似を使って計算できる。すなわち、

$$a = \frac{1}{2} \ln \frac{1+r}{1-r} - \frac{1}{\sqrt{n-3}} Z(\alpha/2)$$

$$b = \frac{1}{2} \ln \frac{1+r}{1-r} + \frac{1}{\sqrt{n-3}} Z(\alpha/2)$$

と書くことにすると*4、母相関係数の $100 \times (1 - \alpha)\%$ 信頼区間の下限は $(\exp(2a) - 1)/(\exp(2a) + 1)$ 、上限は $(\exp(2b) - 1)/(\exp(2b) + 1)$ である*5。

順位相関係数は、直線に乗らない相関関係を捉えたい場合以外にも、分布が歪んでいたり、外れ値がある場合に使うと有効である。スピアマンの順位相関係数 ρ は*6、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になる。 X_i の順位を R_i 、 Y_i の順位を Q_i と書けば、

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

となる。スピアマンの順位相関係数がゼロかどうかという両側検定は、サンプルサイズが 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。ただし、

```
methods(cor.test)
# の結果、cor.test には default と formula という S3 メソッドがあるので
getS3method("cor.test", "default")
```

によってソースコードを確認すると、R の `cor.test()` 関数では、サンプルサイズが 1290 以下のときは、明示的に `exact=F` というオプションをつけない限り、正確な確率が計算されることがわかる。

*4 $Z(\alpha/2)$ は標準正規分布の $100 \times (1 - \alpha/2)$ パーセント点である。 α を alpha と書けば `qnorm(1-alpha/2,0,1)` で得られる。例えば有意水準 5%、すなわち $\alpha = 0.05$ なら、`qnorm(0.975,0,1)` とする。

*5 なお、 \ln は自然対数、 \exp は指数関数を表す。この式から明らかなように、母相関係数の信頼区間はどんなに広がっても下限は -1 以下にはならず、上限は 1 以上にならない。

*6 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。

ケンドールの順位相関係数 τ は、

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

によって得られる。ここで A は順位の大小関係が一致する組の数、 B は不一致数である。

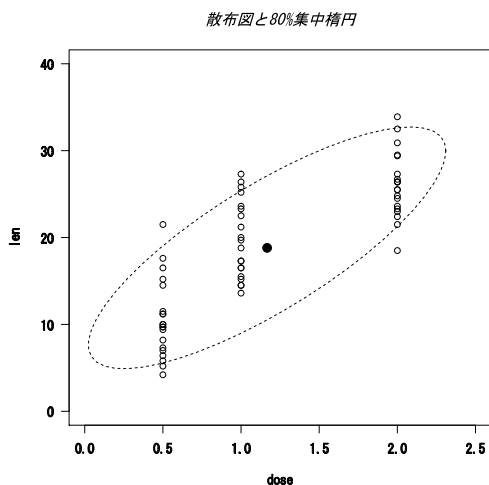
R では `cor.test(X, Y, method="pearson")` とすれば (あるいは `method=オプション` をつけなければ) ピアソンの積率相関係数が、`cor.test(X, Y, method="spearman")` でスピアマンの順位相関係数が、`cor.test(X, Y, method="kendall")` でケンドールの順位相関係数が得られる。同時に、`alternative` を指定しないときは、「相関係数がゼロである」を帰無仮説として両側検定した有意確率と 95% 信頼区間が表示される。なお、例えば `cor.test(X, Y, alternative="greater")` とすれば、ピアソンの積率相関係数が計算され、対立仮説を「正の相関がある」とした片側検定の結果が得られる。なお、ケンドールについても並べ換えによる正確な確率を求めることができ、その場合は `exact=T` というオプションを指定する (ソースコードをみると、`exact` オプションを指定しない場合、サンプルサイズが 50 未満だとデフォルトで正確な確率が計算されるが、同順位のデータがあるときは正確な確率を求めることができない)。

例題

`ToothGrowth` は各群 10 匹ずつのモルモットに、3 段階の用量のビタミン C をアスコルビン酸またはオレンジジュースとして投与したときの象牙芽細胞 (歯) の長さを比較するデータである。変数 `len` が長さ、`supp` が投与方法、`dose` が用量を示す。投与方法の違いを無視して用量と長さの相関関係を調べよ。

まず `attach(ToothGrowth)` して `ToothGrowth` に含まれている変数が見えるようにしてから、用量と長さの関係を概観するために散布図を描く。横軸を `dose`、縦軸を `len` としたプロットをするために、`plot(dose, len)` とする。なんとなく `dose` が増すにつれて `len` が長くなっていくような、正の相関関係があるように見える。集中楕円を重ね描きさせるには、`car` ライブラリの `ellipse()` 関数を用いるのが便利である。この関数は `ellipse(center, shape, radius)` という形で使い、`center` は楕円の中心、即ちデータの重心を示す要素数 2 のベクトル、`shape` は 2×2 の共分散行列、`radius` は楕円を生成する円の半径で、2 変量正規分布の 80% 信頼区間を示す場合は自由度 2 のカイ二乗分布の 80% 点の平方根、即ち `sqrt(qchisq(.8, 2))` を与えればよい。もっと手抜きをすると、`car` ライブラリには `data.ellipse()` という関数があり、散布図と集中楕円を同時に描くことができる。

```
require(car)
data.ellipse(dose, len, levels=0.8, col="black", lty=2, lwd=1,
  xlim=c(0,2.5), ylim=c(0,40), main="散布図と 80% 集中楕円")
```



次に相関係数を算出し、「相関係数がゼロと差がない」という帰無仮説を検定してみるために、`cor.test(dose, len)` と打てば、次の出力が得られる。

```
Pearson's product-moment correlation

data:  dose and len
t = 10.2501, df = 58, p-value = 1.243e-14
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6892521 0.8777169
sample estimates:
      cor
0.8026913
```

この結果をみると、ピアソンの積率相関係数の推定値は 0.80、95% 信頼区間が [0.69, 0.88] (ここでは四捨五入で示しているが、真の区間を含む最小の幅に丸める方がよいという意見もあり、それなら [0.68, 0.88] と記載する) となる。95% 信頼

区間がゼロを含んでいないので、帰無仮説が有意水準 5% で棄却されるのは明らかだが、有意確率をみても、 $p\text{-value}=1.243e-14$ とほとんどゼロであることが確認できる。

練習

method="spearman" とか method="kendall" でも試してみよう。

8.3 回帰

実験によって、あるサンプルの濃度を求めるやり方の 1 つに、検量線の利用がある。検量線とは、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（曲線的な関係になる場合もあるが、通常は何らかの変換をほどこし、線型回帰にして利用する。ここで「線型」は linear の訳語であり、「線形」と訳されることもあるが、本書では「線型」で統一する）。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。

いずれも、量がわかっている「独立変数」（この場合は濃度）を x 、誤差を含んでいる可能性がある測定値である「従属変数」（この場合は吸光度）を y として $y = bx + a$ という形の回帰式の係数 a と b を最小二乗法で推定し、サンプルを測定した値 y から $x = (y - a)/b$ によってサンプルの濃度 x を求める。測定値から濃度を推定するときには、回帰式をそのまま使うのではなく、逆算する形になるので注意が必要である。

回帰直線の適合度の目安としては、相関係数の 2 乗が 0.98 以上あることが望ましい。また、データ点の最小、最大より外で直線関係が成立する保証はない。したがって、サンプル測定値が標準物質の測定値の最小より低いか、最大より高いときは、測定限界を超えていることになってしまうので、測定をやり直す必要がある。通常、サンプルを希釈するか濃縮し^{*7}、検量線の濃度範囲に収まるようにして測定する^{*8}。希

*7 濃縮にもいろいろあって、測定したい物質が気化しないならば水分を飛ばすだけ（何倍に濃縮したかは容量や重量の変化で把握する）で済むかもしれないし、有機物と結合させ、少量の有機溶媒に溶出させてから分液漏斗で有機溶媒の部分を取り出して測るといった面倒な手続きが必要なものもある。

*8 より広い濃度範囲で直線性が得られれば、標準物質の測定点を増やして検量線を作り直すという手

積の溶媒を何にするか、濃縮した場合にその効率はどうか、化学変化を起こさないか、といった検討が必要になってきて、分析技術としてはなかなか厄介であるが、仕方がない。検量線を利用するには回帰の外挿は禁忌である。

測定点 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ が得られたときに、検量線 $y = bx + a$ を推定するには、図に示した線分の二乗和が最小になるように a と b を設定すればよい、というのが最小二乗法の考え方である*9。つまり、

$$\begin{aligned} f(a, b) &= \sum_{i=1}^n \{y_i - (bx_i + a)\}^2 \\ &= b^2 \sum_{i=1}^n x_i^2 - 2b \sum_{i=1}^n x_i y_i + 2ab \sum_{i=1}^n x_i - 2a \sum_{i=1}^n y_i + na^2 + \sum_{i=1}^n y_i^2 \end{aligned}$$

が最小になるような a と b を推定すればよい。通常、 a と b で偏微分した値がそれぞれ 0 となることを利用して計算すると簡単である。つまり、

$$\frac{\partial f(a, b)}{\partial a} = 2na + 2(b \sum_{i=1}^n x_i - \sum_{i=1}^n y_i) = 0$$

$$i.e. \quad na = \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i$$

$$i.e. \quad a = (y \text{ の平均}) - (x \text{ の平均}) * b$$

$$\frac{\partial f(a, b)}{\partial b} = 2b \sum_{i=1}^n x_i^2 + 2(a \sum_{i=1}^n x_i - \sum_{i=1}^n x_i y_i) = 0$$

$$i.e. \quad b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i$$

を連立方程式として a と b について解けばよい。これを解くと、

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

もある。

*9 なお、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検量線とするには、 $y = bx$ について同じ手順で計算すればよいので、 $b = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2$ となる。なお、原点を通る回帰直線を求めるための R のコードは `lm(Y~X-1)` または `lm(Y~0+X)` である。

が得られる*¹⁰。 b の値を上のに代入すれば a も得られる。検量線に限らず、一般の回帰直線でも、計算方法は原則として同じである。一般に、 $y = bx + a$ という回帰直線について、 b を回帰係数 (regression coefficient)、 a を切片 (intercept) と呼ぶ。回帰係数は直線の傾き (slope) を意味する。

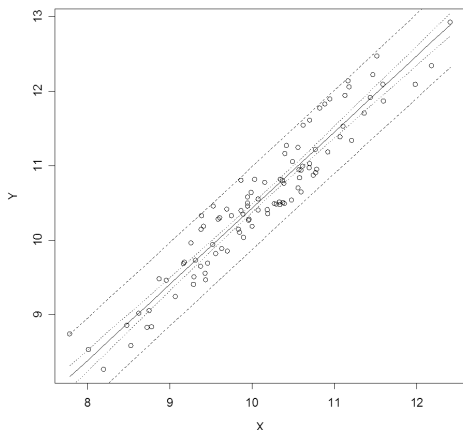
R では、線型回帰を行うための関数は `lm()` である。例えば、`lm(Y~X)` のように用いれば、回帰直線の推定値が得られる。散布図の上に回帰直線を重ね描きするには、`plot(Y~X)` としてから、`abline(lm(Y~X))` とすればよい。また、回帰直線の 95% 予測区間 (期待値の標準誤差が 95% の信頼性をもってこの区間に含まれるであろう範囲) と 95% 信頼区間 (データの 95% がこの範囲に入るであろうという範囲) を点線と破線で重ね描きするには、`predict()` 関数を使って範囲を予測し、重ね描きすればよい*¹¹。例えば、次の例のようになる。最初の 4 行は乱数を使ってデータを作っている部分である。`rnorm()` は正規分布に従う乱数 (正規乱数) を発生させ、`runif()` は一様分布に従う乱数 (一様乱数) を発生させる。`matlines()` は行列またはデータフレームを引数に与えて一度に複数の線を描く関数である。

c08-1.R

```
RNGkind("Mersenne-Twister")
set.seed(1)
X <- rnorm(100,10,1)
Y <- X + runif(100,0,1)
summary(res <- lm(Y~X))
XX <- data.frame(X=seq(min(X),max(X),length=20))
plim <- predict(res, XX, interval="prediction")
clim <- predict(res, XX, interval="confidence")
plot(X,Y)
matlines(XX,plim,col=1,lty=c(1,2,2))
matlines(XX,clim,col=1,lty=c(1,3,3))
```

*¹⁰ 分母分子を n^2 で割れば、 b は $x_i y_i$ の平均から x_i の平均と y_i の平均の積を引いて、 x_i の二乗の平均から x_i の平均の二乗を引いた値で割った形になる。

*¹¹ ただし `predict()` 関数は回帰式の計算値そのものも返すので、これで重ね描きする場合は `abline()` は不要である。



5行目で回帰分析を実行した結果を `res` というオブジェクトに付値すると同時に、決定係数や回帰係数と切片の検定結果を出力している。出力結果は次の枠内の通りである。

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min       1Q   Median       3Q      Max
-0.39958 -0.24095 -0.04863  0.20490  0.57297

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17392    0.31703   0.549   0.585
X            1.02583    0.03124  32.837 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2792 on 98 degrees of freedom
Multiple R-Squared:  0.9167, Adjusted R-squared:  0.9158
F-statistic: 1078 on 1 and 98 DF,  p-value: < 2.2e-16
```

Residuals:の部分は残差を示す。残差とは、回帰による予測値と実測値の差であ

る。独立変数（ここでは X ）の最小値 (Min)，第 1 四分位 (1Q)，中央値 (Median)，第 3 四分位 (3Q)，最大値 (Max) に対応する従属変数の値から，回帰式にそれらの独立変数の値を代入して得られる値（これが回帰による予測値）を引いた値を意味する。0 に近いほど回帰式のデータへの当てはまりは良いと考えられる。

次の Coefficients: のところに表示されるのが，さまざまな係数とその検定結果である。(Intercept) の行は切片を示す。 X の行が変数 X についての情報を与える。Estimate の値が切片と回帰係数の点推定量であり，Std. Error の列はそれぞれの標準誤差を示す。t value は，「切片がゼロと差がない」及び「回帰係数がゼロと差がない」を帰無仮説とする検定を行うための， t 分布に従う検定統計量である。Pr(>|t|) は有意確率を示す。下の方に，Adjusted R-squared とあるのが自由度調整済み相関係数の二乗で，後述するように決定係数とも呼ばれ，従属変数 Y のばらつきのどれくらいの割合が独立変数 X のばらつきによって説明されるかを示す値である。このデータでは 92% 近い値であり，説明力の強い回帰式が得られたといえる。

検量線の例題

血清鉄濃度を Fe-Test Wako というキットで測定するため，鉄の標準希釈系列を 0, 0.5, 1, 2 (mg/L) として作成し，それをこのキットで処理して発色させた溶液の波長 562 nm の吸光度を測った結果が，0.012, 0.058, 0.104, 0.193 として得られた。これから検量線を求めて，測定に使えるかどうか評価せよ。次に，6 人の血清サンプルを同じ方法で処理して発色させた溶液の吸光度が 0.107, 0.075, 0.077, 0.099, 0.096, 0.108 だったときに，この 6 人の血清鉄濃度を求めよ。

まず鉄濃度を conc，吸光度を abs としてデータを入力し，conc を横軸，abs を縦軸にして散布図を描く^{*12}。

c08-2.R(1)

```
conc <- c(0, 0.5, 1, 2)
abs <- c(0.012, 0.058, 0.104, 0.193)
plot(abs~conc)
```

だいたい直線に乗っているように見えるので，回帰分析をしてみる。以下のように，回帰分析の結果をいったん res に保存しておく，描画や表示やその後の計算に便利である。

^{*12} 描画命令は，ここでは plot(abs~conc) としたが，plot(conc, abs) でも同じことである。

```
c08-2.R(2)
```

```
res <- lm(abs~conc)
abline(res)
summary(res)
```

出力される結果から, Adjusted R-squared: 0.9999 なので検量線として使ってよいと判断できる。回帰係数は 0.0904571, 切片は 0.0126000 として得られているが, これらの値はそれぞれ `res$coef[2]`, `res$coef[1]` として参照できるので*13, サンプルの吸光度から濃度を逆算するときには, 次の枠内のように変数名のまま参照した方が間違えない。

```
c08-2.R(3)
```

```
dat <- c(0.107, 0.075, 0.077, 0.099, 0.096, 0.108)
print((dat-res$coef[1])/res$coef[2])
```

8.3.1 決定係数

現実問題として, 回帰直線に完璧にデータが乗ることはない。そこで, 回帰直線の当てはまりのよさを評価する必要が出てくる。 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,

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

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

を計算すると, この Q は回帰直線の当てはまりの悪さを示す尺度となる。 Q を「残差平方和」と呼び, それを n で割った Q/n を残差分散 $\text{var}(e)$ という。残差分散 $\text{var}(e)$ と Y の分散 $\text{var}(Y)$ と相関係数 r の間には, $\text{var}(e) = \text{var}(Y)(1 - r^2)$ という関係が常に成り立つので, $r^2 = 1 - \text{var}(e)/\text{var}(Y)$ 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。

*13 `res$coef` の部分は `coef(res)` でも同じ意味である。

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 Y への「寄与率」と呼ぶこともある。

なお、モデルのデータへの当てはまりを評価する指標は、残差分散と決定係数の他にも、AIC や BIC や Deviance などいろいろある。これらについては、一般化線型モデルのところ（第 12 章）で、その一部を説明する。

8.3.2 回帰直線推定と検定のしくみ

回帰直線は、最小二乗法によって、もっとも残差平方和が小さくなるように選ばわけだが、データの配置によっては、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、独立変数と従属変数（として選んだ変数）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。

その意味で、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ことが大事である。そのために使われるのが、 t 値という統計量である。いま、 Y と X の関係が $Y = a_0 + b_0X + e$ という線型モデルで表されるとして、誤差項 e が平均値 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 の平均値と分散）の正規分布に従い、残差平方和 Q を誤差分散 σ^2 で割った Q/σ^2 が自由度 $(n - 2)$ のカイ二乗分布に従うことから、

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

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

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

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。既に示したとおり、これらの検定結果は `summary(lm())` で表示される。

8.3.3 独立変数・従属変数と因果の向き

実は、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、どちらかを独立変数、どちらかを従属変数、とみなすことに問題がある。一般には、身長によって体重が決まってくるというように方向性（因果の向き）が仮定できれば、身長を独立変数と見なしてもよいことになっているが、回帰分析をしてしまうと、独立変数に測定誤差がある可能性が排除されてしまうことには注意しておくべきである。つまり、測定誤差が大きい可能性がある変数を独立変数とした回帰分析は、できれば避けたほうが良い^{*14}。また、最小二乗法の説明から自明なように、独立変数と従属変数を入れ替えた回帰直線は一致しないので、どちらを従属変数とみなし、どちらを独立変数とみなすか、因果の向きに基づいてきちんと決めねばならない。

8.3.4 回帰式を予測に用いる際の留意点

回帰式を使って予測をするとき、外挿には注意が必要である。前述の通り、検量線は、原則として外挿は禁忌である。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）。しかし、外挿による予測は、実際にはかなり行われている。例えば世界人口の将来予測とか、河川工学における基本高水計算式とか、感染症の発症数の将来予測は、回帰の外挿による場合が多い（これらの場合は逆算ではなく、実際にデータを得られていない横軸の値を代入したときに縦軸の値がいくつになるかを予測している）。このやり方が妥当性をもつためには、その回帰関係が (1) かなり説明力が大きく、(2) 因果関係がある程度認められ、(3) それぞれの変数の分布が端の切れた分布でない (truncated distribution でない) という条件を満たす必要がある。そうでない場合は、その予測結果が正しい保証はどこにもない。

^{*14} そうもいかないのが実際には行なわれている。

例題

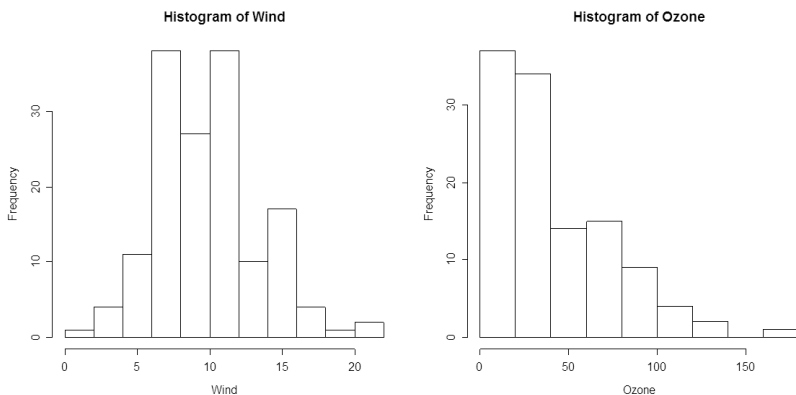
airquality は、1973 年 5 月 1 日から 9 月 30 日まで 154 日間のニューヨーク市の大気環境データである。含まれている変数は、Ozone (ppb 単位でのオゾン濃度), Solar.R (セントラルパークでの 8:00 から 12:00 までの 4000 から 7700 オングストロームの周波数帯の太陽放射の強さを Langley 単位で表した値), Wind (LaGuardia 空港での 7:00 から 10:00 までの平均風速, マイル/時), Temp (華氏での日最高気温), Month (月), Day (日) である。

このデータから、オゾン濃度と風速の関係について検討し、もし風速からオゾン濃度を予想できるとしたら、風速 15 マイル/時の日のオゾン濃度はどうなるか、また、もしも 25 マイル/時の日があったとしたらどうなるか、期待値とその 95% 信頼区間 (標準誤差から計算される、母集団における期待値の信頼区間) を計算せよ。

データフレーム airquality を attach し、まず風速とオゾン濃度の分布の正規性について Shapiro-Wilk の検定をし、ヒストグラムによってそれぞれの分布の様子をみる。

```
c08-3.R(1)
```

```
attach(airquality)
shapiro.test(Wind)
shapiro.test(Ozone)
layout(t(1:2))
hist(Wind)
hist(Ozone)
```



風速は正規分布に従っているがオゾン濃度は正規分布に従っていない

ので、

```
c08-3.R(2)
cOzone <- log(Ozone+10)
```

と変数変換する^{*15}。そこで、次に大雑把に風速とオゾン濃度の関係を見るために散布図をプロットしてみる。

散布図をみると、確かに風速が大きくなるほどオゾン濃度が下がる関係がありそうに見える。そこで、独立変数を Wind、従属変数を cOzone とする回帰分析をしてみる。95% 予測区間（データの 95% がそこに入るであろう範囲）付きで、回帰直線をさっきの散布図に重ね描きしてみると、風速が中程度のところで回帰直線よりオゾン濃度が低い方に外れた値がいくつかあり、回帰関係全体がそれによって歪んでいる可能性が考えられる。したがって、この図から厳しく判断するなら、この回帰関係は予測に使うべきではない。残差プロット（右側）をみても、それほど残差は大きくないものの、やはり中央付近が凹んでいるように見える^{*16}。

^{*15} このデータの場合、ただの対数変換では正規分布に従うという帰無仮説はまだ棄却されるので、別の変換が必要となる。例えば、ここで採用したように 10 を加えた後で対数変換するか、または立方根変換すれば、正規分布に従うという帰無仮説が棄却されなくなる。メカニズムを考えると、風速が無限大になればオゾン濃度は限りなくゼロに近づきそうだが、風速ゼロのときでもオゾン濃度が無限大になることはなさそうなので、 $\log(Ozone+10)$ とした（これが最適とは限らない）。なお、変換するかどうかは、正規性の検定で有意なら必ずするわけではなく、変換によるひずみと正規分布に従わないことによるひずみを勘案して決定する。

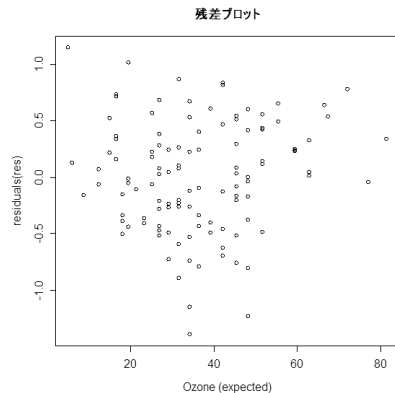
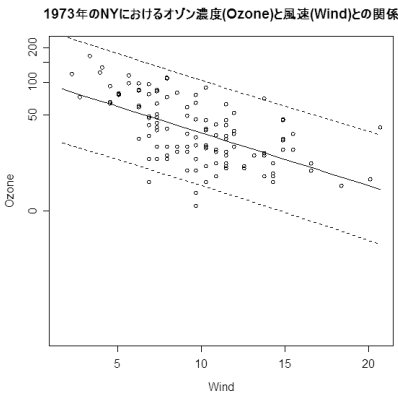
^{*16} 残差プロットの横軸は `fitted.values(res)` だと回帰式による従属変数の推定値そのものになるが、ここでは変数変換しているのだから、それを元に戻す（逆変換）ために $\exp(\text{fitted.values}(\text{res}))-10$ とした。横軸は `res$model$X` として独立変数の値をとる場合もある。また、残差プロットの縦軸 `residuals(res)` は、個々のデータと従属変数の推定値の差になり、これも変数変換後の値だが、対数をとって差を出すと、比の対数になるので、指数をとって元に戻しても残差にならない（残比とでもいうべきか）ため、逆変換していない。

```

c08-3.R(3)

layout(t(1:2))
cOzone <- log(Ozone+10)
shapiro.test(cOzone)
plot(cOzone~Wind,yaxt="n",ylab="Ozone",ylim=c(0,log(210)),
     main="1973年のNYにおけるオゾン濃度(Ozone)と風速(Wind)との関係")
yi <- 0:4*50
axis(2,log(yi+10),yi)
res <- lm(cOzone~Wind)
X <- data.frame(Wind=seq(min(Wind),max(Wind),length=20))
Y <- predict(res,X,interval="predict")
matlines(X,Y,col=1,lty=c(1,2,2))
plot(exp(fitted.values(res))-10,residuals(res),main="残差プロット",
     xlab="Ozone (expected)")
exp(predict(res,list(Wind=15),interval="confidence"))-10
exp(predict(res,list(Wind=25),interval="confidence"))-10
summary(res)
detach(airquality)

```



そうはいつでも、この程度のズレなら、データのある範囲内なら回帰式を使えないこともない。回帰式は

$$\log(\text{Ozone} + 10) = 4.74 - 0.0985 \cdot \text{Wind}$$

であり、「回帰係数がゼロと差がない」帰無仮説の検定の有意確率は 6.8×10^{-12} なので帰無仮説は棄却されるし、決定係数は 0.33 なので、オゾン濃度のばらつきの

33% は風速のばらつきによって説明されると考えられる。この程度では予測には十分ではないが、それでも強引に Wind が 15 (マイル/時) のときのオゾン濃度の期待値と期待値の 95% 信頼区間を求めたところ、16.2[12.3, 20.7] (ppb) であった。

しかし Wind が 25 のときの期待値と 95% 信頼区間は $-0.2[-3.4, 4.5]$ (ppb) となり、期待値さえ負の値というありえない結果になってしまう。これは (1) 回帰式のデータへの当てはまりが不十分かつ (2) データのない範囲への外挿なので式が成り立つ保障がそもそもないという 2 点を考えれば当然の結果である。したがって、風速 25 マイル/時の日の予測は不可能といえる。

8.4 課題

<http://phi.med.gunma-u.ac.jp/msb/data/p08.txt> は、ソロモン諸島のある村に居住する成人女性 17 人の身体計測データで、含まれている変数は身長 (HT, 単位は cm), 体重 (WT, 単位は kg), Body Mass Index (BMI, 単位は kg/m^2), タニタの体脂肪計つき体重計で測定した体脂肪割合 (FAT, 単位は %), 収縮期血圧 (SBP, 単位は mmHg) である。なお、言うまでもないが、BMI は身長と体重からの計算値である。

このデータから、以下のどちらかに答えよ (両方でもよい)。

- (1) BMI と体脂肪割合の相関について検討せよ。
- (2) 身長を独立変数、体重を従属変数とした回帰分析を行って、もし次に測定した人の身長が 155 cm だったら、その人の体重は何 kg と予想されるか、95% 信頼区間をつけて推定せよ。

第 9 章

計数データと比率の解析

9.1 母比率を推定する方法

既に、量的な変数が正規分布に従うとして、1つの変数の標本平均と既知の母平均との差の検定、2つの変数の平均値の差の検定、多群の平均値の差の検定、と説明を進めてきた。前章では2つの量的な変数の関係を分析した。本章と次章では、同じような分析を、カテゴリ変数に対して行う方法を説明する。

名義尺度や順序尺度をもつカテゴリ変数1つがもっている情報は、データ数と、個々のカテゴリが占める割合（標本比率）である。したがって、このデータから求める統計的な指標は、母比率、即ち個々のカテゴリが母集団で占めるであろう割合である。ランダムサンプルであれば、標本比率と一致することが期待される。

例えば、手元の容器の中に、数百個の白い碁石があるとする。この概数を手っ取り早く当てるために、数十個の黒い碁石を混ぜる。よくかき混ぜてから20個程度の石を取り出して（標本）、その中で黒い石が占めていた割合（標本比率）を求め、それが母比率と等しいと仮定して加えた黒い碁石の数を割って総数を求め、黒い碁石の数を引けば、元々の白い碁石の数が得られる。生態学で、野原のバッタの数を調べたいときに全数を調べるわけにはいかないので、捕まえて塗料でマークして放して暫く経ってからまた捕まえてマークされているバッタの割合を求めて、マークした数をそれで割って総数を推定する、というリンカーン法（Capture-Mark-Recapture; 略してCMRともいう）のやり方と同じである。

例題

最初に混入した黒い石の数が40個、かき混ぜてから20個の石を取り出してみたら黒石2個、白石18個だった場合、元の白石の数はいくつと推定されるか？

元の白石の数を x とすると、母比率と標本比率が一致するなら、

$$40/(40+x) = 2/(2+18)$$

となるので、これを x について解けば、 $x = 360$ が得られる。したがって 360 個と推定される。この程度は R を使うまでもないが、 $40/(2/(2+18))-40$ と電卓のように打てば、360 が得られる。

9.2 推定値の確からしさ

ここで、このようにして求めた推定値がどれほど確からしいか？ を考えよう。例えば、黒石の割合（母比率）が p である容器から 20 個の石を取り出したときに、黒石がちょうど 2 個である確率を考えると、これは 2 項分布に従う*1。

つまり、復元抽出で考えれば、確率 p の現象が 20 回中 2 回起こり、残りの 18 回は確率 $(1-p)$ の現象が起こったわけだから、その確率をすべて掛け合わせ、20 回中どの 2 回で起こるのかという組み合わせの数だけパターンがありうるので ${}_{20}C_2$ 回だけそれを足し合わせた確率になる。

R では、この確率は、母比率 p を与えると、`choose(20,2)*p^2*(1-p)^18` あるいは `dbinom(2,20,p)` で得られる。

逆に考えれば、この「母比率 p の現象が 20 回中ちょうど 2 回得られる」確率を最大にするような p が真の母比率として、一番尤もらしいと考えられる。0.01 刻みでこの確率を最大にする p を探索するには次の枠内のようにする。1 行目は x という変数に、`seq()` 関数を使って 0 から 1 まで 0.01 刻みで連続する数を付値する。2 行目で、 x のそれぞれの値を母比率とする現象が 20 回中ちょうど 2 回得られる確率を 2 項分布の確率密度関数 `dbinom()` で得て、 y という変数に保存している。3 行目は `plot()` 関数の `type="l"` を使って、 x を横軸、 y を縦軸にとって、これらの値を線でつないだグラフを表示する。4 行目のコメントにある通り、グラフ表示だけでなく `curve()` を使う方が簡単である。5 行目は、 y のうち最大値となるのが何番目の値を `which.max()` で得ておき、その順番の x の値を表示させれば、求める母比率が得られるというわけである。実行すると 0.1 が得られる。

*1 個々の抽出を考えると復元抽出でないとも 2 項分布に従わないが、すべての場合の確率の合計を考えれば非復元抽出でもそうなる。

```
c09-1.R
```

```
x<-seq(0,1,by=0.01)
y<-dbinom(2,20,x)
plot(x,y,type="l")
# 曲線を描くだけなら, 上 3 行の代わりに curve(dbinom(2,20,x),0,1) で OK
x[which.max(y)]
```

40 個入れて全体の 0.1 を占めるのだから, $40/0.1=400$ が全体の数で, $400-40=360$ が元の白石の数だと推定できる。ただし, 図を見ればわかるように, $p = 0.09$ だろうが $p = 0.11$ だろうが, 黒石がちょうど 2 個である確率には大した差はない。だから, 360 個という点推定値は, 404 個 ($p = 0.09$ の場合) とか 324 個 ($p = 0.11$ の場合) に比べて, それほど信頼性は高くない。

9.3 母比率の信頼区間

ある程度の信頼性が見込める範囲を示すためには, 平均値の場合と同様, 信頼区間を用いることができる。母比率が $p = 0.1$ のときに, 20 個のサンプル中の黒石がちょうど 2 個である確率は, $\text{dbinom}(2,20,0.1)$ より, 約 28.5% に過ぎない (もちろんこれは, 母比率が $p = 0.7$ のときに 20 個のサンプル中の黒石がちょうど 2 個になる確率である約 3.6×10^{-8} よりもずっと大きい)。ここはやはり, 95% くらい確からしさをもって, 母比率はここからここまでの範囲に入るという形で説明したいと考え, 95% 信頼区間を計算するのが自然だろう。

平均値の場合は正規分布や t 分布を使ったが, 比率の場合は 2 項分布を用いればよい。つまり, サンプルサイズ N のうち, ある事象が観察された個体数が X だったとすると, 母比率 p の点推定量は $p = X/N$ で与えられるので, 平均値の場合から類推して, 95% 信頼区間の下限は $\text{qbinom}(0.025, N, p)/N$ で, 上限は $\text{qbinom}(0.975, N, p)/N$ と考えるのがもっともシンプルである。しかし, $p \neq 0.5$ のときの 2 項分布は左右対称ではなく, 分位点関数が整数値しかとれないので*3, あらゆる可能性のうち少なくとも 95% を含む最短の区間を 95% 信頼区間として求めるには, 別の考え方をしなくてはならないだろう。

R では, Clopper CJ, Pearson ES: The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26: 404-413, 1934. に記載されて

*2 $\text{dbinom}(2,20,0.7)$ より得られる。

*3 もっとも, N がある程度大きくて, それほど稀でない事象ならば, 後述するように 2 項分布は正規分布に近づいていくから悪い近似ではない。

いるアルゴリズムでこの信頼区間を計算する関数実装済みである。このアルゴリズム^{*4}を使っても最短であることは保証されないが、少なくとも95%を含むことは保証するとされている。`binom.test(X,N,p)` とすれば、「 N 個体中 X 個体に観察される事象の母比率が p と差がない」という帰無仮説の検定結果が表示される (p が X/N に等しいとき p-value が 1 となる) とともに、Clopper and Pearson の方法による95% 信頼区間が計算される。

9.3.1 正規近似

観察数 n , 母比率 p の 2 項分布 $B(n,p)$ は, n が大きいときは平均値 np , 分散 $np(1-p)$ の正規分布 $N(np, np(1-p))$ で近似できる。次の枠内を打てば, $n = 100, p = 0.2$ の場合について, 両者がほとんど重なっていることがグラフで確認できる。1 行目が `barplot()` を使って 2 項分布の確率密度関数 `dbinom()` の棒グラフを描き, 2 行目が `lines()` を使って正規分布の確率密度関数 `dnorm()` を曲線 (厳密にみれば折れ線だが) で重ね描きしている。

```
ii <- barplot(dbinom(0:40,100,0.2))
lines(ii,dnorm(0:40,20,4),col="red")
```

正規分布は左右対称なので, 95% のサンプルは, 平均 \pm 標準偏差 $\times 1.96$ (正確には $qnorm(0.975)$ であり 1.9599... となる) に含まれると考えてよく, 下式が成立する。

$$\Pr[-1.96 \leq (X - Np) / \sqrt{Np(1-p)} \leq 1.96] = 0.95$$

これから $p^* = X/N$ を使って式変形すると,

$$\Pr[p^* - 1.96\sqrt{p^*(1-p^*)/N} \leq p \leq p^* + 1.96\sqrt{p^*(1-p^*)/N}] = 0.95$$

となるので, 母比率 p は 95% の確率で下限 $p^* - 1.96\sqrt{p^*(1-p^*)/N}$, 上限 $p^* + 1.96\sqrt{p^*(1-p^*)/N}$ の範囲にあるといえる。即ちこれが, 母比率 p の 95% 信頼区間となる。

例題

25 匹のマウスに毒物 A を一定量経口投与したところ, 十分な観察期間内に 5 匹が死亡した。この毒物のその用量によるマウスの致命率の点推定量と 95% 信頼区間を求めよ。

^{*4} R で, 括弧をつけずに `binom.test` とすると, どういう計算をしているのが確認できるので試してみたい。

点推定量は、 $5/25$ より 20% であることは自明である。シンプルな考え方で 95% 信頼区間を求めると、下限が $qbinom(0.025, 25, 5/25)/25$ から 0.04、上限が $qbinom(0.975, 25, 5/25)/25$ から 0.36 となる。 $binom.test(5, 25, 0.2)$ によれば $[0.068, 0.407]$ となって、シンプルな考え方よりも上側にずれる。正規近似によれば、 $0.2 - qnorm(0.975) * \sqrt{0.2 * 0.8 / 25}$ が約 0.043、 $0.2 + qnorm(0.975) * \sqrt{0.2 * 0.8 / 25}$ が約 0.357 となるので、 $[0.043, 0.357]$ が 95% 信頼区間となり、ちょっと幅が狭くなってしまう。基本的には $binom.test()$ の結果を使っておけば問題ない。

9.4 カテゴリ 2 つの場合の母比率の検定

あらかじめ母比率について何らかの期待があるときには (50% であるとか)、標本から推定された比率がそれと違ってないかどうかを調べたい、ということが起こる。カテゴリが 2 つしかない場合は、上で説明した 2 項分布による推定の裏返しでよい。つまり、「サンプル N 個体中 X 個体に観察された事象の母比率が p と差がない」という帰無仮説を検定するには、 $binom.test(X, N, p)$ とすればよい。丁寧に考える方法を下の例題で示すが、実際には $binom.test()$ を使えば十分である。

例題

ある病院で生まれた子ども 900 人中、男児は 480 人であった。このデータから、(1) 男女の生まれる比率は半々であるという仮説、(2) 出生性比が 1.06 である (= 男児 1.06 に対して女児 1 という割合で生まれる) という仮説、は支持されるか? (出典: 鈴木義一郎「情報量基準による統計解析入門」、講談社サイエンティフィク、1995 年)

- (1) 母比率が 0.5 であるとして、得られているデータよりも外れたデータが偶然得られる確率 (両側に外れることを考えなくてはいけないので、480 人以上になる確率と 420 人以下になる確率の合計) がきわめて小さければ、「男女の生まれる比率は半々である」という仮説はありそうもないと考えてよいことになる。母比率 0.5 で起こる現象が、900 人中ちょうど 480 人に起こる確率は $dbinom(480, 900, 0.5)$ で与えられ、480 人以上になる確率は、 $dbinom(480, 900, 0.5) + dbinom(481, 900, 0.5) + \dots + dbinom(900, 900, 0.5)$ となるが、これは分布関数を使えば、 $1 - pbinom(479, 900, 0.5)$ で計算できる。420 人以下になる確率も同様に分布関数を使って書けば、 $pbinom(420, 900, 0.5)$ である。したがって、求める確率はこれらの和、即ち、

$$(1 - \text{pbinom}(479, 900, 0.5)) + \text{pbinom}(420, 900, 0.5)$$

である。計算してみると 0.04916... となるので、有意水準 5% で仮説は棄却されることがわかる*5。

(2) 同じように考えれば、

$$1 - \text{pbinom}(479, 900, 1.06/(1.06+1)) + \text{pbinom}(446, 900, 1.06/(1.06+1))$$

でよいはずであり(446 は帰無仮説の下での母比率の値である $900 * 1.06 / (1 + 1.06)$ が約 463 なので、480 と反対側に同じだけ外れた人数を考えた値である)、約 0.271 となる*6ので、有意水準 5% で帰無仮説は棄却されない。つまり仮説はとりあえず支持されるといえる。

9.5 カテゴリが 3 つ以上ある場合の母比率の検定

注目しているカテゴリ変数のカテゴリは 2 つとは限らず、3 つ以上あるかもしれない。そのうち 1 つの事象に着目して、それが起こるか起こらないかだけを分析することもあるが、それぞれのカテゴリの出現頻度のデータをすべて分析することを考えてみる。こういう場合の基本的な考え方としては、標本データのカテゴリごとの度数分布が、母集団について期待される分布と差がないという帰無仮説の下で観察データよりも外れたデータが偶然得られる確率を調べて、それが統計的に意味があると考えられるほど小さい場合に帰無仮説を棄却することになる。

具体的には、カテゴリ数が全部で n 個あって、 i 番目のカテゴリの観測度数が O_i 、期待度数が E_i であるとき、 $\chi^2 = \sum (O_i - E_i)^2 / E_i$ が*7、自由度 $n - 1$ のカイ二乗分布に従うことを利用して検定する(ただし、期待度数を計算するために不明な母数をデータから推定したときは、その数も自由度から引く。 E_i が 1 未満のときはカテゴリ分けをやり直す。また、度数は整数値だけれどもカイ二乗分布は連続分布なので、 χ^2 を計算する際に連続性の補正と呼ばれる操作をすることがある)。このような χ^2 が大きな値になることは、観測された度数分布が期待される分布と一致している可能性が極めて低いことを意味する。一般に、 χ^2 が自由度 $n - 1$ のカイ二乗分布の

*5 `binom.test(480, 900, 0.5)` の結果得られる p-value と一致する。

*6 `binom.test(480, 900, 1.06/(1.06+1))` の結果と一致する。

*7 χ は「カイ」と発音する。英語では chi-square と書かれるので、英文を読むときに間違っ「チ」と読んでしまうと大変恥ずかしい。

95% 点よりも大きいときは、統計学的に有意であるとみなして、帰無仮説を棄却する。この検定方法をカイ二乗適合度検定と呼ぶ。

R で自由度 1 のカイ二乗分布の確率密度関数を図示するには、

```
curve(dchisq(x,1),0,5)
```

とすればよい。 χ^2 値が 1 より大きくなる確率は `1-pchisq(1,1)` より得られ、約 0.317 である。参考までに、自由度 n のカイ二乗分布の確率密度関数 (R では `dchisq(x,n)` で得られる) は、 $x > 0$ について、

$$f_n(x) = 1/(2^{(n/2)}\Gamma(n/2))x^{(n/2-1)} \exp(-x/2)$$

であり、平均値 n 、分散 $2n$ である。なお、自由度 (degrees of freedom; d.f.) とは、既に説明した通り、カテゴリ数から、前もって推定する母数の数を引いた値である。この例なら推定する必要がある母数は $\sum E_i$ だけなので自由度は $n - 1$ となる ($\sum E_i$ が決まれば、そこから E_1 から E_{n-1} を引けば E_n が決まることになり、自由に決められるカテゴリ数は $n - 1$ 個といえる。なお、 $\sum E_i$ は、 $\sum O_i$ に等しいものとして推定する)。

このやり方は、カテゴリが2つときのデータについても適用できる。上の例題に適用してみると、(1) の場合、 χ^2 は、 $X \leftarrow (480-450)^2/450 + (420-450)^2/450$ として計算される。この値が自由度 1 のカイ二乗分布に従うので、R で `1-pchisq(X,1)` とすれば、男女の生まれる比率が半々である場合に 900 人中男児 480 人よりも半々から外れた観察値が得られる確率、つまり有意確率が計算できる。実行してみると、0.0455... となる。したがって、有意水準 5% で「男女の生まれる母比率は半々である」という帰無仮説は棄却される。(2) の場合は、

```
c09-2.R(1)
```

```
EM <- 900*1.06/2.06
EF <- 900*1/2.06
X <- (480-EM)^2/EM+(420-EF)^2/EF
1- pchisq(X,1)
```

より、有意確率は約 0.26 となるので、帰無仮説の下で偶然、男児が 900 人中 480 人以上になる確率は約 26% あると解釈され、この帰無仮説は棄却されない。

ちなみに、出生 900 中男児が 480 人観察されたとき、母集団における出生性比の 95% 信頼区間を考えてみると、R Console に次の枠内を入力すれば、`[1.0005,1.3059]` となることがわかる。`binom.test()` の結果が付値された `res` という変数の構造の

中には `conf.int` という要素で信頼区間が含まれている。最終行は、この値が全体の中での男児割合の信頼区間なので、女児に対する男児の比の信頼区間にする換算をしている。

c09-2.R(2)

```
res <- binom.test(480,900,480/900)
res$conf.int/(1-res$conf.int)
```

9.5.1 少し複雑な例

例題

1日の交通事故件数を155日間について調べたところ、0件の日が79日、1件の日が61日、2件の日が13日、3件の日が1日、4件の日が1日だったとする。このとき、1日あたりの交通事故件数はポアソン分布に従うと言えるか？（豊川裕之、柳井晴夫（編著）「医学・保健学の例題による統計学」、現代数学社、1982、より改変）^a

^a 一般に、稀な事象についてベルヌーイ試行を行うときの事象生起件数がポアソン分布に従うことが知られている。交通事故は稀な事象であり、ある日に交通事故が起こる件数と翌日に交通事故が起こる件数は独立と考えられるので、交通事故件数はポアソン分布に従うための条件を満たしている。

Rでは、ポアソン分布の確率関数（離散分布の場合は、確率密度関数と言わずに確率関数というのが普通）は、`dpois(件数, 期待値)` で与えられる。この例題ではポアソン分布の期待値（これは母数である）がわからないので、データから推定すれば、

$$\frac{(0 \times 79 + 1 \times 61 + 2 \times 13 + 3 \times 1 + 4 \times 1)}{155}$$

で得られる。この値を `Ehh` として計算し、観測度数の分布をプロットするスクリプトを下枠内に示す。

c09-3.R(1)

```
cc <- 0:4
hh <- c(79,61,13,1,1)
names(hh) <- cc
print(Ehh <- sum(cc*hh)/sum(hh))
barplot(hh)
```


従って、1日の交通事故件数が期待値 E_{hh} のポアソン分布に従うとしたときの、交通事故件数 0~4 の期待日数 e_{pp} は、 $e_{pp} \leftarrow dpois(cc, E_{hh}) * \sum(hh)$ で得られる。

こうなれば、 $X \leftarrow \sum((hh - e_{pp})^2 / e_{pp})$ としてカイ二乗値を求め、これが自由度 3 (件数の種類が 5 種類あって、ポアソン分布の期待値が母数として推定されたので、 $5 - 1 - 1 = 3$ となる) のカイ二乗分布に従うとして $1 - pchisq(X, 3)$ が 0.05 より小さいかどうかで適合を判定すれば良さそうなものだが、そうはいかない。

$e_{pp}[5]$ (この場合、 $e_{pp}[cc==4]$ と同じものを指すことになるので、以後、この記法を用いる) が 1 より小さいので、カテゴリを併合しなくてはならないのである*8。そこで、 $e_{pp}[5]$ を $e_{pp}[4]$ と併合する。

即ち、

c09-3.R(2)

```
ep <- epp[cc<4]
ep[4] <- ep[4]+epp[5]
```

として期待度数の分布 ep を得、

c09-3.R(3)

```
h <- hh[cc<4]
h[4] <- h[4]+hh[5]
```

として観測度数の分布 h を得る。

後は、 $XX \leftarrow \sum((h - ep)^2 / ep)$ としてカイ二乗値を求め、 $1 - pchisq(XX, 2)$ を計算すると (カテゴリが 1 つ減ったので自由度も 1 減って 2 となる)、約 0.187 となることがわかる。即ち、1日の交通事故件数がポアソン分布に従っているという仮定の下でこのデータよりも偏ったデータが得られる確率は約 19% あり、「一日の事故件数がポアソン分布に従っている」という帰無仮説は棄却されない。

R にもカイ二乗適合度検定をやってくれる関数を用意されていて、もし自由度の調整がなければ、

```
chisq.test(as.table(h), p=ep/sum(ep), correct=F)
```

とすればカイ二乗値とその有意確率が計算できるのだが、カイ二乗分布は自由度 2 の場合と自由度 3 の場合では大きく違うので、この場合のように自由度を減らさなく

*8 もっとも、併合した分布は元の分布と等価ではないので、併合の際にも本当は慎重な検討が必要である。

てはいけないうきには使えない。なお、2つの分布が一致しているという帰無仮説を検定する方法としては、コルモゴロフ=スミルノフ検定(KS検定)という方法もあり、これなら、`ks.test(h,ep)`で検定できる。なお、このデータについては、ここで示したどのやり方で分析しても、帰無仮説が有意水準5%で棄却されない(つまり、「1日の事故件数はポアソン分布に従っている」と考えられる)という結論は変わらない。

9.6 サイコロの正しさの検定

この特別な場合として、どのカテゴリも出現頻度が等しいという帰無仮説を検定することが考えられる。例えば、サイコロを900回振って出た目の回数が下表のようであったとき、このサイコロの各目の出やすさに差はないと考えていいかという問題である。

目	1	2	3	4	5	6
回数	137	163	137	138	168	157

上と同じように考えれば、

```
c09-4.R
```

```
h <- c(137,163,137,138,168,157)
```

```
X <- sum((h-150)^2/150)
```

```
1-pchisq(X,4)
```

により、どの目の出やすさにも差がないという(つまり、900回振ったときの各目の期待頻度は150回ずつという)帰無仮説を検定すると、有意確率は0.145...となるので、有意水準5%で帰無仮説は棄却されず、このサイコロの各目の出やすさには差がないといえる。

9.7 2群間の比率の差

この話をもっと一般化して、1つのカテゴリ変数のカテゴリ間の頻度の差ではなく、独立した事象の観察頻度に差があるかどうかを考えてみる。もっとも単純な場合として、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。これは、独立2群間の比率の差の検定と呼ばれる。カイ二乗適合度検定でもいい(ただし特性をもたない者についても期待度数と観測度数の差を考えなくてはいけない)のだが、

以下では、2乗しないで正規近似によって検定してみる。

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、

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - E(\hat{p}_1 - \hat{p}_2)}{\sqrt{V(\hat{p}_1 - \hat{p}_2)}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}} \sim N(0, 1)$$

によって*9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、

c09-5.R(1)

```
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))
print(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 群間で有意差があるといえる。

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、

*9 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、確率を 2 倍して有意確率を得る。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。

```
c09-5.R(2)
```

```
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。なお、最下行の `cat` はコンソールに値を表示する関数である。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^{*10}。

```
c09-5.R(3)
```

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

母比率の推定と、その差があるかどうかの検定^{*11}、差の 95% 信頼区間を一気に出力してくれる。上で段階ずつ計算した結果と一致すること確かめてみよう。

9.8 3 群以上の比率の差

`prop.test()` 関数は、3 群以上の間で「どの群でも事象の生起確率に差がない」という帰無仮説の検定にも使える。その帰無仮説が棄却されるときに、どの群間で差があるのかをみるには、検定の多重性が生じるので、平均値の差の場合と同様、第 1 種の過誤を調整する必要がある。ボンフェローニの方法やホルムの方法を用いることができる。R の関数は `pairwise.prop.test()` である。もちろん、平均値の比較の場合に一元配置分散分析をしたときと同じように、多群間の比較というフレームにしないで、群分け変数が事象生起確率に有意な効果を持つか、言い換えると、「これら 2

^{*10} `prop.test()` の結果は、コンソールで実行すれば表示されるが、`source()` で読み込んで実行すると表示されないで、`print()` で括ってある。なお、`source()` 関数の引数として `echo=T` を付ければ、`print()` で括らなくても結果が表示される。

^{*11} 連続性の補正済み、事象が生起しない場合についても考慮してカイ二乗適合度検定をしているのだが、この操作は次章で説明する 2 つの変数の独立性のカイ二乗検定と数学的に等価である。

つの変数が独立」という帰無仮説を検定する戦略もありうる。

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、パプアニューギニア高地辺縁部の3つの村で皮膚疾患の検査結果が次の表の通りだった場合、

村	A	B	C
検査人数	120	143	160
皮膚疾患あり	7	19	24
水浴び場アクセス難度	1	2	3

次の枠内のように分析を実行することができる。

```
disease <- c(7,19,24)
total <- c(120,143,160)
prop.test(disease,total)
pairwise.prop.test(disease,total)
score <- c(1,2,3)
prop.trend.test(disease,total,score)
```

この例では、通常の `prop.test()` では3群間に皮膚疾患割合に差がないという帰無仮説が有意水準5%で棄却されないが(個々の対比較においても有意差のある村はない)、コクラン＝アーミテージの検定により、有意水準5%で皮膚疾患割合には一定の傾向があるといえる。情報が増えることによって、より検出力が高い検定をすることができる(両側検定よりも片側検定の方が検出力が高いのと同じ理屈である)。なお、傾向を示すためのスコアは外的基準に基づいて各群に割り振る。とくに外的基準がない場合は、1から連続した整数値を割り振ることもある。なお `prop.trend.test` とすれば詳細な説明が表示されるので参考にされたい^{*12}。

例題

あるIT企業の健診時に得たアンケートを集計したところ、部別の喫煙頻度は、総務部が214人中42人、営業部が658人中242人、開発部が327人中122人だった。この企業の喫煙割合は部によって差があるといえるか？

^{*12} なお、スコアを独立変数、事象生起確率を従属変数とした線型回帰を行って、回帰係数が有意ならば、回帰式から予測される各スコアごとの事象生起確率が、実際に観測された事象生起確率に適合しているかどうか、カイ二乗適合度検定を行うことも論理的には不可能ではない。けれども、傾向があることを言いたい場合、回帰式が適合しているという仮説が棄却されないよりも、「傾向がない」が棄却される方が強い論証になるし、おそらく実質的な意味がないので、通常、そういう分析は行われたい。

次の枠内のように入力すれば、部によって差がないという帰無仮説の検定の結果、得られる有意確率は約 7.5×10^{-6} なので有意水準 5% で帰無仮説は棄却され、部によって統計学的に有意差があったといえる。さらに、ホルムの方法で第1種の過誤を調整した多重比較の結果から、総務部と営業部、総務部と開発部の喫煙割合はそれぞれ有意水準 5% で有意な差があるが、営業部と開発部の喫煙割合には差があるとはいえない。なお、枠内のスクリプトで生成される図では割合と実数のグラフを並べて示してあるが、ここで検討している割合の差の比較をする目的であれば割合のグラフだけ表示すれば十分である（割合のグラフに人数を数値として書き込むこともある）。

c09-6.R

```
smoker <- c(42,242,122)
names(smoker) <- c("総務","営業","開発")
pop <- c(214,658,327)
crosstab <- rbind(smoker,pop-smoker)
rownames(crosstab) <- c("喫煙者","非喫煙者")
print(crosstab)
layout(t(1:2))
barplot(crosstab,legend=T,main="部門別喫煙者数")
barplot(crosstab/rbind(pop,pop),legend=T,main="部門別喫煙割合")
print(prop.test(smoker,pop))
print(pairwise.prop.test(smoker,pop))
```

9.9 課題

バブアニューギニアのある地方の、内陸、川沿い、海沿いの3つの村で、住民の悉皆調査によってマラリア原虫が血液中に検出される割合を調べた結果、内陸では180人中6人、川沿いでは220人中10人、海岸では80人中18人が原虫陽性だったとする。マラリア原虫陽性割合を村ごとに図示し、それらの割合の間に差があるか検討せよ。

付加的な情報としては、マラリア原虫を媒介するハマダラカの相対密度が、内陸を1とすると川沿いでは2、海沿いでは4程度になるということがわかっている。余裕があれば、ハマダラカの密度が高くなるほどマラリア原虫陽性割合が上昇する傾向があるかどうか検討せよ。

第 10 章

クロス集計

10.1 複数のカテゴリ変数を分析するために

本章では、複数のカテゴリ変数の関係を分析する方法を扱う^{*1}。カテゴリデータの分析、とくに関連性についての分析には、`vcd` ライブラリや `epitools` ライブラリを導入しておくとても非常に便利である^{*2}。これらの追加ライブラリ内の関数を使いたいときは、`library(vcd)` あるいは `require(vcd)` としてライブラリをメモリに読み込むことで、そこに含まれる関数やデータが使える状態になる。余談だが、`vcd` ライブラリには `goodfit()` という適合度検定を行う関数が含まれていて、前章の「複雑な例題」も以下のように簡単に実行することができる^{*3}。

*1 実は前章の課題も村落とマラリア原虫陽性 / 陰性という 2 つのカテゴリ変数の関係の分析とみなせるが、敢えてそういう見方をしなかった。

*2 自分のコンピュータに管理者権限でインストールした R に `vcd` ライブラリを導入したい場合は、`install.packages("vcd", dep=T)` として、出てくるウィンドウで適当なミラーサーバ（日本国内では Japan(Tsukuba) または Japan(Tokyo) を推奨）を選ぶだけでよい。ブロードバンドなら 1 分もかからないだろう。

*3 ただし分布の推定法に ML 法と MinChisq 法があり、前章で説明した期待値を出すのは ML 法であるが、その場合、適合度の検定法も尤度比検定になってしまうので、前章の方法とまったく同じ結果にはならない。なお、期待値推定のコード（内部処理）はプロンプトに `goodfit` と打てば（関数名の後の括弧をつけずに注意）わかるし、適合度検定のコードは `getS3method("summary", "goodfit")` とすれば見える。UseMethod を使って定義された関数（`generic.class` という形式）のコードは、一般に、ただ `generic.class` と打つのでは見えず、`getS3method("generic", "class")` とする必要がある。

```
library(vcd)
hh <- as.table(c(79,61,13,2))
names(hh) <- 0:3
print(res <- goodfit(hh,"poisson","ML"))
summary(res)
plot(res)
```

10.2 2つのカテゴリ変数の独立性の検定

まずは2つのカテゴリ変数が独立である（つまり、関係がない）という帰無仮説を検定する方法について説明する。

例えば、肺がんと判明した男性患者100人と、年齢が同じくらいの健康な男性100人を標本としてもってきて、それまで10年間にどれくらい喫煙をしたかという聞き取りを行うという「症例対照研究 (case control study)*⁴」を実施したとする。喫煙の程度を「一度も吸ったことがない」から「ずっとヘビースモーカーだった」まで何段階かのスコアを振れば、200人の一人一人について喫煙状況という変数に値が割り振られることになる。喫煙状況という変数と肺がんの有無という変数の組み合わせを考え、クロス集計することによって、それらが独立であるかどうか（関連がないかどうか）を検討することになる*⁵。

10.2.1 クロス集計とは？

前章でみたとおり、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関係について検討したいときには、まずそれ

*⁴ 患者対照研究ともいう。

*⁵ ただし、むしろこのデザインは、肺がん患者は健康な人に比べて、どれくらい喫煙していた割合が高いか、を評価するためのデザインである。独立性の検定は、一時点の断面研究（英語では cross-sectional study で、「横断的研究」ともいう）で調べた属性変数間でなされるのが普通である。症例対照研究では、既に亡くなっている人が除かれてしまっているので、注目している要因によってその疾患が起こりやすくなる程度が過小評価されるかもしれない。逆に、喫煙者而非喫煙者を100人ずつ集めて、その後の肺がん発生率を追跡調査する前向きのコホート研究 (cohort study) では、非喫煙群に比べて、喫煙者ではどれくらい肺がんの発生率が高くなるかを評価できる。「……に比べてどれくらい高いか」を示すためには、リスク比とかオッズ比のような「比」を用いるのが普通である。これらの「比」については本章後半で扱う。

らの組み合わせの度数を調べた表を作成する。これをクロス集計表と呼ぶ。例えば、パプアニューギニアのある村で、巡回健診を受けた13人の性別と病気の有無の調査結果が下表の通りであったとする*6。

人	1	2	3	4	5	6	7	8	9	10	11	12	13
性別	男	男	男	男	男	男	女	女	女	女	女	女	女
病気	有	有	有	無	無	無	有	有	有	有	無	無	無

この生データをもっとも簡単にRに入力し、性別と病気のクロス集計表を作るには次の枠内を打つ。1行目が個人番号の定義、2行目が性別（最初の6人が男性で後の7人が女性なのを、1の6回繰り返し rep(1,6) と、2の7回繰り返し rep(2,7) の組み合わせで表現し、as.factor() によって要因型に変換している）、3行目で levels() を使って性別の水準に「男」「女」と名前をつけている。4行目で病気の有無を、有を1、無を2として与えてから as.factor() で要因型に変換し、5行目でその水準に名前をつけている。6行目の table() という関数が、生のカテゴリ変数値の組み合わせをカウントしてクロス集計表を作ってくれ、最下行の mosaicplot() という関数が、それをグラフ表示してくれる。

c10-1.R

```
pid <- 1:13
sex <- as.factor(c(rep(1,6),rep(2,7)))
levels(sex) <- c("男","女")
disease <- as.factor(c(1,1,1,2,2,2,1,1,1,1,2,2,2))
levels(disease) <- c("有","無")
print(ctab <- table(sex,disease))
mosaicplot(ctab,main="2 × 2 クロス集計表のモザイクプロット例")
```

クロス集計表としては、次の枠内の結果が得られる。

```
      disease
sex  有  無
  男  3  3
  女  4  3
```

とくに、2つのカテゴリ変数が、この例のようにともに2値変数のとき、そのクロス集計は2×2クロス集計表(2×2分割表)と呼ばれ、その統計的性質が良く調べられている。

*6 このデザインは断面研究である。

10.2.2 独立性のカイ二乗検定の原理

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もっとも有名である（実はカイ二乗適合度検定と同じ原理である）。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一致を調べてもいいが、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ので、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い場合に「統計的に有意な関連があった」と判断するのである。

	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た結果が上表として得られたとき、母集団の確率構造が、

	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、期待される度数は*7、

	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、一般に 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

*7 ただし $N = a + b + c + d$ である。

ならば $\Pr(A \cap B) = \Pr(A) \Pr(B)$ と考えられることを使って, $\Pr(A)$ と $\Pr(B)$ を母数として推定する*8。2つの母数をデータから推定したため, 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。 $\Pr(A)$ 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, 母集団の各組み合わせの確率は下式で得られる。

$$\begin{aligned}\pi_{12} &= (b+d)(a+b)/(N^2) \\ \pi_{21} &= (a+c)(c+d)/(N^2) \\ \pi_{22} &= (b+d)(c+d)/(N^2)\end{aligned}$$

これらの値を使えば,

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

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, Yates 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。

もちろん, Rにはこの検定を簡単に行う関数を実装されている。例えば $a=12$, $b=8$, $c=9$, $d=10$ なら次の通り*9。

*8 $\Pr(X)$ はカテゴリ X の出現確率を示す記号である。

*9 連続性の補正をしたくない場合は, 2行目が `chisq.test(x, correct=F)` となるが, 通常その必要はない。

```
c10-2.R
x <- matrix(c(12,9,8,10),nr=2)
chisq.test(x)
```

カテゴリ変数 A と B について各個人の生データが名義尺度として得られているときは、`table(A,B)` とすればクロス集計表ができて、`chisq.test(table(A,B))` とすれば、独立性のカイ二乗検定ができる^{*10}。

例題

肺がんの患者 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^a、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。肺がん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。

^a この操作をペアマッチサンプリングという。ただし、このような症例対照研究でマッチングをすると、却ってバイアスが生じる場合があるので注意されたい。

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

	肺がん患者群	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。肺がんと喫煙が無関係という帰無仮説の下で期待される各カテゴリの人数は、

	肺がんあり	肺がんなし
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。したがって、Yates の連続性の補正を行なったカイ二乗統計量は、

^{*10} 実は `chisq.test(A,B)` でもカイ二乗検定はできてしまうが、表を与える形にしておく方がよい。なお、R の `chisq.test()` 関数では、`simulate.p.value=TRUE` というオプションを使えば、そのカイ二乗値より大きなカイ二乗値が偶然得られる確率を、シミュレーションによって計算させることも可能である。一般に、カイ二乗分布による近似的な検定よりも正確な p 値が得られるが、遅いコンピュータだと計算時間が長くなるのが欠点である。

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

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\dots` となり、有意水準5%で帰無仮説は棄却される。つまり、肺がんの有無と過去の喫煙の有無は独立とはいえない。`chisq.test()` 関数を使って、

```
X <- matrix(c(80,20,55,45),nr=2)
chisq.test(X)
```

と入力すれば、次の枠内の結果が得られる。

```
Pearson's Chi-squared test with Yates' continuity correction

data: X
X-squared = 13.1282, df = 1, p-value = 0.0002909
```

この検定は、前章で説明した `prop.test()` を使って、肺がん群と対照群の間で、過去の喫煙者の割合に差があるかどうかを検定することと数学的に同値である。次の枠内のコードを実行すれば、まったく同じ有意確率が得られる。

```
smoker <- c(80,55)
pop <- c(100,100)
prop.test(smoker,pop)
```

ただし、カイ二乗検定はあくまで正規近似なので、各カテゴリの組み合わせごとの期待度数が小さすぎると近似が悪くなってしまう。一般に、期待度数が5以下の組み合わせを検討すべき組み合わせ数の20%以上あるときは^{*11}カイ二乗検定は適当でないとされる。

10.2.3 フィッシャーの直接確率（正確な確率）

期待度数が低い組み合わせがあるときには、前章で述べたようにカテゴリを併合して変数を作り直す方法もあるし、シミュレーションで有意確率を求めることもできる

*11 例えば 2×2 クロス集計表なら1つでも期待度数5以下のセルがあれば該当する。

が、実はもっといい手がある。

ここで調べたいのは組み合わせの数なので、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、現実に行われているクロス集計表が得られる確率よりも低い確率でしか得られないクロス集計表の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の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、フィッシャーの直接確率、あるいは、フィッシャーの正確な確率（検定）という^{*12}。これは近似ではないので、期待度数が低い組み合わせがあっても問題ない。

もう少し丁寧に説明すると、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 （1 でない個体数が $n_2 = N - n_1$ ）という状況を考え、この n_1 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」という帰無仮説の下で、現実に行われたデータと同じあるいはそれより偏ったデータが偶然得られる確率（有意確率）になる^{*13}。

フィッシャーの正確な確率は、R では、`fisher.test(table(A,B))` で実行できる。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。

^{*12} 英語では Fisher's exact probability test という。

^{*13} 有限母集団からの非復元抽出になるので、平均値 $E(a)$ 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。多くの組み合わせについて計算せねばならないため、実際には手計算で実行することはまずありえず、ソフトウェア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、まったく同じ確率の表があった場合に、それを足し算に含めるのかどうかは難しい点である。これを乱数によって決める「ランダム検定」という手法もあるが、あまり一般的ではない。

例題

上記の肺がんの有無と過去の喫煙の有無のデータでフィッシャーの正確な確率を計算せよ。

既に X にクロス集計表が付値されているので、`fisher.test(X)` を実行すると、有意確率は 0.0002590 と得られ、有意水準 5% で「肺がんの有無と過去の喫煙の有無は独立」という帰無仮説は棄却される。なお、このように 2×2 分割表を分析する場合は、`fisher.test()` 関数は、後で説明するオッズ比とその 95% 信頼区間も同時に計算してくれる。

サンプルサイズが小さい場合について、実際に数値を使って説明しておく。フィッシャーの正確な確率は仮定が少ない分析法なので、動物実験などでは重宝する。

例題

15 人の健康なボランティアに数値計算をしてもらったところ、得点が高得点群と低得点群の 2 群にきれいに分かれたとする。この人たちに、その日に朝食を食べてきたかどうかを尋ねた結果、食べてきた人とこなかった人がいたとする。個人別のデータは下表の通りであったとする。朝食を食べたかどうかと数値計算の得点が独立かどうか検定せよ。

ID 番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
得点(高:H, 低:L)	H	H	L	H	L	L	H	H	L	L	L	L	L	L	L
朝食(有:B, 無:N)	B	B	B	N	N	N	B	B	N	N	N	B	N	B	N

この検定を実行するための R のコードは次の枠内の通り。

c10-3.R

```
calc <- as.factor(c(1,1,2,1,2,2,1,1,2,2,2,2,2,2))
levels(calc) <- c("高得点", "低得点")
bf <- as.factor(c(1,1,1,2,2,2,1,1,2,2,2,1,2,1,2))
levels(bf) <- c("朝食あり", "朝食なし")
print(X <- table(bf, calc))
fisher.test(X)
```

なお、ここではラベル付き要因型変数を `as.factor()` と `levels()` を使って 2 つの文で与えているが、`factor()` を使って 1 つの文で一気に定義することも可能である。例えば最初の 2 行は

```
calc <- factor(c(1,1,2,1,2,2,1,1,2,2,2,2,2,2),
  levels=c(1,2), labels=c("高得点", "低得点"))
```

とできる。c10-3.R の最下行の `fisher.test()` の結果、p-value が 0.1189 なので、5% 水準で帰無仮説は棄却されない。つまり、このデータは、数値計算の得点と朝食を食べたかどうかが独立であることを示唆する（もっとも、データが少ないので、検出力が足りずに第 2 種の過誤が起きている可能性がある）。この計算結果は、以下のように考えて導かれる。まず、下から 2 行目の結果からクロス集計表を書いてみると、次の通りである^{*14}。

	高得点	低得点	合計
朝食あり	4	3	7
朝食なし	1	7	8
合計	5	10	15

15 人のうち 5 人が高得点で、7 人が朝食あり、という条件が決まっているとき^{*15}、偶然この表が得られる確率は、15 人のうち高得点の 5 人の内訳が、朝食を食べた 7 人から 4 人と、食べていない 8 人から 1 人になる確率となる。つまり、15 から 5 を取り出す組み合わせのうち、7 から 4 を取り出し、かつ残りの 8 から 1 を取り出す組み合わせをすべて合わせたものが占める割合になるので、 ${}_{7}C_4 \cdot {}_{8}C_1 / {}_{15}C_5 \approx 0.0932$ である。

つまり、上のクロス集計表が、偶然（2 つの変数に何も関係がないとき）得られる確率は 0.0932 ということである。これだけでも既に 5% より大きいので、「2 つの変数が独立」という帰無仮説は棄却されず、得点の高低と朝食の有無は関係がないと判断していいことになる。

しかし、有意確率、つまり第 1 種の過誤を起こす確率は、得点の高低と朝食の有無には関係がないと判断した場合にそれが間違っている確率なので、この表だけではなく、この表よりも偶然得られる確率が低い表が得られる確率をすべて足さねばならない。周辺度数が上の表と同じ表は、

(1)	高得点	低得点	(2)	高得点	低得点	(3)	高得点	低得点	合計
朝食あり	0	7		1	6		2	5	7
朝食なし	5	3		4	4		3	5	8
合計	5	10		5	10		5	10	15

^{*14} `library(vcd)` として `vcd` ライブラリが使えるようにしておいて、`mar_table(X)` とすれば、このような周辺度数を含めた集計表の作成も可能である。epitools ライブラリの `table.margins()` 関数も同じ機能をもつ。

^{*15} 「各変数については母比率が決まっているとき」ということで、このことを「全ての周辺度数が固定されているとき」ともいうのである。

(4)	高得点	低得点	(5)	高得点	低得点	(6)	高得点	低得点	合計
朝食あり	3	4	4	3	5	2	7		
朝食なし	2	6	1	7	0	8	8		
合計	5	10	5	10	5	10	15		

の計 6 種類しかない（説明するまでもないが、(5) がこの例題で得られている表である）。(1) や (6) の表よりもさらに稀な場合を考えると、(1) の先は高得点かつ朝食ありの人の数がマイナスになってしまうし、(6) の先は高得点かつ朝食なしの人の数がマイナスになってしまうので、そういうありえない表は考えなくていい。

そこで、すべての表について、偶然得られる確率を計算すると^{*16}、以下の数値が得られる^{*17}。

$$(1) \quad {}_7C_0 \cdot {}_8C_5 / {}_{15}C_5 \simeq 0.0186$$

$$(2) \quad {}_7C_1 \cdot {}_8C_4 / {}_{15}C_5 \simeq 0.1632$$

$$(3) \quad {}_7C_2 \cdot {}_8C_3 / {}_{15}C_5 \simeq 0.3916$$

$$(4) \quad {}_7C_3 \cdot {}_8C_2 / {}_{15}C_5 \simeq 0.3263$$

$$(5) \quad {}_7C_4 \cdot {}_8C_1 / {}_{15}C_5 \simeq 0.0932 \text{ (上で計算した通り)}$$

$$(6) \quad {}_7C_5 \cdot {}_8C_0 / {}_{15}C_5 \simeq 0.0070$$

元の表 (= (5)) より得られる確率が低い（つまりより偶然では得られにくい）表は (1) と (6) なので、それらを足せば、元の表の両側検定（どちらに歪んでいるかわからない場合）での有意確率が、 $0.0932 + 0.0186 + 0.0070 = 0.1188$ となることがわかる（fisher.test() の結果と小数第 4 位で 1 違うのは丸め誤差のせいである）。

10.3 研究デザインと疫学指標

独立とはいえないなら、次に調べることは、どの程度の関連性があるのかということである。カテゴリ変数間の関連については、従来より疫学分野で多くの研究が蓄積されてきた。疫学研究では、研究デザインによって、得られる関連性の指標が異なることに注意しなければならない。その意味で、具体的な解析方法に入る前に、疫学の

^{*16} 既出の通り、R で組み合わせ計算を行う関数は choose() である。例えば ${}_7C_3$ は、choose(7,3) で計算できる。

^{*17} これらの確率をすべて足すと 1 になる。ここで計算値として書いた値を使うと 0.9999 となるが、これは丸め誤差のせいであり、厳密に計算すれば 1 になる。

基礎知識が必要なのでまとめておく*¹⁸。

集団内の疾病の状況を表すためには、たんに患者数だけでは不十分である。集団全体のうちどれくらいの人に疾病が観察されるのか、という意味で、分母と分子を厳密に規定する必要がある。まず、どのくらいの規模のどういう集団をどのくらいの期間観察したのか、という意味で、分母の定義が本質的に重要である。一方、分子の定義には、どういう診断基準で判定したのかということと、その診断基準の信頼性 (reliability)・妥当性 (validity)・正確さ (accuracy)・精度 (precision) を把握し高めることが重要である。大雑把に言えば、信頼性は同じ条件で繰り返し測定した場合の再現性が確保されているかを意味する。妥当性は測りたいものがきちんと測れているかどうかを意味する。正確さは系統的なズレ (バイアス) がいないかどうかを示す。精度は偶然誤差の小ささを示す (例えば小数点以下何桁目まで測れているかということ)。

10.3.1 頻度の指標

具体的な指標としては、まず、次の 3 つを区別する必要がある。これらはすべて疾病の発生状況を示すための頻度の指標である。

有病割合 (prevalence)

有病率と呼ばれることもあるが、時間概念を含まないため「率 (rate)」ではないので、割合と呼ぶ方が紛れがない。一時点での人口に対する患者の割合で無次元である。一時点でのということを示すには、point prevalence という。急性感染症で prevalence が高いなら患者が次々に発生していることを意味するが、慢性疾患の場合はそうとは限らない。行政施策として必要な医療資源や社会福祉資源の算定に役立つ。日本では、高血圧や高コレステロール血症の prevalence が高く、対策がとられているが、なかなか奏効しない。

*¹⁸ 疫学データについて関連性の指標を説明する都合上、以下では、何らかのリスクファクターへの曝露の有無と疾病の有無の関連性の分析について説明するが、数学的にはそれに限らず、2 つのカテゴリ変数間の関連性の程度について広く適用可能な指標である。

累積罹患率 (cumulative incidence)

通常、たんにリスク (risk) といえ、この累積罹患率を指す。この指標も時間概念を含まないので「率」ではないのだが、慣習的にこのような呼び名となっている。期首人口のうち観察期間中に病気になった人数の割合であり、無次元である。当然、観察期間が短ければ小さい値になるし、観察期間が長ければ大きい値になるので、「20 年間のがんの発症リスク」のように、期間をつけて表現しなくては意味が無い。観察期間中に転居などで脱落した者は、通常、分母から除外する（脱落分を正しく扱うためには第 13 章で説明する生存時間解析という手法を用いる）。

罹患率 (incidence rate)

発生率ともいう。個々の観察人年の総和で患者の発生数を割った値であり、観察期間によらない値になる。次元は 1 / 年であり、時間当たりの罹患患者発生速度を意味する。International Epidemiological Association の Last JM [Ed.] “A Dictionary of Epidemiology, 4th Ed.” (Oxford Univ. Press, 2001) に明記されているように、incidence だけだと発生数（罹患数）を意味する。通常は、感受性をもつ人の中で新たに罹患する人が分子であり、一度罹患した人は観察対象から除外する。したがって、発生数に再発を含む場合はそのように明記する必要がある。意味としては、瞬間における病気へのかかりやすさ。つまり疾病罹患の危険度（ハザード）を示す。疾病発生状況と有病期間が安定していれば、平均有病期間は、有病割合を罹患率で割った値にほぼ等しくなる。無作為化比較試験 (Randomized Controlled Trial; RCT)^a でよく使われる指標である。

^a 研究対象を乱数を使って無作為に 2 群に分け、片方の群を実験群、もう一方を対照群に割り付け、実験群と対照群の間で疾病・死亡・回復など適切な帰結を厳密に比較することにより実験の効果を評価する研究方法。

さらに、オッズ (odds) という概念を押さえておく必要がある。オッズとは、ある事象が起きる確率の起きない確率に対する比である。一時点での非患者数に対する患者数の比を疾病オッズ (disease-odds) と呼ぶ。また、症例対照研究などで、過去に何らかの危険因子に曝露した人数の、曝露していない人数に対する比を曝露オッズ (exposure-odds) と呼ぶ。

参考までにまとめておくと、以下の指標も疫学的には重要である。

死亡率 (mortality rate) 人口のうち、ある一定期間に死亡した人数の割合。1 年間の死亡数を 1 年間の観察人数で割るのが普通なので、次元は 1/年となる。分母分子ともカテゴリ分けしてカテゴリごとに計算した死亡率をカテゴリ別死亡率 (category-specific mortality rate) という。例えば、性・年齢別死亡率 (age-sex specific mortality rate) はカテゴリ別死亡率の一例である。死因別

死亡率 (disease-specific mortality rate) を計算する際は、分母は共通で、分子のみカテゴリ別になる。観察年人としては、本来なら 1 人ずつの観察人日を積み上げて 365 日か 366 日で割るべきだが、通常は 1 年間の半ばの人口が 1 年間ずっと観察されたと考え、その値を用いる。この「1 年間の半ばの人口」を年央人口と呼ぶ(日本の人口統計では 10 月 1 日人口を用いる)。疾病がもたらす結果を示す指標の 1 つといえるが、年齢によって大きく異なるので、年齢で標準化することが多い。

致命率 (case-fatality rate) ある疾病に罹患した人のうち、その疾病で死亡した人の割合 (通常は % で表される無次元の量であり、「率」ではない)。一般に、致命率 = 死亡率 / 罹患率という関係が成り立ち、疾病の重篤度を示す指標といえる。ただし、慢性疾患では有病期間が長いので、観察期間の設定が重要である。この場合、罹患してから観察を開始し、観察年人を分母として、観察された死亡数を分子とすれば、年当たりの致命「率」が得られる。

死因別死亡割合 (proportional mortality rate; PMR) ある特定の死因による死亡が全死亡に占める割合。これも時間概念を含まないので rate ではないが、慣例的にこう呼ばれる。増減はその疾患の増減だけでなく、他の疾患の増減とも連動する。

PMI (proportional mortality indicator) 50 歳以上死亡割合と訳す。全死亡数に対する 50 歳以上死亡数の占める割合を % で表示した値である。計算に必要なのは年齢 2 区分の死亡数のみなので、統計資料が整備されていない途上国でも信頼性が高い値を得ることができるのが最大の利点である。

10.3.2 効果の指標

頻度の指標を押さえた上で、何らかの危険因子への曝露があると、曝露がなかった場合に比べて何倍くらい病気に罹りやすくなる効果をもつかといったことを推論することになる。効果の指標としては、以下の 4 つを区別しておこう。とくにリスク比やオッズ比はよく使われる指標である。

相対危険 (Relative Risk)

以下 3 つを総称して相対危険 (relative risk) という。どれも同次元の指標の比なので無次元である。リスク比を指している場合が多い。

リスク比 (risk ratio) 累積罹患率比 (cumulative incidence rate ratio) ともいう。曝露群のリスクの非曝露群のリスクに対する比である。

罹患率比 (incidence rate ratio) 曝露群の罹患率の非曝露群の罹患率に対する比をいう。

死亡率比 (mortality rate ratio) 曝露群の死亡率の非曝露群の死亡率に対する比をいう。罹患率比と死亡率比を合わせて率比 (rate ratio) という。

オッズ比 (odds ratio)

読んで字のごとく、オッズの比である。2 種類のオッズ比 (odds ratio), 即ちコホート研究における、曝露群の疾病オッズの非曝露群の疾病オッズに対する比である疾病オッズ比 (disease odds ratio) と、症例対照研究における症例群の曝露オッズの対照群の曝露オッズに対する比である曝露オッズ比 (exposure odds ratio) は、数値としては一致する。断面研究では、曝露群・非曝露群という分け方も症例群・対照群という分け方も、どちらにしても後付けになってしまうが、疾病オッズ比と曝露オッズ比が数値として一致するため、どちらで考えても差し支えない。

寄与危険 (attributable risk)

危険因子への曝露による発症増加を累積罹患率(リスク)または罹患率の差で表した値を寄与危険 (attributable risk) という。つまり、累積罹患率差 = リスク差 (risk difference), または罹患率差 (incidence rate difference) である。超過危険 (excess risk) ともいう。

寄与割合 (attributable proportion)

曝露群の罹患率のうち、その曝露が原因となっている割合を寄与割合 (attributable proportion) と呼ぶ。つまり罹患率差を曝露群の罹患率で割った値になる。罹患率比から 1 を引いて罹患率比で割った値とも等しい。

参考までにまとめておくと、その他の効果の指標としては以下のものがある。

相対差 要因ももたず発症もしていない者のうち、要因をもった場合のみ発症する割合を相対差という。罹患率差を、対照群の罹患率を 1 から引いた値で割った値になる。

母集団寄与率 母集団において真に要因の影響によって発症した者の割合を母集団寄与率という。曝露群と非曝露群を合わせた集団全体の罹患率を π , 非曝露群の罹患率を π_2 として、 $(\pi - \pi_2)/\pi$ である。

要因の有無と病気の有無がまったく関係がなければ、リスク比もオッズ比も 1 になることが期待される。それぞれ信頼区間を計算して、例えば 95% 信頼区間が 1 を含まなければ、5% 水準で有意な関連があるといえる^{*19}。

ところで、病気のリスクは、全体（期首人口）のうちで病気を発症する人の割合であったから、まず全体を把握していないと定義できない。つまり、まず観察対象全体で曝露群と対照群を把握しておいて、経時的に追跡調査して、それぞれの群で何人ずつ発症するかを調べるといって、「前向き研究 (prospective study)」（この意味ではコホート研究 (cohort study) とかフォローアップ研究 (follow-up study) と言ってもいい）でないと、リスク比（に限らず相対危険すべて）は計算できない。

つまり、症例対照研究 (case-control study)^{*20}とか断面研究 (cross-sectional study)^{*21}では、曝露時点での全体が未知なので、原理的にリスクを計算できないことになる。大きな曝露を受けた人は調査時点以前に病気を発症して死んでしまった可能性があるので、症例対照研究や断面研究から無理にリスクを見積もろうとするとリスクを過小評価してしまうことになる。これらの研究デザインでは、オッズ比を計算するのが普通である。

では、クロス集計表から、これらの値を計算してみよう。次の表（表 とする）を考える。

^{*19} ただし、重要なのは 95% 信頼区間が 1 を含むかどうかという意思決定だけではなく、むしろリスク比やオッズ比の点推定量と信頼区間の値そのものである。知りたいのは、非曝露群に比べて曝露群のオッズやリスクが何倍になっているかということである。Rothman や Greenland に代表される現代の疫学者は、オッズ比やリスク比の有意性をみる仮説検定は、せっかく関連性の程度が得られているのに、それを有無という 2 値に還元してしまうので情報量の損失が大きく、あまり意味がないと言っている。それゆえ、疫学研究では検定結果よりも 95% 信頼区間そのものの方が重要である。Rothman は関連の程度に応じた有意確率の変化を示すという意味で、p-value 関数（リスク比やオッズ比を横軸にとり、「真の値が横軸の値と差が無い」帰無仮説の検定の有意確率 (=p-value) を縦軸にとり、とりうるすべてのリスク比やオッズ比について線で結んだグラフ）を求めるべきだと主張している（Rothman KJ 著、矢野栄二・橋本英樹監訳『ロスマンの疫学』篠原出版新社、pp.150-156）。

^{*20} 調査時点で、症例（患者）を何人サンプリングすると決め、同数でもいいが通常はその何倍かの人数の対照（その病気でないことだけが症例と違って、それ以外の条件はすべて患者と同じことが望ましい。ただし原則としてマッチングに使った変数で層別解析しなくてはならない）を選んで、それぞれが過去に受けた曝露要因や、現在の生活習慣、態度などを調べることによって、その病気の原因を探る方法論。

^{*21} 調べてみないと患者かどうかわからないような場合や、因果の向きがはっきりしない変数間の関係を見たいときは、全体で何人サンプリングすると決めて一時点で調査する。こういう方法論を断面研究という。

	疾病あり	疾病なし	合計
曝露あり	a	b	m_1
曝露なし	c	d	m_2
合計	n_1	n_2	N

10.3.3 リスク比とオッズ比の点推定量

点推定量の計算は簡単である。この表でいえば、リスク比は

$$\frac{a/m_1}{c/m_2} = \frac{am_2}{cm_1}$$

となる。疾病オッズ比は

$$\frac{a/b}{c/d} = \frac{ad}{bc}$$

である。曝露オッズ比は

$$\frac{a/c}{b/d} = \frac{ad}{bc}$$

となり、疾病オッズ比と一致するので、結局、「オッズ比は $(ad)/(bc)$ 」といえる。

ただし、R の `fisher.test()` で計算されるオッズ比は、 $(ad)/(bc)$ という単純な計算式から得られる値と異なっている。`fisher.test()` では周辺度数をすべて固定したクロス集計表の最初の要素に対して、非心度パラメータがオッズ比で与えられるような非心超幾何分布を仮定して最尤推定がなされる。

`vcd` ライブラリの `oddsratio()` 関数で `log=F` オプションを付けると、上述の通り $(ad)/(bc)$ を計算してくれる。ただし、 a から d のどれかが 0 のときは、

$$\frac{(a + 0.5)(d + 0.5)}{(b + 0.5)(c + 0.5)}$$

が計算される。`log=F` オプションを付けなければ対数オッズ比が計算される。対数オッズ比を求めた場合のみ、`summary(oddsratio())` により、「オッズ比が 1」を帰無仮説とする検定の有意確率が得られる。一方、95% 信頼区間は、`log=F` オプションを付けても付けなくても、`confint(oddsratio())` とすれば推定できる (`log=F` を付けた場合は対数オッズ比の 95% 信頼区間、付けなければオッズ比の 95% 信頼区間が得られる)。また、`epitools` ライブラリにある同名の `oddsratio()` 関数は、もっと複雑な計算をしているので、注意が必要である。

オッズ比が重要なのは、稀な疾患の原因を研究するとき、リスク比のよい近似になるためと言われている。例えば、送電線からの高周波が白血病の原因になるという仮説を検証するために、送電線からの距離が近い場所に住んでいる人（曝露群）と、遠いところに住んでいる人（対照群）をサンプリングして、5年間の追跡調査をして、5年間の白血病の累積罹患率（リスク）を調査することを考えよう。白血病は稀な疾患だし、高周波に曝露しなくても発症することもあるので、このデザインでリスク比を計算するためには、莫大な数のサンプルをフォローアップする必要があり、大規模な予算とマンパワーが投入される必要があるだろう。

例えば曝露群と対照群それぞれ 10 万人をフォローアップした調査結果が下表のようであったとすると、

	白血病発症	発症せず	合計
送電線近くに居住	4	99996	100000
送電線から離れて居住	2	99998	100000
合計	6	199994	200000

$(4/100000)/(2/100000) = 2$ から、リスク比が 2 なので、送電線の近くに住むことで白血病を発症するリスクは、送電線から離れて住む場合の 2 倍になったとすることができる（疾病オッズ比をみると、 $(4 * 99998)/(2 * 99996) \approx 2.00004$ と、ほぼリスク比と一致している）。こうして得られるリスク比は、確かに原理的に正しくリスクを評価するのだが、稀なリスクの評価のためには大規模な調査が必要になるので、効率が良いとはいえない。そこで通常は、前向きのコホート研究ではなく、症例対照研究を行って、過去の曝露との関係を見る。この場合だったら、白血病患者 100 人と対照 100 人に対して、過去に送電線の近くに居住していたかどうかを聞くわけである。それで得られた結果が、仮に下表のようになったとしよう。

	白血病	白血病でない	合計
送電線近くに居住した経験あり	20	10	30
送電線から離れて居住	80	90	170
合計	100	100	200

この場合、白血病かつ送電線の近くに居住した経験がある 20 人は、送電線の近くに住んだ経験がある人からのサンプルではなく、白血病患者からのサンプルなので、リスク（累積罹患率）が定義できず、リスク比も計算できない。形の上から無理やり計算しても意味はない。しかし、曝露オッズは計算できる。白血病の人の送電線の近くに居住した経験の曝露オッズは 0.25 となり、白血病でない人の曝露オッズが 0.111... となるので、これらの曝露オッズの比は 2.25 となる。この値は母集団におけ

るリスク比のよい近似になることが知られているので、このように稀な疾患の場合は、大規模コホート研究をするよりも、症例対照研究で曝露オッズ比を求める方が効率が良い（もちろん、コストさえかければ大規模コホート研究の方が強い証拠となるデータが得られるが、問題は得られる結果がコストに見合うかどうかである。金銭的コストばかりではなく、調査に協力してくれる人の負担や、その疾患の社会的インパクトも考慮して判断されねばならない）。

原理的に前向き調査ができない場合もある。とくに、薬害と呼ばれる現象は、妙な病気が見つかったときに、後付けで原因を探ることになるので、症例対照研究にならざるを得ない。例えば、スモンとかサリドマイドは、そうやって原因がわかった問題である。腕が短く生まれた子どもの母親と、そうでない子どもの母親に、妊娠中に飲んだ薬の有無を尋ねて、特定の時期にサリドマイドを飲んだという曝露による疾病オッズ比が有意に大きい結果が得られたのである。

10.3.4 リスク比とオッズ比の 95% 信頼区間

次に、リスク比とオッズ比の 95% 信頼区間を考えよう。まずリスク比の場合から考えるために、曝露群と非曝露群をそれぞれ m_1 人、 m_2 人フォローアップして、曝露群で X 人、非曝露群で Y 人が病気を発症したとする。得られる表は、

	発症	発症なし	合計
曝露あり	X	$m_1 - X$	m_1
曝露なし	Y	$m_2 - Y$	m_2
合計	$X + Y$	$N - X - Y$	N

となる。このとき、母集団でのリスクの点推定量は、曝露があったとき $\pi_1 = X/m_1$ 、曝露がなかったとき $\pi_2 = Y/m_2$ である。リスク比の点推定量は $RR = \pi_1/\pi_2 = (Xm_2)/(Ym_1)$ となる。

リスク比の分布は N が大きくなれば正規分布に近づくので、正規分布を当てはめて信頼区間を求めることができるが、普通は右裾を引いているので対数変換または立方根変換 (Bailey の方法) をしなくてはならない。対数変換の場合、95% 信頼区間の下限と上限はそれぞれ、

$$RR \cdot \exp(-\text{qnorm}(0.975) \sqrt{1/X - 1/m_1 + 1/Y - 1/m_2}) \quad (\text{下限})$$

$$RR \cdot \exp(\text{qnorm}(0.975) \sqrt{1/X - 1/m_1 + 1/Y - 1/m_2}) \quad (\text{上限})$$

となる。なお、 RR が大きくなると、対数変換ではうまく近似できないので立方根変換しなくてはならないが、複雑なのでここでは説明しない。R のコードは以下の通

り。 p 値は帰無仮説 $RR = 1$ の検定の有意確率である。

```
c10-4.R
riskratio2 <- function(X,Y,m1,m2) {
  data <- matrix(c(X,Y,m1-X,m2-Y,m1,m2),nr=2)
  colnames(data) <- c("疾病あり","疾病なし","合計")
  rownames(data) <- c("曝露群","対照群")
  print(data)
  RR <- (X/m1)/(Y/m2)
  n1 <- X+Y; T <- m1+m2; n2 <- T-n1
  p.v <- 2*(1-pnorm(abs((X-n1*m1/T)/sqrt(n1*n2*m1*m2/T/T/(T-1))))))
  RRL <- RR*exp(-qnorm(0.975)*sqrt(1/X-1/m1+1/Y-1/m2))
  RRU <- RR*exp(qnorm(0.975)*sqrt(1/X-1/m1+1/Y-1/m2))
  cat("リスク比の点推定量:",RR,"(p=",p.v,
      ") 95% 信頼区間=[",RRL,"",RRU,""]\n")
}
riskratio2(4,2,100000,100000)
```

結果は以下の通り。

	疾病あり	疾病なし	合計
曝露群	4	99996	1e+05
対照群	2	99998	1e+05

リスク比の点推定量: 2 (p= 0.4142103) 95% 信頼区間=[0.3663344 , 10.91899]

ちなみに `epitools` ライブラリには `riskratio()` という関数があり、先に非曝露、発症なしのデータを与える仕様なので注意が必要だが、

```
library(epitools)
riskratio(c(99998,2,99996,4))
```

よって、曝露 2(`Exposed2`) の行に点推定量 2 と 95% 信頼区間 (0.37, 10.9) が得られる。

また、率比については別に `rateratio()` という関数があって、分母を観察年とした率比とその信頼区間を計算してくれる。信頼区間の計算は `method` オプションで "midp" または "wald" または "boot" の 3 種類が指定できる。非曝露群のデータを曝露群のデータより先に指定することに注意しなければならないが、使い方は簡単である。なお、この関数は、`method="wald"` オプションをつけないと、点推定量について

も median-unbiased な推定値を計算するので、率比といっても単純な率の比とはやや異なる。簡単のため曝露群でも対照群でも白血病発症時点は観察終了直前だったとすれば、

```
library(epitools)
rateratio(c(2,4,5*100000,5*100000),method="wald")
```

により、率比の点推定量は 2, 95% 信頼区間は (0.37, 10.9) が得られ、リスク比の値と一致する(ただし、median-unbiased な推定結果だと、これよりかなり幅が広がる)。

次にオッズ比の信頼区間を考える。表の a, b, c, d という記号を使うと、オッズ比の点推定量 OR は、 $OR = (ad)/(bc)$ である。オッズ比の分布も右裾を引いているので、対数変換または Cornfield (1956) の方法によって正規分布に近づけ、正規近似を使って 95% 信頼区間を求めることになる。対数変換の場合、95% 信頼区間は、

$$OR \cdot \exp(-\text{qnorm}(0.975) \sqrt{1/a + 1/b + 1/c + 1/d}) \quad (\text{下限})$$

$$OR \cdot \exp(\text{qnorm}(0.975) \sqrt{1/a + 1/b + 1/c + 1/d}) \quad (\text{上限})$$

となる。Cornfield の方法の方が大きなオッズ比については近似がよいが、手順がやや複雑であるため、ここでは扱わない。現在では Exact 法を用いることが推奨されているので、基本的に `fisher.test()` の結果を採用すればよい。R のコードは以下の通り。 p 値は帰無仮説 $OR = 1$ の検定の有意確率である。

c10-5.R

```
oddsratio2 <- function(a,b,c,d) {
  data <- matrix(c(a,b,a+b,c,d,c+d,a+c,b+d,a+b+c+d),nr=3)
  colnames(data) <- c("疾病あり","疾病なし","合計")
  rownames(data) <- c("曝露群","対照群","合計")
  print(data)
  OR <- (a*d)/(b*c)
  N1 <- a+c; M1 <- a+b; N0 <- b+d; M0 <- c+d; T <- a+b+c+d
  p.v <- 2*(1-pnorm(abs((a-N1*M1/T)/sqrt(N1*N0*M1*M0/T/T/(T-1))))))
  ORL <- OR*exp(-qnorm(0.975)*sqrt(1/a+1/b+1/c+1/d))
  ORU <- OR*exp(qnorm(0.975)*sqrt(1/a+1/b+1/c+1/d))
  cat("オッズ比の点推定量:",OR," (p=",p.v,
      ") 95% 信頼区間 = [",ORL,"",",",ORU,""]\n")
}
```

```
oddsratio2(4,2,99996,99998)
```

なお、サンプルサイズが大きいときは、R の `fisher.test()` 関数や、それを内部的に利用している `epitools` ライブラリの `oddsratio.fisher()` 関数^{*22}では `Out of workspace` というエラーを起こして計算できないことがあるが、`vcd` ライブラリの `oddsratio()` 関数は計算方法が異なるため実行できる。ただし、`fisher.test()` 関数の `Out of workspace` エラーはデフォルトで 20 万バイト確保されている計算用メモリでは不足したというエラーなので、呼び出すときに大きめの `workspace` を確保すれば回避可能である。これらを使って計算するためのコードを次の枠内に示す。

c10-6.R

```
X <- matrix(c(4,2,99996,99998),nr=2)
fisher.test(X, workspace=1000000)
require(epitools)
oddsratio(c(4,99996,2,99998),method="fisher")
detach(package:epitools)
require(vcd)
OR <- oddsratio(X,log=F)
ORL <- summary(oddsratio(X))
ORCI <- confint(OR)
M <- c("オッズ比の点推定量"," (p=,") 95% 信頼区間 = [ ", " , " ]\n")
cat(M[1],OR,M[2],ORL[1,4],M[3],ORCI[1],M[4],ORCI[2],M[5])
```

結果を次の表にまとめて示す。

方法	点推定量	有意確率	95% 信頼区間	
			下限	上限
上で定義した <code>oddsratio2()</code>	2.00004	0.4142	0.366	10.9
<code>fisher.test()</code>	2.000022	0.6875	0.2866	22.11
<code>epitools</code> の <code>oddsratio.fisher()</code>	2.000022	0.6875 (midp 0.453)	0.2866	22.11
<code>vcd</code> の <code>oddsratio()</code>	2.00004	0.1898	0.4262	9.386

^{*22} `oddsratio(...,method="fisher")` でも呼び出される。`epitools` ライブラリの `oddsratio()` 関数には `midp`, `fisher`, `wald`, `small` という 4 種類の `method` があり、それぞれ別々の関数を内部的に呼び出している (`S3method` の継承ではない)。`riskratio()` 関数とも `rateratio()` 関数とも引数を与える順序が異なるので注意されたい。この辺り、`epitools` ライブラリは若干思想が良くないと思う。

10.3.5 関連性の指標

2つのカテゴリ変数によってクロス集計表を作る目的としては、曝露の有無が疾病の有無に与える効果を評価する他に、2つのカテゴリ変数の関連の程度を見たい場合もある。関連性の指標としては、ユール (Yule) の Q 、ファイ係数、ピアソンのコンティンジェンシー係数、クラメールの V が良く用いられる。

ユールの Q オッズ比を -1 から 1 の値を取るようにスケーリングしたもの。 $Q = (OR - 1)/(OR + 1)$ 。独立な場合は 0 となる。

ファイ係数 (ϕ) 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である。 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、 $\phi = \sqrt{(\pi_1 - \pi_2)(\theta_1 - \theta_2)}$ 。この値は 2×2 に限らず、一般の $k \times m$ の分割表について計算でき、ピアソンのカイ二乗統計量 χ_0^2 と総人数 n を用いて、 $\sqrt{\chi_0^2/n}$ と定義される。 k と m のどちらか小さな方の値が t だとすると、ファイ係数は 0 から $\sqrt{t-1}$ の範囲をとる。

ピアソンのコンティンジェンシー係数 C ファイ係数はカテゴリ数の影響を受けるので、それを除去したものである。ファイ係数を用いて、 $C = \sqrt{\phi^2/(1+\phi^2)}$ として計算される。取りうる値の範囲は 0 から $\sqrt{(t-1)/t}$ である。

クラメールの V ファイ係数を用いて、 $V = \phi/\sqrt{t-1}$ と表せる。取りうる値の範囲は 0 から 1 となり、変数のカテゴリ数によらないのが利点である。

なお、ファイ係数、ピアソンのコンティンジェンシー係数、クラメールの V (これらは総称して属性相関係数と呼ばれることがある) は `vcd` ライブラリの `assocstats()` 関数で計算できる。この関数は、これらの係数の他、「関連がない」を帰無仮説とする検定を実行して、ピアソンのカイ二乗統計量と尤度比カイ二乗統計量 (ここでは説明しないが、多くの場合にピアソンのカイ二乗統計量を使った通常のカイ二乗検定よりもよいとされる) を計算してくれる。それらの有意確率も計算してくれる。属性相関係数はすべてピアソンのカイ二乗統計量に基づいて計算されるので、その有意性検定はカイ二乗検定の結果と等価と考えてよい。上記白血病のコホート研究の例でこれらを計算するには次の枠内を打てばよいが、これらの係数の値はすべて 0.002 となるので、データからはほとんど関連を見出せないといえる。

```
require(vcd)
assocstats(matrix(c(4,2,99996,99998),nr=2))
```

10.3.6 一致度の指標 κ 係数

2 回の繰り返し調査をしたり、同じ対象を 2 人の評価者が別々に評価したときに、あるカテゴリ変数がどれくらい一致するかを見るには、クロス集計表という形は同じでも、効果や関連性を見るのではなく、「偶然ではありえないくらい一致しているかどうか」を評価しなくてはならない。2 回の繰り返し調査の場合、test-retest reliability (検査再検査信頼性) の指標といえ、同じ対象を 2 人の評価者が評価する場合は inter-rater agreement (評価者間一致度) の指標といえる。そのような一致度の指標として、もっとも有名なものが κ 係数である。カテゴリ変数間の一致度をみるための作図には、vcd ライブラリに含まれている agreementplot() という関数がある。

	2 回目	2 回目 ×	合計
1 回目	a	b	m_1
1 回目 ×	c	d	m_2
合計	n_1	n_2	N

という表から、偶然でもこれくらいは一致するだろうと思われる値は、1 回目と 2 回目の間に関連がない場合の各セルの期待値を足して全数で割った値になるので $P_e = (n_1 \cdot m_1 / N + n_2 \cdot m_2 / N) / N$ 、実際の一致割合 (1 回目も 2 回目も a , 1 回目も 2 回目も b であった割合) は $P_o = (a + d) / N$ とわかる。ここで $\kappa = (P_o - P_e) / (1 - P_e)$ と定義すると、 κ は、完全一致のとき 1、偶然と同じとき 0、それ以下で負となる統計量となる。

κ の分散 $V(\kappa)$ は、 $V(\kappa) = P_e / (N \cdot (1 - P_e))$ となるので、 $\kappa / \sqrt{V(\kappa)}$ が標準正規分布に従うことを利用して、帰無仮説「 $\kappa = 0$ 」を検定したり、 κ の 95% 信頼区間を求めたりすることができる。次の枠内は、2 × 2 クロス集計表を与えたときに、 κ の点推定量と 95% 信頼区間と有意確率を計算する R の関数 kappa.test() を定義してから、 x で回答する項目について 2 回の繰り返し調査をしたときに、1 回目も 2 回目も a であった人数が 10 人、1 回目 a で 2 回目 b であった人数が 2 人、1 回目 b で 2 回目 a であった人数が 3 人、1 回目も 2 回目も b であった人数が 19 人であったときにその計算を実行させるコードである。

```

c10-7.R
kappa.test <- function(x) {
  x <- as.matrix(x)
  a <- x[1,1]; b <- x[1,2]; c <- x[2,1]; d <- x[2,2]
  m1 <- a+b; m2 <- c+d; n1 <- a+c; n2 <- b+d; N <- sum(x)
  Pe <- (n1*m1/N+n2*m2/N)/N
  Po <- (a+d)/N
  kappa <- (Po-Pe)/(1-Pe)
  seK0 <- sqrt(Pe/(N*(1-Pe)))
  seK <- sqrt(Po*(1-Po)/(N*(1-Pe)^2))
  p.value <- 1-pnorm(kappa/seK0)
  kappaL<-kappa-qnorm(0.975)*seK
  kappaU<-kappa+qnorm(0.975)*seK
  list(kappa=kappa,conf.int=c(kappaL,kappaU),p.value=p.value)
}
kappa.test(matrix(c(10,3,2,19),nr=2))

```

vcd ライブラリの `Kappa()` 関数は $m \times m$ のクロス集計表について、重みなしと重みつきで κ 係数を計算してくれる^{*23}。結果を `confint()` 関数に渡せば信頼区間も推定できる。同じデータに適用するには、次の枠内を打つ。先の枠内を実行した時と同じ結果が得られる^{*24}。

```

require(vcd)
print(myKappa <- Kappa(matrix(c(10,3,2,19),nr=2)))
confint(myKappa)

```

^{*23} 重みは、`Po` や `Pe` を計算する際に `weights=オプション` を指定しないとき、あるいは `weights="Equal-Spacing"` にマッチしない任意の文字を指定した場合は、`weights="Fleiss-Cohen"` と指定したのと同じで、カテゴリ数が `nc` として $1 - (\text{abs}(\text{outer}(1:nc, 1:nc, "-")) / (nc-1))^2$ となり、`weights="Equal-Spacing"` を指定したときは $1 - \text{abs}(\text{outer}(1:nc, 1:nc, "-")) / (nc-1)$ が重みとなる。つまり、 \times の一致をみるならカテゴリ数は 2 なので、重みはどちらの方法でも `matrix(c(1,0,0,1),nc=2)` となる。

^{*24} 有意確率は出ないが、 κ 係数は、有意性の検定をするよりも、95% 信頼区間を示すことと、目安としての一致度の判定基準（負だと poor な一致、0-0.2 で slight な一致、0.21-0.4 で fair な一致、0.41-0.6 で moderate な一致、0.61-0.8 で substantial な一致、0.81-0.99 で almost perfect な一致、1 で perfect な一致とする、Landis and Koch, 1977, Biometrics, 33: 159-174 など）を参照して一致度を判定するという使い方が普通らしく、vcd ライブラリでもそのような実装がされているのだと思われる。考えてみれば、一致度を評価する上で $\kappa = 0$ という帰無仮説の検定には意味が乏しいのは当然であろう。

10.4 スクリーニングにおける ROC 分析

10.4.1 ROC 分析とは

ROC 曲線とは、Receiver Operating Characteristic 曲線の略である^{*25}。集団を対象に、すばやく実施可能な方法で、疾病を暫定的に識別することをスクリーニングというが、いくつかのスクリーニング方法があるときに、それらの相対的な有効性を視覚的に判定する基準の一つが ROC 曲線である。

1 つのスクリーニング方法について陽性・陰性の基準値を最小値から最大値まで段階的に変えると、偽陽性率（本当は病気ではない人を誤って陽性と判定してしまう割合、つまり $1 - \text{特異度}$ と一致する）も感度（病気の人を正しく陽性と判定する割合）も 0 から 1 まで変わるので、偽陽性率を横軸に、感度を縦軸にとって線で結ぶと、基準値の変化に対応する曲線を引くことができる。この曲線ができるだけ左上を通る方がスクリーニングとしての有効性は高い方法だといえる。

また、この曲線のもっとも左上の点（理想は偽陽性率 0 で感度 1 だが、現実にはそうなることはまずない）を与える基準値が、陽性・陰性を分けるカットオフポイントとしてももっとも有効性が高いと判断される。

つまり、ROC 曲線は、ある検査値について適切なカットオフポイントを検索するのにも使えるし、複数のスクリーニング方法の優劣を比較することにも使える。

ROC 曲線を描いて視覚的評価をするだけではなく、AUC（Area under curve; 曲線下面積）を計算する、あるいは右下の点からもっとも離れた点を与えるカットオフポイントを最適値とするなどの計算も含めて、ROC 分析と呼ぶ。複数のスクリーニング方法の AUC を比較し、もっとも大きい AUC を与える方法がもっとも優れていると考えるのが普通である。ただし、感度や特異度がもっとも優れていても、他のもっとも廉価に大勢を検査できる方法と大差なければ、高価だったり時間や手間がかか

^{*25} 日本語では、受診者動作特性曲線という訳語がついている教科書と、受信者動作特性曲線という訳語がついている教科書が並立しているが、ROC が何の略であるかを明示して「ROC 曲線」だけを掲載している本も増えてきたので、ここでも敢えて訳さないことにする。手元にある本で調べると、日本疫学会（編）「疫学 基礎から学ぶために」南江堂、能登洋「日常診療にすぐ使える臨床統計学」羊土社などが「受診者」派で、鈴木・久道（編）「シンプル衛生公衆衛生学 2006」南江堂、日本疫学会（訳）「疫学辞典 第 3 版」日本公衆衛生協会、フレッチャー RH、フレッチャー SW、ワグナー EH、福井次矢（監訳）「臨床疫学」メディカルサイエンスインターナショナルなどが「受信者」派であった。稲葉・野崎（編）「新簡明衛生公衆衛生 改訂 4 版」南山堂、丹後俊郎「メタ・アナリシス入門」朝倉書店などは、「ROC 曲線」だけを掲載していた。

る（倫理面も含めて）などの理由で採用されない場合もある。

10.4.2 計算手順を考える

具体例で考えよう。以下のデータ（架空である）が得られたとする。

対象者	質問紙得点	臨床診断
1	20	うつ
5	22	うつ
6	28	うつ
2	13	健康
3	19	健康
4	21	健康
7	11	健康
8	25	健康
9	16	健康
10	19	健康

この質問紙得点が、あるカットオフポイントより高いことを、うつのスクリーニングとして使おうというのが、このデータを得た目的であるとする、問題は、適切なカットオフポイントを見つけることになる。

例えば、カットオフポイントを 18、すなわち、質問紙得点が 18 点以上なら陽性、そうでないなら陰性と判定することになると、以下のクロス集計表ができる。

	うつ	健康
陽性	3	4
陰性	0	3

このとき、感度は $3/(3+0) = 1$ 、特異度は $3/(4+3) = 0.429$ 、偽陽性率は $4/(4+3) = 1 - 0.429 = 0.571$ となる。得点の最小値から最大値 +1 までカットオフポイントをずらしていくと、感度も偽陽性率も 1 から 0 まで変化するので、これをグラフに描けば ROC 曲線となる^{*26}。これを MS-Excel のような表計算ソフトで計算しようとする、AUC を計算したり最適カットオフポイントを見つけることは容易ではないし、いちいち多くのセルを使って計算式を入力するのが甚だ面倒である。

^{*26} 丹後俊郎『メタ・アナリシス入門』朝倉書店に紹介されているように、クロス集計表のどこかのセルが 0 になる場合は各セルに 0.5 ずつ加える Woolf(1955) の修正を薦める教科書もあるが、その場合曲線の端点が (0,0) と (1,1) にならないので、本書では修正しない。

ROC 曲線の変曲点はデータ点であることを考慮し、R で素直に式を書けば、次の枠内のような関数 `roc` を定義することができる。この関数はカットオフポイントをずらしていったときの、感度、偽陽性率、「感度 0、偽陽性率 1」の点からの距離、区間ごとの曲線下面積 (AUC) をリストとして返す*²⁷。

```
roc <- function(values,iscase) {
  cutoffs <- unique(sort(values))
  cutoffs <- c(cutoffs,max(values)+1)
  ns <- length(cutoffs)
  sensitivity <- rep(0,ns)
  falsepositive <- rep(0,ns)
  dist <- rep(0,ns)
  aucp <- rep(0,ns)
  D <- sum(iscase==1)
  H <- sum(iscase==0)
  for (i in 1:ns) {
    cutoff <- cutoffs[i]
    positives <- ifelse(values >= cutoff,1,0)
    PinD <- sum(positives==1 & iscase==1)
    NinH <- sum(positives==0 & iscase==0)
    sensitivity[i] <- PinD/D
    falsepositive[i] <- 1-NinH/H
    dist[i] <- sqrt((PinD/D)^2+(NinH/H)^2)
    aucp[i] <- ifelse(i==1,(1-falsepositive[i])*sensitivity[i],
                      (falsepositive[i-1]-falsepositive[i])*sensitivity[i])
  }
  list(cutoffs,sensitivity,falsepositive,dist,aucp)
}
```

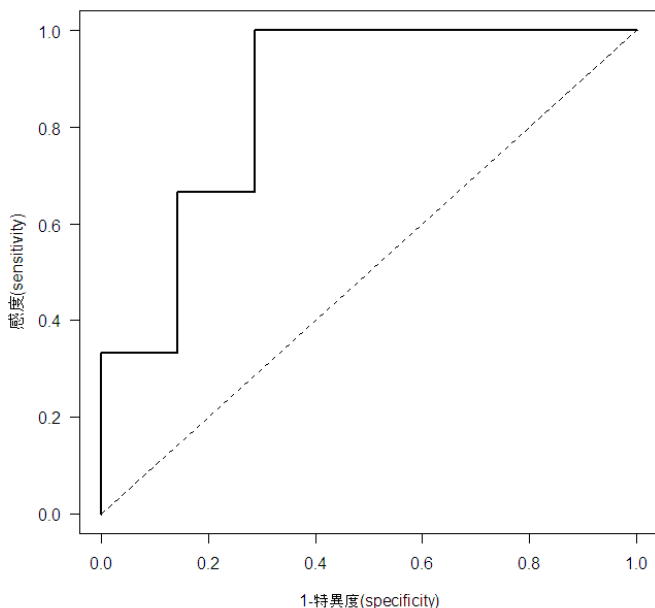
結果として得たリストの値を使って ROC 曲線を描き、最適カットオフポイントを求めるには次のようにラッパー関数を定義すると便利である。

*²⁷ なお、曲線下面積は、このように図からそのまま単純に点推定値を求める方法だけではなく、最尤法でバイアスを補正した計算法も提案されているし、ジャックナイフなどで分散を計算することもできるが、本書では扱わない。詳細は、Zhou XH (2003) Evaluation of diagnostic test's accuracy in the presence of verification bias. In: Lu Y and Fang JQ [Ed.] *Advanced Medical Statistics*, World Scientific Publishing Co. Pte. Ltd.などを参照されたい。

```
rocc <- function(...) {
  res <- roc(...)
  cat("cutoff\ttsensitivity\t1-specificity\tdistance\n",
      sprintf("%5.3f\t%5.3f\t%5.3f\t%5.3f\n",
              res[[1]],res[[2]],res[[3]],res[[4]]))
  mlcs <- "最適カットオフポイント:%5.3f, 感度:%5.3f, "
  mlcs <- paste(mlcs,"特異度 %5.3f\nAUC=%5.3f\n",sep="")
  mlcc <- which.max(res[[4]])
  cat(sprintf(mlcs,res[[1]][mlcc],res[[2]][mlcc],1-res[[3]][mlcc],
              sum(res[[5]])))
  plot(res[[3]],res[[2]],type="l",lwd=2,xlab="1-特異度 (specificity)",
        ylab="感度 (sensitivity)")
  lines(c(0,1),c(0,1),lwd=1,ltty=2)
}
```

最初に示したデータを使って計算するには次の枠内を打つ。最初の2行はデータ入力だが、表計算ソフト上でセルを選んでコピーし、`x <- scan("clipboard")`のようにしても入力可能である。このように関数定義をすれば、実行は `rocc(x,y)` だけで済む。最適カットオフポイント 20、そのときの感度が 1 で特異度が 0.714、曲線下面積が 0.857 とわかる。

```
x <- c(20,22,28,13,19,21,11,25,16,19)
y <- c(rep(1,3),rep(0,7))
rocc(x,y)
```



上記の方法は計算過程を完全に把握できるところはいいが、今ひとつ美しくない。そこで、CRAN から 1 つ、Epi というライブラリをインストールする。インターネットに接続されたコンピュータであれば R のコンソールで `install.packages("Epi",dep=T)` とすればよい。もしレポジトリあるいはミラーを選ぶようにという選択肢がでてきたら、Japan(Tsukuba) または Japan(Tokyo) を選ばせよ。

10.4.3 Epi ライブラリを使う方法

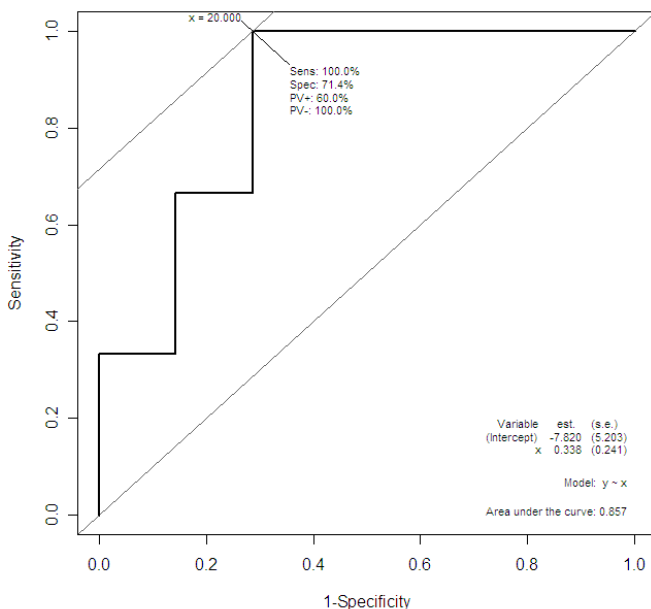
Epi ライブラリは、デンマーク・コペンハーゲン大学の Bendix Carstensen らが開発して CRAN で公開している、慢性疾患の疫学のためのライブラリである^{*28}。ROC の他、age-period-cohort モデルや Lexis diagram を描く関数も含まれている。

Epi ライブラリを使った実行方法は非常に簡単で、次の枠内を打つだけでいい。計

^{*28} 詳しくは <http://staff.pubhealth.ku.dk/~bxc/Epi/> を参照。

算結果も図内にすべて示される。

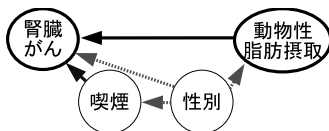
```
require(Epi)
ROC(x,y,plot="ROC")
```



10.5 交絡を考える

量的変数の相関にも見かけの相関や擬似相関があったように、クロス集計でも見かけの関連が出てしまうことがある。大きな原因は、交絡因子があることである。一般に、2つの変数の関係を調べているとき、その両方に影響を与えている第3の変数があれば、それは交絡因子（交絡変数）である。次の図の例では、動物性脂肪摂取と腎臓がん罹患の関係性を調べているのだが、性別が動物性脂肪摂取と腎臓がん罹患の両方に影響しているため、交絡を起こす可能性がある。喫煙は腎臓がん罹患には影響するが、動物性脂肪摂取と関係するとは思われないため、このフレームでは交絡

ではない。



Rothman(2002)*²⁹によると、交絡の条件は3つあって、

1. 交絡因子は疾病と関連してはならない
2. 交絡因子は曝露と関連してはならない
3. 交絡因子は曝露の効果であってはならない

が満たされていないとまとめられている。

10.5.1 シンプソンのパラドックス

第3の変数による交絡があるために、真の関係と見かけの関係が異なる事例としてもっとも有名なものの1つが、シンプソンのパラドックスである。

c10-8.R

```

# 出典: Simpson EH (1951) The interpretation of interaction in
# contingency tables. J. Royal Stat. Soc. Ser. B, 13: 238-241.
males <- matrix(c(4,3,8,5),nr=2)
dimnames(males) <- list(c("生存","死亡"),c("処置なし","処置あり"))
females <- matrix(c(2,3,12,15),nr=2)
dimnames(females) <- list(c("生存","死亡"),c("処置なし","処置あり"))
total <- males+females
require(vcd)
prop.table(males,2)
summary(oddsratio(males))
prop.table(females,2)
summary(oddsratio(females))
prop.table(total,2)
summary(oddsratio(total))
  
```

上枠に示した Simpson の論文に載っている例では、男女別にみれば処置ありの方が生存割合が高いのに、男女をプールしてみると処置による生存割合の差が消失し

*²⁹ Rothman, KJ (2002) *Epidemiology: An Introduction*. Oxford Univ. Press. (矢野栄二, 橋本英樹監訳 (2004) 「ロスマンの疫学: 科学的思考への誘い」, 篠原出版社, として邦訳がでている)

ている。もっとも、消失してしまうといっても、サンプルサイズが小さいこともあって、統計的に有意とはいえない。しかし、現実にもこういうことは頻繁にあって、雑な解析では真の関連を見誤ってしまう危険がある。

例えば、次の枠内は、スティーヴン・セン（松浦俊輔訳）『確率と統計のパラドックス』青土社の p.39 に掲載されている例を分析するプログラムだが^{*30}、年齢をプールすると糖尿病の型と死亡率は独立でない ($p = 0.001$) のに、40 歳以上と 40 歳未満で区切って層別に解析するとどちらの層でも独立性の帰無仮説は棄却されない（それぞれ $p = 0.27$, $p = 1$ ）。しかも、40 歳以上でも 40 歳未満でも IDDM 群の死亡率の方が NIDDM 群の死亡率より高いのに（40 歳以上では IDDM 群 0.46 に対して NIDDM 群 0.41, 40 歳未満では IDDM 群 0.008 に対して NIDDM 群 0）、年齢をプールすると IDDM 群の死亡率（0.29）よりも NIDDM 群の死亡率（0.40）の方が高くなる。これは年齢が交絡しているために、本来はない見かけ上の関連が見えてしまったことを意味する。40 歳未満群の大半が IDDM であって、かつ 40 歳未満群の死亡率が 40 歳以上群の死亡率より遥かに低いために、こうなったのである。

c10-9.R(1)

```
over40 <- matrix(c(311,218,124,104),nc=2)
under40 <- matrix(c(15,0,129,1),nc=2)
dimnames(over40) <- list(c("生存","死亡"),c("NIDDM","IDDM"))
dimnames(under40) <- list(c("生存","死亡"),c("NIDDM","IDDM"))
print(over40)
fisher.test(over40)
print(under40)
fisher.test(under40)
total <- under40+over40
print(total)
fisher.test(total)
```

10.5.2 交絡を制御するには

このような交絡が起こらないようにするには、もちろん、臨床試験を実施する場合のように研究をデザインできる状況であれば、無作為割付を行ってフォローアップ

^{*30} 「生存」は、censored なので観察終了時までにはイベントが起こっていないことを意味し、通常「打ち切り」と訳されるが、この本の訳文は「調査中」となっていて、大胆に意味を酌んでいいなら「生存」と訳してしまってもいいだろうと判断した。

し、率比を分析するなど、デザイン上で交絡を防ぐ工夫をすればよいし、それが王道である。

しかし、実際問題として、研究は実験ばかりではないし、地域調査において完全に交絡を防ぐデザインをすることは、ほぼ不可能である。交絡を制御して真の関連を検討するには、大別して 2 つのアプローチがある。

1 つは、交絡変数も原因となる変数とともに独立変数として投入し、それらの交互作用も考えながら、結果となる変数（多くは疾病発生）を従属変数として説明するようなモデルの当てはめを行う方法である。ロジスティック回帰分析を含むこの方法は、一般化線型モデルというフレームで扱えるので第 12 章で説明する。

もう 1 つは、交絡因子によって層別解析を行うか、または限定を行うことである。層別解析とは、交絡因子のカテゴリによってデータを分割し、それぞれ別々の層として分析を進めることをさす。層別解析をした上で、どの層でも同じ向きに関連がありそうなら、たんにプールするのではなくて、「どの層でも同じ向きに関連がある」を対立仮説として、クロス集計表を併合した分析を行う。具体的な方法としては、マンテル=ヘンツェルの要約カイ二乗検定とか、共通オッズ比といったものがある*³¹。R では `mantelhaen.test()` 関数により、マンテル=ヘンツェルの要約カイ二乗統計量とその検定、さらに各層が 2×2 分割表のときは共通オッズ比とその 95% 信頼区間を計算することができる。ここで得られる共通オッズ比は、層の違いを調整した関連の強さを示す指標となる。

ただし、マンテル=ヘンツェルの要約カイ二乗検定は層別変数との交互作用が存在しないこと（言い換えると、クロス表を作っている変数間の関連がどの層でも同じということ）を前提として行うものなので、それに先立って Woolf の検定（経験的ロジスティック変換を用いて、帰無仮説「どの層でも変数間の関連が共通」を検定する）によってそれを確認しておくべきとされる*³²。Woolf の検定は `vcd` ライブラリの `woolf_test()` 関数で可能である。グラフ表示は `vcd` ライブラリの `fourfold()` 関数を用いるとよい。拡張モザイクプロットも `vcd` ライブラリの `mosaic()` 関数でできるが、引数を与える順序を変えねばならず、かつあまり見やすすくないので、個人的にはお勧めしない。

*³¹ マンテル=ヘンツェルの方法による複数の層の関連の指標の併合は、オッズ比だけでなく、リスク差やユールの Q やファイ係数についても可能である。文献：佐藤俊哉，前田和甫（1987）疫学研究から得られる層別データの要約。日本公衆衛生学雑誌，34(5): 255-260。

*³² なお、向きは異なるかもしれないがともかく何らかの関連があるかどうかを調べたい場合は、自由度 1 のカイ二乗分布する変数 k 個の和が自由度 k のカイ二乗分布に従うことを使って、各層で得られたカイ二乗統計量の総和を出してやれば検定できる。

先に作った2つの 2×2 クロス集計表 `under40` と `over40` は、次の枠内のコードのようにすれば3次元のクロス集計表 `x` にすることができる。

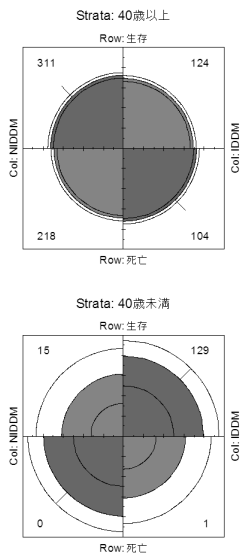
c10-9.R(2)

```
x <- array(c(over40,under40),dim=c(2,2,2))
dimnames(x) <- list(c("生存","死亡"),c("NIDDM","IDDM"),
  c("40歳以上","40歳未満"))
print(x)
```

または、いきなり数値を `array()` 関数に渡して次の枠内のように定義してもよい。Woolfの検定で有意でないので層別変数との交互作用はなく、マンテル=ヘンツェルの要約カイ二乗検定でも有意でないので、どの層でも同じ向きに関連があるとはいえないことがわかる。なお、`fourfold()` は `grid` グラフィックスを使うので `layout()` することができないようだ。

c10-10.R

```
x <- array(c(311,218,124,104,15,0,129,1),dim=c(2,2,2))
dimnames(x) <- list(c("生存","死亡"),c("NIDDM","IDDM"),
  c("40歳以上","40歳未満"))
require(vcd)
woolf_test(x)
mantelhaen.test(x)
y <- array(c(x[1,1,],x[2,1,],x[1,2,],x[2,2,]),dim=c(2,2,2))
dimnames(y) <- list(c("40歳以上","40歳未満"),c("生存","死亡"),
  c("NIDDM","IDDM"))
structable(y,split_vertical=T)
fourfold(x)
mosaic(y)
```



10.6 課題

アルコール摂取と食道がんの症例対照研究を実施したとする。当初、食道がん患者 180 人と対照 575 人のサンプリングを行ったが、年齢が交絡している可能性を考え、年齢群別に集計したところ、55 歳未満では、患者群 46 人のうちアルコール多量摂取者が 30 人、対照群 372 人のうちアルコール多量摂取者が 64 人で、55 歳以上では、患者群 134 人のうちアルコール多量摂取者が 66 人、対照群 203 人のうちアルコール多量摂取者が 45 人だったとする。

このデータから、アルコール摂取と食道がんには関連があるか、あるとしたらどの程度の関連か評価せよ。まず年齢層別に 2 つのクロス集計表を作り、別々に分析してから、その結果から判断して、必要な場合には、どの層でも共通した関連があるかどうか、あるとすればどの程度の関連か評価せよ。

第 11 章

量的データのノンパラメトリックな分析

11.1 2 群の分布の位置の差に関するノンパラメトリックな検定

11.1.1 ノンパラメトリックな検定とは？

パラメータ (parameter) とは母数という意味である。これまで説明してきた検定法の多くは、母数、つまり母集団の分布に関する何らかの仮定をおいていた。その意味で、 t 検定も F 検定もパラメトリックな分析法といえる。パラメトリックな分析法は、原則としては、その分析法が前提とする、母数に関する仮定を満たすようなデータに対して使うべきである。一方、フィッシャーの正確な確率は母数を仮定しないのでパラメトリックでない。ノンパラメトリックな分析とは、パラメトリックでない分析、つまり母数を仮定しない分析をさし、分布がひどく歪んでいたり、上限や下限があったりするようなデータに対しても使うことができるという利点がある*¹反面、理

*¹ 正規分布に従わないデータに出会ったときは、ノンパラメトリックな検定を行う他に、(1) 対数変換などによりデータの分布を正規分布に近づけるか、(2) 外れ値を吟味して、データそのものにエラーがありそうか、あるいは本質的に異なる母集団からの標本が混ざっていると判断されるなら、それを分析から除外することによって正規分布に近づける、といった方法もあるが、これらの操作が逆にデータを歪めてしまう危険もあるので、データの性状を細かくみて、他の知見とも考え合わせ、慎重に行わねばならない。また、論文や報告書にまとめる際にも、これらの操作をしたならしたと明記せねばならない。

想的な場合のパラメトリックな分析に比べると検出力は高くない*2。

2 群の分布の位置の差に関して問題を定式化すると、次のようになる。

1. 標本データ X_1, X_2, \dots, X_n が互いに独立に分布 $F(x)$ に従い、別の標本データ Y_1, Y_2, \dots, Y_n が互いに独立かつ X とも独立で分布 $G(y)$ に従う。
2. F と G には連続分布であるという以外には制約をおかない。
3. このとき、「2 つの分布に差はない」という帰無仮説 ($H_0 : F(x) \equiv G(x)$) を検定する。

つまり、2 群の分布の位置の差についてのノンパラメトリックな検定では、「母数を仮定しない」とは言っても、連続分布であることだけは仮定する。もっとも理想的には分布の形が同じで位置だけがずれているという、「ズレのモデル」が仮定できると話は簡単である。具体的な方法としては、Wilcoxon の順位和検定、符号付順位和検定、符号検定などがある。得られたデータがある種の経験分布関数に一致するかどうかを調べるために良く使われる検定法としてはコルモゴロフ = スミルノフ検定 (KS 検定) がある*3。

11.1.2 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。図示は層別箱ヒゲ図を用いるのが普通である。

データがもつ情報の中で、単調変換に対して頑健なのは順位なので、これを使って検定しようという発想である。以下、Wilcoxon の順位和検定の手順を箇条書きする。

1. 変数 X で表される第 1 群のデータを x_1, x_2, \dots, x_m とし、変数 Y で表される

*2 検定法によってはそれほど低くもない。例えば後述する Wilcoxon の順位和検定は、位置母数だけが異なる 2 つの正規分布の差を検出する効率が t 検定の約 95% である。なお、厳密に考えると区分はそれほど明確ではない。例えば、カイ二乗検定では母集団の分布には特定の仮定は置いていないので、定義からすると、実はノンパラメトリックな分析になる。ただし、カイ二乗統計量がカイ二乗分布に従うためにはデータ数が十分に多いことが必要である。もっとも、そう言ってしまうと正規近似する場合の順位和検定もデータ数が多くなることが必要なので、問題は何を検定の本質と見なすかという話になってくる。一般には、量的変数を分析するのに、量の情報を使わずに大小関係、即ち順位の情報だけを使う分析をノンパラメトリックな解析と呼ぶと考えておけば大過ない。なお、相関については、既にスピアマンとケンドールの順位相関係数を説明済みである。

*3 R では `ks.test(変数 1, 変数 2)` で実行可能である。

第2群のデータを y_1, y_2, \dots, y_n とする。

2. まず、これらをませこぜにして小さい方から順に番号をつける*4。例えば、 $x_8[1], y_2[2], y_{17}[3], \dots, x_4[N]$ のようになる (ただし $N = m + n$)
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 ($i = 1, 2, \dots, m$) 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 H_0 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?*5

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、 $1, 2, 3, \dots, N$ 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、 ${}_N C_m$ 通りある*6。
6. $X > Y$ の場合には、 ${}_N C_m$ 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. $k/({}_N C_m)$ が有意水準 α より小さいときに H_0 を疑う。 N が小さいときは有意になりにくいのが、 N が大きすぎると計算が大変面倒である*7。そこで、正規

*4 同順位がある場合の扱いは後述する。

5 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、数学的に等価な検定である。R では、Wilcoxon の順位和統計量の分布関数が提供されているので、例えばここで得られた順位和を RS と書くことにすると、 $2(1-pwilcox(RS, m, n))$ で両側検定の正確な有意確率が得られる。

*6 R では `choose(N, m)` によって得られる。

*7 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、総当た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

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

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

（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，

$$\text{var}(R) = E(R^2) - (E(R))^2$$

から，

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

となるので*8，

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

と

$$\begin{aligned} E(R_i R_j) &= \frac{1}{N(N-1)} \left\{ \left(\sum_{k=1}^N k \right)^2 - \sum_{k=1}^N k \right\} \\ &= \frac{1}{N(N-1)} \left(\frac{N^2(N+1)^2}{4} - \frac{N(N+1)(2N+1)}{6} \right) \\ &= \frac{(N+1)(3N+2)}{12} \end{aligned}$$

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

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。

*8 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，

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

となる。

9. 標準化^{*9}して連続修正^{*10}し、 $z_0 = \{|R_X - E(R_X)| - 1/2\} / \sqrt{\text{var}(R_X)}$ 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、 $2 * (1 - \text{pnorm}(z0, 0, 1))$ とすればよい。
10. ただし、同順位があった場合は、ステップ 2 の「小さい方から順に番号をつける」ところで困ってしまう。例えば、変数 X が $\{2, 6, 3, 5\}$ 、変数 Y が $\{4, 7, 3, 1\}$ であるような場合には、 X にも Y にも 3 という値が含まれる。こういう場合は、下表のように平均順位を両方に与えることで、とりあえず解決
- | 属する変数 | Y | X | X | Y | Y | X | X | Y | |
|-------|---|---|---|-----|-----|---|---|---|---|
| できる。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^{*11}。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

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

となる^{*12}。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いくつか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）。

^{*9} 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。

^{*10} これも何度も出てくるが、連続分布に近づけるために $1/2$ を加減する操作である。

^{*11} 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 `wilcox.test()` では正確な確率は求められない。`exactRankTests` ライブラリの `wilcox.exact()` 関数は、同順位のデータがあっても正確な確率を求めてくれるので、そちらを使った方がいいかもしれない。

^{*12}

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

とも書ける。

例題 1

以下は、外来と入院患者に対して、それぞれ 3 ヶ月治療したときの胃潰瘍の面積変化割合の表である^a。

	人数	面積変化割合（肩数字はその割合を示した人数を意味する）
X：入院患者	32	-100 ⁽¹²⁾ , -93, -92, -91 ⁽²⁾ , -90, -85, -83, -81, -80, -78, -46, -40, -34, 0, 29, 62, 75, 106, 147, 1321
Y：外来患者	32	-100 ⁽⁵⁾ , -93, -89, -80, -78, -75, -74, -72, -71, -66, -59, -41, -30, -29, -26, -20, -15, 20, 25, 37, 55, 68, 73, 75, 145, 146, 220, 1044

入院患者と外来患者の間で、面積変化割合に差があるといえるだろうか？ Wilcoxon の順位和検定を実行してみよう。

^a 出典：Doll R, Pygott F: Factors influencing the rate of healing of gastric ulcers: admission to hospital, phenobarbitone, and ascorbic acid. *Lancet*, i: 171-175, 1952. から Armitage P, Berry G, Matthews JNS: *Statistical Methods in Medical Research* 4th ed. Blackwell Publishing, 2002, pp.281. に引用されているもの

まず、2 群を併せて順位を計算する。

```
c11-1.R(1)
```

```
X <- c(rep(-100,12), -93, -92, rep(-91,2), -90, -85, -83, -81,
      -80, -78, -46, -40, -34, 0, 29, 62, 75, 106, 147, 1321)
Y <- c(rep(-100,5), -93, -89, -80, -78, -75, -74, -72, -71,
      -66, -59, -41, -30, -29, -26, -20, -15, 20, 25, 37, 55, 68,
      73, 75, 145, 146, 220, 1044)
d <- data.frame(gr=factor(c(rep('X',length(X)),rep('Y',length(Y)))),
               val=c(X,Y))
rnk <- rank(d$val)
dd <- data.frame(d,rk=rnk)
RX <- dd$rk[dd$gr=='X']
RY <- dd$rk[dd$gr=='Y']
```

上枠内のように直接データ入力してもいいが、Microsoft Excel や OpenOffice.org の Calc のような表計算ソフトで、入院患者と外来患者を 1 行ずつ、横に 32 カラム使って面積変化割合を入力し、まず入院患者の全データを選択してコピーしてから、

R に移って `X <- scan("clipboard")` として、次に再び表計算ソフトに戻って外来患者の全データを選択してコピーしてから R に移って `Y <- scan("clipboard")` としてデータ入力の部分を済ませる方が便利だろう。

2群のデータを積み重ねたデータフレームを定義してから `rank()` 関数を使って順位をつけると2群を併せた順位になるので、その後で `[]` を使って群別のデータに分ければ、`RX` に `X` の各値の順位が得られ、`RY` に `Y` の各値の順位が得られる。-100 はすべて9位となる(1位から17位までが-100なので平均順位の9位が与えられる)ので、`RX` や `RY` を表示させても最初の方の値は9.0ばかりである。

`RX` と `RY` それぞれの合計は、`sum(RX)` より $R_X = 858$ 、`sum(RY)` より $R_Y = 1222$ となる。`RY` の合計の方が大きいのでそちらを検定統計量として採用し、同順位がある場合の正規近似から、 R_Y の期待値 $E(R_Y)$ は、 $E(R_Y) = 32 \times (32 + 32 + 1) / 2 = 1040$ となり、分散 $V(R_Y)$ は、 $V(R_Y) = 32 \times 32 \times (32 + 32 + 1) / 12 - 32 \times 32 / \{12 \times (32 + 32 - 1) \times (32 + 32)\} \times \{17^3 - 17 + (2^3 - 2) \times 5\} = 5442.413$ となるので、

$$\frac{R_Y - E(R_Y) - 1/2}{\sqrt{V(R_Y)}} = 2.460259$$

より^{*13} $2 \times (1 - \text{pnorm}(2.460259, 0, 1)) = 0.01388366$ となるので^{*14}、有意水準5%で帰無仮説は棄却され、2群には有意な差があるといえる。この計算部分のRのプログラムは下枠内の通り。

```
c11-1.R(2)
print(sum(RX))
print(sum(RY))
N <- 32+32+1
print(ERY <- 32*N/2)
print(VRY <- 32*32*N/12-32*32/(12*(N-2)*(N-1))*(17^3-17+(2^3-2)*5))
print(z0 <- (sum(RY)-ERY-0.5)/sqrt(VRY))
print(2*(1-pnorm(z0,0,1)))
```

^{*13} この場合は2群のサンプルサイズが等しいので、期待値や分散は `RX` と `RY` のどちらを検定統計量にしても変わらないが、連続修正の効く方向が異なるので合計順位の大きい方を検定統計量とする。

^{*14} `wilcox.test(X,Y,exact=F)` の結果の `p-value` と一致する。`exact=F` は、有意確率の計算を正確な確率でなく正規近似で行うオプション指定である。このオプションがないと、サンプルサイズが50未満の場合、順位の組み合わせを使った正確な確率が計算される。

例題 2

R の組み込みデータ `sleep` は、20 人の患者を 10 人ずつ 2 群に分けて、それぞれ異なる催眠剤を与えたときに睡眠時間が何時間長くなったかという値をもつ。変数 `extra` が睡眠時間の増分を示し、`group` が催眠剤の異なる 2 群を示す要因型の変数である。2 群間で睡眠時間の増加量に差はあるか、Wilcoxon の順位和検定をせよ。

手順は以下の通り。実際に入力して結果をみてみよう。

```
attach(sleep)
boxplot(extra~group)
wilcox.test(extra~group,exact=F)
detach(sleep)
```

ここでも `exact=F` オプションをつけないと、同順位があるので正確な p 値が計算できないという警告メッセージが表示される。サンプルサイズが 50 未満なので正確な確率を計算しようとするのだけれども、同順位の値があるので正確な確率が計算できず、正規近似で計算されているためである。 $W = 25.5$, $p = 0.069$ より、2 群間に睡眠時間の増加量の差がないという帰無仮説は有意水準 5% で棄却できないので、差があるとはいえない。

11.1.3 メディアン検定

検定統計量を計算するために、順位そのものでなくても、大小関係を反映するスコアならば使うことができる。順位が i 番目のオブザーベーションのスコア $s(R_i)$ として、順位そのものを使う代わりに、 $i \geq [(N+1)/2]$ のとき 1, $i < [(N+1)/2]$ のとき 0 を用いるのがメディアン検定である。次のように言い換えることもできる。

m 個のデータからなる X と n 個のデータからなる Y を合わせた $N = m + n$ 個のデータを、全体のメディアン以上かメディアン未満かによって分類すると、以下の 2×2 クロス集計表が得られる。

	X	Y	合計
メディアン以上	H	$(m+n/2) - H$	$(m+n)/2$
メディアン未満	$m - H$	$H + (n-m)/2$	$(m+n)/2$
合計	m	n	$m+n$

帰無仮説の下では H は $m/2$ の周りに分布する (超幾何分布) ので、

$Pr(H = h') = {}_n C_{h'} \cdot {}_n C_{(m+n)/2-h'}/m+n C_{(m+n)/2}$ より, $Pr(H \geq h')$ をすべて合計して2倍すれば, 両側検定での有意確率が得られる。下枠内のように関数定義しておけば, 例えば, `median.test(rnorm(100), rnorm(100))` のように2つのベクトルを与えることによって, 簡単にメディアン検定をすることができる。ただし, 通常はデータが歪んでいても Wilcoxon の順位和検定ができれば十分なので, メディアン検定はあまり使われない。

```
median.test <- function(X,Y) {
  M <- median(c(X,Y))
  fisher.test(cbind(table(X>=M), table(Y>=M))) }
```

例題 3

例題 2 で取り上げたデータ `sleep` の2群間でメディアン検定をせよ。

上で定義した関数では $X \sim C$ 型の引数は与えられないので, 上のように定義した後で `median.test(sleep$extra[sleep$group==1], sleep$extra[sleep$group==2])` とするか, または, 下枠内のようにする。得られる p -value が約 0.18 なので, 有意水準 5% では有意差はないと判断できる。

```
c11-2.R
attach(sleep)
X <- subset(extra, group==1, drop=T)
Y <- subset(extra, group==2, drop=T)
M <- median(extra)
fisher.test(cbind(table(X>=M), table(Y>=M)))
detach(sleep)
```

11.1.4 符号付き順位和検定

2群間の各サンプルに対応がある場合には, 単純な順位和検定よりも切れ味がよい方法がある。符号付き順位和検定あるいは符号化順位検定と呼ばれるこの方法は, 対応のある t 検定の場合と同じような考え方に基づく。

変数 X の任意の i 番目 (i は 1 から n までの整数値) のデータが $x_i = e_i + \theta_i$ のように, 誤差変動 e_i と真の効果 θ_i の和であると捉えれば, もし X と Y が同じ母集団からのサンプルであるならば $X - Y$ により X と Y に共通する真の効果打ち消すことができ, $U_i = x_i - y_i = e_i - e_i'$ が得られる。このとき帰無仮説は, e_i と e_i' の分布が

同じということなので、 U_i は原点に対して対称になるはずである。そこで、 U_i の絶対値が小さい方から順に順位 R_i をつける。さらに、 $\varepsilon_i = 1(U_i > 0)$, $\varepsilon_i = -1(U_i < 0)$ とすれば、帰無仮説の下で $Pr(\varepsilon_i = 1) = Pr(\varepsilon_i = -1) = 1/2$ となる。いま、

$$R^* = \sum_{i=1}^n \varepsilon_i R_i$$

とおけば、 R^* の大きさによって検定ができる。

すべての場合 (ε_i の値が各 i について 2 通りあるので、 2^n 通り) を計算してやれば正確な確率が計算できるが^{*15}、 n が大きくなると計算が大変なので、 $n \geq 15$ ならば近似を行ってよいことになっている。 R^* の期待値は

$$E(R^*) = \sum_{i=1}^n R_i E(\varepsilon_i) = \sum_{i=1}^n R_i (1 \times 1/2 + (-1) \times 1/2) = 0$$

分散は

$$\begin{aligned} \text{var}(R^*) &= \sum_{i=1}^n R_i^2 \text{var}(\varepsilon_i) \\ &= \sum_{i=1}^n R_i^2 (1^2 \times 1/2 + (-1)^2 \times 1/2) \\ &= \sum_{i=1}^n R_i^2 = n(n+1)(2n+1)/6 \end{aligned}$$

となるので、標準化と連続修正をして、

$$\frac{|R^*| - 1/2}{\sqrt{\text{var}(R^*)}}$$

が標準正規分布に従うことを利用して検定する。なお、R では、対応のある 2 群の生データを X と Y に付値しておき、`wilcox.test(X, Y, paired=T)` とすればこの検定ができる^{*16}。

^{*15} この正確な確率の計算法は、R. A. Fisher が考案した「並べ換え検定」(permutation test) と呼ばれている。R で実行するには `exactRankTests` ライブラリの `perm.test()` 関数を用いる。

^{*16} 先に挙げた `exactRankTests` ライブラリの `wilcox.exact()` 関数も、`paired=T` オプションをつけられれば符号付き順位和検定を実行できる。

例題 4

以下は、気分障害の患者 9 人についてのハミルトンの抑うつ尺度の測定値である。治療前と精神安定剤で治療を開始した後の値を示す^a。治療によって抑うつが改善されたといえるか、Wilcoxon の符号付き順位和検定で検定せよ。

治療前	1.83	0.50	1.62	2.48	1.68	1.88	1.55	3.06	1.30
治療後	0.878	0.647	0.598	2.05	1.06	1.29	1.06	3.14	1.29

^a 出典: Hollander M, Wolfe DA (1973) *Nonparametric statistical inference*, New York: John Wiley & Sons, pp.27-33.

下枠内のように入力すればよい。注意すべきことは、治療後の方が抑うつ尺度は改善していると期待されるので、対立仮説を「X が Y より大きい」とする片側検定にすべきだということである。なお、対応のある t 検定と同じく、差を計算して、その期待値がゼロより大きいという対立仮説を検定する形にしても同等である。

c11-3.R

```
X <- c(1.83, 0.50, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.30)
Y <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)
wilcox.test(X, Y, paired=T, alt="greater")
# wilcox.test(X-Y, alt="greater")と同じ
```

11.2 多群間の分布の位置の差の検定

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

ノンパラメトリックな分析でも、多群間の分布の位置の差を調べるためのアプローチとしては、パラメトリックな場合と同じく大別して 2 つある。1 つは、一元配置分散分析と同じく、群分け変数が量的な変数に与える効果のみをみるという形にする方法で、もう 1 つは、2 群ずつすべての組み合わせについて検定の多重性を調整して Wilcoxon の順位和検定を繰り返す方法である。

前者を行う代表的な方法がクラスカル=ウォリス (Kruskal-Wallis) の検定である。量的な変数 X 、群分け変数 C にデータが付値されているとして、`kruskal.test(X~C)` とすれば実行できる。なお、一元配置分散分析を行うときに前もってパートレットの検定によって分散の同等性をしておく必要があったのと同様、厳密に言えば、Fligner-Killeen の検定 (`fligner.test(X~C)`) によって、前もってばらつきの同等性を検定しておかねばならない。以下、Kruskal-Wallis の検定の仕組みを箇条書きで

説明する。

1. 「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。
2. まず 2 群の比較の場合の Wilcoxon の順位和検定と同じく、すべてのデータを込みにして小さい方から順に順位をつける（同順位がある場合は平均順位を与える）。
3. 次に、各群ごとに順位を足し合わせて、順位和 $R_i (i = 1, 2, \dots, k; k \text{ は群の数})$ を求める。
4. 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、

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

により H を補正した値 H' を求める。

5. H または H' から表を使って（データ数が少なければ並べ換え検定によって）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。

例題 5

R の組み込みデータ `chickwts` を使って、餌の種類 (`feed`) によって、6 週間飼育後の鶏の体重 (`weight`) に差がでるかを、ノンパラメトリックな方法で分析せよ。

上で説明したとおり、次の枠内のように入力すればよい。ノンパラメトリックな位置母数の比較のためのグラフ表示としては、`boxplot()` 関数を使って層別箱ヒゲ図を作成するのが普通である（2 群の場合と同様）。

```
c11-4.R  
attach(chickwts)  
boxplot(weight~feed)  
fligner.test(weight~feed)  
kruskal.test(weight~feed)  
detach(chickwts)
```

11.2.2 フリードマン (Friedman) の検定

対応のある多群間の差をノンパラメトリックな方法で調べるには、フリードマン (Friedman) の検定と呼ばれる手法を用いた方が切れ味がよい。簡単に説明すると、まず同じ個体について群間で順位をつける (群といっても、対応がある場合だから、例えば 2005 年の予測値と 2010 年の予測値と 2025 年の予測値というように、個々の個体について順位をつけることが可能である)。次に、群ごとに順位の合計 (順位和) を計算する。順位和の二乗和から順位和の平均の二乗を引いた値を統計量 S として、サンプル数が少ない場合は表によって (あるいはシミュレーションによって) 有意確率を計算し、サンプル数が多い場合は自由度が群数より 1 少ないカイ二乗分布に従う統計量 Q を S の 12 倍を個体数と群数と「群数 + 1」の積で割った値として計算して有意確率を計算する。ただし同順位がある場合は調整が必要であり、煩雑なので、通常はコンピュータソフトウェアに計算させる。R では `friedman.test()` 関数を用いる。使い方は `example(friedman.test)` を参照されたい。

なお、予測値では微妙だが、経時変化のある測定値については、個人ごとに時点間が独立でないことが自明なので、パラメトリックな分析なら繰り返しのある分散分析 (Repeated Measures ANOVA) という方法を用いるのが普通である。Friedman の検定は、ノンパラメトリックな分析において Repeated Measures ANOVA に相当するものと解釈することも可能である。しかし、経時変化のある測定値をパラメトリックに分析する場合、最近ではランダム効果も考慮できる線型混合効果モデル (mixed model) を適用することが多く^{*17}、ノンパラメトリックな分析ではランダム効果を扱うことが出来ないので、よほど分布が歪んでいたり外れ値があるのでなければ、パラメトリックな解析が好ましいと思う。

^{*17} 本書の範囲を超えるので詳しくは説明しないが、nlme ライブラリの `lme()` 関数を用いる方法が、B. エヴェリット (石田基広他訳)『R と S-PLUS による多変量解析』シュプリンガー・ジャパン、2007 年の第 9 章に紹介されている。

11.2.3 多重比較

多群があるときに 2 群ずつのすべての組み合わせについて分布の位置の差を検定するには、パラメトリックな分析でやったのと同様、多重比較を行って第 1 種の過誤を調整する必要がある。第 1 種の過誤の調整法としては、Holm の方法と Bonferroni の方法は問題なく使えるが、Tukey の HSD のように母集団の分布を仮定した方法は使えない。R では量的変数を X 、群分け変数を C として、`pairwise.wilcox.test(X,C)` 関数を用いればよい ($X \sim C$ ではなく X, C であることに注意)。なお、サンプルサイズが小さいときは正確な確率を求めるのがデフォルト動作だが、同順位の値があると正確な確率を求めることができず、警告メッセージとともに近似値が表示される。最初から近似計算を指定するには、`exact=F` オプションをつければよい。

例題 6

例題 5 の `chickwts` データを使って、どの餌とどの餌の間で体重が異なるかを、検定の多重性を調整して検定せよ。

上で説明したとおり、下枠内のように入力すればよい (同順位の値が含まれているので `exact=F` オプションをつけないと警告メッセージがでる)。

c11-5.R

```
attach(chickwts)
pairwise.wilcox.test(weight,feed,exact=F)
detach(chickwts)
```

11.3 課題

<http://phi.med.gunma-u.ac.jp/msb/data/p11.txt> は、ある途上国の 3 つの地域 (変数 `GRP`、離島にある `P` 村、首都から車で約 1 時間離れた `M` 村、首都 `H` 市) の、再生産をおえたカップル 30 組ずつが生涯に産んだ子供の数 (変数 `PARITY`) のデータである (架空のものである)。これら 3 地域間で、生涯子供数には互いに有意差があるか、ノンパラメトリックな方法で調べよ。検定の有意水準は 5% とせよ。作図をした上で適切な検定を行い、その結果を提示して解釈せよ。

第 12 章

一般化線型モデル

12.1 一般化線型モデルとは？

前章では、なるべく仮定なしにデータを分析する方法を説明した。本章では逆に、かなり強い法則性を仮定して立てたモデルを、データに当てはめる。モデルによってデータのすべてが完全に説明されることはまずありえないが、かなりの程度説明されれば、そのモデルはデータに内在する法則性として妥当な解釈を与えることができると考えてもいいだろう。

具体的なモデルとしては、重回帰分析、共分散分析、ロジスティック回帰分析を扱う。一般化線型モデル (Generalized Linear Model) は、基本的には、

$$Y = \beta_0 + \beta X + \varepsilon$$

という形で表される (Y が従属変数群^{*1}, X が独立変数群 (及びそれらの交互作用項), β_0 が切片群, β が係数群, ε が誤差項である)。係数群は未定であり、そのモデルがもっとも良くデータに当てはまるようになる数値を、最小二乗法または最尤法で求めるのが普通である。こうして得られる係数は、通常、偏回帰係数と呼ばれ、互いに他の独立変数の影響を調整した、各独立変数独自の従属変数への影響を示す値と考えられる (なお、相対的にどの独立変数の影響が大きいかをみるときは、独立変数の絶対値の大きさに依存してしまう偏回帰係数で比較することはできず、標準化偏回帰係数を用いる^{*2})。R では、`glm()` という関数を使ってモデルを記述するのが基本だが、外部ライブラリとして、もっと凝ったモデル記述とその当てはめを行うための

*1 変換したものである場合もある

*2 なお、標準化偏回帰係数は、各偏回帰係数に各独立変数の標準偏差を掛け、従属変数の標準偏差で割れば得られる。

パッケージがいくつも開発され、CRAN で公開されている。また、一般化線型モデルとは違うモデルとして、独立変数群の効果が線型結合でない（例えば、ある独立変数の二乗に比例した大きさの効果があるような場合）、いわゆる非線型モデルも `nls()` 関数で扱うことができる。

12.2 モデルの記述法

R の `glm()` 関数における一般化線型モデルの記述は、例えば、

- (1) 独立変数群が X_1 と X_2 で、従属変数が Y であり、 Y が正規分布に従う場合、
- (2) (1) と同じ構造だが切片をゼロに固定して偏回帰係数を推定したい場合、
- (3) `dat` というデータフレームに従属変数 Y とその他すべての独立変数が含まれていて、余分な変数はなく、 Y が 2 値変数である場合、
- (4) 独立変数群がカテゴリ変数 C_1, C_2 と、それらの交互作用項で、従属変数が正規分布に従う量的変数 Y である場合、

について順に示すと、下枠内ようになる。

```
glm(Y ~ X1+X2)
glm(Y ~ X1+X2-1)
glm(Y ~ ., data=dat, family="binomial")
glm(Y ~ C1+C2+C1:C2)
```

`family` のデフォルトは `"gaussian"` なので、上 2 行のように `family` を指定しなければ正規分布を仮定することになる。この場合、モデルとしては単純な線型重回帰モデルとなるため、例えば (1) の場合なら `lm(Y ~ X1+X2)` と同等である。`summary(lm())` ならば自由度調整済み重相関係数の二乗が得られるので、従属変数にも正規分布を仮定できる単純な線型重回帰モデルで済むときは、`lm()` を使うことを薦める。(4) も従属変数が正規分布に従うので、`lm()` の方がよい。また、独立変数が複数のカテゴリ変数であるときに、主効果と交互作用項のすべてを指定するには、*で変数名をつなぐ方法もあり、(4) の右辺は `C1*C2` とも書ける。(4) のモデルは二元配置分散分析なので、結局、`anova(lm(Y ~ C1*C2))` とするのが普通である（交互作用効果がある場合、平方和の求め方にも注意する必要がある。次のコラムも参照せよ）。

コラム：分散分析における平方和

分散分析表にてでくる因子の残差平方和の求め方としては、因子が直交していれば（因子間の交互作用がなければ）、他の因子を加える順序によらず一定になるので、他の因子を含まない単独のモデルで出した平方和をそのままその因子の平方和とみなしてい（これが逐次平方和と呼ばれる Type I の平方和）けれど、因子が直交していないときは別の考え方をする必要があって、そこで出てくるのが、Type II とか Type III の平方和である。

Type II の平方和を計算するには、まずすべての因子の主効果を含むモデルを基準にする。それから 1 つの因子を取り去ったモデルのモデル平方和と元のモデルのモデル平方和の差を、取り去った因子の寄りとみなして、その因子の偏平方和（Type II SS）とする。次に 2 因子交互作用を含むモデルを基準にして、交互作用を取り去ったモデルとの平方和の差を交互作用効果の偏平方和とする。

Type III は繰り返し数不揃いのときにデータ数の少ないセルを他のセルと同等とみなす目的で使う平方和である。同等とみなすと逆にバイアスが生じる可能性もあるので、不揃いでも Type II を使うべきという意見もある。Type IV は SAS には入っているが、あまり使われない。高橋・大橋・芳賀『SAS による実験データの解析』（東大出版会）によると、数量化 I 類をするときや、乱塊法の場合や、MANOVA の場合や欠損値がある場合は Type II の使用が薦められるとあるので、とりあえず Type I と Type II だけ出せば充分ではないかと思う。なお、同書の 16 章には、行列言語 IML で Type III の平方和を計算する方法が載っている。

R で分散分析を実行する場合、標準の `anova()` や `aov()` では Type I の平方和が計算されるが（`anova(lm())` が `aov()` と同じ意味）、`car` ライブラリの `Anova()` では Type II または Type III（後者を出すには `type="III"` という引数をつける）の平方和が計算できる。ただ、`library(car)` してから `help(Anova)` すると、`Anova()` 関数で計算される Type II は SAS の Type II と同じだが Type III は微妙に違うので注意して使えと書かれている。`car` ライブラリの開発者 John Fox の著書 “An R and S-PLUS companion to applied regression.” (SAGE Publications) の p.140 の Type III の説明によると、例えば因子 A の主効果を、因子 B の主効果と因子 A と因子 B の交互作用効果をテストした後でテストしたいような場合に、他の効果のすべてを出した後で因子 A によって加えられる分を Type III として計算するとのことである。

結論としては、R で、因子が直交していなくてセルごとの繰り返し数不揃いの二元配置分散分析をしたいときは、`library(car)` してから、`Anova(lm(Y~C1*C2))` を使えば Type II の平方和、つまり偏平方和が計算されるので、そうすることをお薦めする。

また、これらのモデルの当てはめの結果は、`res <- glm(Y ~ X1+X2)` のように任意のオブジェクト（この例では `res` のこと）に保存しておくことができ、`plot(residuals(res))` として残差プロットをしたり、`summary(res)` として詳細な結果を出力させたり、`AIC(res)` として AIC を計算させたり、`step(res)` として変

数選択をさせたりするのに使える。

12.3 変数の種類と数の違いによる線型モデルの分類

以下のように整理すると、 t 検定、分散分析、回帰分析といった分析法が、すべて一般化線型モデルの枠組みで扱えることがわかる。

分析名	従属変数 (Y)	独立変数 (X)
t 検定 (注 1)	量的変数 1 つ	2 値変数 1 つ
一元配置分散分析	量的変数 1 つ	カテゴリ変数 1 つ
多元配置分散分析 (注 2)	量的変数 1 つ	カテゴリ変数複数
(単) 回帰分析	量的変数 1 つ	量的変数 1 つ
重回帰分析	量的変数 1 つ	量的変数複数 (注 3)
共分散分析	量的変数 1 つ	(注 4)
ロジスティック回帰分析	2 値変数 1 つ	2 値変数, カテゴリ変数, 量的変数複数
正準相関分析	量的変数複数	量的変数複数

(注 1) Welch の方法でない場合。

(注 2) 独立変数となるカテゴリ変数 (因子とも呼ぶ) が 2 つの場合は二元配置分散分析, 3 つなら三元配置分散分析と呼ばれる。独立変数はカテゴリ変数そのものだけでなく, 交互作用項も含めるのが普通である。分散分析をするときには変数ごとに平方和を求めるわけだが, 二元配置以上では平方和の求め方が Type I から Type IV まで 4 通りあるので注意が必要である (前掲コラムを参照)。

(注 3) カテゴリ変数はダミー変数化せねばならない。ただし, 要因型にしておけば, モデルに投入する際は自動的にダミー変数化される。

(注 4) 2 値変数 1 つと量的変数 1 つの場合が多いが, 「2 値変数またはカテゴリ変数 1 つまたは複数」と「量的変数 1 つまたは複数」を両方含めれば使える。

こう考えてみると, t 検定は分散分析の特殊な場合ということができるし, 分散分析は線型モデルの特殊な場合ということができるし, 線型モデルは一般化線型モデルの特殊な場合ということができる。

12.4 重回帰分析についての留意点

重回帰分析が独立変数 1 つの回帰分析よりも優れている点は、複数の独立変数を同時にモデルに投入することにより、従属変数に対する、他の影響を調整した個々の変数の影響をみることができることである。

重回帰分析は、何よりもモデル全体で評価することが大切である。例えば、独立変数が年齢と体重と一日当たりエネルギー摂取量、従属変数が血圧というモデルを立てれば、年齢の偏回帰係数（または偏相関係数または標準化偏回帰係数）は、体重と一日当たりエネルギー摂取量の血圧への影響を調整した（取り除いた）後の年齢と血圧の関係を示す値だし、体重の偏回帰係数は年齢と一日当たりエネルギー摂取量の影響を調整した後の体重と血圧の関係を示す値だし、一日当たりエネルギー摂取量の偏回帰係数は、年齢と体重の影響を調整した後の一日当たりエネルギー摂取量と血圧の関係を示す値である。独立変数が年齢と体重で従属変数が血圧である場合の年齢の偏回帰係数は、独立変数に一日当たりエネルギー摂取量も入っている場合の年齢の偏回帰係数とは異なる。

モデル全体としてのデータへの当てはまりは、重相関係数の 2 乗（決定係数）や、AIC で評価する。

あるモデルの中で、各独立変数が他の独立変数の影響を調整した上でも従属変数に有意な影響を与えているかどうかをみるには、独立変数ごとに、偏回帰係数の有意性検定を行う。ある独立変数の偏回帰係数がゼロという帰無仮説を検定するには、その変数と従属変数の間の偏相関係数がゼロという帰無仮説を t 分布を使って検定すればよい。また、1 つの重回帰モデルの中で、相対的にどの独立変数が従属変数（の分散）に対して大きな影響を与えているかは、偏相関係数の 2 乗の大小によって評価するか、または標準化偏回帰係数^{*3}によって比較することができる。しかし、別の重回帰

^{*3} 前述の通り、偏回帰係数にその独立変数の不偏標準偏差を掛けて従属変数の不偏標準偏差で割れば得られる。R で計算するには、例えば `res <- lm(Y ~ X+Z)` という重回帰分析を実施した場合なら、`coef(res)*c(0,sd(X),sd(Z))/sd(Y)` でよい。ここで `coef(res)` は、切片、 X の偏回帰係数、 Z の偏回帰係数からなるベクトルを返すので、0 は切片用である。ただ、`lm()` は欠損値を勝手に削除して解析してくれるが、`sd()` はそうでないので、欠損値があるとエラーになってしまう。したがって、回帰分析投入前に `subset()` と `complete.cases()` を使って欠損値の無いデータにしておくか、あるいは、`coef(res)*sd(res$model)/sd(res$model[1])` として、回帰分析に使用されたデータだけを不偏標準偏差の計算に用いるべきである。`res$model[1]` は回帰分析に用いられた従属変数の生データを保持し、`res$model[2]` が以降に回帰分析に用いられた独立変数群の生データを保持している。

モデルとの間では、原則として比較不可能である。

12.5 多重共線性 (multicollinearity)

一般に、複数の独立変数がある場合の回帰で、独立変数同士に強い相関があると、重回帰モデルの係数推定が不安定になるのでうまくない。ごく単純な例でいえば、従属変数 Y に対して独立変数群 X_1 と X_2 が相加的に影響していると考えられる場合、 $\text{lm}(Y \sim X_1+X_2)$ という重回帰モデルを立てるとしよう。ここで、実は X_1 が X_2 と強い相関をもっているとすると、もし X_1 の標準化偏回帰係数の絶対値が大きければ、 X_2 による効果もそちらで説明されてしまうので、 X_2 の標準化偏回帰係数の絶対値は小さくなるだろう。まったくの偶然で、その逆のことが起こるかもしれない。したがって、係数推定は必然的に不安定になる。この現象は、独立変数群が従属変数に与える線型の効果を共有しているという意味で、多重共線性 (multicollinearity) と呼ばれる。

多重共線性があるかどうかを判定するには、独立変数間の散布図を 1 つずつ描いてみるなど、丁寧な吟味をすることが望ましいが、各々の独立変数を、それ以外の独立変数の従属変数として重回帰モデルを当てはめたときの重相関係数の 2 乗を 1 から引いた値の逆数を VIF (Variance Inflation Factor; 定訳は不明だが、分散増加因子と訳しておく) として、VIF が 10 を超えたら多重共線性を考えねばならないという基準を使う (Armitage et al., 2002) のが簡便である。多重共線性があるときは、拡張期血圧 (DBP) と収縮期血圧 (SBP) のように本質的に相関があっても不思議はないものだったら片方だけを独立変数に使うとか、2 つの変数を使う代わりに両者の差である脈圧を独立変数として使うのが 1 つの対処法だが、その相関関係自体に交絡が入る可能性はあるし、情報量が減るには違いない。変数を減らさずに調整する方法としては、centring という方法がある。リッジ回帰 (R では MASS ライブラリの `lm.ridge()`) によっても対処可能である。また、DAAG ライブラリ (Maindonald and Braun, 2003) の `vif()` 関数を使えば、自動的に VIF の計算をさせることができる*4。

*4 ただし、Armitage et al. (2002) が説明している方法と若干計算方法が異なり、結果も微妙に異なる。

例題 1

R に標準で入っているデータフレーム `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。
 ニューヨーク市のオゾン濃度を, セントラルパークの日照, LaGuardia 空港の平均風速, 日最高気温によって説明する重回帰モデルを, このデータに当てはめよ。

重回帰モデルの当てはめと, 3 つの独立変数すべてについて Armitage et al. (2002) の方法で VIF の算出を行う R のプログラムを次の枠内に示す。なお, ここでは VIF を計算する関数 `VIF()` を定義したが, DAAG ライブラリを使って VIF を計算する場合は, `lm()` の結果を `res` に付値した後, `require(DAAG)` 実行後に `vif(res)` とすれば, 3 つの独立変数全ての VIF が得られる。

c12-1.R

```
attach(airquality)
res <- lm(Ozone ~ Solar.R+Wind+Temp)
VIF <- function(X) { 1/(1-summary(X)$r.squared) }
VIF(lm(Solar.R ~ Wind+Temp))
VIF(lm(Wind ~ Solar.R+Temp))
VIF(lm(Temp ~ Solar.R+Wind))
summary(res)
coef(res)*sd(res$model)/sd(res$model[1])
AIC(res)
detach(airquality)
```

3 つの独立変数の VIF はすべて 10 より遥かに小さく, 多重共線性の問題はないと考えられる。`summary(res)` の結果は次の枠内の通り得られる。すべての偏回帰係数が 5% 水準でゼロと有意差があり, 3 つの独立変数すべてがオゾン濃度に有意に影響している。偏回帰係数は, 他の独立変数の値が変わらないとして各独立変数の値が 1 単位増えたときに従属変数の値がどれだけ変わるかを示す値なので, 独立変数の単位に依存するし符号も意味をもつ。この結果では, `Solar.R` と `Temp` はオゾン濃度に正の効果をもち, `Wind` は負の効果をもつことがわかる (日照と気温が高いほどオゾン濃度が高くなり, 風速が低いほどオゾン濃度が高くなるのは, 直感的にも明らかだが)。また, Adjusted R-squared (自由度調整済み重相関係数の 2 乗) の値から,

オゾン濃度のばらつきが、これら 3 つの独立変数のばらつきによって約 60% 説明されることがわかる。

```
Call:
lm(formula = Ozone ~ Solar.R + Wind + Temp)

Residuals:
    Min       1Q   Median       3Q      Max
-40.485 -14.219  -3.551   10.097   95.619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Solar.R      0.05982    0.02319   2.580  0.01124 *
Wind        -3.33359    0.65441  -5.094 1.52e-06 ***
Temp         1.65209    0.25353   6.516 2.42e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.18 on 107 degrees of freedom
Multiple R-Squared:  0.6059,    Adjusted R-squared:  0.5948
F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

次の $\text{coef(res)} * \text{sd(res\$model)} / \text{sd(res\$model[1])}$ で得られる標準化偏回帰係数は、次の枠内の通りである。切片のところは回帰式の切片そのものである。この絶対値をみると、オゾン濃度のばらつきに対して、相対的には気温のばらつきの影響がもっとも大きいことがわかる。

(Intercept)	Solar.R	Wind	Temp
-64.3420789	0.1638655	-0.3564122	0.4731461

以上の結果を含めて、重回帰分析（重回帰モデルの当てはめ）の結果は、次のような表の形にまとめることが多い。AIC については後述する。

表. ニューヨーク市のオゾン濃度に寄与する要因の重回帰分析結果

独立変数	偏回帰係数	標準化偏回帰係数	t 値	有意確率
切片	-64.3	—	-2.79	0.006
Solar.R	0.060	0.164	2.58	0.011
Wind	-3.334	-0.356	-5.09	< 0.001
Temp	1.652	0.473	6.52	< 0.001

Adjusted R^2 : 0.59, F 値 54.8 (自由度 3, 107), $p < 0.001$, AIC: 998.7

12.6 モデルの評価

モデルの当てはめで大事なのは、(1) どのモデルがよりよくデータを説明するのか？ (2) そのモデルはどの程度よくデータを説明しているのか？ を評価することである。以下、簡単にまとめてみる。

線型回帰モデルならば決定係数、すなわち自由度調整済み重相関係数の 2 乗が大きいモデルを採用するというのが 1 つの考え方である。しかし、この基準はかなりナイーブである。一般に、モデルの採否を決定するための基準としてよく使われるのは、残差分析、尤度比検定、AIC である。

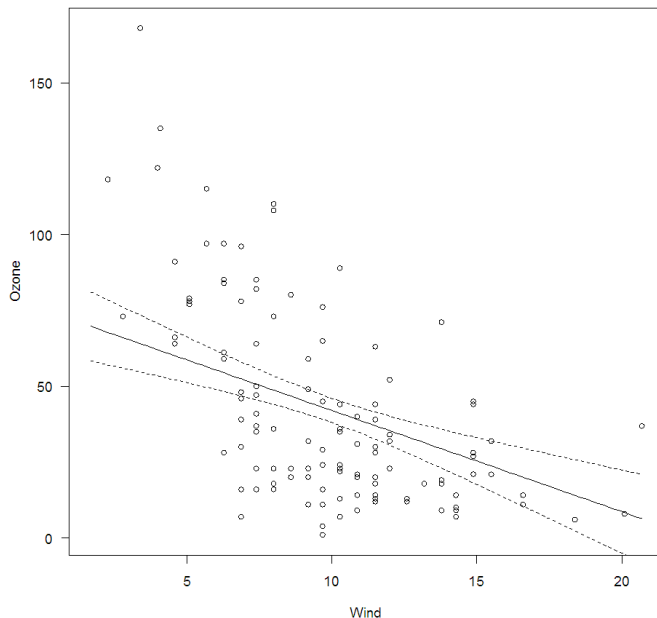
12.6.1 残差分析と信頼区間

残差分析を行うと、モデルがデータから系統的にずれていないかどうかを検査することができる。系統的なズレは、とくにモデルを予測や信頼区間の推定に用いる場合に大きな問題となる（系統的なズレが大きいモデルは、そういう目的には使えない）。回帰モデルの結果を `res` に付値しておけば、例えば、Wind の大小と残差の大小の間に関連があるかどうか見るためには、`plot(residuals(res)~res$model$Wind)` とすることで、回帰の結果から残差を取り出してプロットすることができる（横軸としてはすべての独立変数について試してみるべきである）。横軸の大小によらず、縦軸のゼロの近辺の狭い範囲にプロットが集中していれば、残差に一定の傾向がないことになり、系統的なズレはなさそうだと判断できる。なお、横軸の変数を指定せずに、`plot(residuals(res))` したときの横軸は、オブザーベーションの出現順を意味するインデックス値になる。

残差分析の裏返しのようなイメージになるが、信頼区間の推定も有用である。線型モデルであれば、信頼区間の推定には `predict()` 関数を用いることができる。例えば Wind のとる範囲に対して 95% 信頼区間を得るためには、他の 2 つの変数が平均値で固定されていると仮定して、次の枠内のプログラムを用いれば、Wind を横軸に、

Ozone を縦軸にしたデータそのものがプロットされた上で、重回帰モデルによる推定値が実線で、その 95% 信頼区間が点線で重ね描きされる。ただし、単回帰分析（独立変数が 1 つだけの回帰分析）の場合ほど意味がクリアではない。

```
c12-2.R  
attach(airquality)  
res <- lm(Ozone ~ Solar.R+Wind+Temp)  
EW <- seq(min(Wind),max(Wind),len=100)  
ES <- rep(mean(Solar.R,na.rm=T),100)  
ET <- rep(mean(Temp,na.rm=T),100)  
Ozone.EWC <- predict(res,list(Wind=EW,Solar.R=ES,Temp=ET),  
  interval="conf")  
plot(Ozone~Wind)  
lines(EW,Ozone.EWC[,1],lty=1)  
lines(EW,Ozone.EWC[,2],lty=2)  
lines(EW,Ozone.EWC[,3],lty=2)  
detach(airquality)
```



12.6.2 尤度比検定

次に、モデルの相対的な尤もらしさを考えよう。重回帰分析で独立変数が3つの場合とそのうち1つを除いた2つの場合、あるいは3次回帰と2次回帰のように、一方が他方を一般化した形になっている場合は、これら2つのモデルの尤もらしさを比較できる。

一般に、 $f(x, \theta)$ で与えられる確率密度関数からの観測値を $\{x_1, x_2, \dots, x_n\}$ とするとき、 θ の関数として、 $L(\theta) = f(x_1, \theta)f(x_2, \theta) \cdots f(x_n, \theta)$ を考えると、確率密度関数の値が大きいくところほど観測されやすいため、 $L(\theta)$ の値を最大にするような θ を真の θ の推定値とみなすのが一番尤もらしい。この意味で $L(\theta)$ を尤度関数と呼び、この θ のような推定量のことを最尤推定量 (MLE; Maximum Likelihood Estimator^{*5})

^{*5} MLE は、Maximum Likelihood Estimation、つまり最尤推定法の略として使われる方が普通かもしれない。

と呼ぶ。尤度関数を最大にすることはその対数をとったもの（対数尤度）を最大にすることと同値なので、対数尤度を θ で偏微分した式の値をゼロにするような θ の中から $\ln L(\theta)$ を最大にするものが、最尤推定量となる。例えば、正規分布に従うサンプルデータについて得られる尤度関数を母平均 μ で偏微分したものをゼロとおいた「最尤方程式」を解けば、母平均の最尤推定量が標本平均であることがわかる。詳しくは鈴木 (1995) を参照されたい。

一般に、より一般性の低いモデルをデータに当てはめたときの最大尤度を、より一般的なモデルの最大尤度で割った値の自然対数をとって -2 を掛けた値 λ は、「尤度に差がない」という帰無仮説のもとで、比較するモデル間のパラメータ数の差を自由度とする（つまり 2 パラメータモデルと 3 パラメータモデルの比較なら自由度 1 の）カイ二乗分布に従うので、検定ができる。この検定を尤度比検定と呼ぶ。R では、`logLik()` が対数尤度とパラメータ数を計算する関数なので、この関数を使えばよい。

例題 2

例題 1 と同じデータで、独立変数が日照、風速、気温すべてであるモデルと、独立変数が日照と風速だけのモデルを尤度比検定せよ。

次の枠内のように入力すれば、尤度比検定した有意確率は 10^{-9} のオーダーなので、有意水準 5% で帰無仮説は棄却される。したがって、この場合は 2 変数よりも 3 変数のモデルを採用すべきである。

c12-3.R

```
attach(airquality)
res.3 <- lm(Ozone ~ Solar.R+Wind+Temp)
res.2 <- lm(Ozone ~ Solar.R+Wind)
lambda <- -2*(logLik(res.2)-logLik(res.3))
1-pchisq(lambda,1)
detach(airquality)
```

この例題では線型重回帰の関数 `lm()` を扱ったが、ここで示した尤度比検定の考え方は、2 つのモデルが包含関係にありさえすれば、一般化線型モデル `glm()` でも非線型モデル `nls()` でも、同じように使える。

12.6.3 AIC: モデルの当てはまりの悪さの指標

さて一方, AIC はパラメータ数と最大尤度からモデルの当てはまりの悪さを表すものとして計算される指標で, 数式としては, L を最大尤度, n をパラメータ数として,

$$\text{AIC} = -2 \ln L + 2n$$

で表される。AIC が小さなモデルほど当てはまりがいい = 良いモデルであると考える。

R には, `AIC()` という関数と `extractAIC()` という 2 つの関数がある。前者は “Akaike’s An Information Criterion” となっていて, 後者は “The (generalized) Akaike *A*n *I*nformation *C*riterion for a fitted parametric model” となっている。前者が以前からある汎用関数である。`extractAIC()` は MASS ライブラリに含まれていたのが S4 メソッドとして標準実装されるようになった関数で, 変数選択のために `step()` 関数の中から呼び出されるのが主な用途である。

例えば, 例題 2 に示された 2 つのモデルについて AIC を計算するには, `AIC(res.3)` とすれば,

```
-2*logLik(res.3)+2*attr(logLik(res.3),"df")
```

と同じで 998.7 が得られる。一方, `extractAIC(res.3)` の結果は 681.7 となる。`res.2` についても同様で, `AIC(res.2)` は 1033.8, `extractAIC(res.2)` は 716.8 を返す。この結果から, 独立変数 3 つのモデルの方が AIC が小さく良いモデルと言える。定義通りの AIC を返すのは `AIC()` 関数なのだが, 変数選択に使うためならそれと定数の差があってもいいので, 計算量が少ない `extractAIC()` 関数が `step()` で使われている*6。

*6 <http://www.is.titech.ac.jp/~shimo/class/gakubu200409.html> (東工大・下平英寿さんの講義「R による多変量解析入門」の第 8 回「モデル選択」の資料) に, それぞれが使っている式の説明があり, `AIC()` 関数は $-2 \ln L + 2\theta$ (L は最大尤度, θ はパラメータベクトルの次元) を計算する汎用関数であって, オブザーベーション数 n , パラメータ数 p , 標準偏差 σ として, 線型重回帰の場合は $n(1 + \ln(2\pi\sigma^2)) + 2(p + 1)$ を計算し (正規分布を仮定するから), `extractAIC()` 関数は線型重回帰のときだけ使える関数で, $n \ln(\sigma^2) + 2p$ を計算する。前者から後者を引けば $n(1 + \ln(2\pi)) + 2$ と, オブザーベーション数は含むけれどもパラメータ数には依存しない定数になるので, 変数選択はこちらでやっても問題ないことになる。

12.7 変数選択

このように、重回帰モデルの独立変数の取捨選択を行うことを変数選択と呼ぶ。step(lm(Ozone~Wind+Solar.R+Temp)) のように、step() 関数を使って自動的に変数選択を行わせることができる*7。変数増加法 (direction="forward"), 変数減少法 (direction="backward"), 変数増減法 (direction="both") などがある。減少法の場合、direction="backward" オプションをつけるが、変数選択候補範囲を明示的に与えない場合、step() 関数のデフォルトは減少法になっているので、線型重回帰分析の結果を step() に渡す場合には、direction 指定はしなくても同じ結果になる。

このデータの場合、ress <- step(lm(Ozone~Solar.R+Wind+Temp)) とすると、3 つすべての変数が残った場合の AIC である 682 が最小であることがわかり、採択されたモデルが ress に保存される。ここで表示された AIC は step() 関数が、内部的に extractAIC() 関数を使って得た値なので、通常の AIC を表示するには、採択されたモデルに対して AIC(ress) としなくてはならない。lm() で使われたオブザーベーションが 111 しかないので (Ozone と Solar.R に欠損値が多いため)、 $AIC(ress) - 111 * (1 + \log(2 * \pi)) - 2$ とすると、確かに 681.7127 という結果になり、step() 関数の出力に出てくる値と一致することがわかる。まとめると、変数減少法で変数選択をさせ、最終的に採択されたモデルについての情報を表示させるには、下枠内のように入力すればよい。

```
c12-4.R(1)
```

```
attach(airquality)
res <- lm(Ozone~Solar.R+Wind+Temp)
ress <- step(res)
summary(ress)
AIC(ress)
```

重回帰分析では、たくさんの独立変数の候補から比較的少数の独立変数を選択することが良く行われるが、モデル全体で評価するという観点からは、あまり薦められない。数値以外の根拠により投入する変数を決めて、各々の偏回帰係数 (または偏相関係数) が有意であるかないかを見る方が筋がよい。十分な理由があれば、有意でない

*7 glm() については、以前のバージョンでは残差分析や尤度比検定や AIC の結果を見ながら手作業でモデリングを進めていくしかなかったが、R-2.5.0 では既に適用可能になっている。

変数も含めた重回帰式を作っても良い。とくに疫学研究においては、先行研究によって交絡変数になることが一般に示されている変数は、独立変数に固定的に入れてモデルを作成するのが普通である。

しかし、数値以外の根拠が薄い場合もあるし、偏回帰係数が有意でない（偏相関係数がゼロという帰無仮説が棄却されない）変数を重回帰モデルに含めることを嫌う立場もある。したがって、数値から最適なモデルを求める必要もありうる。そのためには、総当り法が良いとされる。総当り法では、まず、独立変数が 1 個の場合、2 個の場合、3 個の場合、……のそれぞれについて、すべての組み合わせの重回帰モデルを試し、重相関係数の二乗が最大となるモデルをそれぞれ求める。その上で、独立変数が n 個の場合が $n - 1$ 個の場合のすべての変数を含むならば尤度比検定を行って、尤度が有意に大きくなるところまでの $n - 1$ 個を独立変数として採用する。独立変数が n 個の場合が $n - 1$ 個の場合のすべての変数を含まない場合は判断が難しく、一つ一つの変数の意味を先行研究とも比較するなどしてモデルへの採否を吟味すべきだが、すべてのモデルの中で AIC が最小になるモデルを選ぶのは一つの基準となりうる。M.G. ケンドール著（奥野忠一、大橋靖雄訳）『多変量解析』（培風館、1981）でも総当り法が薦められているが、R の `step()` 関数では提供されていない*8。

12.8 採択されたモデルを使った予測

モデルの当てはめがうまくできれば、独立変数群の値から従属変数の値を予測することができる。そのためには、信頼区間の計算で示したように、`predict()` 関数を利用する。例えば、風速も日照も気温も観測値の平均値になった日に、オゾン濃度がいくつになるかを、`res<-lm(Ozone~Solar.R+Wind+Temp)` という回帰式から予測するには、先の枠内のコードを実行させた後に次の枠内を実行する*9。

*8 外部ライブラリとしては、CRAN からダウンロードできる `leaps` ライブラリは総当り法のためのライブラリである。

また、http://aoki2.si.gunma-u.ac.jp/R/All_possible_subset_selection.html に、群馬大学社会情報学部の青木繁伸教授が開発された R コードが公開されている。

*9 ここで、平均値を求めるために `mean(Solar.R)` でなく `mean(res$model$Solar.R)` などとしているのは、データに欠損値が含まれるためである。`mean(Solar.R, na.rm=T)` などでも欠損値のある変数の平均値は計算できるが、それだと個々の変数ごとに欠損値が除かれる。回帰モデルをデータに当てはめるときは、モデルに含まれる変数のどれか 1 つでも欠損があったケースは除外されて回帰式が推定されるので、`res$model$Solar.R` などによりモデルの当てはめに使われたデータを参照せねばならない。

c12-4.R(2)

```
predict(res, list(Solar.R=mean(res$model$Solar.R),
  Wind=mean(res$model$Wind), Temp=mean(res$model$Temp)))
```

他の観測値がわかっている、オゾン濃度だけを測れなかった日の値を推定する（補間）にも、同じ方法が使える。

ただし、単回帰のところでも述べたように、回帰の外挿には慎重でなければならない。こうして推定された偏回帰係数を用いて、Solar.R と Temp がそれぞれこの重回帰分析で使われた値の平均値で、Wind=25 のときのオゾン濃度を点推定すると約-8.1 となってしまって、やはり採用できない（95% 信頼区間はゼロを跨いでいるが）。結局、いくら AIC が小さくなくても、論理的に問題がある線型回帰を適用して予測をしてはいけないということである。

そこで登場するのが非線型回帰である。Wind と Ozone の間に負の相関関係があるので Wind が大きくなると Ozone がマイナスになるという線型回帰の弱点を避けるために、例えば Wind と Solar.R の 2 変数で、係数が負の指数関数の形で風速がオゾン濃度に影響する非線型関係を仮定したモデルを作って回帰分析を行うには、nls() 関数を用いて、次の枠内のようにする（ここで注意すべきは start=list() オプションで指定する未定な係数の初期値である。いい初期値を指定しないと収束しなかったり変な解に収束してしまうことがある）。

c12-4.R(3)

```
resmr.2 <- nls(Ozone ~ a*exp(-b*Wind) + c*Solar.R,
  start=list(a=200,b=0.2,c=1))
summary(resmr.2)
AIC(resmr.2)
```

AIC は独立変数 3 つの線型モデルより大きいですが、独立変数 2 つの線型モデルより小さいので悪くはない。そこで、次の枠内のように Temp も入れて独立変数 3 つの非線型モデルを作って尤度比検定をする。有意確率は約 0.149 となるので、3 変数にしても当てはまりは有意に改善しないことになり、独立変数 2 つの非線型モデルが採用できる^{*10}。

^{*10} もっとも、十分に AIC が小さいとはいえないので、このデータに含まれていない要因の影響が大きいと思われ、本来はもっと別の要因を探索する必要がある。


```
c12-4.R(4)
resmr.3 <- nls(Ozone ~ a*exp(-b*Wind) + c*Solar.R + d*Temp,
  start=list(a=200,b=0.2,c=1,d=1))
summary(resmr.3)
AIC(resmr.3)
lambda <- -2*(logLik(resmr.2)-logLik(resmr.3))
print(1-pchisq(lambda,1))
```

そこで続けて次の枠内を入力すれば、*11

```
c12-4.R(5)
SRM <- mean(subset(Solar.R,complete.cases(Ozone,Solar.R,Wind)))
predict(resmr.2,list(Wind=25,Solar.R=SRM))
detach(airquality)
```

約 16.4 となるので、LaGuardia 空港の平均風速が 25 マイル/時のときのニューヨーク市のオゾン濃度は、太陽放射が平均的な条件なら、約 16.4 ppb になると予測される。

12.9 共分散分析

共分散分析は、典型的には、

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$$

というモデルになる。2 値変数 X_1 によって示される 2 群間で、量的変数 Y の平均値に差があるかどうかを比べるのだが、 Y が量的変数 X_2 と相関がある場合に X_2 の Y への影響を考えたモデルを立て（このとき X_2 を共変量と呼ぶ）、 X_2 と Y の回帰直線の傾き (slope) が X_1 の 2 群間で差がないときに、 X_2 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) に、 X_1 の 2 群間で差があるかどうかを検定する。

*11 Solar.R が欠損値を含むため、モデルの当てはめに使われたデータだけの平均値を求めるために、1 行目で subset() と complete.cases() を使っていることに注意されたい。nls() では lm() と異なり、モデルの当てはめに使われたデータは結果オブジェクト内に保持されていない。

コラム：共分散分析モデルの数式

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、次の 2 つの仮説が考えられる。

1. まず傾きに差があるかどうか？ を考える。つまり、 $H_0 : \beta_1 = \beta_2, H_1 : \beta_1 \neq \beta_2$ である。

2. 次に、もし傾きが等しかったら、 y 切片も等しいかどうかを考える。つまり、 $\beta_1 = \beta_2$ のもとで、 $H'_0 : \alpha_1 = \alpha_2, H'_1 : \alpha_1 \neq \alpha_2$ を検定する。

各群について、 X と Y の平均値 (E_X, E_Y) と変動 (SS_X, SS_Y) と共変動 (SS_{XY}) は、サンプルサイズ N_1 の第 1 群に属する x_i, y_i について、 $E_{X1} = \sum x_i / N_1, SS_{X1} = \sum (x_i - E_{X1})^2, E_{Y1} = \sum y_i / N_1, SS_{Y1} = \sum (y_i - E_{Y1})^2, E_{XY1} = \sum x_i y_i / N_1, SS_{XY1} = \sum (x_i y_i - E_{XY1})^2$ となる。第 2 群も同様に計算できる。

これらの値を使えば、対立仮説 H_1 のもとでの残差平方和

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

と帰無仮説 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、 $F = (d_2 - d_1) / (d_1 / (N - 4))$ 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。

H_0 が棄却されたときは、 $\beta_1 = SS_{XY1} / SS_{X1}, \beta_2 = SS_{XY2} / SS_{X2}$ 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。

H_0 が採択されたときは、共通の傾き β を、 $\beta = (SS_{XY1} + SS_{XY2}) / (SS_{X1} + SS_{X2})$ として推定する。

この場合はさらに y 切片が等しいという帰無仮説 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、 $F = (d_3 - d_2) / (d_2 / (N - 3))$ を求める。 F 値が帰無仮説のもとで第 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する。

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、採択されたら、要するに 2 群間に差がないということになるので、2 群を一緒にして普通の単回帰分析をしていいことになる。

R で共分散分析を実行する手順を次に示す。なお、以下の説明では、 X_1 を示す変数名を C (注： C は factor, つまり要因型の変数である必要がある), X_2 を示す変数名を X とし、 Y を示す変数名を Y とする。

- 2 本の回帰直線がともに有意にデータに適合していて、かつ 2 本の回帰直線の間で傾き (slope) が等しいかどうかを検定して、傾きが等しいことを確かめておかないと、修正平均の比較には意味がない。そこで、まず例えば、

`summary(lm(Y[C==levels(C)[1]]~X[C==levels(C)[1]])` と入力して第 1 のグループでの単回帰分析を行い、次に `summary(lm(Y[C==levels(C)[2]]~X[C==levels(C)[2]])` として第 2 のグループでの単回帰分析を行なう^{*12}。そもそも回帰直線の適合が悪ければその変数は共変量として考慮する必要がない。

2. 2 本の回帰直線の適合が良ければ、`summary(lm(Y~C+X+C:X))` (または `summary(lm(Y~C*X))`) として傾きの差について検討する。傾きに有意差があることは C と X の交互作用項が有意に Y に影響していることと同値なので、結果出力の中で `Coefficients` の `C2:X` と書かれている行の右端を見れば、「傾きに差がない」帰無仮説の検定の有意確率が得られる。傾きに差があればグループによって独立変数と従属変数の関係が異なっているわけだから、2 群を層別して別々に解釈すべきである。
3. 「傾きに差がない」帰無仮説が棄却されなかった場合、`summary(lm(Y~C+X))` とすれば、X の影響を調整した上で、C 間で Y の修正平均 (調整平均) が等しいという帰無仮説についての検定結果が得られる。C2 と表示される行の右端に出ているのがその有意確率である。

例題 3

R の組み込みデータ `ToothGrowth` は、各群 10 匹ずつのモルモットに 3 段階の用量のビタミン C をアスコルビン酸としてあるいはオレンジジュースとして投与したときの象牙芽細胞 (歯) の長さを比較するデータである。変数 `len` が長さ、`supp` が投与方法、`dose` が用量を示す。「相関と回帰」の章では投与方法の違いを無視して用量と長さの関係を調べたが、用量と長さの関係が投与方法によって異なるかどうかを共分散分析を使って調べよ。

まずグラフを描いてみる。共分散分析をするような場面では、通常、次の枠内のように群によってマークを変えて散布図を重ね描きし、さらに線種を変えて群ごとの回帰直線を重ね描きする。あるいは、`coplot(len~dose | supp)` として横に 2 枚の散布図が並べて描かれるようにすることも可能である。

^{*12} `C==levels(C)[1]` は、`as.integer(C)==1` と等価である。要因型変数を整数扱いすると、何番目の水準かを示す整数値になる。

```
c12-5.R(1)
```

```
attach(ToothGrowth)
plot(dose, len, pch=as.integer(supp), ylim=c(0, 35))
legend(max(dose)-0.5, min(len)+1, levels(supp), pch=c(1, 2))
abline(lm1 <- lm(len[supp=='VC'] ~ dose[supp=='VC']))
abline(lm2 <- lm(len[supp=='OJ'] ~ dose[supp=='OJ']), lty=2)
summary(lm1)
summary(lm2)
```

summary(lm1) と summary(lm2) をみると、投与方法別の回帰係数がゼロと有意差があることがわかる。そこで次に 2 本の回帰直線の傾きに有意差がないという帰無仮説を検定する。モデルの右辺に独立変数間の交互作用項を含めればいいので、

```
c12-5.R(2)
```

```
lm3 <- lm(len ~ supp*dose)
summary(lm3)
detach(ToothGrowth)
```

とすると、

```

Call:
lm(formula = len ~ supp * dose)

Residuals:
    Min       1Q   Median       3Q      Max
-8.22643 -2.84625  0.05036  2.28929  7.93857

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.550      1.581    7.304 1.09e-09 ***
suppVC        -8.255      2.236   -3.691 0.000507 ***
dose           7.811      1.195    6.534 2.03e-08 ***
suppVC:dose    3.904      1.691    2.309 0.024631 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.083 on 56 degrees of freedom
Multiple R-Squared:  0.7296,    Adjusted R-squared:  0.7151
F-statistic: 50.36 on 3 and 56 DF,  p-value: 6.521e-16

```

という出力が得られる。suppVC:dose の従属変数 len への効果（交互作用効果）がゼロという帰無仮説の検定の有意確率が 0.024631 なので、有意水準 5% で帰無仮説は棄却される。したがってこの場合は「投与経路によって投与量と長さの関係の傾きが有意に異なる」ことを示した上で、先に計算済みの投与経路別の回帰分析の結果を解釈すればよい。修正平均の差の検定はしても意味がない。

例題 4

`http://phi.med.gunma-u.ac.jp/msb/data/p12.txt` は変数名付きのタブ区切りテキスト形式のデータで、5 つの変数が含まれている（`PREF` が都道府県名、`REGION` が東日本か西日本か、`CAR1990` が 1990 年の 100 世帯当たりの自動車保有台数、`TA1989` が 1989 年の人口 10 万人当たり交通事故死者数、`DIDP1985` が 1985 年の人口集中地区居住者割合である）。次の枠内のコードを実行すると、データフレーム `dat` に読み込むことができる。

```
c12-6.R(1)
```

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p12.txt")
attach(dat)
```

ここで、東日本と西日本で交通事故死者数に差があるかを検討したいとする。しかし、交通事故死者数は、自動車保有台数と関連がありそうなので、もしそうなら、その影響を調整した上でなお、東日本と西日本で差があるかを検定したい。どうしたらよいか。

まず散布図を描いてから、`REGION` ごとに自動車保有台数と交通事故死者数の単相関を検討すると、東日本の回帰係数の有意性検定の p -value は $5.72e-05$ 、西日本では 0.00267 と、ともに 5% よりずっと小さいので、有意な関連があるといえる。

次に 2 つの回帰直線の傾きに有意差があるかどうかを検定すると、`REGIONWest:CAR1990` の p -value は 0.990 なので、傾きに有意差はないといえる。

そこで修正平均の差を検定すると、`REGIONWest` の係数の検定の有意確率は 0.0319 なので、5% 水準では有意である。よって、自動車保有台数を調整しても、東日本と西日本では交通事故死者数に有意差があるといえる。以上の分析を実行するためのコードは次の枠内の通りである。

```

c12-6.R(2)
plot(CAR1990,TA1989,pch=as.integer(REGION))
legend(max(CAR1990)-10,min(TA1989)+1,levels(REGION),pch=c(1,2))
abline(lm1 <- lm(TA1989[REGION=='East']~CAR1990[REGION=='East']))
abline(lm2 <- lm(TA1989[REGION=='West']~CAR1990[REGION=='West']),lty=2)
summary(lm1)
summary(lm2)
lm3 <- lm(TA1989 ~ REGION*CAR1990)
summary(lm3)
lm4 <- lm(TA1989 ~ REGION+CAR1990)
summary(lm4)
detach(dat)

```

12.10 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（ロジスティック回帰分析では反応変数と呼ぶこともある）が 2 値変数であり、正規分布に従わないので `glm()` を使う。

例えば疾病の有無について、量的な変数も含めた複数の変数の影響を調整しながら、ある要因の有無によるオッズ比を計算できるのが利点であり、医学統計ではもつともよく使われる手法の一つである。

この問題は、疾病の有病割合を P とすると、 $\ln(P/(1-P)) = b_0 + b_1X_1 + \dots + b_kX_k$ と定式化できる。 X_1 が要因の有無を示す 2 値変数で、 X_2, \dots, X_k が交絡であるとき、 $X_1 = 0$ の場合を $X_1 = 1$ の場合から引けば、

$$b_1 = \ln(P_1/(1-P_1)) - \ln(P_0/(1-P_0)) = \ln(P_1 * (1-P_0)/(P_0 * (1-P_1)))$$

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仮定すれば、オッズ比の 95% 信頼区間が

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

として得られる。

モデルの当てはまりについては、前述の AIC の他、Nagelkerke の R^2 という値（Nagelkerke, 1991）が使われることがある。これは線型回帰の場合に使われる自由度調整済み重相関係数の二乗（いわゆる決定係数）を一般化したもので、Faraway

(2006) によれば,

$$R^2 = \frac{1 - (\hat{L}_0/\hat{L})^{2/n}}{1 - \hat{L}_0^{2/n}} = \frac{1 - \exp((D - D_{null})/n)}{1 - \exp(-D_{null}/n)}$$

である。 L は尤度 (L_0 は帰無仮説の下での尤度) を示し, D は Deviance (線型回帰における残差平方和のようなもの), D_{null} は帰無仮説の下での Deviance である。 n はサンプルサイズである。決定係数と同じく 0 から 1 の間の値をとり, モデルがデータのどれくらいを説明しているかを表す指標である。

例題として, `require(MASS)` で MASS ライブラリをロードし^{*13}, `birthwt` というデータフレームを使ってロジスティック回帰分析を実行してみる。これは, 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るためのデータであり, 含まれている変数は以下の通りである。

<code>low</code>	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)
<code>age</code>	年齢
<code>lwt</code>	最終月経時体重 (ポンド ^a)
<code>race</code>	人種 (1=白人, 2=黒人, 3=その他有色人種)
<code>smoke</code>	喫煙の有無 (1=あり)
<code>ptl</code>	非熟練労働経験数
<code>ht</code>	高血圧の既往 (1=あり)
<code>ui</code>	子宮神経過敏の有無 (1=あり)
<code>ftv</code>	妊娠の最初の 3 ヶ月の受診回数
<code>bwt</code>	児の出生時体重 (g)

^a 略号 lb. で, 1 lb. は 0.454 kg に当たる。

このデータフレームを使ってロジスティック回帰分析をするには, まず次の枠内を打って変数の型などを整える。従属変数 (反応変数) は疾病の有無を示す 2 値変数でなくてはならない。「疾病あり」が 1 で「疾病なし」が 0 であれば数値型のままでも問題ないが, 通常は `factor()` を使って要因型にする。独立変数については, 共変数としての扱いで良ければ量的な変数でよい。他の変数の影響を調整したオッズ比を出したいときは要因型あるいは論理型にせねばならない。3 つ以上の水準を含む要因型の場合, 自動的にダミー変数化 (複数の 2 値変数の組み合わせに分解) されてモデル

^{*13} `require()` と `library()` は, どちらもインストール済みのライブラリをメモリに呼び出す関数であり, 実用上ほとんど差がない。

に投入される。オッズ比を算出する際のリファレンスカテゴリは、もっとも水準が低いカテゴリになる。つまり、論理型なら FALSE の方が TRUE より水準が低いので、FALSE がリファレンスになるし、要因型の場合は最初の水準がリファレンスになる（要因型変数の各カテゴリの水準を確認するには `levels()` 関数を用いる）。リファレンスカテゴリを変えたい場合は、`relevel()` 関数を使うと便利である（コラム参照）。なお、いきなりロジスティック回帰分析をするのではなく、それより前に独立変数 1 つずつ、従属変数との関係を見るために、クロス集計や図示などしておくべきである。ここではクロス集計をしている。

```
c12-7.R(1)
```

```
require(MASS)
attach(birthwt)
low <- factor(low)
race <- factor(race, labels=c("white","black","other"))
print(table(low,race))
smoke <- (smoke>0); print(table(low,smoke))
ht <- (ht>0); print(table(low,ht))
ui <- (ui>0); print(table(low,ui))
```

続いて次の枠内を入力する。即ち、加工済みの変数を `data.frame()` でまとめて新しいデータフレーム `bw` を作り、先に `attach()` した `birthwt` を `detach()` する。`glm()` にはいくつかのモデル記述法があるが、ここでは `data=` オプションで指定したデータフレームに含まれる、従属変数以外のすべての変数を独立変数とする指定にした。結果を `res`（この名称は任意）というオブジェクトに付値してから、まず `summary()` をとって結果を表示する。次に前述の Nagelkerke の R^2 を計算する（枠内のスクリプトでは関数定義してからそれを適用している）。次に `exp(coef(res))` とすると、対数オッズ比の指数をとることになるので、各変数についてオッズ比の点推定量が得られる。次の行は、その 95% 信頼区間を求めている。さらに、`res2<-step(res)` によりステップワイズで変数選択させた結果を `res2` に付値し、それについても同様な分析を行う^{*14}。変数選択後の結果は、下表のような形でまとめることが多い。人種は 3 つのカテゴリがあるので、自動的にダミー変数化されて処理されている。切片と量的な変数のオッズ比には意味が薄いので、表の中で

^{*14} ただし、ロジスティック回帰分析においては、機械的にステップワイズで変数選択をすることは必ずしも推奨できない。むしろ、従属変数に対する効果をみたい変数と交絡因子となっている変数はすべてモデルに投入するべきである。

なく下に共変量として調整したと書くのが普通である。他の変数の影響を調整しても喫煙者は非喫煙者に比べて約 2.56 倍、低体重出生児をもちやすいと解釈できる。

c12-7.R(2)

```

bw <- data.frame(low,age,lwt,race,smoke,ptl,ht,ui,ftv)
detach(birthwt)
print(summary(res <- glm(low ~ ., family=binomial, data=bw)))
NagelkerkeR2 <- function(rr,n) {
  (1-exp(-(rr$dev-rr$null)/n))/(1-exp(-rr$null/n)) }
print(NagelkerkeR2(res,nrow(bw)))
print(exp(coef(res)))
print(exp(confint(res)))
print(summary(res2 <- step(res)))
print(NagelkerkeR2(res2,nrow(bw)))
print(exp(coef(res2)))
print(exp(confint(res2)))

```

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		p 値
		下限	上限	
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

Nagelkerke の R^2 : 0.223, 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、最終月経時体重と非熟練労働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。

コラム：要因型変数の水準操作

birthwtに含まれているftvという変数は、妊娠後3ヶ月間の受診回数を示す数値型の変数である。今回は数値型のまま処理したが、データが明らかな二峰性を示すなど、カテゴリズしたいこともあるだろう。ftvの度数分布は下枠内の通りなので、例えば0回、1回、2回以上、という3つのカテゴリに分けることを考える。

```
  0  1  2  3  4  6
100 47 30  7  4  1
```

一番簡便な方法は、

```
ftv <- factor(ftv)
levels(ftv)[3:6] <- "2+"
```

である。ftvを要因型に変換すると6つの水準をもつ変数となり、2回以上（最初の水準は0回なので、3番目の水準以降）の水準名をすべて"2+"にしてしまえば、それらは統合される。しかしこの方法は少々トリッキーである。まともにするなら、下枠内のようにする。

```
ftv <- factor(ifelse(ftv>2,2,ftv),labels=c("0","1","2+"))
```

carライブラリのrecode()関数を使えば、相当柔軟なカテゴリ変換が可能になる。この場合なら下枠内のようにする。

```
require(car)
ftv <- recode(factor(ftv),"2:6='2+'")
```

こうして要因型に型変換したftvは3つの水準をもつ。これをロジスティック回帰の独立変数として使うとリファレンスカテゴリは"0"になるのだが、3ヶ月で1回行くのが普通なので、2番目のカテゴリである"1"をリファレンスにしたい場合（つまり、3ヶ月に1回行く人に比べて、一度も行かない人や2回以上行く人のオッズが何倍になるか知りたい場合）は下枠内のようにrelevel()関数を使って水準を変換するとよい。

```
ftv <- relevel(ftv,2)
```

なお、マッチングを行った症例対照研究でロジスティック回帰分析を行うには、条件付ロジスティック回帰分析を行わねばならないため、glm()では実行できない。Rでは、生存時間解析のパッケージsurvivalの中に、コックス回帰を行うための関数coxph()を利用して条件付ロジスティック回帰分析を行うclogit()というラッ

パー関数が提供されており、モデルの右辺に `+strata`(マッチング変数) という項を付加することで条件付ロジスティック回帰分析が可能である。

12.11 課題

MASS ライブラリに含まれている `bacteria` というデータフレームは、オーストラリアのノーザンテリトリーに居住するアボリジニの、中耳炎の乳児 50 人についての Dr. Amanda Leach らの研究結果である。抗生物質 (アモキシリン) あるいはプラセボ (偽薬) を投与する 2 群にランダムに割付け、インフルエンザ菌 (*H. influenzae*) が検出されるかどうかを調べた RCT である。投薬されてもきちんと服薬しない場合があるために、単純に 2 群間で比較することができず、コンプライアンスも含めてデータが取られている。`bacteria` に含まれている変数は以下の通りである。

<code>y</code>	インフルエンザ菌の有無。水準 <code>n</code> と <code>y</code> をもつ要因型変数で、 <code>y</code> がありなので、このまま従属変数にできる。
<code>ap</code>	投薬割付け。水準 <code>a</code> と <code>p</code> をもつ要因型変数で、 <code>a</code> が本当の薬を与える群、 <code>p</code> が偽薬群。
<code>hilo</code>	コンプライアンス。水準 <code>hi</code> と <code>lo</code> をもつ要因型変数で、 <code>hi</code> がコンプライアンスが良い群、 <code>lo</code> が悪い群を意味する。
<code>week</code>	週数。対象者ごとに、実験参加後の週数を示す整数型変数。
<code>ID</code>	対象者の ID。50 の水準をもつ要因型変数。
<code>trt</code>	処置。 <code>placebo</code> , <code>drug</code> , <code>drug+</code> の 3 つの水準をもつ要因型変数。 <code>ap</code> と <code>hilo</code> を再コーディングしたもので、 <code>placebo</code> は偽薬群、 <code>drug</code> は本当の薬を与えられているがコンプライアンスが悪い群、 <code>drug+</code> は本当の薬を与えられていてコンプライアンスも良い群。

ロジスティック回帰分析により、他の要因の影響を調整した上で投薬がインフルエンザ菌検出オッズを何倍にするか (当然 1 より小さくなる) を検討せよ。同じ子供について 0, 2, 4, 6 週目と経過観察があるのでデータ数は 220 あり、本来なら条件付ロジスティック回帰分析をすべきだが、ここでは 220 人の別々の子供とみなして分析すればよいことにする。ロジスティック回帰分析の結果の表と、投薬の効果についての解釈を記せ。

第 13 章

生存時間解析

13.1 生存時間解析概論

生存時間解析の特徴は、期間データを扱うことと、観察打ち切りケースが含まれるデータを扱えることである。

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（観察期間の終点となるイベント）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。しかし、観察期間には限りがあるので観察中には死亡イベントが起こらないケースもある。そのような「観察打ち切り」*1 ケースは相対的に長く生きている可能性が高いので、そのレコードを単純にデータから除去してしまうと全体の生存時間を過小評価してしまって不都合である。このような期間データを扱うには、一般に生存時間解析 (Survival Analysis または Event History Analysis) と呼ばれる分析法を用いる。

なかでもよく知られている指標が Kaplan-Meier の積・極限推定量である。現在では、普通、カプラン = マイヤ推定量と呼ばれている。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*2）あたりのイベント発生数（即ち、イベント発生率）を 1 から引いたものを掛け合わせて得られる関数値である。カプラン = マイヤ推定量が 0.5（つまり 50%）を横切る時点は、イベントを経験せずにいる人が期首人口の半分になるまでの時間を意味する。言い換えると、この点は、生存時間の中央値の、ノンパラメトリックな最尤推定量である。通常、推定と同時に階段状

*1 英語では censoring という。

*2 population at risk の訳語として、大橋・浜田 (1995) に従っておく。

の生存曲線を描き、観察打ち切りレコードは曲線上に縦棒で表す。

コラム：生命表解析

データ数が多い場合は、個々の間隔データを集計して生命表解析を行うこともある。生命表解析の代表的なものは、ヒトの平均寿命を計算するときに行われている（官庁統計としても、まさしく生命表という形で発表されている）。平均寿命とは 0 歳平均余命のことだが、これは、ある時点での年齢別死亡率（後述する q_x ）に従って、ゼロ歳児 10 万人が死んでいったとすると、生まれてから平均してどれくらいの期間生存するのかという値である^a。

一般に x 歳平均余命は、 x 歳以降の延べ生存期間の総和 T_x を x 歳時点の個体数 l_x で割れば得られる。延べ生存期間の総和は、ちょうど x 歳に達した者が $x+1$ 歳に達しないで死亡する確率、すなわち年齢階級 $[x, x+1)$ における死亡率 q_x が変化しないとして、 $l_x(1 - q_x/2)$ によって x 歳から $x+1$ 歳まで生きた人口 L_x （開始時点の人口が決まっていれば死亡率も変化しないので x 歳の静止人口と呼ばれる）を求め、それを x 歳以降の全年齢について計算して和をとることで得られる。

ヒトの人口学では、通常の年齢別死亡率 m_x （ある年に x 歳で死亡した人数 d_x をその年の x 歳中央人口で割った値）から q_x を $q_x = m_x / (1 + m_x/2)$ として求めて生命表を計算するのが普通だが^b、生物一般について考えるときは、同時に生まれた複数個体（コホート）を追跡して年齢別生存数として l_x を直接求めてしまう方法（コホート生命表）とか、たんに年齢別個体数を l_x と見なしてしまう方法（静態生命表、偶然変動で高齢の個体数の方が多い場合があるので平滑化するのが普通）がよく行われる。

l_x から

$$\mu_x = -\frac{1}{l_x} \frac{dl_x}{dx}$$

として求められる、ちょうど x 歳における瞬間の死亡率 μ_x は死力と呼ばれる。これは後で説明するコックス回帰で出てくるハザード関数に他ならない。

^a 誤解されることが多いが、死亡年齢の平均ではないので注意されたい。

^b 5 歳階級で計算するときは死亡の線型性を仮定するのは無理なので別の補正を使う。よく用いられる Greville の方法は、

$${}^5q_x = \frac{{}^5m_x}{1/5 + {}^5m_x \left[1/2 + 5/12 \left[{}^5m_x - \ln({}^5m_x + 5/5 {}^5m_x)^{1/n} \right] \right]}$$

である。

一方、複数の期間データ列の差の比較、すなわち生存時間の差の検定には、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。

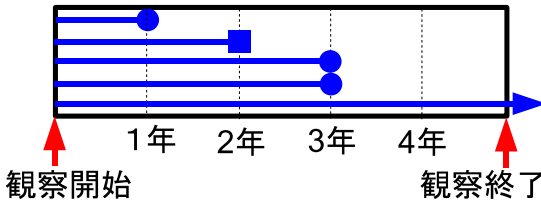
イベントが起こるまでの期間に何らかの別の要因が与える効果を調べたいときはコックス回帰(それらが基準となる個体のハザードに対して $\exp(\sum \beta_i z_i)$ という比例定数の形で掛かるとする比例ハザード性を仮定する方法)と、パラメトリックなモデルに対数線型モデルの独立変数項として入れてしまう加速モデルがある。加速モデルは `survreg()` 関数で実行できるが、本書では詳細には立ち入らない。

R では生存時間解析をするための関数は `survival` ライブラリで提供されており、`library(survival)` または `require(survival)` とすれば使えるようになる^{*3}。

13.2 カプラン = マイヤ法 : survfit() 関数

カプラン = マイヤ推定量について、まず一般論を示しておく。イベントが起こる可能性がある状態になった時点を開始点として、イベントが起こった時点をも t_1, t_2, \dots 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, \dots の直前でリスク集合の大きさを n_1, n_2, \dots で示す。リスク集合の大きさは、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なく、打ち切りによっても減少する。したがって 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。例えば、下図のように 5 人のがん患者をフォローアップしていて、ちょうど 1 年で 1 人死亡くなり、ちょうど 2 年で 1 人転出したためフォローアップ不能になり、ちょうど 3 年で 2 人が同時に死亡し、5 年目まで継続観察してまだ最後の 1 人は生き残っていたとすると、 t_1 が 1 年で t_2 が 3 年となり、 $d_1 = 1$ 、 $d_2 = 2$ 、 $n_1 = 5$ 、 $n_2 = 3$ となる。

^{*3} `survival` は起動時にロードされていないが推奨ライブラリなので、Windows 版では R 本体をインストールするだけで同時にインストールされる。ちなみに、R バイナリに組み込み済みのライブラリは、`base`、`datasets`、`grDevices`、`graphics`、`grid`、`methods`、`splines`、`stats`、`stats4`、`tcltk`、`tools`、`utils` であり、推奨ライブラリで、将来全バイナリに組み込み予定なのは、`survival` の他には、`KernSmooth`、`MASS`、`boot`、`class`、`cluster`、`foreign`、`lattice`、`mgcv`、`nlme`、`nnet`、`rpart`、`spatial` がある。`search()` でロード済みライブラリ一覧、`.packages(all.avail=T)` でインストール済み一覧が表示される。ロード済みのライブラリをアンロードするには `detach(package=survival)` などとする。



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、

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

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、

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

で得られる。なお、カプラン=マイヤ推定では、階段状の生存曲線のプロットを行うのが普通である。

R では、`library(survival)` または `require(survival)` として `survival` ライブラリを呼び出し、`Surv`(生存時間, 打ち切りフラグ) 関数で生存時間型のデータを作る。打ち切りフラグは、1 でイベント発生, 0 が打ち切りを示すが、`TRUE` がイベント発生, `FALSE` が生存としてもいいし, 1 でイベント発生, 2 で打ち切りとすることもできる。また、区間打ち切りレコード^{*4}を扱うときは、`Surv`(最終観察時点, 観察不能になった時点, 打ち切りフラグ) とする。打ち切りフラグは、0 が右側打ち切り, 1 がちょうどイベント発生, 2 が左側打ち切り, 3 が区間打ち切りを示す。打ち切りレコードがないときは、打ち切りフラグそのものを省略することもできる。例えば、1 人目については 81 ヶ月より後で 92 ヶ月より前にイベントが発生した区間打ち切り, 2 人目については 22 ヶ月でイベント発生, 3 人目については 29 ヶ月まで観察して、まだイベントが発生していないという場合、次の枠内のように生存時間型のデータ `dat` を定義する。

^{*4} ある間隔の中で観察打ち切りかイベント発生があったことはわかっているが、正確なその時点が不明のケースは、区間打ち切りレコードとして扱う。例えば、ある年に観察して生存していた人が、翌年行ったら転居してしまっていて追跡不能な場合、最後の観察時点と最初に観察不能になってしまった時点の間のどこかであることは既知だが、その 1 年のどこで打ち切りになったか不明なので、区間打ち切りレコードとなる。


```
time <- c(81,22,29)
time2 <- c(92,22,NA)
event <- c(3,1,0)
dat <- Surv(time,time2,event,type="interval")
```

生存時間型のデータが `dat` に付値されていれば、`res <- survfit(dat)` 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、`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。

コラム：日時を扱う関数

生データとして生存時間が与えられず、観察開始とイベント発生の日付を示している場合、それらの間隔として生存時間を計算するには、`difftime()` 関数や `ISOdate()` 関数を使うと便利である。例えば、次の枠内では、まず `x` というデータフレームに変数 `names` (名前)、`dob` (誕生年月日) と `dod` (死亡年月日) を付値している。次に `difftime()` 関数で 4 人分の死亡年月日と誕生年月日の差 (= 生存日数) を計算し、`[x$names=="Robert"]` で Robert (これは言うまでもなくロベルト・コッホのことである) についてだけの生存日数が得られ、それが `alivedays` に付値される。その次の行のように 365.24 で割れば、生存年数に換算される。`as.numeric()` で数値型に変換しているのは、`difftime()` が返す変数が `difftime` クラスになっていて、単位が日数に固定されているためである。`t.test()` のように `difftime` クラスの変数は受け付けず、数値型にしないと計算できない関数もあるので、一般には `difftime()` の結果は `as.numeric()` で数値型に変換しておくべきだろう。

日数の与え方はロケールによって違う。ロケールは地域情報を意味し、文字コード、通貨記号、日付形式、小数点記号などを決めることができる。R の中では、`Sys.getlocale("LC_ALL")` によってロケールを得ることができるが、日本語 Windows 環境の場合、デフォルトでは "Japanese_Japan.932" となっているはずである。英国環境にしたければ `Sys.setlocale("LC_ALL", "eng")` とすればいいし、日本の設定に戻したいなら `Sys.setlocale("LC_ALL", "jpn")` とすればいい。日本語ロケールの場合、ダブルクォーテーションマークで括って、年、月、日がハイフンまたはスラッシュでつながれた形で与えることもできるし (それが日付であることを明示するには `as.Date()` に渡せばいいし、`as.character()` でも計算は可能)、最終行のように `ISOdate(年, 月, 日)` という形で与えることもできる。なお、`read.delim()` 関数などでファイルから "1749-5-17" などの日付変数の列を読み込んだ場合、通常は要因型として扱われてしまうので、必ず `as.Date()` で型変換する癖をつけておくとうい。ちなみに最終行は、エドワード・ジェンナー、北里柴三郎、ロベルト・コッホ、野口英世の 4 人がもし 2007 年 1 月 22 日に生きていたら満何歳になるかという計算結果を与える。

c13-1.R

```
x <- data.frame(
  names = c("Edward", "Shibasaburo", "Robert", "Hideyo"),
  dob = c("1749-5-17", "1853-1-29", "1843-12-11", "1876-11-9"),
  dod = c("1823-1-26", "1931-6-13", "1910-5-27", "1928-5-21"))
alivedays <- difftime(x$dod, x$dob)[x$names=="Robert"]
as.numeric(alivedays/365.24)
as.numeric(difftime(ISOdate(2007, 1, 22), x$dob)/365.24)
```

例題 1

大橋・浜田 (1995) の p.60-61 に掲載されている Gehan の白血病治療データは、42 人の白血病患者を、6-MP とプラセボを投与するペアにランダムに割り付けて治療し、寛解が続いている週数である (イベントは白血病の再発)。R では MASS ライブラリにも `gehan` というデータフレームとして含まれている。これは有名なデータで、<http://data.princeton.edu/wws509/datasets/gehan.dat> としてプリンストン大学の web サイト内でインターネット上にも公開されている。このデータを用い、6-MP 治療群とプラセボ群の 2 群を別々に、カプラン = マイヤ推定せよ。

プログラムは次の枠内の通り。

c13-2.R

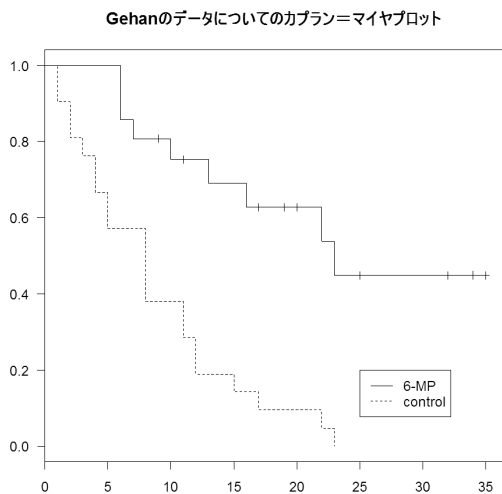
```
require(MASS)
require(survival)
print(res<-survfit(Surv(time,cens)~treat,data=gehan))
par(family="sans",las=1)
plot(res,lty=c(1,2),main="Gehan のデータについてのカプラン = マイヤプロット")
legend(30,0.2,lty=c(1,2),legend=levels(gehan$treat))
summary(res)
```

3 行目により以下が表示される。この結果だけで、2 群別々のメディアン生存時間 (カプラン = マイヤ推定量) とその 95% 信頼区間はわかる。

```
Call: survfit(formula = Surv(time, cens) ~ treat, data = gehan)
```

	n	events	median	0.95LCL	0.95UCL
treat=6-MP	21	9	23	16	Inf
treat=control	21	21	8	4	12

4-6 行目により次のグラフが描かれる (4 行目は表示フォントファミリーを "sans" にし、`las=1` によって縦軸のラベルが 90 回転した形にならないようにする設定で、6 行目の `legend()` は凡例を表示する関数だが、凡例を表示する座標を (30,0.2) としたのは試行錯誤による)。カプラン = マイヤ法でメディアン生存時間を推定した場合、このような生存曲線を描くのが普通である。短い縦棒は観察打ち切りの生じた時点を示す。



7行目の `summary(res)` によって、次の枠内の通り、すべてのイベント発生時点における生存確率と、その 95% 信頼区間が表示される（生存曲線はこの結果を使って描かれている）。

```
Call: survfit(formula = Surv(time, cens) ~ treat, data = gehan)
```

```

      treat=6-MP
time n.risk n.event survival std.err lower 95% CI upper 95% CI
  6    21     3   0.857  0.0764   0.720   1.000
  7    17     1   0.807  0.0869   0.653   0.996
 10    15     1   0.753  0.0963   0.586   0.968
 13    12     1   0.690  0.1068   0.510   0.935
 16    11     1   0.627  0.1141   0.439   0.896
 22     7     1   0.538  0.1282   0.337   0.858
 23     6     1   0.448  0.1346   0.249   0.807

```

```

      treat=control
time n.risk n.event survival std.err lower 95% CI upper 95% CI
  1    21     2   0.9048  0.0641   0.78754   1.000
  2    19     2   0.8095  0.0857   0.65785   0.996
  3    17     1   0.7619  0.0929   0.59988   0.968
  4    16     2   0.6667  0.1029   0.49268   0.902
  5    14     2   0.5714  0.1080   0.39455   0.828
  8    12     4   0.3810  0.1060   0.22085   0.657
 11     8     2   0.2857  0.0986   0.14529   0.562
 12     6     2   0.1905  0.0857   0.07887   0.460
 15     4     1   0.1429  0.0764   0.05011   0.407
 17     3     1   0.0952  0.0641   0.02549   0.356
 22     2     1   0.0476  0.0465   0.00703   0.322
 23     1     1   0.0000    NA         NA         NA

```

例題 2

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、維持化学療法が生存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、生存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。

c13-3.R

```
require(survival)
print(res <- survfit(Surv(time,status)~x, data=aml))
summary(res)
par(family="sans", las=1)
plot(res, lty=c(1,2), main="急性骨髄性白血病の維持化学療法の有無別カプラン = マイヤプロット")
legend(100, 0.8, lty=c(1,2), legend=c("維持群", "非維持群"))
```

上枠内のように入力すると、2 行目の `survfit()` でカプラン = マイヤ推定がなされ、下枠内が表示される。

	n	events	median	0.95LCL	0.95UCL
<code>x=Maintained</code>	11	7	31	18	Inf
<code>x=Nonmaintained</code>	12	11	23	8	Inf

この表は、維持群が 11 人、非維持群が 12 人、そのうち死亡まで観察された人がそれぞれ 7 人と 11 人いて、維持群の生存時間の中央値が 31 週、非維持群の生存時間の中央値が 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。4 行目を入力すると、死亡が観察された 7 人と 11 人について、

それぞれの死亡時点までの生存確率が標準誤差と 95% 信頼区間とともに表示される。最後の 3 行で、例題 1 と同様に生存曲線が描かれる。

13.3 ログランク検定：survdif() 関数

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、生存時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ため、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、死亡イベントが発生したすべての時点では、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。

このラットの例では、死亡イベントが発生した時点 1~8 において各群の期待死亡数を計算し、各群の観測された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 群しかないため、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 の合計スコアの 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、時点の重みについては、ログランク検定ではすべての時点について 1 である。時点の重みを各時点での 2 群を合わせたリスク集合の大きさにした検定方法を、一般化ウィルコクソン検定という（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、イベント発生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*5。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

*5 打ち切りデータは、リスク集合の大きさが変わることを通してのみ計算に寄与する。打ち切り時点

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, $4 \times 4/64 + 4 \times 3/49 + 3 \times 3/36 + 3 \times 2/25 + 2 \times 2/16 + 2 \times 1/9$ で計算すると, 約 1.457 となる。したがって, $\chi^2 = 1.338^2/1.457 = 1.23$ 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 でログランク検定を実行するには, 生存時間を示す変数を time, 打ち切りフラグを event, グループを group とすれば, `survdif(Surv(time,event)~group)` とすればよい。この例の場合なら, 次の枠内の通りである。なお, 一般化ウィルコクソン検定をするには `survdif(Surv(time,event)~group,rho=1)` とすればよい。

c13-4.R

```
require(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)
survdif(Surv(time,event)~group)
```

ではスコアは計算されないことに注意。

出力結果を見ると、 $\chi^2 = 1.2$ 、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。なお、ログランク検定だけでなく、カプラン=マイヤ法により生存時間の中央値を求め、生存曲線も図示するのが普通である。

例題 3

例題 2 のデータで、維持群と非維持群の間に生存時間の有意差はあるか、有意水準 5% でログランク検定せよ。

`survdif(Surv(time,status)~x,data=aml)` と打つだけである。カイ二乗値が 3.4 で、有意確率が 0.0653 と計算される。したがって、有意水準 5% で帰無仮説は棄却されず、ログランク検定では、維持群と非維持群の間の生存時間の差は有意ではないといえる。なお、このデータは打ち切りを含んでいるが、R で計算する限りにおいては、そのことを意識する必要はない（適切に処理してくれる）。

13.4 コックス回帰—比例ハザードモデル：coxph() 関数

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。そのため、比例ハザードモデルとも呼ばれる。

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 $z_i = (z_{i1}, z_{i2}, \dots, z_{ip})$ 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、

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

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。 $\beta_1, \beta_2, \dots, \beta_p$ 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。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。

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、

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

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。

13.4.1 二重対数プロット

ここで生存関数とハザード関数の関係について整理しておこう。まず、 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。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、

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

である。累積ハザード関数は、 $H(t) = \int_0^t h(u) du = -\log S(t)$ となる。これを式変形すると、 $S(t) = \exp(-H(t))$ とも書ける。

そこで、共変量ベクトルが z である個体の生存関数を $S(z, t)$ 、累積ハザード関数を $H(z, t)$ とすれば、

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

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

となる。したがって、比例ハザード性が成立していれば、

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

が成り立つことになるので、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になる。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、コックス回帰するのは不適切といえる。

13.4.2 コックス回帰のパラメータ推定

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*⁶。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 4

例題 2 の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。

*⁶ 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。

```
c13-5.R
require(survival)
summary(res <- coxph(Surv(time,status)~x,data=aml))
loglogplot <- function(X) {
  S <- X$surv
  T <- X$time
  G <- X$ntimes.strata
  GG <- names(X$strata)
  GX <- rep(GG,G)
  xr <- c(0,max(T)*1.5)
  mas <- ifelse(max(S)==1,0.99,max(S))
  mis <- ifelse(min(S)==0,0.01,min(S))
  yr <- c(log(-log(mas)),log(-log(mis)))
  plot(T[GX==GG[1]],log(-log(S[GX==GG[1]])),type="l",lty=1,
       xlim=xr,ylim=yr,xlab="time",ylab="log(-log(S))",
       main="二重対数プロット")
  for (i in 2:length(GG)) {
    lines(T[GX==GG[i]],log(-log(S[GX==GG[i]])),lty=i)
  }
  legend(max(T),-2,legend=GG,lty=1:length(GG))
}
KM <- survfit(Surv(time,status)~x,data=aml)
par(family="sans",las=1,mfrow=c(1,2))
plot(KM,main="aml データのカプラン = マイヤプロット")
loglogplot(KM)
```

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

```

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^{*7}。したがって差はないと解釈される。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% では有意でない。

3 行目以降により、左に 2 群別々に推定した Kaplan-Meier プロットが描かれ^{*8}、右に二重対数プロットがなされる。

^{*7} 最終行に Score (logrank) test とあるが、これは Rao の Score 検定の結果である。survdiff() により実行されるログランク検定の結果とは微妙に異なる。

^{*8} コックス回帰の場合は、通常、群の違いは比例ハザード性を前提として 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ため、plot(survfit(coxph(Surv(time,cens)~treat+pair,data=gehan))) のようにするのだが、推定値は共変量の影響を無視して plot(survfit(Surv(time,cens),data=gehan)) とすると、あまり大きくは変わらないことが多い。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=(treat=="6-MP") 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、この場合は独立変数に群別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のだが、信頼区間まで重ね描きされると見にくいので、あまりお勧めしない。通常は他の共変量とはとりあえず無視して Kaplan-Meier プロットすれば十分であろう。

例題 5

例題 1 にあげた Gehan の白血病治療データで、対照群に対する 6-MP 処置群のハザード比を推定せよ。

プログラムは次の枠内の通りである。

c13-6.R

```
require(MASS)
require(survival)
res <- coxph(Surv(time,cens)~treat,data=gehan)
summary(res)
plot(survfit(res))
```

summary(res) により、次の枠内の結果が出力される。plot(survfit(res)) では、コックス回帰で推定されるベースラインの生存曲線が 95% 信頼区間つきで描かれる。

Call:

```
coxph(formula = Surv(time, cens) ~ treat, data = gehan)
```

n= 42

	coef	exp(coef)	se(coef)	z	p	
treatcontrol	1.57	4.82	0.412	3.81	0.00014	
	exp(coef)	exp(-coef)	lower	.95	upper	.95
treatcontrol	4.82	0.208	2.15	10.8		

Rsquare= 0.322 (max possible= 0.988)

Likelihood ratio test= 16.4 on 1 df, p=5.26e-05

Wald test = 14.5 on 1 df, p=0.000138

Score (logrank) test = 17.3 on 1 df, p=3.28e-05

どの検定結果をみても有意水準 5% で「6-MP 処置が死亡ハザードに与えた効果がない」という帰無仮説は棄却される。exp(coef) の値 4.82 が、2 群間のハザード比の推定値になるので、6-MP 処置群に比べて対照群では 4.82 倍 (95% 信頼区間が [2.15, 10.8]) 死亡ハザードが高いと考えられ、6-MP 処置は有意な延命効果をもつと解釈できる。

13.4.3 コックス回帰における共変量の扱い

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

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、この場合のように、がん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`event`、治療方法を示す群分け変数が`treat`、がんの進行度を表す変数が`stage`であるとき、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,event)~treat+strata(stage))`とすればよい。

なお、コックス回帰はモデルの当てはめなので、一般化線型モデルで説明したのと同様、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。

例題 6

survival ライブラリに含まれているデータ colon は結腸癌の補助的な化学療法の臨床試験の最初の成功例の 1 つである。Levamisole はそれまで動物の寄生虫処理に使われてきた毒性の低い化合物であり、5-FU は中等度の毒性をもつ化学療法の薬剤である。1 人あたり再発と死亡の 2 種類のレコードが含まれている。変数は以下の通りである。

id 個人の id 番号
 study 全員が 1
 rx 処理。3 水準をもつ要因型変数。Obs は経過観察のみ、Lev は Levamisole のみ投与、Lev+5FU は両方投与。
 sex 性別を示す数値型変数。1 が男性、0 が女性。
 age 満年齢を示す数値型変数。
 obstruct 腫瘍による結腸の閉塞を示す数値型変数。1 が閉塞あり、0 が閉塞なし。
 perfor 穿孔の有無を示す数値型変数。1 が穿孔あり、0 が穿孔なし。
 adhere 隣接臓器への癒着の有無を示す数値型変数。1 が癒着あり、0 が癒着なし。
 nodes リンパ節転移数を示す数値型変数。
 status 打ち切りかどうかを示す数値型変数。1 がイベント発生、0 が打ち切り。
 differ 分化の程度を示す数値型変数。1 が高分化、2 が中等度、3 が低分化型。
 extent 局所の広がり程度を示す数値型変数。1 が粘膜下、2 が筋、3 が漿膜、4 が隣接組織。
 surg 手術後の時間の長さを示す。0 が短時間、1 が長時間。
 node4 4 つ以上の癌細胞陽性のリンパ節があるかどうかを示す。1 があり、0 がなし。
 time イベント発生までの時間（日）を示す数値型変数。
 etype イベントの種類を示す数値型変数で、1 が再発、2 が死亡を意味する。

Levamisole 単独投与あるいは 5-FU と共に投与した場合に、経過観察に比べ、死亡のハザード比がいくつになるか、年齢と性別を共変量として調整してコックス回帰せよ。

まず死亡をエンドポイントとするサブセット colon2 を作る。

```
colon2 <- subset(colon, etype==2)
```

とすればよい。年齢と性別を共変量として調整したコックス回帰の実行は次の枠内の通り（`loglogplot()` は c13-5.R で定義済みのもの）。


```
colon2$sex <- factor(colon2$sex)
KM <- survfit(Surv(time,status)~rx,data=colon2)
layout(1:2)
plot(KM)
loglogplot(KM)
res <- coxph(Surv(time,status)~rx+age+sex,data=colon2)
summary(res)
```

Levamisole と 5-FU を両方投与すると、経過観察だけのときに比べて、年齢と性別を調整したコックス回帰でハザード比が 0.688 倍 (95% 信頼区間が [0.545,0.869]) で、ハザード比 1 を帰無仮説とする検定の有意確率は 0.0017 となるので、両方投与は結腸癌治療に有効であることが示された。このコックス回帰モデル自体のデータへの適合度は、 $R^2 = 0.013$ と低いが⁹、尤度比検定で $\chi^2 = 12.5$, $d.f. = 4$, $p = 0.014$ と有意であり、意味はある。モデルの適合度を上げるには分化の程度など共変量を増やすのが常道である。

なお、年齢と性別を調整したコックス回帰のベースラインの生存曲線を 95% 信頼区間付きで処理別に描かせるには、次の枠内のようにすればよい⁹。

```
attach(colon2)
xls <- c(0,max(time))
plot(survfit(coxph(Surv(time,status)~age+sex,subset=(rx=="Obs"))),
     col=1,xlim=xls)
par(new=T)
plot(survfit(coxph(Surv(time,status)~age+sex,subset=(rx=="Lev"))),
     col=2,xlim=xls)
par(new=T)
plot(survfit(coxph(Surv(time,status)~age+sex,subset=(rx=="Lev+5FU"))),
     col=3,xlim=xls)
detach(colon2)
```

⁹ なお、共変量を調整したコックス回帰で推定される生存関数について二重対数プロットを描かせるのはなかなか面倒であるが、余裕があれば試してみると面白いかもしれない。

例題 7

大橋, 浜田 (1995) の付録 A に掲載されている膵臓癌データを入力してタブ区切りテキストファイルとして公開しており, それを R で読み込んで変数の型などを適切に設定するプログラムも作成し公開してあるので, 同書 p.138-150 に掲載されている SAS の解析結果と R の `coxph()` 関数による解析結果を比較せよ。

pcancer.R

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p13.txt")
dat$CENSOR <- 1-dat$CENSOR
dat$SEX <- factor(dat$SEX, labels=c("男性","女性"))
dat$TREAT <- factor(dat$TREAT, labels=c("照射なし","照射あり"))
dat$CH <- ordered(dat$CH,labels=c("CH0","CH1","CH2","CH3"))
dat$STAGE <- ordered(dat$STAGE,labels=c("III","IV"))
dat$PS <- ordered(dat$PS,labels=c("0","1","2","3","4"))
# data from 大橋・浜田 (1995)「生存時間解析 SAS による生物統計」東京大学
# 出版会付録 A.3
# 膵臓がんデータ (Nishimura et al. 1988. の一部)
# 手術中放射線照射の延命効果を見る目的のデータだがランダム割付ではない。
# 入力 Minato Nakazawa 20/1/2007 * R では死亡が 1 なので CENSOR のカ
# テゴリに注意。
# *** 変数リスト ***
# CASENO 患者番号
# TIME 生存時間 (月) 連続量
# CENSOR 打ち切りの有無を示すフラグ (元データでは打ち切りが 1, 死亡が 0 で
# R と逆なので 1 から引く)
# AGE 手術時の年齢 連続量
# SEX 性別 0 が男性, 1 が女性
# TREAT 処置法 0 が術中照射無し, 1 が術中照射あり
# BUI 占拠部位 0 が頭部, 1 が頭部以外
# CH 膵内胆管への浸潤 順序尺度 1: CH0, 2: CH1, 3: CH2, 4: CH3
# P 腹膜播種性転移 0 がなし, 1 があり
# STAGE TNM 分類のステージ 順序尺度 3: III 期, 4: IV 期
# PS Performance Status (活動度) 順序尺度 1: 0, 1, 2: 2, 3: 3, 4: 4
```

pcancer.R を実行後^{*10}, コックス回帰を行うコードは以下の通りである。同時発生イベントに対する当時の SAS のデフォルトは Breslow 法だったので, 同書と結

^{*10} `source("http://phi.med.gunma-u.ac.jp/msb/pcancer.R")` とすればよい。

果を比較するために敢えて `method="breslow"` で実行したが、もちろん Efron 法や Exact 法の方が望ましい。

```
c13-7.R
require(survival)
summary(coxph(Surv(TIME,CENSOR)~AGE+SEX+TREAT, data=dat,
  method="breslow"))
summary(coxph(Surv(TIME,CENSOR)~AGE+SEX+TREAT, data=dat,
  method="efron"))
summary(coxph(Surv(TIME,CENSOR)~AGE+SEX+TREAT, data=dat,
  method="exact"))
res <- step(coxph(Surv(TIME,CENSOR)~AGE+SEX+TREAT+BUI+CH+P+STAGE+PS,
  data=dat, method="breslow"))
summary(res)
```

結果の出力順序はやや異なるが、Breslow 法の結果は SAS とほぼ一致していた（わずかな丸め誤差と思われる違いは存在するが）。変数選択の結果も SAS と一致し、最終的に TREAT, BUI, STAGE が選ばれた。モデルの検定結果や有意確率もほぼ一致した^{*11}。

13.5 課題

`survival` ライブラリに入っているデータ `ovarian` は、卵巣がんに対する 2 種類の治療法を比較する無作為化臨床試験の結果である。Eastern Cooperative Oncology Group の研究であり、含まれている変数は以下の通りである。

^{*11} ただし、なぜか STAGE のリスク比だけが一致しなかった。同じデータについて同じ解析をしたときに SAS と同じ結果が出ることは、ソフトの信頼性の 1 つの根拠になりうるが、筆者はこの程度の違いであればあまり問題はないと考える。この食い違いが何に起因するのか不明だが、他のデータでも検証しておくべきかもしれない。

futime 生存時間または観察打ち切りまでの時間
fustat 打ち切りフラグ
age 年齢
resid.ds 残留疾病の有無 (1 がなし, 2 があり)
rx 治療種類 (処理群別を示す変数)
ecog.ps ECOG 能力状況 (0 が病気がないのとまったく同じく何の制限もなく活動できる, 1 が強い運動はできないが軽い家事労働やオフィスワークならできる, 2 が起きている時間の半分くらいは活動できる, 3 が半分以上ベッドか椅子にいる, 4 がセルフケア不能, 5 が死亡を意味する)

このデータから, 治療種類の違いによって卵巣がんの生存時間に差が出たか, 年齢と残留疾病の有無を共変量として調整して分析せよ。

第 14 章

文献

14.1 R に関する日本語の文献

- 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.
- 岡田昌史 (編) (2004) 「The R Book」九天社.
- 舟尾暢男 (2004) 「The R Tips」九天社.
- U. リゲス (石田基広訳) (2006) 「R の基礎とプログラミング技法」シュプリンガー・ジャパン.
- Peter Dalgaard (岡田昌史監訳) (2007) 「R による医療統計学」丸善.
- B. エヴェリット (石田基広他訳) (2007) 「R と S-PLUS による多変量解析」シュプリンガー・ジャパン.
- 金明哲 (2007) 「R によるデータサイエンス：データ解析の基礎から最新手法まで」森北出版

14.2 R に関する英語の文献

- Faraway JJ (2006) Extending the linear models with R: Generalized linear, mixed effects and nonparametric regression models. Chapman and Hall.
- Maindonald J, Braun J (2003) Data analysis and graphics using R, Cambridge Univ. Press.

14.3 疫学・統計学についての文献

- 大橋靖雄, 浜田知久馬 (1995) 「生存時間解析 SAS による生物統計」東京大学出版会.
- 鈴木義一郎 (1995) 情報量基準による統計解析, 講談社サイエンティフィック.
- 高橋・大橋・芳賀 (1989) SAS による実験データの解析, 東大出版会.
- M.G. ケンドール著 (奥野忠一, 大橋靖雄訳) (1981) 多変量解析, 培風館.
- Armitage P, Berry G, Matthews JNS (2002) *Statistical Methods in Medical Research*, 4th ed., Blackwell Publishing.
- Nagelkerke N (1991) A note on a general definition of the coefficient of determination. *Biometrika*, 78: 691-692.

14.4 R に関するウェブサイト

- <http://phi.med.gunma-u.ac.jp/swtips/R.html> 統計処理ソフトウェア R についての Tips (中澤 港)
- <http://aoki2.si.gunma-u.ac.jp/R/> R による統計処理 (群馬大学社会情報学部・青木繁伸教授)
- <http://www.okada.jp.org/RWiki/> RjpWiki (筑波大学・岡田昌史講師)
- <http://www.r-project.org/> R Project
- <http://cran.r-project.org/> CRAN
- <http://cran.md.tsukuba.ac.jp/> CRAN の筑波大学ミラー

第 15 章

【課題解答例】

第 1 章

`http://phi.med.gunma-u.ac.jp/msb/data/p01.xls` を読み込み、タブ区切りテキスト形式 `p01.txt` として保存し、R で `p01 <- read.delim("p01.txt")` としてから、`str(p01)` とすると、以下の出力が得られるので、オブザーベーションの数は 100 であり、各変数は、`pid` が `int` (整数型)、`sex` が "F", "M" という 2 水準をもつ `Factor` (要因型)、`ht` と `wt` が `num` (数値型) であることがわかる。

```
'data.frame': 100 obs. of 4 variables:
 $ pid: int  1 2 3 4 5 6 7 8 9 10 ...
 $ sex: Factor w/ 2 levels "F","M": 1 2 1 1 2 1 2 2 1 1 ...
 $ ht : num  165 169 160 163 172 ...
 $ wt : num  61.3 65.7 57.6 62.9 58.3 55.2 70.2 60.6 60.3 58.9 ...
```

さらに `summary(p01)` とすれば、以下が得られるので、NA の数までわかり、有効なサンプルサイズは、`pid` と `sex` が 100、`ht` が 99 (欠損 1)、`wt` が 98 (欠損 2) であることがわかる。

pid	sex	ht	wt
Min. : 1.00	F:50	Min. :150.6	Min. :45.60
1st Qu.: 25.75	M:50	1st Qu.:160.2	1st Qu.:58.08
Median : 50.50		Median :165.0	Median :61.95
Mean : 50.50		Mean :165.1	Mean :61.93
3rd Qu.: 75.25		3rd Qu.:169.7	3rd Qu.:66.40
Max. :100.00		Max. :181.3	Max. :76.60
		NA's : 1.0	NA's : 2.00

ただし、このデータは欠損が少ないので、1 つでも欠損があるケースは解析から除いてしまうのが筋である。その場合、

```
p01s <- subset(p01,complete.cases(p01))
```

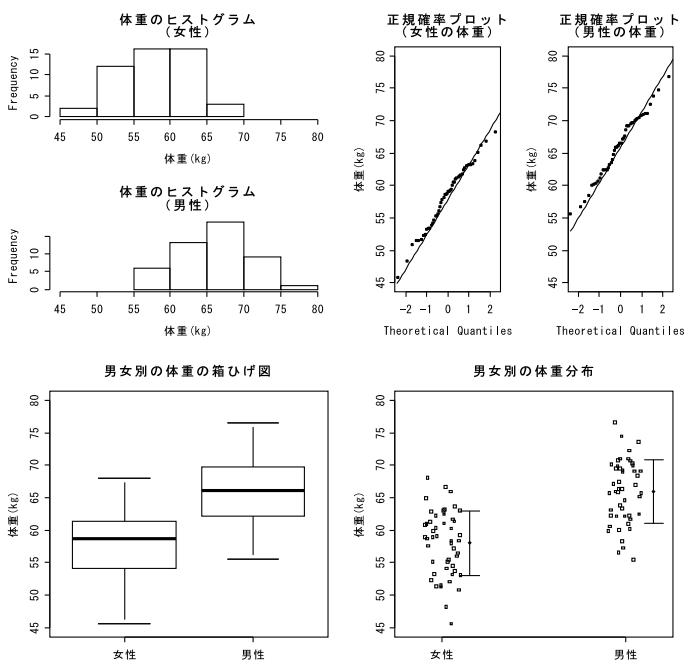
とすれば、欠損値を含まないサブセット `p01s` が得られるので、これに対して `str(p01s)` または `summary(p01s)` を行うことで、全体としての有効なサンプルサイズが 97 であるとわかる（後者の場合は男女別人数がわかるので、その合計でサンプルサイズがわかる）。もちろん、各変数の型はサブセットにする前と同じである。

第 2 章

体重は比尺度をもつ数値型の量的変数なので、分布を示すにはヒストグラム、正規確率プロット、箱ヒゲ図、ストリップチャートなどを用いる。とくに、男女間で分布を比較する目的なら、ヒストグラムや正規確率プロットより箱ヒゲ図やストリップチャートの方が見やすい場合もある。下枠内のコードを実行すれば、その下のグラフ群ができあがる。


```
c02a.R
```

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p01b.txt")
dat2 <- subset(dat, complete.cases(dat))
rm(dat)
attach(dat2)
mwt <- tapply(wt,sex,mean)
swt <- tapply(wt,sex,sd)
IS <- c(1,2)+0.15
layout(matrix(c(1,2,5,5,1,2,5,5,3,3,6,6,4,4,6,6),4,4))
hist(wt[sex=="F"],main="体重のヒストグラム△(女性)",
     xlab="体重 (kg)",xlim=c(45,80))
hist(wt[sex=="M"],main="体重のヒストグラム△(男性)",
     xlab="体重 (kg)",xlim=c(45,80))
qqnorm(wt[sex=="F"],main="正規確率プロット△(女性の体重)",
       ylab="体重 (kg)",ylim=c(45,80))
qqline(wt[sex=="F"],lty=2)
qqnorm(wt[sex=="M"],main="正規確率プロット△(男性の体重)",
       ylab="体重 (kg)",ylim=c(45,80))
qqline(wt[sex=="M"],lty=2)
levels(sex) <- c("女性","男性")
boxplot(wt~sex,main="男女別の体重の箱ヒゲ図",ylab="体重 (kg)",
        ylim=c(45,80))
stripchart(wt~sex,method="jitter",vert=T,main="男女別の体重分布",
           ylab="体重 (kg)",ylim=c(45,80))
points(IS,mwt,pch=18)
arrows(IS,mwt-swt,IS,mwt+swt,code=3,angle=90,length=.1)
detach(dat2)
```

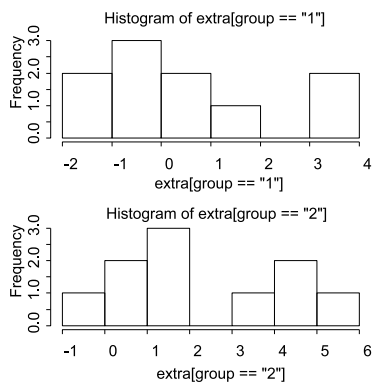


第 3 章

sleep データの分布の位置とばらつきの情報を与える記述統計量を計算する前に、図示により分布の様子を確認する。図示の方法はいろいろあるが、とりえずヒストグラムを作ってみる。

```
attach(sleep)
layout(c(1,2))
hist(extra[group=="1"])
hist(extra[group=="2"])
detach(sleep)
```

により下図ができ、正規分布とは到底言えなそうだとわかる。そこで位置とばらつきの情報としては、平均 \pm 不偏標準偏差よりもむしろ中央値 \pm 四分位偏差を示すべきと判断される。



```
SIQR <- function(x) { (fivenum(x)[4]-fivenum(x)[2])/2 }
attach(sleep)
tapply(extra,group,median)
tapply(extra,group,SIQR)
detach(sleep)
```

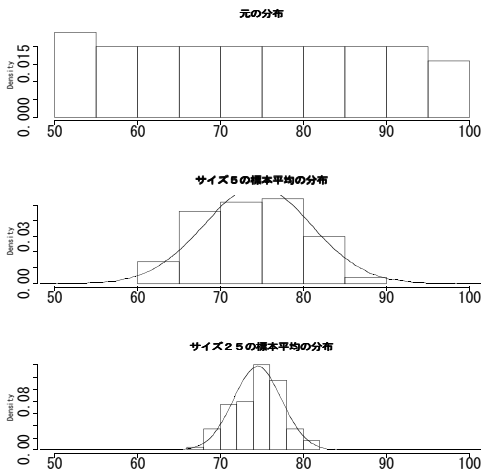
により (tapply は、1 番目の引数について、2 番目の引数ごとに層別して、3 番目の引数で与える関数を適用するという関数なので、group 別の統計量を計算する時に便利)、催眠薬 1 を投与したときの睡眠時間変化の中央値 ± 四分位偏差は 0.35 ± 1.1 、催眠薬 2 投与時の睡眠時間変化の中央値 ± 四分位偏差は 1.75 ± 1.8 とわかる。

第 4 章

中心極限定理を確かめてみるための課題であった。解答例を作るためのプログラムは以下の通りである。母集団のデータがすべてあるので、tsd として母集団での標準偏差を計算する関数を定義し、標本平均の分布が近づくであろう理論分布を赤い点線でヒストグラムに重ね描きするものである。

```
c04a.R
```

```
X <- rep(50:99,1000)
tsd <- function(XX) { sqrt(var(XX)*(length(XX)-1)/length(XX)) }
RNGkind("Mersenne-Twister")
set.seed(1)
layout(1:3)
hist(X,xlim=c(50,100),freq=F,main="元の分布")
Z5 <- rep(0,100)
for (i in 1:100) { Z5[i] <- mean(sample(X,5)) }
hist(Z5,xlim=c(50,100),freq=F,main="サイズ 5 の標本の平均の分布")
curve(dnorm(x,mean(X),tsd(X)/sqrt(5)),add=T,col="red",lty=2)
Z25 <- rep(0,100)
for (i in 1:100) { Z25[i] <- mean(sample(X,25)) }
hist(Z25,xlim=c(50,100),freq=F,main="サイズ 25 の標本の平均の分布")
curve(dnorm(x,mean(X),tsd(X)/sqrt(25)),add=T,col="red",lty=2)
```



標本サイズを大きくすると、標本平均の分布は、平均が母平均で標準偏差が母集団の標準偏差を標本サイズの平方根で割った値の正規分布に近づき、しかもそのばらつきが小さくなるのがわかる。

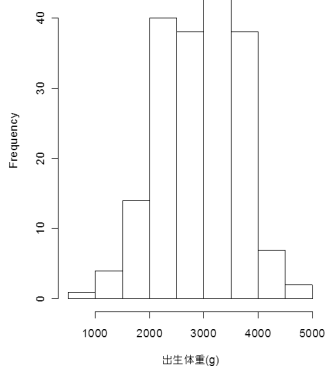
第 5 章

```
c05a.R
```

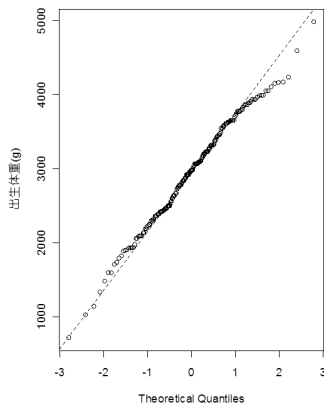
```
library(MASS)
attach(birthwt)
layout(t(1:2))
hist(bwt,main="出生体重のヒストグラム",xlab="出生体重 (g)")
qqnorm(bwt,main="出生体重の正規確率プロット",ylab="出生体重 (g)")
qqline(bwt,lty=2)
shapiro.test(bwt)
source("http://phi.med.gunma-u.ac.jp/msb/msb-funcs.R")
geary.test(bwt)
detach(birthwt)
```

結果は次のグラフと枠内の通り（情報量のない行は削除済み）。グラフを見るだけでも正規分布に近そうだと見当はつく。何度も繰り返すが、検定よりも先に作図をすることは非常に重要である。

出生体重のヒストグラム



出生体重の正規確率プロット



```

Shapiro-Wilk normality test
data:  bwt
W = 0.9924, p-value = 0.4354

Geary's test for normality:
G = 0.8126568 / p = 0.1693836

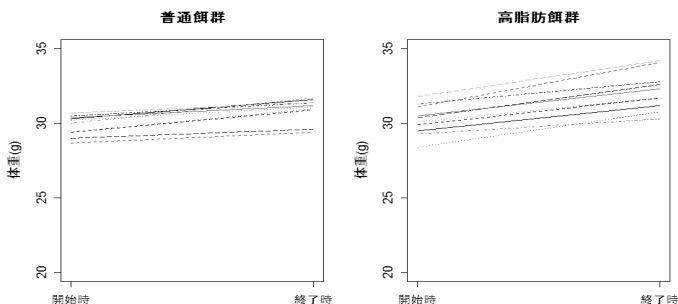
```

「出生体重データは正規分布にしたがう」という帰無仮説を，Shapiro-Wilk の検定と Geary の検定により，有意水準 5% で検定した結果，有意確率 (p 値) が 0.05 よりずっと大きいので，有意水準 5% で帰無仮説は棄却できない。よって，とりあえず正規分布にしたがっているとみなしてよい。

第 6 章

一見，対応のある検定かと思わせるが，2 群にランダムに割り付けたマウスの体重増加を比較するというデザインであり，普通餌摂取群の 1 匹目と高脂肪餌摂取群の 1 匹目には何も関係はないので，これは独立 2 標本の平均値の差の検定になる。

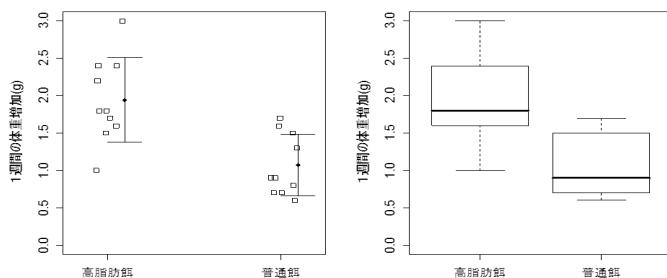
図示は，2 群別々に体重変化の折れ線グラフを作って並べると，情報量損失を最小限に抑えることができる。このグラフだと，開始時に，体重そのものに 2 群間で差があった可能性^{*1}がわかる一方，体重変化量を 2 群間で比較するには向いていない。



むしろ，まず 2 群別々に 1 週間の体重増加量を求め，下図のように，増加量を 2 群間で比較するストリップチャートや箱ヒゲ図を描く方が，この課題で検証すべき

*1 もしそうだとすると，ランダム割付がうまく行っていなかったということになる。その場合，データを救うのは難しいが，体重増加量ではなく体重増加率で検討するのは一案である。

「普通餌摂取群と高脂肪餌摂取群で体重変化に差があるかどうか」を見るには適している。



図では差がありそうに見える。続いて「普通餌摂取群と高脂肪餌摂取群で1週間の体重増加に差がない」という帰無仮説を検定してみる。母分散が未知の場合になるので、まず「2群の分散に差が無い」という帰無仮説を F 検定する。 $F = 0.5306$, num df = 9, denom df = 9, p-value = 0.3591 という結果から、有意水準 5% で帰無仮説は棄却されない(分散には有意差がない)ので、続いて分散に差がない場合の通り、通常の t 検定を実施する。結果として $t = -3.9148$, df = 18, p-value = 0.001015 が得られるので、有意水準 5% で帰無仮説は棄却される。以上より、普通餌群と高脂肪餌群では体重増加に有意な差があると判断できる。以上の作図と検定を実施するプログラムを示す。

c06a.R

```

dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p06.txt")
attach(dat)
layout(matrix(c(1,3,2,4),2,2))
matplot(rbind(rep(1,10),rep(2,10)),rbind(NDS,NDE),type="l",
  ylim=c(20,35),ylab="体重 (g)",main="普通餌群",xaxt="n",xlab="")
axis(1,1:2,c("開始時","終了時"))
matplot(rbind(rep(1,10),rep(2,10)),rbind(HFDS,HFDE),type="l",
  ylim=c(20,35),ylab="体重 (g)",main="高脂肪餌群",xaxt="n",xlab="")
axis(1,1:2,c("開始時","終了時"))
NDD <- NDE-NDS
HFDD <- HFDE-HFDS
WeightGain <- c(NDD,HFDD)
Diet <- as.factor(c(rep("普通餌",10),rep("高脂肪餌",10)))
IX <- c(1.1,2.1)
MWG <- tapply(WeightGain,Diet,mean)
SDWG <- tapply(WeightGain,Diet,sd)
stripchart(WeightGain~Diet,method="jitter",vert=T,ylim=c(0,3),
  ylab="1週間の体重増加 (g)")
points(IX,MWG,pch=18)
arrows(IX,MWG-SDWG,IX,MWG+SDWG,angle=90,code=3)
boxplot(WeightGain~Diet,ylim=c(0,3),ylab="1週間の体重増加 (g)")
print(res<-var.test(NDD,HFDD))
VAREQ <- ifelse(res$p.value < 0.05, FALSE, TRUE)
t.test(NDD,HFDD,var.equal=VAREQ)
detach(dat)

```

第 7 章

以下のようにして、まずデータを読み込み（1行目）、attachしてから（2行目）変数 VIL を要因型に変換する（3行目）ことで、分析の準備が整う。

```

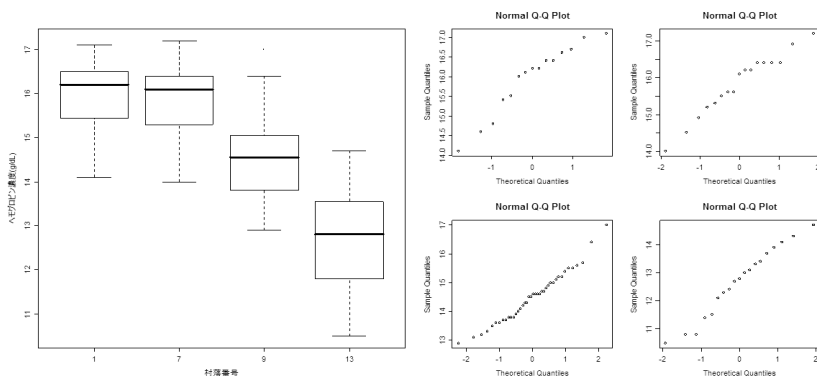
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p07.txt")
attach(dat)
VIL <- as.factor(VIL)

```

次に下枠内のように画面を 5 分割して、左半分には村落ごとの層別箱ヒゲ図を描き、

右半分の4区分に村落ごとの正規確率プロットを描く。層別箱ヒゲ図をみると、村落間でヘモグロビン濃度には差がありそうに見える。正規確率プロットはどれも概ね直線状に見える。

```
layout(matrix(c(1,1,1,1,2,4,3,5),nr=2))
boxplot(HB ~ VIL, ylab="ヘモグロビン濃度 (g/dL)", xlab="村落番号")
tapply(HB,VIL,qqnorm)
```



そこで以下のように、村落ごとに Shapiro-Wilk の検定をする。

```
tapply(HB,VIL,shapiro.test)
```

下枠内のように、どの村でもヘモグロビン濃度が正規分布に従っているという帰無仮説は棄却されない結果になる（関係部分のみ表示した）。

```

$'1' W = 0.9274, p-value = 0.2492
$'7' W = 0.9597, p-value = 0.6255
$'9' W = 0.975, p-value = 0.5104
$'13' W = 0.9637, p-value = 0.6463
```

そこで `bartlett.test(HB,VIL)` として、Bartlett の検定により、「村落間でヘモグロビン濃度の分散に差がない」という帰無仮説を検定すると、Bartlett's K-squared = 3.7251, df = 3, p-value = 0.2927 が得られるので、有意水準 5% で帰無仮説は棄却されない。

`summary(aov(HB~VIL))` により一元配置分散分析を行うと、下枠内の通り、村の

ヘモグロビン濃度への効果は有意である。

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
VIL	3	122.820	40.940	43.478	< 2.2e-16 ***
Residuals	87	81.920	0.942		

どの村とどの村の間で差があるかを検討するため、TukeyHSD(aov(HB~VIL))としてTukeyのHSD法により多重比較する。

	diff	lwr	upr	p adj
7-1	-0.1282353	-1.028646	0.7721758	0.9821751
9-1	-1.4425000	-2.212059	-0.6729415	0.0000248
13-1	-3.2663158	-4.144232	-2.3883998	0.0000000
9-7	-1.3142647	-2.050165	-0.5783645	0.0000616
13-7	-3.1380805	-3.986647	-2.2895139	0.0000000
13-9	-1.8238158	-2.532015	-1.1156170	0.0000000

より*², 1村と7村の間ではヘモグロビン濃度に有意差がないが, それ以外のすべての組み合わせで有意差があることがわかる。

第 8 章

(1) BMI と FAT の相関の分析*³

*² レポートなどに記載する値はこんなに長く書く必要はなく, 有効数字を考えて適切な小数以下数桁を記載すれば十分である。

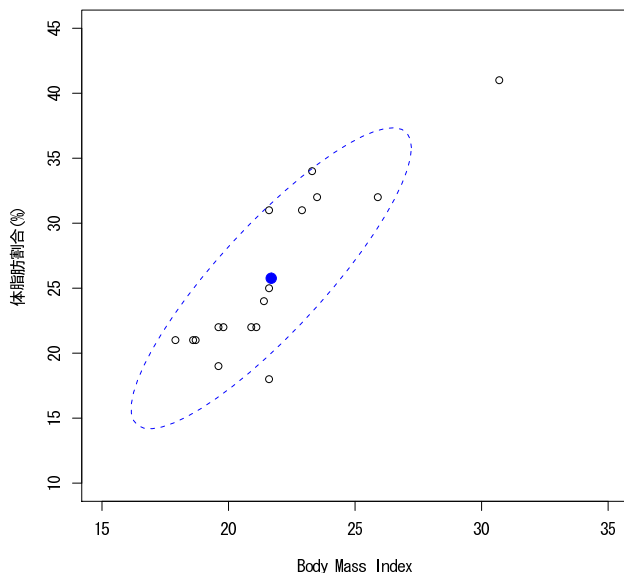
*³ シャピロ=ウィルクの検定の結果からすると, 実はBMIとFATは有意水準5%で正規分布とは言えないので, 本来は変数変換を検討するか順位相関係数を考えるべきである。しかしここでは, 変換しないでそのまま分析してみる。

```

dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p08.txt")
attach(dat)
plot(BMI,FAT,xlab="Body Mass Index",ylab="体脂肪割合 (%)",xlim=c(15,35),
      ylim=c(10,45),main="ソロモン諸島女性の BMI と体脂肪割合の関係と 80% 集中楕円")
require(car)
ellipse(c(mean(BMI),mean(FAT)),cov(cbind(BMI,FAT)),sqrt(qchisq(.8,2)),
        lty=2,lwd=1,col="blue")
cor.test(BMI,FAT)
cor.test(BMI,FAT,method="spearman")
cor.test(BMI,FAT,method="kendall")

```

ソロモン諸島女性のBMIと体脂肪割合の関係と80%集中楕円



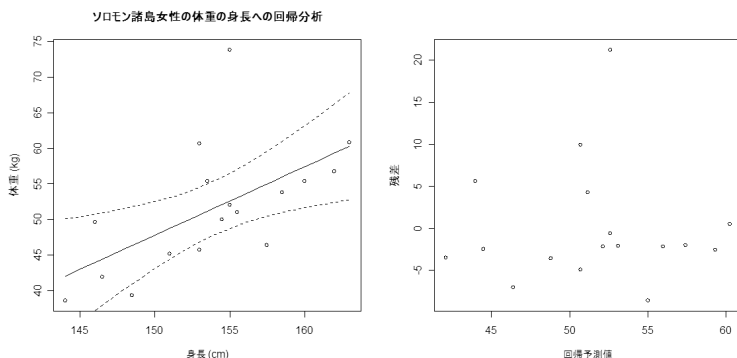
散布図をみると、BMI が大きい人はFAT も大きく、BMI が小さい人はFAT も小さい傾向があるようにみえるので、正の相関がありそうである。相関係数は、Pearson が 0.87 [0.66,0.95]、Spearman が 0.82、Kendall が 0.72 となり、それぞれゼロと差が無いという帰無仮説の検定で得られる有意確率も 10^{-6} から 10^{-5} のオーダーなので、有意水準 5% で帰無仮説は棄却される。相関係数の値そのものから考えて、強い正の相関があるといえる。

(2) 身長を独立変数，体重を従属変数とした回帰分析

```

layout(t(1:2))
plot(WT ~ HT,xlab="身長 (cm)",ylab="体重 (kg)",
      main="ソロモン諸島女性の体重の身長への回帰分析")
res <- lm(WT ~ HT)
LHT <- seq(min(HT),max(HT),length=20)
matlines(LHT,predict(res,list(HT=LHT),interval="confidence"),
         lty=c(1,2,2),col=0)
plot(residuals(res) ~ fitted.values(res),xlab="回帰予測値",
     ylab="残差")
summary(res)
predict(res,list(HT=155),interval="confidence")
detach(dat)

```



体重と身長の間にも直線的な関係がありそうにみえる。回帰分析の結果，

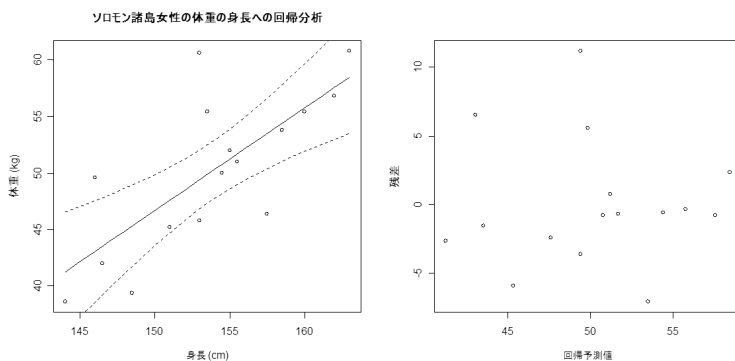
$$\text{体重} = 0.957 \times \text{身長} - 95.7$$

という回帰式が得られる。回帰係数がゼロと差が無いという帰無仮説の検定は有意確率が 0.0116 となるので棄却される。自由度調整済み相関係数の二乗は 0.312 となり、体重のばらつきの約 30% が身長にばらつきによって説明されるといえる。この程度の説明力の回帰式では予測に使うには不十分だが、強引に身長が 155 cm のときの体重の予測値を計算すると、52.6 [48.7, 56.5] kg となる。

ただし、ここで残差プロットをよくみると、1 人だけ大きな外れ値になっている人が見つかる。身長にのりに体重が極端に大きいこの人は、実は体脂肪割合も 41% あり、尿検査の結果、糖尿病であった。そのため、他の健康な人と同じ母集団からのサ

サンプルと考えるべきではない可能性がある。そこで、この人を除外して回帰分析をやり直してみる*4。

```
dat2 <- subset(dat, WT<70, drop=T)
attach(dat2)
res2 <- lm(WT ~ HT)
layout(t(1:2))
plot(WT ~ HT, xlab="身長 (cm)", ylab="体重 (kg)",
     main="外れ値を除くソロモン諸島女性の\n 体重の身長への回帰分析")
LHT <- seq(min(HT), max(HT), length=20)
matlines(LHT, predict(res2, list(HT=LHT)), interval="confidence"),
        lty=c(1,2,2), col=0)
plot(residuals(res2) ~ fitted.values(res2), xlab="回帰予測値",
     ylab="残差")
summary(res2)
predict(res2, list(HT=155), interval="confidence")
detach(dat2)
```



今度の回帰式は

$$\text{体重} = 0.906 \times \text{身長} - 89.2$$

となる。回帰係数がゼロと差が無いという帰無仮説の検定の結果、帰無仮説は有意水準 5% で棄却される (t 値=4.142, p 値=0.001)。自由度調整済み相関係数の 2 乗は 0.519 となり、今度は体重のばらつきの約半分が身長のみによって説明される

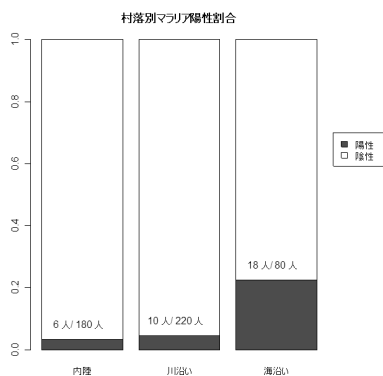
*4 ただし、このように他の情報により根拠づけられればよいが、たんに数値的に外れ値というだけでデータから除外してしまうのは危険である。

ことがわかる。これでもまだ予測に使うには不十分だが、1 回目の回帰分析よりも、かなり当てはまりは改善している。残差プロットも最初のものより均等にばらついているようにみえる。身長 155 cm のときの体重の推定値は、51.2 [48.6, 53.8] kg となる。推定値の 95% 信頼区間の下限はあまり変わらないが、最初より低めである。

第 9 章

下枠内のように、村別のマラリア原虫陽性者人数を変数 `malaria` に、検査総数を変数 `pop` に付値して、まず、「マラリア原虫陽性割合には村落間に差がない」という帰無仮説を検定する。

```
c09a.R
malaria <- c(6,10,18)
pop <- c(180,220,80)
names(malaria) <- c("内陸","川沿い","海沿い")
positive <- malaria/pop
negative <- 1-positive
tab <- rbind(positive,negative)
rownames(tab) <- c("陽性","陰性")
print(tab)
op <- par(mar=c(5,4,4,6.5)+0.1,xpd=NA)
ip <- barplot(tab,main="村落別マラリア陽性割合",col=c("red","white"))
legend(ip[3]+0.7,0.7,legend=rownames(tab),fill=c("red","white"))
text(ip-0.05,positive+0.05,paste(malaria,"人/",pop,"人"))
par(op)
prop.test(malaria,pop)
pairwise.prop.test(malaria,pop)
mosquito <- c(1,2,4)
prop.trend.test(malaria,pop,mosquito)
```



`prop.test()` の結果, 有意確率は 10^{-8} のオーダーなので帰無仮説は有意水準 5% で棄却される。

そこで, `pairwise.prop.test()` を実行すると, 2 村落のペアごとに「原虫陽性割合に差が無い」を帰無仮説とする検定の有意確率 (Holm の方法で検定の多重性を調整済み) は, 内陸と川沿いの間で 0.72, 内陸と海沿いの間で 8.0×10^{-6} , 川沿いと海沿いの間で 1.3×10^{-5} となる。つまり, 有意水準 5% で検定すると, 内陸と川沿いにはマラリア原虫陽性割合に有意差がなく, 海沿いと内陸, 海沿いと川沿いにはそれぞれ有意差があると判断される。

最後に, ハマダラカの相対的な密度のスコアを `mosquito` という変数に与え, この順に原虫陽性割合が大きくなっていく傾向があるかどうかをコクラン = アーミテジジ検定すると, $\chi^2 = 30.043$ で, 有意確率は 10^{-8} のオーダーなので, 対数オッズがスコアと比例して変化する傾向があるという対立仮説が採択される。つまり, ハマダラカの相対的な密度が高いほどマラリア原虫陽性割合が高い傾向が有意にあるといえる。

第 10 章

まず 55 歳未満と以上で別々にクロス表を入力し, Fisher の直接確率を計算する。ともに有意であり (55 歳未満 $p = 3.35 \times 10^{-11}$, 55 歳以上 $p = 3.01 \times 10^{-7}$), 55 歳未満でも 55 歳以上でもアルコール多量摂取と食道がん発生には関連があった。

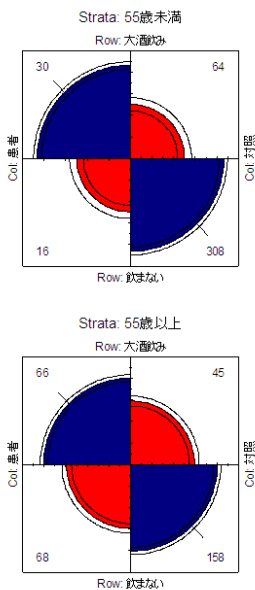
次に Woolf の検定で 3 次の交互作用がない (年齢層が違ってても関連の向きは同じ) という帰無仮説を検討すると, 有意水準 5% で帰無仮説は棄却された ($\chi^2 = 5.47$, $p = 0.019$)。したがって, 年齢層によってアルコール多量摂取と食道がん発生の関連

性は異なると考えられ、前提が満たされないのでマンテル=ヘンツェルの共通オッズ比は求められない*5。

年齢層別にオッズ比と 95% 信頼区間をみると、55 歳未満群で 8.96 (4.43, 18.7)、55 歳以上群で 3.39 (2.07, 5.63) であり、55 歳未満ではアルコール多量摂取すると食道がん罹患リスクが約 9 倍に上昇するのに比べ、55 歳以上では 3 倍余りにとどまっていた。Fourfold プロットを見ても、55 歳未満群の方がアルコール多量摂取がより強く食道がんリスクを高めると判断される。以上の解析をするコードは次の通り。

```
c10a.R
under55 <- matrix(c(30,16,64,308),nc=2)
fisher.test(under55)
over55 <- matrix(c(66,68,45,158),nc=2)
fisher.test(over55)
allbyage <- array(c(under55,over55),dim=c(2,2,2))
dimnames(allbyage) <- list(c("大酒飲み","飲まない"),
  c("患者","対照"),c("55 歳未満","55 歳以上"))
allbyage
library(vcd)
woolf_test(allbyage)
mantelhaen.test(allbyage)
fourfold(allbyage)
detach(package:vcd)
```

*5 それでも強引にやってしまうと、mantelhaen.test(allbyage) の結果の共通オッズ比は 4.6 で 95% 信頼区間は (3.13, 6.76) である。



第 11 章

以下のようにして、まずデータを読み込み (1 行目), attach してから (2 行目) 地域別出生児数分布表を作り (3 行目), 地域別の出生児数別カップル数を別々の変数に保管する (4 行目-6 行め) ことで, 分析の準備が整う。

```
c11a.R(1)
dat <- read.delim("http://phi.med.gunma-u.ac.jp/msb/data/p11.txt")
attach(dat)
X <- table(GRP,PARITY)
HF <- X[1,]
MF <- X[2,]
PF <- X[3,]
```

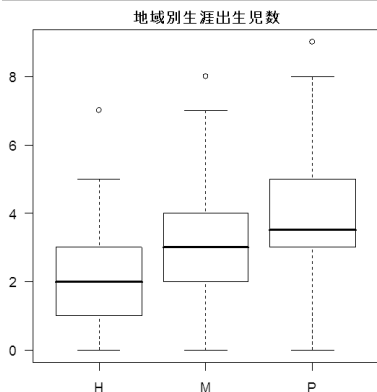
次いで, グラフを描く。個別に barplot(HF) などをした方がよいが, より簡便には, boxplot(PARITY~GRP) により, 1つのグラフィック枠に3地域を層別して箱ヒゲ図が描かれる。外れ値があることからノンパラメトリックな比較を考え, 次に

Fligner-Killeen の検定を行う。 $\chi^2 = 0.75, p = 0.69$ より、3 地域間でばらつきが均質であるという帰無仮説が棄却されないので、次に Kruskal-Wallis の検定を行う。 $\chi^2_{KW} = 11.2, p = 0.0036$ より、出生児数の分布の位置母数に 3 地域で差がないという帰無仮説は棄却される。つまり少なくともどこかの 2 地域間で差があることがわかる。

最後に、帰無仮説族 $\{H \text{ と } M \text{ に差がない}\}, \{M \text{ と } P \text{ に差がない}\}, \{P \text{ と } H \text{ に差がない}\}$ を検定するため、Holm の方法で検定の多重性を調整したウィルコクソンの順位和検定を `pairwise.wilcox.test()` により行うと、調整済み有意確率が、H 市と M 村の比較で 0.0251, H 市と P 村で 0.0054, M 村と P 村で 0.4030 となるので、H 市のカップルの生涯子供数は、M 村とも P 村とも有意水準 5% で統計的有意差があるが、M 村と P 村では生涯子供数に統計的有意差はないといえた。

```
c11a.R(2)
```

```
win.metafile("it11-ans-2006-1.emf",width=6,height=6,pointsize=14)
par(family="sans",mai=c(0.4,0.4,0.4,0.4),las=1)
boxplot(PARITY~GRP,main="地域別生涯出生児数")
dev.off()
fligner.test(PARITY~GRP)
kruskal.test(PARITY~GRP)
pairwise.wilcox.test(PARITY,GRP,exact=F)
```



なお、分布をみるには、

```
layout(1:3)
tapply(PARITY,GRP,hist,xlim=c(0,10),breaks=0:10,main="",right=F)
```

として3地域別々のヒストグラムを描かせるか (right=F として区間の右端を入れないことが重要。また, tapply() の4番目以降の引数は, 3番目の引数である関数にそのまま渡される), あるいは子供数は離散値なので, 以下のように棒グラフにしてもよい。

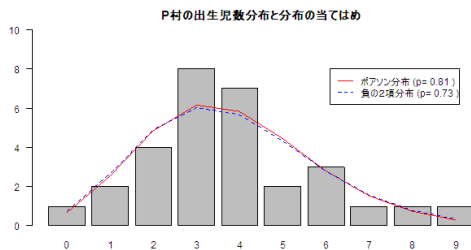
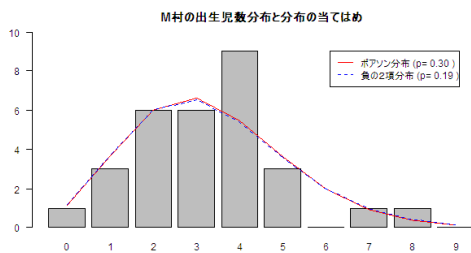
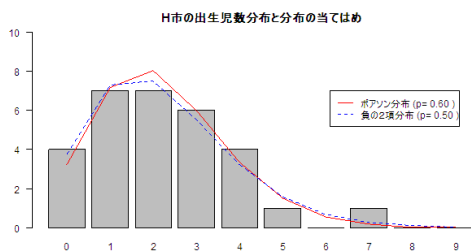
```
layout(1:3)
barplot2 <- function(...) { barplot(table(...)) }
tapply(PARITY,GRP,barplot2)
```

最初のところをもう少し丁寧に分析するには, 以下のように, 地域別に棒グラフを描き, そこに既知の分布を当てはめてみるとよい。子供数の分布については, 自然出生集団では Poisson 分布が, 意図的な出産抑制をしている集団では負の2項分布 (1回につき確率 p で成功する一連のベルヌーイ試行について, 成功が x 回起こるまでの失敗数の分布) が当てはまると言われているので, 両方を試してみる。下枠内のように手計算もできる (H市についてポアソン分布を当てはめた例) が, 第10章で紹介したように vcd ライブラリの goodfit() 関数を使うと, より簡便である。

```
H <- rep(0:9,HF)
ix <- barplot(HF,main="H市の出生児数分布と分布の当てはめ",ylim=c(0,10))
lines(ix,EH<-dpois(0:9,mean(H))*30)
print(mean(H))
print(XH <- sum((HF-EH)^2/EH))
1-pchisq(XH,8)
```

具体的には下枠内のコードで実行できる。検定結果も凡例の形でグラフに書き込んでみた。どの地域においてもポアソン分布が適合しているという帰無仮説も, 負の2項分布が適合しているという帰無仮説も棄却できない結果となった。しかしパラメータは互いに違いがありそうに見えるので, この後で, 最初に示したように Flinger-Killeen, Kruskal-Wallis, 多重性の調整付き Wilcoxon の順位和検定, と進むのがよい。

```
c11a.R(3)
H <- rep(0:9, HF); M <- rep(0:9, MF); P <- rep(0:9, PF)
library(vcd)
win.metafile("it11-ans-2006-2.emf", width=8, height=12, pointsize=14)
par(family="sans", mai=c(0.4, 0.4, 0.4, 0.4), las=1, mfrow=c(3, 1))
XHP <- goodfit(H, "poisson"); SXHP <- summary(XHP); TXHP <- paste("ポア  
ソン分布 (p=", sprintf("%.2f", SXHP[3]), ")")
XHN <- goodfit(H, "nbinom"); SXHN <- summary(XHN); TXHN <- paste("負の  
2項分布 (p=", sprintf("%.2f", SXHN[3]), ")")
ix <- barplot(HF, main="H市の出生児数分布と分布の当てはめ", ylim=c(0, 10))
lines(ix, predict(XHP, newcount=0:9), lty=1, col="red")
lines(ix, predict(XHN, newcount=0:9), lty=2, col="blue")
legend(8, max(HF), lty=c(1, 2), legend=c(TXHP, TXHN), col=c("red", "blue"))
XMP <- goodfit(M, "poisson"); SXMP <- summary(XMP); TXMP <- paste("ポア  
ソン分布 (p=", sprintf("%.2f", SXMP[3]), ")")
XMN <- goodfit(M, "nbinom"); SXMN <- summary(XMN); TXMN <- paste("負の  
2項分布 (p=", sprintf("%.2f", SXMN[3]), ")")
ix <- barplot(MF, main="M村の出生児数分布と分布の当てはめ", ylim=c(0, 10))
lines(ix, predict(XMP, newcount=0:9), lty=1, col="red")
lines(ix, predict(XMN, newcount=0:9), lty=2, col="blue")
legend(8, max(MF), lty=c(1, 2), legend=c(TXMP, TXMN), col=c("red", "blue"))
XPP <- goodfit(P, "poisson"); SXPP <- summary(XPP); TXPP <- paste("ポア  
ソン分布 (p=", sprintf("%.2f", SXPP[3]), ")")
XPN <- goodfit(P, "nbinom"); SXPN <- summary(XPN); TXPN <- paste("負の  
2項分布 (p=", sprintf("%.2f", SXPN[3]), ")")
ix <- barplot(PF, main="P村の出生児数分布と分布の当てはめ", ylim=c(0, 10))
lines(ix, predict(XPP, newcount=0:9), lty=1, col="red")
lines(ix, predict(XPN, newcount=0:9), lty=2, col="blue")
legend(8, max(PF), lty=c(1, 2), legend=c(TXPP, TXPN), col=c("red", "blue"))
dev.off()
```



第 12 章

必要な計算をするプログラムは以下の通りである。変数選択は必要ない。

c12a.R

```

NagelkerkeR2 <- function(rr) {
  print(n <- nrow(rr$model))
  (1-exp((rr$dev-rr$null)/n))/(1-exp(-rr$null/n))
}
require(MASS)
table(bacteria$y,bacteria$week)
table(bacteria$y,bacteria$trt)
res <- glm(y ~ week+trt, family=binomial, data=bacteria)
NagelkerkeR2(res)
summary(res)
exp(coef(res))
exp(confint(res))

```

この課題で立てるモデルは、菌の検出の有無を従属変数、週数と処置を独立変数とするロジスティック回帰である。週数は数値のまま入れ、影響を調整すべき共変量として扱えばよい。処置は3水準あるが、プラセボをリファレンスにしたときにコンプライアンスの悪い服薬群とコンプライアンスの良い服薬群で*6、それぞれどれくらい菌の検出が減るか（オッズ比がいくつになるか）を明らかにすることが目的である。ロジスティック回帰に先立ち、菌の検出の有無と週数、処置をそれぞれクロス集計した結果は次の通りである。

```

y / week
  0  2  4  6 11
n  5  4 11 11 12
y 45 40 31 29 32
y / trt
  placebo drug drug+
n      12  18  13
y      84  44  49

```

0週、2週では大半の子供に菌が検出されていたが4週以降になると検出されない子供が増えてくることがわかる。また、プラセボ群では大半の子供に菌が検出されるが、投薬群では菌が検出されない子供の割合が増えてきているようにみえる。

ロジスティック回帰分析の結果を次の表にまとめる。

*6 本当は、コンプライアンスの良い服薬群とコンプライアンスの悪い服薬群などというものができてしまった時点で、RCTとしては問題がある。

表. アボリジニの中耳炎乳児へのアモキシリン投与が *H. influenzae* の有無に与える影響のロジスティック回帰分析 *

独立変数	オッズ比	95% 信頼区間		p 値
		下限	上限	
プラセボ	1	—	—	—
アモキシリン投与, コンプライアンス低	0.331	0.140	0.752	0.009
アモキシリン投与, コンプライアンス高	0.521	0.215	1.253	0.144

Nagelkerke の R^2 : 0.095, AIC: 211.8, D_{null} : 217.4 (自由度 219), D : 203.8 (自由度 216)

* 投与週数の効果を共変量として調整した (偏回帰係数 -0.116 , $p = 0.009$)

アモキシリン投与かつコンプライアンスが低かった群では菌検出のオッズ比が有意に 1 より小さく, 約 1/3 に減少していた。コンプライアンスが高かった群でも菌検出のオッズ比は 1 より小さいが有意水準 5% で有意ではなかった。コンプライアンスが低い群の方が抗生物質の効きが良いと考えると不思議な結果である。

他の情報がないのでスペキュレーションしかできないが, コンプライアンスが低い群の方が効くのではなく, 良く効いたために少ししか飲まなくても症状が軽快し, 以後飲まなくなった, つまり効いた人の方がコンプライアンスが悪くなった可能性があるかもしれない。

第 13 章

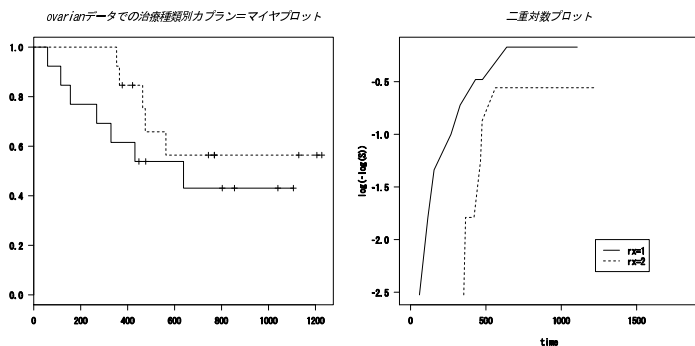
まず他の共変量を無視して, 治療種類ごとに Kaplan-Meier = マイヤ法で生存曲線を求め, 二重対数プロットもしてみるため, 下枠内を実行する。

```

c13a.R(1)
require(survival)
loglogplot <- function(X) {
  S <- X$surv
  T <- X$time
  G <- X$ntimes.strata
  GG <- names(X$strata)
  GX <- rep(GG,G)
  xr <- c(0,max(T)*1.5)
  yr <- c(log(-log(max(S))),log(-log(min(S))))
  plot(T[GX==GG[1]],log(-log(S[GX==GG[1]])),type="l",lty=1,
       xlim=xr,ylim=yr,xlab="time",ylab="log(-log(S))",
       main="二重対数プロット")
  for (i in 2:length(GG)) {
    lines(T[GX==GG[i]],log(-log(S[GX==GG[i]])),lty=i)
  }
  legend(max(T),-2,legend=GG,lty=1:length(GG))
}
#
print(KM <- survfit(Surv(futime,fustat)~rx,data=ovarian))
pdf("p13a.pdf",width=16,height=8,pointsize=14)
par(family="Japan1GothicBBB",las=1,mfrow=c(1,2))
plot(KM,
     main="ovarian データでの治療種別カプラン = マイヤプロット",
     lty=1:2)
loglogplot(KM)
dev.off()

```

すると、カレント作業ディレクトリに p13a.pdf という名前の pdf 形式のファイルとして、下図ができる。治療種類 2 の方が治療種類 1 よりも生存時間が長いことがわかる。カプラン = マイヤ推定による生存時間の中央値は、治療種類 1 では 638 日 (95% 信頼区間は 268 日から無限大)、治療種類 2 では無限大 (95% 信頼区間は 475 日から無限大) である。また、二重対数プロットは、ほぼ平行に見える。



そこで、次の枠内のように、治療種類、年齢、残留疾病の有無を共変量としたコックス回帰を実行する。モデルのデータのばらつきの半分弱を説明し ($R^2 = 0.475$)、有意に当てはまっているが (Rao のスコア検定で $\chi^2 = 20.8, d.f. = 3, p = 0.00012$ など)、生存時間に対して有意に影響しているのは年齢だけであり (ハザード比 1.137, 95% 信頼区間は 1.036-1.25, $p=0.0066$)、残留疾病の有無と治療種類の影響は有意ではなかった。

c13a.R(2)

```
res <- coxph(Surv(futime,fustat)~rx+age+factor(resid.ds),
  data=ovarian)
summary(res)
```

したがって、年齢と残留疾病の有無を共変量として調整したときに、この治療種類 1 と 2 の違いは、卵巣がん患者の生存時間に有意差をもたらすとはいえないという結論になる。

なお、<http://phi.med.gunma-u.ac.jp/msb/c13a2.R> を使うと、2 群別々に共変量として年齢と残留疾病の有無を調整してコックス回帰したときのベースラインハザードを使った二重対数プロットを描くことができるので参考にされたい。

コマンド索引

- < -, 14
- << -, 14
- .Renviron, 13
- .packages, 15, 223
- ;, 56
- ?, 10, 15
- [], 185

- abline, 119, 214
- abs, 63, 162, 163, 167
- agreementplot (vcd), 166
- AIC, 195, 199, 205, 206
- anova, 101, 194, 195
- Anova (car), 195
- aov, 101, 109, 195
- array, 177
- arrows, 42
- as.Date, 226
- as.factor, 145, 151
- as.integer, 23, 211
- as.matrix, 33, 166
- as.numeric, 23, 226
- as.ordered, 23, 25
- as.table, 144
- assocstats (vcd), 165
- attach, 20, 125, 199, 202, 204, 217
- attr, 205
- axis, 97

- barplot, 32, 33, 132, 142
- bartlett.test, 102
- binom.test, 132–134
- bonett.test (moments), 85
- boxplot, 41, 93, 95, 190

- c, 32, 138, 144
- capture.output, 40
- car, 115, 195
- cat, 139, 162–164
- cbind, 33
- chisq.out.test (outliers), 51
- chisq.test, 11, 148, 149
 - correct, 147
 - simulate.p.value, 148
- choose, 79, 153
- clogit (survival), 219
 - +strata, 220
- cm.colors, 46
- cochran.test (outliers), 51
- coef, 122, 197, 199, 217
- colnames, 33, 162, 163
- complete.cases, 19, 41, 197
- confint, 159, 164, 167
- coplot, 211
- cor, 113
- cor.test, 113–116
 - method=kendall, 117
 - method=spearman, 117
- cov, 113
- coxph (survival), 219, 235, 237, 239, 242
 - method=breslow, 243
 - strata, 239
 - subset, 237
- cumsum, 33
- curve, 59, 80–82, 130, 135

- data.ellipse (car), 115
- data.frame, 217
- dbinom, 79, 130, 132, 133
- dchisq, 81, 135
- detach, 15, 20, 164, 199, 202, 204, 217, 223
- df, 82
- difftime, 226
- dimnames, 177
- dixon.test (outliers), 51
- dnorm, 74, 79, 132
- dotchart, 36
- dpois, 136
- dt, 79, 82

- ellipse (car), 115
- Epi, 172
- exact
 - cor.test, 115
- example, 44, 191
- exp, 58, 126, 162, 163, 217

- extractAIC, 205, 206
- factor, 19, 92, 151, 216, 217, 219
- FALSE, 19
- fisher.test, 150–152, 159, 163, 164
- fitted.values, 126
- fiveum, 61
- figner.test, 189
- for, 33, 71
- fourfold (vcd), 176, 177
- friedman.test, 191
- function, 14, 162, 163, 166, 170, 199
- geary (moments), 85
- geary.test (本書で定義), 86
- getS3method, 114, 143
- glm, 193, 194, 204, 215, 217, 219
 - data=, 217
 - family=gaussian, 194
- goodfit, 143, 144
- grubbs.test (outliers), 51
- gstem (本書で定義), 40
- heat.colors, 46
- help, 195
- help.search, 10
- hist, 38, 70, 71, 74
- identify, 43
- ifelse, 78, 170, 219
- install.packages, 14, 143, 172
- integer, 19
- ISOdate, 226
- Kappa (vcd), 167
 - weights, 167
- kruskal.test, 189
- ks.test, 138, 180
- layout, 70, 71, 73, 142, 177
- legend, 214, 227
- length, 33, 40, 51, 84, 113, 170
- levels, 145, 151, 211, 217, 219
- library, 109, 143, 144, 152, 162, 195, 216, 223, 224
- lines, 74, 96, 132, 202
 - lty=, 202
- list, 166, 177, 202
- lm, 101, 119, 124, 194, 197, 199, 202, 204, 206, 214
- lm.ridge (MASS), 198
- lme (nlme), 191
- log, 58, 126, 206
- logical, 19
- logLik, 204, 205
- mantelhaen.test, 176
- maptools, 45
- mar.table (vcd), 152
- MASS, 38
- matlines, 119
- matplot, 43
- matpoints, 43
- matrix, 33, 149, 162–164
- max, 94, 170, 202, 214
- mean, 14, 16, 51, 202
 - na.rm=T, 202, 207
- median, 53
- median.test (本書で定義), 187
- methods, 114
- min, 202, 214
- mosaic, 176
- mosaicplot, 145
- NA, 18
- names, 32, 33, 142, 144
- nls, 194, 204, 208
 - start=list(), 208
- numeric, 19
- oddsratio (epitools), 159, 164
 - method, 164
- oddsratio (vcd), 159, 164
 - log=F, 159
- oddsratio.fisher (epitools), 164
- oneway.test, 102
 - var.equal, 102
- ordered, 19
- outer, 167
- p.value
 - var.test, 93
- pairs, 43
- pairwise.prop.test, 107, 140–142
- pairwise.t.test, 107
 - p.adjust.method=bonferroni, 107
 - p.adjust.method=holm, 107
- pairwise.wilcox.test, 107, 192
 - exact=F, 192
- par, 31, 237
 - las=1, 227
- paste, 33, 34
- pbinom, 133, 134
- pchisq, 81, 137, 138, 149, 204
- perm.test (exactRankTests), 188
- pf, 82, 91

- pi, 206
- pie, 37
- plot, 43, 96, 119, 121, 130, 144, 195, 214, 225, 237, 238
 - pch, 43
 - xaxt, 97
- pnorm, 80, 162, 163, 166
- points, 42, 43
- predict, 119, 201, 207
 - interval=conf, 202
- print, 140, 142, 144, 162, 163, 177
- prod, 58
- prop.test, 140–142, 149
- prop.trend.test, 141
- pt, 82, 84

- q, 14
- qbinom, 131, 133
- qchisq, 81, 115
- qf, 82
- qnorm, 80, 114, 133, 139, 162, 163, 166
- qqline, 39
 - lty, 39
- qqnorm, 39
- qt, 75, 82
- qtukey, 109
- quantile, 61

- rainbow, 46
- rank, 185
- rateratio (epitools), 162, 164
 - method, 162
- rbind, 142
- read.delim, 20, 226
- read.xls (xlsReadWrite), 17
- recode (car), 219
- relevel, 217, 219
- rep, 70, 145, 170, 202
- require, 143, 164, 199, 216, 217, 219, 223, 224
- residuals, 126, 195, 201
- riskratio (epitools), 162, 164
- RNGkind, 67, 70
- rnorm, 67, 90, 119
- roc (本書で定義), 170
- rocc (本書で定義), 171
- rownames, 33, 142, 162, 163
- RSiteSearch, 10
- runif, 119

- sample, 70
- scan, 171, 185
- sd, 14, 42, 64, 197, 199

- search, 15, 223
- seq, 130, 202
- set.seed, 67, 70
- shapiro.test, 84
- simtest (multcomp), 109
- sort, 170
- source, 14, 31, 140
 - echo, 140
- sqrt, 113, 115, 133, 139, 162, 163, 166, 170
- stars, 44
- stem, 40
- step, 195, 205–207, 217
- str, 19, 100
- stripchart, 41, 93, 95
- structable (vcd), 177
- subset, 19, 41, 197
- sum, 33, 51, 137, 138, 166, 170
- summary, 19, 101, 124, 144, 159, 164, 194, 195, 199, 214, 217, 225, 238
- Surv (survival), 224, 232, 233, 237, 239
- survdiff (survival), 232, 233, 237
 - rho, 232
- survfit (survival), 225, 230, 237, 238
- survreg (survival), 223
- symbols, 43
- Sys.getlocale, 226
- Sys.setlocale, 226


- t, 142
- t.test, 15, 76, 84, 90, 93, 95, 226
 - alternative, 93
 - paired, 96
 - var.equal, 92, 93
- table, 31, 32, 145, 148, 151
- table.margins, 152
- text, 33, 34, 40, 43
- topo.colors, 46
- TRUE, 19
- truehist (MASS), 38
- truemedian (本書で定義), 55
- TukeyHSD, 109

- unique, 170
- UseMethod, 143

- var, 28, 63, 113
- var.test, 91–93
- vif (DAAG), 198, 199
- VIF (本書で定義), 199

- which.max, 130
- wilcox.exact (exactRankTests), 183, 188
- wilcox.test, 96, 183, 185

paired=T, 188
woolf_test (vcd), 176



事項索引

- F* 検定, 91, 93, 179
F 比, 91
F 分布, 82, 91, 102, 210
 α エラー, 83
 α 係数, 26–28
 β エラー, 83
 χ^2 分布, 82
 κ 係数, 166, 167
t 検定, 15, 96, 104, 179, 180, 196
t 値, 123
t 分布, 75, 79, 82, 90, 96, 108, 113, 121, 123, 131, 197
 0 歳平均余命, 222
 2 × 2 クロス集計表, 145, 149, 186, 231
 2 × 2 分割表, 145, 150, 151
 2 項係数, 78
 2 項分布, 78–80, 130–133
 2 値変数, 145, 194, 209, 215, 216
 50 歳以上死亡, 156
 50 歳以上死亡割合, 156
 65 歳以上人口比率, 45
 95% 信頼区間, 131–133, 135, 139, 151, 159, 161–163, 166, 208, 215, 217, 227, 228, 230, 231, 237, 238, 241

 accuracy, 154
 AIC, 123, 195, 197, 200, 201, 205–208, 215, 239
 airquality (データフレーム), 125, 199
 aml (データフレーム), 230
 ANACOVA, 共分散分析
 ANOVA, 分散分析
 APC モデル, 172
 ArcExplorer, 45
 attributable proportion, 157
 attributable risk, 157
 AUC, 168, 170

 bacteria (データフレーム), 220
 BIC, 123
 birthwt (データフレーム), 87, 216
 BMI, 90, 128
 Bonett-Seier 検定, 85

 Bonferroni, 104
 Bonferroni の方法, 192
 Breslow 法, 242

 C.V., 64
 car (ライブラリ), 219
 case-control study, 158
 case-fatality rate, 156
 censored, 175
 censoring, 221
 centring, 198
 chickwts (データフレーム), 100
 Clopper and Pearson の方法, 132
 CMR, 129
 Cochran-Armitage 検定, 183
 cohort study, 158
 Cornfield の方法, 163
 correlation, 111
 CP932, 31
 CRAN, 11–15, 45, 109, 172, 194, 246
 cross-sectional study, 158
 cumulative incidence, 155
 CV, 変動係数

 DAAG (ライブラリ), 199
 Deviance, 123, 216
 difftime クラス, 226
 disease odds ratio, 157
 disease-odds, 155
 DIVA-GIS, 45

 Efron 法, 235, 243
 epitools (ライブラリ), 143, 152, 159, 162, 164
 EPS, 103
 ESRI, 45
 event history analysis, 221
 Exact 法, 163, 235, 243
 excess risk, 157
 exposure odds ratio, 157

 FALSE, 217
 Fisher の制約つき LSD 法, 104
 Fligner-Killeen の検定, 189

- follow-up study, 158
 for ループ, 71, 97
 Friedman の検定, 191
- Geary の検定, 85
 Gehan, 227
 gehan (データフレーム), 227
 Generalized Linear Model, 193
 GIS, 45
 Greville の方法, 222
 grid グラフィックス, 177
- Holm, 105
 Holm の方法, 192
 HSD, 104
- ICR マウス, 97
 incidence, 155
 incidence rate, 155
 incidence rate difference, 157
 incidence rate ratio, 157
 intercept, 119
 IQR, 四分位範囲
 IT 企業, 141
- Kaplan-Meier の積・極限推定量, 221
 Kendall の τ , 113
 Kendall の S 検定, 180
 Kruskal-Wallis の検定, クラスカル=ウォリスの検定, 189
 KS 検定, 138, 180
- leaps (ライブラリ), 207
 Lexis 図, 172
 linear, 117
- Mann-Whitney の U 検定, 92, 180
 MASS (ライブラリ), 205, 216, 227
 median, 41
 median-unbiased, 163
 Microsoft Excel, 184
 MLE, 203
 moments (ライブラリ), 85
 mortality rate, 155
 mortality rate ratio, 157
 multcomp (ライブラリ), 109
 multicollinearity, 198
- Nagelkerke の R^2 , 215, 217
 nlme (ライブラリ), 191
- odds, 155
 odds ratio, 157
- OpenOffice.org, 184
 Out of workspace, 164
 outliers (ライブラリ), 51
- p-value, 85, 158
 p-value 関数, 158
 paired- t 検定, 96
 Peritz, 105
 PMI, 156
 PMR, 156
 point prevalence, 154
 population at risk, 221
 precision, 154
 prevalence, 154
 proportional mortality indicator, 156
 proportional mortality rate, 156
 prospective study, 158
 proxy, 13
- Q1, 61
 Q2, 61
 Q3, 61
- Randomized Controlled Trial, 155
 rank sensitive, 53
 Rao の Score 検定, 237
 rate ratio, 157
 Rcmdr (ライブラリ), 20
 RCT, 155
 regression coefficient, 119
 relative risk, 157
 reliability, 154
 Repeated Measures ANOVA, 191
 residual, 122
 risk, 155
 risk difference, 157
 risk ratio, 157
 ROC, 168
 ROC 分析, 168
 R コマンダー, 20
- S4 メソッド, 205
 SAS, 195, 242, 243
 SBP, 128
 Scheffé, 104
 SD, 標準偏差
 SE, 標準誤差
 Shapiro-Wilk の検定, 84
 SIQR, 四分位偏差
 slope, 119
 Spearman の ρ , 113
 survival analysis, 221
 survival (ライブラリ), 219, 223, 224

- TFR, 94
 tie, 54
 ToothGrowth (データフレーム), 115
 TRUE, 217
 Tukey, 104
 Type III の平方和, 195
 Type II の平方和, 195
 Type IV の平方和, 195
 Type I の平方和, 195

 URL, 19

 validity, 154
 value sensitive, 53
 Variance Inflation Factor, 198
 vcd (ライブラリ), 143, 152, 164–167, 176
 VIF, 分散増加因子

 Welch の拡張による一元配置分散分析, 102, 110
 Welch の方法, 92, 93, 96, 102
 Wilcoxon の順位和検定, 92, 180, 184, 186, 189, 190
 Wilcoxon の符号付き順位和検定, 189
 Windows 拡張メタファイル, 102
 Woolf の検定, 176, 177
 Woolf の修正, 169
 workspace, 164

 xls, 17
 X 軸, 111

 Yates の連続性の補正, 11, 147, 148
 y 切片, 210

 アスコルビン酸, 115
 当てはまり, 122
 アルコール, 178
 アルゴリズム, 52, 53, 132
 アンケート, 141
 安定性, 123
 アンロード, 223

 意思決定, 83, 89, 113
 一元配置分散分析, 99–105, 110, 140, 189, 196, 257
 位置母数, 180, 190
 一樣乱数, 119
 一致度の判定基準, 167
 一般化, 203
 一般化ウィルコクソン検定, 222, 231
 一般化線型モデル, 123, 176, 193, 194, 196, 204, 239
 イベント, 221–223, 231

 イベント発生順位, 231
 イベント発生率, 221
 医療資源, 154
 因果関係, 124
 因果の向き, 124
 因子, 195, 196
 インストール, 143, 223
 インターネット, 19
 インデックス値, 201

 ウィリアムズ, 104
 ウィルコクソンの順位和検定, 180, 231
 ウィルコクソンの符号順位検定, 96
 ウェルチの方法, 92
 打ち切り, 57, 175, 223, 227, 231, 233
 打ち切りフラグ, 224, 232, 239
 うつのスクリーニング, 169

 疫学, 89, 153
 疫学研究, 158, 207
 エディティング, 15
 エラー, 164, 179
 エラーバー, 41, 94
 円グラフ, 37
 エンドポイント, 221

 オゾン濃度, 125
 オッズ, 155, 157, 158, 219
 オッズ比, 144, 151, 156–159, 161, 163, 215–217
 帯グラフ, 35
 オブザーベーション, 206
 重み, 167

 回帰, 111
 回帰係数, 119, 121–123, 141
 回帰式, 117, 121, 141, 200
 回帰直線, 111, 117, 119, 123, 124, 209, 210
 原点を通る, 117
 切片のある, 117
 回帰の外挿, 118, 208
 回帰分析, 121, 196
 回帰モデル, 201
 階級幅, 56
 回収率, 18, 94
 有効, 18
 階段関数, 112
 カイ二乗検定, 11, 146, 149, 150, 165, 180
 カイ二乗値, 137, 233
 カイ二乗適合度検定, 135, 137, 138, 140, 141, 146
 カイ二乗統計量, 147, 148, 165, 180, 231

- カイ二乗分布, 81, 82, 123, 134, 135, 137, 147,
 149, 180, 190, 191, 204, 231, 232
 科学的仮説, 112
 拡張期血圧, 198
 拡張子, 17
 拡張モザイクプロット, 176
 確率, 130, 131, 135, 139
 確率関数, 136
 確率構造, 146
 確率楕円, 111
 確率分布, 113
 確率変数, 50, 69, 79, 81, 234
 確率母関数, 79, 80, 90
 確率密度関数, 79–82, 85, 132, 135, 203
 下限, 114, 139, 179
 過小評価, 144, 158, 221
 仮説, 160
 仮説検定, 77, 89, 158
 河川工学, 124
 加速モデル, 223
 片側検定, 83, 90, 92, 113, 115
 型変換, 219, 226
 傾き, 119, 209, 210
 カットオフポイント, 168
 カテゴリー, 219
 カテゴリー, 219
 カテゴリー数, 134, 135
 カテゴリー変数, 19, 23, 24, 58, 129, 134, 138,
 143, 144, 153, 165, 194, 196
 カテゴリー別死亡率, 155
 カプラン=マイヤ推定, 224, 230
 カプラン=マイヤ推定量, 221, 223, 224, 227
 カプラン=マイヤプロット, 237
 カプラン=マイヤ法, 225, 227, 233
 間隔データ, 222
 観察打ち切り, 221, 222
 観察期間, 155, 221
 観察値, 135
 観察人年, 155, 162
 患者, 138, 139, 148
 患者数, 154
 患者対照研究, 症例対照研究, 144
 感受性, 53, 155
 関数定義, 171
 観測度数, 134, 136–138, 146
 官庁統計, 222
 感度, 168, 169
 幹葉表示, 40
 管理者権限, 143
 関連性, 153, 158
 関連性の指標, 153, 165
 関連の程度, 165
 関連の向き, 111
 外挿, 124
 外的基準, 141
 頑健, 53
 ガンマ関数, 81, 82
 幾何平均, 58
 期間, 154
 期間データ, 221
 棄却, 83, 102, 113, 135, 139, 149, 151, 237
 棄却限界, 107
 棄却楕円, 111
 危険因子, 156
 希釈, 117
 期首人口, 155, 158, 221
 記述統計量, 49
 基準値, 168
 基準ハザード関数, 233, 234
 期待死亡数, 231
 期待値, 69, 74, 128, 137, 143, 182, 183, 185,
 188
 期待度数, 134, 137, 138, 146, 149, 150
 期待日数, 137
 期待頻度, 138
 喫煙, 139, 141, 144, 148
 喫煙割合, 142
 帰無仮説, 82–84, 90, 91, 93, 100, 102, 105,
 113, 115, 121, 132, 134, 135,
 137–139, 141, 142, 144, 148–152,
 158, 159, 162, 166, 181, 183, 186,
 190, 197, 204, 207, 210, 216, 231,
 232, 237, 238, 241
 帰無仮説族, 105–108
 級間変動, 100, 101
 吸光度, 117, 121, 124
 急性感染症, 154
 共分散, 113
 共分散行列, 115
 共分散分析, 193, 196, 209, 210
 共変動, 113, 210
 共変量, 209, 211, 216, 218, 233, 234, 237,
 239, 241
 共有, 198
 極限, 80
 曲線下面積, 168, 170
 共通オッズ比, 176
 寄与危険, 157
 寄与率, 123
 寄与割合, 157
 近似, 132, 147, 150, 188
 ギャリーの検定, 85
 擬似相関, 112, 173
 逆関数法, 70

- 逆算, 117
- 逆変換, 126
- 行列, 119
- 行列言語, 195
- 偽陽性率, 168, 169
- ギリシャ文字, 50

- クイックソート, 53
- 区間, 54
- 区間打ち切り, 224
- 区間推定, 74, 89
- 組み合わせ, 130, 150, 153
- 組み込み済みのライブラリ, 223
- クラスカル=ウォリスの検定, 99, 102, 189
- クラメールの V, 165
- 繰り返し制御文, 71
- 繰り返し調査, 166
- 繰り返しのある分散分析, 191
- クロス集計, 24, 143, 144, 148, 150-152, 158, 159, 165, 166, 169, 173, 183, 217
- クロンバックの α 係数, α 係数

- 偶然, 150
- 偶然誤差, 154
- グラデーション, 46
- グラフィックデバイス, 31, 102
- グリーンウッドの公式, 224
- 群間分散, 級間分散
- 群間変動, 級間変動, 100, 101
- 群別変数, 95, 237
- 群分け変数, 93, 99, 189, 239

- ケースコントロール研究, 症例対照研究
- 経験的ロジスティック変換, 176
- 経験分布, 180
- 傾向, 141
- 警告メッセージ, 186
- 計算用メモリ, 164
- 経時変化のある測定値, 191
- 係数, 121, 193, 198
- 継続行, 13
- 系統的なズレ, 201
- 桁落ち, 63
- 血清鉄, 96, 121
- 欠損値, 17, 19, 41, 197, 206, 207
- 決定係数, 121-123, 197, 201, 215, 216
- 血液, 142
- 研究デザイン, 158
- 検査再検査信頼性, 166
- 検出限界, 18
- 検出力, 83, 104, 180
- 健診, 141

- 検定, 57, 76, 77, 82, 89, 113, 132, 139, 144, 149, 159, 162, 204, 209, 210, 231, 241
- 検定結果, 121
- 検定統計量, 84, 106-108, 113, 121, 181, 185
- 検定の多重性, 99, 110, 140, 189, 192
- 検量線, 117-119, 121, 122

- 原虫陽性, 142
- 限定, 112, 176

- コーディング, 15
- コード, 143
- コード表, 16
- コアチーム, 11
- 効果, 156
- 効果の指標, 156
- 高血圧, 154
- 高コレステロール血症, 154
- 交互作用, 193, 194, 211
- 高脂肪飢餓, 97
- 高周波, 160
- 厚生科学研究, 94
- 交通事故件数, 136, 137
- 交絡, 173, 175, 176, 207, 215
- 交絡因子, 217
- 効率, 180
- 国民栄養調査, 90
- コクラン=アーミテージの検定, 141
- コクラン=マンテル=ヘンツェル, 231
- コックス回帰, 219, 222, 223, 233, 235, 237-239, 241
- コホート, 222
- コホート研究, 144, 157, 158, 160, 165
- コホート生命表, 222
- コルモゴロフ=スミルノフ検定, 138, 180
- コロブレス図, 45
- コントロール群, 対照群

- 礫石, 129
- 合計出生率, 94
- 誤差, 193
- 誤差項, 123
- 誤差自由度, 108
- 誤差分散, 108, 109, 123
- 誤差変動, 100, 102
- 五数要約値, 61
- ゴセット, 82

- 再現性, 154
- サイコロ, 138
- 最小値, 60, 121
- 最小二乗法, 117, 118, 123, 124, 193
- 最小有意差法, 104

- 再生産, 192
 最大値, 60, 121
 最大尤度, 204, 205
 最適値, 168
 最頻値, 51, 57
 最尤推定, 159
 最尤推定量, 203, 221
 最尤法, 170, 193
 最尤方程式, 204
 最良線型不偏推定量, 84
 作業ディレクトリ, 13, 19
 サブセット, 41
 三元配置分散分析, 196
 三次曲線, 112
 算術平均, 50
 散布図, 43, 111, 117, 121, 126, 198, 234, 235
 サンプルング, 158
 サンプル, 標本, 17, 49, 71, 91, 160
 サンプルサイズ, 50, 69, 71, 114, 131, 139, 150, 164, 175, 192, 216, 232, 235

 残差, 120, 122, 201
 残差分散, 122, 123
 残差分析, 201, 206, 239
 残差プロット, 126, 195
 残差平方和, 122, 195, 210, 216

 死因別死亡率, 155
 死因別死亡割合, 156
 シェイプファイル, 45
 シェフェ, 104
 シェルソート, 53
 視覚的評価, 168
 シグモイド, 112
 指数, 126, 217
 指数関数, 114, 208
 指数分布, 223
 自然対数, 114, 204
 市町村コード, 45
 悉皆調査, 142
 疾病, 154
 疾病オッズ, 155, 157
 疾病オッズ比, 157, 159, 160
 四分位数, 60
 四分位範囲, 41, 60–62, 92
 四分位偏差, 60–62, 64, 92, 250, 251
 死亡, 155
 死亡数, 156
 死亡率, 155
 死亡率比, 157
 シミュレーション, 70, 148, 149, 191
 社会福祉資源, 154
 尺度
 間隔, 23
 順序, 23
 比, 23
 名義, 23
 シャピロ=ウィルクの検定, 125
 試薬ブランク, 117
 集合論, 50
 収縮期血圧, 128, 198
 修正平均, 209–211
 集団, 154
 集中楕円, 111, 115
 周辺度数, 150, 152, 159
 主効果, 194
 出生率, 94
 出生性比, 133, 135
 出現順, 201
 出現頻度, 138
 瞬間イベント発生率, 233
 生涯子供数, 192
 少子化, 94
 小数点記号, 226
 将来予測, 124
 症例, 157, 158
 症例対照研究, 144, 148, 155, 157, 158, 160, 161, 219
 初期値, 67, 208
 食道がん, 178
 死力, 222
 診断基準, 154
 身長, 128
 真の中央値, 55
 シンプソンのパラドックス, 174
 信頼区間, 75, 90, 111, 114, 117, 119, 128, 131, 132, 158, 161–163, 167, 201, 207
 信頼性, 27, 131, 154, 156

 時間依存性共変量, 233
 時系列, 112
 次元, 155
 事象生起確率, 141
 実線, 39
 実測値, 120
 時点の重み, 231
 ジャックナイフ, 170
 重回帰式, 207
 重回帰分析, 193, 197, 200, 206
 重回帰モデル, 197–199, 206, 207
 重心, 46, 115, 123
 重相関係数, 194, 197, 198, 207, 215, 239
 自由度調整済み, 199, 201
 従属変数, 111, 117, 121, 123, 124, 128, 141, 176, 193, 194, 197, 198, 207, 215–217
 重篤度, 156

- 受信者動作特性曲線, 168
- 受診者動作特性曲線, 168
- 順位, 180, 190
- 順位和, 181, 190, 191
- 順序, 164
- 順序型, 19
- 順序尺度, 129
- 順序統計量, 84
- 純水, 117
- 自由度, 63, 75, 76, 81, 82, 93, 96, 102, 108, 113, 123, 134, 135, 137, 146, 147, 149, 190, 191, 231, 232
- 自由度調整済み, 194
- 自由度調整済み相関係数の二乗, 121
- 条件付き確率, 69
- 条件付ロジスティック回帰分析, 219, 220
- 上限, 114, 139, 179
- 上側確率, 83, 90, 96
- 情報量, 158
- 人口, 155
- 人口学, 222
- 人口統計, 156
- 人口ピラミッド, 38
- 推奨ライブラリ, 223
- 水準, 211, 216, 217, 219
- 推定, 74, 129, 135, 201
- 推定値, 123, 130
- 推定量, 203
- 数学的に等価, 181
- 数値型, 19, 51, 216, 219, 226
- スクリーニング, 168
- スコア, 141, 186, 231
- ステップダウン法, 105
- ステップワイズ, 217
- ステューデント化された範囲の分布, 108
- ステューデントの t 分布, 81
- ストリップチャート, 41
- スピアマン・ブラウンの公式, 27
- ズレのモデル, 180
- 正確さ, 154
- 正確な確率, 114, 115, 186, 188, 192
- 生活習慣, 158
- 生起確率, 77
- 正規確率プロット, 39
- 正規近似, 114, 133, 139, 149, 163, 180, 181, 183, 185, 186
- 正規分布, 39, 59, 74, 79, 80, 84, 89, 102, 108, 110, 123, 125, 129, 131, 132, 139, 161, 163, 179, 194, 204, 215, 235
- 正規乱数, 67, 95, 119
- 生残確率, 228
- 静止人口, 222
- 整数型, 19, 51
- 生存関数, 234, 237
- 生存曲線, 222, 224, 227, 228, 230, 231, 237, 238, 241
- 生存時間, 221, 230–232, 234, 235, 237, 239
- 生存時間解析, 155, 219, 221
- 生存時間型, 224
- 生存時間の差の検定, 222
- 生存時間分布, 235
- 生態学的条件, 110
- 静態生命表, 222
- 精度, 154
- 正二十面体サイコロ, 78
- 性・年齢別死亡率, 155
- 正の相関, 111
- 生命表, 222
- 生命表解析, 222
- 世界人口, 124
- 折半法, 27
- 切片, 119, 121–123, 193, 194, 200
- 説明変数, 111
- 説明力, 121, 124
- セミパラメトリック, 234
- 線型, 198
- 線型回帰, 117, 141, 215
- 線型混合効果モデル, 191
- 線型重回帰, 204
- 線型重回帰モデル, 194
- 線型モデル, 123, 196
- 先行研究, 83
- 線種, 39
- 尖度, 85
- 絶対値, 62, 83, 123, 139
- 絶対モーメント, 85
- ゼロ歳平均余命, 222
- ゼロ点調整, 117
- 全国市町村境界データ, 45
- 全死亡, 156
- 全体, 158
- 総当り, 181
- 相関, 111, 128, 173, 198, 208, 209
- 相関関係, 111
- 相関係数, 20, 27, 112, 113, 115–117, 122, 165, 259–261
- ケンドールの順位, 113–115
- 順位, 114, 180, 258
- スピアマンの順位, 113–115
- 属性, 165
- ピアソンの積率, 112–116

- 相対危険, 157, 158
 相対差, 157
 相対度数, 69
 相対密度, 142
 送電線, 160
 総平均, 52
 層別, 41, 234
 層別因子, 239
 層別化, 112
 層別解析, 158, 176
 層別箱ヒゲ図, 180, 190
 測定限界, 117
 測定誤差, 64, 111, 124
 測定精度, 64
 ソロモン諸島, 128

 象牙芽細胞, 115

 対応がある場合の図示, 96
 対応のある多群, 191
 対応のある t 検定, 96, 187
 対応のある 2 標本, 96
 大気環境データ, 125, 199
 体脂肪割合, 128
 対照, 138, 139, 148, 149, 157, 158
 体重, 128
 対数, 126, 215, 234
 対数オッズ比, 159, 215, 217
 対数正規分布, 58
 対数線型モデル, 223
 対数変換, 112, 126, 161, 163, 179
 対数尤度, 204
 対立仮説, 83, 84, 91, 93, 115, 189, 210
 多群, 110
 多群間の分布の位置の差, 189
 多元配置分散分析, 196
 多重共線性, 198, 199
 多重比較, 99, 100, 104, 105, 109, 142, 192
 縦軸, 121, 201
 タブ区切りテキスト形式, 17
 単回帰, 208
 単純ソート, 53
 単調変換, 180

 第 1 四分位, 41, 60, 121
 第 1 種の過誤, 83, 99, 104, 142, 152, 192
 第 3 四分位, 41, 60, 121
 大小関係, 180, 186
 代数, 50
 第 2 種の過誤, 83, 152
 代表性, 18, 57
 代表値, 49
 脱落, 155

 妥当性, 154
 ダネット, 104, 109
 ダミー変数, 19, 25, 196, 239
 ダミー変数化, 216
 ダンカン, 104
 団体コード, 45
 断面研究, 145, 157, 158

 地域情報, 226
 地域相関, 112
 逐次棄却型検定, 105
 逐次平方和, 195
 地図情報, 45
 致命率, 132, 156
 中央値, 41, 51, 52, 61, 64, 92, 121, 221, 230
 抽出, 67
 中心極限定理, 64, 73, 74, 89
 中心傾向, 49, 51
 中心性, 51
 中点, 54
 超過危険, 157
 超幾何分布, 150, 186
 調整, 193, 218
 調整平均, 209, 211
 調和平均, 58
 直線的, 112
 直交, 195

 追跡調査, 144, 158, 160
 通貨記号, 226

 定義域, 24, 79, 80
 定数, 205
 低体重出生, 216
 適合, 210
 適合度, 117, 241
 適合度検定, 143
 手続き, 52
 テューキー, 104
 テューキーの HSD, 108, 110
 点推定量, 74, 121, 131, 132, 158, 159, 161, 162, 217

 データ数, 129
 データフレーム, 19, 119, 194, 216, 217
 デザイン, 144, 153, 175

 統計学的に有意, 83
 統計資料, 156
 統計的に有意な関連, 146
 統計量, 50, 57, 67, 191
 特異度, 168, 169
 途上国, 156

- ドーナツグラフ, 37
 同時散布図, 43
 同時点, 224
 同順位, 54, 181, 183, 185, 186, 190
 毒物, 132
 独立, 82, 136, 138, 144, 153
 独立性のカイ二乗検定, 140, 148
 独立性の検定, 146, 150
 独立 2 標本, 96
 独立変数, 111, 117, 121, 123, 124, 128, 141, 176, 193, 194, 197-199, 201, 205-208, 216, 217, 219, 223, 237
 度数, 57, 144
 度数分布, 134, 219
 度数分布図, 32
 度数分布表, 55
 ドットチャート, 36

 内部処理, 143
 生データ, 148
 並べ換え, 52
 並べ換え検定, 188, 190

 二元配置分散分析, 109, 194-196
 二次曲線, 112
 二次元正規分布, 112
 二重対数プロット, 235, 237, 241
 二乗和, 63, 122
 二峰性, 57, 219
 日本語文字コード, 31
 日本語ロケール, 226
 入力フォーム, 16

 塗り分け地図, 45

 年央人口, 156, 222
 年齢, 156
 年齢階級, 55
 年齢 2 区分, 156
 年齢別死亡率, 222
 年齢別人口, 45

 濃縮, 117
 濃度, 117, 121
 延べ生存期間, 222
 ノンパラメトリック, 92, 102, 105, 112, 179, 180, 189, 190, 192, 221, 222
 ノンパラメトリックな解析, 180

 肺がん, 144, 148, 149
 ハザード, 155, 223
 ハザード関数, 222, 233, 234
 ハザード比, 233, 234, 237, 238, 241

 外れ値, 51, 57, 60, 179
 外れ値の検定, 51
 破線, 39
 波長, 121
 白血病, 160, 165
 発生数, 155
 発生率, 144, 155
 ハマダラカ, 142
 範囲, 60
 反応変数, 215

 パートレットの検定, 102, 110, 189
 バイアス, 51, 148, 154, 170
 曝露, 154, 156-158, 161, 221
 曝露オッズ, 155, 157, 160
 曝露オッズ比, 157, 159
 バブルソート, 53
 ばらつき, 49, 60, 121
 ばらつきの同等性, 189

 パーセンタイル, 61
 バプアニューギニア, 50, 62, 110, 142
 高地辺縁部, 141
 パラメータ, 123, 179, 235
 パラメータ数, 204, 205
 パラメトリック, 179, 180, 222, 223, 235
 パラメトリックモデル, 234

 非該当, 17, 18
 引数, 164
 非心超幾何分布, 159
 非心度パラメータ, 159
 ヒストグラム, 38, 70, 125
 非線型回帰, 208
 非線型モデル, 194, 204
 左側打ち切り, 224
 日付, 226
 日付形式, 226
 非曝露, 157
 非復元抽出, 130, 150
 非復元抽出, 69
 皮膚疾患, 141
 評価者, 166
 評価者間一致度, 166
 表計算ソフト, 17, 184
 表示単位, 54
 標準化, 139, 156, 183, 188
 標準化偏回帰係数, 193, 197, 198, 200
 標準希釈系列, 121
 標準誤差, 64, 69, 74, 121, 224, 231
 標準正規分布, 74, 80, 81, 89, 91, 139, 166, 182, 183, 188
 標準物質, 117

- 標準偏差, 14, 19, 28, 41, 59, 60, 64, 69, 73,
74, 80, 86, 89, 92, 95, 132, 193, 205,
251, 252
不偏, 27, 60, 64, 71, 74, 75, 94, 197, 250
標本, 49, 89, 129, 179
標本サイズ, 52, 63
標本抽出, 17, 67, 70, 73
標本統計量, 67
標本比率, 129, 130, 139
標本分布, 71
標本平均, 50, 69, 74, 89, 90, 204
比率, 131
比率の差の検定, 138
比例定数, 223
比例ハザード性, 223, 233–235, 237
比例ハザードモデル, 233–235
頻度, 52
頻度の指標, 154, 156
- ピアソンのコンティンジェンシー係数, 165
ピアソンの積率相関係数, 相関係数
ピアソンの相関係数, 相関係数
- ファイ係数, 165
フィッシャーの正確な確率, 149–151, 179
フィッシャーの直接確率, 149, 150
風速, 125
フォローアップ, 175
フォローアップ研究, 158
復元抽出, 69, 77, 130
符号化順位検定, 187
符号検定, 180
符号付き順位和検定, 187
符号付順位和検定, 180
付値, 19, 51
普通餉, 97
負の相関, 111
フリードマン (Friedman) の検定, 191
- 物理法則, 111
部分帰無仮説, 105
部分尤度, 235
ブランク, 18
分位数, 60, 61
分位点関数, 80, 131
分散, 26, 49, 60, 63, 64, 67, 69, 71, 72, 91–94,
96, 108, 110, 122, 123, 132, 135, 139,
150, 166, 170, 181–183, 185, 188,
189, 197, 231, 232, 235, 255, 257
級間, 102
群間, 102
誤差, 102
- 不偏, 49, 60, 63, 64, 69, 71, 72, 76,
89–91, 94, 96, 102
分散共分散行列, 232
分散増加因子, 198
分散分析, 195
分散分析表, 101, 109, 195
分子の定義, 154
分布, 59, 143, 161, 163, 179, 222
正規, 49
分布関数, 79, 113, 133, 181
分布の位置, 49
分布の正規性, 125
分布の正規性の検定, 84
分布の広がり, 49
分母の定義, 154
- プログラム, 53
プロット, 201
プロンプト, 13
- 平滑化, 222
平均, 181, 183
重み付き, 52
平均寿命, 222
平均順位, 183, 190
平均値, 19, 28, 41, 49, 50, 64, 67, 71, 80, 89,
91, 92, 94–96, 108, 110, 113, 123,
131, 132, 135, 139, 150, 209, 210
平均値からの距離, 51
平均平方和, 102
平均偏差, 60, 62, 86
平均有病期間, 155
併合, 137, 231
平方根, 182, 183
平方和, 63, 101
平方和の求め方, 194
ヘモグロビン濃度, 110
偏回帰係数, 193, 194, 197, 199, 206–208
変換, 117
変曲点, 170
偏差, 51, 62
変数, 219
離散, 23
連続, 23
変数減少法, 206
変数選択, 195, 205, 206, 217
変数の型, 216
変数変換, 126
偏相関係数, 197, 206, 207
変動, 100, 113, 210
変動係数, 28, 64
偏微分, 118, 204
偏尤度, 235

- ベースライン, 237, 238, 241
 ベースラインハザード, 239
 ベータ関数, 82
 ベルヌーイ試行, 77, 136

 ベアマッチサンプリング, 148
 ベリ, 105

 包括的帰無仮説, 105
 法則性, 193
 飽和, 124
 補間, 208
 保健医療, 89
 補集合, 50
 保留, 107
 ホルム, 105, 107, 140
 ホルムの方法, 142

 棒グラフ, 32, 94, 132
 積み上げ, 33
 母集団, 17, 49, 67, 69, 91, 108, 135, 146, 161, 179
 母集団寄与率, 157
 母集団統計量, 63
 母数, 49, 67, 84, 134, 135, 137, 147, 179
 位置, 49
 尺度, 49
 母相関係数, 113, 114
 母標準偏差, 67
 母比率, 129-133, 138, 150, 152
 真の, 130
 母比率の推定, 140
 母分散, 49, 68, 89-92, 102, 108, 255
 母平均, 67, 69, 74, 89, 90, 105, 204
 母平均値, 100
 ボンフェローニ, 104, 107, 109, 140
 ボンフェローニの不等式, 106

 ポアソン分布, 79, 136
 ポストスクリプト, 103

 マウス, 132
 前向き研究, 158
 マッチング, 148, 158, 219
 マラリア, 142
 マルチンゲール理論, 235
 丸め誤差, 153
 稀な疾患, 160
 慢性疾患, 154, 156
 マンテル=ヘンツェル, 176
 マンテル=ヘンツェルの要約カイ二乗検定, 176, 177
 マン=ホイットニーの U 検定, 180

 見かけの相関, 112, 173
 右側打ち切り, 224
 脈圧, 198
 ミラーサーバ, 143

 無回答, 17, 18
 無作為, 67
 無作為化比較試験, 155
 無作為抽出, 70
 無作為割付, 175
 無次元, 155
 無制約 LSD 法, 104

 名義尺度, 129, 148
 メタアナリシス, 43
 メタファイル, 31
 メディアン検定, 186
 メディアン生存時間, 225, 227
 メルセンヌツイスター, 70

 モーメント, 85
 目的変数, 111
 文字コード, 226
 モデル, 193, 206, 207, 216, 239
 モデル選択, 239
 もでるのあてはめ
 モデルの当てはめ, 176

 葉害, 161

 ユールの Q, 165
 有意, 206, 207
 有意確率, 83, 85, 91, 96, 101, 107, 113, 115, 117, 121, 135, 138, 139, 142, 149-152, 158, 159, 167, 181, 183, 187, 191, 204, 208, 233, 241
 有意差, 83, 95, 139, 141, 192
 有意水準, 83, 90, 104-107, 113, 134, 135, 138, 139, 141, 142, 149, 151, 183, 186, 204, 232, 233, 237, 238

 有機溶媒, 117
 有限母集団, 150
 有効性, 168
 尤度, 216, 235
 尤度関数, 203, 204
 尤度比カイ二乗統計量, 165
 尤度比検定, 143, 201, 204, 206-208, 239, 241
 有病期間, 155, 156
 有病割合, 154, 215

 要因, 215
 要因型, 19, 145, 196, 210, 216, 217, 219, 226
 溶媒, 118

- 要約, 19
 要約カイ二乗検定, 176
 横軸, 111, 121, 201
 余事象, 77
 予測, 111, 124, 201, 207
 予測区間, 111, 119, 126
 予測値, 120, 121
- ライブラリ, 45, 143, 172
 ラベル付き要因型変数, 151
 乱数, 70, 80, 150
 乱数発生, 67
 ランダム検定, 150
 ランダム効果, 191
- 罹患数, 155
 罹患率, 155, 157
 罹患率差, 157
 罹患率比, 157
 離散分布, 79
 離散分布, 136
 離散法, 235
 離散ロジスティックモデル, 235
 リスク, 155, 158, 160
 リスク因子, 216
 リスク差, 157
 リスク集合, 221, 223, 231, 235
 リスク比, 144, 156-161
 リスクファクター, 154
 リッジ回帰, 198
 立方根変換, 126, 161
 率比, 157, 162, 163, 176
 リファレンスカテゴリ, 217-219
 両側検定, 83, 90-92, 96, 108, 113-115, 181, 183, 187
 量的変数, 91, 93, 95, 99, 173, 180, 194, 209
 理論分布, 77
 リンカーン法, 129
 臨床試験, 175
- 累積度数, 56
 累積ハザード関数, 234
 累積罹患率, 155, 160
 累積罹患率差, 157
 累積罹患率比, 157
- レーダーチャート, 44
 例示, 44
 連続修正, 183, 185, 188
 連続性の補正, 134, 139, 140
 連続分布, 80, 134, 139, 147, 180
- ロード, 223
 ログランク検定, 222, 231-233, 237
 ロケール, 226
 ロジスティック回帰, 219
 ロジスティック回帰分析, 176, 193, 215-217, 219
 ロバスト, 53
 論理型, 19, 216, 217
 論理的整合性, 56
- ワイブル分布, 223
 割合, 154, 155