保健・医療研究の進め方入門
—R と EZR を用いて—
（保健学研究共通特講 IV, VIII テキスト Rev. 1.1.0）

神戸大学大学院保健学研究科教授：中澤　港
<minatonakazawa@gmail.com>

2024 年 1 月 13 日

このテキストの目的は、保健・医療分野において、主として学位論文取得を目指す大学院生を対象に、どのように研究計画をデザインし、どのように実験や調査によって生データを得て、どのようにデータファイルを作成し、どのようにデータの性質を確認し、どのように統計解析を進め、その結果を解釈して論文にまとめるか、という一連の流れのガイドラインを示すことである。神戸大学大学院保健学研究科で 2012 年度から担当している『エビデンスベーストヘルスケア特講』（2019 年度から『保健学研究共通特講 IV, VIII』）のテキストとして開発した。

なお、このテキストで統計解析において用いるソフトウェアは、2003 年にピアソン・エデュケーションから『R による統計解析の基礎』を出版した頃から考えると信じられないほど普及している **R** と、R の代表的な GUI フロントエンドである **Rcmdr** を医療統計向けに自治医科大学の神田善伸先生がフルカスタマイズされた **EZR** である。

目 次

問い合わせ先：神戸大学大学院保健学研究科パブリックヘルス領域・教授　中澤　港
e-mail: minatonakazawa@gmail.com

改版履歴

2013 年 8 月 9 日： 第 0.5 版、2013 年度講義資料から、概要版として作成（まだ修正すべき点は多数あり）

2013 年 8 月 23 日： 第 0.6 版、一応最後まで修正とスクリーンキャプチャ完了。

2013 年 8 月 27 日： 第 0.7 版、順序の入れ替えと整理。

2014 年 4 月 6 日： 第 0.8 版、細かいアップデート。

2014 年 4 月 30 日： 第 0.8.1 版、実験計画法のセクションを「研究のデザイン」と変更し、加筆修正。

2014 年 5 月 21 日： 第 0.8.2 版、「パッケージ」と書くべきところが「ライブラリ」になっていたのを修正。順序のあるカテゴリ変数間の関係の分析について加筆。

2014 年 7 月 27 日： 第 0.8.3 版、途中で切れていたり英語のままだったところを何ヶ所か修正。

2015 年 3 月 6 日： 第 0.8.4 版、2 群間の分布の位置の差の検定について修正加筆。

2015 年 4 月 20 日： 第 0.8.5 版、ブック形式に変更。何ヶ所か書式修正。

2015 年 4 月 21 日： 第 0.8.5.1 版、第 3 章 2 節に追加した表の書式修正。サンプルサイズ設計の英文例を 2 つに分割。

2015 年 5 月 23 日： 第 0.8.6 版、EZR の作図や作表機能が大幅に向上していることに気づいたので第 6 章を大幅修正。第 2 章に「他のフロントエンド」を追記。

2015 年 5 月 24 日： 第 0.8.6.1 版、日本語や体裁がおかしかったところを微修正。

2015 年 6 月 26 日： 第 0.8.7 版、相関のところで Hotelling の T2 と集中楕円を追記。

2015 年 7 月 8 日： 第 0.8.7.1 版、反復測定分散分析のところに説明を追加。

2015 年 7 月 27 日： 第 0.8.7.2 版、院生からの指摘によりサンプルサイズの計算のところのミスを修正

2015 年 8 月 5 日： 第 0.8.7.3 版、Holm の方法による検定の多重性の補正の式に誤記があったので修正。比率の差についてのサンプルサイズの計算式も符号を修正。

2015 年 8 月 13 日： 第 0.9 版、cran ミラー情報を更新、相互参照と索引を追加

2015 年 8 月 14 日： 第 0.9.1 版、タイプミスなどを微修正。索引を若干追加。

2016 年 2 月 24 日： 第 0.9.2 版、多項ロジスティック回帰とポアソン回帰の説明を追加。

2016 年 3 月 14 日： 第 0.9.3 版、順位相関係数の信頼区間について説明を追加。

2016 年 7 月 26 日： 第 0.9.4 版、時空間データを扱うためのデータベースについての記述を追加。

2017 年 4 月 12 日： 第 0.9.4.1 版、新年度なので各種更新。

2017 年 5 月 24 日： 第 0.9.4.2 版、整合性のため章立てを若干変更。

2017 年 5 月 29 日： 第 0.9.4.3 版、日付のミスタイプ修正、mran 情報更新。

2017 年 6 月 20 日： 第 0.9.4.4 版、EZR による相関係数の計算について若干更新。

2017 年 8 月 13 日： 第 0.9.4.5 版、重回帰分析における偏相関係数の二乗の求め方についての記述を追加。

2018 年 4 月 6 日： 第 0.9.5 版、インストール関連を微修正。

2018 年 4 月 8 日： 第 0.9.5.1 版、妥当性関連について若干加筆。

2018 年 5 月 30 日： 第 0.9.5.2 版、Games-Howell 法の実行方法について追加。

2018 年 6 月 5 日： 第 0.9.5.3 版、生存時間解析で廃止された組み込みデータ名を修正。

2018 年 7 月 4 日： 第 0.9.5.4 版、生存時間解析に Statistics in Medicine にかつて掲載された素晴らしいチュートリアル論文の紹介を追加。

2018 年 7 月 5 日： 第 0.9.5.5 版、メタアナリシスの章と生存時間解析の章を入れ替え、生存時間解析に若干追記。

2019 年 3 月 29 日： 第 0.9.5.6 版、講義名変更。

2019 年 4 月 7 日： 第 0.9.6 版、jamovi の情報を追加。

2019 年 5 月 22 日： 第 0.9.7 版、Brunner-Munzel 検定と Bhapkar の検定について情報追加。

2019 年 6 月 14 日： 第 0.9.8 版、3 群以上の比較のところを整理。

2019 年 6 月 19 日： 第 0.9.8.1 版、重回帰分析のところに stargazer による整形出力を紹介。

2019 年 6 月 20 日： 第 0.9.8.2 版、重回帰分析のところに多重共線性の説明を追加（『R による保健医療データ解析演習』から採録し一部修正）。

2019 年 6 月 26 日： 第 0.9.8.3 版、多項ロジスティック回帰分析について例示を追加。

2019 年 7 月 6 日： 第 0.9.9.0 版、R-3.6.1 が最新である記載、反復測定分散分析の例 2 に大幅に説明追加（主として EZR でなく素の R での分析方法について）。

2019 年 7 月 7 日： 第 0.9.9.1 版、レイアウトを微修正。

2019 年 7 月 15 日： 第 0.9.9.2 版、第 8 章（2 つのカテゴリ変数間の関係）の説明順序がおかしいところがあったので修正。

2019 年 7 月 19 日： 第 0.9.9.3 版、Bland-Altman プロットについて blandr パッケージを使う方法を追加。

2019 年 7 月 24 日： 第 0.9.9.4 版、効果量についてとメタアナリシスの Z 変換を使う方法について追加。

2020 年 4 月 6 日： 第 0.9.9.5 版、バージョンなど微妙に更新。

2020 年 5 月 27 日： 第 0.9.9.6 版、順序ロジスティック回帰の尤度比による順序尺度の 2 群間の比較法を追加。

2020 年 6 月 13 日： 第 0.9.9.7 版、サーバ URL を http から https に変更。rcompanion パッケージを使った効果量について追加。

2020 年 7 月 24 日： 第 0.9.9.8 版、MethComp について 2019 年にあった不具合が解消していたので、その部分の記述を削除。

2020 年 7 月 29 日： 第 0.9.9.9 版、生存時間解析についてサンプルデータの説明を追加。

2021 年 1 月 6 日： 第 1.0.0 版、R-4.0.3 対応。スタイル変更。欠損値処理加筆。

2021 年 6 月 16 日： 第 1.0.1 版、spearman.ci.sas() が欠損値処理をしていないことを注記。

2022 年 1 月 6 日： 第 1.0.2 版、メタアナリシスについて meta パッケージの用例と PRISMA ガイドラインを追記。

2022 年 1 月 7 日： 第 1.0.3 版、R の記述を最新版に合わせた。

2022 年 1 月 30 日： 第 1.0.4 版、posthocTGH() 関数の userfriendlyscience パッケージから rosetta パッケージ移行に対応。

2022 年 3 月 17 日： 第 1.0.5 版、2 群の分布の位置の差の検定の効果量について追記。

2022 年 7 月 20 日： 第 1.0.6 版、survival パッケージの aml データを EZR でアクティブにする方法を注記。

2022 年 12 月 4 日： 第 1.0.7 版、サンプルサイズの計算に G*Power による方法を追記。R の最新版情報を更新。講義名修正。

2023 年 4 月 8 日： 第 1.0.8 版、冒頭の倫理指針の記述をアップデート。

2023 年 7 月 18 日： 第 1.0.9 版、R のバージョン更新、定性検査の性能比較について追記。

2024 年 1 月 13 日： 第 1.1.0 版、R のバージョン更新、グラフの色について追記。

第1章　研究の基本

研究成果を学問として発表するためには、２つの本質的条件を満たす必要がある。第 1 に、これまで誰も発表したことがない、新しい発見を含んでいることである。第 2 に、既存の学問体系の中で、その研究を位置づけることである。誰も言ったことがないことであっても、既存の学問体系の中に位置づけることができないと、エッセイ（あるいは小説）に過ぎず、学問にはならない。どんな研究でも先行研究をレビューすることが必要なのは、このためである。

また、人を研究対象とする場合は、倫理的な問題がつきまとうため、事前に所属機関の倫理審査委員会[1]に研究計画書などを提出し、審査を受けて、その研究に倫理的な問題がないことを保証してもらう必要がある。本学保健学研究科では、詳細は http://www.ams.kobe-u.ac.jp/for_staff/rinri/ を参照されたい (ただしこのページは学内からしか見えない)。2023 年 4 月 8 日現在、国の倫理指針の最終改訂は、http://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hokabunya/kenkyujigyou/i-kenkyu/ から見ることができる。2023 年 3 月 27 日付けで、人を対象とする生命科学・医学系研究に関する倫理指針[2]（ただし、2022 年 6 月 6 日付けで 1 つ前の版のガイダンス[3]も出ていて、こちらの方が読みやすい）が出ている。

長い歴史から、保健医療分野に限らず、研究の進め方には一定のスタイルが確立している。どんなテーマであれ、大きく分けると、２つの型のどちらかに含まれるだろう。

1.1　研究の 2 つの型

研究を大別すると、問題発見型と問題解決型に分けられる（もちろん、別の分け方もあるだろうが)。

問題発見型の研究は、そのターゲットに対する研究の初期段階で行われる。例えば、パイロットスタディ、ケースレポート、記述調査、問題の定式化のための研究は、この型をとる。

それに対して、問題解決型の研究は、ある程度の研究の蓄積により、問題の所在が明確になった後で行われる。通常は標本調査で、検出力分析を用いたサンプルサイズの設計を含む適切な研究デ

[1] もし所属機関に存在しない場合は所属学会などの倫理審査委員会。海外調査の場合は、通常、相手国当局の倫理審査も通す必要がある。

[2] https://www.mhlw.go.jp/content/001077424.pdf

[3] https://www.mhlw.go.jp/content/000946358.pdf

ザインが本質的に重要である。多くの場合、仮説が明確であり、**サンプリング→データ→図示→区間推定や検定→有意差や相関の検出、モデルの当てはめ**というプロセスによって仮説検証を行う。

ある程度仮説が確からしいことがわかった上で、因果関係を明らかにするために、集団を対象として行うのが介入研究である。典型的なのが、Randomized Controlled Trial (RCT：無作為化統制試験）である。例えば、新薬の有効性を調べたい時は、患者をランダムに 2 群に分け、片方は新薬を、もう一方は従来の標準的な薬を投与して、効果を比較する。この際、研究対象者も薬を投与する医師も、その薬が新薬なのか従来薬なのかがわからないようにして与えられる二重盲検（Double Blind）を行うのが普通である。

1.2　データへのアプローチ

データを得る方法には、通常は問題発見型研究で行われるインタビューまたは質問紙（構造化／半構造化／非構造化（自由回答型）がある)、観察（測定を含む)、実験（動物実験や RCT を含む）に加えて、先行研究をまとめて再分析するシステマティックレビューないしメタアナリシスがある。

1.3　全数調査と標本調査

全数調査（悉皆調査）は、問題発見型研究に多い。母集団の基礎データを得るために実施されることがある。典型的な全数調査は国勢調査。国民健康・栄養調査のような標本調査をするための母集団の基礎情報は国勢調査から得られる。一般に用いられる統計学的手法は使えない。ただし、全数調査の中から標本を抽出して詳細な集計を行ったり関連性を調べたりすることは行われている（国勢調査の場合だと、約 1%の調査票を使った速報集計や、抽出詳細集計がそれに当たる)。

統計解析の対象になるデータは、大抵の場合、標本調査によって得る。仮説検証型研究、動物実験、介入研究では適切な標本抽出（サンプリング）が必須である。臨床研究では、ある期間内に集まった症例数で妥協するしかない場合があるが、あくまで妥協と考えるべきである。原則として標本サイズはきちんと設計する必要がある（詳しくは後述)。動物実験や介入研究ではとくにクリティカル（サンプルサイズが小さくて検出力が足りなかったために有意な差が検出できなかったという言い訳は通用しない)。

最近流行のビッグデータは、（多くの場合自動的に記録される、従来のリレーショナルデータベースでは処理しきれないほど）大量のデータのことだが、全数調査ではない。かといって計画された標本調査でもない。本来、母集団が何かということに注意を払うべきだろうが、あまり考慮されていないようにみえる。しかし、手に入った範囲の大量のデータを短時間で高速に処理することで、限定的な特性の概略の傾向を掴むことがビッグデータ解析の目的であることが多いので、それでも実用上の意味はあるのだろう。

第2章　Rの基本

こうしてテーマに対するアプローチが決まったら、次にすることは、研究計画を立てることである。その後で研究を実施し、信頼できるデータを得て、データ解析をして、先行研究と比較しながら結果を解釈することになる。

大学院生の研究は、多くの場合、テーマを絞った問題解決型の標本調査になると思われるので、研究計画を立てる上で、サンプルサイズの計算が必須になる。そこで、先に進む前に、統計ソフトについて紹介しておく。現在では、SAS、JMP、SPSS 等さまざまな統計解析ソフトが利用できるが、このテキストでは代表的なフリーソフトウェアである R と、それを GUI で操作でき、医学統計解析向けの関数を整備した EZR を使った操作について説明する。**神戸大学では図書館のコンピュータの SPSS のライセンス契約は 2016 年度末で切れたため、研究科としても R または EZR を使うことが推奨されている。**

R は MS Windows、Mac OS、Linux など、さまざまな OS で動作する。中間栄治さんが早い段階で開発に参加してくださったおかげで、テキスト画面でもグラフィック画面でも日本語の表示が可能だし、岡田昌史さんや間瀬茂さんを中心に組織されたユーザグループの協力によって、インターフェースの多くの部分で日本語に翻訳されたメッセージが利用可能である[1]。Windows 版や Mac OS 版は、通常、実行形式になっているものをダウンロードしてインストールする。Linux では tar で圧縮されたソースコードをダウンロードして、自分でコンパイルすることも難しくないが、ubuntu などではコンパイル済みのバイナリを提供してくれている人もいるので、それを使う方が容易にインストールできるかもしれない。

R はフリーソフトなので、自分のコンピュータにインストールすることも自由にできる。R 関連のソフトウェアは CRAN (The Comprehensive R Archive Network) からダウンロードすることができる。CRAN のミラーサイトが世界中に存在し、ダウンロードは国内のミラーサイトからすることが推奨されているので、日本では統計数理研究所のサイト[2]を利用すべきだろう。

[1] この翻訳作業は、R の大きなバージョンアップの際には毎回必要になる。日本語への翻訳チームは、RjpWiki の岡田さんが組織している（http://www.okadajp.org/RWiki/index.php?RJapaneseTranslation）。

[2] http://cran.ism.ac.jp/

2.1 R のインストール方法

2024 年 1 月 13 日現在、R の最新版は R-4.3.2（コード名 “Eye Holes”）である。R はバージョンごとに開発コード名が付いているが、その出典が漫画 Peanuts!なのは有名である。“Eye Holes” の出典もそうで[3]、リリース日がハロウィン近かったので、子どもがハロウィンでお化けの仮装をするのに布に目出し穴を開けるのを忘れて見えない、というネタであった。ちなみに 4.3.1 のコード名は “Beagle Scouts” で、初出から来年で 50 周年を迎える特設サイト[4]ができているくらい有名なテーマだった。R-4.1.*からネイティブなパイプ記法が導入されるなど大きく機能追加された。2022 年 4 月にリリースされ R-4.2.0 でも大きな変化があり、とくに Windows 版では、UTF-8 をネイティブサポートしたため、R-4.1.*以前に Shift-JIS で書いていたコードを R-4.2.*のコンソールで開くと文字化けする[5]。Rtools も 4.2 に変わった。MSVCRT ではなく UCRT に依存しているため Windows10 (1903) 以上でない場合、R より先に UCRT のランタイムをインストールする必要がある。32 ビット版もサポートしなくなるなど変化が大きいが、もはや大半の Windows ユーザは Windows 10 か Windows 11 で、64 ビットプロセッサであろうから、個人的には良いと思う。グラフィックデバイスの扱いが変わり[6]、これまで入っていたパッケージもすべて Rtools42 以降を使ってビルドし直して再インストールする必要がある。2024 年 1 月 13 日現在の Rtools の最新版は Rtools43[7]である。

Windows CRAN ミラーから R-4.3.2 のインストール用ファイル (`R-4.3.2-win.exe`) をダウンロードし、ダブルクリックして実行する。インストール途中で、スタートアップオプションをカスタマイズするかどうか尋ねるダイアログが表示されるので、ここは**いいえ（デフォルト）でなく、はい（カスタマイズする）の方をマークして**「次へ」をクリックすることをお薦めする[8]。次に表示されるウィンドウで **SDI (separate windows) にチェックを入れて**「次へ」をクリックするのが重要である。他のオプションは好みに応じて選べば良い。**なお、後で Rcmdr/EZR をインストールしたい場合は、Windows 自体のログインユーザ名に全角文字（2 バイトコード）を使わない方が良い。**

Macintosh CRAN ミラーからダウンロードしてインストールできるが、Mac OS X のバージョンとハードウェアの CPU によってインストールすべきバージョンが異なるので注意が必要である（Rcmdr や EZR を使うには X11 が必須なので，XQuartz のリンク先（`https://www.`

[3] `https://www.gocomics.com/peanuts/1996/11/01`

[4] `https://www.snoopy.co.jp/beaglescouts50/`

[5] 日本語コード指定できるエディタで変換するか、Chrome などのブラウザで開いて文字化けしていない状態でコピーし、コンソールやスクリプトウィンドウにペーストすれば問題ない。

[6] `https://developer.r-project.org/Blog/public/2021/12/14/updating-graphics-devices-for-r-4.2.0/index.html` に書かれているように、グラフィックデバイスを提供しているパッケージもすべて deviceVersion を 15 にしないと動かなくなる。

[7] `https://cran.r-project.org/bin/windows/Rtools/rtools43/rtools.html`

[8] スタートアップオプションがデフォルトでは、R を起動した後のすべてのウィンドウが、1 つの大きなウィンドウの中に表示される MDI モードになってしまうのだが、それだと Rcmdr/EZR が非常に使いにくくなるからである。

xquartz.org/"）から XQuartz-2.8.5.dmg もダウンロードしてインストールすること）。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[9]があるので参照されたい。

Linux Debian、RedHat、ubuntu など、メジャーなディストリビューションについては有志がコンパイルしたバイナリが CRAN にアップロードされているので、それを利用すればインストールは容易であろう。例えば ubuntu の場合は、ソフトウェアセンターから R を選んで「Install」をクリックするだけで良い。ただし最新版を使いたい場合は若干の手間が必要である。https://cran.ism.ac.jp/bin/linux/ubuntu/#installation が参考になるだろう。

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-4.3.2.tar.gz をダウンロードして展開して自力でコンパイルする。最新の環境であれば、./configure と make してから、スーパーユーザになって make install で済むことが多いが、場合によっては多少のパッチを当てる必要がある。

なお、マルチコアの CPU に対応した Revolution R が開発されていたが、開発していた Revolution Analytics 社が Microsoft に買収された後、引き続き Microsoft R Open として MRAN (Microsoft R Application Network)[10]からダウンロードできる状態が暫く維持されていたものの、2023 年 7 月 1 日にサイトが廃止された[11]。

2.2 R の使い方の基本

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

Windows では、インストールが完了すると、デスクトップまたはクイック起動メニューに R のアイコンができている。Rgui を起動するには、デスクトップの R のアイコンをダブルクリックす

[9] http://aoki2.si.gunma-u.ac.jp/R/begin.html
[10] https://mran.microsoft.com/download/
[11] https://techcommunity.microsoft.com/t5/azure-sql-blog/microsoft-r-application-network-retirement/ba-p/3707161

るだけでいい[12]。ウィンドウが開き、作業ディレクトリの.Rprofile が実行され[13]、保存された作業環境.RData が読まれて、

```
>
```

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

入力した命令や関数は、「ファイル」メニューの「履歴の保存」で保存でき、後で「ファイル」の Source で呼び出せば再現できる。プロンプトに対して source("プログラムファイル名") としても同じことになる（但し、Windows ではファイルパス中、ディレクトリ（フォルダ）の区切りは/または\\で表すことに注意[14]。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、キーボードの↑を押せば既に入力したコマンドを呼び戻すことができる。

なお、R をインストールしたディレクトリの bin にパスを通しておけば、Windows 8/8.1/10/11 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。もっといえば、Makefile を書いておき、make を使って R を実行することもできる。下枠内のように書いたバッチファイル（make.cmd とか）を作っておき、INPUT.R に R のコードを書き、バッチファイルをダブルクリックして実行させ、結果が output.txt に保存されるように設定することもできる。

```
Rterm --vanilla < ./INPUT.R > output.txt
```

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

[13]R-4.1.*以前に日本語を含む.Rprofile が保存されている場合に、R-4.2.*を新規インストールして同じ作業ディレクトリを使うと、.Rprofile に含まれている Shift-JIS の日本語が文字化けして正常動作しないという問題が生じるので、.Rprofile ファイルの文字コードには注意を払うべきである。

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

2.3 Rgui プロンプトへの基本操作

終了 q()

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

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

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

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

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

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

とする。最初のダウンロード利用時には、パッケージをどのミラーサーバからダウンロードするかを聞いてくるので、通常は国内のミラーサーバを指定すればよいだろう。クラウドを指定しても良い。筆者は国立情報学研究所のサーバを利用することが多い。dep=TRUE は dependency（依存）が真という意味で、Rcmdr が依存している、Rcmdr 以外のパッケージ（かなりたくさんある）も自動的にダウンロードしてインストールしてくれる。なお、TRUE は T でも有効だが、誤って T を変数として別の値を付値してしまっていると、意図しない動作をしてしまい、原因を見つけにくいバグの元になるので、できるだけ TRUE とフルスペル書いておくことが推奨されている。

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

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

2.4 R Commander / EZR を使う

このようなコマンドベースの使い方に習熟するには一定の時間が必要である。世界各地で R ユーザが開発した追加機能パッケージが多数公開されているが、なかでもカナダ・マクマスタ大学の

John Fox 教授が開発した Rcmdr（R Commander）は、**メニュー形式で R を操作できる**パッケージとして有名である。Rcmdr のメニューはカスタマイズすることができるし、プラグインという仕組みで機能追加もできるので、自治医科大学の神田善伸教授が医学統計向けにフルカスタマイズし機能追加したものが EZR である[15]。

R Commander や EZR をインストールすれば、メニューから選んでいくだけで多くの R の機能を使うことができるので便利である。ただし、メニューに入っていない機能も多いので、必要に応じて、スクリプトウィンドウに直接 R の関数を打ち、実行したい範囲を選択して「実行」ボタンをクリックする必要がある。

EZR をインストールするには、Rcmdr をインストールした後で、

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

と打てばよい[16]。

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

そこから EZR を呼び出すには、メニューの「ツール」から、「Rcmdr プラグインのロード」を選び、プラグインとして RcmdrPlugin.EZR を選んで OK ボタンをクリックする。少し待つと、「R コマンダーを再起動しないとプラグインを利用できません。再起動しますか？」と尋ねるダイアログが表示されるので、「はい (Y)」をクリックすると EZR が起動する[17]。

なお、オリジナルの Rcmdr メニューも「標準メニュー」として残っているので、EZR であっても Rcmdr としての標準的な使い方ができる。

---

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

[16] Windows の場合、自治医大のサイトで公開されている、EZR 組込み済みの R をダウンロードして利用するとインストールが簡単である。2024 年 1 月 13 日時点では、R-4.3.1 ＋ Rcmdr2.9-1 ベースになっているが、R 本体のバージョンが最新版よりは古くなることが多いので注意が必要である。R は別々のディレクトリに複数のバージョンを併存させることが可能である。MacOS 版のインストール後、あるいは Windows でも R 本体のインストール後にパッケージとして Rcmdr と RcmdrPlugin.EZR をインストールした場合についても、自動的に EZR 起動までいくようにする設定方法も、自治医大のサイトで解説されている。

[17] 実は、EZR を呼び出すところまでやってからツール＞オプションでフォントサイズの設定などをしてから、ツール＞オプションの保存で.Rprofile を作業フォルダ（R の起動アイコンのプロパティで設定）に保存しておけば、R を起動後に library(Rcmdr) をするだけで、フォントサイズなども設定済の状態で EZR が起動するようになる。EZR しか使わない人なら、保存した.Rprofile をエディタで開いて下の方の#を４つ消すだけで、R を起動すると勝手に EZR が起動するようになって便利である。その後、この.Rprofile を C:/EZRDATA（別の作業フォルダでも良い）に入れておき、普通の R の起動アイコン（ショートカット）をコピーして名前を EZR としておき、プロパティで作業フォルダを C:/EZRDATA（さっきコピーした.Rprofile が存在するフォルダ）にすれば、R 本体は１つしかインストールしなくても、普通に素の R を呼び出すアイコンと EZR を呼び出すアイコンを使い分けられる。ここまでの手順はインストーラを実行する方法に比べると若干面倒だが、EZR を最新版の R 環境で使うには、こちらの方が良い。

2.5 他のフロントエンド等

R には、Rcmdr/EZR の他にもいくつかのフロントエンドとなるソフトウェアが存在する。統計解析の機能としては R を使うのだけれども、操作するためのフロントエンドとして、R コンソールよりも多機能なソフトをかぶせることによって操作性を改善するものである。

2.5.1 RStudio

統計解析の専門家やパッケージ開発者に人気なのが、RStudio[18]である。R 本体と同様に、Windows 版、MacOS 版、Linux 版が存在する。RStudio の利点はいろいろあるが、プロジェクトという単位でコードを管理できるのが大きい。パッケージ作成者がメンテをするには、ほぼ必須のツールといえる。オブジェクトの一覧が常時得られていて、それらの中身を確認する際も R コンソールより遙かに見やすい。ただ、Rcmdr/EZR とは組み合わせにくいので、たぶん初心者は Rcmdr/EZR で、中級者以上になったら RStudio を使うのがいいだろう。筆者は R コードとそれを Rterm を使って実行するためのスクリプトを書いてファイラからバッチ実行するという使い方をすることも多いが、普通は RStudio で十分である。

2.5.2 R AnalyticFlow

日本で開発されたフロントエンドで、フローチャートとして分析プロセスを操作するとコードが生成される点に特徴があるのが、R AnalyticFlow[19]である。分析の流れを可視化してくれるのは便利だし、他にも便利な機能は多い。これも R 本体と同様、Windows 版、MacOS 版、Linux 版が存在する。Rcmdr/EZR よりもコード実行を意識するし、RStudio よりもコードの流れはわかりやすいので、初心者が中級者に移行するときに役に立つかもしれない。デバッグ機能が優れているので、中級者以上でも RStudio より R AnalyticFlow が好きだという人も珍しくない。

2.5.3 jamovi

jamovi[20]も計算のバックエンドには R を使うことができるが（R のコードを出力させることができる）、SPSS のようにデータの属性をビジュアルに指定し、表を見ながら統計手法を選んで使うソフトのようである。Rz に似ている思想と思われるが、パッケージではなく RStudio 同様、独立し

[18] http://www.rstudio.com/products/RStudio/
[19] http://www.ef-prime.com/products/ranalyticflow/
[20] https://www.jamovi.org/

たソフトウェアとしてインストールする。Windows でも MacOS でも Linux でも動作させることができる。

特筆すべきは、“Learning Statistics with jamovi”[21]というフリーテキストが提供されていることである。このテキストには、芝田征司氏による日本語訳『jamovi で学ぶ心理統計』[22]も存在する。元の “Learning Statistics with jamovi” は、David Foxcroft 氏が、Danielle Navarro 氏によるテキスト “Learning Statistics with R”[23]の使用ソフトを R から jamovi に置き換えて作成したものであり、心理統計の入門書としても大変参考になる。

2.5.4　Ωnyx、Stan など

最近では、SAS や SPSS からも R を呼び出して使うことができるようになった。また、構造方程式モデリングのために R 本体への追加パッケージとして `sem` や `lavaan` を用い、モデル自体を GUI で操作できるフロントエンドとして Ωnyx[24]を使うとか、階層ベイズモデルのために R には `RStan` パッケージをインストールして、Stan[25]と連動させて使うなど、高度な統計解析のために他のソフトウェアと連動させて使うことも広く行われている。このあたりが、R のオープンさゆえの長所だと思う。

2016 年に入り、2015 年に Revolution 社を買収した Microsoft が、C#, Visual Basic, F#, C++, JavaScript, TypeScript, Python など多くの言語の開発環境である Visual Studio を R に対応させる R Tools for Visual Studio[26]をリリースした。他の言語で Visual Studio を使い慣れているプログラマにとっては便利かもしれない。

[21] `https://sites.google.com/brookes.ac.uk/learning-stats-with-jamovi`
[22] `https://bookdown.org/sbtseiji/lswjamoviJ/`
[23] `https://compcogscisydney.org/learning-statistics-with-r/`
[24] `https://onyx-sem.com/`
[25] `http://mc-stan.org/`、解説としては `http://tjo.hatenablog.com/entry/2014/01/27/235048` など参照
[26] `https://www.visualstudio.com/en-us/features/rtvs-vs.aspx`

第3章　測定と疫学調査の基礎知識

研究計画に入る前に、もう１つ説明しておかねばならない。統計解析を要する研究においては測定が必須であり、その測定は「正しく」なければならない。この章では、まず測定の正しさについて説明し、保健・医療の分野で集団を対象として正しい測定をするために必要な、疾病量と疾病量への効果を把握する疫学的な方法論の基礎について簡単に解説する。詳細は Rothman (2012) などを参照されたい。

3.1　正しい測定とは？

測定の正しさを考えるとき、少なくとも 3 種類の異なる正しさが存在することに注意したい。Validity (妥当性)、Accurary（正しさ・正確性）、Precision（精度）の 3 つである。

Validity とは、測りたいものを正しく測れていることである。例えば、ELISA で抗体の特異性が低いと、測定対象でないものまで測ってしまうことになるので、測定の妥当性が低くなる。途上国で子供の体重を調べる研究において、靴を履いた子供がいてもそのまま体重計に載せて、表示された値を使ったという研究があったが、そういうときの靴は往々にしてブーツなので 1 kg 近くの重さがある場合もあり、表示値そのままでは体重が正しく測れていない。また、測定すること自体が測定値に影響を与えてしまうと、妥当性は損なわれてしまう。例えば、心理的ストレスに対して何らかの物質の血中濃度が鋭敏に反応するとしても、採血自体がストレスを与えてしまう可能性があるので、その物質の測定値が高いという結果が出ても、元々ストレスが掛かっていたのか、採血がもたらしたストレスのせいで高値になったのか判別不能なので、短期的ストレス評価の指標として血液を含む侵襲的なサンプル採取を伴うと妥当性は損なわれる。もちろん、侵襲度が低ければ良いというものでもない。ストレス評価の場合、POMS-J のような質問紙は非侵襲、心拍や唾液や光トポグラフィや脳波は低侵襲、血液は高侵襲だが、高侵襲な指標ほど真に測りたいものに近いので妥当性が高いことが多い。長期的なストレスを判定したい場合ならば、血液中の物質で、長期にわたる心理的ストレスに応じて高濃度になるが、短期的なストレスに対しての応答速度が速くない物質を測定すれば、妥当性が高くなる。ただし一般に、そこまで高い妥当性が必要でなければ低侵襲な方が倫理的に良い。

Accuracy とは、バイアス（系統誤差）が小さいことである。モノサシの原点が狂っていると（目

盛幅が狂っているときも)、正しさが損なわれる。例えば、何も載っていない状態で 1 kg と表示されている体重計で体重を測定すると、真の体重が 60 kg の人が載ったときの表示は 61 kg となるだろう。すべての人の体重が 1 kg 重く表示されたら、正しいデータは得られない。

Precision は、確率的な誤差（ランダムな誤差）が小さいことである。信頼区間が狭い、CV（変動係数）が小さいこととも同義である。一般に感度の低い測定は精度が低くなる。例えば、手の大きさを測るのに、通常のものさし（最小表示目盛が 1 mm）で測るのと、ノギス（最小表示目盛 0.1 mm）で測るのでは、最小目盛の 1/10 まで読むとしても、精度が 1 桁異なる。

3.1.1　信頼性と妥当性

信頼性は安定性、再現性 (test-retest reliability) や測定者間一致度 (inter-observer concordance) や項目間一致度 (Cronbach の α 係数等）で示される。系統誤差がない場合、信頼性が高い測定ができれば、一般に妥当性もある。

本質的に直接測定が不可能な場合（質問紙によるストレス評価など)、３種の妥当性を確保する必要があると言われてきた。

- 内容的妥当性 (content validity)：専門家の判定により妥当であるとされること。
- 基準関連妥当性 (criterion validity)：併存/予測について既存指標との相関が高いこと。
- 構成概念妥当性 (construct validity)：収束的妥当性 (convergent validity) や弁別的妥当性 (discriminant validity) とも関連するが、クロンバックらが考えたもので、すべての測定の背後に測定不可能な構成概念が存在し、構成概念間及び構成概念と測定の間に決まった対応関係からなるネットワーク（法則定立的ネットワーク）があることを仮定し、データによってこのネットワークが検証されることにより測定と構成概念の対応関係が明確化されることで構成概念妥当性の証拠が得られるとした。

しかし、村山 (2012) は、妥当性とは構成概念妥当性のことであり、他の「○△妥当性」は構成概念妥当性を検証するための方法・証拠のタイプと述べている。つまり、内容的証拠（専門家が妥当と判断するということ)、収束的証拠（類似概念を知るためのテスト指標と高い相関があること)、弁別的証拠（異なる概念を知るためのテストと低い相関しかないか無相関なこと)、のように考えるべきとしている。

3.1.2　正確さを保つために

正確な測定には、ゼロ点調整が重要である。例えば、電子天秤で秤量するとき、試薬皿を載せた状態で tare しておかないと、試薬皿の重さだけ少なく量ってしまう。原点を通る検量線を描きたい

とき、ブランクで吸光度がゼロになるように調整する（このときのブランクは水ではなく、対象物質の濃度がゼロで試薬は入っている、試薬ブランク）。

正確さを保証するには、例えば、標準物質の測定結果が certified range に入っているかどうかを確認する（certified range 自体の正しさは複数の reference labo での測定で相互保証）。

正しい検量線を作る方法としては、以下２つが知られている。

標準添加法： 共存物質が影響するとき、試料溶液を分けて標準希釈系列を添加し混合したものの吸光度を測定し、添加濃度を横軸にとって検量線を描く。ゼロ点調整は水ブランク。吸光度ゼロに相当する添加濃度（マイナスになっている）が試料の濃度（試料は最初からそれだけその物質を含んでいると考える）。

内標準法： これも共存物質の影響を除くため、測定物質と似ていて測定対象でない物質を内標準物質として標準試料と未知試料に添加し、内標準物質と標準試料の吸光度比を検量線の縦軸にとる

3.1.3 精度を保つために

精度を保つ方法としては、同一サンプルを繰り返し測定（または duplicate や triplicate で同時に測定）して、CV が小さいことを確かめることが、よく行われる。CV (Coefficient of Variation) とは、標準偏差を平均値で割った値（通常は 100 を掛けて％表記）である。

発想としては、測定値が真値±測定誤差の結果であり、測定誤差が平均ゼロの正規分布に従うと考え、誤差の標準偏差が測定値そのものに比べて十分小さい（例えば５％未満）なら測定値は信用できると考える。

異なるサンプルの測定値のばらつきを示すのに CV を使うのは誤用であり、その目的なら標準偏差そのものを見るべきである。また、サンプルから母集団におけるデータのばらつきを推定するには不偏標準偏差（不偏分散の平方根）を用いる。

サンプルから母集団の平均値を繰り返し推定した場合、その平均値がどの程度ばらつくか（即ち平均値の標準偏差）を示すのが標準誤差。サンプルから得られた不偏標準偏差をサンプルサイズの平方根で割った値になる。平均値が欲しいなら、標準誤差が小さいほど精度は高くなる。つまり、サンプルサイズが大きいほど精度は上がる。

3.2 リスク因子への曝露が疾病発生に与える影響をどう評価するか?

疫学調査では、通常、「疾病／非疾病」と「曝露／非曝露」の関連性を調べるため、２つのカテゴリ変数間の関連性の程度をクロス集計により評価するのが普通である。

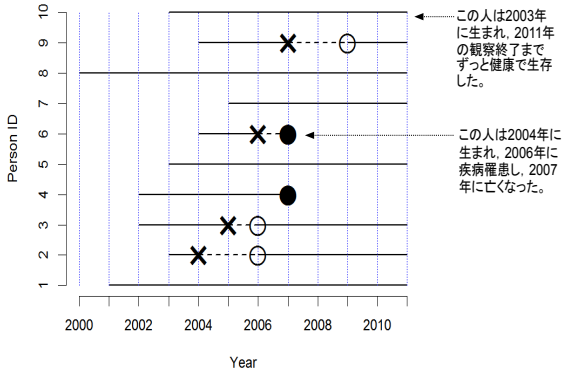
これを別の角度からみると、曝露群と非曝露群の間で、疾病量を比較することに相当する。疾病量の指標としては、prevalence（有病割合）または odds（オッズ）、risk（リスク）、incidence rate（罹患率）を区別する必要がある。

疾病と曝露の関連性の程度【＝ effect（効果）】の評価法としては、difference（差）でみるか、ratio（比）でみるかを区別して考えるべきである。どちらも一長一短であり、目的に応じて使い分けるべきである（下表）。

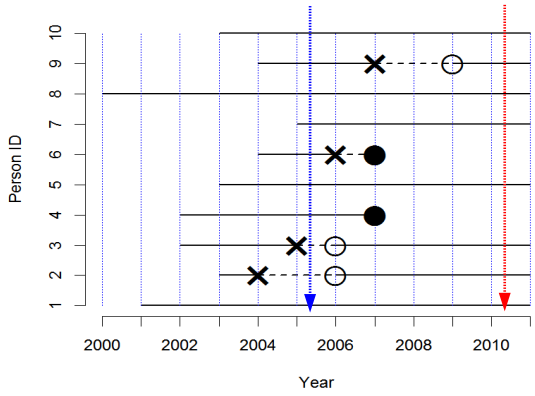
疾病量の指標	差 (difference)	比 (ratio)
罹患率 (incidence rate)	罹患率差（＝率差）＝曝露群の罹患率 − 非曝露群の罹患率	罹患率比（＝率比）＝曝露群の罹患率÷非曝露群の罹患率
リスク (risk)	リスク差＝曝露群のリスク − 非曝露群のリスク	リスク比＝曝露群のリスク÷非曝露群のリスク
オッズ (odds)	（なし）	オッズ比＝要因あり群の疾病オッズ÷要因なし群の疾病オッズ＝症例群の曝露オッズ÷対照群の曝露オッズ

3.3 疾病量をどうやって把握するか—有病割合とリスクと罹患率の違い

疾病発生の実際の状況を考えてみよう。下図では、横線のそれぞれが１人の個人を意味し、実線は健康、破線は疾病である状態を示す。×は疾病罹患、○は治癒、●は死亡を意味する。横軸は観察年を意味する。



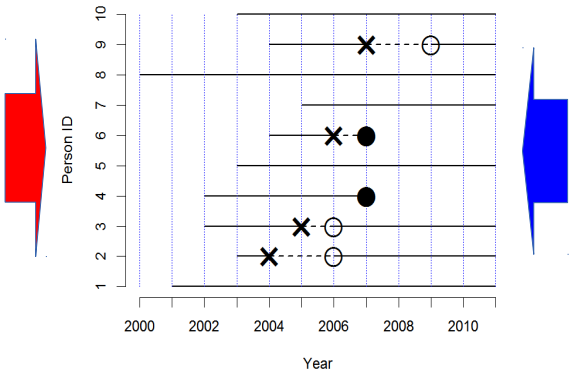
まず、**有病割合 (prevalence)** とは、ある時点で、全体の中でどれくらいの割合の人が病気かを意味し、以下のようにして求める。例えば、2005 年半ばに横断研究をすると（下図青矢印）、10 人の人がいて、うち 2 人が病気なので、有病割合は 2/10 で 0.2 となる（なお、似て非なる概念である**疾病オッズ**[1]は 2/8 で 0.25 である）。調査が簡単なのが利点だが、2010 年半ばに横断研究をすると、有病割合は 0/8 で 0 となってしまうこと（下図赤矢印）からわかるように、ある瞬間の情報しか与えてくれないという欠点がある。



次に、**リスク (risk)** は、累積罹患率 (cumulative incidence rate) ともいい、最初にいた観察対象人数を分母、観察期間内に病気を発症した人数を分子として求めた、罹患の確率を意味する。当然、観察期間が長いほど、大きい値になる傾向がある。2011 年に生き残っている 8 人の子供について親に過去の罹患について思い出してもらい（下図青矢印）、3 人について疾病罹患が報告されたなら、この 11 年間のリスクは 3/8=0.375 と推定される。このような後ろ向きの研究は簡単で安価にできるが、既に亡くなった子供の情報を聞き逃してしまう欠点をもつ。逆に、2000 年から 11 年間

[1]疾病有りの人数の疾病無しの人数に対する比をいう。

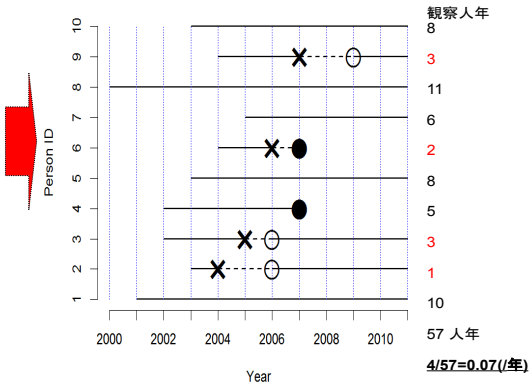
のコホート研究をした場合は（下図赤矢印）、10 人の子供のうち疾病罹患は 4 人が経験したので、11 年間のリスクは 0.4 となる。ただし生後 1 年間に同じ病気に罹るリスクは 1/10=0.1 となる。この例から、リスクは観察期間に依存することが良くわかる。



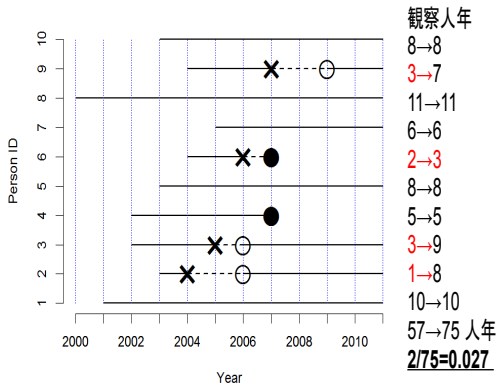
観察期間に依存しない指標を得るには、分母を人数でなく、延べ観察人年にすればよい。この発想で得られる指標が、**罹患率 (incidence rate)** である。上図赤矢印と同じく 2000 年からコホート研究すれば全人年データを観察できる[2]。下図のように数えた感受性のある観察期間の合計人年を分母、疾病発生数を分子にした値が**「罹患率」(incidence rate)** となる。罹患率は、(1/年）という次元をもつ（時間当たりの件数、即ち発生速度を意味する)。実際に計算してみると、57 観察人年のうち 4 例発症しているから、4/57 で約 0.07（/年）となる。

なお、何度も罹患する疾病について、集団の罹患率を求めるには、年央のリスク人口を分母、その年の疾病発生数を分子とすると、罹患率が得られる（通常、10 万人年当たりで計算する)。感染症サーベイランス事業で医師が診断したときに全数報告することになっている疾患については、報告数を年央人口で割ることによって毎年の罹患率が計算できるが、定点報告疾患や、あるいは罹患しても医療機関を受診しないような軽い疾患については、この方法で罹患率を求めることはできない。

[2]ただし、注目している疾病が一生に一度しか罹らないものなら、一度罹患した人は感受性を失うので、患者はリスク人口 (population at risk) から除去されることに注意されたい。



観察のエンドポイントを疾病発生から死亡に変えると、罹患率の代わりに死亡率が計算できる。この例では、下図のように 0.027/年となる。罹患率と同様に、大集団についての指標としては、年間死亡数を年央人口で割るとその年の死亡率が得られる（通常、1000 人年または 10 万人年当たりで表す）。この例は小標本だから不適切だが、仮に下図で計算すると、2007 年に 0.2/年、他の年は 0 となる。



3.4 曝露と疾病の関係を調べるには＝曝露と非曝露の間で疾病量を比較する

典型的な比較方法には、差（絶対比較）と比（相対比較）がある。両方ともそれぞれ意味がある。

比較する疾病量による違いがあり、リスクの場合は、リスク差またはリスク比を求める。罹患率の場合は、罹患率差または罹患率比を求める。また、死亡率の場合は死亡率差または死亡率比を求

める（これらを総称して率差とか率比とかいう）。しかし、有病割合の場合は、オッズ比を用いるのが普通であり、有病割合の差を比べることはあまりない。

3.5　絶対比較

絶対比較は、リスク差または率差による。疫学では寄与危険とか超過危険と呼ぶ。

架空の例を考えてみる。高圧線の近くの住民 10 万人（曝露群）を 5 年間追跡調査して、毎年 2 人ずつ白血病患者が見つかったとする。高圧線から離れたところの住民 10 万人（非曝露群）を同じ期間追跡調査して、毎年 1 人ずつ白血病患者が見つかったとする。この人たちに居住地以外の（教育歴とか年収とか人種とか気候とか）違いはないものとする。

このとき、リスク差は、

$$10/100000 - 5/100000 = 5/100000 (= 0.00005)$$

罹患率差は、

$$10/(100000 + 99998 + 99996 + 99994 + 99992) - 5/(100000 + 99999 + 99998 + 99997 + 99996)$$

$$\fallingdotseq 0.0000100006 (/\text{year})$$

罹患率自体が非常に小さいので、差は小さく見える。

3.6　相対比較 (1)

同じ例で、リスク比と罹患率比を考えてみる。リスク比は、

$$(10/100000)/(5/100000) = 2$$

罹患率比は、

$$\frac{10/(100000 + 99998 + 99996 + 99994 + 99992)}{5/(100000 + 99999 + 99998 + 99997 + 99996)}$$

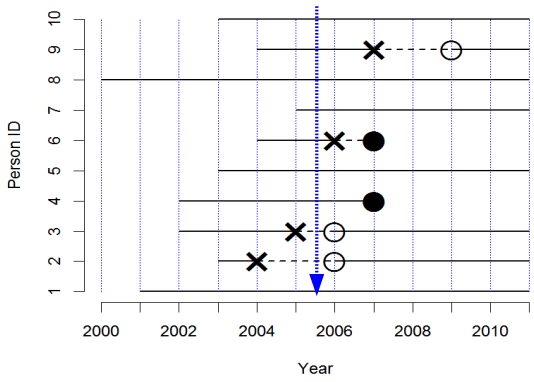
$$\fallingdotseq 2$$

どちらの値も、「高圧線の近くに住むと白血病のリスクが約 2 倍になる」ことを意味する。統計学的に有意かどうかを調べるには、「母集団における比が 1」という帰無仮説を検定する。R (EZR), SAS, JMP のようなソフトを使えば簡単で、p 値は 0.2 より少し小さい程度なので有意水準 5%では有意でない（`fmsb` パッケージを使って、下枠内のように R コンソールで打てば、点推定量とともに、95%信頼区間や p 値も示される）。

```
library(fmsb)
riskratio(10, 5, 100000, 100000)
rateratio(10, 5, 100000+99998+99996+99994+99992,
          100000+99999+99998+99997+99996)
```

3.7 相対比較 (2)

先に示した例で、2005 年に横断的研究をしたときのオッズ比を考えてみる。下図の状況である。



横断的研究では有病割合が疾病量の指標になる。しかし通常、有病割合そのものの差や比は評価しない。代わりにオッズ比を計算する。ここで、偶数番号の人が喫煙者、奇数番号が非喫煙者だったとする。喫煙者の疾病オッズは 1/4、非喫煙者の疾病オッズも 1/4 となるので、オッズ比は (1/4)/(1/4) = 1 となる。

症例対照研究では、オッズ比は症例の曝露オッズ（曝露有りの人数の曝露無しの人数に対する比）を対照の曝露オッズで割った値になる。もし 2005 年に 2 番と 3 番の人が症例として見つかった人で、他の 8 人が対照として標本抽出された人だったとすると（かつ偶数番号の人が喫煙者）、オッズ比は (1/1) / (4/4) = 1 となり、これら 2 つのオッズ比は一致する。

白血病の例では、有病割合が低いので、横断的研究で十分な検出力を得るためには大きなサンプルサイズが必要になるし、コホート研究で高圧線の近くに住むことの影響を調べるデザインも、比較的大きなサンプルサイズと長期間の観察を必要とするので効率が悪い。そのため、症例対照研究で、ある病院を一定期間に受診した白血病患者全員を症例とし、同じ病院を外傷で受診した人を対照としてリクルートし、居住場所が高圧線の近くかどうかを尋ねるのが常道である。

例えば、白血病患者 100 人のうち、高圧線の近くに住んでいる人が 20 人、外傷患者 200 人のう

ち、高圧線の近くに住んでいる人が 20 人だったとすると、オッズ比は以下のように計算できる。

$$(20/80)/(20/180) = 18/8 = 2.25$$

このオッズ比の統計学的有意性を検定するのに、R では以下のようにする。fisher.test() 関数と fmsb パッケージの oddsratio() 関数では若干のずれがあるが、いずれも p 値は 0.05 より小さく、高圧線の近くに住むことが、有意水準 5%で統計学的に有意に白血病罹患と関連しているといえる。

```
> fisher.test(matrix(c(20, 80, 20, 180), 2)) #オッズ比は最尤法での反復推定値
> library(fmsb)
> oddsratio(20, 80, 20, 180) #定義通りのオッズ比、正規近似による p 値、信頼区間
```

EZR の場合は、メニューの「統計解析」→「名義変数の解析」→「分割表の直接入力と解析」と進んで、表示されたウィンドウの中に 2 × 2 の分割表があるので、それぞれに 20, 80, 20, 180 と入力して OK ボタンをクリックすれば、`fisher.test()` と同じ結果が得られる。

第4章　サンプルサイズの問題

4.1　仮想的な例

大学院在籍期間中に研究に応じてくれた患者の数が限られていたため、要因 Y が結果 Z（通常、その病気であること）と関連がないという特定の帰無仮説 X を検定するために、10 人の患者（とそれに対応する１～３倍程度の対照群）しか調査できなかったとする。検定の結果、有意水準 5% で帰無仮説 X は棄却されなかったとする。

公表バイアス (publication bias) を避けるためには、こういう、「統計学的に有意でない」結果も、“Lack of association ...” のようなタイトルを付けて投稿すべきである。しかし、おそらく査読者は、この研究結果が「有意でなかった」のは、サンプルサイズが小さかった（言い換えると、検出力が足りなかった）からと判定する。これは研究デザインの致命的な欠点なので、論文はリジェクトされる。

大学院生としては学位が欲しいので、ここで統計学者に助けを求めることが多いのだが、この段階ではもはや手遅れ（せいぜい言い訳の仕方を教えるくらいしかできない)。良くある悲劇である。

では、どうすれば良かったのだろうか？　この研究は、仮説検定という枠組みで行われた。サンプルサイズを大きくすれば統計的検出力が大きくなることは既知なので、研究開始前に、十分な統計的検出力を得るために必要なサンプルサイズを決定することが可能なはずである。

つまり、サンプルサイズの検討をせずに研究を開始したのが、そもそもの間違いだった。データを得てしまった後でできるのは、次のような言い訳を書くことだけである。

- 稀な疾患のため適当な研究期間内に同意を得られた患者が少なかった
- 資金面での制約からサンプルサイズを大きくできなかった
- その研究分野では伝統的にこの程度のサンプルサイズで研究が行われてきた、等。

こういう研究であっても、結果は将来のメタアナリシスに貢献するし、査読を通ることは時々ある。しかし、本当は、事前の計算から必要なサンプルサイズはこれくらいと予想されるけれども、稀な疾患なので期間内に研究対象とできる患者数はこれくらいと予想されるため、これだけの患者が集まったら分析する、と事前にデザインしておき、研究ノートあるいは短報の形で投稿すべきで

あろう[1]。

4.2　医学統計のテキストにはどう書かれているか？

最近の医学統計のテキストをみると、サンプルサイズが小さすぎる研究は、価値のある相関や差を示すのに十分な検出力をもっていないので**非倫理的**であるし、サンプルサイズが大きすぎる研究も、研究対象者に対して、既に劣っていることが明らかになっている治療（処置）を受けさせる可能性があるので、同様に**非倫理的**であると書かれている。

多くの学術雑誌が今では査読のチェックリストをもっていて、その中には、方法のセクションにサンプルサイズを決定する過程が含まれているかどうかについての項目が入っているし、サンプルサイズの決定が研究の前に行われていて、事後的になされたのではないことを確認するように書かれていることが多い。

有名な学術雑誌の一つである英国医学雑誌 (British Medical Journal) に、Altman らが 2000 年に示した統計学ガイドラインでは、“Authors should include information on … the number of subjects studied and why that number of subjects was used.”（著者は、……研究対象者の人数と、どうしてその対象者数が用いられたのかについての情報を論文中に含めるべきである）と書かれている。

4.3　サンプルサイズを計算しない理由付け

しかし、テキストの中には、サンプルサイズを計算しなくてよい、という言説もある。以下のような理由付けが与えられている。

- 皮肉な書き方をする統計家の中には、サンプルサイズの計算は、たんなる推定に数学という仮面を被せるだけだと言った人がいる。サンプルサイズの計算をするためには、結果として得られる数値の標準偏差のような要因についての情報を必要とすることがしばしばあり、もしその値が先行研究から得られない場合は、もっともらしい値を仮定するしかない。そして、サンプルサイズの計算結果は、そこで仮定される値のいくつかが少しでも変わると大きく違ってしまうことがある。

- サンプルサイズによらず、どんな研究であっても、なにがしかの情報を提供する。だから、いくつかのサンプルサイズの小さな研究をプールしてメタアナリシスをすることができるし、その方が、大きなサンプルサイズの１つの研究結果よりも一般化可能性が高い。

[1] きわめて重要なテーマなら原著にもなりうる。

- しばしばあるのは、研究対象とするサンプルサイズが、実用性によって決まってしまうことである。稀な病気について研究する場合は、研究対象として同意を得られる患者数が少ないかもしれないし、資金や時間による制約がある場合もある。

- 臨床試験も含めて、研究対象となっているアウトカムが複数ある（例えば、治療によって得られる利益と有害副作用等）場合、それぞれのアウトカム測定値が必要とするサンプルサイズは異なっていることがある[2]。

4.4 本当にサンプルサイズの計算が不要な研究もある

質的研究やケースレポート、あるいは予備的研究やパイロットスタディでは、統計学的な検定や推定をしない場合があって、そういう研究ではサンプルサイズの計算は不要である。

記述的な研究では、測定値についての事前情報が存在しないのが普通であるため、サンプルサイズの計算が不可能なことも多い（通常、サンプルサイズ計算には、測定値の標準偏差が必要）。

実験研究の場合、経験則として、各群 12 個体という目安はある（3 個体とか 6 個体にする場合もある。動物実験では純系の動物が使われるため、遺伝的多様性による個体差は無視できることが多く、比較的ばらつきが小さいからである）。主な要因によるクロス集計表を考え、セル（条件の組合せ）ごとの個体数が十分大きくなるような総個体数を確保せねばならない。

4.5 探索的研究では……

研究には、大別すると 2 つのタイプがある。既に説明したように、仮説検証では、研究前にサンプルサイズを計算することは常に必要である。

しかし、隠れた仮説を見つけ出したり、95%信頼区間を推定する記述的研究では、事前のサンプルサイズの計算は必ずしも要求されない。しかし、その場合は、サンプリングの適切性を評価するためのパワーアナリシス（検出力分析）を事後に実施することは可能である。

記述的研究では、小標本からの有病割合の推定は精度が低いか、ミスリーディングになることがある。例えば、有病割合を求めたいとき。20 人を調べて 2 人が病気だった場合、有病割合は 10%となるけれども、この 10%は信頼性が低い。病気の人が 1 人増減しただけで、有病割合が 5%も変わってしまう。サンプルサイズの計算をすることにより、十分に狭い信頼区間を得るためには何人調べれば良いかがわかる。

[2]通常は、そのうち最大のサンプルサイズを採用することで対応するが。

4.6　探索的研究の例

計算に必要な値は先行研究から得る。

ある狭さの信頼区間をもった平均値を推定するためのサンプルサイズを決定するのに必要な情報は次の 3 つである。

1. 測定する値の標準偏差 (SD)
2. 欲しい信頼区間の幅 (d)
3. 信頼水準（通常、90%、95%、あるいは 99%とする。$1-\alpha$ である）

このとき、必要なサンプルサイズ (n) は次の式で得られる（ただし、qnorm$(1-\alpha/2)$ は、正規分布の $1-\alpha/2$ パーセント点を意味する。通常、95%信頼区間を求めるには、97.5%点が必要になることに注意)。

$$n = \text{qnorm}(1-\alpha/2)^2 \times 4 \times \text{SD}^2/d^2$$

（例）ある患者群において、収縮期血圧の平均値を、95%信頼区間が 10 mmHg または 5 mmHg に収まるように推定したいとする。先行研究から、標準偏差としては 11.4 mmHg を使うことが妥当と考えられたとする。

95%信頼区間が 10 mmHg 幅でよければ、

$$n = 1.96^2 \times 4 \times 11.4^2/10^2 = 19.97... = 20$$

しかし 95%信頼区間を 5 mmHg 幅に収めたければ、

$$n = 1.96^2 \times 4 \times 11.4^2/5^2 = 79.88... = 80$$

つまり、精度を 2 倍にするには、サンプルサイズを 4 倍にする必要がある。

次に、割合を推定する場合を示す。必要な情報は先行研究と割合を推定する目的から決まる、次の 3 つである。

1. 母集団において期待される割合 (p)
2. 欲しい信頼区間の幅 (d)
3. 信頼水準 ($1-\alpha$)

必要なサンプルサイズを推定するための近似式は、

$$n = \mathrm{qnorm}(1-\alpha/2)^2 \times 4 \times p \times (1-p)/d^2$$

（例）成人集団における喘息の有病割合を、95%信頼区間が 10%幅に収まるように推定したいとする。母集団における喘息の有病割合が 10%とすると、$p = 0.1$、$d = 0.1$、$\alpha = 0.05$ なので、

$$n = 1.96^2 \times 4 \times 0.1 \times (1-0.1)/0.1^2 = 1.96^2 \times 36 = 138$$

4.7 仮説検定の原理

仮説検定においてサンプルサイズの計算に必要な情報は、次の 5 種類である。

1. 検定方法（帰無仮説と、片側か両側かを含む）
2. 第一種の過誤（α、帰無仮説が正しいのに誤って棄却してしまう確率）
3. 第二種の過誤（β、誤った帰無仮説を棄却するのに失敗する確率。$1-\beta$ が検出力）
4. 先行研究から期待される測定値（標準偏差や割合）
5. 臨床的（科学的）に意味があると判断される最小の差

検定方法ごとに（もっといえば、どれも近似式なので教科書あるいはソフトウェアによっても）推定式は異なっている。

例えば、両側 t 検定で平均値を比べるときは、推定される標準偏差を SD、意味のある差を d として（z_α は正規分布の 100α パーセント点を意味する。片側なら $z_{\alpha/2}$ のところが z_α となる）、

$$n = 2 \times (z_{\alpha/2} - z_{1-\beta})^2 \times \mathrm{SD}^2/d^2 + z_{\alpha/2}^2/4$$

カイ二乗検定で 2 つの標本比率の差を比べるときは、2 つの集団において期待される比率を p_1、p_2 として、

$$n = (z_{\alpha/2} - z_{1-\beta})^2 \times \{p_1(1-p_1) + p_2(1-p_2)\}/(p_1-p_2)^2$$

これらはまったく異なる式であることが一目瞭然である。式で手計算するよりも、サンプルサイズ計算に特化したソフト（nQuery、PASS、PS など）あるいは、一般的な統計解析ソフト（SAS、SPSS、STATA、EZR、R 等）を使う方が便利である。

4.8　仮説検定におけるサンプルサイズ計算の例

リハビリ中の患者を 2 群に分けて、電気刺激をしたときに、しないときに比べて、肘の屈曲角が大きくなるかどうか、平均値の比較をしたい例を考える。

先行研究から、屈曲が 4 度大きくなれば臨床的に重要な意味があると考えられ、屈曲角の増加の標準偏差は 5 度と考えられたとする。

「肘の屈曲角の増加に差は無い」という帰無仮説に対して有意水準 5%、検出力 90%の**片側** t 検定をするために必要なサンプルサイズは、

$$n = 2 \times (-1.64 - 1.28)^2 \times 5^2/4^2 + (-1.64)^2/4 = 27.3174$$

この結果から、各群 28 人いれば十分と考えられる（四捨五入でなく、それを上回る最小の人数であるべきなので、切り上げにすることに注意）。

実際には、諸事情により研究に参加した患者は、各群 28 人にはわずかに足りなかった。それでも、26 人の患者に電気刺激をし、25 人の患者に電気刺激をせずに、屈曲角の増加を測定したところ、彼らの屈曲角の増加は、それぞれ、16 ± 4.5 と 6.5 ± 3.4 であった。2 群の平均的な差は 9.5 (95%CI は 7.23 から 11.73) で、**片側** t 検定の結果は、t=8.43, df=49, $p < 0.001$ となり、有意水準 5% で統計学的に有意な差があったといえる。

4.9　このプロセスを英語論文に書くには

Methods セクションには以下のように書き、

> We designed the study to have 90% power to detect a 4-degree difference between the groups in the increased range of elbow flexion. Alpha was set at 0.05.

Results セクションには

> Patients receiving electrical stimulation (n=26) increased their range of elbow flexion by a mean of 16 degrees with a standard deviation of 4.5, whereas patients in the control group (n=25) increased their range of flexion by a mean of only 6.5 degrees with a standard deviation of 3.4. This 9.5-degree difference between means was statistically significant (95%CI = 7.23 to 11.73 degrees; one-sided Student's t test, t=8.43; df=49; $p < 0.001$).

のように書く (出典は 2 つの例文とも、Lang and Secic, 2006, pp.47；一部改変)[3]。

この英文には、期待される屈曲角の増加の標準偏差が 5 度であることや、計算に使われた式は書かれていない。ともに暗黙のうちに想定されるのが普通である。通常、計算に使ったソフトウェ

[3] デザインや題材や数字が独自になるのでコピーではないが、表現形式の作法があるため、このプロセスの英文表現がある程度似てしまうことは不可避である。iThenticate や Turnitin のような剽窃 (plagiarism) チェックソフトにかけると、既存論文との類似性を指摘される可能性があるが、この部分に関しては、たいていの場合、問題ない。

アとオプション指定を書いておけば、使った計算式も決まるので、式そのものを論文に書く必要は無い。

4.10 PS による計算

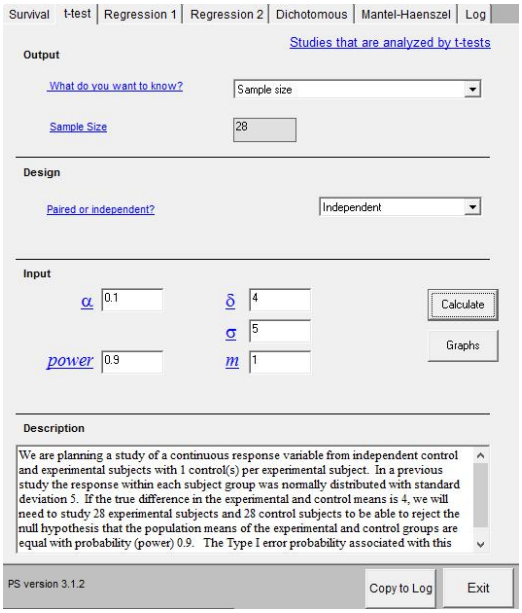
PS: Power and Sample Size Calculator は、ヴァンダービルト大学で開発、公開されている、サンプルサイズの計算に特化したフリーソフトである。

`https://biostat.app.vumc.org/wiki/Main/PowerSampleSize` からダウンロードできる。内蔵されている統計解析方法は、`Survival`（ログランク検定）、`t-test`、`Regression1`、`Regression2`、`Dichotomous`（カイ二乗検定またはフィッシャーの正確確率検定）、`Mantel-Haenszel`（マンテル＝ヘンツェルの要約カイ二乗検定）である。

特徴として挙げられるのは、サンプルサイズの計算を示す例文がテキストとして自動生成されることである[4]。なお、このソフトで指定できるオプションには、実験群と対照群のサイズの比 (m) も含まれている。m は 1 にすることもあるが、2 とか 3 にすることも珍しくない。

先に示した例でサンプルサイズを計算するには、PS を起動後、検定手法を選ぶタブの中から `t-test` を選び、求めるものは Sample Size、サンプルの独立性は Independent という順番で選び、δ に 4、σ に 5、α に 0.1（片側 5%なのだが、PS には片側検定の場合がないので両側 10%で計算する）、検出力に 0.9、m に 1 と入力してから、「Calculate」ボタンをクリックすると、必要サンプルサイズが 28 と計算され、例文も自動生成される（下図）。

[4] この場合、PS が生成する英文表現は、手法ごとに極めて似たものになるが、PS を引用しておけば問題ない。



4.11　EZR による計算

「統計解析」「必要サンプルサイズの計算」から条件を選ぶだけで完了する。この例では、メニューに表示される「2 群の平均値の比較のためのサンプルサイズの計算」を選んで、下図のように、「2 群間の平均値の差」のテキストボックスに 4、「2 群共通の標準偏差 (SD)」のテキストボックスに 5、「α エラー (0.0-1.0)」のテキストボックスに 0.05、「検出力 (1-β エラー）(0.0-1.0)」のテキストボックスに 0.9、「グループ 1 と 2 のサンプルサイズの比 (1:X)」のテキストボックスに 1 と入力し、解析方法のラジオボタンを One-sided にして「OK」ボタンをクリックすると、出力ウィンドウに各群 27 人という結果が得られる（One-sided のところが両側ならば 33 人となる)。

4.12 R コンソールでは

```
> power.t.test(delta=4, sd=5, sig.level=0.05, power=0.9,
   alt="one.sided")

     Two-sample t test power calculation

              n = 27.46584
          delta = 4
             sd = 5
      sig.level = 0.05
          power = 0.9
    alternative = one.sided

NOTE: n is number in *each* group
```

よって、必要なサンプルサイズは、各群 28 人となる。

4.13 G*Power による計算

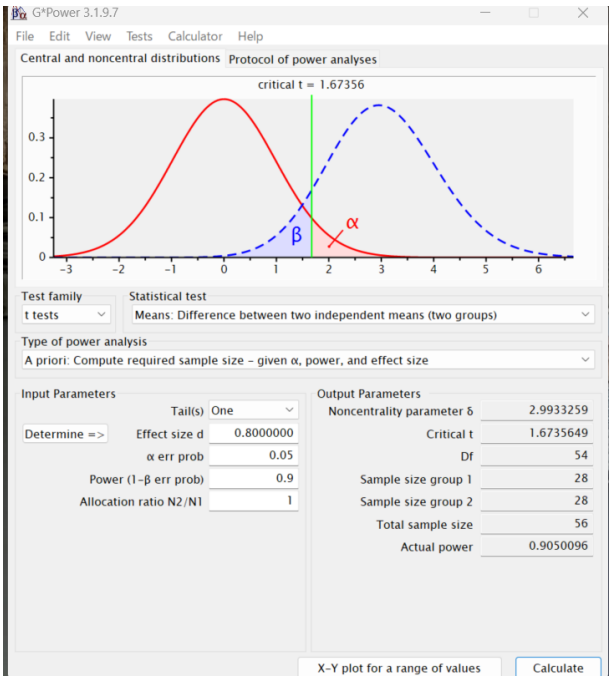
サンプルサイズや検出力の計算専用のソフトとして、上述の PS は最近更新されていないので、G*Power というソフトも便利である。ドイツのキール大学の Franz Faul によって開発されたソフトだが、多くの人が共同で維持管理しており、最新版は 3.1.9.7 である[5]。ここで取り上げられている例を G*Power で実行するには、起動後に、**Tests** というメニューの **Means** から **Two Independent Groups** を選択し、Type of power analysis というメニューが **A priori Compute required sample size - given α, power, and effect size** となっているのを確認し、Tail(s) というメニューで **One** を選択する。

次に Effect size d を入力しなくてはいけないのだが不明なので、それを計算するために、その行の左側に表示されている **Determine** というボタンをクリックする。すると、効果量を計算するためのダイアログが開く。この場合、平均値の差が 4、標準偏差が 2 群とも 5 と考えられるので、Mean group 1 に **0**、Mean group 2 に **4**、SD σ group 1 に **5**、SD σ group 2 に **5** と打って、左下の **Calculate** というボタンをクリックする。Effect size d として **0.8** と表示されるので、それを元のウィンドウの Effect size d の枠に手入力しても良いし、**Calculate and transfer to main window** というボタンをクリックすると自動入力される。

α err prob には想定している検定の有意水準 **0.05**、Power (1-β err prob) には想定している検出力 **0.9**、Allocation ratio N2/N1 には 2 群のサンプルサイズの比 **1** を入力し、右下の **Calculate** というボ

[5] https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower からマニュアルや本体がダウンロードできる。

タンをクリックすると結果が得られる。下図に示すように、Sample size group 1、Sample size group 2 とも 28 となっている。



G*Power では、PS や EZR ではサポートされていない、ロジスティック回帰分析のために必要なサンプルサイズの計算などもサポートされているので、使いこなせるようになると便利と思われる。

第5章　研究のデザイン

既に説明したとおり、研究には、大別すると、研究者が条件を設定しない研究（観察研究）と条件を設定する研究（実験研究や介入研究）がある。観察研究には、現状の記述を目的とする記述的研究と、仮説検証型の研究がある。

5.1　記述的観察研究のデザイン

- 母集団から適切なサンプリングを行うことに尽きる
- サンプルサイズが大きいほど精度が高まるがコストがかかる
- 稀な疾患の場合はある程度観察期間が必要（過去の記録を含む）

5.2　仮説検証型観察研究のデザイン

横断的研究 断面研究ともいう。英語では Cross Sectional Study。この場合も記述研究と同じく、適切なサンプリングをすることが肝要である。調べる項目が多いほどサンプルサイズも大きくするべき。

症例対照研究 患者対照研究とか、英語をそのままカタカナにしてケースコントロール研究 (Case Control Study) ともいう。原則として症例は定点で全数把握。いかに適切な対照を選択してサンプリングするかが鍵（病院対照、一般母集団対照等あるが一長一短。サンプリングには密度依存サンプリングなど。詳細は疫学の専門書参照）。基本的に、症例群と対照群で過去の曝露を比較するデザイン。

コホート研究 英語では Cohort Study。１つの曝露イベントに焦点を当て、曝露群と非曝露群を追跡し、疾病発生などの（複数ある場合もある）アウトカムを比較するデザインが典型的。大規模コホート研究と組み合わせたコホート内症例対照研究をすると効率が良いが、修士論文のように研究期間が短い場合は、コホート研究をデザインから実施することは困難と思われる。

5.3　実験研究のデザイン

実験的研究では、動物実験でも臨床試験でも、研究者が条件設定できるので、注意深くデザインされねばならない。実験研究のデザインは、実験計画法として発展してきた。

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

5.4　Fisher の三原則

繰り返し（Replication）： 各処理について最低 2 回以上の繰り返し測定が必要。

無作為化（Randomization）： 実験の順序や区画（農業試験の場合）は無作為に割り付けねばならない。

局所コントロール（＝ブロック化：blocking）： 大規模な実験の場合、サンプル全体の無作為化は不適切であり、代わりにいくつかのやり方で局所のブロックを作り、各ブロック内で無作為割り付けをすることで、ブロック間変動として、偏りを除去することができる。

5.5　実験計画の起源についての伝説

ケンブリッジのある晴れた日、多くの教授がアフタヌーンティーを楽しんでいた。と、ある婦人が、自分はミルクティーを飲めば、それがミルクが先か紅茶が先のどちらで淹れたものか判定できると主張した。世間一般の普通の人々は、そんなのどうでもいいじゃないかと思うかもしれないが、大学教授という人種は、そういうことに拘ることを生きがいにしている人が普通である。その場にいた教授たちの間で、当然のように、彼女の主張を廻って大激論が勃発した。先行研究があるわけでもなく、当然、収拾がつかない。

そこで R.A. Fisher が、「実験したらどうだい？」と言った。Fisher がいうには、この能力は、ミルクを先に入れて作ったミルクティーと紅茶を先に淹れて作ったミルクティーを何杯か用意して、無作為な順番で飲んで貰えば、確率論的に、どれくらい偶然ではありそうもないことかを判定できるというのだ。ここで大事なことは、能力の判定条件を考える必要があるということと、何回の繰り返しが必要かということだ。それを体系化した考え方が、上記 3 原則を含む実験計画法だというのである。

これは、「本当かどうかはわからないが」との注釈つきで、David Salsburg という人が、20 世紀において、どのようにあらゆる科学に統計学が影響を与えてきたかについて書いた、“The lady tasting tea.” に掲載されているエピソードである。

5.6 ミルクと紅茶の順番は本当に味に影響する？

ミルクティーの味の話がでてきたので、ここでちょっと余談。ロシア革命の混乱を風刺した小説『動物農場』や管理社会のディストピア小説として知られる『1984 年』の作者であるジョージ・オーウェルは、無類の紅茶好きであり、「完璧な紅茶を淹れる 11 の法則」(Perfect Cuppa)[1]を次のように書いている（ストレートのダージリンに代表される、普通の「おいしい紅茶」の淹れ方とは随分違うが)。

George Orwell 「完璧な紅茶を淹れる 11 の法則」 (Perfect cuppa)

1. Use tea from India or Ceylon (Sri Lanka), not China（茶葉はインドかスリランカ。中国茶ではいけない）
2. Use a teapot, preferably ceramic（ティーポットを使うこと。セラミックが望ましい）
3. Warm the pot over direct heat（ポットを直接の熱で温める）
4. Tea should be strong - six spoons of leaves per 1 litre（紅茶は濃くなくてはいけない。1 リットルの湯に対してティースプーン 6 杯の茶葉を使う）
5. Let the leaves move around the pot - no bags or strainers（茶葉はポットの中で自由に動けるようにする。ティーバッグや浸漬式の茶漉しは使わない）
6. Take the pot to the boiling kettle（ポットは沸騰しているケトルのところに持っていって、沸騰している湯を注げるようにする）
7. Stir or shake the pot（ポットの中をかきまぜ、揺する）
8. Drink out of a tall, mug-shaped tea cup（高さのある、マグ型のティーカップで飲む）
9. Don’t add creamy milk（クリーミーなミルクは入れない。ミルクとしては必ず新鮮な牛乳を使う）
10. Add milk to the tea, not vice versa（ミルクを紅茶に加える。逆ではいけない）
11. No sugar!（砂糖は入れない！）

しかし、英国の王立化学会が George Orwell の生誕 100 年を記念するパーティを開いたとき、Dr. Andrew Stapley (2003) は、次のように述べた。「冷えたミルクをカップの底に入れておいてから、熱い紅茶を注ぐのが良い。こうするとミルクが紅茶を冷ますことができる。逆だと熱い紅茶がミルクの温度を急に上げるのでミルクの風味が損なわれる」[2]。オーウェル推奨の順番は間違っているというわけである。

[1]出典：http://news.bbc.co.uk/2/hi/uk/3016342.stm で公開されている BBC News
[2]出典：前掲 BBC News

本当はどちらが先だとより美味しいのか、試してみた日本人ブロガーがいた。130cc の紅茶と 30cc のタカナシ低温殺菌牛乳を使用した（高温殺菌とかロングライフのミルクでは違いが分からないらしい）。この方の主観的判断では、「ミルクが先」が美味だったとのことである[3]。

5.7　白黒付けるには何杯飲めばいい？

1 杯ずつ増やして確率計算をしてみよう。

- 1 杯飲んでそれが偶然正しい判定になる確率は、50%である。
- 2 杯飲んで 2 杯とも偶然正しい判定になる確率は、25%である。
- 3 杯飲んで 3 杯とも偶然正しい判定になる確率は、12.5%である。
- 4 杯飲んで 4 杯とも偶然正しい判定になる確率は、6.25%である。
- 5 杯飲んで 5 杯とも偶然正しい判定になる確率は、3.125%である。

というわけで、本当は判別能力が無いのに偶然５回連続で正解する確率が、3.125% とわかる。この値は、偶然で片付けるには稀すぎる。通常、この判定基準は 5%を切るかどうかにおく。これが有意水準（Fisher 流）である。5 回とも正しく判定されれば、帰無仮説「彼女は判定能力をもっていない」が有意水準 5%で棄却される。

つまり、この仮説の統計学的検定には、少なくとも 5 杯のミルクティーを飲む必要がある。

5.8　有名な実験計画デザイン

実験計画法には、目的に応じて様々なデザインがある。有名なデザインをいくつか列挙しておく。

- 単一群、事前-事後デザイン
- 平行群間比較試験（完全無作為化法）
- 乱塊法
- 要因配置法
- ラテン方格法
- クロスオーバー法

[3] http://blog.livedoor.jp/teatime312/archives/cat_123365.html

5.8.1 単一群、事前-事後デザイン

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

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

EZR での対応のある *t* 検定の実行例

EZR で、対応のある *t* 検定を実行する例を示す。対応のある *t* 検定自体は、事前-事後デザインに限らず、同一対象者に対して同じ精度で測定した値があれば適用可能なので、ここでは、R には標準で含まれている MASS パッケージの、survey データフレームの、左右の手の大きさを比べる例を示す（このデータはアデレード大学で統計学を受講している学生 237 人に対する質問紙による横断的研究で得られた値であり、事前-事後デザインではない。出典：Venables and Ripley, 1999）。
このデータに含まれる変数のうち、Wr.Hnd は、字を書く方の手を広げたときの大きさ、即ち親指の先端と小指の先端の距離であり、NW.Hnd は、反対側の手を広げたときの大きさである。
EZR での手順としては、まずデータをアクティブにする。「ツール」の「パッケージの読み込み」で MASS を選んでから、「ファイル」の「パッケージに含まれるデータを読み込む」から、パッケージとして MASS をダブルクリックし、データフレームとして survey をダブルクリックして「OK」ボタンをクリックすればよい。
次に、「統計解析」「連続変数の解析」から「対応のある 2 群間の平均値の比較 (paired *t* 検定)」を選ぶ。第 1 の変数として Wr.Hnd、第 2 の変数として NW.Hnd を選び（逆でも良い）、「OK」ボタンをクリックすると、下枠内の結果が表示される。文字を書く手の方が、反対側の手よりも平均して 0.086cm 大きく、この差は統計学的に有意水準 5%で有意であるとわかる。

```
Paired t-test

data:  survey$Wr.Hnd and survey$NW.Hnd
t = 2.1268, df = 235, p-value = 0.03448
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.006367389 0.166513967
sample estimates:
mean of the differences
              0.08644068
```

既に述べたように、対応のある *t* 検定は差の平均がゼロという一標本 *t* 検定と数学的にまったく同じである。数式でも簡単に説明しておく、サンプルサイズ n の標本の i 番目の人について、対応のある 2 つの変数 X と Y の値を X_i、Y_i と書けば、X と Y に差がないという帰無仮説は、$Z_i = X_i - Y_i$ **として計算される「差の変数」Z の標本平均が母平均 0 と差がない**という帰無仮説に帰着する。このとき、検定統計量 t_0 は、

$$E(Z) = \frac{\sum_{i=1}^{n} Z_i}{n}$$

として、

$$t_0 = \frac{E(Z)}{\sqrt{\frac{\sum_{i=1}^{n}(Z_i - E(Z))^2/(n-1)}{n}}}$$

が自由度 $n-1$ の t 分布に従うことを使って検定できる。

練習問題

珈琲を飲む前後で、計算問題の得点が変わるかどうか比べる。EZR では、まずデータを作る。「ファイル」「新しいデータ」を選んでデータセット名を入力し（例えば coffee などと名付ける）、表計算のようなウィンドウでデータを入力する（1 列目に珈琲を飲む前の得点を入力し、2 列目に飲んだ後の得点を入力する）。各列の変数名の部分をクリックすると、変数名を入力したり、型が数値 (numeric) か文字 (character) かを指定することができる。
または、コンソールに

```
> scores <- data.frame(
+  precoffee=c(6, 5, 7, 6, 6, 7, 4, 5, 6, 7),
+  postcoffee=c(7, 8, 6, 7, 7, 8, 5, 6, 7, 7))
```

と入力してもよい。
EZR の「統計解析」から「連続変数の解析」「対応のある 2 群間の平均値の比較 (paired *t* 検定)」と選んで、第 1 の変数として precoffee、第 2 の変数として postcoffee を選べば、[t = -2.862, df = 9, p-value = 0.01872] と結果が出る。有意水準 5%で統計学的に有意な差があるといえる。

5.8.2 平行群間比較試験（完全無作為化法）

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

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

しかし、今ではコンピュータソフトが使えるので、紙の表を使わなくても、簡単にランダム割り付けはできる。R の場合だと、例えば 45 人の対象者に対して 3 種類の処理を割り付けたいなら、次のようにする。

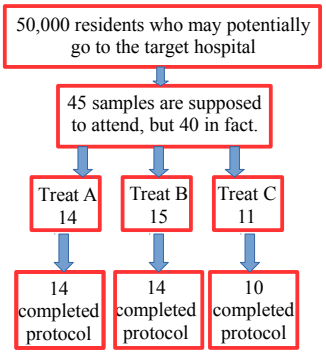
```
> matrix(sample(1:45, 45, replace=FALSE), 3)
```

このままだと数字の出現順がバラバラで見にくいので、いったん x というオブジェクトに結果を付値して、それを行方向に小さい順に並べ替えて表示させるには以下のようにする。

```
> x <- matrix(sample(1:45, 45, replace=FALSE), 3)
> x[1,]
[1]  31 34 23  5 17 44 27 14  7 30 16  1 33  4 19
> apply(x, 1, sort)
      [,1] [,2] [,3]
 [1,]    1    2    8
 [2,]    4    3   10
 [3,]    5    6   11
 [4,]    7    9   12
 [5,]   14   13   20
 [6,]   16   15   21
 [7,]   17   18   22
 [8,]   19   24   26
 [9,]   23   25   28
[10,]   27   29   35
[11,]   30   32   39
[12,]   31   36   41
[13,]   33   37   42
[14,]   34   38   43
[15,]   44   40   45
```

しかし、実際の研究では脱落が起こったりして、予定の人数に満たない場合がある。上の例でも、40 人しか対象者が集まらないと、第 2 の処理を受ける人は 15 人いるが、第 1 の処理を受ける人は 14 人、第 3 の処理を受ける人はたった 11 人になってしまい、サンプルサイズがアンバランスになって検出力が落ちる。

このデザインによる研究の流れを下図に示す。このダイアグラムでは、研究対象者から得られる量的なデータを一元配置分散分析 (oneway ANOVA) で分析し、割合のデータをカイ二乗検定で分析する。



5.8.3 乱塊法 (randomized block(s) design

何らかの事情で研究が完了しない場合、完全無作為化法では、既に述べたようにグループ間のサンプルサイズがアンバランスになる可能性がある。

この欠点を克服するのが乱塊法である。上述の例では、3 種類の処理を 15 人ずつに適用する。3 種類の処理を実施する順番は 6 通りなので[4]、1 人ずつランダムに選ぶ代わりに、この 6 通りのブロックをランダムに 15 回選べば良いと考えられる。

そうすることによって、もし研究が途中で終わってしまっても、群間のサンプルサイズの差は最大 1 に抑えられる。記述と分析は完全無作為化法と基本的に同じだが、ブロックの効果を考慮した分析も可能である。

サイズバランスを保つ方法はもう一つあり、「最小化法」と呼ばれる。サンプリング時点ごとに群間のサンプルサイズの違いが最小になるような制約条件のもとでランダム割り付けを行う（あまりお薦めしない）。

5.8.4 要因配置法 (Factorial design)

2 × 2 の要因配置法の例を示す。McMaster ら (1985) は、乳がんの自己触診の教材としての小冊子とテープ／スライドを評価する無作為化試験を実施した。2 種類の教材があるので、2 × 2 の組

[4]例えば処理が A, B, C の 3 種類であれば、[A, B, C], [A, C, B], [B, C, A], [B, A, C], [C, A, B], [C, B, A] の 6 通りのブロックを考えることができる。

合せで、4 種類の平行群間試験としての実験をデザインできる。即ち、

1. 小冊子もテープ／スライドも使わない（対照群）
2. 小冊子を提示する
3. テープ／スライドを使った教育プログラムを使う
4. 両方を使う

このデザインでは、2 種類の教材の教育効果は二元配置分散分析 (Two-way ANOVA) を使って評価できる。

5.8.5　ラテン方格法 (Latin-square design)

実験において、2 つ以上の水準 p がある処理 A の効果を評価したいとき、ともに同じ水準数 p をもつ交絡因子 B と C の影響を調整する必要がある場合に、このデザインが有効である。ラテン方格という名前は古代のパズルに起源がある。

ここでは p が 3 であると仮定して説明する。ラテン方格は下図のようになる。最初のグループとしては、n1 人の対象者について、a1b1c1 という組合せの処理を行う。次の n2 人は、グループ 2 として、a1b2c3 という処理を受ける。残りも同じように図示した順番で処理を行う。そうすれば、アウトカム変数に対する B と C の効果を、分散分析 (ANOVA) によって調整（除去）して、要因 A の効果を評価することができる（デザインの工夫によって B と C の効果は打ち消せると期待されるので、ターゲットである A の効果だけを一元配置分散分析すれば良くなる）。

	c1	c2	c3
b1	a1	a2	a3
b2	a2	a3	a1
b3	a3	a1	a2

5.8.6　クロスオーバー法 (cross-over design)

クロスオーバー法では、それぞれの対象者が 2 種類の処理を受ける。このとき、適当な間隔（ウォッシュアウト期間と呼ぶ。前の処理の影響のキャリーオーバーを避けるために設ける）をおくこと

と、処理の順番が違う２つのグループを設定することが必要である。

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

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

統計解析は、プレテストとして (1) キャリーオーバー効果が無視できるか検定（２群それぞれの２つの測定値の和の平均値の差の検定)、それが確認できたら、(2) ２群それぞれの差の平均値の差の検定、と実施すると良い (参考：Wellek S, Blettner M: *Dtsch Arztebl Int.* Apr 2012; 109(15): 276-281. doi: 10.3238/arztebl.2012.0276[5])

5.9 結果の評価のタイプ

週刊医学界新聞の記事[6]が大変参考になるので読まれたい。

臨床試験を含む実験研究では、アウトカム（結果）の評価にいくつかのタイプがあり、それぞれ検定すべき帰無仮説が異なるため、サンプルサイズの設計法も異なる。

優性試験は、新しい処理群が対照群に比べて統計学的に有意に良い効果を示すかどうかを調べる。新薬の臨床試験などの場合に用いられる。「差がない」帰無仮説を検定して、p 値が有意水準より小さければ帰無仮説を棄却する。

同等性試験は、新しい処理群が対照群と同じような効果を示すかどうかを調べる。検定でなく、「十分なサンプルサイズ」で正確に同等だというために信頼区間を用いる。事前に決めた「±○×%の差であれば臨床的に同等」とみなす同等性の許容範囲（同等性マージン、研究計画書にも記載）内なら同等とみなす。

非劣性試験は同等性試験の特殊な場合で、新しい処理が対照群に比べて劣っていないことを示せばいいとき、信頼区間を片側にすることでサンプルサイズを節約できるというアイディアに基づく。例えば、安価なジェネリック医薬品が従来薬に比べて薬効が劣っていなければいいという場合が典型的である。

[5] http://dx.doi.org/10.3238/arztebl.2012.0276
[6] http://www.igaku-shoin.co.jp/paperDetail.do?id=PA02971_04

5.10　効果量

実験計画法に基づく研究結果は、これまで書いてきたとおり、p 値と帰無仮説の棄却、あるいは、差の信頼区間で示すのが普通である。しかし、1994 年に Cohen が “The earth is round ($p < .05$)” という論文を書いて有意水準批判をしたことから、アメリカ心理学会 (APA) の推測統計に関する専門委員会が、統計解析の結果を報告する際のガイドラインを検討し、1999 年に発表された提案において、信頼区間を使用した区間推定を示すことに加えて、重要な知見あるいは p 値を報告するときは必ず効果量を報告することを含めたことから、心理学分野では統計改革が起こった。

APA Publication Manual 6th Edition (2009) では、具体的に、信頼区間をブラケットで示すことや、p 値の後に効果量を記載することが APA が発行する論文誌では最低限の要求であると書かれている。統計改革の流れは他分野にも波及したが、医学・保健学では信頼区間の重要性が強調され、生態学などフィールド生物学では統制が難しくサンプルサイズが小さいことが多いため検定力についての記載が求められることが多いので、効果量を記載することが必須になったのは、現状では、心理学分野と、質問紙調査で心理的尺度を用いる教育学や社会調査、疫学の一部にほぼ限られている[7]。

大久保・岡田 (2012) による効果量 (effect size) の定義は、「効果の大きさをあらわす統計的な指標のこと」で、「p 値や検定統計量とは異なり、帰無仮説が正しくない程度を量的に表す指標」で、「帰無仮説が完全に正しい場合、一般に効果量は 0、帰無仮説が正しくない場合は、正しくない程度が大きいほど効果量の絶対値も大きくなる」とのことである。仮説検定で帰無仮説が棄却できなくても効果量が 0 とは限らない点に注意が必要である。

一般に、**検定統計量はサンプルサイズの関数と効果量の関数の積として表される**ことから、サンプルサイズが同じなら、効果量が大きいほど検定統計量は大きくなり、p 値が小さくなることと、効果量が同じならば、サンプルサイズが大きいほど検定統計量が大きくなり p 値が小さくなることが言える。サンプルサイズが大きくなると、ほとんど無視できるほど小さい効果量であっても、p 値が 0.05 より小さくなることは普通にありうるので、数万人以上のサンプルサイズで有意水準を 5%にした仮説検定などしても、実際的な意味はない。

効果量はサンプルサイズに依存しない（検定統計量からサンプルサイズに依存する部分を除去したものが効果量であるともいえる）のは、重要な性質である。以下、典型的な効果量をいくつか示す。

[7] 詳しくは、大久保街亜・岡田謙介 (2012)『伝えるための心理統計：効果量・信頼区間・検定力』勁草書房、ISBN978-4-326-25072-1 を参照されたい。水元篤・竹内理「研究論文における効果量報告のために—基礎的概念と注意点—」英語教育研究, 31: 57-66, 2008.http://www.mizumot.com/files/EffectSize_KELES31.pdf も参考になる。

5.10.1 *d* 族の効果量

群間差についての効果量を ***d* 族の効果量**と呼ぶ。三重大学奥村教授の解説[8]が参考になるが、２群間の平均値の差についての **Cohen の** *d* と **Hedges の** *g* が有名である。

Cohen の *d* は、標本の平均値の差を、標本のプールした標準偏差[9]で割って得られる、記述的な効果量である。言い換えると、２群の平均値の差が、プールした標準偏差の何倍かという値が Cohen の *d* である。Hedges の *g* は、Cohen の *d* を求める式において、標本のプールした標準偏差の代わりに、２群に共通な母分散を標本から推定する際の不偏推定量[10]を用いて得られる値である。s_p は不偏推定量ではないため、*g* も不偏推定量とはならないが、母標準偏差との差は s_p の方が S_p よりも小さいことが知られているので、*g* の方が *d* より推奨されている。２群のサンプルサイズが等しい場合は、*g* に補正係数を掛けることによって得られる、バイアス補正した *g*（$\hat{\delta}$）を用いることもできる。補正係数 $J(n_1 + n_2 - 2)$ は次の式で得られる。

$$J(n_1 + n_2 - 2) = \frac{\Gamma((n_1 + n_2 - 2)/2)}{\sqrt{(n_1 + n_2 - 2)/2}\Gamma((n_1 + n_2 - 3)/2)}$$

Hedges(1981) は $J(n_1 + n_2 - 2)$ の近似式として

$$1 - \frac{3}{4(n_1 + n_2) - 9}$$

を提案している。

R では、`effsize` パッケージの `cohen.d()` 関数で、２群の値を示すベクトルをコンマで区切って与えるか、量の変数を第１引数、群の変数を第２引数としてコンマで区切って与えるかすれば、Cohen の *d* または Hedges の *g* が得られ（後者が欲しいときはオプションとして `hedges.correction=TRUE` を指定する）、それらの効果量の信頼区間も得られ、効果量の大きさの目安として、絶対値が 0.2 未満だと無視できる (negligible)、0.2 以上 0.5 未満だと小さい (small)、0.5 以上 0.8 未満だと中程度 (medium)、0.8 以上だと大きい (large) という判定を表示してくれて便利である。ただし、`cohen.d()` が Cohen's d として表示するのは大久保・岡田 (2012) でいう Hedges の g であり、Hedges's g として表示されるのがバイアス補正された Hedges の g であることに注意されたい。

---

[8] `https://oku.edu.mie-u.ac.jp/~okumura/stat/effectsize.html`

[9] 第１群、第２群の標本分散をそれぞれ ${S_1}^2$、${S_2}^2$ と書き、それぞれのサンプルサイズを n_1、n_2 と書くと、

$$S_p = \sqrt{\frac{n_1 {S_1}^2 + n_2 {S_2}^2}{n_1 + n_2}}$$

となる。

[10] 第１群、第２群の不偏分散をそれぞれ ${s_1}^2$、${s_2}^2$ と書き、それぞれのサンプルサイズを n_1、n_2 と書くと、

$$s_p = \sqrt{\frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}}$$

となる。

https://minato.sip21c.org/ebhc/dfamefs.R

```
dfamefs <- function(G1, G2) {
 varp <- function(X) { sum((X-mean(X))^2)/length(X) }
 J <- function(alpha) {
  gamma(alpha/2)/(sqrt(alpha/2)*gamma((alpha-1)/2)) }
 n1 <- length(G1)
 n2 <- length(G2)
 Sp <- sqrt((n1*varp(G1)+n2*varp(G2))/(n1+n2))
# Sp2 <- sqrt((n1*var(G1)+n2*var(G2))/(n1+n2))
# Sp3 <- sqrt(((n1-1)*(sd(G1)^2)+(n2-1)*(sd(G2)^2))/(n1+n2))
 d <- (mean(G1)-mean(G2))/Sp
# d2 <- (mean(G1)-mean(G2))/Sp2
# d3 <- (mean(G1)-mean(G2))/Sp3
 sp <- sqrt(((n1-1)*var(G1)+(n2-1)*var(G2))/(n1+n2-2))
 g <- (mean(G1)-mean(G2))/sp
 gadj <- g*(1-3/(4*(n1+n2)-9))
 gadj2 <- g*J(n1+n2-2)
 delta <- (mean(G1)-mean(G2))/sd(G2) # G2 has to be control
 deltaadj <- (1-3/(4*length(G2)-5))*delta
 return(list(Cohend=d, Hedgesg=g, gadj=gadj, gadjexaxt=gadj2,
  Glassdelta=delta, deltaadj=deltaadj))
}

library(effsize)
G1 <- sleep$extra[sleep$group==1]
G2 <- sleep$extra[sleep$group==2]

cohen.d(G1, G2)
# cohen.d(sleep$extra, sleep$group)

cohen.d(G1, G2, hedges.correction=TRUE)
# cohen.d(sleep$extra, sleep$group, hedges.correction=TRUE)

dfamefs(G1, G2)

# Okubo, Okada (2012) Table 3.4
Exp <- c(59, 48, 51, 41, 39, 84, 95, 56, 86, 74)
Ctl <- c(47, 24, 38, 28, 39, 74, 77, 48, 40, 60)

cohen.d(Exp, Ctl)
cohen.d(Exp, Ctl, hedges.correction=TRUE)
dfamefs(Exp, Ctl)
}
```

結果は以下の通り。

```
> cohen.d(G1, G2)
Cohen's d

d estimate: -0.8321811 (large)
95 percent confidence interval:
     lower      upper
-1.8115649  0.1472027

> cohen.d(G1, G2, hedges.correction=TRUE)
Hedges's g

g estimate: -0.7970185 (medium)
95 percent confidence interval:
     lower      upper
-1.7731697  0.1791327

> dfamefs(G1, G2)
$Cohend
[1] -0.8771959

$Hedgesg
[1] -0.8321811

$gadj
[1] -0.7970185

$gadjexaxt
[1] -0.7969352

$Glassdelta
[1] -0.7891127

$deltaadj
[1] -0.7214745
```

```
> # Okubo, Okada (2012) Table 3.4
> Exp <- c(59, 48, 51, 41, 39, 84, 95, 56, 86, 74)
> Ctl <- c(47, 24, 38, 28, 39, 74, 77, 48, 40, 60)
>
> cohen.d(Exp, Ctl)
Cohen's d

d estimate: 0.8321811 (large)
95 percent confidence interval:
     lower      upper
-0.1472027  1.8115649

> cohen.d(Exp, Ctl, hedges.correction=TRUE)
Hedges's g

g estimate: 0.7970185 (medium)
95 percent confidence interval:
     lower      upper
-0.1791327  1.7731697

> dfamefs(Exp, Ctl)
$Cohend
[1] 0.8771959

$Hedgesg
[1] 0.8321811

$gadj
[1] 0.7970185

$gadjexaxt
[1] 0.7969352

$Glassdelta
[1] 0.8831702

$deltaadj
[1] 0.8074699
```

5.10.2 *r* 族の効果量

変数間の関係の大きさを表す効果量を ***r* 族の効果量**と呼ぶ。ピアソンの積率相関係数やポリシリアル相関係数、ポリコリック相関係数、回帰分析における決定係数、偏相関係数の二乗、分散分析における η^2 や ${\eta_p}^2$ などは、すべて *r* 族の効果量である。

第6章　データ入力・記述統計・図示

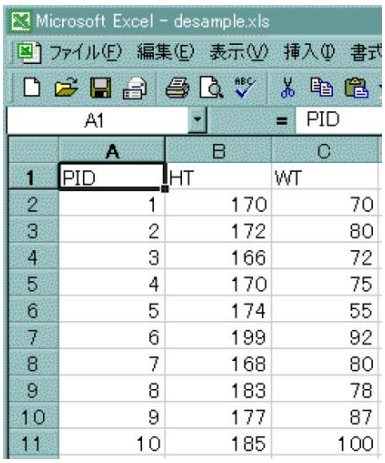
6.1　データ入力

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

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

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

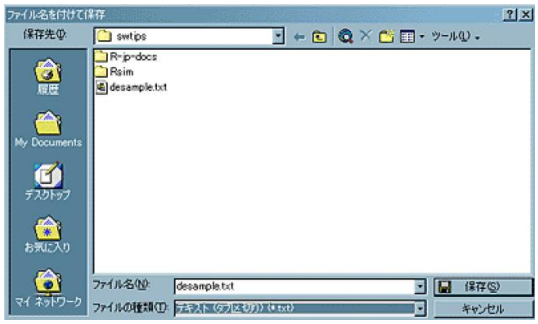
対象者 ID	身長 (cm)	体重 (kg)
1	170	70
2	172	80
3	166	72
4	170	75
5	174	55
6	199	92
7	168	80
8	183	78
9	177	87
10	185	100



まずこれを表計算ソフトに入力する。一番上の行には変数名を入れる。日本語対応 R なら漢字

やカタカナ、ひらがなも使えるが、半角英数字（半角ピリオドも使える）にしておくのが無難である。入力が終わったら、一旦、そのソフトの標準の形式で保存しておく（Excel ならば*.xlsx 形式、LibreOffice の Calc ならば*.ods 形式）。入力完了した状態は、右の画面のようになる。

次に、この表をタブ区切りテキスト形式で保存する。Excel の場合は、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 xlsx から txt に変わるので、「保存 (S)」ボタンを押せば OK である（下のスクリーンショットを参照）[1]。複数のシートを含むブックの保存をサポートした形式でないとかいう警告が表示されるが無視して「はい」を選んでよい。その直後に Excel を終了しようとすると、何も変更していないのに「保存しますか」と聞く警告ウィンドウが現れるが、既に保存してあるので「いいえ」と答えてよい（「はい」を選んでも同じ内容が上書きされるだけであり問題はない）。この例では、desample.txt ができる。



R コンソールを使って、このデータを Dataset という名前のデータフレームに読み込むのは簡単で、次の 1 行を入力するだけでいい（ただしテキストファイルが保存されているディレクトリが作業ディレクトリになっていなくてはいけない）。

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

[1] LibreOffice Calc の場合は、メニューバーの「ファイル (F)」から「名前を付けて保存 (A)」を選び、現れるウィンドウで「ファイルの種類 (T)」から「テキスト csv」を選び、適切なファイル名を付けて（拡張子も含めて）から「保存 (S)」を選ぶ。警告メッセージが表示されるが OK すると「テキストファイルのエクスポート」というダイアログが表示されるので、ここで「フィールドの区切り記号 (F)」として {タブ} を選び、「テキストの区切り記号 (T)」を削除してから「OK」ボタンをクリックする。

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

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

6.1.1 表形式では扱いにくいデータ

断面研究が典型的だが、１つの時点で多くの対象者から得たデータは、上述のように表形式にするのが容易である。複数のグループがあっても、グループを示すカテゴリ変数を作り、ともかく 1 人が 1 行に収まるようにすればよい。

コホート研究や繰り返し測定を含むデータの場合は、異なる時点で得たデータをどのように保持するかによって表の形が変わるため（あくまで 1 人を 1 行として入力する方が、後述する反復測定分散分析やフリードマンの検定をするには便利だし、1 時点での測定値を 1 行とし、同じ人が複数行に出てくるように入力した方が作図や解析の目的次第では便利である）、表形式にするよりもリレーショナルテータベースとしてデータベース構造を定義し、解析目的に応じてさまざまな表を出力できるようにした方が便利である。表計算ソフトを使う場合でも、時点ごとに別々のシートに入力すれば、ある程度この目的に対応できる。リレーショナルデータベースとしては、Microsoft Office ならば Access、LibreOffice ならば base として含まれている。米国 CDC が開発して無料で公開している EpiInfo では[2]、[Enter data] というメニューから、Access 形式のデータベースを設計して入力・編集することができるようになっている。

経時的にとったデータの推移や、複数の時系列の相互作用を扱いたい場合は、データ構造が複雑になる。国によって暦法など時間情報を扱う単位が異なっていたり時差があったりするため、時間データベースが標準化されていて、時間情報解析ソフトというものも開発されている。HuTime と

[2] https://www.cdc.gov/epiinfo/index.html

いうソフト[3]は無料で利用できる。

地理情報もまた、表形式では扱いにくい情報である。バックグラウンドとなる地理情報と関連づけるためには、地理情報解析システム (GIS; Geographic Information System) を利用すると良い（詳細は非常勤でご出講いただく谷村先生の講義を参考にされたい）。高度な機能を使いたい場合は ArcInfo というきわめて高価なソフトが標準的なソフトとして広まっているが、基本的な解析だけならば、QGIS[4]などのフリーソフトでも十分に実行できる。

6.2　入力ミスを防ぐためのデータ入力の原則

なお、データ入力は、入力ミスを防ぐために、2 人以上の人が同じデータを入力し、それを比較するプログラムを実行して誤りをチェックする方法がよいとされる。Excel のワークシートが 2 枚できたときに、それらを比較するには、1 つのブックの Sheet1 と Sheet2 にそれらを貼り付けておき、Sheet3 の一番左上のセル（A1）に、

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

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

しかし、現実には 2 人の入力者を確保するのが困難なため、1 人で 2 回入力して 2 人で入力する代わりにするか、あるいは 1 人で入力してプリントアウトした結果を元データと見比べてチェックするといった方法が使われることも多い。生データを自分で読み上げて録音し、再生音を聞きながら入力したデータをチェックする方法も、比較的効率は良い。

6.3　欠損値の扱い

ここで注意しなければならないのは、欠損値の取扱いである。一般に、統計処理をする対象のデータは、母集団から標本抽出したサンプルについてのものである。サンプルデータを統計解析して、母集団についての情報を得るためには、そのサンプルが正しく母集団を代表していることが何より大切である。質問紙調査の場合でも、実験研究の場合でも、欠損値（質問紙なら無回答、非該当、わからない、等、実験研究なら検出限界以下、測定用の試料の量の不足、測定失敗等）をどのように扱うかによって、サンプルの代表性が歪められてしまうことがある。欠損が少なければあま

[3] http://www.hutime.jp
[4] http://www.qgis.org/ja/site/

り気にしなくていいが、たとえば、健診の際の食生活質問等で、「甘いものが好きですか」に対して無回答の人は、好きだけれどもそれが健康に悪いと判断されるだろうから答えたくない可能性があり、その人たちを分析から除くと、甘いもの好きの人の割合が、全体よりも少なめに偏った対象の分析になってしまう。なるべく欠損が少なくなるような努力をすべきだけれども、どうしても欠損のままに残ってしまった場合は、結果を解釈する際に注意する。

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

欠損コードの変更

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

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[5]（もちろん、非該当は欠損値ではあるが外してはならない）。その場合、統計ソフトに渡す前の段階で、そのケースのデータ全体（Excel 上の１行）を削除してしまうのが簡単である（もちろん、元データは別名で保存しておいて、コピー上で行削除）。質問紙調査の場合、たとえば 100 人を調査対象としてサンプリングして、調査できた人がそのうち 80 人で、無回答項目があった人が 5 人いたとすると、回収率 (recovery rate) は 80%（80/100) となり、有効回収率 (effective recovery rate) が 75%（75/100）となる。調査の信頼性を示す上で、これらの情報を明記することは重要である。目安としては有効回収率が 80%程度は欲しい。

もう少し厳密に考えると、上述のごとくランダムでない欠損は補正のしようがないが、欠損がランダムな場合でも 2 通りの状況を分けて考える必要がある。即ち、MISSING COMPLETELY AT RANDOM (MCAR) の場合は単純に除去しても検出力が落ちるだけでバイアスはかからないが、MISSING AT RANDOM (MAR)、つまり欠測となった人とそうでない人の間でその変数の分布に

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

は差が無いが他の変数の分布に差がある場合には、単純に除去してしまうとバイアスがかかるのである。そのため、multiple imputation（多重代入法）という欠損値を補う方法がいろいろ開発されている[6]。R では、`mitools`[7]と `mice`[8]という 2 つのパッケージがあり、後者のメインテナは Dr. Stef van Buuren というオランダの方で、実は Multiple Imputation Online というサイト[9]の Head をされている専門家である。ネット上の情報を眺めていると、たぶん `mice` パッケージを使うのが良いと思われるが、まだ評価は定まっていないようである。最近では `Amelia` というパッケージがあり、高橋将宜・渡辺美智子『欠測データ処理：R による単一代入法と多重代入法』共立出版、ISBN978-4-320-11256-8 で丁寧に使い方が紹介されている。

例えば、欠損のあるデータフレーム名を withmiss とすると、`library(mice)` の後で、

```
imp <- mice(withmiss)
```

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

無料で読める日本語による欠損値処理の説明としては、覚え書きとのことだが、村山航 (2011) 欠損データ分析 (missing data analysis) ―完全情報最尤推定法と多重代入法―[10]がわかりやすいと思う。英語だと cran の Missing Data の Task View[11]や finalfit というパッケージの vignette[12]がコンパクトでわかりやすい。

MAR なのか MCAR なのかを調べる Little の MCAR 検定については、`mvnmle` パッケージや `BaylorEdPsych` パッケージに含まれていた `LittleMCAR()` という関数があったが、どちらのパッケージも既にメンテナンスされなくなって久しく、cran から削除されている（Github の cran アーカイヴに入っているので、`devtools` パッケージの `install_github()` 関数を使って `cran/mvnmle` のような指定をすればインストールできるが、メンテされていないパッケージを使うより、Little RJA (1988) A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83: 1198-1202.[13]を読んで自分でコーディングした方が良い）。

[6]欠損でない人の平均値を代入するなどの単純代入法と呼ばれる方法もあるが、お勧めしない。
[7]http://cran.r-project.org/web/packages/mitools/index.html
[8]http://cran.r-project.org/web/packages/mice/index.html
[9]http://www.multiple-imputation.com/
[10]https://koumurayama.com/koujapanese/missing_data.pdf
[11]https://cran.r-project.org/web/views/MissingData.html
[12]https://cran.r-project.org/web/packages/finalfit/vignettes/missing.html
[13]https://doi.org/10.2307/2290157

Task View に書かれている通り、RBtest パッケージ[14]でも MCAR 性の検定ができるが、典拠となる文献が書かれていないし、回帰に基づく方法ということで Little の方法とは違い、信頼できるかまだわからない。

6.4 図示

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

何よりも先に図示をすべき

といえる。変数が表す尺度の種類によって、さまざまな図示の方法がある。離散変数の場合は、度数分布図、積み上げ棒グラフ、帯グラフ、円グラフが代表的であり、連続変数の場合はヒストグラム、正規確率プロット、箱ひげ図、散布図が代表的である。

以下、MASS パッケージに含まれている survey というデータフレームを使って作図する例を示す。

6.4.1 グラフの色について

実際の作図に入る前に、色について簡単にまとめておく。修論などを見ていると、時々、カラー印刷される前提で、多数の色を使った図を見かけることがある。しかし、印刷物においてはモノクロになっても判別できるようにしておくことが重要である。投稿論文ではカラー図版には追加料金を請求されることが多いし、大学等のレギュレーションによって、正式に保存される論文はモノクロコピーであることが珍しくないからである。

したがって、印刷物（学会の抄録なども含めて）では、グラフの線は色分けではなく、実線、破線、一点鎖線のような線種で見分けがつくようにすることが望ましい。線だけでなく、データ点に打つシンボルの種類を変えることも有効な場合がある。領域も色による塗り分けではなく、角度や密度を変えたハッチングや粗さを変えた点描によって塗り分けるべきである。R のコードでは、lty=というオプションで線種、lwd=で線の太さを指定できるようになっている関数が多い。図の中に凡例を明示することも必要である。

一方、学会発表や修論報告会などのプレゼンテーションにおいては、色分けが有効である。多くの関数で、col=というオプションにベクトル（整数または文字列。文字列は"black"のような

[14] https://cran.r-project.org/web/packages/RBtest/index.html

色名でも有効だし、"#ffffff"のように RGB に 16 進数 2 桁ずつを割り当てた文字列でも良い。"#ffffff77"のように最後の 2 桁で透過性を示すアルファチャンネルを指定することもできる）を与えることで色指定ができる。ただし、整数と色の対応関係のデフォルトの、1: 黒、2: 赤、3: 緑、4: 青、……という配色は、色覚多様性を考慮すると、必ずしも見分けやすくない。カラーユニバーサルデザインで推奨される配色セットを有効にするためには、palette("Okabe-Ito") を一度実行すれば良い。配色パレットが変わるので、それ以後ほとんどのグラフで使われる色が変わってくれる。グラフによっては領域塗りつぶしを重ねる必要がある場合があるが、その際はハッチングかアルファチャンネル指定を使うと良い。起動時に自動的にロードされる grDevices 関数に含まれている adjustcolor() 関数を使えば、色名文字列と透過性を簡単に組み合わせることができる。

```
palette("Okabe-Ito") # カラーユニバーサルデザインの配色に
palette(adjustcolor(palette(), alpha.f=0.3)) # 配色そのままで透過度 70 %に
palette("R3") # R-3.*までのデフォルトパレットに
palette("default") # R-4.*のデフォルトパレットに
```

6.4.2　survey データフレームの読み込み

10 人の身長・体重のデータでは作図の例示には向かないため、MASS パッケージに含まれている survey というデータフレームを使って説明する。MASS パッケージは推奨パッケージとして Windows 版のインストーラには元々含まれているし、Rcmdr や EZR も依存しているので、別途インストールする必要はない。R コンソールでは library(MASS) とするだけでが MASS パッケージがメモリにロードされ、survey データフレームが使える状態になる。

R コマンダーでは、メニューの [ツール] の [パッケージのロード] を選んで表示されるウィンドウの中で、MASS を選ぶ。次に [データ] の [パッケージ内のデータ] の [アタッチされたパッケージからデータセットを読み込む] を選び、表示されるウィンドウの左の枠で MASS をダブルクリックし、次に右の枠で survey をダブルクリックし、[OK] ボタンをクリックする。
EZR でも、まずメニューの [ツール] の [パッケージのロード] を選んで表示されるウィンドウの中で、MASS を選ぶ[a]。「ファイル」メニューの「パッケージに含まれるデータを読み込む」から、パッケージとして MASS をダブルクリックし、データセットとして survey をダブルクリックしてから「OK」ボタンをクリックするだけでよい。

[a] 古いバージョンでは起動時から MASS がロードされていたが、2015 年 5 月現在のバージョンではこの操作が必要になっている。

survey というデータは、アデレード大学の学生 237 人の調査結果であり、含まれている変数は以下の通りである。

survey の変数

Sex 性別、Male（男性）または Female（女性）（要因型）

Wr.Hnd 字を書く利き手の親指と小指の間隔、cm 単位（数値型）

NW.Hnd 利き手でない方の親指と小指の間隔、cm 単位（数値型）

Wr.Hnd 利き手、Left（左利き）または Right(右利き）（要因型）

Fold 腕を組んだときにどちらが上になるか ?: R on L（右が上）、L on R（左が上）、Neither（どちらでもない）、の 3 水準（要因型）

Pulse 心拍数／分（整数型）

Clap 両手を叩き合わせた時、どちらが上にくるか ?: Right（右）、Left（左）、Neither（どちらでもない）、の 3 水準（要因型）

Exer 運動頻度：Freq（頻繁に）、Some（時々）、None（しない）、の 3 水準（要因型）

Smoke 喫煙習慣：Heavy（ヘビースモーカー）、Regul（定期的に吸う）、Occas（時々吸う）、Never（決して吸わない）、の 4 水準（要因型）

Height 身長：cm 単位（数値型）

M.I 身長の回答が Imperial（インペリアル＝フィート／インチ）でなされたか、Metric（メトリック＝ m ／ cm）でなされたか（要因型）

Age 年齢：年単位（数値型）

6.4.3 離散変数（カテゴリデータ）からの作図

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

上記 survey データで Smoke のカテゴリごとの度数分布図を描くには、

```
barplot(table(survey$Smoke))
```

でよい（下図左）。

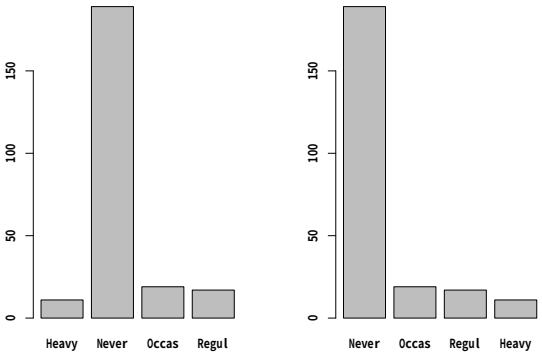
ただしこれだと（Rcmdr/EZR の場合も同様だが）、カテゴリ名のアルファベット順に棒が並んでしまうので、そうしたくない場合は若干工夫が必要で、

```
barplot(table(survey$Smoke)[c("Never", "Occas", "Regul", "Heavy")])
```

または、

```
survey$Smoke <- factor(survey$Smoke,
 levels=c("Never", "Occas", "Regul", "Heavy"))
barplot(table(survey$Smoke))
```

のようにすれば、棒を並べる順序を指定することができる（下図右）。



Rcmdr では上記 survey データフレームをアクティブにした状態で、「グラフ」の「棒グラフ」を選び（EZR では「グラフと表」「棒グラフ（頻度）」を選ぶ）、「変数（1 つ選択）」の中から Smoke を選んで「OK」をクリックすると、喫煙習慣ごとの人数がプロットされる。
EZR でカテゴリ名のアルファベット順でなく、指定した順番でカテゴリを並べたいときには、予め因子水準の順番を決めておく。この場合なら、「アクティブデータセット」「変数の操作」「因子水準を再順序化」を選び、「因子（1 つ選択）」から Smoke を選び、因子の名前は＜元と同じ＞となっているので変えず、「順序のある因子の作成」の右側のボックスにチェックを入れてから「OK」ボタンをクリックする。「変数 Smoke が既に存在します。上書きしますか？」というダイアログが表示されるので OK し（もし元のカテゴリ順序の情報がなくなるのが嫌なら、因子の名前のところに既存の変数名と重複しない新しい名前を入力しておけば、このダイアログは出ない。その場合は新しい変数が作られる）、「新しい順序」に表示したい順番を入力して「OK」ボタンをクリックする。この操作をしてから、「グラフと表」「棒グラフ（頻度）」でグラフを作ればよい。

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。単純に survey データフレームの喫煙習慣について積み上げ棒グラフを描くには、R コンソールでは

```
barplot(as.matrix(table(survey$Smoke)))
```

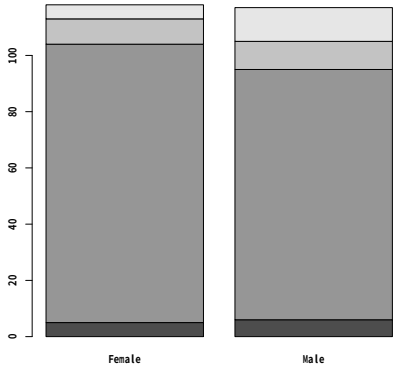
で描画される。しかし、積み上げにするのは、複数カテゴリ間で比べるためなので、例えば男女間で喫煙習慣を比較するために、

```
barplot(table(survey$Smoke, survey$Sex))
```

あるいは同じことだが、

```
barplot(xtabs(~Smoke+Sex, data=survey))
```

のようにするのが現実的な利用法である。**Rcmdr ではサポートされていない。**



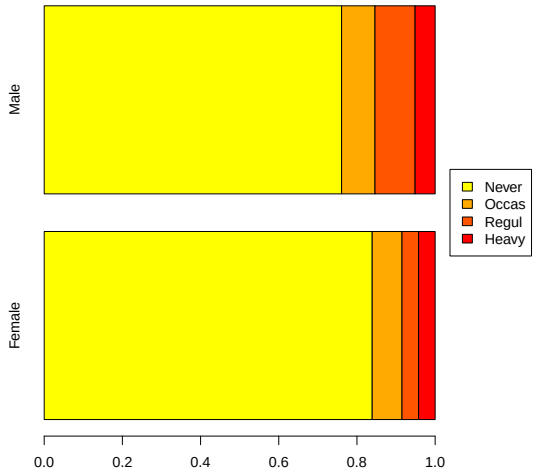
EZR では、上記 survey データフレームをアクティブにした状態で、「グラフと表」「棒グラフ（頻度）」を選び、「変数 (1 つ選択)」のところで Smoke を選ぶまでは度数分布を描くときと同じだが、「群別化変数 1（0〜1 つ選択）」の中からも Sex を選んで「OK」ボタンをクリックすることにより、層別積み上げ棒グラフを描くことができる。

帯グラフ 横棒を全体を 100%として各値の割合にしたがって区切って塗り分けた図である。R では積み上げ棒グラフをグループごとに割合にして 90 度回転して表示すればよいので、

```
barplot(prop.table(table(survey$Smoke, survey$Sex), 2), horiz=TRUE)
```

とする。もう少し見やすく描き、キャプションを付けるには、例えば以下のコードを R コンソールに打つと、下図が描かれる。

```
par(oma=c(0,0,0,6), xpd=NA)
Extent <- c("Never","Occas","Regul","Heavy")
FC <- rev(heat.colors(5)[1:4])
barplot(prop.table(table(survey$Smoke, survey$Sex)[Extent, ], 2),
 horiz=TRUE, col=FC)
par(oma=c(0,0,0,0))
legend("right", fill=FC, legend=Extent)
```


Male
Female
Never
Occas
Regul
Heavy
0.0
0.2
0.4
0.6
0.8
1.0

EZR では、90 度回転して表示するオプションがサポートされていないが、上記 survey データフレームをアクティブにした状態で、「グラフと表」「棒グラフ（頻度）」を選び、「変数 (1 つ選択)」のところで Smoke を選び、「群別化変数 1（0〜1 つ選択)」の中からも Sex を選ぶところまで層別積み上げ棒グラフと同じで、さらに「群間の比較の場合に各群の中の割合で描画する」の左側のチェックボックスにチェックを入れてから「OK」ボタンをクリックすると、全体が 100%になる。
90 度回転したければ、描画されたグラフをクリップボードにコピーしてからパワーポイントなどに貼り付け、描画オブジェクトとして回転させるという方法もある。なお、その際、日本語を含むグラフをメタファイルとして貼り付けて描画オブジェクトに変換すると Windows 環境では文字化けする可能性があり、それを防ぐためには、描画前に R スクリプトウィンドウに

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

の 2 行を打って（windowFont() の中は JP2, JP3 などコンマで区切って複数のフォントを指定可能)、選択してから「実行」ボタンをクリックし、出現するグラフィックウィンドウをそのままにして描画すれば、そこに描かれたグラフ要素はフォントが MS ゴシックになるので、クリップボード経由でメタファイルとしてコピーペーストしても日本語が文字化けしなくなる。

円グラフ（ドーナツグラフ・パイチャート）円全体を 100%として、各値の割合にしたがって中心から区切り線を引き、塗り分けた図である。ドーナツグラフでは 2 つの同心円にして、内側の円内を空白にする。R では pie() 関数を用いる。ただし錯覚しやすいのでお勧めしない。度数分布図の方が良い。

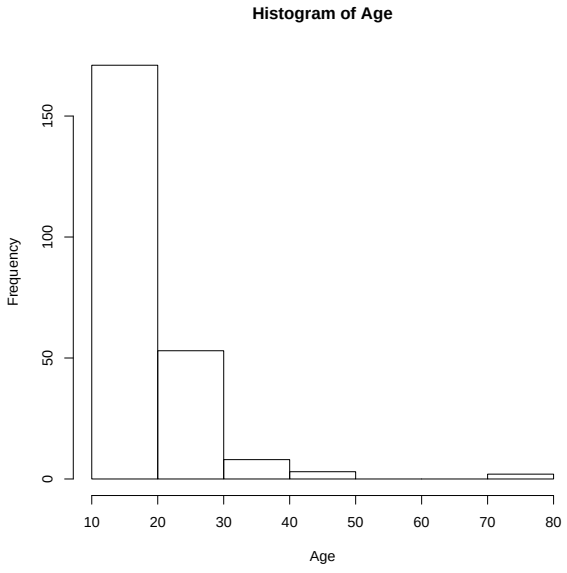
Rcmdr では「グラフ」の「円グラフ」（EZR では「グラフと表」の「円グラフ（頻度)」）を選ぶ。survey データフレームをアクティブにした状態で、変数として Smoke を選ぶと、喫煙習慣ごとの人数の割合に応じて円が分割された扇形に塗り分けられたグラフができる。

6.4.4 連続量データからの作図

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

```
hist(survey$Age, breaks=1:8*10, right=FALSE, xlab="Age",
 main="Histogram of Age")
```

とする。



Rcmdr では「グラフ」の「ヒストグラム」、EZR では「グラフと表」の「ヒストグラム」を選ぶ。survey データでは、変数として Age を選べば、年齢のヒストグラムが描ける（アデレード大学の学生のデータのはずだが、70 歳以上の人や 16.75 歳など、大学生らしくない年齢の人も含まれている）。Rcmdr や EZR では「〜以上〜未満」にはできない（裏技的には、予め「〜以上〜未満」のカテゴリデータに変換しておき、「棒グラフ（頻度）」で描画することはできるが、バーの間に隙間があるのはヒストグラムとしては正当でないのでお勧めしない）。

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ためのグラフである。データを小さい順に並べて、縦軸に生データをとり、横軸に対応する標準正規分布のパーセント点をとって描いた散布図であり、縦軸に正規分布に当てはまっていれば点が直線上に並ぶ。R コンソールでは qqnorm() 関数を用いる（getS3method("qqnorm","default") とすると実際のコードの中身がわかる。中で使われている ppoints() 関数の定義を見るには、ppoints と打てばよい）。

例えば、survey データフレームの心拍数 (Pulse) について正規確率プロットを描くには、qqnorm(survey$Pulse) とする。以下のように打つと同じグラフになることを確認されたい。

```
rPulse <- sort(survey$Pulse)
n <- length(rPulse)
tNorm <- qnorm((1:n-ifelse(n<=10, 3/8, 1/2))/n)
plot(tNorm, rPulse, main="Normal QQ plot")
```

Rcmdr では「グラフ」の「QQ プロット」を選び、EZR では「グラフと表」の「QQ プロット」を選ぶ。survey データフレームでは、変数として Age を選ぶと、まったく正規分布でないので直線状でないし、Pulse を選ぶと、やや歪んでいるけれども概ね直線に乗るので正規分布に近いことがわかる。また、EZR では、「統計解析」「連続変数の解析」「正規性の検定 (Kolmogorov-Smirnov 検定)」でも、正規性の検定が実施されるのと同時にヒストグラムに正規分布の曲線を重ね描きしてくれるので、正規分布と見なせるかどうかは見当がつく。

幹葉表示 (stem and leaf plot) 大体の概数（整数区切りとか 5 の倍数とか 10 の倍数にすることが多い）を縦に並べて幹とし、それぞれの概数に相当する値の細かい部分を葉として横に並べ

て作成する図。R では stem() 関数を用いる。同じデータで心拍数の幹葉表示をするには、stem(survey$Pulse) とする。

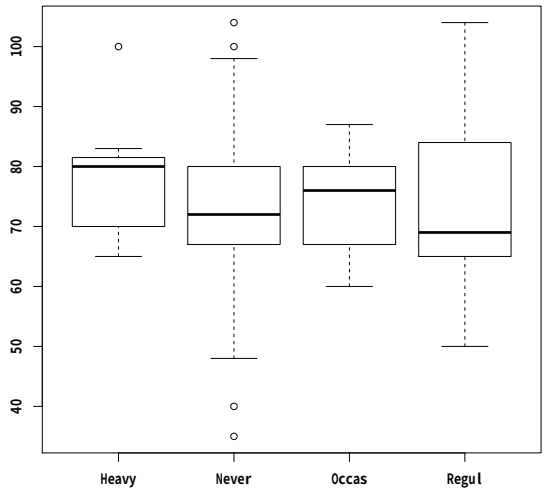
```
The decimal point is 1 digit(s) to the right of the |

 3 | 5
 4 | 0
 4 | 88
 5 | 004
 5 | 569
 6 | 00000000000012222344444444
 6 | 55555566666678888888888888889
 7 | 000000000000011222222222222344444
 7 | 55555666666666666668888999
 8 | 00000000000000000001333344444
 8 | 55556667788889
 9 | 00000000222222
 9 | 66678
10 | 0044
```

Rcmdr では「グラフ」の「幹葉表示」を選ぶ。fmsb パッケージの gstem() 関数を使えばグラフィック画面に出力することもできる。Rcmdr には含まれていない。EZR では「グラフと表」「幹葉表示」を選び、「変数（１つ選択）」の枠内から Pulse を選んで「OK」ボタンをクリックすれば Output ウィンドウにテキスト出力される。さまざまなオプションが指定可能である。

箱ヒゲ図 (box and whisker plot) 縦軸に変数値をとって、第１四分位を下に、第３四分位を上にした箱を書き、中央値の位置にも線を引いて、さらに第１四分位と第３四分位の差（四分位範囲）を 1.5 倍した線分をヒゲとして第１四分位の下と第３四分位の上に伸ばし、ヒゲの先より外れた値を外れ値として○をプロットした図である。現在の日本の高校では五数要約値のグラフ化として教えられているため、ヒゲが箱の上端から最大値までと、箱の下端から最小値までとして教えられているが、Tukey の元々のアイディアとしては既述の通り、四分位範囲の 1.5 倍をヒゲとしている。単独で箱ヒゲ図を描くよりも、カテゴリ変数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 コンソールでは boxplot() 関数を用いる。例えば、survey データで喫煙状況 (Smoke) 別に心拍数 (Pulse) の箱ヒゲ図を描くには次の 2 行のどちらかを打つ。

```
boxplot(survey$Pulse ~ survey$Smoke)
boxplot(Pulse ~ Smoke, data=survey)
```



Rcmdr では「グラフ」の「箱ひげ図」、EZR では「グラフと表」の「箱ひげ図」を選ぶ。survey データで喫煙状況別に心拍数の箱ひげ図を描かせるには、Rcmdr の場合は変数として Pulse を選び、[層別のプロット] というボタンをクリックして表示されるウィンドウで、層別変数として Smoke を選んで [OK] ボタンをクリックしてから、戻ったウィンドウで再び [OK] をクリックすればいい（EZR の場合は「群別する変数（0〜1 つ選択）で Smoke を選び、上下のひげの位置として「第 1 四分位数-1.5x 四分位範囲、第 3 四分位数+1.5x 四分位範囲」の左のラジオボタンをチェックして「OK」をクリックするだけでよい）。

似た用途のグラフとして、層別の平均とエラーバーを表示して折れ線で結ぶことも Rcmdr の場合「グラフ」の「平均のプロット」でできる（EZR では「グラフと表」の「棒グラフ（平均値）」または「折れ線グラフ（平均値）」）。survey データで喫煙習慣ごとに心拍数の平均値とエラーバーを表示して折れ線で結びたいなら、「因子」として Smoke、「目的変数」として Pulse を選べばよい。エラーバーとしては標準誤差（デフォルト）、標準偏差、信頼区間から選択できる。

レーダーチャート 複数の連続変数を中心点から放射状に数直線としてとり、データ点をつないで

表される図である。それら複数の変数によって特徴付けられる性質のバランスをみるのに役立つ。別の言い方をすれば、個人ごとのプロファイルを見るのに適している。解析でというよりは、調査で得られたデータを対象者に返すときに役立つ（とくに質問紙調査から複数のスコアを計算して返すような場合）。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 でレーダーチャートを描くには、plotrix パッケージか fmsb パッケージをインストールする必要がある。どちらも CRAN のミラーサイトからダウンロードしてインストールできる。インターネットにつながったいる環境ならば、前者は

install.packages("plotrix")、

後者は

install.packages("fmsb")

とすればインストールできる。その上で、例えば後者の場合なら、library(fmsb) としてから example(radarchart) とすれば使い方がわかる[15]。**Rcmdr や EZR のメニューには入っていない。**

散布図 (scatter plot) 2つの連続変数の関係を2次元の平面上の点として示す、基本的な図である。関連があるかもしれない量的な変数が複数あるときは、まずは散布図を作ってみるのが原則である。R では plot() 関数を用いる（ただし、実は R の plot() 関数は総称的関数なので、縦軸が数値変数で、横軸がカテゴリ変数であるときは自動的に層別箱ひげ図になる）。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。

例えば、MASS パッケージの survey データフレームに含まれる Height を横軸に、Wr.Hnd を縦軸にして散布図を描きたい場合は以下 2 行のどちらかを打つ。

```
plot(Wr.Hnd ~ Height, data=survey)
plot(survey$Height, survey$Wr.Hnd)
```

男女別に違うマークでプロットしたい場合は、

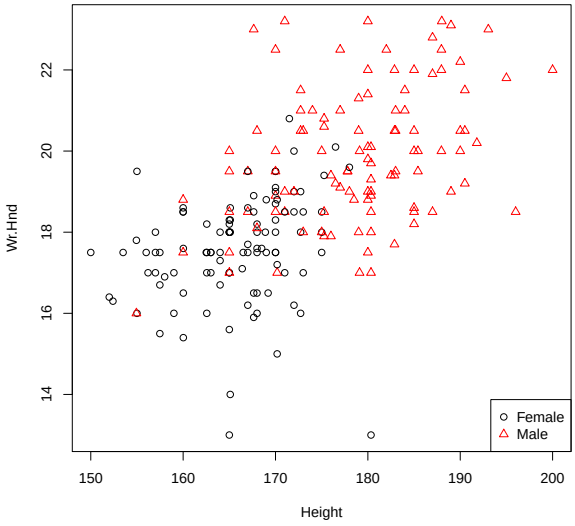
```
plot(Wr.Hnd ~ Height, data=survey, pch=as.integer(Sex),
 col=as.integer(Sex))
```

とする。この図に凡例を追加したければ、

[15] https://minato.sip21c.org/demography/how-to-make-pref-charts.html も利用例として参照されたい。

```
legend("bottomright", pch=1:2, col=1:2, legend=c("Female", "Male"))
```

とすればよい。points() 関数を使って重ね打ちすることもできる。点ごとに異なる情報を示したい場合（異なる大きさの円でプロットするなど）は symbols() 関数を用いることができる。



Rcmdr では「グラフ」「散布図」、EZR では「グラフと表」「散布図」を選び、「x 変数（１つ選択）」から Height を、「y 変数（１つ選択）」から Wr.Hnd を選んで、「周辺箱ひげ図」と「最小 2 乗直線」（Rcmdr では「平滑線」）の左のボックスのチェックを外してから「OK」をクリックすれば、上と同じような図が描ける。オプション指定で層別にマークを変えることもできる。[層別のプロット] ボタンをクリックして層別変数として Sex を選べばよい。できあがった散布図の点をマウスでクリックして値を確認したい時は、x 変数や y 変数を指定するウィンドウで、「点を確認する」にチェックを入れておく。確認したい点の上で左クリックするとレコード番号が表示され、右クリックするまで繰り返すことができる。

複数の連続変数間の関係を調べたい場合は、matplot() 関数と matpoints() 関数を使って重ね描きすることもできるが、別々のグラフとして並べて同時に示す「散布図行列」を描画するのが便利である。pairs() 関数を用いる。例えば、MASS パッケージの survey データフレームに含まれるすべての数値型変数（Wr.Hnd、NW.Hnd、Pulse、Height、Age）について

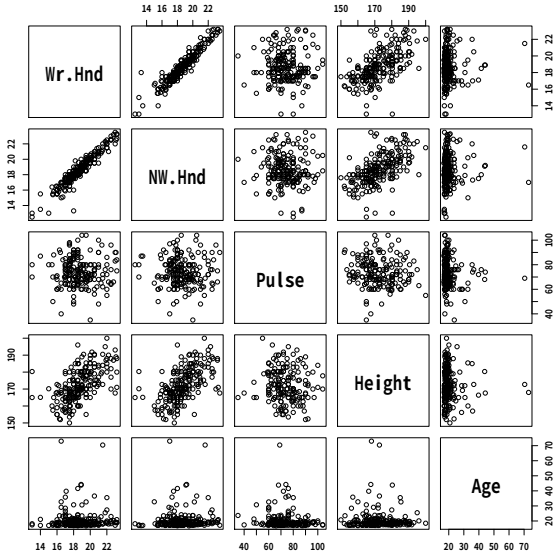
散布図行列を作りたいときは、

```
pairs(subset(survey, select=sapply(survey, is.numeric)))
```

とすれば良い。下の図ができる。すべての変数でなく、例えば Wr.Hnd、Pulse、Age の 3 つの関係だけを調べたいなら、

```
pairs(~ Wr.Hnd+Pulse+Age, data=survey)
```

とすればよい。



EZR では「グラフと表」「散布図行列」メニューで pairs() の機能が実装されている。オプション指定により、対角線上に個々の変数についてのさまざまなグラフを表示できるのが便利である。

6.5 記述統計・分布の正規性・外れ値

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

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

6.5.1　中心傾向 (central tendency)

中心傾向の代表的なものは以下の３つである。

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

母集団の平均値 μ（ミューと発音する）は、

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

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

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

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

である。n は、もちろん標本サイズである[16]。

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

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

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

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

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

平均値は、(1) 分布のすべての値を考慮した値である、(2) 同じ母集団からサンプリングを繰り返した場合に一定の値となる、(3) 多くの統計量や検定で使われている、という特長をもつ。標本調査値から母集団の因果関係を推論したい場合に、もっとも普通に使われる。しかし、(1) 極端な外れ値の影響を受けやすい、(2) 打ち切りのある分布では代表性を失う場合がある[17]、という欠点があり、外れ値があったり打ち切りがあったりする分布では位置の指標として中央値の方が優れている。最頻値は、標本をとったときの偶然性の影響を受けやすいし、もっとも頻度が高い値以外の情報はまったく使われない。しかし、試験の点で何点の人が多かったかを見たい場合は最頻値が役に立つし、名義尺度については最頻値しか使えない。

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

6.5.2 ばらつき (Variability)

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

四分位範囲 (Inter-Quartile Range; IQR) 四分位範囲について説明する前に、分位数について説明する。値を小さい方から順番に並べ替えて、４つの等しい数の群に分けたときの 1/4, 2/4, 3/4 にあたる値を、四分位数 (quartile) という。1/4 の点が第１四分位、3/4 の点が第３四分位である（つまり全体の 25%の値が第１四分位より小さく、全体の 75%の値が第３四分位より小さい）。2/4 の点というのは、ちょうど順番が真中ということだから、第２四分位は中央値に等しい。ちょっと考えればわかるように、ちょうど４等分などできない場合がもちろんあって、上から数えた場合と下から数えた場合で四分位数がずれる可能性があるが、その場合はそれらを平均するのが普通である。また、最小値、最大値に、第１四分位、第３四分位と中央値を加えた５つの値を五数要約値と呼ぶことがある（R では `fivenum()` 関数で五数要約値を求めることができる）。第１四分位、第２四分位、第３四分位は、それぞれ Q1, Q2, Q3 と略記

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

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

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

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

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

で定義される[18]。標本分散 $V(X)$ は、標本平均 $\bar{X}$ を使って、

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

という式で得られる。ただし、標本平均がそのまま母平均の推定値となるのに対して、標本分散は母分散の推定値にならない。母分散の推定値としては、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 では `var()` で得られる）。

標準偏差 (standard deviation) 分散の平方根をとったものが標準偏差である。**平均値と次元を揃える意味**をもつ。不偏分散の平方根をとったものは、不偏標準偏差と呼ばれる（R では `sd()` で得られる）[19]。もし分布が正規分布ならば、Mean±2SD[20]の範囲にデータの 95%が含まれるという意味で、標準偏差は便利な指標である。なお、名前は似ているが、「標準誤差」はデータのばらつきでなくて、推定値のばらつきを示す値なので混同しないように注意されたい。例えば、平均値の標準誤差は、サンプルの不偏標準偏差をサンプルサイズの平方根で割れば得

[18] 実際に計算するときは 2 乗の平均から平均の 2 乗を引くとよい。

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

[20] 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。

られるが、意味は、「もし標本抽出を何度も繰り返して行ったとしたら、得られる標本平均のばらつきは、一定の確率で標準誤差の範囲におさまる」ということである。ちなみに、データそのものがどのような分布であっても、標本抽出回数を増やしていくと、標本平均の分布は正規分布に近づくことが中心極限定理によって証明されている。例えば、区間 (0, 1) の一様分布から、サイズ 10 の無作為標本を 1000 回抽出すると、その平均値の分布がほぼ正規分布していることは、以下のコードで確かめることができる。

```
x <- runif(10000)
y <- matrix(numeric(1000*10), 10)
for (i in 1:1000) { y[, i] <- sample(x, 10, replace=FALSE) }
hist(colMeans(y))
```

上記の記述統計量を計算するには、Rcmdr からは、メニューバーの「統計量」の「要約」から「数値による要約」を選べばよいし、EZR では、「統計解析」の「連続変数の解析」の「連続変数の要約」を選べばよい。選びたい変数の上にマウスカーソルを移動し、CTRL を押しながら左クリックすることで、複数の変数を選択することができる。オプションをとくに指定しなければ、平均値、標準偏差、最小値、第１四分位（25 パーセンタイル）、中央値（50 パーセンタイル）、第３四分位（75 パーセンタイル）、最大値、有効標本サイズ、欠損値の数が、選択した変数すべてについて表示される。

6.5.3 分布の正規性と外れ値の検定

分布の正規性の検定は、コルモゴロフ＝スミルノフ検定あるいはシャピロ＝ウィルク検定が有名である。fmsb パッケージにはギアリーの検定も実装してある。もちろん検定をする前に、ヒストグラムか正規確率プロットによって分布の様子を確認しておくことは必須である。例えば、survey データフレームに含まれている NW.Hnd という変数（利き手でない方の手を開いたときの親指と小指の先の距離）について分布の正規性の検定を行うには以下のように打つ。

```
mu <- mean(survey$NW.Hnd[!is.na(survey$NW.Hnd)])
psd <- sd(survey$NW.Hnd[!is.na(survey$NW.Hnd)])
ks.test(survey$NW.Hnd, "pnorm", mean=mu, sd=psd)
shapiro.test(survey$NW.Hnd)
```

EZR では、「統計解析」の「連続変数の解析」の「正規性の検定」を選び、変数として NW.Hnd を選んで OK ボタンをクリックするだけでできる。自動的にコルモゴロフ＝スミルノフ検定の（サンプルサイズが 5000 以下のときはシャピロ＝ウィルク検定も）結果が表示され、ヒストグラムに正規分布を重ね描きしたグラフも作成される。コルモゴロフ＝スミルノフ検定とシャピロ＝ウィルク検定の結果が異なることがあるが、これらの検定は分布の正規性へのアプローチが異なるので、結果が一致しないこともある。また、多くの検定手法がデータの分布の正規性を仮定しているが、これらの検定で正規性が棄却されたからといって機械的に変数変換やノンパラメトリック検定でなくてはいけないとは限らない。独立 2 標本の平均値の差がないという仮説を検定するための t 検定（等分散性を仮定しない Welch の方法も含めて）は頑健な手法なので、正規分布に従っているといえなくても、そのまま実行してもいい場合も多い。ただし、明らかな外れ値がある場合は検出力が落ちるので、Wilcoxon の順位和検定のようなノンパラメトリックな検定の方が良い場合もある。

外れ値についても、グラフで確認することは必須である。統計学的に有意に正規分布から外れているかどうかの検定にも何種類かあるが、スミルノフ＝グラブス検定は、outliers パッケージに含まれている grubbs.test() 関数や群馬大学の青木繁伸教授が web サイトで提供している SG() 関数[21]によって実行可能である。

EZR では、「統計解析」の「連続変数の解析」の「外れ値の検定」を選んで、変数として NW.Hnd を指定して OK ボタンをクリックするだけでいい。外れ値を NA で置き換えた新しい変数を作成することも右のオプション指定から容易にできる。しかし、既に書いた通り、この結果だけで機械的に外れ値を除外することはお薦めできない。有意な外れ値がない場合は、"No outliers were identified." と表示される。

6.5.4　研究対象の基本属性情報のまとめを作る

論文を書く場合、結果の最初に Table 1 としてサンプルの基本属性の集計結果をまとめて表示することが多い。EZR は、バージョンアップによって、この機能を内蔵するようになった。メニューの「グラフと表」の「サンプルの背景データのサマリー表の出力」を選び、群別する変数（0〜1 つ選択）、カテゴリー変数（名義変数、順序変数）、連続変数（正規分布）、連測変数（非正規分布）[22]を選び、正規分布しない連続変数の範囲表示をデフォルトの「最小値と最大値」にするのか「四分位範囲 (Q1-Q3)」にするのかを選んでから「OK」をクリックする。

[21] http://aoki2.si.gunma-u.ac.jp/R/SG.html
[22] 連続変数（非正規分布）の typo か？

cran には table1 というパッケージが存在し、名前の通り、Table 1 を見やすく作るための機能を提供している。詳細は https://cran.r-project.org/web/packages/table1/vignettes/table1-examples.html に説明されている。同じ目的で tableone というパッケージも存在し、https://cran.r-project.org/web/packages/tableone/vignettes/introduction.html に使い方が説明されている。他にもたくさんのパッケージが似た目的で開発されており、比較記事[23]が書かれていて参考になる。

[23] https://thatdatatho.com/2018/08/20/easily-create-descriptive-summary-statistic-tables-r-studio/

第7章　２群間の差の検定

医学統計でよく使われるのは、伝統的に仮説検定である。仮説検定は、意味合いからすれば、元のデータに含まれる情報量を、仮説が棄却されるかどうかという２値情報にまで集約してしまうことになる。これは情報量を減らしすぎであって、点推定量と信頼区間を示す方がずっと合理的なのだが、伝統的な好みの問題なので、この演習でも検定を中心に説明する。もっとも、Rothman とか Greenland といった最先端の疫学者は、仮説検定よりも区間推定、区間推定よりも p 値関数の図示（リスク比やオッズ比については、`fmsb` パッケージに `pvalueplot()` 関数として実装済み）の方が遥かによい統計解析であると断言している。

最も単純な検定の１つは、独立２標本間の分布の位置の差の検定であり、非常によく使われているので、まずはここから検定の考え方を見てみよう。

7.1　独立２標本間の平均値の差の検定

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

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

独立２群間の統計的仮説検定の方法は、以下のようにまとめられる。

1. 量的変数の場合

   (a) 正規分布に近い場合（`shapiro.test()` で Shapiro-Wilk の検定ができるが、その結果を機械的に適用して判断すべきではない）：Welch の検定（R では `t.test(x,y)`）[1]

   (b) 正規分布とかけ離れている場合：Wilcoxon の順位和検定[2]が標準的に使われてきたが（Mann-Whitney の U 検定という言い方もあるが、数学的にまったく同じ検定である）、これらは分布の形は仮定しないけれども、2 群が連続分布であることと、分布の形に差がない「ズレのモデル」を前提としているため、それさえも仮定しない、Brunner-Munzel 検定が最近提案され、広まりつつある。

2. カテゴリ変数の場合：母比率の差の検定（R では `prop.test()`）。次章で扱う。

7.2　等分散性についての F 検定

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（サンプルサイズが各々 n_X と n_Y とする）の比較を考える。

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

この結果、2 群間で分散が統計的に有意に異なっていたら、その情報自体が、異なる母集団からのサンプルであるとか、サンプリングが偏っていた可能性を示唆する。かつては、等分散性の検定結果によって平均値の差の t 検定において等分散性を仮定したり Welch の方法にしたりといったことをしていたが、現在では、等分散と考えられるかどうかによらず、Welch の方法で検定すれば良い。

---

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

[2] R では wilcox.test(x,y) で実行できる。

Rcmdr では「統計量」の「分散」から「分散の比の F 検定」を選び、グループ (Group variable) として C を、目的変数 (Response variable) として X を選ぶ。ただし、グループ変数は要因型になっていないと候補として表示されないので、もし 0/1 で入力されていたら、予め「データ」の「アクティブデータセット内の変数の操作」で「数値変数を因子に変換」を用いて要因型にしておく（字面は 0/1 のままでも OK）。survey データで、「男女間で身長の分散に差がない」という帰無仮説を検定するには、グループとして Sex を、目的変数として Height を選んで [OK] ボタンをクリックする。デフォルトでは両側検定されるが、仮説によっては片側検定をすることもあり、その場合は「対立仮説」の下のラジオボタンのチェックを変えればいい。男女それぞれの分散と、検定結果が「出力ウィンドウ」に表示される。
EZR では、「統計解析」の「連続変数の解析」から「2 群の等分散性の検定（F 検定）」を選び、目的変数（１つ選択）の枠から X を、グループ（1 つ選択）の枠から C を選び（EZR では要因型にしなくても選べる）, OK ボタンをクリックする。survey データで、「男女間で身長の分散に差がない」という帰無仮説を検定するには、目的変数として Height を、グループとして Sex を選び、OK ボタンをクリックすると、男女それぞれの分散と検定結果が Output ウィンドウに表示される。

7.3 Welch の方法による t 検定

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

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

R では、`t.test(X,Y,var.equal=F)` だが、`var.equal` の指定を省略した時は等分散でないと仮定して Welch の検定がなされるので省略して `t.test(X,Y)` でいい。量的変数 `X` と群分け変数 `C` という入力の仕方の場合は、`t.test(X~C)` とする[3]。survey データで「男女間で平均身長に差がない」という帰無仮説を検定したいときは、`t.test(Height ~ Sex, data=survey)` とする。

なお、t 検定の効果量については、第 5 章「研究のデザイン」で既に述べたので、そちらを参照されたい。

[3] この書き方からわかるように、平均値の差の t 検定は、見方を変えれば、群分け変数を独立変数、比較される量の変数を従属変数として、量のばらつきが誤差に比べてどれくらい群間の違いで説明されるかをみる分散分析であり、線型モデルの当てはめといえる。

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

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

ちなみに、survey データで年齢（Age）に性差があるかどうかを検定したい場合に Welch の t 検定をすると、年齢は明らかに正規分布から大きく外れており、かつ外れ値が多いので検出力が落ちる。実際にやってみるとわかるが有意でない。

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

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

または

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

のように変換しておく必要がある[4]。プロットするには次のように入力すればよい[5]。

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

7.4 対応のある 2 標本の平均値の差の検定

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

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

survey データで「親指と小指の間隔が利き手とそうでない手の間で差がない」という帰無仮説を検定するには、R コンソールでは、

```
t.test(survey$Wr.Hnd, survey$NW.Hnd, paired=TRUE)
```

と打てばよい。グラフは通常、同じ人のデータは線で結ぶので、例えば次のように打てば、差が 1 cm 以内の人は黒、利き手が 1 cm 以上非利き手より大きい人は赤、利き手が 1 cm 以上非利き手より小さい人は緑で、人数分の線分が描かれる。

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

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

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

Rcmdr では「統計量」の「平均」の「対応のある *t* 検定」を選ぶ（EZR では「統計解析」「連続変数の解析」から「対応のある 2 群間の平均値の検定（paired *t* 検定）」を選ぶ）。第１の変数として `Wr.Hnd` を、第２の変数として `NW.Hnd` を選び、[OK] ボタンをクリックすると、出力ウィンドウ（EZR では Output ウィンドウ）に結果が得られる。有意水準 5%で帰無仮説は棄却され、利き手の方がそうでない手よりも親指と小指の間隔が有意に広いといえる。
なお、対応のある t 検定の意味を考えれば、EZR でも「アクティブデータセット＞変数の操作＞計算式を入力して新たな変数を作成する」として、新しい変数名として `Diff.Hnd` として、計算式として `Wr.Hnd-NW.Hnd` として OK をクリックすることで手の大きさの差の変数を作り、この `Diff.Hnd` の母平均がゼロかどうかを「統計解析＞連続変数の解析＞ 1 標本の平均値の t 検定」で検定することもできる。その場合は結果として差の期待値と信頼区間が得られる。

例題

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

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

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

本当は因果を逆に考えてロジスティック回帰またはポアソン回帰する方が筋がいいと思うが、ここでは敢えて平均値の差の検定をしてみる。2 群間で分布が異なるし対照群では正規分布から明らかに外れているが、それにも目をつぶって平均値の差の検定を行う。2 回以上というのを 2 回と

扱っていいのかという点にも問題があるが、ここでは目をつぶる。R コンソールでは、この操作は単純である。必要な *t* 検定をするには、次のように打てば良い。

(1) `t.test(spontaneous ~ case, data=infert)` とする。もし患者群と対照群の間で分散が等しいという帰無仮説を検定したいなら、

```
var.test(spontaneous ~ case, data=infert)
```

と打つ。

(2) `t.test(infert$induced, infert$spontaneous, paired=TRUE)`

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

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

(2) 「統計量」の「平均」の「対応のある *t* **検定」を指定し（EZR なら「統計解析」「連続変数の解析」から「対応のある 2 群間の平均値の検定（paired** *t* **検定）」を選ぶ）、第１の変数として** spontaneous**、第２の変数として** induced **を選んで [OK] ボタンをクリックすれば実行できる。**

7.5　Wilcoxon の順位和検定

Wilcoxon の順位和検定は、パラメトリックな検定でいえば、*t* 検定を使うような状況、つまり、独立２標本の分布の位置に差がないかどうかを調べるために用いられる。Mann-Whitney の U 検定と（これら２つほど有名ではないが、Kendall の S 検定とも）数学的に等価である。Rcmdr では、「統計量」の「ノンパラメトリック検定」を選んで実行する。

データがもつ情報の中で、単調変換に対して頑健なのは順位なので、これを使って検定しようという発想である。以下、Wilcoxon の順位和検定の手順を箇条書きする。

1. 変数 X のデータを $x_1, x_2, ..., x_m$ とし、変数 Y のデータを $y_1, y_2, ..., y_n$ とする。

2. まず、これらをまぜこぜにして小さい方から順に番号をつける[6]。例えば、$x_8[1], y_2[2], y_{17}[3], ..., x_4[N]$ のようになる（但し $N = m + n$）。

3. ここで問題にしたいのは、それぞれの変数の順位の合計がいくつになるかということである。ただし、順位の総合計は $(N + 1)N/2$ に決まっているので、片方の変数だけ考えれば残りは引き算でわかる。そこで、変数 X だけ考えることにする。

4. X に属する x_i（$i = 1, 2, ..., m$）の順位を R_i と書くと、X の順位の合計は

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

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

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

5. もし X と Y に差がなければ、X は N 個のサンプルから偶然によって m 個取り出したものであり、Y がその残りである、と考えることができる。順位についてみると、$1, 2, 3, ..., N$ の順位から m 個の数値を取り出すことになる。同順位がなければ、ありうる組み合わせは、${}_NC_m$ 通りある（`choose(N,m)` によって得られる）。

6. $X > Y$ の場合には、${}_NC_m$ 通りのうち、合計順位が R_X と等しいかより大きい場合の数を k とする（$X < Y$ の場合は、合計順位が R_X と等しいかより小さい場合の数を k とする）。

[6] 同順位がある場合の扱いは後述する。

7. $k/{}_N C_m$ が有意水準 α より小さいときに $H0$ を疑う。N が小さいときは有意になりにくいが、N が大きすぎると計算が大変面倒である（もっとも、R で `wilcox.test(X,Y,exact=T)` とすれば、サンプル数の合計が 50 未満で同順位の値がなければ、総当りして正確な確率を計算してくれるので簡単である。が、つい 15 年くらいまではコンピュータは誰もが使える道具ではなかったし、総当りをするには計算時間がかかりすぎた。今のコンピュータでもサンプルサイズが大きいと、総当りでは計算時間がかかりすぎて実用的でない）。そこで、正規近似を行う（つまり、期待値と分散を求めて、統計量から期待値を引いて分散の平方根で割った値が標準正規分布に近似的に従うという関係を用いて検定する）。

8. 帰無仮説 H_0 のもとでは、期待値は

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

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

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

から、

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

となるので[7]、

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

と

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

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

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

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

9. 標準化[8]して連続修正[9]し、$z_0 = \{|R_X - E(R_X)| - 1/2\}/\sqrt{var(R_X)}$ を求める。m と n が共に大きければこの値が標準正規分布に従うので、例えば $z_0 > 1.96$ ならば、両側検定で有意水準 5%で

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

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

となる。

[8]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。

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

有意である。R で有意確率を求めるには、z_0 を `z0` と書けば、`2*(1-pnorm(z0,0,1))` とすればよい。

10. ただし、同順位があった場合は、ステップ 2 の「小さい方から順に番号をつける」ところで困ってしまう。例えば、変数 X が $\{2,6,3,5\}$、変数 Y が $\{4,7,3,1\}$ であるような場合には、X にも Y にも 3 という値が含まれる。こういう場合は、下表のように平均順位を両方に与えることで、とりあえず解決できる。

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

11. ただし、このやり方では、正規近似をする場合に分散が変わる[10]。帰無仮説の下で、$E(R_X) = m(N+1)/2$ はステップ 8 と同じだが、分散が

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

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

例として、`survey` データで、身長 (`Height`) の分布の位置が男女間で差がないという帰無仮説を検定してみよう。

R コンソールでは簡単で、`library(MASS)` してあれば、

```
> wilcox.test(Height ~ Sex, data=survey)
```

だけで良い。

Wilcoxon の順位和検定の効果量については、`https://core.ecu.edu/wuenschk/docs30/Nonparametric-EffectSize.pdf` で説明されているように、p 値から Z 値を求めて（両側検定の p 値から求める場合は 2 で割ることを忘れないように）サンプルサイズの平方根で割るという正規近似による方法と、`effectsize` パッケージを使って `rank_biserial()` 関数を使う方法が知られている（得られる効果量は Cliff's delta といい、負の値も取り得るが、絶対値を解釈すれば良い）。この効果量 r は、`https://cran.r-project.org/web/packages/statsExpressions/vignettes/stats_details.html` に書かれ

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

ているように、0.1-0.3 で小さい、0.3-0.5 で中程度、0.5 以上で大きいと判断するのが目安である。上記の例であれば、コードは以下。

```
> library(MASS)
> res <- wilcox.test(Height ~ Sex, data=survey)
> abs(qnorm(res$p.value/2))/sqrt(sum(complete.cases(survey[,c("Height", "Sex")])))
> library(effectsize)
> rank_biserial(Height ~ Sex, data=survey)
```

Rcmdr では「統計量」の「ノンパラメトリック検定」で「2 標本ウィルコクソン検定」を選び、グループ変数として `Sex` **を選び、応答変数として** `Height` **を選んで [OK] ボタンをクリックする。**
EZR では「統計解析」「ノンパラメトリック検定」から「2 群間の比較（Mann-Whitney U 検定）」を選び、目的変数として `Height` **を、比較する群として** `Sex` **を選んで「OK」ボタンをクリックする。検定結果が Output ウィンドウに表示されるだけでなく、箱ひげ図も同時に描かれる。身長の性差の検定結果は Welch の t 検定と同じく有意差があるといえる。**
同じ `survey` **データフレームで年齢（**`Age`**）に性差があるかどうかについては、Welch の t 検定では有意でなかったが、Wilcoxon の順位和検定をすると 5%水準で有意である。これは** `Age` **の分布が大きく右裾を引いた歪んだ分布になっていて、しかも外れ値が多いため、Welch の t 検定の検出力が低くなっているためである。**

7.6 Brunner-Munzel 検定

詳しくは、web 上の三重大学奥村晴彦教授の解説記事[11]や hoxo__m さんの解説記事[12]を参照されると良いと思うが、「独立 2 群から一つずつ値を取り出したとき、どちらが大きい確率も等しい」という帰無仮説を検定するために[13]、Brunner と Munzel が提案した比較的新しい方法である。残念ながら現在のところ EZR には入っていない。

元論文は、Brunner E, Munzel U (2000) The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, 42: 17-25. であり、R では、lawstat パッケージの `brunner.munzel.test()` で実行できる（使うためには、予め `install.packages("lawstat", dep=TRUE)`

[11] `http://oku.edu.mie-u.ac.jp/~okumura/stat/brunner-munzel.html`
[12] `http://d.hatena.ne.jp/hoxo_m/20150217/p1`
[13] ただし、裏 RjpWiki さんの記事（`https://blog.goo.ne.jp/r-de-r/e/2c2f187d4975cc0928e6f4a0710d6191`）で、Mann Whitney の U 検定や Wilcoxon の順位和検定と実質的に同じで、検定に使う分布が違うだけで、並べ替え検定にしたら完全に一致すると批判されている。しかしサンプルサイズが小さい場合は近似に使う分布が違えば得られる p 値は異なるので、違う検定と考えても良いであろう。少なくともズレのモデルを仮定していないとは言える。

により、各種依存パッケージとともに lawstat パッケージをインストールしておく必要がある)。

上の例と同様に、MASS パッケージに入っている survey データで、身長の分布の位置が男女で差がないという帰無仮説を Brunner-Munzel 検定するには、

```
> library(lawstat)
> HT <- tapply(survey$Height,survey$Sex,print)
> brunner.munzel.test(HT$Male,HT$Female)
```

とすれば良い。下枠内の結果が得られる。**Age についても同様に計算できるので確かめられたい。**

```
        Brunner-Munzel Test

data:  HT$Male and HT$Female
Brunner-Munzel Test Statistic = -17.6824, df = 157.29, p-value < 2.2e-16
95 percent confidence interval:
 0.0615193 0.1496350
sample estimates:
P(X<Y)+.5*P(X=Y)
       0.1055771
```

なお、サンプルサイズが極めて小さい場合は、Neubert K, Brunner E (2007) A studentized permutation test for the non-parametric Behrens-Fisher problem. *Computational Statistics and Data Analysis*, 51: 5192-5204. に示されているように、Brunner-Munzel 統計量に並べ替え検定を適用すれば良いが(前掲した三重大学奥村晴彦教授の解説記事にコードが示されている)、相当な計算時間がかかる。

最近開発された、brunnermunzel パッケージの brunnermunzel.test() 関数の perm=TRUE オプションを使えば(強引にさせたい場合は force=TRUE オプションもつければ)、高速な並べ替え検定が可能であるが、欠損値が含まれているとエラーが出るので、予め欠損値のない subset にしておくか、多重代入法などで欠損値を推定してから適用し結果を統合するという面倒なプロセスを踏まねばならないし、サンプルサイズが大きい場合もエラーが出て計算できない。

install.packages("brunnermunzel", dep=TRUE) によって brunnermunzel パッケージをインストールした後、以下のコードを実行すると Brunner-Munzel 検定ができるが、サンプルサイズが大きすぎて並べ替え検定の結果は得られない。

```
> library(brunnermunzel)
> mHT <- subset(survey, !is.na(Height)&Sex=="Male")$Height
> fHT <- subset(survey, !is.na(Height)&Sex=="Female")$Height
> brunnermunzel.test(mHT, fHT, perm=FALSE)
> brunnermunzel.test(mHT, fHT, perm=TRUE, force=TRUE) # エラーになる
```

そこで、2 種類の催眠剤 (group) を 10 人の被験者 (ID) に使った場合の睡眠時間増加 (extra) データである sleep を使って試すには(このデータは同じ人の 2 回の測定値であり、対応があるため、

Brunner-Munzel 検定には向いていないが、適当なサンプルサイズであるため使ってみた）以下の通り。

```
d1sleepextra <- subset(sleep, group=="1")$extra
d2sleepextra <- subset(sleep, group=="2")$extra
brunnermunzel.test(d1sleepextra, d2sleepextra, perm=FALSE)
brunnermunzel.test(d1sleepextra, d2sleepextra, perm=TRUE, force=TRUE)
```

この結果の p 値をみると、近似的な検定では 0.04682 と 5%水準では有意だが、並べ替え検定では 0.05513 と有意でない。

7.7　Wilcoxon の符号付き順位検定

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

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

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

```
> wilcox.test(survey$Wr.Hnd, survey$NW.Hnd, paired=TRUE)
```

とすればよい。

Rcmdr では、「統計量」、「ノンパラメトリック検定」、「対応のあるウィルコクソン検定」と選択し（EZR では、「統計解析」「ノンパラメトリック検定」から「対応のある 2 群間の比較（Wilcoxon の符号付き順位和検定）」を選び）、第 1 の変数として左側のリストから `Wr.Hnd` を選び、第 2 の変数として右側のリストから `NW.Hnd` を選んで [OK] ボタンをクリックするだけである。順位和検定のときと同じく検定方法のオプションを指定できるが、通常はデフォルトで問題ない。
ちなみに、この検定結果は `p=0.0825` となり、5%水準で有意でない。対応のある t 検定では有意だったのにこうなるのは、データが正規分布に近い分布である場合には、t 検定の方が検出力が大きいためである。`Diff.Hnd` は分布の正規性の検定をすると Shapiro-Wilk の検定で 5% 水準で有意に正規分布と差があるけれども（1 人だけある外れ値を除外しても有意差あり）、外れ値を除外してヒストグラムを描くと正規分布に近いように見える。こういう場合は、一般に t 検定の方が検出力が高い。

7.8 2 群間での順序尺度の比較

食物摂取頻度調査の結果のように、頻度を順序尺度で示すデータについて、独立 2 群間での比較をしたい場合、(1) 各頻度カテゴリを 1 つの数値で代表させ、量として扱うことで t 検定する、(2) 順序なので Wilcoxon の順位和検定（Mann-Whitney の U 検定）を行う、(3) 順序の情報を無視してカテゴリとしてカイ二乗検定する、といった方法が良く行われているが、(1) は代表のさせ方が妥当である保証がなく、(2) は同順位が多くなると考えられるため正しい p 値が得られない可能性が高く、(3) は順序の情報を捨ててしまうため検出力が下がるという問題がある。

このような場合、グループで順序を説明するという順序ロジスティック回帰と、定数で順序を説明する（つまり順序間に差が無い）という順序ロジスティック回帰の結果の尤度比をとり、尤度比検定をすることができる。EZR のメニューにはないが、以下に例を示す。**y** が 6 段階の順序尺度をもつ変数で、**x** が A と B の 2 群を示す変数である。

```
set.seed(123)
y1 <- c(sample(1:3, 20, rep=TRUE), sample(4:6, 30, rep=TRUE))
y2 <- c(sample(1:3, 30, rep=TRUE), sample(4:6, 20, rep=TRUE))
y <- as.ordered(c(y1, y2))
x <- as.factor(c(rep("A", 50), rep("B", 50)))
table(y, x)
library(MASS) # polr() 関数を使うので
anova(polr(y ~ x), polr(y ~ 1)) # 尤度比検定
chisq.test(table(y, x)) # カイ二乗検定でもやってみる
```

第8章　2つのカテゴリ変数間の関係

2つのカテゴリ変数間の関係としては、独立かどうか、独立でないとしたらどの程度の関連があるかを調べることになるが、2つのカテゴリ変数の関係が独立であるという帰無仮説は、第1のカテゴリ変数で示されるグループ間で、第2のカテゴリ変数で示される構成比に差がないという帰無仮説と同等である。これらのカテゴリ変数がともに2値変数である場合は、これは2群の母比率の差の検定に帰着する。

8.1　2群の母比率の差の検定

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

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

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

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

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

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

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

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

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

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

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

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

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

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

母比率の推定と、2 群間でその差がないという帰無仮説の検定[1]、差の 95%信頼区間を一気に出力してくれる。survey データフレームで「利き手が左である割合に男女で差が無い」という帰無仮説を検定するには、

```
> prop.test(table(survey$Sex,survey$W.Hnd))
```

とすれば良い。

[1]連続性の補正済み。事象が生起しない場合についても考慮し、カイ二乗適合度検定をしているので、次節に示す 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。

Rcmdr では、「統計量」の「比率」から「２標本の比率の検定」を選ぶ。survey **データフレームで、「利き手が左である割合に男女で差がない」という帰無仮説を検定するには、グループとして** Sex**、目的変数として** W.Hnd **を指定し、検定のタイプとして「連続修正を用いた正規近似」にチェックを入れて [OK] ボタンをクリックすればよい。**
EZR では、「統計解析」「名義変数の解析」から「分割表の作成と群間の比率の比較」を選ぶ。この例では、行の選択の枠から Sex**、列の変数の枠から** W.Hnd **を選び、「仮説検定」のなかから「カイ２乗検定」の左のボックスにチェックを入れ、「カイ２乗検定の連続性補正」の下のラジオボタンを「はい」にして「OK」ボタンをクリックする。**

8.2 独立性の検定

２つのカテゴリ変数の関係を考えるとき、一般に、もっともよく行われるのは、それらが独立であるという帰無仮説を立てて検定することである。本節ではその仕組みについて説明する。

カテゴリ変数のもつ統計的な情報は、カテゴリごとの度数だけである。そこで、２つのカテゴリ変数の間に関係について検討したいときには、まず**それらの組み合わせの度数**を調べた表を作成する（R では `table()` や `xtabs()` という関数が使える）。これを**クロス集計表**と呼ぶ。とくに、２つのカテゴリ変数が、ともに２値変数のとき、そのクロス集計は２×２クロス集計表 [2 by 2 cross tabulation]（２×２分割表 [2 by 2 contingency table]）と呼ばれ、その統計的性質が良く調べられている。**以下では２×２分割表の例で独立性のカイ二乗検定とフィッシャーの正確確率検定を説明するが、これらはカテゴリが３つ以上あっても通用する。**同じ目的で使われる検定に、別名 G 検定とも呼ばれる**尤度比検定**があるが、これは R では `vcd` パッケージの `assocstats()` 関数で計算できる。

なお、２つのカテゴリ変数の両方が順序のあるカテゴリである場合の独立性の検定としては、「線形連関の検定」（出典：藤井良宜『R で学ぶデータサイエンス 1　カテゴリカルデータ解析』共立出版, pp.65-68、英語では Agresti A (2002) *Categorical Data Analysis.* Hoboken, New Jersey: John Wiley & Sons. で、“linear-by-linear association test” と呼ばれている。SPSS の日本語版では「線型と線型による連関検定」と意味の通じにくい訳語になっているので注意）を使う方法もある。これは、２つの順序のあるカテゴリ変数の各個体に対して順序を整数のスコアとして与え、スコア間で計算したピアソンの積率相関係数を２乗した値にサンプルサイズから１を引いた値を掛けた統計量が、近似的に自由度１のカイ二乗分布に従うことから検定を実行するものであり、R では `coin` パッケージの `lbl_test()` 関数で計算できる。

8.2.1　カイ二乗検定

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

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

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

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

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

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

であるから、

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

として、自由度 3 のカイ二乗検定をすればよいことになる。しかし、普通、π は未知である。そこで、$Pr(\bar{A}) = 1 - Pr(A)$ かつ、この 2 つのカテゴリ変数が独立ならば $Pr(A \cap B) = Pr(A)Pr(B)$ と考えれば良い[3]ことを使って、$Pr(A)$ と $Pr(B)$ を母数として推定する[4]。$Pr(A)$ 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。

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

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

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

同様に考えれば、母集団の各セルの確率は下式で得られる。

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

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

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

これらの値を使えば、

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

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

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

分子の中括弧の中は N^2 なので、結局、

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

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

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

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

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

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

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

例題

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

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

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

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

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

となる。従って、**連続性の補正を行なった**カイ二乗統計量は、

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

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

R コンソールでは次の 1 行を打つだけで上の結果を得ることができる。

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

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

例題

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

R コンソールでは次の２行を打てばできる。

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

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

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

8.2.2 フィッシャーの正確確率

期待度数が低い組み合わせがあるときには、カイ二乗検定での正規近似が非常に悪い近似になる。そういう場合、カテゴリを併合して変数を作り直し、組み合わせの種類を減らして、各組み合わせの頻度を上げる方法もあるが、恣意的なカテゴリの併合は結果を歪める可能性もあるし、元々カテゴリが 2 つずつだったらそれ以上減らせないので、もっといい方法が考案されている。それがフィッシャーの正確確率（検定）である。

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

もう少し丁寧に言うと、サイズ N の有限母集団があって、そのうち変数 A の値が 1 である個体数が m_1、1 でない個体数が m_2 あるときに、変数 B の値が 1 である個体数が n_1 個（1 でない個体数が $n_2 = N - n_1$ 個）あるという状況を考え、この n_1 個のうち変数 A の値が 1 である個体数がちょ

うど a である確率を求めることになる。これは、m_1 個から a 個を取り出す組み合わせの数と m_2 個から $n_1 - a$ 個を取り出す組み合わせの数を掛けて、N 個から n_1 個を取り出す組み合わせの数で割った値になる。これと同じ周辺度数をもつ 2 × 2 分割表のうち、確率がこれと同じかこれよりも小さい表の確率をすべて足し合わせたものが、「変数 A と変数 B が独立」という帰無仮説が成り立つ確率になる。

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

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

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

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

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

例題

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

R コンソールでは次の 2 行を打てばよい。

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

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

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

8.3 カテゴリ変数間の関連性の指標

ここまで説明した検定は、２つのカテゴリ変数が独立であるという帰無仮説の検定であって、得られる p 値が有意水準より小さくて帰無仮説が棄却された場合、p 値の小ささは関連がない可能性がどれほど低いかを示す意味しか無く、どの程度の関連があるかは示さない。ただし、フィッシャーの正確確率を得る関数の出力をよく見ると、オッズ比が表示されており、このオッズ比は関連の大きさを示す指標の一つである。

「独立でない」場合について、関連の強さを評価したいときに用いる指標について述べよう。「相関と回帰」の章に示すように、量的な変数（連続変数）の場合は相関係数が関連の強さの指標となるが、カテゴリ変数の場合は、既に述べたオッズ比やリスク比のほか、ファイ係数（記号は ϕ を用いるのが普通）と呼ばれる指標は、要因の有無、発症の有無を 1,0 で表した場合のピアソンの積率相関係数と同じ計算式で得られる。正の関連性だけ考えるなら、π_1、π_2 を要因ありのうちの発症者割合、要因無しのうちの発症者割合、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、$\phi = \sqrt{(\pi_1 - \pi_2)(\theta_1 - \theta_2)}$ である。実は二乗して N を掛けると連続修整無しのカイ二乗値に一致する。以下のコードを試されたい。

```
phi <- function(X) { # X は 2 × 2 の行列
M <- rowSums(X)
N <- colSums(X)
phi0 <- (X[1,1]/M[1]-X[2,1]/M[2])*(X[1,1]/N[1]-X[1,2]/N[2])
phi1 <- ifelse(phi0<0, -sqrt(-phi0), sqrt(phi0))
phi2 <- (X[1,1]*X[2,2] - X[1,2]*X[2,1])/sqrt(M[1]*M[2]*N[1]*N[2])
return(c(phi1, phi2))}

XX <- matrix(c(80, 20, 55, 45), 2, 2)
phi(XX)
phi(XX)^2*sum(XX)
chisq.test(XX, correct=FALSE)$statistic
```

同様の目的でピアソンのコンティンジェンシー係数や Cramer の V という係数も用いられる。以上すべては vcd パッケージの assocstats() 関数にクロス集計表を与えると計算してくれる。

また、カテゴリ変数または順序変数間の相関関係をみるには、ポリコリック相関係数を用いる方法もある（関連して書いておくと、順序変数と量的変数の間の関連は、ポリシリアル相関係数でみることができる）。これらは、polycor パッケージに含まれており、前者は polychor() 関数（2 つのカテゴリ変数を与える。標準誤差が欲しいときは std.err=TRUE オプションを付ける）、後者は polyserial() 関数で求めることができる。

ポリコリック相関係数の計算例

```
> library(MASS) # survey データフレームを含む
> library(polycor) # 未インストールの際は一度だけ以下を実行
>                  # install.packages("polycor", dep=TRUE)
 要求されたパッケージ mvtnorm をロード中です
 要求されたパッケージ sfsmisc をロード中です
> polychor(survey$Smoke, survey$Exer, std.err=TRUE)

Polychoric Correlation, 2-step est. = -0.042 (0.1006)
Test of bivariate normality: Chisquare = 5.628, df = 5, p = 0.3441
> # カテゴリ順序を付け直す。
> # EZR では「アクティブデータセット」「変数の操作」「因子水準を再順序化する」
> survey$Smoke <- factor(survey$Smoke,
   levels=c("Never", "Occas", "Regul", "Heavy"), ordered=TRUE)
> survey$Exer <- factor(survey$Exer,
   levels=c("None","Some","Freq"), ordered=TRUE)
> xtabs(~Smoke+Exer, data=survey)
        Exer
Smoke    None Some Freq
  Never    18   84   87
  Occas     3    4   12
  Regul     1    7    9
  Heavy     1    3    7
> polychor(survey$Smoke, survey$Exer, std.err=TRUE)

Polychoric Correlation, 2-step est. = 0.1269 (0.1011)
Test of bivariate normality: Chisquare = 4.274, df = 5, p = 0.5107
```

正しい順序の順序変数にすると、ポリコリック相関係数は、順序を無視したときとは結果が変わることがわかる（もちろん、後述するようにカイ二乗検定やフィッシャーの正確検定の結果は、順序の情報を使わないので変わらない）。順序変数の間では、量的変数としての扱いを強制し、スピアマンの順位相関係数を求めることもできるが、同順位が多いため p 値が正しく得られないという問題を生じるので、ポリコリック相関係数を使う方が良い。

第9章　3群以上の比較

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

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

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

9.1　一元配置分散分析

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

例えば、南太平洋の３つの村 X, Y, Z で健診をやって、成人男性の身長や体重を測ったとしよう。このとき、データは例えば次のようになる（架空のものである）。このデータは `https://minato.sip21c.org/grad/sample2.dat` として公開してあり、R から `read.delim()` 関数で読み込める。

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

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

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

すると、次の枠内に示す「分散分析表」が得られる。

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

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

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

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

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

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

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

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

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

Rcmdr ではメニューにないので、スクリプトウィンドウで `aov` の部分を `oneway.test` に書き直して「実行」するしかなかったが、EZR は 1.21 から一元配置分散分析のメニューウィンドウに「Welch 検定」として等分散を仮定しないオプションが追加された。

9.1.1　一元配置分散分析の効果量

ケンブリッジ大学の統計 Wiki の効果量についての FAQ[1] によると、一元配置分散分析の効果量は、通常 η^2[2] という記号で示される「効果の偏差平方和を、それ自身とその効果に関連した誤差の偏差平方和の和で割った値」になる。これは、分散分析表で得られる分散比である F、グループ数 k、サンプルサイズを N とすると、

$$\eta^2 = \frac{F \cdot (k-1)}{F \cdot (k-1) + (N-k)}$$

で得られ[3]、目安として 0.01 以上 0.06 未満で弱い効果、0.06 以上 0.14 未満で中程度の効果、0.14 以上で大きな効果であるとされている。

Rcmdr や EZR のメニューには含まれていないが、R では、`lsr` パッケージの `etaSquared()` 関数や、`rstatix` パッケージの `eta_squared()` 関数に、`aov()` 関数の出力オブジェクトを与えれば計算できる。これらの関数には、等分散を仮定しない `oneway.test()` 関数の出力オブジェクトは与えることができないが、F 値とグループ数とサンプルサイズを使った式が等分散を仮定しない場合にも成り立つとすれば、上の式から計算できる。

6 種類の異なる餌をヒヨコに与えて育てた後の鶏の体重データである `chickwts` を使った実行例を以下に示しておく。どの計算方法でも、等分散を仮定した ANOVA については $\eta^2 = 0.542$ となることがわかる。等分散を仮定しない Welch の ANOVA については $\eta^2 = 0.602$ となる（正しいかどうかは確認が十分に取れていないので注意）。いずれにせよ、与える餌の種類が鶏の体重に与える効果は大きいと言える。

[1] `http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize`
[2] イータ二乗とかイータスクウェアドと読む。
[3] 当該 FAQ ページの式は間違っていて、正しい式は、そのリンク先の下の方に説明されているので注意。

```
# 普通の ANOVA
res <- aov(weight ~ feed, data=chickwts)
summary(res)
# 偏差平方和から計算
SS <- summary(res)[[1]]$"Sum Sq"
SS[1] / (SS[1]+SS[2]) # eta squared
# F 値とグループ数とサンプルサイズから計算
.k <- length(levels(chickwts$feed))
.N <- length(chickwts$feed)
.F <- summary(res)[[1]]$"F value"[1]
.F * (.k-1) / (.F * (.k-1) + .N - .k) # eta squared
# lsr パッケージ
library(lsr)
etaSquared(res, anova=TRUE)
# rstatix パッケージ
library(rstatix)
eta_squared(res)
# 等分散を仮定しない Welch の ANOVA
res2 <- oneway.test(weight ~ feed, data=chickwts)
res2
..F <- res2$statistic
..F * (.k-1) / (..F * (.k-1) + .N - .k) # eta squared?
```

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

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

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

き、各群について統計量 B_i を $B_i = n_i\{R_i/n_i - (N+1)/2\}^2$ として計算し、

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

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

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

により H を補正した値 H' を求める。

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

上の例で村落の身長への効果をみるには、R コンソールでは、

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

と打てば結果が表示される。

Rcmdr では、「統計量」、「ノンパラメトリック検定」、「クラスカル－ウォリスの検定...」と選び、「グループ」として VG を、「目的変数」として HEIGHT を選び、[OK] をクリックする。
EZR では、「統計解析＞ノンパラメトリック検定＞ 3 群以上の間の比較 (Kruskal-Wallis 検定)」と選び、目的変数（1 つ選択）の枠で HEIGHT を、グループ（1 つ選択）の枠で VG を選び、[OK] をクリックする。

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

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

9.2.1 Kruskal-Wallis 検定の効果量

Rcmdr や EZR には含まれていないが、`https://rcompanion.org/rcompanion/` で"An R Companion for the Handbook of Biological Statistics"という素晴らしいテキスト（デラウェア大学の John H. McDonald によるテキスト"Handbook of Biological Statistics"[4]の例を使って R による解析方法を示すために、ラトガース大学の Salvatore S. Mangiafico 准教授が書いたもの）が公開されており、そのために作られた `rcompanion` パッケージを使うと、`epsilonSquared()` 関数や `ordinalEtaSquared()` 関数で効果量が推定できる。

`epsilonSquared()` 関数で得られる ϵ^2 は、目安として 0.01 以上 0.08 未満で弱い効果、0.08 以上 0.26 未満で中程度の効果、0.26 以上で大きな効果とされている。`ordinalEtaSquared()` 関数で得られるのは η^2 なので、効果の大きさの目安は ANOVA の場合と同様である。

ANOVA の場合と同じく、`chickwts` データを使った実行例を示しておく。$\epsilon^2 = 0.533$、$\eta^2 = 0.498$ となるので、与える餌の種類は鶏の体重に大きな効果をもっていると言える。なお、`epsilonSquared()` 関数に `ci=TRUE` オプションを付けることにより、ブートストラップ法で求めた η^2 の 95%信頼区間が得られる。

```
kruskal.test(weight ~ feed, data=chickwts)
library(rcompanion)
epsilonSquared(chickwts$weight, chickwts$feed, ci=TRUE)
ordinalEtaSquared(chickwts$weight, chickwts$feed)
```

9.3 検定の多重性の調整を伴う対比較

多重比較の方法にはいろいろあるが、良く使われているものとして、ボンフェローニ (Bonferroni) の方法、ホルム (Holm) の方法、シェフェ(Scheffé) の方法、テューキー (Tukey) の HSD、ダネット (Dunnett) の方法、ウィリアムズ (Williams) の方法がある。最近は、FDR(False Discovery Rate) 法もかなり使われるようになった。ボンフェローニの方法とシェフェの方法は検出力が悪いので、特別な場合を除いては使わない方が良い。テューキーの HSD またはホルムの方法が薦められる。なお、ダネットの方法は対照群が存在する場合に対照群と他の群との比較に使われるので、適用場面が限定されている[5]。ウィリアムズの方法は対照群があって他の群にも一定の傾向が仮定される場合には最高の検出力を発揮するが、ダネットの方法よりもさらに限られた場合にしか使えない。

テューキーの HSD は平均値の差の比較にしか使えないが、ボンフェロー二の方法、ホルムの方法、FDR 法は位置母数のノンパラメトリックな比較にも、割合の差の検定にも使える。R コンソー

[4] `http://www.biostathandbook.com/HandbookBioStatThird.pdf` として無償公開されている。

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

ルでは、pairwise.t.test()、pairwise.wilcox.test()、pairwise.prop.test() という関数で、ボンフェローニの方法、ホルムの方法、FDR 法による検定の多重性の調整ができる。

fmsb パッケージを使えば、pairwise.fisher.test() により、Fisher の直接確率法で対比較をした場合の検定の多重性の調整も可能である。

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

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

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

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

とすれば、2 村落ずつのすべての組み合わせについてボンフェローニの方法で有意水準を調整した p 値が表示される（"bonferroni"は"bon"でも良い。また、pairwise.*系の関数では data=というオプションが使えないので、データフレーム内の変数を使いたい場合は、予めデータフレームを attach() しておくか、またはここで示したように、変数指定の際に一々、“データフレーム名$” を付ける必要がある）。

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

とすれば、ボンフェローニの方法で有意水準を調整した、すべての村落間での身長の差を順位和検定した結果を出してくれる。これらの関数で、p.adjust.method を指定しなければホルムの方法になるが、明示したければ、p.adjust.method="holm"とすればよい。FDR 法を使うには、p.adjust.method="fdr"とすればよい。R でもボンフェローニが可能なのは、一番単純な方法であるという理由と、ホルムの方法に必要な計算がボンフェローニの計算を含むからだと思われる。なお、R を使って分析するのだけれども、データがきれいな正規分布をしていて、かつ古典的な方法の論文しか accept しない雑誌に対してどうしても投稿したい、という場合は、TukeyHSD(aov(HEIGHT ~ VG, data=sp)) などとして、テューキーの HSD を行うことも可能である。

Rcmdr の場合は、「統計量」の「平均」から「１元配置分散分析」を選んで実行するときに、「2 組ずつの平均の比較（多重比較）」の左のボックスにチェックを入れておけば、自動的に Tukey の HSD で検定の多重性を調整した対比較を実行してくれるし、同時信頼区間のグラフも表示される。しかし、第一種の過誤を調整する方法は、まだサポートされていない。

EZR では、一元配置分散分析メニューのオプションとして実行できる。「統計解析」「連続変数の解析」「3 群以上の平均値の比較（一元配置分散分析 one-way ANOVA)」を選んで、「目的変数」として HEIGHT を、「比較する群」として VG を選んでから、下の方の「↓ 2 組ずつの比較 (post-hoc 検定) は比較する群が１つの場合のみ実施される」から欲しい多重比較法の左側のボックスにチェックを入れてから「OK」ボタンをクリックする。ただし、等分散を仮定しない「一般化 Welch 検定」の場合は、多重比較オプションが使えない。
これは、Rmcdr や EZR に実装されている、Bonferroni や Holm の方法で多重性を調整した t 検定や Tukey の方法が等分散を仮定しているためである。

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

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

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

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

実は、等分散を仮定しない多重比較も存在する。Games-Howell 法（元論文：Games PA, Howell JF (1976) Pairwise Multiple Comparison Procedures with Unequal N's Pairwise Multiple Comparison Procedures with Unequal N's and/or Variances: A Monte Carlo Study. *Journal of Educational Statistics*, 1(2) : 113-125.）と呼ばれる。

群馬大学青木先生の web サイトの Tukey 法のページ[6]に説明と R コードが掲載されている。そのため、青木先生のコードを利用している弘前大学の「改変 R コマンダー」[7]では Games-Howell 法が実行可能である。青木先生のコードについては、英語だが、http://www.gcf.dkf.unibe.ch/BCB/files/BCB_10Jan12_Alexander.pdf の説明もわかりやすい。他にオリジナルなコードとして、https://gist.github.com/aschleg/ea7942efc6108aedfa9ec98aeb6c2096 も存在するが、おそらく rosetta パッケージの posthocTGH() 関数を使うのが簡単だろう[8]。R の組み込みデータである chickwts（6 種類の餌 [feed] で育てた鶏の体重 [weight] の比較）を使って実行する例を示しておく。ちなみに、このコードで games-howell となっているところを tukey にすれば、TukeyHSD() 関数を使った場合と同じ結果が得られる。

```
if (require(rosetta)==FALSE) {
 install.packages("rosetta", dep=TRUE)
 library(rosetta) }
posthocTGH(chickwts$weight, chickwts$feed, method="games-howell")
```

[6] http://aoki2.si.gunma-u.ac.jp/R/tukey.html

[7] http://www.hs.hirosaki-u.ac.jp/~pteiki/research/stat/S/

[8] この関数は、以前は userfriendlyscience パッケージに入っていたが、ufs という名前に変わったときに、posthocTGH() 関数は rosetta パッケージに移ったようだ。弘前大学のサイトの説明を見た限りでは、R-4.0.5 以降の変更らしい。

9.4 Dunnett の多重比較法

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

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

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

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

9.5　3 群間の比率の差の検定、少なくとも 1 つの変数が 3 水準以上ある場合の 2 × 2 クロス集計表

`prop.test()` 関数は、3 群以上の間でも、「どの群でも事象の生起確率に差がない」という帰無仮説を検定するのに使える。例えば、A, B, C という 3 つの処理を実施したときの脱落が、A: 0/15, B: 2/13, C: 1/14 であったとき、「3 つの処理間で脱落率に差が無い」という帰無仮説の下で期待される脱落率は、

$$\frac{0+2+1}{15+13+14} = 3/42$$

これが正しければ、期待されるクロス表は

	A	B	C
脱落	15*(3/42)	13*(3/42)	14*(3/42)
完了	15*(39/42)	13*(39/42)	14*(39/42)

すべてのセルについて、観測度数と期待度数の差の2乗を期待度数で割った値を合計したものがカイ二乗値χ^2

$$\chi^2 = (0 - 15 \times (3/42))^2/(15 \times (3/42)) + ... = 2.485...$$

χ^2は自由度2（3処置、2種類の結果だから、それぞれから1を引いて積をとったものが自由度）のカイ二乗分布に従う。`1-pchisq(X2, 2)`を計算すると（注：`pchisq(X2, 2)`は自由度2のカイ二乗分布の`X2`までの確率密度の積分値）、0.2886となる。Rコンソールでは、

```
prop.test(c(0, 2, 1), c(15, 13, 14), correct=FALSE)
```

と打てば、`X-squared = 2.4852, df = 2, p-value = 0.2886`と結果が得られる。

この例では脱落数が少ないので、カイ二乗検定では近似が悪い。そういう場合は、「処理間で脱落率に差が無い」という帰無仮説は、「脱落するかどうかは処理の種類と無関係（＝独立）」という帰無仮説と同値なので、以下のクロス表を想定し、フィッシャーの正確な検定 (Fisher's exact test) を実行する。

	A	B	C	Total
Dropout	0	2	1	3
Complete	15	11	13	39
Total	15	13	14	42

「脱落するかどうか」と「処理の種類」が偶然この組合せになっている確率を計算するには、この表と同じ周辺度数をもつクロス表（例えば下記）をすべて考え、それぞれが偶然得られる確率を計算する。下表の確率は、${}_{15}C_1 \times {}_{13}C_1 \times {}_{14}C_1 / {}_{42}C_3$で、約0.238となる。実際の脱落データの表は、${}_{15}C_0 \times {}_{13}C_2 \times {}_{14}C_1 / {}_{42}C_3$で、約0.095となる。

	A	B	C	Total
Dropout	1	1	1	3
Complete	14	12	13	39
Total	15	13	14	42

実際の表より偶然得られる確率が小さな表の確率を合計すると、フィッシャーの正確な確率が得られる。つまり、0.095+0.040+0.025+0.032 = 0.192となる。Rコンソールでは、`fisher.test(matrix(c(0, 15, 2, 11, 1, 13)`で0.1914となる（違いは丸め誤差）である。

2群ずつ比べて、どの群間で差があるのかをみようとすると、平均値の場合と同様に検定の多重性が生じるので、平均値の差の場合と同様、第一種の過誤を調整する必要があり、ボンフェローニの

方法やホルムの方法、あるいは FDR 法を用いることになる。R の関数は pairwise.prop.test() である。

なお、3 群以上の間で事象の生起確率に一定の傾向がみられるかどうかを調べたい場合には、コクラン＝アーミテージの検定という手法がある。例えば、漁師 100 人、農民 80 人、事務職 30 人について便の検査をして、日本住血吸虫虫卵陽性者が 60 人、30 人、8 人だったとしたとき、職業的な貝との接触リスクに対して勝手に漁師を 4、農民を 2、事務職を 1 とスコアリングして、陽性割合の増加傾向が、このスコアと同じかどうかを調べることができる。この場合なら、R コンソールのコマンドは以下のようになる。

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

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

EZR では、「統計解析」「名義変数の解析」から「分割表の直接入力と解析」を選び、列数バーを一つ右にずらして 2 行 3 列の分割表にしてから数値を入力する。行ラベルと列ラベルの欄には整数が入っているが文字列に書き換えることができる。日本語も入力可能である。「仮説検定」のところで独立性のカイ２乗検定（連続補正有り）の左のボックスにチェックを入れれば、Yates の連続性の補正をした検定結果が得られる。コクラン＝アーミテージ検定は、生データから計算するメニューはあるが、集計後データからは計算できないので、実行したい場合はそのような生データを生成する必要がある（詳細は省略）。

生データから計算する場合について、MASS パッケージの survey データセットを使って例示しよう。Clap という変数は、両手を叩き合わせたときにどちらが上に来るかを意味し、左、右、ど

ちらでもない、という3つのカテゴリからなる。W.Hnd は字を書く手がどちらか、つまり利き手を意味する。両手を叩き合わせたときに上に来る手の3タイプ間で、左利きの割合に差が無いという帰無仮説を検定する。次いで、3タイプ中のすべての2タイプの組み合わせについて、左利きの割合に差が無いという帰無仮説を検定し、第一種の過誤をホルムの方法で調整してみる。R コンソールでは次の2行を打つだけで済む（多重性の調整をホルム以外の方法でやりたければ、p.adjust.method="bon"などとオプションで指定する）。

```
> prop.test(table(survey$Clap, survey$W.Hnd))
> pairwise.prop.test(table(survey$Clap, survey$W.Hnd))
```

Rcmdr では、「統計量」「分割表」「２元表」を選び、行の変数として Clap **を選び、列の変数として** W.Hnd **を選んで [OK] ボタンをクリックすると、カイ二乗検定の結果が表示される。検定の多重性の調整は Rcmdr ではサポートされていない。**
EZR では、「統計解析」「名義変数の解析」から「分割表の作成と群間の比率の比較（Fisher の正確検定）」を選び、「行の選択」から W.Hnd**、「列の変数」から** Clap **を選ぶ。「仮説検定」としてカイ２乗検定とフィッシャーの正確検定の左のボックスにチェックを入れ、カイ２乗検定の連続性補正の下のラジオボタンは「はい」にし、「↓ 2 組ずつの比較（post-hoc 検定）は比較する群が 1 つの場合のみ実施される」の下の「2 組ずつの比較（Holm の多重比較）」の左のボックスにチェックを入れてから「OK」ボタンをクリックすると、カイ二乗検定だけではなく、フィッシャーの正確検定でも対比較して検定の多重性の調整を Holm の方法で実施した結果が得られる。なお、R コンソールでも** fmsb **パッケージに含まれる** pairwise.fisher.test() **関数を使えば実行できる。**

第10章　２つの量的な変数間の関係

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

MASS パッケージの survey データフレームで、身長と利き手の大きさ（親指の先端と小指の先端の距離）の関係を調べるには、R コンソールでは、require(MASS) として MASS パッケージをメモリに読み込んだ後であれば、

```
plot(Wr.Hnd ~ Height, data=survey)
```

とするだけである。もし男女別にプロットしたければ、pch=as.integer(Sex) というオプションを指定すれば良い。

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

10.1　相関と回帰の違い

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

10.2　相関分析

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

散布図で相関関係があるように見えても、見かけの相関関係 (apparent correlation) であったり[1]、擬似相関 (spurious correlation) であったり[2]することがあるので、注意が必要である。

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

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

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

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

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

が自由度 $n-2$ の t 分布に従うことを利用して求められる。

散布図を描いた `survey` データフレームの身長と利き手の大きさの間でピアソンの相関係数を計算し、その有意性を検定するには、R コンソールでは次の 1 行を打てばよい（スピアマンの順位相

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

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

関について実行したい時は、method=spearman を付ける）。

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

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

```
Pearson's product-moment correlation

data:  survey$Height and survey$Wr.Hnd
t = 10.7923, df = 206, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5063486 0.6813271
sample estimates:
      cor
0.6009909
（中略）
correlation coefficient = 0.601, 95% CI 0.506-0.681, p.value = 8.23e-22
```

EZR の出力は最下行がまとめなので、そこだけ見れば良い。身長と利き手の大きさの関係について求めたピアソンの積率相関係数は、$r = 0.601$（95%信頼区間が $[0.506, 0.681]$）であり[3]、p-value < 2.2e-16（有意確率が 2.2×10^{-16} より小さいという意味）より、「相関が無い」可能性はほとんどゼロなので、有意な相関があるといえる。なお、相関の強さは相関係数の絶対値の大きさによって判定し、伝統的に 0.7 より大きければ「強い相関」、0.4〜0.7 で「中程度の相関」、0.2〜0.4 で「弱い相関」とみなすのが目安なので、この結果は中程度の相関を示すといえる。

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

[3] 95%信頼区間の桁を丸めて示す場合、真の区間を含むようにするために、四捨五入ではなく、下限は切り捨て、上限は切り上げにするのが普通だが、EZR では四捨五入で示されている。散布図をみて明らかに相関がありそうな場合、「相関がない」という帰無仮説の下で当該データが偶然得られた可能性を示す p 値はきわめて小さな値になるのが当然で、むしろ 95%信頼区間を示す方が情報量は多くなる。

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

EZR では、相関係数を求めるメニューの下の方にあるボックスに `Sex=="Male"` **と入力してから [OK] ボタンをクリックすれば男性の身長と利き手の大きさについてピアソンの積率相関係数を求めて有意性の検定をすることができるし、女性についてもボックスにタイプする文字列を、** `Sex=="Male"` **の代わりに** `Sex=="Female"` **とするだけで良い。**

もちろん、相関についてだけ男女別の分析をするなら、データセットの分割をしなくてもいい。R コンソールなら、

```
isMale <- (survey$Sex=="Male")
isFemale <- (survey$Sex == "Female")
cor.test(survey$Height[isMale], survey$Wr.Hnd[isMale])
cor.test(survey$Height[isFemale], survey$Wr.Hnd[isFemale])
```

のようにしてもいいし、第一の引数として与える行列やデータフレームを、第二の引数として与える要因型変数で層別したリストとして返す機能をもつ `split()` 関数を使って、

```
z <- split(survey[, c("Height", "Wr.Hnd")], survey[, "Sex"])
cor.test(z[[1]]$Height, z[[1]]$Wr.Hnd) # Female
cor.test(z[[2]]$Height, z[[2]]$Wr.Hnd) # Male
```

のようにしてもいい。あるいは、

```
cor.testm <- function(MM) {
 k <- NCOL(MM)
 for (i in 1:(k-1)) {
  for (j in (i+1):k) {
   print(cor.test(MM[, i], MM[, j]))
  }
 }
}
lapply(split(survey[, c("Height", "Wr.Hnd")], survey$Sex), cor.testm)
```

のように、行列を引数として、すべての 2 列ずつの組合せについて相関係数を計算する関数 `cor.testm()` を定義してから、`lapply` と `split` を使って 1 行で分析を完了することもできる。このやり方なら、`survey$Sex` のところを、例えば `survey$Smoke` にするだけで喫煙状況の層別に相

関係数を計算することができる。

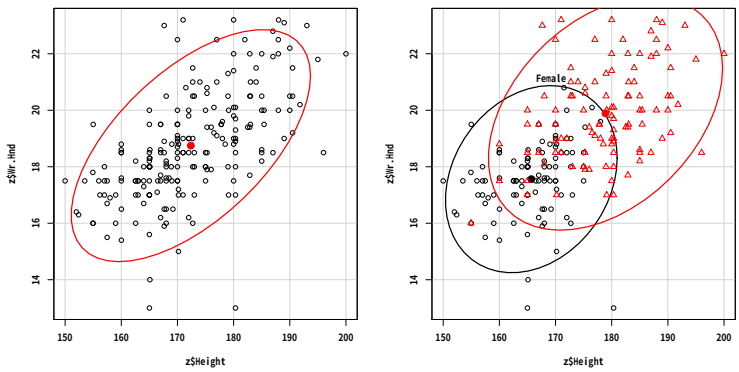
10.2.1 集中楕円と Hotelling の T^2

Rcmdr や EZR のメニューにはまだ入っていないが、相関関係を図示したい場合、回帰直線を引くのはミスリーディングになるので、集中楕円とか確率楕円と呼ばれる楕円を描画するのが普通である。一般に楕円が細長ければ相関関係が強く、真円に近ければ相関関係はないか、あったとしても弱い（真横や完全に上下に細長い＝ばらつきが小さい場合でも、相関関係がないときは、適切にスケーリングすれば散布図は真円に近くなるはずである）。

集中楕円を描く関数は、car という標準パッケージに含まれている `dataEllipse()` である（散布図を描いて楕円を重ね描きしてくれる）。層別に色を変えて描くこともできる。例えば、身長と利き手の大きさであれば、

```
library(car)
z <- survey[, c("Height", "Wr.Hnd", "Sex")]
z <- subset(z, complete.cases(z)) # 欠損値があると dataEllipse() ができない
layout(t(1:2))
dataEllipse(z$Height, z$Wr.Hnd, levels=0.95)
dataEllipse(z$Height, z$Wr.Hnd, z$Sex, levels=0.95)
```

と打てば、身長と利き手の大きさの関係が散布図及び 95%の集中楕円として示される。右の図では 3 番目の引数として性別を与えたために男女層別に楕円が描かれている。



ここで気になってくるのは、「2 つの楕円に差が無いかどうか」である。もちろん、この例では身長も利き手の大きさも、明らかに男性の方が女性よりも値が大きいので、検定などするまでもなく、２つの楕円が異なることは明らかだが、もっと微妙な場合もあるだろう。この問題は、Hotelling の

T^2 検定という方法で分析できる。Excel で実行する方法[4]や SAS で実行する方法[5]もインターネット上に公開されているが、簡単に言えば、1 つの量的変数について 2 群間で平均値の差が無いという帰無仮説を検定する *t* 検定（実際は Welch の方法を使うのが普通）を 2 つ以上の量的変数の比較に拡張しようというアイディアである。数式はミネソタ大学の統計学の資料が見やすい[6]。

集中楕円は 2 つの変数について描くので、「2 つの集中楕円に差が無いか」をみるために Hotelling の T^2 検定を使うなら、比べる量的変数は 2 つで、それらの変数のどちらについても 2 群間で平均値に差が無いというのが帰無仮説になる。身長と利き手の大きさに性差がないという帰無仮説を検定するには、既に上述のコードで集中楕円を描くために z という欠損値を除去したデータフレームができていれば、続けて次のコードを実行すればよい。

```
if (require(Hotelling)==FALSE) {
 install.packages("Hotelling", dep=TRUE)
 library(Hotelling)
}
Z <- split(z[, c("Height", "Wr.Hnd")], z[, "Sex"])
print(hotelling.test(Z[[1]], Z[[2]]))
```

結果は、検定統計量である Hotelling の T^2 が 108.91、分子の自由度が 2（2 変数の比較なため）、分母の自由度が 204（サンプルサイズから比較する変数の個数を引いてさらに 1 を引いた値）で、最後に `P-value:  0` と表示されるので、p 値が 0 に近いきわめて小さい値であり、2 つの集中楕円には統計学的に有意な差があったといえる。

なお、有意差がない場合もついでに示しておく。survey データで男女別に年齢と心拍の関係をプロットし、集中楕円を重ね描きした上で、年齢と心拍について性差がないか Hotelling の T^2 検定をしてみるには、既に MASS、 car、 Hotelling を読み込んだ後であれば、以下のコードを打てば良い。

```
z2 <- survey[, c("Age", "Pulse", "Sex")]
z2 <- subset(z2, complete.cases(z2)) # 欠損値除去
dataEllipse(z2$Age, z2$Pulse, z2$Sex, levels=0.95)
Z2 <- split(z2[, c("Age", "Pulse")], z2[, "Sex"])
print(hotelling.test(Z2[[1]], Z2[[2]]))
```

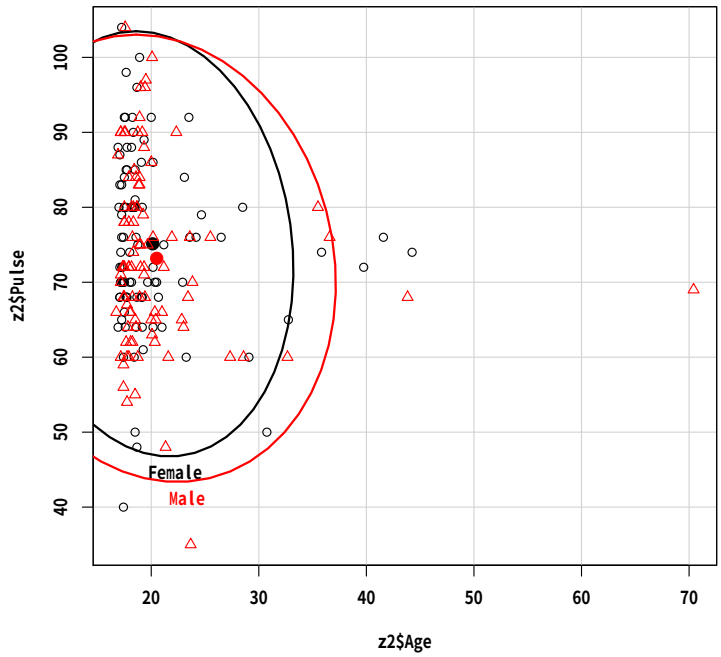
次のグラフが描かれた後に、Test stat: 0.68793、Numerator df: 2、Denominator df: 188、P-value: 0.5039 と表示され、このデータについては男女間に統計学的に有意な差があるとはいえないことがわかる。

---

[4] http://www.real-statistics.com/multivariate-statistics/hotellings-t-square-statistic/hotellings-t-square-independent-samples/

[5] http://sites.stat.psu.edu/~ajw13/stat505/fa06/11_2sampHotel/01_2sampHotel.html

[6] http://users.stat.umn.edu/~gary/classes/5401/handouts/11.hotellingt.handout.pdf



10.2.2 順位相関係数

散布図をみて明らかに外れ値がある場合や、関連が直線的でない場合などは、順位相関係数の適用も検討すべきである。

スピアマンの順位相関係数 ρ は[7]、値を順位で置き換えた（同順位には平均順位を与えた）ピアソンの積率相関係数と同じである。X_i の順位を R_i、Y_i の順位を Q_i とかけば、

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

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

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

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

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

[7] 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。

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

R コンソールで順位相関係数を計算するには、cor.test() 関数のオプションとして、method="spearman" または method="kendall"を指定すれば良い。

EZR では、「統計解析」「ノンパラメトリック検定」「相関係数の検定（Spearman の順位相関係数）」から、解析方法のところで Spearman か Kendall の横のラジオボタンを選んで OK ボタンをクリックすれば計算できる。

ただし、cor.test() 関数や Rcmdr/EZR では、順位相関係数の信頼区間は表示されない。x と y の相関を求めるとき、cor.test(x, y, method="spearman") は、ρ そのものは cor(rank(x), rank(y)) で求めている[8]。けれども、ρ を z 変換したものが近似的に正規分布に従うと仮定しないと cor.test(rank(x), rank(y)) の結果表示される信頼区間が ρ の信頼区間にはならないので、そう言い切ってしまうのは抵抗を感じる[9]。おそらく bootstrap で求めるのが原理的には正しいやり方であり、追加パッケージをインストールすることで実行可能になる。スピアマンの順位相関係数の信頼区間は、RVAideMemoire パッケージの spearman.ci() 関数で求めることができる。bootstrap の回数は nrep=オプションで指定するが、通常は 1000 もあれば良い。ケンドールの順位相関係数 τ の信頼区間は、NSM3 パッケージの kendall.ci() 関数で得られる。bootstrap=FALSE オプションを与えると（それがデフォルト）近似値が得られ、bootstrap するときは、bootstrap=TRUE オプションを付け、B=オプションで回数を指定する。以下に、アメリカ桜のサイズデータである trees における、樹高 Height と体積 Volume の相関を検討する例を示す。bootstrap は数値シミュレーションなので、実行するたびに若干違う結果になるが、近似計算よりも幅が広めに出るようである。

なお、DescTools パッケージの KendallTauB() 関数で conf.level=オプションを付けて使えば、ケンドールの順位相関係数の点推定量とフィッシャーの Z 変換を使って得られる近似的な信頼区間が同時に得られる。欠損値を含むデータにも対応しているし、動作が軽いので、通常はこれで十分だと思う。

[8] getS3method("cor.test", "default") で内部コードを見るとわかる。

[9] その方法がとられているソフトウェアもある。後述するように、SAS の PROC CORR ではバイアスの補正なしの結果がこれに一致する。

```
data(trees)
plot(trees$Height, trees$Volume) # 散布図は、やや下に凸な正の相関にみえる
cor.test(trees$Height, trees$Volume, method="spearman")
cor.test(rank(trees$Height), rank(trees$Volume), method="pearson")
library(RVAideMemoire)
spearman.ci(trees$Height, trees$Volume, nrep=1000)
cor.test(trees$Height, trees$Volume, method="kendall")
library(NSM3)
kendall.ci(trees$Height, trees$Volume, bootstrap=FALSE)
kendall.ci(trees$Height, trees$Volume, bootstrap=TRUE, B=1000)
```

ちなみに、SAS の PROC CORR の方法[10]でスピアマンの順位相関係数と信頼区間を計算する関数は、次のように定義できる。既に fmsb パッケージに入れてあるので、打たなくても使える。

```
spearman.ci.sas <- function(x, y, adj.bias=TRUE, conf.level=0.95) {
 NAMEX <- deparse(substitute(x))
 NAMEY <- deparse(substitute(y))
 xx <- subset(x, !is.na(x)&!is.na(y))
 y <- subset(y, !is.na(x)&!is.na(y))
 x <- xx
 n <- length(x)
 rx <- rank(x)
 ry <- rank(y)
 mx <- mean(rx)
 my <- mean(ry)
 rho <- sum((rx-mx)*(ry-my))/sqrt(sum((rx-mx)^2)*sum((ry-my)^2))
 adj <- ifelse(adj.bias, rho/(2*(n-1)), 0)
 z <- 1/2*log((1+rho)/(1-rho))
 gg <- qnorm(1-(1-conf.level)/2)*sqrt(1/(n-3))
 ge <- z - adj
 gl <- ge - gg
 gu <- ge + gg
 rl <- (exp(2*gl)-1)/(exp(2*gl)+1)
 re <- (exp(2*ge)-1)/(exp(2*ge)+1)
 ru <- (exp(2*gu)-1)/(exp(2*gu)+1)
 cat(sprintf("Spearman's rank correlation between %s and %s\n", NAMEX, NAMEY))
 cat(sprintf("N= %d, rho = %5.3f, %2d%% conf.int = [ %5.3f, %5.3f ]\n",
  n, re, conf.level*100, rl, ru))
 return(list(X=NAMEX, Y=NAMEY, N=n, rho=re, rho.ll=rl, rho.ul=ru, adj.bias=adj.bias))
}
```

[10] http://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/corr_toc.htm

10.3　回帰モデルの当てはめ

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

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

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

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

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

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

これらの a と b の値と、未知の濃度のサンプルについて測定された吸光度（例えば 0.67 としよう）から、そのサンプルの濃度を求めることができる。注意すべきは、サンプルについて測定された吸光度が、標準希釈系列の吸光度の範囲内になければならないことである。回帰モデルが標準希釈系列の範囲外でも直線性を保っている保証は何もないのである[11]。

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

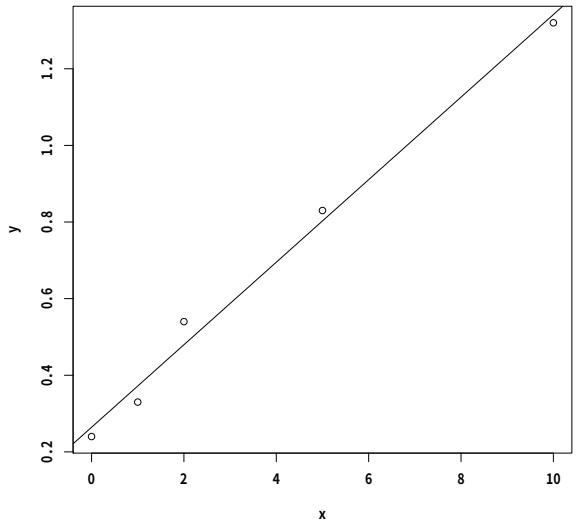
R コンソールでは、lm()（linear model の略で線形モデルの意味）を使って、次のようにデータに当てはめた回帰モデルを得ることができる。

```
y <- c(0.24, 0.33, 0.54, 0.83, 1.32)
x <- c(0, 1, 2, 5, 10)
res <- lm(y ~ x) # 線形回帰モデルを当てはめ
summary(res) # 詳しい結果表示
plot(y ~ x) # 散布図を表示
abline(res) # 回帰直線を重ね描き
AB <- coef(res) # 切片と回帰係数を取り出す
(0.67 - AB[1])/AB[2] # 吸光度 0.67 に対応する濃度を計算
```

上枠内の「詳しい結果」は次のように得られる。

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

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

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

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

推定された切片は $a = 0.26417$、回帰係数は $b = 0.10773$ である。また、このモデルはデータの分散の 98.75% (0.9875) を説明していることが、Adjusted R-squared からわかる。また、p-value は、吸光度の分散がモデルによって説明される程度が誤差分散によって説明される程度と差が無いという帰無仮説の検定の有意確率である。



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

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

検量線以外の状況でも、同じやり方で線形回帰モデルを当てはめることができる。`survey` デー

タフレームに戻ってみよう（もちろん、survey データセットを使う前には、MASS パッケージをロードしておく必要がある）。もし利き手の幅の分散を身長によって説明したいなら、線形回帰モデルを当てはめるには、R コンソールでは次のようにタイプすればいい。

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

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

10.4 推定された係数の安定性を検定する

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

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

が自由度 $(n-2)$ の t 分布に従うことになる。

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

回帰係数についても同様に、

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

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

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

第11章　回帰モデルの応用

11.1　重回帰モデル

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

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

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

EZR では、まず「アクティブデータセット」の下の枠をクリックして survey を選び直してから、「統計解析」「連続変数の解析」「線形回帰（単回帰、重回帰）」を選び、“目的変数” として Wr.Hnd をクリックし、“説明変数” として Height をクリックしてからキーボードの Ctrl キーを押しながら NW.Hnd もクリックし、「モデル解析用に解析結果をアクティブモデルとして残す」の左のチェックボックスにチェックを入れてから [OK] ボタンをクリックすると、結果がアウトプットウィンドウに示される。
アウトプットウィンドウを少し上にスクロールすると、後述する多重共線性の指標値である VIF の値も計算されているので、10 を超えるような値になっていないか確認するべきである。

重回帰モデルでは、個々の説明変数について推定される回帰係数は、他の説明変数の目的変数への影響を調整した上で、その変数独自の目的変数への影響を示す「偏回帰係数」である。しかし偏回帰係数の値は、各変数の絶対的な大きさに依存しているので、各説明変数の目的変数への影響の相対的な強さを示すものにはならない。そうした比較をしたければ、R コンソールで次のようにタイプして stb として得られる「標準化偏回帰係数」が利用できる。結果をみると、Height の標準化偏回帰係数が 0.058、NW.Hnd の標準化偏回帰係数が 0.929 なので、利き手の大きさは大部分、利き手でない手の大きさによって説明されることがわかる。

```
# res <- lm(Wr.Hnd ~ Height + NW.Hnd, data=survey) # の実行後に
csd <- sapply(res$model, sd) # res$model の各変数の sd を計算
print(coef(res)*csd/csd[1]) # 標準化偏回帰係数
```

EZR には、メニューアイテムとしては、この機能は提供されていない。しかし、重回帰モデルを「残して」あるので、そのオブジェクトを使ったコマンドをスクリプトウィンドウに打つ。右上に「モデル：」とあるところの右側に、最後に生成されたモデル名が表示されている。これが RegModel.1 だとすると、上の 3 行の前に、`res <- RegModel.1` という 1 行を追加すれば良い。その 4 行を選んでから、「実行」ボタンをクリックすれば、結果がアウトプットウィンドウに表示される。

なお、標準化偏回帰係数の代わりに、偏相関係数の二乗を使っても、それぞれの説明変数の目的変数への影響の相対的な寄与を示すことができる（Residuals の項を除く、すべての説明変数の偏相関係数の二乗を合計すると、次節で説明する重相関係数の二乗—ただし当然ながら自由度調整する前の値—になる）。最近の論文では、重回帰分析の結果の表としては、各説明変数に対して、偏回帰係数、標準誤差、95%信頼区間の下限と上限、偏相関係数の二乗を示すことが多いように思う。偏相関係数の二乗を求めるには、RegModel.1 に重回帰分析結果が付値されているとすると、下記の 2 行を実行すれば良い。

```
SS <- anova(RegModel.1)["Sum Sq"]
SS/sum(SS)
```

以下の結果が得られる。

```
              Sum Sq
Height    0.36119012
NW.Hnd    0.56830022
Residuals 0.07050966
```

偏相関係数の二乗は、他の変数の影響を調整した上で、それぞれの変数のばらつきが目的変数のばらつきのどれくらいの割合に寄与しているか（説明するか）を意味していると解釈することができるので、標準化偏回帰係数よりも意味がわかりやすいかもしれない。

なお、r 族の効果量として触れた η_p^2 は、ウィスコンシン大学で開発されていた lmSupport パッケージに含まれている modelEffectSizes() 関数で計算させることができたが、既に cran から消滅しているので、現在では、以下のように lsr パッケージの etaSquared() 関数か、heplots パッケージの etasq() 関数を使うと良い（コードと結果をまとめて示す）。

```
> library(lsr)
> etaSquared(res)
            eta.sq eta.sq.part
Height 0.002251836  0.03094818
NW.Hnd 0.568300223  0.88962341
> library(heplots)
> etasq(res, anova=TRUE)
Anova Table (Type II tests)

Response: Wr.Hnd
          Partial eta^2 Sum Sq  Df  F value  Pr(>F)
Height          0.03095   1.69   1    6.547 0.01123 *
NW.Hnd          0.88962 426.25   1 1652.278 < 2e-16 ***
Residuals                52.88 205
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

11.1.1 多重共線性 (multicolinearity)

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 X1 と X2 が相加的に影響していると考えられる場合、lm(Y ~ X1+X2) という重回帰モデルを立てるとしよう。ここで、実は X1 が X2 と強い相関をもっているとすると、もし X1 の標準化偏回帰係数の絶対値が大きければ、X2 による効果もそちらで説明されてしまうので、X2 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 (multicolinearity) 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 パッケージの lm.ridge()）によっても対処可能である。

また、DAAG パッケージ（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[1]。既に述べた通り EZR では重回帰分析を行うと自動的に VIF は計算されている。fmsb パッケージの VIF() 関数を使うこともできるが、vif() 関数とは使い方が違うので注意が必要である。

11.2　当てはまりの良さの評価

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

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

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

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

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

という関係が常に成り立つので、

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

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

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

[1] 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。

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

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

EZR で AIC を求めるには、標準化偏回帰係数を求めたときと同様に、スクリプトウィンドウに必要な関数を打ち、それを選択した上で「実行」ボタンをクリックする。モデルが `RegModel.1` であれば、必要な関数は、`AIC(RegModel.1)` である。

11.3 回帰モデルを当てはめる際の留意点

最小二乗推定の説明から自明なように、回帰式の両辺を入れ替えた回帰直線は一致しない。身長と体重のように、どちらも誤差を含んでいる可能性がある測定値である場合には、一方を説明変数、他方を目的変数とすることは妥当でないかもしれない[2]。どちらを目的変数とみなし、どちらを説明変数とみなすか、因果関係の方向性に基づいて（先行研究や臨床的知見から）きちんと決めることは重要である。

回帰を使って予測をするとき、外挿には注意が必要である。とくに検量線は外挿してはいけない。実際に測った濃度より濃かったり薄かったりするサンプルに対して、同じ関係が成り立つという保証はどこにもない[3]。サンプルを希釈したり濃縮したりして、検量線の範囲内で定量しなくてはいけない。

例題

組み込みデータ `airquality` は、1973 年 5 月 1 日から 9 月 30 日まで 154 日間のニューヨーク市の大気環境データである。含まれている変数は、`Ozone`（ppb[＝ 10 億分の 1] 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 7:00 と 10:00 に測定した平均風速、マイル／時）、`Temp`（華氏での日最高気温）、`Month`（月）、`Day`（日）である。日照の強さを説明変数、オゾン濃度を目的変数として回帰分析せよ。

[2] この場合は、身長によって体重が決まるという方向性が仮定できるので、通常、身長を説明変数にしてもよいことになっている。

[3] 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。

R コンソールでは、次の 4 行で良い。

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

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

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

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

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

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

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

当てはまりを改善するには、説明変数を追加することが有効な場合がある。この例では、Wind あるいは Temp を説明変数に加えて重回帰モデルにすれば、当てはまりが改善する。R コンソール

では、次の 3 行を打てば重回帰モデルの当てはめができる。自由度調整済み重相関係数の二乗が約 60%にまで改善していることがわかる。

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

変数を追加していったときに有意に当てはまりが改善したかどうかは、尤度比検定によって判定できる。この場合なら、以下のコードを実行すれば良い。

```
res1 <- lm(Ozone ~ Solar.R, data=airquality)
res2 <- lm(Ozone ~ Solar.R + Wind, data=airquality)
res3 <- lm(Ozone ~ Solar.R + Wind + Temp, data=airquality)
anova(res1, res2, res3)
```

Ozone を Solar.R だけで説明するモデルに比べ、Wind を加えたモデルの F 値は 89.1 で p 値は 9.5×10^{-16}、Solar.R と Wind で説明するモデルに比べ、Temp を加えたモデルの F 値は 42.5 で p 値は 2.4×10^{-9} と、それぞれ有意に当てはまりが改善していることがわかる。

EZR では、「統計解析」「連続変数の解析」「線形回帰」を選ぶ。"目的変数" として Ozone を選び、"説明変数" として Solar.R をクリックしてからキーボードの Ctrl キーを押しながら Wind と Temp もクリックし、「モデル解析用に解析結果をアクティブモデルとして残す」の左のチェックボックスにチェックを入れてから [OK] ボタンをクリックすると、同じ結果が得られる。

11.3.1 複数のモデルを整形表示する

目的変数が同じで説明変数が異なる複数の回帰分析の結果をまとめて表示したいことがある。

経済学や社会学の論文を読んでいると頻繁に見かける。この表示をするために便利なパッケージが stargazer である。html 形式や LaTeX 形式でも出力させることができるが、プレインテキストでも可能である。

それぞれ、オプションとして type="html"、type="html"、type="text"を付ければ良い。デフォルトは LaTeX 形式である。

この大気中オゾン濃度の回帰分析の場合、以下のコードを実行すると、その下の表を生成するための LaTeX コードが表示される。

```
res1 <- lm(Ozone ~ Solar.R, data=airquality)
res2 <- lm(Ozone ~ Solar.R + Wind, data=airquality)
res3 <- lm(Ozone ~ Solar.R + Wind + Temp, data=airquality)
library(stargazer)
stargazer(res1, res2, res3)
```

表 11.1:

	Dependent variable:		
	Ozone		
	(1)	(2)	(3)
Solar.R	0.127***	0.100***	0.060**
	(0.033)	(0.026)	(0.023)
Wind		−5.402***	−3.334***
		(0.673)	(0.654)
Temp			1.652***
			(0.254)
Constant	18.599***	77.246***	−64.342***
	(6.748)	(9.068)	(23.055)
Observations	111	111	111
R^2	0.121	0.449	0.606
Adjusted R^2	0.113	0.439	0.595
Residual Std. Error	31.335 (df = 109)	24.917 (df = 108)	21.181 (df = 107)
F Statistic	15.053*** (df = 1; 109)	44.092*** (df = 2; 108)	54.834*** (df = 3; 107)

Note: *p<0.1; **p<0.05; ***p<0.01

11.4 共分散分析 (ANACOVA/ANCOVA)

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

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

となる。ここで、**C** はグループを示す２値変数、**X** と **Y** は量的な変数（連続変数）である。**C** の２群間で **Y** の平均値に差があるかどうかを比べたいのだが、**Y** が **X** によって影響を受ける場合（**X** が共変量である場合)、**X** と **Y** の回帰直線の傾き (slope) が **C** の２群間で差がないなら、**X** による影響を調整した **Y** の修正平均（adjusted mean; 調整平均ともいう）を、**C** の２群間で比べる[4]。ただし、２本の回帰直線がともに充分な説明力をもっていて、かつ２本の回帰直線の間で傾きに差がない場合でないと、修正平均の比較には意味がない。そもそも回帰直線の説明力が低ければその変数は共変量として考慮する必要がないし、傾きが違っていれば群分け変数と独立変数の交互作用が従属変数に関して有意に影響しているということなので、２群を層別して別々に解釈する方が良い。

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

(1) 傾きに差がないという帰無仮説の検定 $H_0 : \beta_1 = \beta_2$、$H_1 : \beta_1 \neq \beta_2$ を検定する。群ごとの X と Y の平均と変動と共変動を使って[5]、仮説 H_1 のもとでの残差平方和

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

と仮説 H_0 のもとでの残差平方和

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

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

(2) 傾きに差がないとき、y 切片に差がない帰無仮説の検定 $\beta_1 = \beta_2$ のもとで（即ち、共通の傾き β を、$\beta = (SS_{XY1} + SS_{XY2})/(SS_{X1} + SS_{X2})$ として推定し)、$H_0' : \alpha_1 = \alpha_2$、$H_1' : \alpha_1 \neq \alpha_2$ を検定する。帰無仮説 H_0' のもとで全部のデータを使った残差平方和 $d_3 = SS_Y - (SS_{XY})^2/SS_X$ を計算して、$F = (d_3 - d_2)/(d_2/(N-3))$ が第 1 自由度 1、第 2 自由度 $N-3$ の F 分布に従うことを

[4]修正平均は **C** の各変数についての係数（２群の場合、基準にする変数の係数はゼロ）に、共変量の平均に共変量の係数を掛けたものを加え、さらに切片を加えることによって計算できる。

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

使って検定できる。H_0' が棄却された場合は各群の平均を共通の傾きに代入すれば各群の切片が求められるし、棄却されない場合は 2 群を一緒にして普通の単回帰分析をすることになる。

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

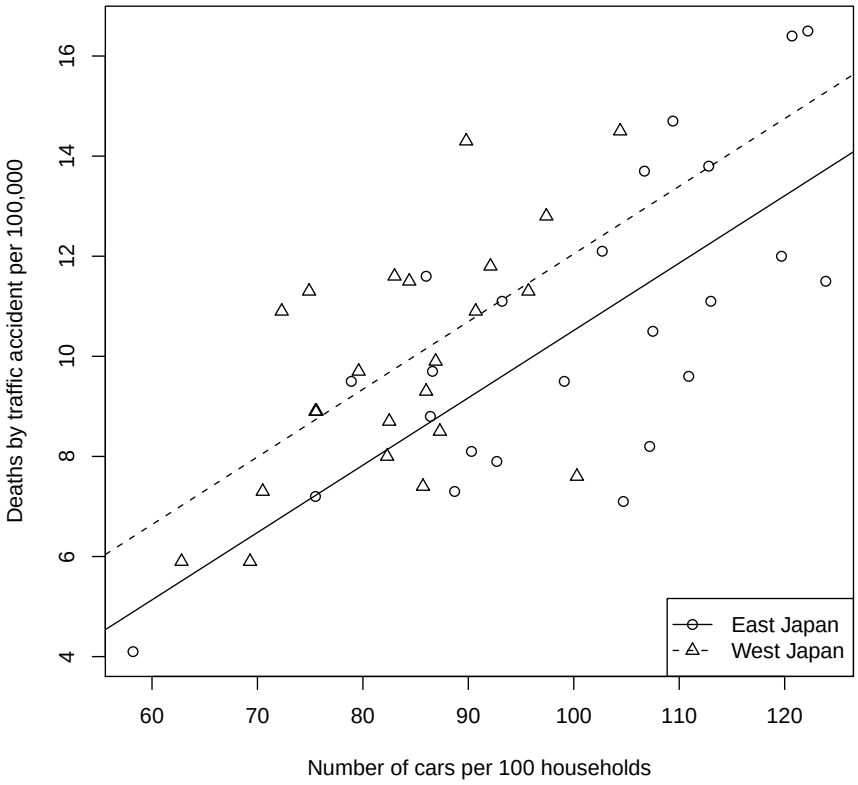
例題

https://minato.sip21c.org/grad/sample3.dat は、都道府県別のタブ区切りテキストデータファイルである。変数としては、都道府県名 (PREF)、日本の東西（REGION）、1990 年の 100 世帯当たり乗用車台数（CAR1990)、1989 年の人口 10 万人当たり交通事故死者数 (TA1989)、1985 年の国勢調査による人口集中地区居住割合 (DIDP1985) が含まれている（REGION の 1 は東日本、2 は西日本を意味する)。このデータについて、東日本と西日本で、100 世帯当たり乗用車台数で調整した人口 10 万人当たり交通事故死者数に差があるか、共分散分析によって検討せよ[a]。

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

R コンソールでの実行例を示す。

```
sample3 <- read.delim("https://minato.sip21c.org/grad/sample3.dat")
plot(TA1989 ~ CAR1990, pch=as.integer(REGION), data=sample3,
 xlab="Number of cars per 100 households",
 ylab="Deaths by traffic accident per 100,000")
east <- subset(sample3,REGION=="East")
regeast <- lm(TA1989 ~ CAR1990, data=east)
summary(regeast)
west <- subset(sample3,REGION=="West")
regwest <- lm(TA1989 ~ CAR1990, data=west)
summary(regwest)
abline(regeast,lty=1)
abline(regwest,lty=2)
legend("bottomright",pch=1:2,lty=1:2,legend=c("East Japan","West Japan"))
summary(lm(TA1989 ~ REGION*CAR1990, data=sample3))
anacova <- lm(TA1989 ~ REGION+CAR1990, data=sample3)
summary(anacova)
cfs <- dummy.coef(anacova)
cfs[[1]] + cfs$CAR1990 * mean(sample3$CAR1990) + cfs$REGION
```



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

```
    East     West
 9.44460 10.96650
```

Rcmdr では、まずデータセット名 sample3 としてインターネットからデータを読み込むため、「データ」「データのインポート」「テキストファイル、クリップボードまたは **URL** から読み込み**...**」を選び、**"**データセット名を入力**"** の右にあるテキストボックスに sample3 と打ち、**"**インターネット **URL"** の隣のラジオボタンをチェックし、**"**フィールド区切り**"** の「タブ」の隣のラジオボタンをチェックし、**[OK]** ボタンをクリックする。表示されるウィンドウで **https://minato.sip21c.org/grad/sample3.dat** と打ち、**[OK]** をクリックすると、インターネットからデータが読み込まれ、sample3 という名前のデータフレームがアクティブになる。
次に REGION で層別した散布図を描く。「**x** 変数」を CAR1990、「**y** 変数」を TA1989 とし、**"**平滑線**"** の隣のチェックを外して **[OK]** し、「グループ別にプロット」をクリックして REGION を選んで **[OK]** し、元のウィンドウでも **[OK]** すると、散布図と **2** 本の回帰直線が描かれ、**2** 本の回帰直線はほぼ平行に見える（丁寧にやるには、ここで東西日本別々に層別して、CAR1990 によって TA1989 が説明されるかをみるため、単回帰分析を行う。東日本のサブセットと西日本のサブセットを作って分析すればよい。CAR1990 の係数は東西どちらでも有意にゼロと異なる。したがって、その影響を調整することに意味はあると思われることが確認できる）。
次に、傾きに差があるかを解析する。「統計量」「モデルへの適合」「線形モデル」でモデル名 LinearModel.3 として左辺の目的変数として TA1989 を、右辺の説明変数群として REGION+CAR1990+REGION:CAR1990 を指定する。結果をみると、REGIONWest:CAR1990 の行に示されている交互作用効果の **p** 値は **0.990** である。この値は、**2** 本の回帰直線の傾きに統計学的な有意差がないことを意味する。
そこで今度は、乗用車所有台数で調整した交通事故死者数の修正平均に差があるかどうかをみるため、交互作用項を除いて回帰を行う。再び「統計量」「モデルへの適合」「線形モデル」で、モデル名を LinearModel.4 とし、左辺はそのまま TA1989 で、右辺の説明変数を CAR1990+REGION に変えて線形モデルの当てはめを実行する。この結果、REGIONWest の行の **p** 値は **0.0319** なので、REGION という変数は、有意水準 **5%** で、CAR1990 の影響を調整しても TA1989 に対して統計学的に有意な影響をもっていることが示された。ただし、修正平均は **R** コンソールと同じコマンドをスクリプトウィンドウに打って選択し、「**Submit**」ボタンをクリックしなくては計算できない。単純な平均値は東日本が **10.5**、西日本が **9.87** であるが、乗用車保有台数の影響を調整した修正平均は、東日本が **9.44**、西日本が **11.0** と逆転し、かつ有意水準 **5%** で統計学的な有意差があるといえた。

EZR には共分散分析のメニューがあって、「統計解析」「連続変数の解析」から「連続変数で補正した **2** 群以上の間の平均値の比較（共分散分析 **ANCOVA)**」を選び、「目的変数（１つ選択）」として TA1989、「**Grouping Variables (pick one)**」として REGION、「補正に用いる連続変数（１つ選択）」として CAR1990 を選び、「モデル解析用に解析結果をアクティブモデルとして残す」の左のボックスにチェックを入れて「**OK**」ボタンをクリックすれば、ほぼ自動的に以上をやってくれる。ただし、修正平均の値だけは自動的には計算されない。保存されたモデルが Model.1 という名称だとすると、次のように **R Script** ウィンドウに打ち、選択した上で「実行」ボタンをクリックする（欠損値があった場合を考えると、元データフレームの CAR1990 ではなく、このように計算に使われたデータを使う方が良い）。

```
cfs <- coef(Model.1)
cfs[[1]] + cfs[[3]]*mean(Model.1$model$CAR1990) + c(0, cfs[[2]])
```

11.5 ロジスティック回帰分析

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

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

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

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

$$\ln(P/(1-P)) = b_0 + b_1 X_1 + ...b_k X_k$$

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

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

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

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

として得られる。

例題

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

low 低体重出生の有無を示す 2 値変数（児の出生時体重 2.5 kg 未満が 1）

age 年齢

lwt 最終月経時体重（ポンド単位。略号 lb. で、1 lb. は 0.454 kg に当たる）

race 人種（１＝白人、２＝黒人、３＝その他）

smoke 喫煙の有無（１＝あり）

ptl 早期産経験数

ht 高血圧の既往（１＝あり）

ui 子宮神経過敏の有無（１＝あり）

ftv 妊娠の最初の３ヶ月の受診回数

bwt 児の出生時体重 (g)

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

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

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

Rcmdr では、「データ」の「パッケージ内のデータ」の「アタッチされたパッケージからデータセットを読み込む」を選んで開くウィンドウで、パッケージの枠から MASS をダブルクリックし、データセットの枠から birthwt をダブルクリックした後に、「データ」「アクティブデータセット内の変数の管理」「数値変数を因子に変換」を選び、まず変数として low を選び、新しい変数名を clow として [OK] ボタンをクリックする。数値 0 が水準 1 となり（NBW と名付ける）、数値 1 が水準 2 となる（LBW と名付ける）。次に race を選び、新変数名を crace として [OK] ボタンをクリックし、出てくるウィンドウで第 1 水準に"white"、第 2 水準に"black"、第 3 水準に"others"とカテゴリ名を指定し、[OK] ボタンをクリックする。smoke、ht、ui についても同様にカテゴリ変数 csmoke、cht、cui に変換する。
EZR では、「ファイル」「パッケージからデータを読み込む」を選び、パッケージとして MASS をダブルクリック、次いでデータセットとして birthwt をダブルクリックして「OK」ボタンをクリックする。次に「アクティブデータセット」の「変数の操作」から「連続変数を因子に変換する」を選ぶ。あとは Rcmdr と同様に操作する。

ロジスティック回帰モデルをこのデータに当てはめるには、R コンソールでは次のようにする。

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

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

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

Rcmdr では、「統計量」「モデルへの適合」「一般化線型モデル」で、式の左辺に clow（因子）をクリックして代入し（たんに clow と入る）、右辺に crace+csmoke+cht+cui+lwt+ptl と打つ（またはクリックして選ぶ）。リンク関数族を binomial にして、リンク関数を logit にして [OK] する。Rcmdr では Nagelkerke の R^2 を求めるオプションはない。
EZR では、「統計解析」「名義変数の解析」「二値変数に対する多変量解析（ロジスティック回帰）」を選ぶ。まず「目的変数」の枠をアクティブにしてから clow [因子] をダブルクリックすると「目的変数」の枠に clow が入る。自動的に「説明変数」の枠がアクティブになるので、そこに crace+csmoke+cht+cui+lwt+ptl と打つ（または変数をダブルクリック、＋の記号ボタンをクリックして選ぶ）。他にもオプションを選べるが、基本的にはこれだけで「OK」ボタンをクリックすればロジスティック回帰分析の結果が得られる。

R コンソールでも Rcmdr でも、表示される結果は次の通りである。

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

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

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

(Dispersion parameter for binomial family taken to be 1)

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

Number of Fisher Scoring iterations: 4
```

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

EZR では、上の出力の後に、各変数についてのオッズ比と 95%信頼区間、p 値も自動的に表示される。Nagelkerke の R^2 は出力に含まれていないので、必要な場合は fmsb パッケージをロードし、`NagelkerkeR2()` 関数にアクティブモデル名を与える。

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

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

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

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

11.6 ポアソン回帰分析

ポアソン回帰分析は、ロジスティック回帰分析では二項分布に従う 2 値変数だった応答変数が、ポアソン分布に従う整数（計数値）である場合に当てはめるモデルである[6]。ロジスティック回帰分析と同じく `glm()` を使い、リンク関数を変えるだけで実行可能である。EZR では Rcmdr のオリジナルメニューである「標準メニュー」の「統計量」の「モデルへの適合」の「一般化線型モデル」を選び、`family` として `poisson` をダブルクリックし、リンク関数として適切なもの（デフォルトは `log`）を選べばよい。

11.6.1 実行例 — Faraway (2006) Chapter 3 より

`faraway` パッケージに含まれている `gala` というデータフレームは、Johnson and Raven (1973) と Weisberg (2005) で提示されているデータで[7]、ガラパゴス諸島の 30 の島のそれぞれの植物の種数[8]（変数名 `Species`）とその島の固有種の数（`Endemics`）に加えて、島の面積（km^2 単位、`Area`）、最高地点の標高（m 単位、`Elevation`）、最寄りの島までの距離（km 単位、`Nearest`）、サンタクルス島までの距離（km 単位、`Scruz`）、隣接する島の面積（km^2 単位、`Adjacent`）という 5 つの地理的変数を含んでいる。

[6] 理論的な詳細は Faraway(2006) の Chapter 3 に詳しい。なお、当然ながら、十分に大きな計数値であって近似的に正規分布が当てはまるとみなして良ければ、通常の線形（重）回帰分析で済む。

[7] `http://www.statsci.org/data/general/galapagos.html` でも公開されている。

[8] Faraway (2006) にはカメの種数と書かれているのだが、2016 年現在公開されている gala データフレームを確認すると、植物の種数と書かれている。Johnson and Raven (1973) のデータは植物とダーウィンフィンチという小鳥の種数しか含んでいないので、カメが何かの間違いと思われる。

Faraway (2006) は、このデータを使って、カメの種の数を、地理的変数でポアソン回帰する例を示している。Faraway (2006) に示されている通り[9]、実は種数の平方根を従属変数にした線形重回帰でも自由度調整済み重相関係数の 2 乗が 0.737 あり、それほど適合は悪くないが、平方根変換の意味づけが難しいし、元々何種かあった（それが切片となるはず）植物がそれぞれの島で独立に種分化したとしたら、種数はポアソン分布に従うだろうと考えるのは合理的な仮定であろう。ポアソン回帰のコードは以下のようになる。

```
if (require(faraway)==FALSE) {
 install.packages("faraway", dep=TRUE)
 library(faraway)
}
data(gala)
gala <- gala[, -2] # delete the number of endemic species column
res <- glm(Species ~., data=gala, family=poisson) # Poisson regression
summary(res)
```

実行すると、次の枠内の結果が得られる。

[9] https://stat.ethz.ch/education/semesters/as2012/asr/Scriptum_Update2012_PoissonRegression.pdf も参照されたい。

```
Call:
glm(formula = Species ~ ., family = poisson, data = gala)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2752  -4.4966  -0.9443   1.9168  10.1849

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68

Number of Fisher Scoring iterations: 5
```

結果を見るとAICが約890と大きく、`Residual deviance`の値から、p値として、`1-pchisq(717, df=24)`を計算させるとほぼゼロとなって当てはまりが良くないことがわかる。通常、こういうときは外れ値の影響を疑うものだが、残差を半正規確率プロット (half-normal plot) させてみると、とくに大きな外れ値はなさそうである。また、このモデルによって説明されるデータのばらつきは、`1-3511/717`を計算すると0.796となり、種数を平方根変換したときの線形重回帰と大差ない。

しかし、もし本当に応答変数である植物の種数がポアソン分布に従っているなら、ポアソン分布は平均と分散が等しいはずなので、このモデルによる予測値（平均値）を横軸に[10]、分散の近似値として残差の二乗を縦軸に[11]とって散布図を描いてみたら確認できるはずである。実行して得られるグラフから、平均と分散には正の相関関係はあるけれども、分散の方が大きな値をとることがわかる。応答変数がポアソン分布に従うというポアソン回帰の仮定が崩れているけれども、リンク関数と独立変数群の選択が間違っていないとしたら、各独立変数の係数は（結果の中で`Estimate`となっているカラム）正しいが、標準誤差が間違った値になる。そうなると、どの独立変数が統計的

[10] リンク関数が対数関数なので、対数をとることに注意。
[11] こちらも対数をとる。

に有意なのかということが、上記の結果からは決定できないことになってしまう。

```
halfnorm(residuals(res))
plot(log(fitted(res)), log((gala$Species-fitted(res))^2),
 xlab=expression(hat(mu)), ylab=expression((y-hat(mu))^2))
```

この過拡散 (overdispersion) という問題を解決するためのアプローチの 1 つとしては、ポアソン過程で発生する応答がそれ自体ガンマ分布に従う確率変数 λ で起こるとしてしまう手がある。λ の期待値がポアソン分布の期待値 μ と同じ、分散が μ/ϕ であると仮定することで、応答変数は負の二項分布に従うようになり、平均はポアソン分布と同じで μ、分散は $\mu(1+\phi)/\phi$ となる。拡散パラメータ ϕ の推定値は次の枠内のようにすれば dp として得られ、それを使って summary() を取り直せば標準誤差が正しく推定できる。ただし、この場合には正規分布を使った検定は信頼性が低いので、F 検定を使うべきである（と Faraway (2006) は書いている）。

```
dp <- sum(residuals(res, type="pearson")^2 / res$df.res)
print(dp)
summary(res, dispersion=dp)
drop1(res, test="F")
```

こうして得られた結果から、島の面積、最高地点の標高、隣接する島の面積の 3 つの変数がガラパゴス諸島の島ごとの植物種数を有意に説明しているといえた。表としては、以下のようにまとめられる。

表. ガラパゴス諸島の島々の地理条件を説明変数群、植物種数を応答変数としたポアソン回帰分析

説明変数	係数	標準誤差	分散比	p 値
切片	3.155	0.292		
その島の面積	−0.00058	0.00015	16.3	0.00048
最高地点の標高	0.0035	0.00049	56.0	1.0×10^{-7}
最寄りの島への距離	0.0088	0.01026	0.76	0.393
サンタクルス島への距離	−0.00571	0.00353	3.24	0.084
最寄りの島の面積	−0.00066	0.00017	20.9	0.00012

AIC: 889.68、D_{null}: 3510.73（自由度 29）、D: 716.85（自由度 24）、拡散パラメータ: 31.7

11.7　多項ロジスティック回帰分析

多項ロジスティック回帰分析は、ロジスティック回帰分析の拡張である。通常のロジスティック回帰分析では応答変数が二項分布に従う 2 値変数だったが、多項ロジスティック回帰分析では多項

分布に従う 3 水準以上のカテゴリ変数である。glm() ではなく、追加パッケージを使うのが普通である。EZR のメニュー上では「標準メニュー」として残っている Rcmdr のオリジナルメニューの、「統計量」の「モデルへの適合」の「多項ロジットモデル」を使って実行可能である（なお、このメニューで使われている関数は、後述する nnet パッケージの multinom() 関数である）。モデルの当てはめ後に変数選択するとか、係数の信頼区間を求めるなどの手続きも、「標準メニュー」の「モデル」から可能である。

Faraway (2006) の Chapter 5 でも nnet パッケージの multinom() 関数を使う方法が解説されているが、ドイツのクリスティアン・アルブレヒト大学キールの心理学教室のサイトにある多項ロジスティック回帰分析のパッケージを比較した記事[12]を読む限りでは、nnet パッケージの multinom() 関数、mlogit パッケージの mlogit() 関数[13]、VGAM パッケージの vglm() 関数の中では、vglm() 関数が多機能であるように思われた。また、RPubs に掲載されている多項ロジスティック回帰分析の記事 (http://rpubs.com/kaz_yos/VGAM) が、VGAM パッケージの vglm() 関数の使い方をわかりやすく説明してくれているので、これを参考にするとよいと思う。また、カリフォルニア大学ロサンゼルス校の記事[14]も参考になる。効果のビジュアル化については、John Fox による https://core.ac.uk/download/pdf/6287961.pdf がわかりやすい。

Faraway (2006) の Chapter 5 の例を挙げておく。同書に含まれている関数やデータは faraway というパッケージをインストールして呼び出せば使えるようになる。Faraway (2006) には同じ 1996 年の米国選挙における支持政党のデータを使って、民主党、どちらでもない、共和党をただのカテゴリと扱う場合と、順序付きカテゴリとして扱う場合のやり方が書かれているが、本稿では前者のみ示す。

なお、欠損値や「わからない」と答えた人や、ビル・クリントンとボブ・ドール以外の支持者はデータから予め除き、結果として 944 人についての 10 個の変数が含まれている。

変数は以下の通りで、元々水準が多い順序付き要因型変数である PID や income をそのまま分析すると結果が解釈しにくいので、PID は 3 水準に再カテゴリ化した変数 sPID を作成し、income は所得の幅のほぼ中央値をとって数値化した変数 nincome を作成して分析する。

[12] http://www.uni-kiel.de/psychologie/rexrepos/posts/regressionMultinom.html
[13] https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf に使い方が説明されている。
[14] https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

popul 回答者の居住地の人口（1000 人単位）。

TVnews 前の週に TV でニュースを見た日数。

selfLR 自分は右翼か左翼か。極端なリベラル（“extLib”）から極端な保守（“extCon”）まで 7 段階の順序付き要因型。

ClinLR ビル・クリントンは右翼か左翼か。スケールは selfLR と同じ。

DoleLR ボブ・ドールは右翼か左翼か。スケールは selfLR と同じ。

PID 支持政党。強く民主党支持（“strDem”）、弱く民主党支持（“weakDem”）、どちらでもないがどちらかといえば民主党支持（“indDem”）、どちらでもない（“indind”）、どちらでもないがどちらかといえば共和党支持（“indRep”）、弱く共和党支持（“weakRep”）、強く共和党支持（“strRep”）という 7 段階の順序付き要因型。

age 回答者の年齢（年）。

educ 回答者の教育歴。7 段階の順序付き要因型。

income 回答者の世帯所得。年収 3000 ドル未満（“$3Kminus”）から 10 万 5000 ドルを超える（“$105Kplus”）までの順序付き要因型。

vote 1996 年大統領選挙での投票予定。“Clinton” と “Dole” の 2 水準の要因型。

多項ロジスティック回帰分析をするコードは以下の通りである。

```
if (require(faraway)==FALSE) {
 install.packages("faraway", dep=TRUE)
 library(faraway)
}
library(nnet)
data(nes96)
sPID <- nes96$PID
levels(sPID) <- c("Democrat","Democrat","Independent","Independent",
 "Independent","Republic","Republic") # reduce levels from 7 to 3
inca <- c(1.5, 0:2*2+4, 9:14+0.5, 0:3*2.5+16, 0:4*5+27.5,
 55, 67.5, 82.5, 97.5, 115)
nincome <- inca[unclass(nes96$income)] # conv ordered factor to numeric
res <- multinom(sPID ~ age + educ + nincome, data=nes96)
summary(res)
```

下枠内の結果が得られる。

```
Call:
multinom(formula = sPID ~ age + educ + nincome, data = nes96)

Coefficients:
            (Intercept)          age     educ.L     educ.Q    educ.C
Independent   -1.197260 0.0001534525 0.06351451 -0.1217038 0.1119542
Republic      -1.642656 0.0081943691 1.19413345 -1.2292869 0.1544575
                 educ^4     educ^5      educ^6    nincome
Independent -0.07657336  0.1360851  0.15427826 0.01623911
Republic    -0.02827297 -0.1221176 -0.03741389 0.01724679

Std. Errors:
            (Intercept)         age    educ.L    educ.Q    educ.C    educ^4
Independent   0.3265951 0.005374592 0.4571884 0.4142859 0.3498491 0.2883031
Republic      0.3312877 0.004902668 0.6502670 0.6041924 0.4866432 0.3605620
               educ^5    educ^6     nincome
Independent 0.2494706 0.2171578 0.003108585
Republic    0.2696036 0.2031859 0.002881745

Residual Deviance: 1968.333
AIC: 2004.333
```

同じデータを VGAM パッケージの vglm() 関数で分析するコードは以下の通り。

```
library(VGAM) # to use vglm()
res <- vglm(sPID ~ age + educ + nincome, data=nes96, family=multinomial)
summary(res)
```

下枠内の結果が得られる。Faraway (2006) には各係数についてカイ二乗検定する方法が書かれているが、vglm() 関数では p 値も自動的に表示される。応答変数のリファレンスカテゴリは 3 番目の水準と書かれているので、これらの係数は共和党支持よりも民主党支持（1）かどちらでもない (2) へのなりやすさを示す。age や nincome は係数が負なので、年齢や所得が高いほど共和党支持になりやすいことを意味する（年齢は 5%水準で有意でないが）。応答変数のリファレンスカテゴリが違うだけで、結果としては nnet パッケージを使った場合と同様である。

```
Call:
vglm(formula = sPID ~ age + educ + nincome, family = multinomial,
    data = nes96)

Pearson residuals:
                      Min      1Q  Median    3Q   Max
log(mu[,1]/mu[,3]) -2.986 -0.8454 -0.3919 1.000 2.267
log(mu[,2]/mu[,3]) -2.318 -0.7826 -0.2850 1.198 2.283

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept):1  1.642650   0.331255   4.959 7.09e-07 ***
(Intercept):2  0.445389   0.372285   1.196   0.2316
age:1         -0.008194   0.004903  -1.671   0.0946 .
age:2         -0.008041   0.005619  -1.431   0.1524
educ.L:1      -1.194114   0.650009  -1.837   0.0662 .
educ.L:2      -1.130593   0.713072  -1.586   0.1128
educ.Q:1       1.229265   0.603936   2.035   0.0418 *
educ.Q:2       1.107555   0.660372   1.677   0.0935 .
educ.C:1      -0.154448   0.486465  -0.317   0.7509
educ.C:2      -0.042485   0.531864  -0.080   0.9363
educ^4:1       0.028269   0.360477   0.078   0.9375
educ^4:2      -0.048311   0.397628  -0.121   0.9033
educ^5:1       0.122118   0.269581   0.453   0.6506
educ^5:2       0.258208   0.296246   0.872   0.3834
educ^6:1       0.037413   0.203183   0.184   0.8539
educ^6:2       0.191694   0.226809   0.845   0.3980
nincome:1     -0.017247   0.002882  -5.985 2.17e-09 ***
nincome:2     -0.001008   0.002917  -0.345   0.7297
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])
Residual deviance: 1968.332 on 1870 degrees of freedom
Log-likelihood: -984.1663 on 1870 degrees of freedom
Number of Fisher scoring iterations: 4
No Hauck-Donner effect found in any of the estimates
Reference group is level  3  of the response
```

次に MASS パッケージに含まれている minn38 というデータを使って説明する。このデータは、1938 年のミネソタ州における高校の卒業生の進路についての集計値である。変数は以下の通り。

hs 高校のランク。“L”、“M”、“U” が、それぞれ下 1/3、中 1/3、上 1/3 を示す。

phs 高校卒業後の状況。“C” は大学進学、“N” は大学以外の進学、“E” はフルタイム雇用、“O” はその他。

fol 父親の職業水準 7 段階。“F1”、“F2”、...、“F7”。

sex 性別。女性 (“F”) か男性 (“M”) かの要因型。

f 頻度。

高校卒業後の進路が性別、高校のランク、父親の職業水準によって説明されるというモデルを当てはめるには、下のコードを打てば良い。

```
library(MASS)
minn38$hs <- as.ordered(minn38$hs)
minn38$fol <- as.ordered(minn38$fol)
res <- multinom(phs ~ hs + fol + sex, weights=f, data=minn38)
summary(res)
```

得られる結果は下枠内の通りである。

```
Call:
multinom(formula = phs ~ hs + fol + sex, data = minn38, weights = f)

Coefficients:
  (Intercept)       hs.L        hs.Q     fol.L      fol.Q     fol.C      fol^4
E  -0.6546827 -0.2624524 -0.04730188 0.8747934 -0.7402795 0.4002617 -0.4295513
N  -1.2038733 -0.4176488 -0.12268126 0.7545555 -0.6906311 0.5286980 -0.3284744
O   0.9958172 -1.1253635 -0.08545117 1.5057952 -0.6788387 0.4125037 -0.5293829
       fol^5      fol^6       sexM
E -0.5471209 0.46682830 -0.9904394
N -0.2194172 0.04308761 -1.3272296
O -0.6055241 0.41878445 -0.2059234

Std. Errors:
  (Intercept)       hs.L       hs.Q      fol.L     fol.Q      fol.C      fol^4
E  0.05077846 0.06746142 0.05795464 0.12210938 0.1127652 0.11361974 0.10893022
N  0.06306977 0.08679242 0.07356078 0.15800767 0.1446081 0.14894973 0.14095890
O  0.03463955 0.04207937 0.03619215 0.07511298 0.0682409 0.07199713 0.07019344
       fol^5      fol^6       sexM
E 0.10451289 0.08968776 0.07255349
N 0.13940822 0.11562495 0.10183668
O 0.06869718 0.05738971 0.04328839

Residual Deviance: 26546.35
AIC: 26606.35
```

同じデータを VGAM パッケージの vglm() 関数で分析するには、下のコードを打つ（上に示した順序付き要因型への変換を済ませた後で実行する）。

```
library(VGAM) # to use vglm()
res <- vglm(phs ~ hs + fol + sex, weights=f, data=minn38,
 family=multinomial)
summary(res)
```

下枠内の結果が得られる。

```
Call:
vglm(formula = phs ~ hs + fol + sex, family = multinomial, data = minn38, weights = f)

Pearson residuals:
                      Min     1Q  Median    3Q   Max
log(mu[,1]/mu[,4]) -16.19 -4.109 -1.3042 1.285 33.33
log(mu[,2]/mu[,4]) -13.44 -2.489 -0.7190 1.651 33.00
log(mu[,3]/mu[,4])  -8.86 -1.615 -0.4686 1.161 33.33

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -0.99583    0.03464 -28.749  < 2e-16 ***
(Intercept):2 -1.65049    0.04489 -36.767  < 2e-16 ***
(Intercept):3 -2.19976    0.05841 -37.659  < 2e-16 ***
hs.L:1         1.12539    0.04208  26.745  < 2e-16 ***
hs.L:2         0.86296    0.06065  14.229  < 2e-16 ***
hs.L:3         0.70800    0.08186   8.649  < 2e-16 ***
hs.Q:1         0.08544    0.03619   2.361 0.018241 *
hs.Q:2         0.03818    0.05237   0.729 0.466038
hs.Q:3        -0.03738    0.06951  -0.538 0.590800
fol.L:1       -1.50581    0.07511 -20.047  < 2e-16 ***
fol.L:2       -0.63088    0.11306  -5.580 2.41e-08 ***
fol.L:3       -0.75116    0.15079  -4.982 6.31e-07 ***
fol.Q:1        0.67889    0.06824   9.949  < 2e-16 ***
fol.Q:2       -0.06142    0.10595  -0.580 0.562109
fol.Q:3       -0.01176    0.13904  -0.085 0.932570
fol.C:1       -0.41241    0.07200  -5.728 1.02e-08 ***
fol.C:2       -0.01218    0.10345  -0.118 0.906266
fol.C:3        0.11633    0.14101   0.825 0.409380
fol^4:1        0.52943    0.07019   7.542 4.61e-14 ***
fol^4:2        0.09981    0.09757   1.023 0.306298
fol^4:3        0.20084    0.13206   1.521 0.128294
fol^5:1        0.60551    0.06870   8.814  < 2e-16 ***
fol^5:2        0.05839    0.09298   0.628 0.530044
fol^5:3        0.38604    0.13067   2.954 0.003133 **
fol^6:1       -0.41882    0.05739  -7.298 2.93e-13 ***
fol^6:2        0.04804    0.08210   0.585 0.558451
fol^6:3       -0.37580    0.10956  -3.430 0.000604 ***
sexM:1         0.20593    0.04329   4.757 1.96e-06 ***
sexM:2        -0.78456    0.06805 -11.530  < 2e-16 ***
sexM:3        -1.12139    0.09880 -11.350  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Names of linear predictors: log(mu[,1]/mu[,4]), log(mu[,2]/mu[,4]), log(mu[,3]/mu[,4])
Residual deviance: 26546.35 on 474 degrees of freedom
Log-likelihood: -13273.17 on 474 degrees of freedom
Number of Fisher scoring iterations: 5
No Hauck-Donner effect found in any of the estimates
Reference group is level  4  of the response
```

結果の最下行に書かれている通り、リファレンスグループが応答変数の第 4 水準ということなの

で、例えば、sexM:1 の係数が 0.20593 と正であることは、進路「その他」に比べて男性は大学進学しやすいということを意味すると解釈される。

順序のある多項ロジスティック回帰分析をしたい場合は、MASS パッケージの polr() 関数を使えば、比例オッズモデルを当てはめることができる。詳細は藤井 (2010) の pp.98-100 を参照されたいが、簡単に使い方だけ紹介しておく。MASS パッケージに含まれている housing というデータフレームを使って説明されている。このデータフレームはデンマーク建造物研究所とデンマークメンタルヘルス研究所によって共同で実施された住宅調査の結果である。調査参加者はコペンハーゲンの 12 の地区の住民である。建物はすべて 1960 年から 1968 年の間に建てられたもので、12 の地区は住民の社会的水準ができるだけばらつかないように選ばれた。各地区から 1〜6 のセクションを選び、各セクションから約 50 人の住民を対象に調査した。含まれている変数は以下の通りである。

Sat 現在の住宅の環境についての満足度（“High”, “Medium”, “Low” の順序付き要因型）。

Infl アパート管理への家主の影響の強さの認知（“High”, “Medium”, “Low” の要因型）。

Type 賃貸居住のタイプ（“Tower”、“Atrium”、“Apartment”、“Teracce” の要因型）。“Tower” は 5 階建て以上のタワーマンションで “Apartment” は 5 階建て未満の集合住宅。

Cont 他の住民との接触の程度（“Low” と “High” の要因型）。

Freq 頻度。各カテゴリに該当する住民の人数

満足度が他の住民との接触の程度によって説明されるというモデルを考える。比例オッズモデルでは、満足度中と満足度高の合計に対する満足度低の比の対数が、他の住民との接触の程度の違いによって説明されるという線型モデルの係数が、満足度高に対する満足度中と満足度低の合計の比の対数が、他の住民との接触の程度の違いによって説明されるという線型モデルの係数と共通であると仮定して、係数を推定する。

```
library(MASS)
res <- polr(Sat ~ Cont, weight=Freq, data=housing)
summary(res)
```

上枠内のコードを実行すると、下枠内の結果が得られる。

```
Call:
polr(formula = Sat ~ Cont, data = housing, weights = Freq)

Coefficients:
          Value Std. Error t value
ContHigh 0.1651    0.09131   1.808

Intercepts:
            Value   Std. Error t value
Low|Medium  -0.5804  0.0735    -7.8980
Medium|High  0.5131  0.0732     7.0119

Residual Deviance: 3645.608
AIC: 3651.608
```

ContHigh の係数が 0.165 で正なので、他の住民との接触が多いと感じている方が住居への満足度が高くなることがわかる（但し統計的に 5%有意ではない、と藤井 (2010) に書かれている。p 値の求め方は書かれていないが、t 値が 1.808 で、サンプルサイズが 1651 なので、自由度 1649 と考えたら p 値は 0.07 程度であろうと思われる）。

同じデータを VGAM パッケージの vglm() 関数で分析するコードは以下の通り。

```
library(MASS) # to load housing data
library(VGAM) # to use vglm function
res <- vglm(Sat ~ Cont, weights=Freq, data=housing, family=propodds)
summary(res)
```

で、実行すると下枠内の結果が得られる。これには p 値も示されていて、確かに 0.07 程度のようである。比例係数の指数をとった値も最後に示されていて、他の住民との接触が多いと満足度が 1.18 倍になると解釈される。

```
Call:
vglm(formula = Sat ~ Cont, family = propodds, data = housing,
    weights = Freq)

Pearson residuals:
                       Min     1Q Median    3Q    Max
logitlink(P[Y>=2]) -12.248 -3.759  1.740 3.451  7.511
logitlink(P[Y>=3])  -8.463 -3.613 -1.689 4.078 10.486

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  0.58042    0.07305   7.946 1.93e-15 ***
(Intercept):2 -0.51307    0.07273  -7.055 1.73e-12 ***
ContHigh       0.16507    0.09114   1.811   0.0701 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: logitlink(P[Y>=2]), logitlink(P[Y>=3])
Residual deviance: 3645.608 on 141 degrees of freedom
Log-likelihood: -1822.804 on 141 degrees of freedom
Number of Fisher scoring iterations: 3
No Hauck-Donner effect found in any of the estimates

Exponentiated coefficients:
ContHigh
1.179476
```

第12章　反復測定データの解析

３つ以上のグループの間で比較をするとき、それらが互いに独立で、位置母数を比較するためなら、一元配置分散分析またはクラスカル＝ウォリスの検定を行うことは、既に示した。

それに対して、同じ対象者について３つ以上の測定値がある場合の、測定値間での比較は？　というのが、この節での主題となる反復測定分散分析またはフリードマンの検定である。手作業でやるのは大変だが、EZR では簡単に実行できる[1]。ただし、注意しなければならない点が２つある。

第一に、データが横長形式になっている必要がある（異なる時点の測定値は異なる変数とする。1 行が 1 人を表す）ということである。

第二に、時間に依存する変数の名前をアルファベット順に与える必要がある（T2 と T10 では T10 が先と判定されるので、T02 などと 0 を入れなくてはいけない）ということである。そうなっていない場合は、「アクティブデータセット」「変数の操作」「変数名を変更する」で名前を変える。

以上 2 点さえ注意すれば手順は簡単である。

12.1　分析の流れ

大雑把に言えば、データ読み込み→グラフを描く→統計解析という流れになる。統計解析の内容は以下３つである。

1. 分散分析表からグループと時点の主効果、それらの交互作用効果を見る
2. 球面性検定（帰無仮説：どの時点でも分散が等しい）
3. もし球面性検定の結果が統計学的に有意なら、G-G または H-F 補正する

12.2　例 1. 8 人の対象者について、さまざまな心理的刺激後の皮膚電位 (mV)

EZR での手順を示す。

[1] https://www.uvm.edu/~dhowell/StatPages/R/RepeatedMeasuresAnovaR.html が参考になる。

1. 「ファイル」「データのインポート」「ファイル、クリップボード、または URL からテキストデータを読み込む」を選んで、データセット名を psycho とし、データファイルの場所をインターネット URL からとし、フィールド区切りをタブとして「OK」ボタンをクリックする

2. URL として、`https://minato.sip21c.org/hypno-psycho01.txt` と入力すると、4 種類の心理刺激に対する、8 人の被験者の皮膚電位ポテンシャルのデータを読み込める。同じ人への異なる刺激に対する 4 種類の測定値の比較なので、通常の一元配置分散分析は不適切である。このデータは既に横長形式で入っている。

3. 生データのグラフを描く。「グラフと表」「反復測定データの折れ線グラフ」を選び、出てくるウィンドウで、データを示す変数として、`calmness`、`despair`、`fear`、`happiness` の４つを選ぶ（複数選ぶ時はキーボードの Cntl を押しながらクリック）。群別はしないので、そのまま「OK」ボタンをクリックすると、同じ人を線で結んだ、心理的刺激に対する皮膚電位ポテンシャルのグラフが表示される。個人差が大きく、どの心理刺激でも高い電位を示しがちな人が 2 人いることがわかる

4. 明らかに正規分布ではないし、個人差も大きく、かつ、個人内変動要因は時間ではないので、反復測定分散分析も不適切である。. ここでの帰無仮説は「皮膚電位ポテンシャルは心理刺激の種類が違っても差が無い」である。これを調べる統計解析は、フリードマンの検定になる。

5. 「統計解析」「ノンパラメトリック検定」「対応のある 3 群以上の比較（Friedman 検定）」と選び、グラフを書いた時と同じ４つの変数を選んで「OK」ボタンをクリックすると（注：ここで 2 組ずつの比較で Bonferroni か Holm のチェックボックスをチェックしておくと、検定の多重性を調整した対比較も自動的にやってくれる）、次の結果が得られる。各心理刺激に対する中央値が示された後に検定結果が表示されている。p 値が 0.09166 と 0.05 より大きいので、有意水準 5%で統計学的に有意ではなく、帰無仮説は棄却されない（ので、対比較はしない）。

```
calmness   despair      fear happiness
    18.20     18.70     22.05     21.15
Friedman chi-squared = 6.45, df = 3, p-value = 0.09166
```

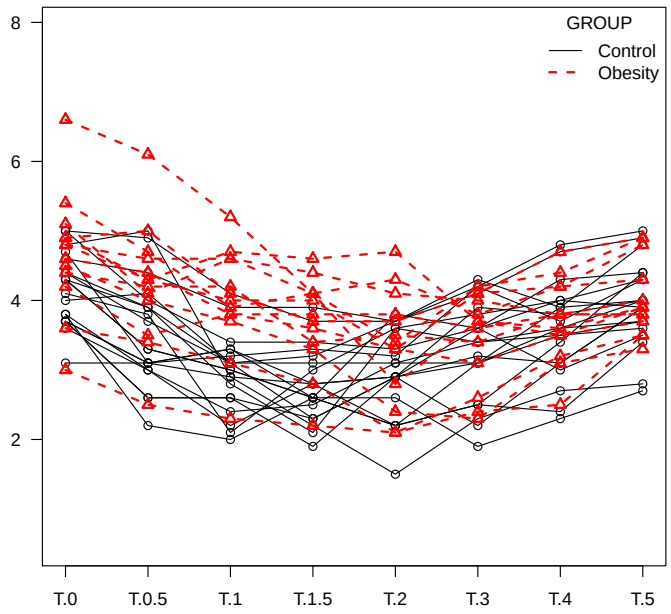
12.3 例２. 33 人について、経口糖負荷試験後血漿無機リン酸塩濃度の変化

データの出典は、B. エヴェリット（著）、石田基宏他（訳）(2007)『R と S-PLUS による多変量解析』シュプリンガー・ジャパンの第９章である。

1. データ読み込み：「ファイル」「データのインポート」「ファイル、クリップボード、または URL からテキストデータを読み込む」を選んで、データセット名を ogtt02 とし、データファイルの場所をインターネット URL からとし、フィールド区切りをタブとして「OK」ボタンをクリックする

2. URL として、`https://minato.sip21c.org/ogtt02.txt` と入力する

3. 生データのグラフを描く。「グラフと表」「反復測定データの折れ線グラフ」を選び、出てくるウィンドウで、データを示す変数として、`T.0`、`T.0.5`、…、`T.5` を選び（連続する複数の変数を選ぶには、まず 1 番上の変数のところでクリックし、Shift キーを押しながら一番下の変数のところでクリックする）、群別変数として `GROUP` を選んでから「OK」ボタンをクリックする



4. `GROUP` は２つのカテゴリからなる因子型の変数で、“Control” が対照、“Obesity” が肥満である。

5. 反復測定分散分析で、時間 (TIME) の効果、群 (GROUP) の効果、それらの交互作用効果を調べるため、「統計解析」「連続変数の解析」「反復測定分散分析」と選んで、データを示す変数として T.0、T.0.5、…、T.5 を選び、群別変数として GROUP を選んで「OK」ボタンをクリックすると以下の結果が得られる。

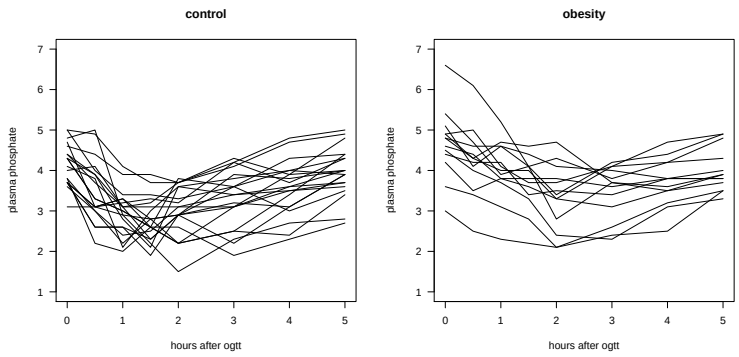
```
Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

                   Sum Sq num Df Error SS den Df   F value       Pr(>F)
(Intercept)        3356.2      1   71.407     31 1457.0243    < 2.2e-16 ***
Factor1.GROUP        15.3      1   71.407     31    6.6592      0.01482 *
Time                 39.2      7   38.098    217   31.9359    < 2.2e-16 ***
Factor1.GROUP:Time    7.7      7   38.098    217    6.2887 0.0000009947 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Mauchly Tests for Sphericity

                   Test statistic        p-value
Time                     0.051567 0.000000098002
Factor1.GROUP:Time       0.051567 0.000000098002


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

                    GG eps Pr(>F[GG])
Time               0.55044  < 2.2e-16 ***
Factor1.GROUP:Time 0.55044  0.0001546 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                       HF eps   Pr(>F[HF])
Time               0.6383131 3.752241e-20
Factor1.GROUP:Time 0.6383131 5.710639e-05
```

そもそも群の効果が p=0.015 で有意だが、球面性検定の結果が時間の効果についても交互作用効果についても有意なので、これらについては G-G または H-F 補正後の数値を見て、どちらも 5% 水準で有意であることがわかる。このデータについても、ノンパラメトリックなフリードマンの検定をすることも可能である。

EZR は内部的にこの処理を car パッケージの Anova() 関数を使って行っている。EZR を使わずに実行するためのコードは以下の通り。

```
ogtt <- read.delim("https://minato.sip21c.org/ogtt02.txt")
times <- c(0, 0.5, 1, 1.5, 2:5)
control <- subset(ogtt, GROUP=="Control")
obesity <- subset(ogtt, GROUP=="Obesity")

layout(t(1:2))
matplot(times, t(control[, 3:10]), type="l", col=1, lty=1, ylim=c(1, 7),
 main="control", xlab="hours after ogtt", ylab="plasma phosphate")
matplot(times, t(obesity[, 3:10]), type="l", col=1, lty=1, ylim=c(1, 7),
 main="obesity", xlab="hours after ogtt", ylab="plasma phosphate")

library(car)
contrasts(ogtt$GROUP) <- "contr.sum" # To calculate type3 SS correctly
rma <- lm(cbind(T.0, T.0.5, T.1, T.1.5, T.2, T.3, T.4, T.5) ~ GROUP,
 data=ogtt)
summary(rma)
inhour <- ordered(times)
idata <- data.frame(inhour)
res.anova <- Anova(rma, idata=idata, idesign=~inhour, type=3)
summary(res.anova, multivariate=FALSE) # Slightly different from EZR
```

エヴェリット (2007) に掲載されているように、対照群と肥満群のグラフを別々に描くと以下のようになる（EZR でもオプション指定で可能である。ただし EZR は散布図でなく折れ線グラフなので間隔が 30 分でも 1 時間でも同じ幅でプロットされている）。



表示される解析結果は、おそらく options() の digits=か scipen=の指定が違うせいで浮動小数点表示方法に違いがあるものの、基本的に同じものである。

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

             Sum Sq num Df Error SS den Df   F value    Pr(>F)
(Intercept)  3356.2      1   71.407     31 1457.0243 < 2.2e-16 ***
GROUP          15.3      1   71.407     31    6.6592   0.01482 *
inhour         39.2      7   38.098    217   31.9359 < 2.2e-16 ***
GROUP:inhour    7.7      7   38.098    217    6.2887 9.947e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Mauchly Tests for Sphericity

             Test statistic    p-value
inhour             0.051567 9.8002e-08
GROUP:inhour       0.051567 9.8002e-08


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

              GG eps Pr(>F[GG])
inhour       0.55044  < 2.2e-16 ***
GROUP:inhour 0.55044  0.0001546 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                HF eps   Pr(>F[HF])
inhour       0.6383131 3.752241e-20
GROUP:inhour 0.6383131 5.710639e-05
```

エヴェリット (2007) は、この解析を car パッケージの Anova() ではなく、混合効果モデルを使って実行するため、nlme パッケージの lme() 関数を用いる方法を紹介している。個人をグループ変数と考えれば、個人差を切片の差、あるいは傾きと切片の両方の違いという形で混合効果モデルに吸収させることが可能である。現在では混合効果モデルならば lmerTest パッケージの lmer() 関数を使うのが便利である。ただし混合効果モデルを当てはめるには、まずデータを縦長形式に変換する必要があることに注意が必要である。縦長形式への変換には reshape() 関数を使うと便利である（ただし、おそらく最近は tidyverse 系の機能を使う方が普通であろう）。コードと結果を以下に示す。

```
 # Reading data from internet
ogtt <- read.delim("https://minato.sip21c.org/ogtt02.txt")
# define times when data were obtained
times <- c(0:3*0.5, 2:5)
ogttlong <- reshape(ogtt, varying=sprintf("T.%g", times), v.names="pp",
 timevar="hours", idvar="ID", direction="long")
ogttlong$hours <- times[ogttlong$hours]

# usual linear regression, with times-squared as an independent variable
resind <- lm(pp ~ hours + I(hours*hours) + GROUP, data=ogttlong)
summary(resind)

# mixed model
library(lmerTest)
reslme0A <- lmer(pp ~ hours + GROUP + (1|ID),
 data=ogttlong, REML=FALSE)
reslme0B <- lmer(pp ~ hours + GROUP + hours:GROUP + (1|ID),
 data=ogttlong, REML=FALSE)
summary(reslme0A)
summary(reslme0B)
anova(reslme0A, reslme0B)

reslme1 <- lmer(pp ~ hours + I(hours*hours) + GROUP + (1|ID),
 data=ogttlong, REML=FALSE)
reslme2 <- lmer(pp ~ hours + I(hours*hours) + GROUP + hours:GROUP + (1|ID),
 data=ogttlong, REML=FALSE)
summary(reslme1)
summary(reslme2)
anova(reslme1, reslme2)

reslme3 <- lmer(pp ~ hours + I(hours*hours) + GROUP + (hours|ID),
 data=ogttlong, REML=FALSE)
reslme4 <- lmer(pp ~ hours + I(hours*hours) + GROUP + hours:GROUP + (hours|ID),
 data=ogttlong, REML=FALSE)
summary(reslme3)
summary(reslme4)
anova(reslme3, reslme4)
```

まず時刻の二乗を独立変数として含む線形重回帰分析の結果は以下の通りである（この結果はエヴェリット (2007) の表 9.8 と一致するはずだが、なぜか微妙に異なる）。個人差を無視しても、時刻、時刻の二乗、グループのすべてが血漿無機リン酸濃度に影響していると言える。なお、グラフから明らかにわかる通り、経時変化が単調でないため、二乗の項を入れないと時刻の効果も有意でなくなる。

```
Call:
lm(formula = pp ~ hours + I(hours * hours) + GROUP, data = ogttlong)

Residuals:
     Min       1Q   Median       3Q      Max
-1.61076 -0.51352  0.02139  0.47504  2.11457

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.99211    0.09921  40.237  < 2e-16 ***
hours            -0.83115    0.09520  -8.730 3.14e-16 ***
I(hours * hours)  0.16361    0.01842   8.881  < 2e-16 ***
GROUPObesity      0.49332    0.08647   5.705 3.16e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6865 on 260 degrees of freedom
Multiple R-squared:  0.3008,Adjusted R-squared:  0.2928
F-statistic: 37.29 on 3 and 260 DF,  p-value: < 2.2e-16
```

混合効果モデルの結果は以下の通りである。まずランダム切片モデルで見てみる。

```
> summary(reslme0A)
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: pp ~ hours + GROUP + (1 | ID)
   Data: ogttlong

     AIC      BIC   logLik deviance df.resid
   562.5    580.4   -276.3    552.5      259

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.33724 -0.68569 -0.08461  0.73489  2.99361

Random effects:
 Groups   Name        Variance Std.Dev.
 ID       (Intercept) 0.2227   0.4719
 Residual             0.3822   0.6183
Number of obs: 264, groups:  ID, 33

Fixed effects:
              Estimate Std. Error        df t value Pr(>|t|)
(Intercept)    3.43885    0.12638  45.80778  27.211   <2e-16 ***
hours         -0.01740    0.02328 231.00000  -0.748   0.4555
GROUPObesity   0.49332    0.18528  33.00000   2.662   0.0119 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) hours
hours       -0.391
GROUPObesty -0.578  0.000
```

個人差を考えても、時刻の二乗を説明変数に入れないと時刻の効果は有意でない。

```
> summary(reslme0B)
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: pp ~ hours + GROUP + hours:GROUP + (1 | ID)
   Data: ogttlong

     AIC      BIC   logLik deviance df.resid
   552.4    573.8   -270.2    540.4      258

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.33492 -0.63852 -0.06215  0.64658  2.70796

Random effects:
 Groups   Name        Variance Std.Dev.
 ID       (Intercept) 0.2252   0.4745
 Residual             0.3627   0.6022
Number of obs: 264, groups:  ID, 33

Fixed effects:
                    Estimate Std. Error        df t value Pr(>|t|)
(Intercept)          3.30165    0.13174  53.72590  25.062  < 2e-16 ***
hours                0.04716    0.02913 231.00002   1.619 0.106744
GROUPObesity         0.84160    0.20989  53.72590   4.010 0.000189 ***
hours:GROUPObesity  -0.16390    0.04640 231.00002  -3.532 0.000498 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) hours  GROUPO
hours       -0.470
GROUPObesty -0.628  0.295
hrs:GROUPOb  0.295 -0.628 -0.470
```

時刻とグループ（対照／肥満）の交互作用を考えると、グループの主効果と交互作用効果は有意なので、グループ間で血漿無機リン酸塩濃度に差があり、経時変化パタンもグループによって異なるといえる。これら２つの結果について尤度比検定を行うと、次に示す通り有意な差があるので、交互作用効果を加えた方が有意にあてはまりが良くなると言える。

```
> anova(reslme0A, reslme0B)
Data: ogttlong
Models:
reslme0A: pp ~ hours + GROUP + (1 | ID)
reslme0B: pp ~ hours + GROUP + hours:GROUP + (1 | ID)
         Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
reslme0A  5 562.51 580.39 -276.26   552.51
reslme0B  6 552.37 573.82 -270.18   540.37 12.149      1  0.0004911 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

次に時刻の二乗を説明変数群に加えてみると、時刻の二乗はもちろんだが、その影響を調整した時刻の効果も有意になる（係数はマイナスなので、時間が経って経口糖負荷の効果が消える部分が二乗の項で説明されると考えれば、この時刻の効果がマイナスなのも納得がいく）。

```
> summary(reslme1)
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: pp ~ hours + I(hours * hours) + GROUP + (1 | ID)
   Data: ogttlong

     AIC      BIC   logLik deviance df.resid
   438.3    459.8   -213.1    426.3      258

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.62326 -0.67286  0.00958  0.54806  2.77540

Random effects:
 Groups   Name        Variance Std.Dev.
 ID       (Intercept) 0.2428   0.4928
 Residual             0.2213   0.4705
Number of obs: 264, groups:  ID, 33

Fixed effects:
                  Estimate Std. Error        df t value Pr(>|t|)
(Intercept)        3.99211    0.12947  50.29035  30.833   <2e-16 ***
hours             -0.83115    0.06524 231.00000 -12.739   <2e-16 ***
I(hours * hours)   0.16361    0.01262 231.00000  12.959   <2e-16 ***
GROUPObesity       0.49332    0.18528  33.00000   2.662   0.0119 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) hours  I(*hr)
hours       -0.396
I(hors*hrs)  0.330 -0.962
GROUPObesty -0.564  0.000  0.000
```

交互作用を加えても各説明変数の血漿無機リン酸塩濃度への効果はそれほど変化しないが、交互作用効果も有意である。

```
> summary(reslme2)
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: pp ~ hours + I(hours * hours) + GROUP + hours:GROUP + (1 | ID)
   Data: ogttlong

     AIC      BIC   logLik deviance df.resid
   418.9    443.9   -202.4    404.9      257

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.58613 -0.61469 -0.01583  0.61708  3.13886

Random effects:
 Groups   Name        Variance Std.Dev.
 ID       (Intercept) 0.2453   0.4952
 Residual             0.2017   0.4492
Number of obs: 264, groups:  ID, 33

Fixed effects:
                    Estimate Std. Error        df t value Pr(>|t|)
(Intercept)          3.85491    0.13159  53.50053  29.294  < 2e-16 ***
hours               -0.76658    0.06376 231.00006 -12.022  < 2e-16 ***
I(hours * hours)     0.16361    0.01205 231.00006  13.574  < 2e-16 ***
GROUPObesity         0.84160    0.19935  44.06226   4.222 0.000119 ***
hours:GROUPObesity  -0.16390    0.03461 231.00006  -4.735 3.81e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) hours  I(*hr) GROUPO
hours       -0.411
I(hors*hrs)  0.310 -0.940
GROUPObesty -0.597  0.079  0.000
hrs:GROUPOb  0.220 -0.214  0.000 -0.369
```

尤度比検定をすると交互作用を加えたモデルの方が有意に当てはまりが改善しているといえる。

```
> anova(reslme1, reslme2)
Data: ogttlong
Models:
reslme1: pp ~ hours + I(hours * hours) + GROUP + (1 | ID)
reslme2: pp ~ hours + I(hours * hours) + GROUP + hours:GROUP + (1 | ID)
        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
reslme1  6 438.29 459.75 -213.15   426.29
reslme2  7 418.89 443.92 -202.45   404.89 21.402      1  3.725e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

最後に切片だけでなく傾きにも個人差があるモデルを当てはめてみる。この結果はエヴェリット (2007) の表 9.7 と一致するはずだが、やはり微妙に合わない。

```
> summary(reslme3)
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: pp ~ hours + I(hours * hours) + GROUP + (hours | ID)
   Data: ogttlong

     AIC      BIC   logLik deviance df.resid
   422.2    450.8   -203.1    406.2      256

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.85820 -0.57655 -0.01046  0.56493  2.73302

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 ID       (Intercept) 0.35528  0.5961
          hours       0.01523  0.1234   -0.56
 Residual             0.17481  0.4181
Number of obs: 264, groups:  ID, 33

Fixed effects:
                  Estimate Std. Error        df t value Pr(>|t|)
(Intercept)        4.01613    0.13881  41.02529  28.933   <2e-16 ***
hours             -0.83115    0.06184 227.61654 -13.441   <2e-16 ***
I(hours * hours)   0.16361    0.01122 197.99322  14.582   <2e-16 ***
GROUPObesity       0.43236    0.18448  32.99979   2.344   0.0253 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) hours  I(*hr)
hours       -0.453
I(hors*hrs)  0.273 -0.902
GROUPObesty -0.524  0.000  0.000
convergence code: 0
Model failed to converge with max|grad| = 0.00202844 (tol = 0.002, component 1)
```

最終行に示されているように、微妙に収束していないが、この程度なら許容可能であろう。この結果も時刻、時刻の二乗、グループのすべてが血漿無機リン酸塩濃度に有意に影響していることを示している。個人ごとの傾きの推定値は –0.56 であり、固定効果より絶対値が小さいが負の関係が示されている。さらに交互作用効果を見たのが次の結果である。エヴェリット (2007) の表 9.9 と一致するはずだが微妙にずれている。

```
> summary(reslme4)
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: pp ~ hours + I(hours * hours) + GROUP + hours:GROUP + (hours |
    ID)
   Data: ogttlong

     AIC      BIC   logLik deviance df.resid
   413.6    445.8   -197.8    395.6      255

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.79396 -0.57037 -0.03002  0.56644  2.88139

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 ID       (Intercept) 0.315351 0.56156
          hours       0.008822 0.09393  -0.48
 Residual             0.174800 0.41809
Number of obs: 264, groups:  ID, 33

Fixed effects:
                    Estimate Std. Error        df t value Pr(>|t|)
(Intercept)          3.85491    0.14194  38.22608  27.159  < 2e-16 ***
hours               -0.76658    0.06296 221.72055 -12.176  < 2e-16 ***
I(hours * hours)     0.16361    0.01122 198.00548  14.582  < 2e-16 ***
GROUPObesity         0.84160    0.21791  32.99104   3.862 0.000497 ***
hours:GROUPObesity  -0.16390    0.04645  32.99590  -3.528 0.001254 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) hours  I(*hr) GROUPO
hours       -0.474
I(hors*hrs)  0.267 -0.886
GROUPObesty -0.605  0.155  0.000
hrs:GROUPOb  0.322 -0.291  0.000 -0.532
```

次に示すように、この場合も、尤度比検定の結果、交互作用効果を含めた方が、有意に当てはまりが良くなっているといえる。

```
> anova(reslme3, reslme4)
Data: ogttlong
Models:
reslme3: pp ~ hours + I(hours * hours) + GROUP + (hours | ID)
reslme4: pp ~ hours + I(hours * hours) + GROUP + hours:GROUP + (hours |
reslme4:     ID)
        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
reslme3  8 422.21 450.82 -203.11   406.21
reslme4  9 413.65 445.83 -197.82   395.65 10.564      1   0.001153 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12.4 例３. 降圧剤投与後の収縮期血圧 (mmHg) の変化

https://minato.sip21c.org/sbp01.txt は、降圧剤投与後の収縮期血圧の変化を示すデータである。有意な経時変化があるかどうかを検討する。

```
subj T.1  T0  T1  T2  T3  T4  T5  T6  T7  T8
1    112 119 113 105 114 110 115 114 110 111
2    116 110 115 110 112 107 116 115 120 118
3    122 123 126 114 111 113 119 123 119 124
4    124 130 127 110 100 127 130 134 120 124
5    126 121 115 122 124 117 124 132 128 120
6    129 135 125 122 115 110 114 124 133 131
```

まず、これを読み込む。EZR では、「ファイル」「データのインポート」「ファイル、クリップボード、または URL からテキストデータを読み込む」を選んで、データセット名を sbp01 とし、データファイルの場所をインターネット URL からとし、フィールド区切りをタブとして「OK」ボタンをクリックすると、URL 入力画面になるので、URL として、https://minato.sip21c.org/sbp01.txt と入力する。
次に、変数名 T.1 を S1 に変更する（EZR では、「アクティブデータセット」「変数の操作」「変数名を変更」で可能)。次に生データの折れ線グラフを描く。「グラフと表」「反復測定データの折れ線グラフ」を選び、出てくるウィンドウで、データを示す変数として、S1, T0, ..., T8 を選んで OK すると 6 本の折れ線グラフが重ね描きされる。
反復測定分散分析は、すべての時点を使うと人数より時点数が多いため解が得られない。敢えてやるなら、時点を絞る必要がある。例えば、「統計解析」「連続変数の解析」「反復測定分散分析」と選び、反復測定データとして、変数 T0, T1, ..., T5 を指定すれば結果が得られる。
ノンパラメトリックなフリードマンの検定は全データを使っても可能である。「統計解析」「ノンパラメトリック検定」「対応のある 3 群以上のデータの比較（Friedman 検定)」と選び、変数として S1, T0, ..., T8 を選んで OK すると、検定結果として p=0.029 が得られる。従って、降圧剤投与後の時間経過に伴い、収縮期血圧は有意水準 5%で統計学的に有意に変化したといえる。

第13章　繰り返し測定または複数の評価者による分割表

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

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

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

13.1　カッパ統計量

検査再検査信頼性を評価するために次の表が得られたとしよう。

	再検査	
	陽性	陰性
検査結果 陽性	a	b
検査結果 陰性	c	d

もし２回の検査結果が完璧に一致していたら $b = c = 0$ となるが、通常は $b \neq 0$ かつ／または $c \neq 0$ である。ここで、２回の検査結果の一致確率は、$P_o = (a + d)/(a + b + c + d)$ と定義できる。

完全な一致の場合、$b = c = 0$ から $P_o = (a + d)/(a + d) = 1$ となる。逆に完全な不一致の場合、$a = d = 0$ から $P_o = 0$ となる。一致の程度が偶然と同じならば、期待される一致確率 P_e は、次の式で計算できる。 $P_e = \{(a + c)(a + b)/(a + b + c + d) + (b + d)(c + d)/(a + b + c + d)\}/(a + b + c + d)$

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

追加パッケージ vcd には、Kappa() という関数があって、κ の点推定量を計算できるし、confint() 関数を適用すれば 95%信頼区間も計算できる。また、筆者も κ を計算する関数 Kappa.test を開発し公開している。この関数は fmsb パッケージに含まれている。

EZR では、分割表のすべての組合せの度数がわかっていれば、「統計解析」「検査の正確度の評価」「２つの定性検査の一致度の評価（Kappa 係数）」から入力することで、点推定量と 95%信頼区間を求めることができるが、Kappa.test() とは異なり一致性の目安は表示されない。

数値計算をしてみよう。次の表を考える。

	再検査	
	陽性	陰性
検査陽性	12	4
検査陰性	2	10

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

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

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

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

ここで“Judgement”（一致度の判定）は、下表の Landis JR, Koch GG (1977) *Biometrics*, 33: 159-174 の基準によっている。

κ の値	一致度の判定
負	"No agreement"（不一致）
0-0.2	"Slignt agreement"（微かな一致）
0.2-0.4	"Fair agreement"（多少の一致）
0.4-0.6	"Moderate agreement"（中程度の一致）
0.6-0.8	"Substantial agreement"（かなりの一致）
0.8-1.0	"Almost perfect agreement"（ほぼ完全な一致）

大雑把なガイドラインに過ぎないが、実用的な基準である。

13.2 マクネマーの検定

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

	介入後	
	あり	なし
介入前 あり	a	b
介入前 なし	c	d

マクネマーの検定では、次のように定義する χ_0^2 を計算する。この χ_0^2 統計量は、帰無仮説（２回

の測定結果の一致の程度が偶然と差が無い）の下で自由度 1 のカイ二乗分布に従う。

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

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

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

拡張マクネマー検定は、M × M クロス集計に適用できるようにしたものである（同じ測定を繰り返したときの変化をみるので、必ず行と列のカテゴリ数は一致する）。セル [i,j] に入る人数を n_{ij}（i, j = 1, 2, ..., M）とすると、次の式で χ_0^2 を計算することができ、この χ_0^2 統計量は帰無仮説（[i,j] に入る確率と [j,i] に入る確率が同じ）の下で自由度 M(M-1)/2 のカイ二乗分布に従う。

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

R コンソールでは、既に関数が提供されていて、対応する 2 変数間の分割表を `TABLE` と書くことにすると、マクネマーの検定は、`mcnemar.test(TABLE)` とするだけでできる。

EZR では、2 × 2 で度数を直接入力する場合は、前述の方法で κ 係数の計算と同時に実行可能である。生データから計算するには、「統計解析」「名義変数の解析」「対応のある比率の比較（二分割表の対称性の検定、McNemar 検定）」を選ぶ。

なお、通常の 2x2 の McNemar の検定の場合は、N12（1 回目はカテゴリ 1 で 2 回目はカテゴリ 2 に変化した人数；以下同様）と N21 の両方がゼロだったら 2 回の間でまったく変化していないということだから、検定する必要がなくなり問題はないが、カテゴリ 3 つ以上の場合は、例えば、N13 はゼロでないが N23 と N32 がゼロ、という状態だと、(N23+N32) が分母になる項で “Division by zero” エラーがでてしまって、カイ二乗値の計算結果自体が NaN になってしまう。この問題への対処としては、その項をスキップするか（カイ二乗検定をするときにスキップした項数だけ自由度を減らせば良いはず）、あるいは全部のセルに極めて小さい数字を足すかすれば計算できるし、実際に"Division by zero"を避けるために全部のセルに 0.01 を足したと書いてある文献が少なくとも 1 つは存在する。

共立出版「R で学ぶデータサイエンス」シリーズの藤井良宜『カテゴリカルデータ解析』には、丁寧に拡張マクネマーの考え方が書かれているけれども分母がゼロの項をどうすべきかについては触れられていないし、「Wonderful R」シリーズの奥村晴彦『R で楽しむ統計』では、拡張マクネマーについては触れられておらず、そもそも連続量をカットオフで二値化したものならば連続量のまま t 検定した方が良いという説明がされている。もちろんその通りなのだけれども、最初から 3 水準以上の名義尺度の変数だと不可能である。カテゴリを併合して 2x2 にできれば問題解決するが、併合しがたい場合もある。

全部のセルに 0.01 を足すのは、検定したい行列オブジェクトを x として、mcnemar.test(x+0.01) で済む。分母がゼロになる組合せをスキップする最低限の関数定義も難しくはないので以下示しておく。使い方は mct(x) とするだけである。

mct()

```
mct <- function(x) {
    L <- NROW(x)
    X <- 0
    N <-0
    for (i in 1:(L-1)) {
    for (j in (i+1):L) {
    s <- (x[i,j]+x[j,i])
    if (s>0) {
    N <- N+1
    X <- X + (x[i,j]-x[j,i])^2/s }}}
    return(list(X2=X, df=N, p=(1-pchisq(X, N))))
}
```

13.2.1 バプカー (Bhapkar) の検定

もう少し深く考えると、拡張マクネマー検定の帰無仮説が、本質的に、周辺度数の均質性（というか変化の対称性）であると考えれば、別のアプローチが可能になる。これについては、さまざまな方法が提案されていて、Sun and Yang (2008) によって SAS のフォーラムで発表された文書[1]によると、SAS ではその 1 つである Bhapkar's test が CATMOD プロシージャの REPEATED ステートメントに実装されていると書かれている。一般論としては Stuart-Maxwell よりも Bhapkar の方が検出力が高いとのことである。

R では irr パッケージに bhapkar() という関数として実装されている。典拠となる論文は Bhapkar (1966)[2]である。bhapkar() の引数は集計後の正方行列ではなく、集計前の 2 つの変数（本質的には順序付きカテゴリ変数であるべきと思われるが、ファクター型でも整数型も構わない）からなるデータフレームか行列でなければならない点に注意が必要だが、これを使えば、対称の位置にある変化の和がゼロである場合が含まれていても、問題なく計算できる。

Sun and Yang (2008) の表 4 に示されているデータを使って分析してみる。元の出典は Walker (2002) の例 18.2 とのことである。ある種の高脂肪食品への渇望を感じる頻度を 3 段階（まったく、時々、頻繁に）で尋ね、2 週間の試験食治療を行う前後で比較した結果である。

[1] http://www2.sas.com/proceedings/forum2008/382-2008.pdf
[2] https://www.jstor.org/stable/2283057

	2 週間の試験食提供後		
試験食提供前	まったく	時々	頻繁に
まったく	14	6	4
時々	9	17	2
頻繁に	6	12	8

これを使って拡張 McNemar 検定や Bhapkar の検定をするコードは以下である。

```
# https://works.bepress.com/zyang/17/download/
# (In Table 4, SAS's gMcNemar gives the results below)
# GMN=6, DF=2, PROBCHI=0.0497871, QCHI95=5.9914645, QCHI99=9.2103404
# (Bhapkar's test in CATMOD give the results below)
# Intercept: DF=2, Chi-Sq=379.45, Pr>Chisq=<.0001
# symp: DF=2, Chi-Sq=6.50, Pr>Chisq=0.0388
# http://www.people.vcu.edu/~dbandyop/BIOS625/GenMcNemar.pdf
# in http://www.people.vcu.edu/~dbandyop/BIOS625.18.html (Lecture 17)

dat <- matrix(c(14, 9, 6, 6, 17, 12, 4, 2, 8), 3, 3)
rownames(dat) <- c("Never", "Occasional", "Frequent")
colnames(dat) <- c("Never", "Occasional", "Frequent")

mcnemar.test(dat)

# Bhapkar's test can be done for raw data.
if (require(irr)==FALSE) {
 install.packages("irr", dep=TRUE)
 library(irr)
}
BEFORE <- rep(1:NCOL(dat), dat[,1])
for (i in 2:NCOL(dat)) {
 BEFORE <- c(BEFORE, rep(1:NCOL(dat), dat[, i]))
}
dx <- data.frame(
 after = factor(rep(1:NROW(dat), colSums(dat)), labels=colnames(dat)),
 before = factor(BEFORE, labels=rownames(dat))
)
bhapkar(dx)
```

結果は下枠内の通りに得られる。

```
> mcnemar.test(dat)

McNemar's Chi-squared test

data:  dat
McNemar's chi-squared = 8.1429, df = 3, p-value = 0.04315

> bhapkar(dx)
 Bhapkar marginal homogeneity

 Subjects = 78
   Raters = 2
    Chisq = 6.5

 Chisq(2) = 6.5
  p-value = 0.0388
```

マクネマー検定の結果が SAS と違うのは、R の mcnemar.test() 関数がデフォルトで連続性の補正をしているためである。SAS と同じ結果を得るには correct=FALSE オプションを付ければ良い。Bhapkar の検定結果は SAS と一致している。

ここで、仮に「頻繁に→まったく」と「まったく→頻繁に」がともにゼロであった場合を考えてみる。コードは以下。

```
# if Both Frequent -> Never and Never -> Frequent were 0
dat[3,1] <- dat[1,3] <- 0
mcnemar.test(dat)
mcnemar.test(dat+0.01)
mct(dat)

BEFORE <- rep(1:NCOL(dat), dat[,1])
for (i in 2:NCOL(dat)) {
 BEFORE <- c(BEFORE, rep(1:NCOL(dat), dat[, i]))
}
dx <- data.frame(
 after = factor(rep(1:NROW(dat), colSums(dat)), labels=colnames(dat)),
 before = factor(BEFORE, labels=rownames(dat))
)
bhapkar(dx)
```

対角成分の和がゼロになる組み合わせが存在するため、マクネマー検定のカイ二乗値は NaN となり、p 値が計算できないが、Bhapkar の検定は普通にできる。なお、すべてのセルに 0.01 を加えてマクネマー検定を実行すると p 値が 0.05 よりわずかに大きくなるが、ゼロになる組み合わせを

スキップする mct() 関数では p 値が 0.02 前後であり、Bhapkar の検定結果に近い。この結果を踏まえれば、全セルに小さな定数を加える方法は妥当でない可能性があり、スキップするか Bhapkar の検定を使うべきと考えられる。

```
> mcnemar.test(dat)

McNemar's Chi-squared test

data:  dat
McNemar's chi-squared = NaN, df = 3, p-value = NA

> mcnemar.test(dat+0.01)

McNemar's Chi-squared test

data:  dat + 0.01
McNemar's chi-squared = 7.7319, df = 3, p-value = 0.05189

> mct(dat)
$X2
[1] 7.742857

$df
[1] 2

$p
[1] 0.02082859

> bhapkar(dx)
 Bhapkar marginal homogeneity

 Subjects = 68
   Raters = 2
    Chisq = 8.74

 Chisq(2) = 8.74
  p-value = 0.0127
```

第14章　検査性能の評価

新しい検査方法を開発する際は、感度 (sensitivity) と特異度 (specificity) が優れていることが必要である。

感度と特異度の計算に必要なデータは、元々カテゴリデータの場合は、その検査で陽性か陰性かと、信頼性が確立した標準的な方法 (gold standard) による確定診断として真にその病気かどうかである。

データさえあれば、感度や特異度、診断正確度 (Diagnostic Accuracy) などの計算は簡単で、下記の定義により、割り算するだけで求められる。また、陽性的中率 (Positive Predictive Value: PPV)、陰性的中率 (Negative Predictive Value: NPV) は、純粋な検査性能の指標ではなく、実際のスクリーニングをしたとき（つまり、真の疾病の有無が未知であるとき）に、陽性あるいは陰性という判定結果がどれくらい当たっているかを意味するため、有病割合に依存して変わる、検査実施の有効性の指標といえる。

	疾病	健康
検査陽性	a 人	b 人
検査陰性	c 人	d 人

感度 $\frac{a}{(a+c)}$ 真にその病気の人を測ったときに検査陽性となる割合

特異度 $\frac{d}{(b+d)}$ その病気ではないとわかっている人（上表では健康）を測ったときに検査陰性となる割合

陽性尤度比 $\frac{(a/(a+c))}{(b/(b+d))}$ 感度を（1 − 特異度）で割った値

陰性尤度比 $\frac{(d/(b+d))}{(c/(a+c))}$ 特異度を（1 − 感度）で割った値

診断正確度 $(a+d)/(a+b+c+d)$ 検査結果が当たっている割合

陽性的中率 $\frac{a}{(a+b)}$ スクリーニングをして陽性だった人のうち、真に病気だった割合

陰性的中率 $\frac{d}{(c+d)}$ スクリーニングをして陰性だった人のうち、真に病気でなかった割合

感度や特異度の点推定は、母比率の推定そのものなので、信頼区間の推定も正規近似や二項分布を使った正確な推定が可能である。手計算では面倒だが、EZR では「定性検査の診断への正確度

の評価」メニューから、分割表に直接入力すれば計算してくれる。epiR パッケージの epi.tests() 関数を使っているので、正規近似でなく Clopper-Pearson 法による正確な 95%信頼区間が自動的に表示される。

数式を示しておくと、サイズ N のうち事象ありの数を X とすると、母比率の点推定値 $\hat{p}$ は $\hat{p} = X/N$ であり、その $(1-\alpha)$ 信頼区間（例えば α が 0.05 であれば 95%信頼区間）は、正規近似では[1]、正規分布の a パーセント点を Z_a と書くことにすれば、

$$\hat{p} - Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \sim \hat{p} + Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

となり、Clopper-Pearson 法による正確な信頼区間は、第 1 自由度 d1、第 2 自由度 d2 の F 分布の a パーセント点を $F_a(d1, d2)$ と書くことにすれば、

$$\frac{X}{X + (N-X+1)F_{1-\alpha/2}(2(N-X+1), 2X)} \sim \frac{(X+1)F_{1-\alpha/2}(2(X+1), 2(N-X))}{(N-X) + (X+1)F_{1-\alpha/2}(2(X+1), 2(N-X))}$$

となる。R で関数定義をすれば以下の通り。

```
ss <- function(X, N, .Exact=FALSE, .CI=0.95) {
 pp <- X/N
 ppl <- ifelse(.Exact, X / (X+(N-X+1)*qf(1-(1-.CI)/2, 2*(N-X+1), 2*X)),
  pp-qnorm(1-(1-.CI)/2)*sqrt(pp*(1-pp)/N) )
 ppu <- ifelse(.Exact, (X+1)*qf(1-(1-.CI)/2, 2*(X+1), 2*(N-X)) /
  ((N-X)+(X+1)*qf(1-(1-.CI)/2, 2*(X+1), 2*(N-X))),
  pp+qnorm(1-(1-.CI)/2)*sqrt(pp*(1-pp)/N) )
 return(sprintf("%4.3f (%4.3f-%4.3f)", pp, ppl, ppu))
}
```

近年、2 つの定性検査の性能を比較する論文[2]がいくつか出ていて、感度、特異度、診断正確度についての比較をマクネマーの検定、陽性的中率と陰性的中率についての比較をフィッシャーの正確な検定でしているものが多い。手順を工夫すれば EZR でもできるが、コードを書いてしまう方が楽であろう[3]。

データが元々連続量の場合に必要なのは、検査値と gold standard で真にその病気かどうかという情報の 2 つである。しかし、通常、連続量である検査値がいくつだったら陽性と判定したら良いのかはわからない。そこで、その閾値を統計学的に根拠のある決め方をするために使える方法が ROC 分析である。

[1]注：初等統計のテキストで示されているこの式は、DescTools パッケージの BinomCI() 関数の method="wald"、あるいは epitools パッケージの binom.approx() 関数に実装されている Wald の方法であり、prop.test() 関数の correct=FALSE で得られる信頼区間は、正規近似とはいえ、epitools パッケージの binom.wilson()、または BinomCI() 関数のデフォルトであり method="wilson"で明示的に指定できる結果とも一致する通り Wilson の方法によっているので結果が異なる

[2]例えば https://doi.org/10.1016/j.jcmg.2019.06.028

[3]https://minato.sip21c.org/t2ct.R として開発中

ROC (Receiver Operating Characteristic) 分析とは、検査で陽性／陰性を判別する閾値を段階的に変え、感度が１に、（１－特異度）が０に最も近い結果を与える値を最適閾値として求めるものである。通常、（１ー特異度）を横軸、感度を縦軸にとって閾値を変えて得られる値を曲線で結ぶ（ROC 曲線と呼ばれる）。

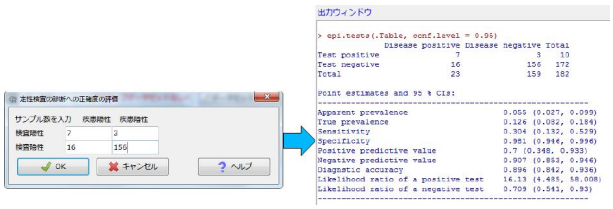
複数の検査方法を ROC 分析で比較することもでき、ROC 曲線の下の面積（AUC; Area Under the Curve）が最も大きい方法が、最も性能が良い方法と判定される。実際にその方法を採用するかどうかは、性能が優れているだけでは不十分で、コストや実施のしやすさなども考慮される

14.1 例 1. 原虫感染強度が低いときのマラリア迅速診断キットの性能評価

マラリアには何種類も RDT（迅速診断キット）がある。元々、マラリア患者の他の熱病患者と区別するために開発されたもので、熱のある患者は血中の原虫感染強度が強いことから、特異度が高いことが重要であり、感度は中程度で良かった。

しかし近年では、原虫感染強度が弱いときの積極的疫学調査（症状がない一般住民を対象とした検査）にも用いられるようになってきた。例えば、ソロモン諸島での三日熱マラリアについて Pan-R malaria を使った検査結果は、以下のように得られたので、感度が不十分であることがわかった。

EZR で実施するには、「統計解析」「検査の正確度の評価」「定性検査の診断への正確度の評価」を選び、表示される分割表に該当する数値を入力すればよい。



14.2 例 2. 診断のために数値の基準値を決定

質問紙に基づいたうつ得点により、うつ病のスクリーニングを行う際に必要なことは、臨床診断でうつ病とわかっている患者と、うつ病でないとわかっている患者（または健康なボランティア）の両方を対象にし、対象者全員について、同じ質問紙によって、尺度得点の合計としての「うつ得点」を得て（下表で 2 行目がうつ得点、3 行目が臨床診断）、ROC 分析を実行することである。

1	2	3	4	5	6	7	8	9	10
20	13	19	21	22	28	11	25	16	19
うつ	非うつ	非うつ	非うつ	うつ	うつ	非うつ	非うつ	非うつ	非うつ

この質問紙得点におけるうつの診断基準が、「18 点以上を陽性と判定する」とすると、診断のためのクロス集計表は以下のようになり、感度は 1 (3/3) で、特異度は約 0.43 (3/7) となることがわかる。

	うつ	非うつ
陽性	3	4
陰性	0	3

この基準値では感度は高いが、特異度が低いことがわかる。基準値を変えることにより、感度と特異度がどちらも高くなるような点を探索することができるというのが ROC 分析の考え方である。

EZR で実行するには、まず「ファイル」「新しいデータセットを作成する（直接入力）」から下図のようにデータを入力する[4]。アルファベット順で先に出現する方が陰性になるため、[Dep] と [Norm] だと [Dep] が陰性扱いになってしまう。そのため、ここでは [1.Dep] と [0.Norm] 入力として [Norm] が陰性として扱われるようにした。

「統計解析」［検査の正確度の評価」「定量検査の診断への正確度の評価（ROC 曲線）」から変数を選んで [OK] をクリック「20 以上がうつ」という基準で最適な感度 (1.0) と特異度 (0.714) が得られた。曲線下面積 (AUC) は 0.8571、95%CI が 0.6044 から 1 まで (DeLong 法) とわかる。

[4]もちろん、Excel などの表計算ソフトで入力し、「ファイル」「インポート」から読み込んでも良い。

14.3 例 3. 複数の方法を ROC 分析で比較

同じモノを評価するための２つの異なった検査の結果は異なりうる。ROC 分析の結果として計算される AUC により、性能を比較することが可能である（AUC が大きいほど高性能といえる）。

EZR では、「ファイル」「データのインポート」「ファイルまたはクリップボード、URL からテキストデータを読み込む」で、データセット名を comptwo、データファイルの場所をインターネットの URL、フィールド区切りをタブにして OK し、URL として `https://minato.sip21c.org/ROC1.txt` を指定する。

注意すべき点は、データセット名として ROC1 と ROC2 は使ってはいけないということである（ファイル名は OK）。もし使うと計算途中でデータセットが上書きされてエラーになる。

「統計解析」「検査の正確度の評価」「２つの ROC 曲線の AUC の比較」を選び、Marker1 と Marker2 を選んで OK ボタンをクリックすると、以下の結果が得られる。5%水準で統計学的に有意な差があるとはいえない。

```
Z = -0.0981, p-value = 0.9218
AUC of roc1 AUC of roc2
 0.8928571   0.9017857
```

第15章　同じ量の2種類の測定結果の一致度の検討

新しく安価あるいは迅速な測定方法を開発したとき、その測定方法が信頼できるかどうかを検討するには、同じ対象者を、従来 gold standard であると考えられてきた方法で測定し、一致しているかをみる。

15.1　検討の方法

３つの方法が可能である。

1. 対応のある *t* 検定→絶対値の大きさとの交互作用を検出できない
2. 相関係数を求め、散布図を描き、x=y の直線も描いて、直線から系統的に点がずれていないかをチェックする→ヒトの認知は斜めの程度を見るのは得意ではない
3. BA プロット（Bland-Altman plot）→ Bland JM と Altman DG の論文が Lancet に発表[1]されて以来必須となった。

BA プロットをするには、「変数の操作」で、新しい変数として、検討したい２つの変数の差の変数 D と、２つの変数の平均値の変数 M を作成し、横軸に M、縦軸に D をとった散布図を描けば良いが、R には BA プロットを実行する機能を関数として実装したパッケージがいくつもある。

15.2　MethComp パッケージを使う

`MethComp` パッケージの `BA.plot()` 関数は便利である。`MethComp` パッケージに含まれているデータ `ox` は、子供 61 人の血中酸素飽和度を血液ガス測定 (CO) とパルスオキシメータ測定 (pulse) で測定した結果である。用例は下記の通り。

[1] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1986; i: 307-10.

```
> library(MethComp)
> data(ox)
> BA.plot(ox)
```

15.3　blandr パッケージを使う

MethComp 以外のパッケージとしては、blandr がお薦めできる。GitHub 上の https://github.com/deepankardatta/blandr/ で開発されており、グラフィクスとして base だけでなく grid 系もサポートされているようである（blandr.draw() 関数では、plotter="rplot"オプションを指定しないと ggplot2 を使った描画になる）。

Bland 自身が管理しているサイト[2]で公開されている、Bland and Altman (1986) にも使われているデータがいくつかあり、このパッケージにはそれらをダウンロードするための関数が含まれている。MethComp と同じく酸素飽和度のデータを読み込む関数は blandr.dataset.o2sats() であり、血中酸素飽和度をより簡便なパルスオキシメータと酸素飽和度計で測定した結果とあるが[3]、MethComp パッケージに入っている ox とは人数が異なる。

Bland-Altman プロットを実行させる関数は、

```
blandr.draw(method1, method2)
```

という形で実行する。method1、method2 は、それぞれの方法で測定した結果の数値ベクトルである。また、2 つの方法の相関係数を計算し、対応のある t 検定の結果を求めると同時に散布図を描き x=y の直線と回帰直線を描く関数は、

```
blandr.method.comparison(method1, method2)
```

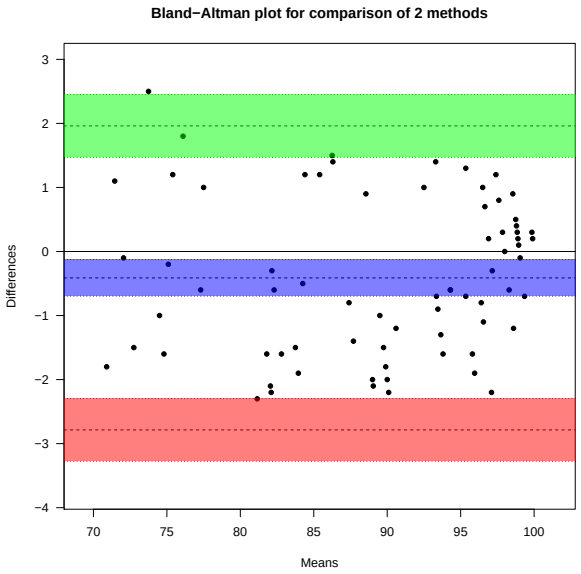
である。また、

```
blandr.output.text(method1, method2)
```

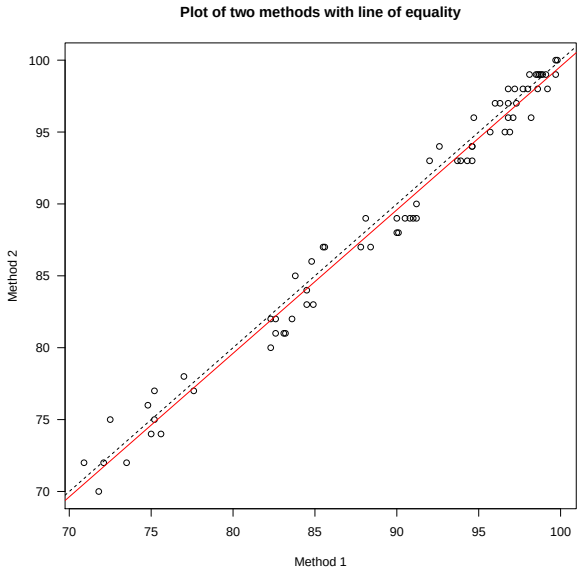
関数を使うと、平均値やバイアスなどの情報を数値として表示させることができる。これらの関数を Bland の酸素飽和度データに適用するコードと結果を示す。

[2] https://www-users.york.ac.uk/~mb55/datasets/datasets.htm
[3] https://www-users.york.ac.uk/~mb55/datasets/sealey.dct

```
if (require(blandr)==FALSE) {
 install.packages("blandr", dep=TRUE)
 library(blandr)
}
o2 <- blandr.dataset.o2sats() # read data via internet
# draw Bland-Altman plot
blandr.draw(o2$pos, o2$osm, plotter="rplot")
# calculate correlation coefficient and paired t-test, draw scattergram
blandr.method.comparison(o2$pos, o2$osm)
# show the values of Bland-Altman plot
blandr.output.text(o2$pos, o2$osm)
```



```
> blandr.method.comparison(ox$pos, ox$osm)

Note as per Bland and Altman 1986 linear correlation DOES NOT mean agreement.
Data which seem to be in poor agreement can produce quite high correlations.
Line of equality in dashed black, linear regression model in solid red.

Paired T-tests evaluate for significant differences between the means of two
sets of data. It does not test agreement, as the results of a T-test can be
hidden by the distribution of differences. See the references for further reading.
Paired T-test p-value:  0.005096854

Correlation coefficients only tell us the linear relationship between 2 variables
and nothing about agreement.
Correlation coefficient: 0.9904082

Linear regression models, are conceptually similar to correlation coefficients,
and again tell us nothing about agreement.
Using method 1 to predict the dependent method 2, using least squares regression.
Regression equation: method 2 =  0.9980026 x method 1 +  -0.2337408
```



```
> blandr.output.text(ox$pos, ox$osm)
Number of comparisons:  72
Maximum value for average measures:  99.9
Minimum value for average measures:  70.9
Maximum value for difference in measures:  2.5
Minimum value for difference in measures:  -2.3

Bias:  -0.4125
Standard deviation of bias:  1.210859

Standard error of bias:  0.1427011
Standard error for limits of agreement:  0.245005

Bias:  -0.4125
Bias- upper 95% CI:  -0.1279621
Bias- lower 95% CI:  -0.6970379

Upper limit of agreement:  1.960784
Upper LOA- upper 95% CI:  2.44931
Upper LOA- lower 95% CI:  1.472258

Lower limit of agreement:  -2.785784
Lower LOA- upper 95% CI:  -2.297258
Lower LOA- lower 95% CI:  -3.27431

Derived measures:
Mean of differences/means:  -0.4711187
Point estimate of bias as proportion of lowest average:  -0.5818054
Point estimate of bias as proportion of highest average -0.4129129
Spread of data between lower and upper LoAs:  4.746567
Bias as proportion of LoA spread:  -8.690491

Bias:
 -0.4125  ( -0.6970379  to  -0.1279621 )
ULoA:
 1.960784  ( 1.472258  to  2.44931 )
LLoA:
 -2.785784  ( -3.27431  to  -2.297258 )
```

なお、BA プロットの詳細な解説としては、Bland JM, Altman DG: Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 1999; 8 (2): 135-160.[4]がわかりやすい。

[4] http://smm.sagepub.com/content/8/2/135.full.pdf+html から全文ダウンロードできる。

第16章　メタアナリシスとシステマティックレビューの方法

非常に難しい。統計学的な考え方の非常に洗練された理解を要する。修得するには、おそらくかなりの努力が必要であろう。

そのため、本テキストでは概要だけ説明するが、メタアナリシスの結果は、多くの研究で共通してみられる知見を明らかにするので、エビデンスベーストメディシン (EBM) で最高レベルのエビデンスを提供すると言われている。

教科書としては、丹後俊郎『メタアナリシス入門：エビデンスの統合を目指す統計手法』朝倉書店、2002 年をお薦めするが、論文を書くには、PRISMA ガイドライン[1]に沿っていることが重要である。2009 年版チェックリストは邦訳[2]もあるが、最新の 2020 年版（チェックリスト[3]、各項目の説明がついた拡張版チェックリスト[4]）を英語で読むのが良いだろう。

16.1　定義

まず、「メタ」は何を意味するか？　について考えてみよう。言語としては、後で発生し、より包括的な何かで、オリジナルのものを批判的に取り扱う、新しいけれども関連した専門分野を名付けるときにしばしば使われるとされる（Egger ら, 1997）。コミュニケーションに対してメタコミュニケーション、文字に対してメタ文字、言語に対してメタ言語、マーケティングに対してメタマーケティング等は好例である。

16.2　概要

統計解析において、メタアナリシスとは、多くの先行研究の結果を統合する手法の１つであり、概ね次のステップを踏む。

[1] http://www.prisma-statement.org/
[2] http://www.prisma-statement.org/documents/PRISMA%20Japanese%20checklist.pdf
[3] http://prisma-statement.org/documents/PRISMA_2020_checklist.pdf
[4] http://prisma-statement.org/documents/PRISMA_2020_expanded_checklist.pdf

1. PubMed や Google Scholar、Web of Science のような文献データベースを使用。適切なキーワードを用い、系統的かつ網羅的に文献検索
2. 適切な基準を決めて、不適切な文献を除外
3. それ以外の文献を精読し、共通して使えるデータを抽出（以上の過程も明示）
4. 抽出したデータを再分析し、対象の違いを超えて共通する知見を見いだす

16.3　歴史

先行研究の結果を統合／総合しようという試み自体は新しくない。Wright 卿 (1896) は、チフスに対する新しいワクチンを開発し、いくつかの異なる集団において同じワクチンの有効性を検査したが、Karl Pearson (1904) は、それまでに使われたワクチンの有効性をレビューして再評価した。これがメタアナリシスの先駆けと言われている。

StudyName	RecovV	DiedV	TotalV	RecovNV	DiedNV	TotalNV
HospitalSA	30	2	32	63	12	75
GarrisonLadysmith	27	8	35	1160	329	1489
SpecialRegimenSA	63	9	72	61	21	82
SpecialHospitalSA	1088	86	1174	4453	538	4991
MilitaryHospitalSA	701	63	764	2864	510	3374
IndianArmy	73	11	84	1052	423	1475

データは `https://minato.sip21c.org/Pearson1.txt` にアップロードしてある。各研究について、四分相関係数を計算する（注：EZR メニューにはないが、`library(psych)` の `tetrachoric()` か、`library(polycor)` の `polychor()` を使用すれば可能)。

```
> tetrachoric(matrix(c(30,2,63,12),2,2))  # 結果は 0.31
> polychor(matrix(c(30,2,63,12),2,2)) # 結果は 0.3069727
```

６つの研究の四分相関係数の平均を計算すると、0.193 となる

```
mean(c(0.307, -0.010, 0.300, 0.119, 0.194, 0.248))
```

Pearson の結論「この効果はワクチンとして推奨するには小さすぎ」

16.4 フィッシャーのZ変換を使い、サンプルサイズで重み付けする

通常、相関係数をまとめるには単純平均ではなく、フィッシャーのZ変換とサンプルサイズによる重み付けをする[5]。この事例は普通の相関係数ではなく四分相関係数なので、Z変換で良いのかは定かでないが、以下方法を示す。

固定効果モデル、即ち相関係数の母数はすべての研究に共通で、研究間の差はランダムエラーであると仮定するモデルでは、

$$Z = 0.5 \ln \frac{(1+r)}{(1-r)}$$

で変換すると、Z 相関係数の分散が $1/(n-3)$ となる。分散の逆数を重みとした Z 相関係数の重み付き平均 M を求め、その標準誤差 SE_m が分散の逆数の和の逆数の平方根として得られることから、95%信頼区間の上限と下限を $M \pm 1.96\mathrm{SE}_m$ として求め、これらをZ変換の逆変換

$$R = \frac{\exp 2Z - 1}{\exp 2Z + 1}$$

を使って元に戻す。

https://minato.sip21c.org/ebhc/pooledr.R

```
pooledr <- function(rs, Ns, pCI=0.975) {
 Zconv <- function(r) { 0.5*log((1+r)/(1-r)) }
 revZ <- function(Z) { (exp(2*Z)-1)/(exp(2*Z)+1) }
 R <- Zconv(rs)
 W <- Ns-3
 M <- sum(R*W)/sum(W)
 SE <- sqrt(1/sum(W))
 LLM <- M-qnorm(pCI)*SE
 ULM <- M+qnorm(pCI)*SE
 R <- revZ(M)
 LLR <- revZ(LLM)
 ULR <- revZ(ULM)
 return(list(R=R, pCI=pCI, CI=c(LLR, ULR)))
}

r <- c(0.307, -0.010, 0.300, 0.119, 0.194, 0.248)
N <- c(32+75, 35+1489, 72+82, 1174+4991, 764+3374, 84+1475)
pooledr(r, N)
```

結果として得られる値は以下の通りであり、相関は0ではないがごく弱いと考えられる。ただ、この事例では四分相関係数がマイナスの研究も含まれており、おそらく固定効果モデルは正しくない。

[5]詳細は、岡田涼・小野寺孝義 (2018)『実践的メタ分析入門：戦略的・包括的理解のために』ナカニシヤ出版、ISBN978-4-7795-1255-1 を参照されたい。逆数を 2 回とる操作で丸め誤差のために元に戻らないなど奇妙な計算をしているが理屈はわかりやすい。

```
> pooledr(r, N)
$R
[1] 0.1462674

$pCI
[1] 0.975

$CI
[1] 0.1297990 0.1626551
```

変量効果モデルでは、研究間の分散にランダムエラーでは説明できない分散を仮定して計算する。理論も計算式もかなり面倒なので省略するが、R では metafor パッケージを使えば計算できると前掲書（岡田・小野寺, 2018）に書かれている。

16.5　オッズ比のメタアナリシス

16.5.1　meta パッケージを使う

それぞれの研究結果は、オッズ比を用いても評価可能である。例えば、(30/2)/(63/12) は 2.86 となる。これは、最初の研究ではワクチン接種の結果生存可能性が 2.86 倍になったことを意味する。なお、fisher.test(matrix(c(30, 2, 63, 12), 2, 2)) の出力するオッズ比は 2.83 だが、これは計算が最尤推定によるためである。

６つの研究結果からマンテル＝ヘンツェルの要約オッズ比を計算するには、meta パッケージを使うのが簡単である。コードは下記の通り（実行するとカレントディレクトリにグラフを含む pdf ファイルができてしまうので注意）。

forestplot.R

```
dat <- read.delim("https://minato.sip21c.org/Pearson1.txt")
library(meta)
print(res <- metabin(RecovV, TotalV, RecovNV, TotalNV,
 studlab=StudyName, data=dat, sm="OR"))
pdf("meta-analysis-vaccine.pdf", width=12, height=8)
forest(res)
drapery(res, type="pval")
dev.off()
```

meta パッケージの metabin() 関数により、以下の結果が得られる。固定効果モデルによるマンテル＝ヘンツェルの要約オッズ比は 1.77（95 %信頼区間は 1.50-2.08）で、I^2 が 26.4%、コクランの Q 検定の結果の p 値が 0.24 と有意でないので、研究間の有意な異質性はなく、固定効果モデル

で解析しても良いことになる。先にフィッシャーの Z 変換を使って四分相関係数をサンプルサイズで重み付けしたときは、固定効果モデルはおそらく正しくないと議論したが、オッズ比で考えると、偶然のばらつきと考えても良いレベルの異質性であったと考えられる。なお、このコードではフォレストプロットの補足的に使って良いとされる Drapery プロットも描画される。各研究の p 値関数が灰色で、要約オッズ比の p 値関数が青（固定効果モデル）または赤（変量効果モデル）で示されている。

```
Number of studies combined: k = 6
Number of observations: o = 13647
Number of events: e = 11635

                         OR           95%-CI    z  p-value
Common effect model  1.7659 [1.4979; 2.0820] 6.77 < 0.0001
Random effects model 1.7854 [1.4258; 2.2356] 5.05 < 0.0001

Quantifying heterogeneity:
 tau^2 = 0.0186 [0.0000; 0.8188]; tau = 0.1362 [0.0000; 0.9049]
 I^2 = 26.4% [0.0%; 69.4%]; H = 1.17 [1.00; 1.81]

Test of heterogeneity:
    Q d.f. p-value
 6.80    5  0.2362

Details on meta-analytical method:
- Mantel-Haenszel method
- Restricted maximum-likelihood estimator for tau^2
- Q-profile method for confidence interval of tau^2 and tau
```

要約統計量としてマンテル＝ヘンツェルの要約オッズ比だけなら、fmsb パッケージの ORMH() 関数でも計算できるが、meta パッケージの metabin() 関数は、sm="OR"による要約オッズ比だけではなく、sm="RR"と指定すれば要約リスク比を、sm="RD"とすれば要約リスク差を計算できる（もちろん元データがコホート研究か RCT でなくては無意味だが）。

プール化の方法も、ORMH() はマンテル＝ヘンツェルの方法だけだが、metabin() では method="Inverse" オプションにより逆分散重み付け法（Fleiss, 1993[6]）、method="Peto"で Peto の方法（Yussuf et al., 1985[7]）、method="SSW"でサンプルサイズ法（Bakbergenuly et al., 2020[8]）が適用される。

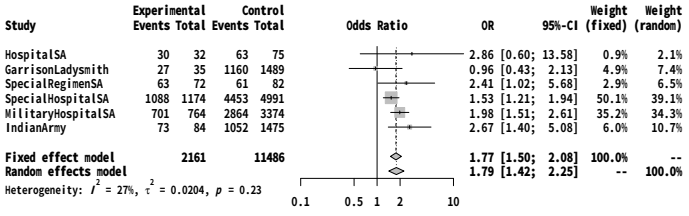
なお、Huedo-Medina T et al. (2006)[9]によると、コクランの Q 統計量を使った Q 検定でも I^2 指標を使っても、研究数が少ない場合の欠点である異質性の検出力の小ささは解決できないとのことで

[6] https://doi.org/10.1177/096228029300200202
[7] https://doi.org/10.1016/s0033-0620(85)80003-7
[8] https://doi.org/10.1002/jrsm.1404
[9] https://psycnet.apa.org/doi/10.1037/1082-989X.11.2.193

ある。

16.5.2　EZR を使う

EZR では「ファイル」「データのインポート」から URL として `https://minato.sip21c.org/Pearson1.txt` を入力してデータを読み込んでおき、「統計解析」「メタアナリシスとメタ回帰」「比率の比較のメタアナリシスとメタ回帰」から「研究の名前を示す変数」として `StudyName`、「テスト群のイベント発生数を示す変数」として `RecovV`、「テスト群の総サンプル数を示す変数」として `TotalV`、「コントロール群のイベント発生数を示す変数」として `RecovNV`、「コントロール群の総サンプル数を示す変数」として `TotalNV` を選び、統合する項目の指定を「オッズ比」にして「OK」をクリックすると、フォレストプロットとともに 6 つの研究結果を統合したオッズ比や異質性指標がグラフィックウィンドウに表示される。固定効果モデルで 1.77、ランダム効果モデルで 1.79 とわかる（ともに有意水準 5%で統計学的に有意に 1 より大きい）。



フォレストプロットは、結果を一覧するのに便利であり、論文投稿にも含めるのが普通である。また、EZR では、「比率の比較のメタアナリシスとメタ回帰」の他、「平均値の比較のメタアナリシスとメタ回帰」と「ハザード比のメタアナリシスとメタ回帰」メニューが提供されている。

第17章　生存時間解析

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

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

生存時間解析を十分説明するには、それだけで1冊の本が必要だが、Rのsurvivalパッケージでは、生存時間解析を実行するための多くの関数が提供されており、かなり高度な解析まで実行できる。

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

17.1　生存時間解析とは

実験においては、化学物質などへの１回の曝露の影響を時間を追ってみていくことが良く行われる。時間ごとに何らかの量の変化を追うほかに、エンドポイントを死亡とした場合、死ぬまでの時

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

間を分析することで毒性の強さを評価することができる。このような期間データを扱う方法としては、一般に生存時間解析 (Survival Analysis または Event History Analysis) と呼ばれるものがある。なかでもよく知られているものが Kaplan-Meier の積・極限推定量である（現在では一般に、カプラン＝マイヤ推定量と呼ばれている）。カプラン＝マイヤ推定量は、イベントが起こった各時点での、イベントが起こる可能性がある人口（リスク集合）あたりのイベント発生数を 1 から引いたものを掛け合わせて得られる、ノンパラメトリックな最尤推定量である。また、複数の期間データ列があったときに、それらの差を検定したい場合は、ログランク検定や一般化ウィルコクソン検定が使われる。

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

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

17.2　カプラン＝マイヤ法

まず、カプラン＝マイヤ推定量についての一般論を示す。イベントが起こる可能性がある状態になってから、イベントが起こった時点を $t_1, t_2, ...$ とし、t_1 時点でのイベント発生数を d_1、t_2 時点でのイベント発生数を d_2、以下同様であるとする。また、時点 $t_1, t_2, ...$ の直前でのリスク集合の大きさを $n_1, n_2, ...$ で示す。リスク集合の大きさとは、その直前でまだイベントが起きていない個体数である。

観察途中で死亡や転居などによって打ち切りが生じるために、リスク集合の大きさはイベント発生によってだけではなく、打ち切りによっても減少する。従って n_i は、時点 t_i より前にイベント発生または打ち切りを起こした個体数を n_1 から除いた残りの数となる（実際は、t_{i-1} から t_i の間に

起こった打ち切り数を $n_{i-1} - d_{i-1}$ から除いたものが n_i となり、それを順次繰り返す)。なお、イベント発生と打ち切りが同時点で起きている場合は、打ち切りをイベント発生直後に起きたと見なして処理するのが慣例である。

このとき、カプラン＝マイヤ推定量 $\hat{S}(t)$ は、

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

として得られる。その標準誤差はグリーンウッドの公式により次の式で分散が得られるので、その平方根となる。

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

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

R コンソールでは、`library(survival)` としてパッケージを呼び出し、生存時間を示す数値型変数を `Time`、1 がイベント発生、0 が打ち切りである（ただし区間打ち切りの場合は 2 とか 3 も使う）整数型の打ち切りフラグを示す変数を `Flag` とすると、

```
dat <- Surv(Time, Flag)
```

のように `Surv()` 関数で生存時間データを作り、

```
res <- survfit(dat ~ 1)
```

でカプラン＝マイヤ法によるメディアン生存時間が得られる。群分け変数 `C` によって群ごとにカプラン＝マイヤ推定をしたい場合は、

```
res <- survfit(dat ~ C)
```

とすればよい。`plot(res)` とすれば階段状の生存曲線が描かれる。イベント発生時点ごとの値を見るには、`summary(res)` とすればよい。

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

例題

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

time 生存時間あるいは観察打ち切りまでの時間（週）

status 打ち切り情報（0 が観察打ち切り、1 がイベント発生）

x 維持化学療法が行われたかどうか（Maintained が維持群、Nonmaintained が非維持群）

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

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

R コンソールでは以下のように実行する。

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

2 行目でカプラン＝マイヤ法での計算がなされ、下枠内が表示される。

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

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

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

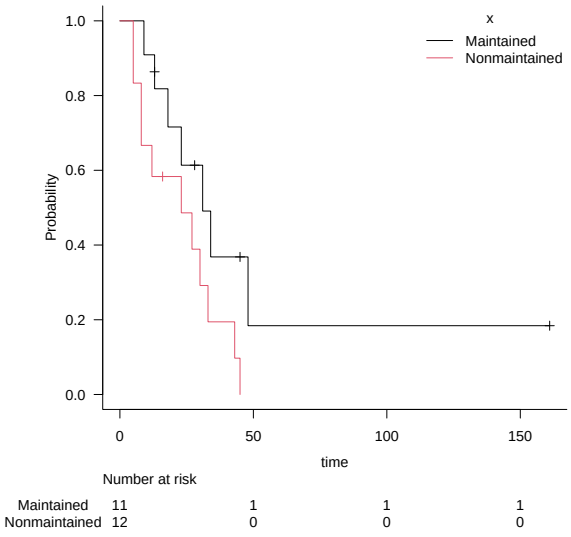
Rcmdr+RcmdrPlugin.survival では、まず survival パッケージ内の leukemia をアクティブにする。「データ」の「アタッチされたパッケージからデータを読み込む」で leukemia の方を選ぶだけで良い。EZR でも、ツールから survival パッケージを読み込み、ファイルの「アタッチされたパッケージからデータを読み込む」で leukemia を選べば良い。

Rcmdr でのカプラン＝マイヤ推定は、「統計量」＞「**Survival analysis**」＞「**Estimate survival function**」と進み、**"Time or start/end times"** として **[time]** を選び、**"Event indicator"** として **[status]** を選ぶ。次に **Strata** だが、leukemia データで **Maintained** 群と **Non-maintained** 群別々にカプラン＝マイヤ推定したいときは、ここに表示されている **[x]** をクリックする。全データを一括して推定したい場合は何もしない[a]。

このメニューは生存曲線も描いてくれる。**"Confidence Intervals"** として **"Log"**、**"Log-log"**、**"plain"**、**"none"** を指定できる。デフォルトは **"Log"** である（つまり、survfit() 関数で、conf.type=オプションを指定しない場合は、summary() で出力される各イベント発生時点の信頼区間は、conf.type="log"として計算される）。大橋・浜田 **(1995)** の **p.68** の出力を見ると、**SAS** バージョン **6** の計算は、conf.type="plain"とした場合と一致していたし、**Statistics in Medicine** に **1997** 年に掲載されていたチュートリアル論文での計算（**https://minato.sip21c.org/swtips/survival.html** を参照のこと）は、conf.type="log-log"とした場合と一致した。**"Plot confidence intervals"** は、常に生存曲線に信頼区間をつけて描画する **[Yes]** と常に信頼区間は描画しない **[No]** の他に、**[Default]** が指定できる。**[Default]** は、**Strata** が指定されている場合は **[No]**、**Strata** が無い場合は **[Yes]** が指定された場合と同じ動作をする（なお、描画しない場合でも、アウトプットウィンドウには、**summary()** の結果として、各イベント発生時点での生存確率の信頼区間の値は表示される）。**"Confidence level"** は信頼水準であり、デフォルトは**.95**、つまり **95** %である。ここを**.99** とすれば **99** %信頼区間が求められる。**"Mark censoring times"** にはデフォルトでチェックが入っており、生存曲線のグラフの打ち切り発生時点で短いティックマークがプロットされる。このチェックを外すと打ち切りレコードは描画されない。**"Method"** は **[Kaplan-Meier]** の他にも **2** つ選べ、**"Variance Method"** も **[Greenwood]** でない方法も選べるが、ここはデフォルトのままでいいと思う。**"Quantiles to estimate"** のボックスに **[.25,.5,.75]** と入っているが、アウトプットウィンドウで**.5** のところに表示される値が生存時間の中央値として推定されるカプラン＝マイヤ推定量である。日本語環境で使う場合に重要なのは、一番下の"部分集合の表現"のボックスに表示されている [＜全ての有効なケース＞] という文字列を消してから **[OK]** ボタンをクリックすることである。おそらく、このプラグインパッケージのバグと思われるが、文字列を消さないで **[OK]** をクリックすると、[＜全ての有効なケース＞] という文字列そのものがオプションとして渡されてしまいエラーを生じる。

[a] 間違って一度クリックしてしまうと、たぶん解除できないので、その場合は諦めてウィンドウを閉じ、やり直すのが無難である。

EZR ではデータをアクティブにした後（注：2022 年 7 月現在、survival パッケージに aml や leukemia というデータフレームは入っているが、なぜか EZR ではデータフレームとして認識されないので、スクリプトウィンドウの 1 行目に library(survival)、2 行目に aml2 <- as.data.frame(aml) と打ってから選んで実行ボタンをクリックすると、aml2 というデータを選べるようになる）、メニューの「統計解析」「生存期間の解析」から「生存曲線の記述と群間の比較（ログランク検定）」を選び、「観察期間の変数」として time、「イベント、打ち切りの変数」として status、「群別する変数」として x を選ぶと、カプラン＝マイヤ推定とログランク検定を両方やってくれて、生存期間の中央値が維持療法群 31ヶ月（95%信頼区間 13ヶ月〜上限無限大）、非維持療法群 23ヶ月（95%信頼区間 5ヶ月〜33ヶ月）、ログランク検定の結果は p 値が 0.0653 で 5%水準では有意な差があるとはいえない。



なお、2022 年 7 月現在、survival パッケージには cancer という大きながん患者の生存時間データが含まれているが、status という打ち切りフラグを示す変数が、打ち切りが 1、死亡が 2 となっているので、EZR で解析するためには、予め変数の操作で新しい打ち切りフラグの変数 flag を、例えば status - 1 あるいは ifelse(status==2, 1, 0) のようにして、イベント発生が 1、打ち切りが 0 であるように変換しなくてはならない。

17.3 ログランク検定

次に、ログランク検定を簡単な例で説明する。

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

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

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

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

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

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

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

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

となる。上の例では、

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

である。これを計算すると約 1.338 となる。

分散は、分散共分散行列の対角成分を考えればいいので、

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

となる。この例の数値を当てはめると、

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

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

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

この例では、R コンソールでは次のようになる。

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

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

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

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

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

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

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

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

Rcmdr+RcmdrPlugin.survival でログランク検定をしたい場合は、leukemia データをアクティブにしてから、統計量＞ Survival analysis ＞ Compare Survival Functions でできる。"Time or start/end times"を [time]、"Event indicator"を [status]、"Strata"を [x] にして（色が反転した選択された状態にして）、"rho"を 0 のまま、"部分集合の表現"ボックスの中を削除して [OK] ボタンをクリックすると、ログランク検定の結果がアウトプットウィンドウに表示される（"rho"を 1 にすると、Peto-Peto 流の一般化ウィルコクソン検定の結果が得られる）。
EZR では、これも「統計解析」「生存期間の解析」から「生存曲線の記述と群間の比較（ログランク検定）」を選べば実行できる。

17.4 コックス回帰

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

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

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

という特徴をもつ。つまり、個体 1 と個体 2 で時点 t のハザードの比をとると、基準ハザード関数 $h_0(t)$ が分母分子からキャンセルされるので、ハザード比は常に、

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

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

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

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

である。累積ハザード関数 $H(t)$ は、

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

となり、式変形すると、

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

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

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

が成り立ち、それを代入すると、後者について

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

となる。両辺の対数を取って符号を逆にして再び両辺の対数を取ると、

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

となる。この式から、共変量で層別して、横軸に生存時間をとり、縦軸に生存関数の対数の符号を逆にしてもう一度対数をとった値をとって散布図を描くと、層間で βz だけ平行移動したグラフが描かれることがわかる。これを**二重対数プロット**と呼ぶ。逆に考えれば、二重対数プロットを描い

てみて、層ごとの散布図が平行になっていなければ、「比例ハザード性」の仮定が満たされないので、コックス回帰は不適切といえる。

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

`leukemia` 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。

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

2 行目で得られる結果は、下枠内の通りである。

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

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

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

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

p 値が 0.074 なので、有意水準 5%で「維持化学療法の有無が生存時間に与えた効果がない」という帰無仮説は棄却されない（ここで、最下行に Score (logrank) test とあるが、これは Rao の Score 検定の結果であり、survdiff() により実行されるログランク検定の結果ではない）。exp(coef) の値 2.5 が、2 群間のハザード比の推定値になるので、維持群に比べて非維持群では 2.5 倍死亡ハザードが高いと考えられるが、95%信頼区間が 1 を挟んでおり、有意水準 5%では統計学的に有意な違いではない。

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

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

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

1. ステージごとに別々に分析する。

2. 他の共変量の影響はステージを通じて共通として、ステージを層別因子として分析する

3. ステージも共変量としてモデルに取り込む

3 番目の仮定ができれば、ステージも共変量としてイベント発生への影響を定量的に評価できるメリットがあるが、そのためには、ステージが違ってもベースラインハザード関数が同じでなければならず、やや非現実的である。また、ステージをどのように共変量としてコード化するかによって結果が変わってくる（通常はダミー変数化することが多い）。

2 番目の仮定は、ステージによってベースラインハザード関数が異なることを意味する。R の `coxph()` 関数で、層によって異なるベースラインハザードを想定したい場合は、`strata()` を使ってモデルを指定する。例えば、がん患者の生存時間データで、生存時間の変数が `time`、打ち切りフラグが `status`、治療方法を示す群分け変数が `x` であるときに、がんの進行度を表す変数 `stage` があったとすると、進行度によってベースラインハザード関数が異なることを想定して、治療方法によって生存時間に差が出るかどうかコックス回帰で調べたければ、`coxph(Surv(time,status)~x+strata(stage), data=leukemia)` とすればよい（但し leukemia データには stage は含まれていないので注意）。

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

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

EZR では、統計解析＞生存時間の解析＞生存時間に対する多変量解析（Cox 比例ハザード回帰）を選んで表示されるウィンドウで、変数（ダブルクリックして式に入れる）の枠にある変数名で、[time]、[status]、[x[因子]] と順にダブルクリックすれば、モデル式：の下の枠の「時間」のところに [time]、「イベント」のところに [status]、「説明変数」のところに [x] が入る。これで [OK] をクリックすればコックス回帰が実行できるが、[OK] をクリックする前に、「比例ハザード性の分析を行う」の左のボックスをチェックしておけば `cox.zph()` による比例ハザード性の仮定が満たされているかどうかの検定と図示もやってくれるし[a]、「ベースラインの生存曲線を示す」の左のボックスをチェックしておけばベースラインの生存曲線と共変量の影響を調整した 95%信頼区間が描かれる。

[a] 検定の帰無仮説は比例ハザード性の仮定が満たされているということである。変数ごとのシェーンフィールド残差プロットの上にスプライン平滑化された曲線と標準誤差の 2 倍の信頼区間が示されたグラフの見方は、明らかな時間傾向がなければ比例ハザード性の仮定には問題が無いとみなす。このグラフの方が、二重対数プロットで平行かどうかを見るよりも見やすいと思う。詳細説明は `https://www1.doshisha.ac.jp/~mjin/R/Chap_37/37.html` を参照されると良い。

参考

https://minato.sip21c.org/swtips/surv.txt は、Bull and Spiegelhalter (1997)[a]で生存時間解析の練習用に公開されているデータ（先天性心疾患である肺動脈閉鎖症の患者 218 人のデータから 30 人を抜き出したもの）を R で読み込みやすいように、欠損値を NA に変えて（元々は –1）タブ区切りテキスト形式で保存したものである。
大雑把な日本語解説と R コードを https://minato.sip21c.org/swtips/survival.html で公開しているので、そこに掲載されている変数の説明や R による解析方法を参照しながら試行錯誤すれば、生存時間解析の方法が身につくであろう。
変数名とその説明だけ採録しておく。

変数名	説明
agepres	来院時日齢
agelast	生存していれば最終来院時日齢、死亡した場合は死亡時日齢
ageopl	最初に手術した日齢（未手術なら欠損）
dead	生存なら 0、死亡していたら 1
sex	男性が 0、女性が 1
paanat	来院時心膜内肺動脈が存在しないか小さければ 0、正常かほぼ正常なら 1
adfol	適切な追跡かどうか。来院から 1 年以内に追跡終了したら 0、1 年以上追跡していたら 1
followup	追跡日数（agelast – agepres）
opfpres	来院から最初の手術までの間隔（日）（未手術なら欠損）
unopage	最初の手術までの追跡日数。手術した人は ageopl−agepres、未手術なら lastage−agepres
unopfpre	来院から最初の手術または最終来院時までの間隔（手術を受けた人には opfpres と同じ）
preopded	手術前に死亡していたら 1、生存していたら 0
hadop	手術を受けていれば 1、未手術なら 0
dedlyrpp	来院から 1 年以内に死亡していたら 1、生存していたら 0、適用できなければ 2（adfol が 0 の人）
agepresx	来院時日齢のカテゴリ：365 日未満なら 0、365 日以上なら 1

[a] https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<1041::AID-SIM506>3.0.CO;2-F

参考 2

Kartsonaki C (2016) Mini-Symposium: Medical Statistics: Survival analysis. *Diagnostic Histopathology*, 22(7): 263-270. https://doi.org/10.1016/j.mpdhp.2016.06.005 も、コンパクトにまとめられた生存時間解析の説明になっている。
この説明では、survival パッケージに入っている colon という 929 人の大腸がん患者のデータが使われているので、EZR でも同じデータを使った解析ができる。

第18章　課題（解答は敢えて提示しない）

https://minato.sip21c.org/grad/worldfactbook2011.txt は、米国 CIA の web サイトで zip 形式に圧縮して公開されている「The world factbook 2011」[1]から作ったタブ区切りテキスト形式のデータである。

このデータに含まれている変数は、次の通りである。

COUNTRY 国名

IMR 乳児死亡率（0 歳での死亡数/出生 1000）

LIFEEXP ゼロ歳平均余命（いわゆる平均寿命）

TFR 合計出生率

NDAIDS HIV/AIDS による年間死亡数

APHIVAIDS HIV/AIDS の成人有病割合 (%)

GDPPCUSD 米ドル換算購買力平価ベース 1 人当たり GDP（国内総生産）

PUNEMP 失業者割合 (%)

GINI 国ごとの家族の所得の不平等度を示す Gini の集中係数（完全な平等のとき 0、完全な不平等で 100）

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

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

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

文献

- Rothman KJ (2012) Epidemiology: An Introduction. 2nd Ed. Oxford University Press, Oxford.
- “Chapter 14. Sample size issues.” in Machin D, Campbell MJ, Walters SJ (2007) Medical Statistics, 4th ed., Wiley, pp. 261-275.
- “Chapter 4. Comparing groups with p values: Reporting hyothesis tests.” in Lang TA, Secic M (2006) How to report statistics in medicine: Annotated guidelines for authors, editors, and reviewers. 2nd ed., American College of Physicians., pp.45-60.
- “Chapter 1. Research design” in Peacock JL, Peacock PJ (2011) Oxford handbook of medical statistics. Oxford Univ. Press, pp.1-73 (especially 56-73).
- Faraway JJ (2006) Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regrassion Models. Chapman and Hall.
- http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/Getting-Started-with-the-Rcmdr.pdf（作成者である John Fox 自身による R Commander の入門テキスト）
- http://www.ec.kansai-u.ac.jp/user/arakit/documents/Getting-Started-with-the-Rcmdr-ja.pdf（日本語版メニュー作成者である荒木孝治さんによる邦訳）
- 青木繁伸 (2009) R による統計解析. オーム社
- 大橋靖雄、浜田知久馬 (1995) 生存時間解析: SAS による生物統計. 東京大学出版会.
- 奥村晴彦 (2016) Wonderful R 1: R で楽しむ統計. 共立出版.
- 神田善伸 (2015) EZR でやさしく学ぶ統計学　改訂 2 版〜EBM の実践から臨床研究まで〜、中外医学社. http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmed.html
- 神田善伸 (2014) 初心者でもすぐにできるフリー統計ソフト EZR(Easy R) で誰でも簡単統計解析, 南江堂. http://www.nankodo.co.jp/g/g9784524261581/
- 神田善伸 (2016) ゼロから始めて一冊でわかる！　みんなの EBM と臨床研究. 南江堂.http://www.nankodo.co.jp/g/g9784524255481/

- 新谷　歩 (2011) 今日から使える医療統計学講座【Lesson 3】サンプルサイズとパワー計算. 週刊医学界新聞、2937 号 (http://www.igaku-shoin.co.jp/paperDetail.do?id=PA02937_06)
- 新谷　歩 (2016) みんなの医療統計：12 日間で基礎理論と EZR を完全マスター！. 講談社サイエンティフィク.http://www.kspub.co.jp/book/detail/1563148.html
- 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.https://minato.sip21c.org/statlib/stat-all-r9.pdf
- 中澤　港 (2007) R による保健医療データ解析演習. ピアソン・エデュケーション.https://minato.sip21c.org/msb/medstatbookx.pdf
- 永田　靖 (2003) サンプルサイズの決め方. 朝倉書店.
- 藤井良宜 (2010) カテゴリカルデータ解析（R で学ぶデータサイエンス 1）. 共立出版.
- 古川俊之 [監修]・丹後俊郎 [著] (1983) 医学への統計学. 朝倉書店.
- 村山　航 (2012) 妥当性概念の歴史的変遷と心理測定学的観点からの考察. 教育心理学年報, 51: 118-130.https://www.jstage.jst.go.jp/article/arepj/51/0/51_118/_article/-char/ja
- Bull K, Spiegelhalter DJ (1997) Tutorial in biostatistics: Survival analysis in observational studies. *Statistics in Medicine*, 16: 1041-1074.https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<1041::AID-SIM506>3.0.CO;2-F（大雑把な日本語解説と R コードを https://minato.sip21c.org/swtips/survival.html で公開している）
- Kartsonaki C (2016) Mini-Symposium: Medical Statistics: Survival analysis. *Diagnostic Histopathology*, 22(7): 263-270. https://doi.org/10.1016/j.mpdhp.2016.06.005

索引