EZR による医学統計入門

第 2.0 版

下川敏雄

和歌山県立医科大学附属病院 臨床研究センター

目次

前章：本資料の概要

0.1 EZR の概要とインストール方法

0.1.1 EZR の概要

EZR とは，自治医科大学附属さいたま医療センター 血液科 神田善伸教授が，R の GUI 環境の一つである R コマンダーを医学統計用にカスタマイズしたものである．そのため，解析自体は，統計学でのデファクトスタンダードである，統計解析環境 R が行っている．

0.1.2 EZR のインストール：Windows の場合

EZR は，自治医科大学埼玉さいたま医療センター 血液科のホームページ

http://www.jichi.ac.jp/saitama-sct/

からダウンロードできる．なお，ブラウザー(例えば，google)から EZR を検索すると，トップページに上記の HP が出てくるようになっている．

図 0.1 は，自治医科大学さいたま医療センター血液科のホームページである．ダウンロードまでの手順を以下に示す．

図 0.1：EZR のダウンロード

[STEP.0] 「自治医科大学埼玉さいたま医療センター 血液科のホームページ」に移動する.

[STEP.2] 「ダウンロード(〇〇版)」を左クリックする. ここで, 〇〇はインストールするパソコンの OS である. 統計解析環境 R がプラットフォーム非依存なので, EZR についても OS に関係なく利用することができる.

[STEP.3] 「〇〇版はここをクリックしてダウンロードいてください(Ver. X.X 20XX/X/X)」を左クリックする. ここで, XX は, バージョンおよび公開日である.

STEP.3 までの作業を行うと, 「EZRsetup.exe」(Windows の場合)という実行ファイルのダウンロードと保存先について聞かれるので, 適当な場所(例えば, デスクトップ)に保存する.

そして, 保存したファイルをダブルクリックして実行する. ダブルクリックをすると, 「EZR をインストールしてます」という画面が表示される. ここで, インストール先(デステネーションフォルダ)を設定するが, とくにこだわりがなければ, そのまま OK ボタンを押しても問題ない.

0.1.3 EZR のインストール：MacOS の場合

EZR 及び R コマンダー(EZR)を MacOS で動作せるためには, X11 ウィンドウシステムが必要になる. しかしながら, Mountain Lion 以降の MacOS では, X11 がプリインストールされていないことから, EZR のインストールに先立って, X11 をインストールしなければならない.

X11 は, 以下の XQuartz のサイト(https://www.xquartz.org/)から Mac 用のイメージファイル「Xquartz-X.X.XX.dmg」(X はバージョンを表す数字)ダウンロードしたうえで, インストールすればよい.

また, MacOS 版は, インストーラーが存在しないことから, 「Step.1：統計解析環境 R のインストール」, 「Step.2：R を起動したうえで, R コマンダーおよび EZR をインストールする」の手順でインストールしなければならない. 詳細な手順を以下に示す.

Step.1	統計解析環境 R をインストールする. 統計解析環境 R は, CRAN(Comprehensive R Archive Network)のサイト https://cran.r-project.org/ からインストールできる. — 上記ホームページの「Download and Install R」のなかの「Download R for (Mac) OS X」をクリックすると, MacOS 用のダウンロードサイトに移動する. — MacOS の統計解析環境 R のインストーラーは, 「Lasted release」の下側にある「R-X.X.X.pkg」(X はバージョンを表す数字)である. これをクリックすればインストールが開始される.
Step.2	統計解析環境 R を起動して, R コマンダー及び EZR をインストールする(一度実施すれば, 統計解析環境 R を再インストールしない限り, 改めて行う必要はない). — 統計解析環境 R を起動すると, 「R Console」というウィンドウが表示されるので, 赤色のコマンドプロンプト「＞」のところで, ＞ install.packages("RcmdrPlugin.EZR", dep=T) と入力したうえで, Enter キーを押す. すると, 「Secure CRAN mirrors」という新しいウィンドウが表示される. これは, CRAN のミラーサイトを選択することを意味する. 基本的には, どれを選択しても構わないが, 日本のミラーサイトを選択する場合には, 東京大学のサイト「Japan (Tokyo)[https:]」を選択すればよい. — 上記の代わりに, 「パッケージとデータ」→「パッケージのインストール」から RcmdrPlugin.EZR を選択しても同じである.

(a) Rコンソール画面　　(b) Rコマンダー(EZR)画面

図 0.2：EZR の画面

0.1.4 EZR の起動

Windows の場合には，EZR のインストール後に R のアイコンと EZR のアイコン(アイコン画像は同じである)の 2 種類が作成され，EZR のアイコンをクリックすれば，EZR が起動する．

一方で，MacOS の場合には，EZR のアイコンが作成されないため，統計解析環境 R を起動したうえで，EZR を読み込まなければならない．以下に，起動の方法を示す．

Step.1	統計解析環境 R を起動する．
Step.2	統計解析環境 R を起動すると，「R Console」というウィンドウが表示されるので，赤色のコマンドプロンプト「＞」のところで， ＞ library(Rcmdr) と入力したうえで Enter キーを押す．この作業でエラーが表示される場合には，library(“Rcmdr”, dep=T)と入力する．あるいは，「パッケージ」→「パッケージの読み込み」から，Rcmdr を選択してもよい．
Step.3	Step.2 を実行すると，R コマンダーが起動するので，メニューの中の「ツール」→「R.app のための Mac OS X の app.nap の管理」で app nap の設定をオフに設定する．
Step.4	「ツール」→「Rcmdr プラグインのロード」として，「RcmdrPlugin.R」を選択する．すると，「再起動しますか？」という問いが出るので，「はい」を選択して R コマンダーを再起動させると，R コマンダーが EZR に変更される．

0.2 EZR の基本操作

0.2.1 操作画面の概要

EZR を起動すると，2 画面(R の画面，EZR の画面)が表示される(図 0.2)．ここで，R コンソール画面(図 0.2(a))は，とくに触る必要はない(EZR を終了する場合に，この画面右上の「×」ボタンを押すか，あるいは「ファイル」→「終了」を選択するのみである)．

EZR の実行は，R コマンダー(EZR)画面(図 0.2(b))で実行する．EZR では，R のスクリプト(プログラム)を自動生成することで統計解析を実行する．この画面の上側(R スクリプト)には，自動生成された R のスクリプトが表示される．通常

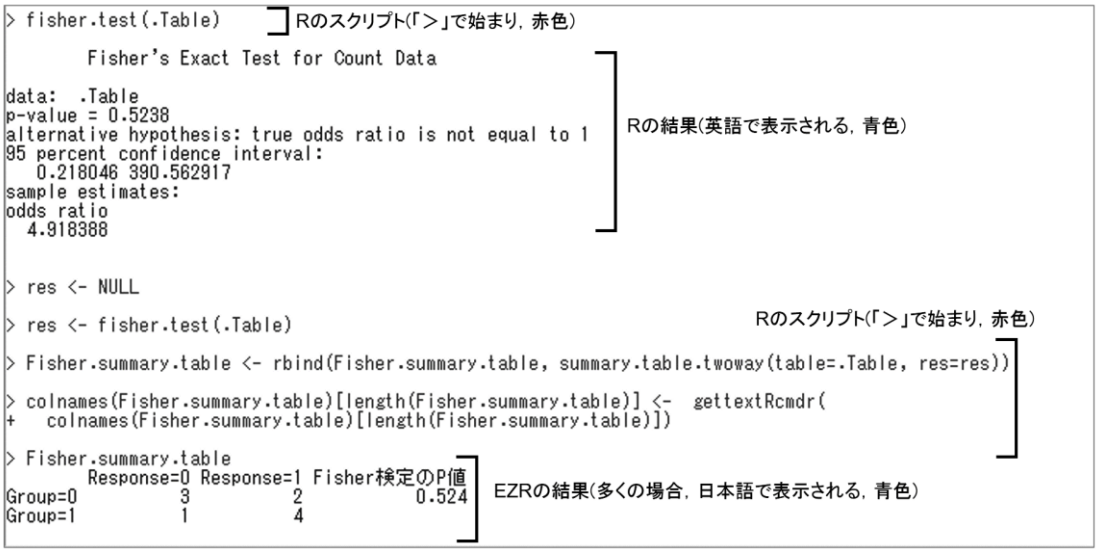


図 0.3：EZR の解析結果表示例

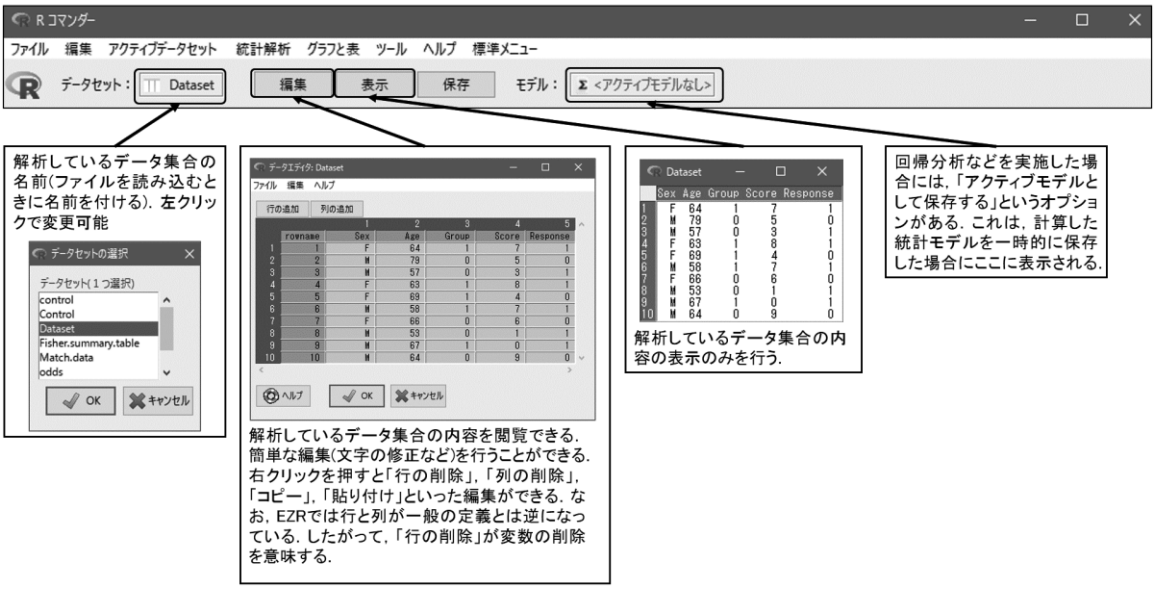


図 0.4：EZR のメニュー下の説明

は，用いなくてよい．下側(出力)には，実行された R のスクリプト及び結果が表示される．このとき，赤色の文字が R のプログラムを表しており，青色の文字が結果を表している．

図 0.3 は，EZR の実行例を表している(出力の部分)．「＞」で始まる赤色の文字は，EZR により自動生成された R のスクリプトであり，無視してかまわない．

青色の文字は，R あるいは EZR の出力を表している．R の出力はすべて英語表記になっているのに対して，EZR の出力の多くは日本語で表記される．また，R のスクリプトの関係で，R での出力が先に表示され，次いで，EZR の出力が表示される．EZR では，先に出力される R の出力の抜粋になっており，必要に応じて R の出力を見なければならないが，多くの場合には，EZR での出力のみを見ればよい(R の出力を見なければいけない場合については，1 章以降で説明する)．

図 0.5：CSV ファイルの読み込み

0.2.3 データの閲覧・簡単な編集

ここでは，メニュー下のボタンの簡単な説明および，単純な編集の方法について述べる(図 1)．「データセット」横の文字は，現在計算しているデータ集合を表している(図 0.4 の場合には，Dataset である)．「編集」は，データ集合を編集可能な状況で表示させるボタンである．編集の仕方は，多くの統計パッケージと同じである．また，セル上で右クリックすると

・現在の行の削除(変数を削除することを意味する)

・現在の列の削除(被験者を削除することを意味する)

・セルの削除　　・セルの切り取り　・セルのコピー　　・セルの貼り付け

が選択できる．

0.2.2 ファイル処理

EZR では，テキストファイル，CSV ファイルだけでなく，Excel ファイルなど，様々なファイルフォーマットを扱うことができる．

図 0.5 は，CSV ファイルの読み込み方法である．読み込みは，「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」を選択する．このとき，Excel のデータの場合には，「データのインポート」から「Excel データをインポート」を選択する．

次いで，読み込むファイルの形式を設定する．データセット名のデフォルトは「Dataset」だが，名称を変更する場合には，ここに入力する．

CSV ファイルの最初の列には，変数名すなわち，

	A	B	C	D	E
1	Sex	Age	Group	Score	Response
2	F	64	1	7	1
3	M	79	0	5	0

のように入力することが推奨される. もし, 入力していない場合には,「ファイル内に変数名あり」のチェックボックスを外す. チェックボックスを外した場合の変数名は, V1,V2,…のようになる.

フィールドの区切り線は, CSV ファイルの場合には,「カンマ」(デフォルト)になる. また, テキストファイルの場合には, 適切な区切り文字を選択する.

1 章：量的データにおける統計解析

1.1 統計学序論

1.1.1 データの形式

データの種類は，量的データと質的データの 2 種類に大別される．量的データとは，個々の観測値が数量で表されるデータであり，平均値あるいは中央値を用いて要約される．量的データには，計量データと計数データの 2 種類がある．計量データとは，血圧，腫瘍径，出血量などのように，数値に単位があるようなデータである．計量データは，小数点以下の値をとり，連続的に切れ目がないため，連続データと呼ぶこともある．一方で，計数データとは，ポリープの個数やリンパ節転移個数のように，個数あるいは回数として計測されたデータである．

質的データは，2 値データと多値データに分けられ，多値データは，更に名義カテゴリカル・データと順序カテゴリカル・データに分けられる．2 値データとは，奏効の有無，疾患の有無，治療の改善・非改善のように，アウトカムが 2 カテゴリで表されるデータである．これに対して，多値データは，3 個以上のカテゴリで表される．名義カテゴリカル・データとは，カテゴリが被験者の状態を表すラベルとして扱われるデータであり，疾患の種類や血液型がこれに該当する．一方で，疾患の進行程度を軽度，中程度，重度のカテゴリで測る場合，疾患の進行には，軽度＜中程度＜重度の順序関係が成り立つ．このように，カテゴリに順序関係が存在する場合を順序カテゴリカル・データという．

1.1.2 量的データの要約

本節では，量的データのなかでも，とくに計量データを要約する方法について略説する．これに対して，計数データの場合には，級分け(例えば，0 個，1-2 個，3 個以上など)を実施したうえでクロス集計表を作成するか，あるいは中央値を用いることが多いため，ここでは割愛する．

(1) 平均値と中央値

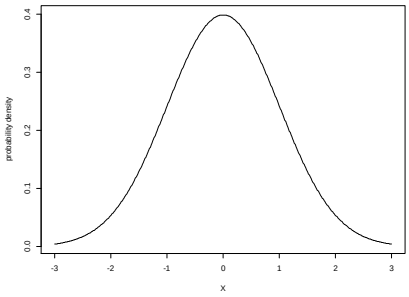
臨床試験の結果を報告するとき，被験者背景を要約する必要がある．このとき，量的データの要約に平均値と中央値のどちらを用いるかを選択する必要がある．医学論文における統計的方法の報告をまとめた SAMPL ガイドライン[1]では，「データが正規分布に従っていると考えられる場合には平均値，そうでない場合には中央値を用いる」ことが記載されている．ただし，背景因子をまとめた表において，ある項目が平均値であるにもかかわらず，別の項目が中央

[1] Lang, T.A. and Altman, D.G.：Reporting Basic Statistical Analyses and Methods in the Published Literature: The SAMPL Guidelines for Biomedical Journals, http://www.equator-network.org/wp-content/uploads/2013/07/SAMPL-Guidelines-6-27-13.pdf.

値であるというのは，非常にわかりにくい．また，データが正規分布に従っているのであれば，平均値と中央値がおおよそ等しい値をとることが期待されるため，中央値を背景因子に用いることが多いように思われる．

(忘記録) 正規分布とは

正規分布とは，統計学の最も基本的な確率分布(統計学では個々のデータは，ある確率によって得られると考えている．このとき，得られたデータとその確率の対応関係のことを確率分布という)であり，自然現象や社会現象の多くの事象は正規分布に従っていると考えている．



また，統計学の多くの方法は，正規分布に基づいている．因みに，正規分布は，平均と標準偏差によって成り立っている．我々がデータを要約して評価する場合には，平均値を用いることが多いが，このことは，暗黙裡に正規分布を想定しているといえる(例えば，期末テストの成績を平均点で評価するなど)．正規分布は，左図のような釣り鐘型の左右対称な形状を示している．

(2) バラツキの要約

バラツキ(データの散らばり具合)の要約は，データの代表値(平均値，中央値)に何を利用するかによって異なり，平均値を用いる場合には標準偏差(あるいは標準誤差)，中央値を用いる場合には四分位範囲(あるいは範囲)を用いなければならない．

被験者数(N)のデータの標準偏差(SD)に対して，標準誤差 SE は $SE=SD/\sqrt{N}$ であるため，標準誤差のほうが小さくなる．そのため，「見栄え」の観点から標準誤差が用いられることがある．ただし，これは標準誤差に対する誤用である．標準偏差とはデータのバラツキ具合を表しており，標準誤差とは平均値のバラツキ(いいかえれば，平均値の信頼性)を表している．被験者背景を要約する場合，被験者にどの程度の個人差があるのかを示すことが重要であるため，標準偏差を用いることが推奨される．一方で，エンドポイントの評価では，平均値にどの程度の信頼性があるかを見る必要があるため，標準偏差を利用するよりも標準誤差のほうが適切である．ただし，SAMPL ガイドラインでは，標準誤差を利用せずに信頼区間を用いたほうが良いと記載されている．なぜなら，データが正規分布に従っているとき，標準誤差は約 68%信頼区間を表しており，バラツキを過小評価しているためである．そのため，SAMPL ガイドラインでは，可能な限り 95%信頼区間を用いることが推奨されている．

また，平均値と標準偏差を「平均値±標準偏差」の形式で記載している論文が散見されるが，先述したように，標準偏差はデータのバラツキを表すことから適切でなく，「平均値(標準偏差)」による記載が本来は適切である(学会誌によっては，±による表記を推奨している場合があるので．注意が必要である)．

四分位範囲は，第3四分位点と第 1 四分位点によって構成される．第 3 四分位点とは，最大値と中央値のあいだの中央の値であり，第 1 四分位点とは最小値と中央値のあいだの最大の値である．すなわち，四分位範囲は，中央値まわりの 50%のデータが含まれる領域として定義される．これに対して，範囲は，最大値と最小値によって構成されるため，100%のデータが含まれる範囲として定義される．範囲は，当該試験の被験者がすべて適格性を満たしていることを示すのに有利であり，一方で，四分位範囲は，外れ値等の影響を受けずに中央値まわりでのバラツキを表すことができる．SAMPLE ガイドラインでは，四分位範囲あるいは範囲のいずれか，あるいは両方を記載することを求めている．

(3) 信頼区間とは

A 病院の月曜日に来院する患者 100 名の臨床検査値の平均値を計算し，このときの検査値の平均値を病院の代表値と決めたとする．このとき，火曜日の来院患者 100 名に同じように平均値を計算しても同じになることは殆どない．このような研究では，研究対象は A 病院の患者の臨床検査値(母集団)であり，月曜日の 100 名の患者の臨床検査値は，母集団を構成する 1 部(標本)である．つまり，月曜日の患者 100 名から計算した平均値は母集団での平均(母平均)の類推であるといえる．これを推定値といい，月曜日の患者から計算した平均値のように，単一の数値で表す推定値を点推定値(point estimator)という．

これに対して，母平均を区間で推定するものを区間推定値という．医学統計学で良く用いられる 95%信頼区間とは，「100 回同じ研究を実施して 95%信頼区間を構成したときに，95 回の研究で母平均が含まれる区間」として定義される．

(4) 仮説検定とは

いま，抗がん剤治療中の胃癌患者に対して，術後補助化学療法開始時から栄養介入を実施した 53 名(栄養介入群)と実施しなかった 47 例(栄養非介入群)での治療後 6 カ月間での体重減少率を比較する研究を実施した．その結果，栄養介入群での体重減少率の平均値は 4.86%(標準偏差 :3.72)であり，栄養非介入群での体重減少率の平均値は 6.60%(標準偏差 :4.90)であった．このとき，「栄養介入が術後補助化学療法を抑制したと判断してよいだろうか」．このことを統計学的に判断する方法が，仮説検定(検定)である．

仮説検定では，2 種類の仮説(帰無仮説 H_0，対立仮説 H_1)を設定する．帰無仮説 H_0 とは，言いたいことと反対の仮説(栄養介入の有無によって体重の平均減少率に違いがない)であり，対立仮説 H_1 とは，本来言いたい仮説(栄養介入の有無によって体重の平均減少率に違いがある[2])である．そして，帰無仮説 H_0 の「確からしさ」が小さいときに，帰無仮説 H_0 が誤っている(棄却される)と判断し，その逆仮説である対立仮説 H_1 が正しい(有意である)と判断する．

帰無仮説 H_0 が正しいとしたもとで，今回の研究結果が「どれぐらいの確率で生じるのか」を計算するとき，この確率は，p 値(有意確率)と呼ばれ「帰無仮説 H_0 の確からしさを表す確率(厳密には，帰無仮説 H_0 が正しいと仮定したときに，研究の結果がどれぐらいの確率で生じ得るか)」として解釈される．事例での p 値は 0.047 であることから，帰無仮説 H0 の確からしさは 4.7%であることがわかる．

このとき，「帰無仮説 H_0 が誤っている(統計用語では，「棄却される」，「有意である」と呼ばれる)」と判断するには，p 値に対する閾値(通常は 0.05)を予め規定しなければならない．この閾値が有意水準 α である．有意水準 α=0.05 とするとき，この研究での p 値は，有意水準 α よりも小さいことから，帰無仮説 H_0 が棄却される．したがって，「栄養介入の有無によって体重の平均減少率に違いがある」と解釈できる．

(5) 量的データにおける統計的方法

ここでは，単群(単アーム)研究及び 2 群比較における位置を表す測度に対する仮説検定の種類について述べる(3 群以上の比較については，次項で触れる)．図 1.1 は，本章で取り上げる検定手法の取捨選択のフローチャートである．単群研究とは，ヒストリカル・コントロール(既存論文やこれまでの臨床成績)と臨床試験での結果を比較する場合である．2 群比較とは，2 種類の治療，あるいは 2 水準の要因によるアウトカムの違いを比較する場合である．2 群比較で

[2] 対立仮説には，両側対立仮説と片側対立仮説が存在する．今回の場合には，両側対立仮説(違いがある)と判断する場合である．これに対して，片側対立仮説では，「栄養介入があるほうが栄養介入がないよりも体重減少量が高くなる」あるいは「栄養介入があるほうが栄養介入がないよりも体重減少量が低くなる」になる．書く検定での両側対立仮説と片側対立仮説は，各検定の略説において解説する．

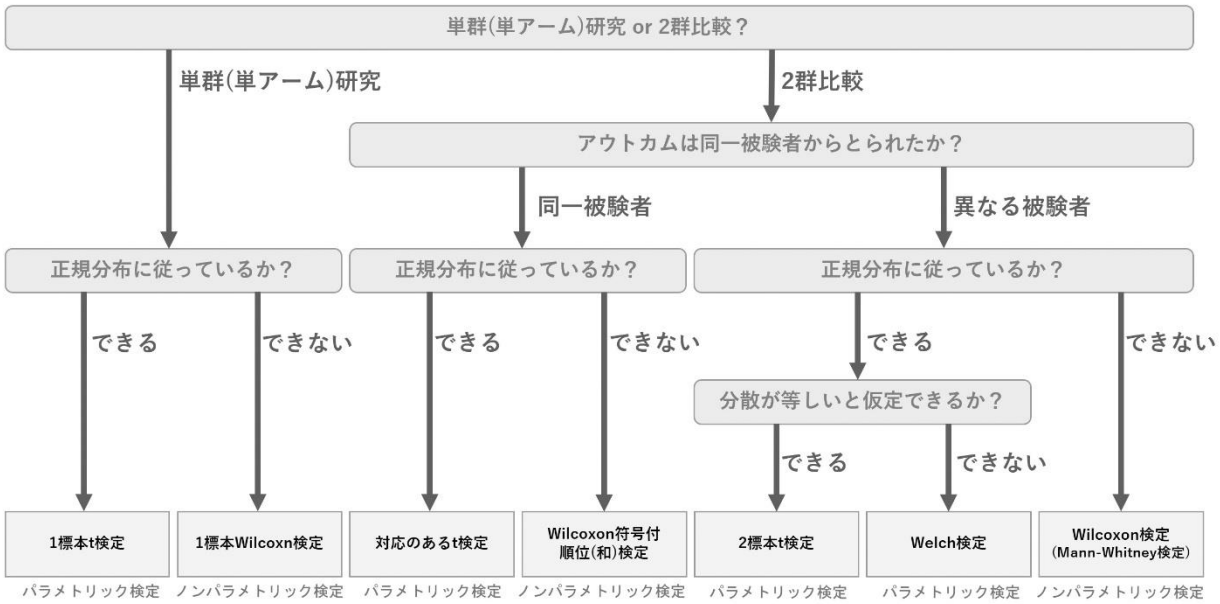


図 1.1：量的データにおける検定の取捨選択

は，アウトカムの取得方法で仮説検定の選択方法が異なる．アウトカムが同一被験者からとられる場合には，治療前後でのアウトカムの比較，或いはクロスオーバー試験がある．因みに，アウトカムが同一被験者からとられることを対応のある場合，あるいはマッチドペアという．一方で，アウトカムが異なる被験者からとられるとは，無作為化比較試験あるいはケース・コントロール研究のように，異なる介入或いは要因をもつ群間のアウトカムを比較する場合であり，独立 2 標本と呼ばれる．

単群研究，2 群比較(対応がある場合，独立 2 標本)のいずれにおいても，アウトカムが正規分布に従っているかどうかによって検定方法が異なる．アウトカムが正規分布に従っている場合には，平均によってアウトカムの相対的な位置関係を要約できる．すなわち，母集団における平均を評価する検定が採用される．正規分布に基づく検定方法のことをパラメトリック検定という．

一方で，アウトカムが正規分布に従っていない場合(例えば，アウトカムの分布形状が歪んでいる場合)，アウトカムの「順位」を用いることで，アウトカムの分布における相対的な位置関係を検討する．正規分布に拠らない検定方法のことをノンパラメトリック検定という．

1.2 ヒストリカル・コントロールとの比較 (1 標本における統計的推測)

(1) データの概要：腎機能患者の血清クレアチニン濃度データ

病院Aに通院する，腎機能障害の患者 6 名の血清クレアチニン濃度(mg/dl)を測定したところ

4.0　3.9　3.8　4.0　4.4　3.9

という観測値が得られた．これに対して，病院Bにおける，同じ腎機能障害の血清クレアチニン濃度の平均値は 4.3(mg/dl)であった．病院Aと病院Bを受診した患者層が異なるといえるかを検討しなさい．このデータのファイルは，One_sample_t.csv である．

(2) 1 標本における統計的方法

単群の臨床研究では，ヒストリカル・コントロールとの比較を行うことがある．このとき，ヒストリカル・コントロールが平均値の場合は 1 標本 t 検定，ノンパラメトリック検定の場合は 1 標本 Wilcoxon 検定がある．ただし，1 標本 Wilcoxon

検定は，中央値を代表値としているわけではなく，設定した任意の値に対して，分布が相対的にずれているか否かを評価するため，解釈が困難な場合がある．そのため，1 標本における統計的評価には 1 標本 t 検定を用いるのが一般的である．因みに，EZR では，1 標本 t 検定のみが実装されている．そのため，ここでは 1 標本 t 検定のみを取り上げる．

1 標本 t 検定では，帰無仮説 H_0「母平均は μ_0(ヒストリカル・コントロール)に等しい」に対する評価を行う．このとき，対立仮説には以下の 3 種類が存在する．

両側対立仮説 H_{1a}：母平均は μ_0(ヒストリカル・コントロール)と異なる．

片側対立仮説 H_{1b}：母平均は μ_0(ヒストリカル・コントロール)よりも大きい．

片側対立仮説 H_{1c}：母平均は μ_0(ヒストリカル・コントロール)よりも小さい．

因みに，臨床試験における第 II 相試験では，片側対立仮説を用いることが多いが，一般的な適用場面では両側対立仮説が用いられる．

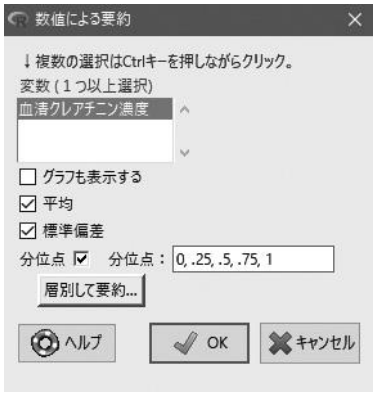
(3) EZR による 1 標本 t 検定の計算

ここでは，血清クレアチニン濃度のデータ(One_sample_t.csv)を用いて，EZR での計算方法について述べる．

なお，仮想データは，以下の手順で読み込むことができる．

「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」を選定し，ファイル(One_sample_t.csv)を選択する．

先ず，データの傾向を捉えるために，記述統計量を計算する．

量的データの要約(1)
1:　「統計解析」→「連続変数の解析」→「連続変数の要約」を選択する． 2:　次のようなメニューが表示される．    このとき， ・「変数(1 つ以上選択)」で「血清クレアチニン濃度」を選択する． 3:　「OK」ボタンを押す

ここで，分位点の数字(0, .25, .5, .75, 1)は(本来は)パーセント点と呼ばれるものであり，以下を意味する．

0.00：最小値，　0.25：第 1 四分位点(四分位範囲の下限値)，　0.50：中央値(第 2 四分位点)

0.75：第 3 四分位点(四分位範囲の上限値)，　1.00：最大値

また，「グラフも表示する」をチェックした場合には，ドットプロット(1 次元散布図)が表示される．

このとき，次のような出力が表示される．

```
      平均   標準偏差  0%    25% 50% 75% 100% n
 4.083333 0.2562551 3.8 3.925    4 4.3   4.4 6
```

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される. 赤色が R のコマンドであるが, 無視してかまわない(EZR では, 出力情報は, すべて青色で表示される).

出力結果より, つぎのことがわかる.

・平均値は, 4.08 である, ・標準偏差は 0.256 である, ・最小値は 3.8 である.

・最大値は 4.4 である. ・四分位範囲は, [3.93, 4.3]である, ・被験者数(n)は 6 名である.

したがって, 病院 A に通院する腎機能障害患者 6 名の血清クレアチニン濃度の平均値(4.08mg/dl)は, 病院 B の平均値(4.3mg/dl)よりも低いことが伺える. 因みに, 平均値と標準偏差を,「4.08±0.256」で表す場合があるが, 標準偏差は, データのバラツキを表すものであり, 平均値の信頼性を表すものではない. そのため, SAMPLEガイドラインでは, このような記述ではなく,「4.08(0.256)」で表すことが推奨されている.

次いで, 1 標本 t 検定により評価する. ここでは, 病院 A に通院する腎機能障害患者の血清クレアチニン濃度が病院 B の患者の平均値(4.3mg/dl)と異なるか否かを評価する(したがって, 両側対立仮説になる).

1 標本 t 検定の実行
1: 「統計解析」→「連続変数の解析」→「1 標本の平均値の t 検定」を選択する. 2: 次のようなメニューが表示される.    このとき, ・「変数(1 つ選択)」で「血清クレアチニン濃度」を選択する. ・「対立仮説」で「母平均 $\mu \neq \mu 0$」を選択する. ・「帰無仮説 $\mu = \mu 0$」で「$\mu 0$」横に「4.3」と入力する. 3: 「OK」ボタンを押す

ここで, 対立仮説は, 3 種類の対立仮説を表しており,

「母平均 $\mu \neq \mu 0$」: 両側対立仮説(病院 A に通院する腎機能障害患者の血清クレアチニン濃度が病院 B の患者の平均値(ヒストリカル・コントロール:4.3mg/dl)と異なる)

「母平均 $\mu < \mu 0$」: 片側対立仮説(病院 A に通院する腎機能障害患者の血清クレアチニン濃度が病院 B の患者の平均値(ヒストリカル・コントロール:44.3mg/dl)よりも低い)

「母平均 $\mu > \mu 0$」: 片側対立仮説(病院 A に通院する腎機能障害患者の血清クレアチニン濃度が病院 B の患者の平均値(ヒストリカル・コントロール:44.3mg/dl)よりも高い)

また,「帰無仮説 $\mu = \mu 0$」横の箱は, ヒストリカル・コントロールの数値を入力するためのものである. さらに, 信頼水準(デフォルト 0.95)とは, 信頼区間の信頼係数を表しており, 0.95 の場合には, 母平均に対する 95%信頼区間が描写される.

このとき, 次のような出力が表示される.

平均 = 4.083333, 95%信頼区間 3.81441-4.352257, P 値 = 0.0931

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される. 赤色が R のコマンド, 青色が R での出力であ. ただし, 上記の EZR の出力と同様の情報が重複して表示されているだけであることから, 改めて見る必要がない.

その結果, 平均は 4.08, 信頼区間は[3.81, 4.35]であった. また, p 値が 0.0931 であることから, 有意水準 α=0.05 のもとで有意でなかった. したがって, 病院 A に通院する腎機能障害患者の血清クレアチニン濃度がヒストリカル・コントロールの 44.3mg/dl と異なる(病院 B と異なる)という根拠は得られなかった.

(4) 余禄:有意でない場合に, 帰無仮説 H_0 が正しいと言ってよいか？

仮説検定において有意でない場合(帰無仮説 H_0 が棄却できない場合), 帰無仮説 H_0 が正しいと解釈してはならない. なぜなら, 仮説検定とは, 帰無仮説 H_0 と対立仮説 H_1 の二者択一の評価を実施しているわけではなく,「帰無仮説 H_0 が棄却できない」とは, 帰無仮説 H_0 を棄却する根拠がないことを主張しているに過ぎないためである.

病院 A に通院する腎機能障害患者の血清クレアチニン濃度のデータでは, p 値が 0.0931 であり有意でなかった. このことは,「病院 A に通院する腎機能障害患者の血清クレアチニン濃度がヒストリカル・コントロールの 44.3mg/dl と同じである」ことを示しているわけでなく, 病院 A に通院する腎機能障害患者の血清クレアチニン濃度がヒストリカル・コントロールの 44.3mg/dl と異なるという根拠が得られなかった」と解釈すべきである.

1.3 2 標本における統計的推測

1.3.1 データの概要：神経障害性疼痛データ

神経障害性疼痛患者を対象に, 2 種類の除痛薬(新薬, 既存薬)投与後の VAS (mm)の減少量を評価している.

新　薬 (n=14)	31	25	28	29	23	25	30	25	29	27	30	20	20	24
既存薬 (n=12)	23	23	20	27	19	15	25	29	15	13	28	21		

新薬と既存薬で VAS の減少量が異なるといえるかを検討しなさい. このデータのファイルは, VAS_comp.csv である.

1.3.2 2 標本における母平均の比較(2 標本 t 検定, Welch 検定)

(1)2 標本 t 検定及び Welch 検定

2 標本における母平均を比較するための方法には, 2 標本 t 検定と Welch 検定の 2 種類がある. いずれの方法でも, 仮説は同じであり, 帰無仮説 H_0「2 つの母平均 μ_1, μ_1 は等しい」に対して, 3 種類の対立仮説は

両側対立仮説 H_{1a}:2 つの母平均 μ_1, μ_1 は異なる($\mu_1 \neq \mu_2$).

片側対立仮説 H_{1b}:母平均 μ_1 のほうが母平均 μ_2 よりも大きい($\mu_1 > \mu_2$).

片側対立仮説 H_{1c}:母平均 μ_1 のほうが母平均 μ_2 よりも小さい($\mu_1 < \mu_2$).

である. 2 標本 t 検定及び Welch 検定では, 母集団が正規分布に従うことを仮定する. 正規分布は, 母平均と母分散(平方根をとると母標準偏差)から構成されるが, 2 標本 t 検定では 2 つの母集団における母分散が等しいことを仮定し, Welch 検定では, 等しいと仮定しない.

ただし, Welch 検定の利用については, 批判的な意見が報告されている. 2 標本の検定の関心は, (1)母集団の違いに差があるのか, (2)平均値の差にあるのか, に大別される. 関心の対象が(1)である場合には, 不等分散であることを示すことができれば(等分散性の検定), Welch 検定を用いる必要は必ずしも存在しない. 関心の対象が(2)である場合においても, 試験結果の分散(標準偏差)に明らかな違いがなければ, 2 標本 t 検定で十分であることがいくつかの文

献で指摘されている．また，母集団の分散が明らかに異なる場合には，母集団が正規分布に従っていないことが想定されるため，Mann-Whitney 検定(Wilcoxon 検定)などのノンパラメトリック検定を用いることが推奨される[3]．

さらに，「等分散性の検定」の結果で，有意であれば「Welch 検定」，有意でなければ「2 標本 t 検定」という取捨選択を推奨する文献があるが，このような作業は，検定を 2 回繰り返すことから，後述する多重比較を行っていることと同じであり，この取捨選択は誤りである．

(2) EZR による 2 標本 t 検定の計算

ここでは，神経障害性疼痛のデータ(VAS_comp.csv)を用いて，EZR での計算方法について述べる． なお，このデータは，以下の手順で読み込むことができる．

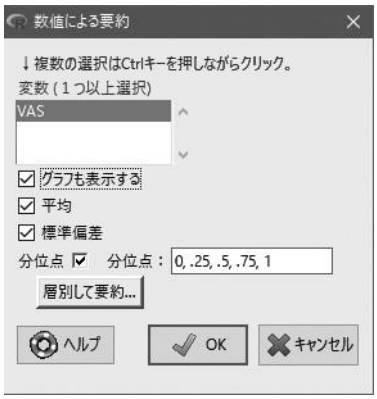
「ファイル」→「データのインポート」→「ファイルまたはクリップボード、URL からテキストデータを読み込む」を選定し，ファイル(VAS_comp.csv)を選択する．

先ず，データの傾向を捉えるために，記述統計量を計算する．

量的データの要約(2)

1: 「統計解析」→「連続変数の解析」→「連続変数の要約」を選択する．

2: 次のようなメニューが表示される．



このとき，

- 「変数(1 つ以上選択)」で「血清クレアチニン濃度」を選択する．
- 「グラフも表示する」にチェックを入れる．
- 「層別して要約. .」を押すと，次のメニューが表示される．

- 「Group」を選択し，「OK ボタンを押す」．

3: 「OK」ボタンを押す

ここで，「層別して要約. .」にチェックしたのは，グループ毎で要約統計量を計算するためである．つまり，ここではグループ毎(Active, Control)に要約統計量を計算することを意味する．

このとき，次のような出力が表示される．

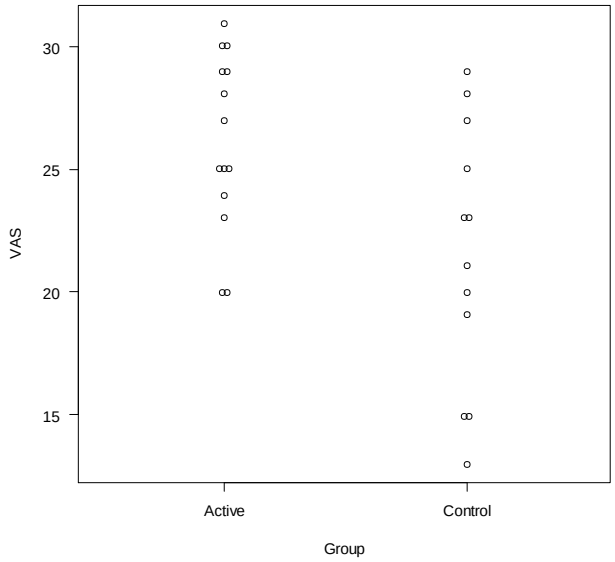
```
            平均 標準偏差 0%   25% 50%  75% 100% data:n
Active  26.14286  3.59181 20 24.25  26 29.0   31     14
Control 21.50000  5.31721 13 18.00  22 25.5   29     12
```

[3] 下川敏雄：実践のための基礎統計学，講談社，2016.

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される. 赤色が R のコマンドであるが, 無視してかまわない(EZR では, 出力情報は, すべて青色で表示される).

その結果, Active(新薬)のほうが, Control(既存薬)に比べて, VAS 減少量の平均値(Active=26.1, Control=21.5)及び中央値(Active=26, Control=22)ともに高いことが伺える.

このときのドットチャートの結果は, 別のウィンドウで下図のように表示される.



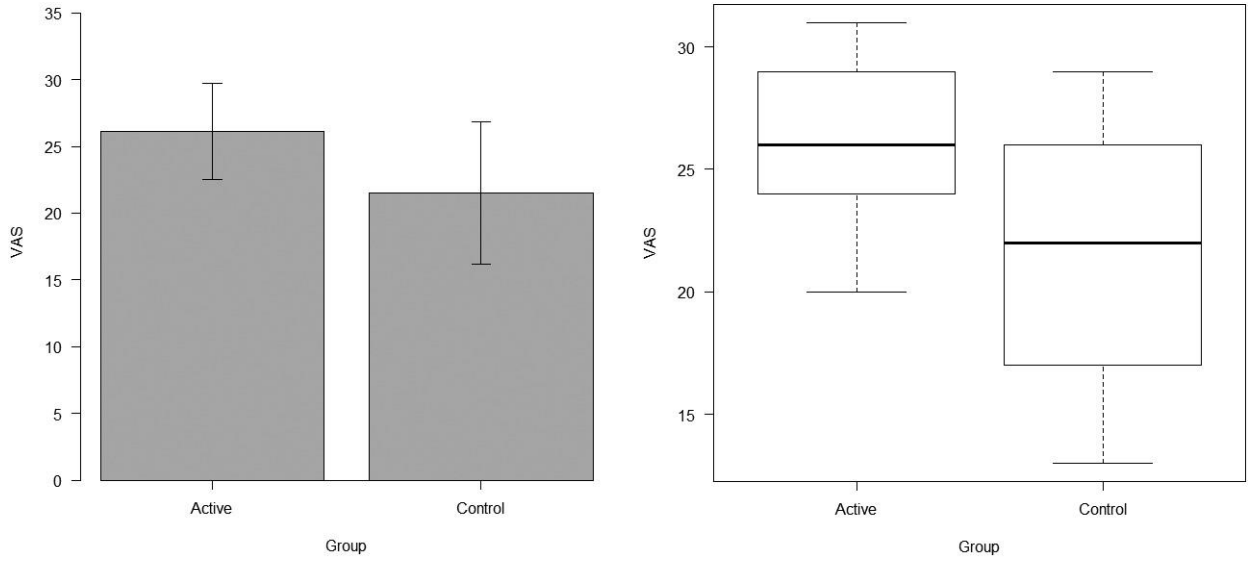
このグラフからも, Active 被験者のほうが Control に比べて, VAS 減少量が高いことが伺える. また, 2 群比較に用いることができる二つのグラフの描写方法(棒グラフ, 箱ひげ図)は, 以下のとおりである.

棒グラフの描写
1: 「グラフと表」→「棒グラフ(平均値)」を選択する.
2: 「棒グラフ」メニューから ・「目的変数(1 つ選択)」のなかで「VAS」を選択する. ・「群別化変数 1(0~1 つ選択)」のなかで「Group」を選択する. ・「群別化変数 2(0~1 つ選択)」は何も選択しない. ・「エラーバー」で「標準誤差」を選択する(今回は平均値を比較するため).
3: 「OK」ボタンを押す

因みに標準誤差とは, 平均値のバラツキを表すものであり, 平均値の信頼性の一つの指標である. 一方で, 標準偏差とは, データのバラツキを表すものであり, このデータの場合には, VAS 減少量の個人差を表している.

箱ひげ図(ボックス・プロット)の描写
1: 「グラフと表」→「箱ひげ図」を選択する.
2: 「棒グラフ」メニューから ・「変数(1 つ選択)」のなかで「VAS」を選択する. ・「群別する変数(0~1 つ選択)」のなかで「Group」を選択する. ・「上下のヒゲの位置」で「第 1 四分位点-1.5x 四分位範囲、第 3 四分位点+1.5x 四分位範囲」を選択する.
3: 「OK」ボタンを押す

箱ひげ図の「ヒゲ」の目的は異常値(あるいは外れ値)を検出することである. 一方で, 「10、90 パーセンタイル」では, データの上下 10 パーセント, 「5、95 パーセンタイル」では, データの上下 5 パーセントが異常値として「必ず」表示され



(a) 棒グラフ　　(b) 箱ひげ図

図 1.2:2 群比較におけるグラフ表示

る. これらの表示形式では,「可能であれば存在してほしくない」異常値(外れ値)を「必ず」表示させるため, 好ましい表示方法ではない. そのため, 一般的には今回の設定方法を用いるほうが多い[4].

なお, 過度な異常値(外れ値)の存在が確認されたからといって, 勝手にデータを削除することは「データの改ざん」になるため, 行ってはならない. 異常値(外れ値)の取扱い方法は, 以下のとおりである.

- 異常値(外れ値)が生じた合理的な理由(単位が異なっていた, 記載ミスだった)があった場合には, 適切な数値に修正する.
- 異常値(外れ値)の影響を受けないノンパラメトリック検定(Mann-Whitney U 検定(Wilcoxon 検定)など)を用いる.

なお, 異常値(外れ値)を削除する合理的な理由がある場合には, 当該解析だけでなく, 研究対象から外し, その理由を論文・発表等で公表するほうがよい.

図 1.2 は, このときのグラフを表している. 棒グラフ(図 1.2(a))は平均値に基づいているため, 今回の母平均を比較するための検定, 箱ひげ図は(図 1.2(b))は中央値に基づいているため, 1.3.3 節のノンパラメトリック検定に用いることが推奨される.

次いで, 2 標本 t 検定により評価する. ここでは, 新薬(Active)と既存薬(Control)で VAS の減少量の母平均が異なるか評価する(したがって, 両側対立仮説になる).

2 標本 t 検定の実行
1:　「統計解析」→「連続変数の解析」→「2 群間の平均値の比較(t 検定)」を選択する.
2:　次のようなメニューが表示される.

[4] 統計検定(日本統計学会)3 級では, 最小値, 最大値を髭に用いている.

このとき，

- 「目的変数(1 つ選択)」で「VAS」を選択する．
- 「比較する群(1 つ以上選択、ただし 2 種類の値だけを持つこと)」で「Group」を選択する．
- 「対立仮説」で「両側」を選択する．
- 「等分散と考えますか」で「はい(t 検定)」を選択する．

3: 「OK」ボタンを押す

ここで，「対立仮説」は，3 種類の対立仮説を表しており(「目的変数(1 つ選択)」の下側の「差」は平均の差を表している)，

「両側」：両側対立仮説(新薬(Active)と既存薬(Control)で VAS 減少量の母平均が異なる)．

「差＜0」：片側対立仮説(新薬(Active)の母平均のほうが既存薬(Control)の母平均よりも VAS 減少量が小さい)．

「差＞0」：片側対立仮説(新薬(Active)の母平均のほうが既存薬(Control)の母平均よりも VAS 減少量が大きい)．

また，信頼水準(デフォルト 0.95)とは，信頼区間の信頼係数を表しており，0.95 の場合には，母平均の差に対する 95%信頼区間が描写される．さらに「等分散と考えますか？」は，2 標本 t 検定と Welch 検定を選択できる．なお，先述したように，Welch 検定の適用は推奨されないため，ここでは省略する(Welch 検定を実行したい場合には，「等分散と考えますか？」を「いいえ(Welch 検定)」にすればよい)．

このとき，次のような出力が表示される．

```
                  平均 標準偏差     P 値
Group=Active  26.14286  3.59181 0.0143
Group=Control 21.50000  5.31721
```

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される．p 値が 0.0143 であることから，有意水準 0.05 のもとで有意である．したがって，新薬と既存薬のあいだで VAS 減少量の平均の差に違いが認められた．

なお，平均値の差(Active の平均値ーControl の平均値)に対する 95%信頼区間(メニューから信頼水準(信頼係数)を 0.95 としている)は，「出力」を上にスクロールしたときの R での出力

```
        Two Sample t-test

data:  VAS by factor(Group)
t = 2.6425, df = 24, p-value = 0.01426
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.016647 8.269067
sample estimates:
 mean in group Active mean in group Control
             26.14286              21.50000
```

の太字部分に表示されている. すなわち, 平均値の差は, 26.14286−21.50000=4.64286 であり, その 95%信頼区間は, [1.016647, 8.269067]である. この信頼区間が 0 を含まないことからも, 新薬と既存薬の平均のあいだで VAS 減少量に有意な違いがあることがわかる.

なお, これらの出力とは別に, 棒グラフが表示されるが, 図 1.2(a)と同じ出力結果なので割愛する.

1.3.3 2 標本における等分散性の検定

(1) 等分散性の検定

2 標本における等分散性を比較するための方法には, 等分散性の検定がある. 等分散性の検定では, 帰無仮説 H_0「2 つの母分散 σ_1^2, σ_2^2 は等しい」に対して, 3 種類の対立仮説は

両側対立仮説 H_{1a}: 2 つの母分散 σ_1^2, σ_2^2 は異なる($\sigma_1^2 \neq \sigma_2^2$).

片側対立仮説 H_{1b}: 母分散 σ_1^2 のほうが母分散 σ_2^2 よりも大きい($\sigma_1^2 > \sigma_2^2$).

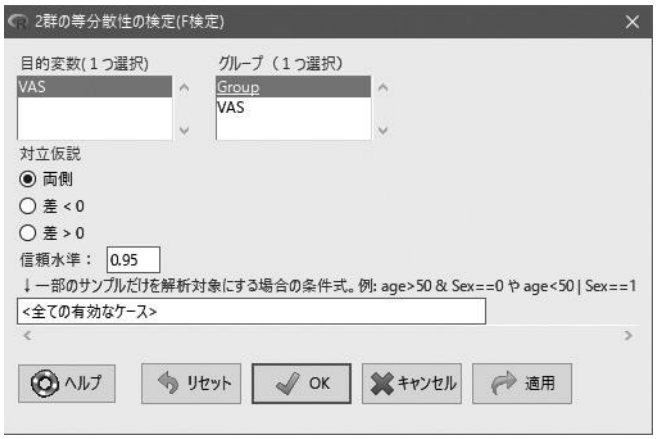
片側対立仮説 H_{1c}: 母分散 σ_1^2 のほうが母分散 σ_2^2 よりも小さい($\sigma_1^2 < \sigma_2^2$).

である.

(2) EZR による等分散性の検定の計算

ここでは, 1.3.1 節で説明した神経障害性疼痛のデータ(VAS_comp.csv)を用いて, EZR での計算方法について述べる. このとき, 新薬(Active)と既存薬(Control)で VAS の減少量の母分散が異なるか評価する(したがって, 両側対立仮説になる).

等分散性の検定の実行
1: 「統計解析」→「連続変数の解析」→「2 群の等分散性の検定(F 検定)」を選択する.
2: 次のようなメニューが表示される.



このとき,

- 「目的変数(1 つ選択)」で「VAS」を選択する.
- 「グループ(1 つ選択)」で「Group」を選択する.
- 「対立仮説」で「両側」を選択する.

3: 「OK」ボタンを押す

ここで,「対立仮説」は, 3 種類の対立仮説を表しており[5],

「両側」: 両側対立仮説(新薬(Active)と既存薬(Control)で VAS 減少量の母分散が異なる).

「差＜0」: 片側対立仮説(新薬(Active)の母分散ほうが既存薬(Control)の母分散よりも VAS 減少量が小さい).

「差＞0」: 片側対立仮説(新薬(Active)の母分散のほうが既存薬(Control)の母分散よりも VAS 減少量が大きい).

[5] EZR では, 母分散を差で表していたが, F 検定は母分散の比を検定する方法であり, EZR の記載は誤りである.

である. また, 信頼水準(デフォルト 0.95)とは, 信頼区間の信頼係数を表しており, 0.95 の場合には, 母分散の比に対する 95%信頼区間が描写される. 等分散性の検定(F 検定)では, 母分散の比を用いるため, その信頼区間も分散の比に対して構成される.

このとき, 次のような出力が表示される.

```
F 検定 P 値 = 0.18
```

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される p 値が 0.18 なので, 有意水準 0.05 のもとで有意でない. よって, 新薬と既存薬のあいだで VAS 減少量の分散に違いが認められるとはいえなかった.

なお, 分散の差(Active の分散／Control 分散)に対する 95%信頼区間(メニューから信頼水準(信頼係数)を 0.95 としている)は,「出力」を上にスクロールしたときの R での出力

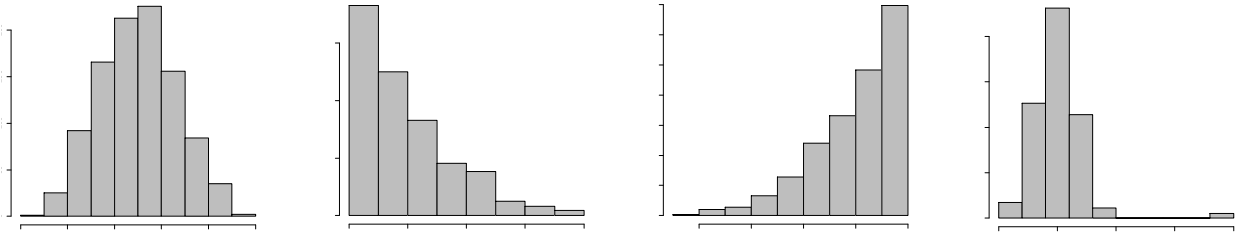
```
          F test to compare two variances

data:  VAS by Group
F = 0.45631, num df = 13, denom df = 11, p-value = 0.1801
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1345358 1.4590462
sample estimates:
ratio of variances
          0.456309
```

の太字部分に表示されている. 分散の比は,「ratio of variances」(因みに, F(F 値) も同じである)で表されており, 0.45631 であり, その 95%信頼区間は, [0.1345358, 1.4590462]である. この信頼区間が 1 を含むことからも, 新薬と既存薬のあいだで VAS 減少量の母分散に有意な違いがないことがわかる.

1.3.4 2 標本におけるノンパラメトリック検定 (Mann-Whitney U 検定)

(1) Mann-Whitney U 検定(Wilcoxon 検定)

下図は, ヒストグラムに対する幾つかのパターンである.

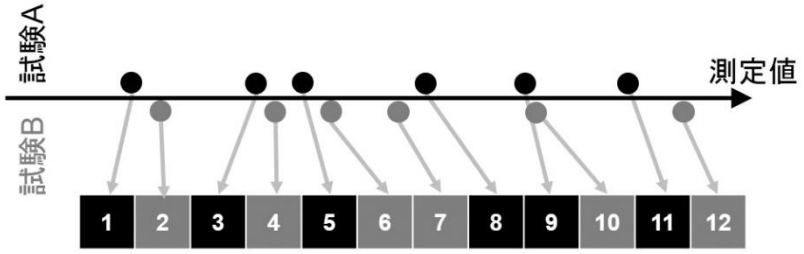


(a)　(b)　(c)　(d)

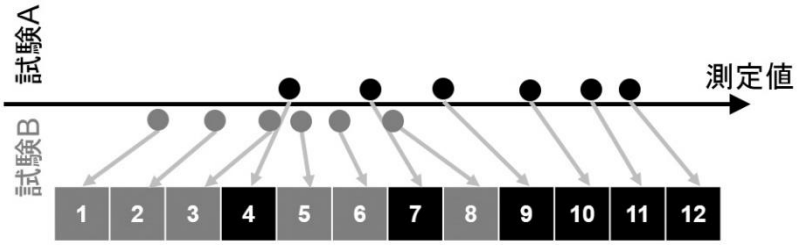
(a)は左右対称な分布形状を示しており, 正規分布に従っていることが示唆される. これに対して, (b)及び(c)は著しく歪んだ分布形状を示している. また, (d)は左右対称な分布形状を示しているものの, 外れ値(異常値)が示唆される. つまり, (b)～(c)では正規分布に従っていない可能性が高く, (d)では平均値が外れ値の影響を受ける可能性がある(つまり, 平均値がデータ全体を代表する値とは言えない). このような場合には, 正規分布に従わない場合に用いることができる検定, すなわち, ノンパラメトリック検定[6]を利用する.

2 群を比較する場合に用いられるノンパラメトリック検定のなかで, 最も代表的なものが Mann-Whitney U 検定(Wilcoxon 検定, Mann-Whitney-Wilcoxon 検定ともいう)である. 図 1.3 は, 2 試験(試験 A, 試験 B)に対する Mann-Whitney U 検定のイメージ図である. ここで, 丸印はアウトカムの位置を表しており, 下側の四角形のなかの数字は 2

[6] 正規分布に基づく検定(厳密には何らかの確率分布に基づく検定)をパラメトリック検定, そうでない場合をノンパラメトリック検定という.



(a) 有意でない場合の例示



(b) 有意である場合の例示

図 1.3: Mann-Whitney U(Wilcoxon)検定のイメージ図(四角印はアウトカムを表している)

試験の結果を併合して昇順に並べ替えた場合の順位を表している. Mann-Whitney U 検定の結果が有意でない場合, 順位を表す四角形のなかの, 試験 A のアウトカム(黒)と試験 B(灰色)が交互に出現している. これに対して, 有意である場合, 左側に試験 B(灰色)が並んでおり, 右側に試験 A(黒)が並んでいる. Mann-Whitney U 検定とは, この順位のコントラストに基づいて検定している.

すなわち, Mann-Whitney U 検定とは, 中央値を比較しているのではなく, 2 つの母集団の相対的な位置関係を比較している. したがって, 帰無仮説 H_0「2 つの母集団は同じである」に対して, 3 種類の対立仮説は

両側対立仮説 H_{1a}:2 つの母集団は異なる.

片側対立仮説 H_{1b}:母集団 1 の相対的な位置関係は, 母集団 2 よりも大きい.

片側対立仮説 H_{1c}:母集団 1 の相対的な位置関係は, 母集団 2 よりも小さい.

である.

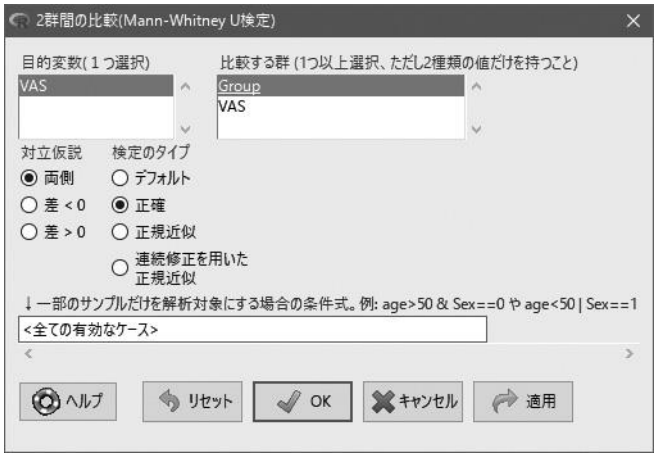
(2) 余禄:ノンパラメトリック検定における p 値

ノンパラメトリック検定には, 数学的な近似を用いて p 値を計算する方法(近似法)と検定統計量から確率的に正確に計算する方法(正確法)の 2 種類が存在する. 被験者数 n が少数の場合には, 正確法による p 値(exact p-value)を用いるべきであるが, 被験者数が増加するにつれて近似法と正確法の p 値はほぼ一致する. (統計ソフトウェアによって異なるが)被験者数が 200 以上になると, 正確法による計算負荷が膨大になるため, コンピュータがオーバーフロー(計算不可能)になる恐れがある. そのため, 近似法の結果を用いたほうが良い.

(3) EZR による Mann-Whitney U 検定(Wilcoxon 検定)の実行

ここでは, 1.3.1 節で説明した神経障害性疼痛のデータ(VAS_comp.csv)を用いて, EZR での計算方法について述べる. このとき, 新薬(Active)と既存薬(Control)で VAS の減少量の分布の相対的な位置関係が異なるか評価する(したがって, 両側対立仮説になる).

Mann-Whitney U 検定の実行
1:　「統計解析」→「ノンパラメトリック検定」→「2 群間の比較(Mann-Whitney U 検定)」を選択する.
2:　次のようなメニューが表示される.



このとき,

- 「目的変数(1 つ選択)」で「VAS」を選択する.
- 「グループ(1 つ以上選択、ただし 2 種類の値だけを持つこと)」で「Group」を選択する.
- 「対立仮説」で「両側」を選択する.
- 「検定のタイプ」で「正確」を選択する.

3:　「OK」ボタンを押す

ここで,「対立仮説」は, 3 種類の対立仮説を表しており,

「両側」：両側対立仮説(新薬(Active)と既存薬(Control)で母集団が異なる).

「差＜0」：片側対立仮説(新薬(Active)の母集団ほうが既存薬(Control)の母集団よりも相対的な位置が小さい).

「差＞0」：片側対立仮説(新薬(Active)の母集団ほうが既存薬(Control)の母集団よりも相対的な位置が大きい.

また,「検定のタイプ」は, p 値の計算方法を表しており, 症例数が小さい場合には,「正確」, それ以外の場合には,「正規近似」あるいは「連続修正を用いた正規近似」を選択したほうがよい(連続修正とは, 正規分布での近似を補正したものであるが, 症例数が多い場合にはほぼ同じになる).

このとき, 次のような出力が表示される.

```
                最小    25% メディアン   75% 最大     P 値
Group=Active    20   24.25         26 29.0   31 0.0215
Group=Control   13   18.00         22 25.5   29
```

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される. p 値が 0.0215 なので, 有意水準 0.05 のもとで有意であった. よって, 新薬と既存薬のあいだで VAS 減少量の相対的な位置関係に違いが認められた.

なお, これらの出力とは別に, 箱ひげ図が表示されるが, 図 1.2(b)と同じ出力結果なので割愛する.

1.3.5 パラメトリック検定とノンパラメトリック検定の取捨選択

臨床試験では, 平均値に基づいて試験デザイン(症例設計)を行うことが多い. そのため, パラメトリック検定を用いて評価することが原則になる. 一方で, 観察研究では, アウトカムが著しく正規分布から外れた場合にはノンパラメトリック検定の選択が考えられる. 研究論文では, 仮説検定による主解析の後続解析として, 多変量解析(重回帰分析等)を用いることがある. ただし, 重回帰分析は, アウトカムが正規分布に従うことを仮定しているため, ノンパラメトリック検定でアウトカムを比較したあとで重回帰分析を用いるのは理論的に整合性がとれない. したがって, ノンパラメトリック検定を用いる場合には, 各要因に関して, アウトカムへの影響を個別に評価を行うことになる.

また，研究結果を 2 標本 t 検定と Mann-Whitney U 検定の両方で検定した場合，2 標本 t 検定では有意であるにも関わらず，Mann-Whitney U 検定では有意でないことがある．このような状況が起こり得ることとしては，(1) 外れ値が存在する場合，(2) アウトカムが著しく歪んでおり正規分布に従わない場合，が考えれれる．これらの場合には，Mann-Whitney U 検定での p 値を採用すべきである．一方で，上記(1)(2)でない場合には，2 標本 t 検定の結果を採用することが推奨される．なぜなら，2 標本 t 検定のほうが Mann-Whitney U 検定に比べて検出力(群間に違いがあるときに正しく違いがあると示すことができる確率)が一般的に高く，解釈がしやすいためである．

1.4 対応があるデータに対する統計的推測

医学系研究において「比較」を考えるとき，2 種類のデータの取得方法がある．一つは，被験者をランダムに 2 群に分け，それぞれの群に対して異なる介入を行なう場合(無作為化比較試験)や，あるいは，暴露要因が異なる 2 群を比較する場合(コホート研究等)などである．この場合には，それぞれの群を構成する被験者が異なる．このようなデータを独立 2 標本といい，1.3 節で述べた統計手法を用いて比較を行う．

もう一つは，介入前後でのアウトカム(検査値やアンケート調査)の変化を比較する場合や，2 種類の検査を同一被験者に実施して，検査結果の違いを比較する場合である．アウトカムが同一被験者からとられることを，対応のある場合，あるいはマッチドペアという．ここでは，対応のある場合の評価方法について述べる．

1.4.1 データの概要：助産師に対するアンケート・データ

助産師が 5 年間の経験で分娩介助についてどのような意識の変革を起こすかを調べるため，資格取得直後と 5 年後に，分娩介助に関する 20 項目を自己評価してもらう研究が行われた(柳川他，2011[7])．

直後	84	78	76	82	68	64	78	66	72	64
	74	78	78	88	78	82	84	82	88	78
5 年後	88	70	80	94	72	68	82	78	72	70
	78	76	76	98	76	94	82	82	90	72

資格取得後と 5 年間の経験後で，助産師の意識の差に違いがあるだろうか．このデータは，Midwife.csv である．

1.4.1 対応のある t 検定

(1) 対応のある t 検定の概要

対応のある t 検定は，対応のある場合(マッチドペア)のアウトカムを比較する場合に用いられる．対応のある t 検定では，被験者毎のアウトカムの差の平均が 0 であるか否かを検討する．すなわち，対応のある t 検定とは，アウトカムの差が観測値である場合の 1 標本 t 検定と見做すことができる．

いま，被験者 i の 2 つのアウトカム(アウトカム 1: x_{1i}，アウトカム 2: x_{2i})の差 Δ_i を $\Delta_i = x_{1i} - x_{2i}$ とする．このとき，対応のある t 検定では，帰無仮説 H_0「アウトカムの差の母平均 $\bar{\Delta}$ は 0 である(2 つのアウトカム間に違いはない)」に対する評価を行う．このとき，対立仮説は以下の 3 種類

両側対立仮説 H_{1a}:アウトカムの差の母平均 $\bar{\Delta}$ は 0 ではない．

片側対立仮説 H_{1b}:アウトカムの差の母平均 $\bar{\Delta}$ は 0 よりも大きい(アウトカム 1 のほうが大きい)．

片側対立仮説 H_{1c}:アウトカムの差の母平均 $\bar{\Delta}$ は 0 よりも小さい(アウトカム 1 のほうが小さい)．

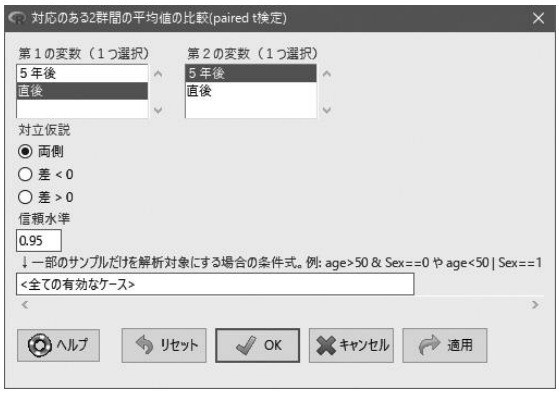
である．

[7] 柳川 堯・西 晃央・椛 勇三郎・堤 千代:看護・リハビリ・福祉のための統計学，近代科学社，2011.

(2) EZR による対応のあるｔ検定の計算

助産師に対するアンケート・データでの関心は，「資格取得直後と 5 年後の分娩介助アンケートのスコア(以下，スコア)に変化があるか否か)」にある．つまり，個々の被験者に対して，資格取得直後と 5 年後のスコアの差を計算し，その平均値が 0 に近くなければ変化したと考えることができる．従って，対応のあるｔ検定における帰無仮説「資格取得直後と 5 年後のスコアの差の母平均は 0 である(資格取得直後と 5 年後のスコアに変化がない)」に対して「資格取得直後と 5 年後のスコアの差の母平均は 0 でない(資格取得直後と 5 年後のスコアに変化ある)」を計算する．対応のあるｔ検定の手順を以下に示す．

対応のあるｔ検定の検定の実行
1:　「統計解析」→「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)」を選択する．
2:　次のようなメニューが表示される．



このとき，
- 「第 1 の変数(1 つ選択)」で「直後」を選択する．
- 「第 2 の変数(1 つ選択)」で「5 年後」を選択する．
- 「対立仮説」で「両側」を選択する．
- 「信頼水準」で 0.95 を入力する．

なお，変数の差は，「第 1 の変数」―「第 2 の変数」で計算される．

3:　「OK」ボタンを押す

ここで，「対立仮説」は，3 種類の対立仮説を表しており，

「両側」：両側対立仮説(直後と 5 年後で助産師の意識の差に違いがある)．

「差＜0」：片側対立仮説(直後のほうが 5 年後よりも助産師の意識が低い)．

「差＞0」：片側対立仮説(直後のほうが 5 年後よりも助産師の意識が高い)．

である．

また，信頼水準(デフォルト 0.95)とは，信頼区間の信頼係数を表しており，0.95 の場合には，アウトカムの差の母平均に対する 95%信頼区間が描写される．

このとき，次のような出力が表示される．

```
        平均 標準偏差     P 値
直後    77.1 7.239511  0.0421
５年後 79.9 8.837123
```

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される．．p 値が 0.0421 なので，有意水準 0.05 のもとで有意である．よって，直後と 5 年後で助産師の意識が変化していることがわかった．

なお，直後と 5 年後での助産師の意識の差(直後－5 年後)に対する 95%信頼区間(メニューから信頼水準(信頼係数)を 0.95 としている)は，「出力」を上にスクロールしたときの R での出力

```
        Paired t-test

data:  Dataset$直後 and Dataset$５年後
t = -2.1794, df = 19, p-value = 0.04209
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.4889668 -0.1110332
sample estimates:
mean of the differences
                  -2.8
```

の太字部分に表示されている. 直後と 5 年後での助産師の意識の差は,「mean of the differences」で表されており, -2.8 であった. したがって, 5 年後のほうが直後に比べて意識が上昇していた. このときの, 95%信頼区間は, [-5.4889668, -0.1110332]である. この信頼区間が 0 を含むことからも, 直後と 5 年後での助産師の意識に有意な変化が認められる.

1.4.2 Wilcoxon 符号付き順位検定

(1) Wilcoxon 符号付き順位検定の概要

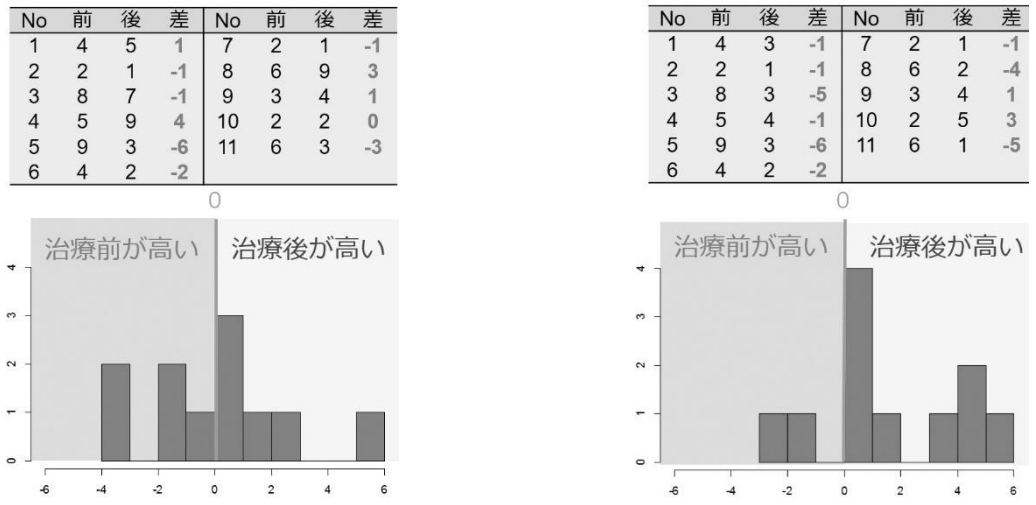
対応のある t 検定では, 被験者毎のアウトカム(アウトカム 1: x_{1i}, アウトカム 2: x_{2i})の差 $\Delta_i = x_{1i} - x_{2i}$ (事例の場合には, 資格取得 5 年後と取得直後でのスコアの変化の大きさ)を計算したうえで, その平均値が 0 であるか否かを評価している. すなわち, 被験者毎のアウトカムの差 Δ_i が正規分布に従っていることが仮定される. 一方で, 正規分布に従っていない場合には, ノンパラメトリック検定の一つである, Wilcoxon 符号付き順位検定(Wilcoxon 符号付き順位和検定)を利用することができる.

Wilcoxon 符号付き順位検定では, 被験者毎の 2 つのアウトカム間の差 Δ_i の正負を用いる. このとき, 帰無仮説 H_0:「2 つのアウトカムのあいだに違いがない」に対して, 対立仮説は以下の 3 種類である:

両側対立仮説 H_{1a}:2 つのアウトカムのあいだには違いがある.

片側対立仮説 H_{1b}:アウトカム 1 のほうがアウトカム 2 よりも高い.

片側対立仮説 H_{1c}:アウトカム 1 のほうがアウトカム 2 よりも低い.



(a) 有意でない場合　　(b) 有意な場合

図 1.4:Wilcoxon 符号付順位検定の概念図

(上側:データ, 下側:ヒストグラム(X 軸は治療前後の差, Y 軸は被験者数である))

図 1.4 は，11 名の被験者に対する治療前後でのアウトカムを比較した仮想例である(上側:データ，下側:アウトカムの差に対するヒストグラム)．ここでのアウトカムの差 Δ_i は，治療前−治療後を表している．したがって，ヒストグラムにおいて，0 よりも左側の被験者は治療前のほうが治療後に比べて高く，右側の被験者は治療後のほうが治療前に比べて高い．

Wilcoxon 符号付き順位検定が有意でないとき(棄却できないとき)，アウトカムの差 Δ_i が負値の被験者数と正値の被験者数がほぼ同じになる(図 1.4(a))．一方で，棄却できるとき(有意であるとき)，アウトカムの差 Δ_i が負値の被験者数と正値の被験者数がアンバランスになる(図 1.4(b))．Wilcoxon 符号付き順位検定の p 値は，負値の被験者数と正値の被験者数のアンバランス具合に基づいて計算される．

(2) EZR による対応のある t 検定の計算

助産師に対するアンケート・データでの関心は，「資格取得直後と 5 年後の分娩介助アンケートのスコア(以下，スコア)に変化があるか否か)」(両側対立仮説)にある．Wilcoxon 符号付き順位検定の手順を以下に示す．

Wilcoxon 符号付順位検定の実行
1: 「統計解析」→「ノンパラメトリック検定」→「対応のある 2 群間の平均値の比較 (Wilcoxon 符号付順位和検定)」を選択する． 2: 次のようなメニューが表示される．    このとき， ・「第 1 の変数(1 つ選択)」で「直後」を選択する． ・「第 2 の変数(1 つ選択)」で「5 年後」を選択する． ・「対立仮説」で「両側」を選択する． ・「検定のタイプ」で「正確」を選択する． 3: 「OK」ボタンを押す

ここで，「対立仮説」は，3 種類の対立仮説を表しており，

「両側」：両側対立仮説(直後と 5 年後で助産師の意識に違いがある)．

「差＜0」：片側対立仮説(直後のほうが 5 年後よりも助産師の意識が低い)．

「差＞0」：片側対立仮説(直後のほうが 5 年後よりも助産師の意識が高い)．

である．「検定のタイプ」は，p 値の計算方法を表しており，症例数が小さい場合には，「正確」，それ以外の場合には，「正規近似」あるいは「連続修正を用いた正規近似」を選択したほうがよい(連続修正とは，正規分布での近似を補正したものであるが，症例数が多い場合にはほぼ同じになる)．

このとき，次のような出力が表示される．

```
対応のある 2 群間の比較(Wilcoxon 符号付順位和検定) P 値 = 0.0414
```

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される. . また, 上側に青色のアウトプット(英語表記)があるが, これは, R での解析結果であり, 同じことを意味することから, 無視してよい. p 値が 0.0414 なので, 有意水準 0.05 のもとで有意であった. よって, 直後と 5 年後で助産師の意識に違いが認められた.

1.5 分散分析

ここでは, 様々な分散分析の方法について述べる. 図 1.5 は, 様々な試験デザインと分散分析モデルの関係を表している. 図 1.5(a)は, アウトカム(痛みの程度)に対して要因(薬剤)が 1 個である. このような場合には, 一元配置の分散分析, あるいはそのノンパラメトリック検定である Kruskal-Wallis 検定を用いる.

図 1.5(b)は, ある薬剤を投与した時の経時的な痛みの程度の変化を評価している. このときの関心は, 薬剤投与によって痛みの程度が経時的に変化しているかどうかを評価することにある. このような場合には, 繰り返し測定の分散分析, あるいはそのノンパラメトリック検定である Friedman 検定を用いる.

図 1.5©は, 図 1.5(b)と同様に痛みの程度の経時的な変化を評価している. ただし, この場合には, 2 種類の薬剤の効果を比較している. そのため, 2 つの関心, すなわち, (1)痛みの程度が薬剤によって異なるか, (2)痛みの程度の経時的変化が薬剤によって異なるか, がある. このような場合においても, 繰り返し測定の分散分析を用いることができる. 一方で, ノンパラメトリック検定は存在しない.

図 1.5(d)は, アウトカム(痛みの程度)に対して, 複数の要因(薬剤, 年齢)が存在する場合である. このような場合には, 2 元配置の分散分析を用いる. 一方で, ノンパラメトリック検定は存在しない.

1.5.1 一元配置の分散分析

(1) データの概要:3 種類の疼痛薬のデータ

いま, 14 名の疼痛患者が服薬した除痛薬(A,B,C)毎にグループに分け分け, それぞれの群での投与後の痛みの程度を測定した.

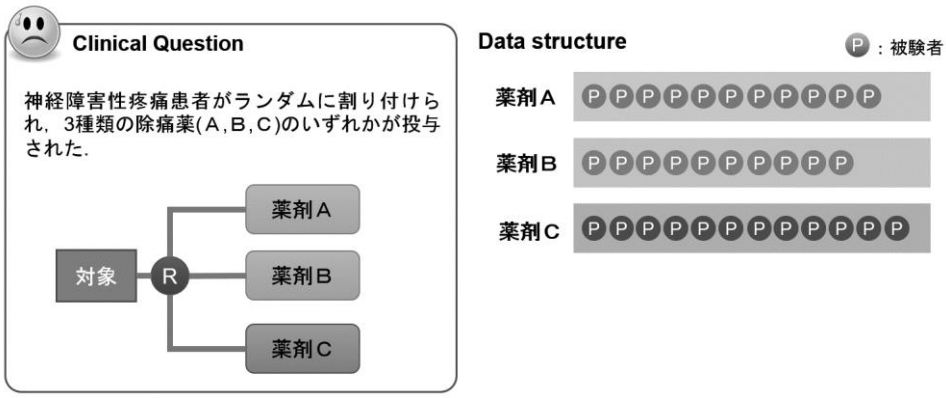
薬 A	7.69	9.69	8.89	6.94	2.13	7.26	5.87	7.20	8.18
	7.24	6.81	6.67	6.98	7.07	7.00	7.00	5.00	8.00
薬 B	12.90	16.60	8.35	9.81	7.84	3.84	9.42		
薬 C	12.40	14.00	11.60	12.20	13.90	9.41	11.20	2.40	

除痛薬によって痛みの程度に違いがあるだろうか. このデータは, Analgesics.csv である.

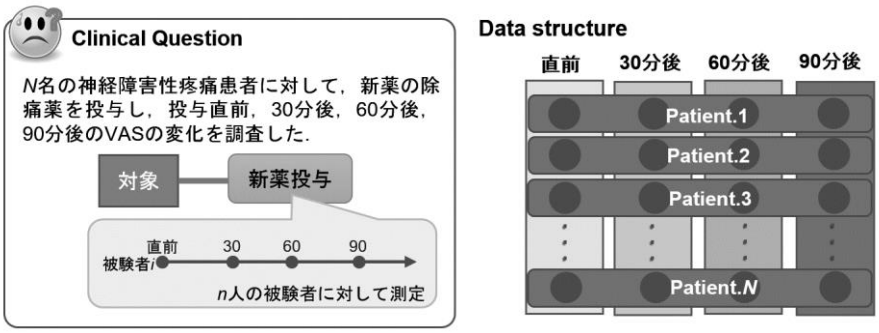
(2) 一元配置の分散分析の概要

4 種類の薬剤に対する臨床試験の例を挙げる. この臨床試験では, 被験者を 4 群に分け, それぞれに対して 4 種類の薬剤(薬剤 A, 薬剤 B, 薬剤 C, 薬剤 D)のいずれかを投与しており, 投与前後での検査値がアウトカムとしてとられている. このとき, 分散分析では, 要因(薬剤)のことを因子(factor)と呼び, 因子を分ける条件(薬剤の種類)を水準(level)という. つまり, 本事例は, 1 因子 4 水準の分散分析である. そして, 1 因子の場合に用いる分散分析法が, 一元配置の分散分析である.

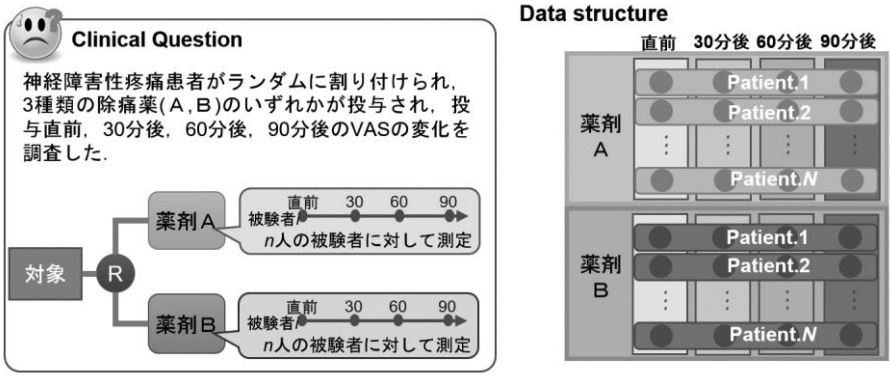
一元配置の分散分析では, 帰無仮説 H_0「水準間(群間)の平均がすべて等しい」に対して, 対立仮説 H_1「帰無仮説 H_0 ではない」が評価される. 図 1.6 は, 一元配置の分散分析における対立仮説 H_1 が正しい(有意である)場合に想定される状況である. ここで, 図 1.6 の μ は各群の母集団における平均である. いずれの状況も平均のバラツキが, 各群の観測値のバラツキに比べて大きいことがわかる.



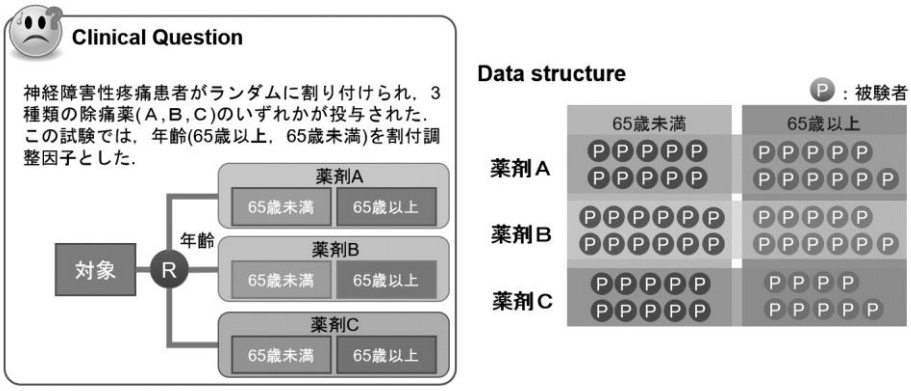
(a)一元配置の分散分析(ノンパラメトリック検定は Kruskal–Wallis 検定)



(b)繰り返し測定の分散分析(ノンパラメトリック検定は Friedman 検定)

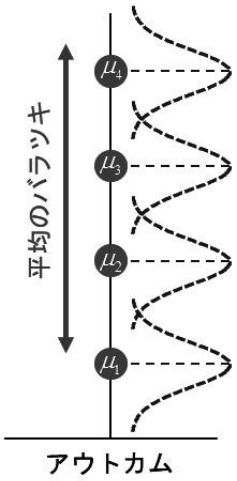


(c)繰り返し測定の分散分析(ノンパラメトリック検定は存在しない)

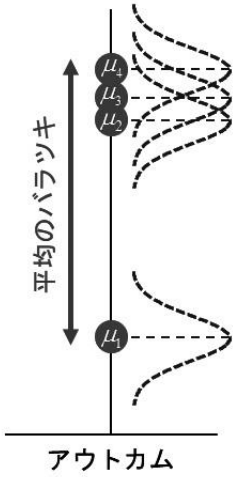


(d)2 元配置の分散分析(ノンパラメトリック検定は存在しない)

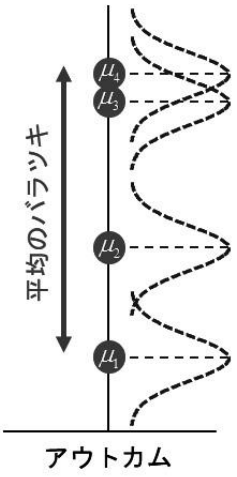
図 1.5：様々なシチュエーションでの分散分析の諸型



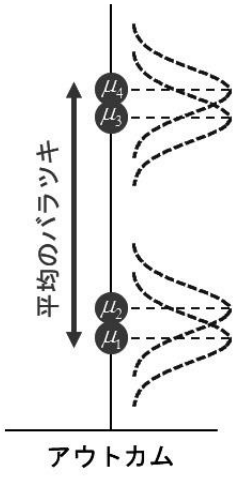
(a)すべて群が離れている



(b)1 群のみ離れている



(c)2 群が離れている



(d)2 群づつ分かれている

図 1.6:一元配置の分散分析において，対立仮説 H_1 が正しい(有意である)状況

一元配置の分散分析では，平均の分散が観測値の分散に対して大きいときに有意である(帰無仮説 H_0 を棄却できる)と判断される．ここで，一元配置の分散分析では，平均の違いを取り扱うことから，すべての群(水準)が同じ分散の正規分布に従うことが仮定されることに注意されたい．因みに，2 標本の場合の平均の分散は，平均の差と同じであり，2 標本 t 検定は，2 水準の一元配置の分散分析と同じになる．

(3) 多重比較の方法

図 1.7 は，128 症例をランダムに 2 群に分け，同じ薬剤を投与する臨床試験をシミュレーションによって 200 回実施したときの試験番号(Trial Number)と 2 標本 t 検定の p 値(p-value)を表している(X 軸:試験番号，Y 軸:p 値)．ここで，横方向の点線は有意水準 0.05 を表している．2 群には同じ薬剤が投与されているので，本来は効果に違いがない．それにも関わらず，10 個の試験で有意差が認められている．

有意水準 0.05 は，帰無仮説が真実であったとしても，5%の確率で有意であると誤ってしまうことを意味する(そのため，有意水準 α は第 1 種の過誤あるいは α エラーと呼ばれる)．図 1.7 において，同じ薬剤を投与した臨床試験であるにも関わらず，5%の確率(10/200)で有意差が認められたのはそのためである．

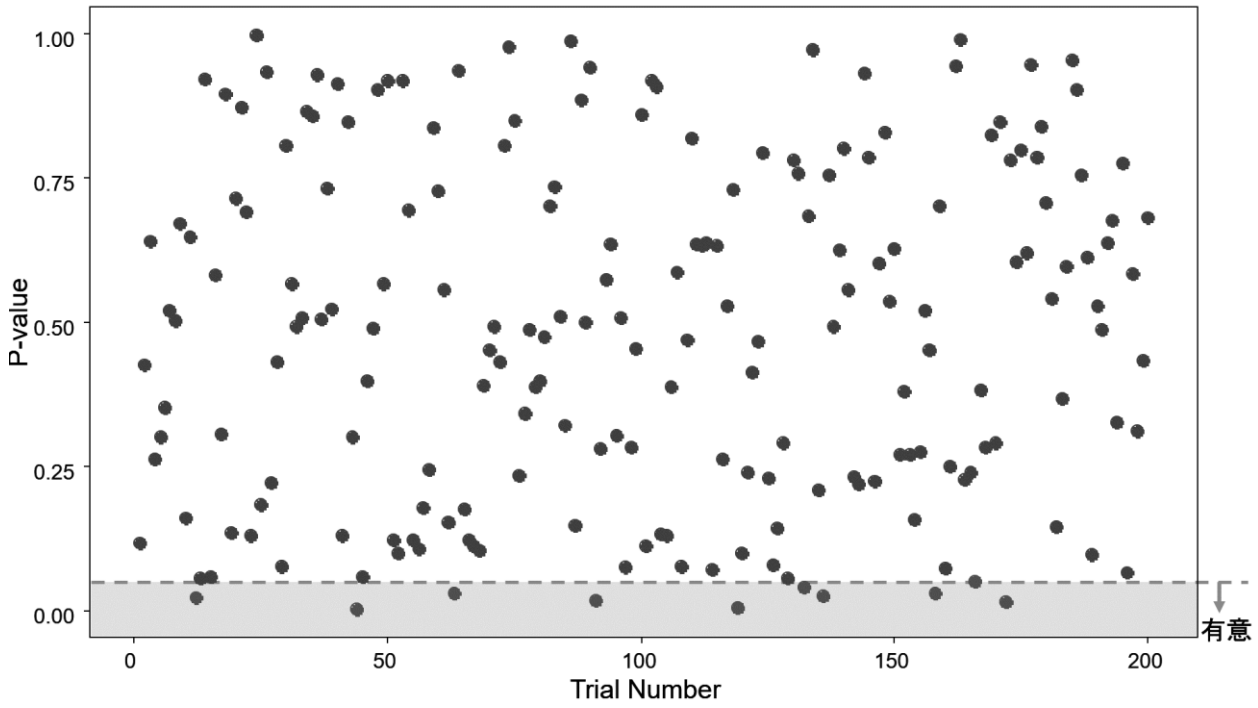


図 1.7：臨床試験のシミュレーションにおける結果

先ほどの 4 剤の効果を比較するとき，全てのパターンで対比較するには，6 回の検定(A vs B，A vs C，A vs D，B vs C，B vs D，C vs D)が必要になる．この比較を有意水準 0.05 で検定した場合，4 剤における(真実の)平均効果が同じであったとしても，26.5%の確率でいずれかの検定が有意になる．つまり，もともとは有意水準 0.05 で比較していたとしても，「下手な鉄砲も数打てば当たる」効果で誤りの確率が増加している．このような状況に対処するための方法が多重比較である．

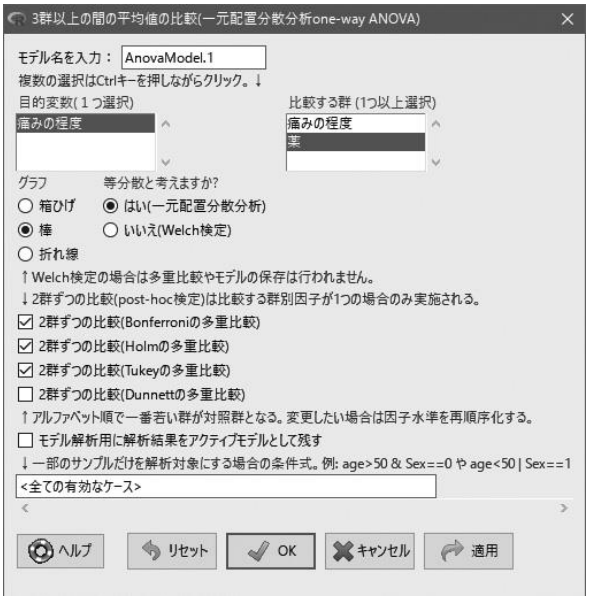
分散分析に対する多重比較には，p 値を調整する方法と分散分析の結果を数理的に展開する方法がある．前者は，検定(一元配置の分散分析では，2 標本 t 検定)で得られた p 値を調整するだけなので，ざまざまな検定に適用することができる．EZR では，Bonferroni の多重比較，Holm の多重比較がこれに該当する．Bonferroni の多重比較は，最も有名な方法の一つである．Bonferroni の多重比較の利点は，調整 p 値(多重比較によって調整された p 値)が，「各検定の p 値×比較回数」で計算できることから，非常に柔軟で単純なことにある．一方で，検定回数(すなわち，群数)が多くなるほど，調整 p 値が有意になりにくくなる傾向にある．p 値が有意になりにくくなる傾向を改善したものが Holm の方法である．

後者の方法では，分散分析の結果と多重比較が対応付けられている方法や，あるいは特定のシチュエーションを想定した方法などがある EZR では，Tukey の多重比較，Dunnett の多重比較がこれに該当する．Tukey の多重比較は，ペアワイズに多重比較のもとで母平均を比較する場合である．Bonferroni の多重比較あるいは Holm の多重比較では，一元配置の分散分析で有意であるものの，多重比較では有意な結果が得られないことがある．このような場合には，どの群間に違いがあるかを判断できない．これに対して，Tukey の多重比較では，一元配置の分散分析が有意だった場合に，いずれかの群間に有意差が認められる．したがて，Tukey の多重比較と一元配置の分散分析を対応付けて解釈できる．

Dunnett の多重比較とは，コントロール群に対して，2 剤(治療)以上の試験群が存在する場合に用いられる．そこでは，コントロール群と(複数の)試験群間の違いを検定することができる．

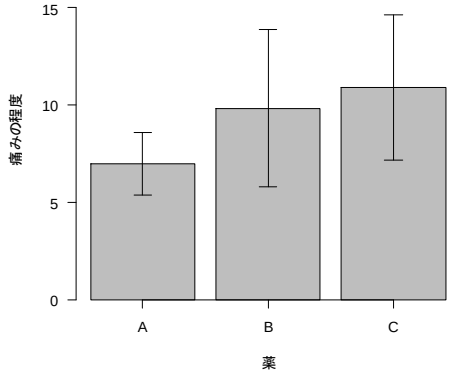
(4) EZR による一元配置の分散分析及び多重比較の実行

一元配置の分散分析の関心は,「3 種類の薬剤(A,B,C)の除痛効果の平均に違いがあるか」にある. 因みに, 一元配置の分散分析には, 両側対立仮説, 片側対立仮説はない. 一元配置の分散分析の手順を以下に示す.

一元配置の分散分析の実行
1: 「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)」を選択する. 2: 次のようなメニューが表示される.  このとき, ・「目的変数(1 つ選択)」で「痛みの程度」を選択する. ・「比較する群(1 つ以上選択)」で「薬」を選択する. ・「グラフ」で「棒」を選択する. ・「等分散と考えますか？」を「はい(一元配置分散分析)」を選択する. ・「2 群づつの比較(Bonferroni の多重比較)」, 「2 群づつの比較(Holm の多重比較)」, 「2 群づつの比較(Tukey の多重比較)」にチェックを入れる. 3: 「OK」ボタンを押す

ここで,「モデル名を入力」とは複数の分散分析モデルを比較するのに用いることができるが, 自動的に名前が割り振られるため, 無視して問題ない.「等分散を仮定しますか？」とは, 1.3.2 節の場合と同様に, Welch 検定を用いるか否かを表している. ただし, 2 標本 t 検定と同様に, 3 群以上の場合にも Welch 検定を用いることは殆どなく, そのような場合には, ノンパラメトリック検定である, Kruskal–Wallis 検定(1.5.2 節)を用いる.

上記では, 複数の多重比較を選択しているが, いずれもペアワイズ比較であり, その傾向を評価するのに複数を選択している. これに対して, Dunnett の多重比較では, コントロール群と複数の治療群の比較を実施するのに用いる. 例えば, コントロール群と 2 種類の新薬(新薬 A, 新薬 B)の場合, Dunnett の多重比較では, コントロール群 vs. 新薬 A, コントロール群 vs. 新薬 B の 2 種類の比較が評価される. EZR では, 変数名のアルファベットの頭文字が一番若いものをコントロール群と認識する. このときの EZR の結果では, 以下の棒グラフ



が表示される(箱ひげ図で表示したい場合には,「グラフ」で「箱ひげ」をチェックすればよい). ここで, エラーバーは, 標準偏差を表している. その結果, 薬剤 A の痛みの程度が最も低かった.

EZR の出力では, 様々な出力が表示される. 表示された青色の箇所毎に説明する.

Output.1

```
            Df Sum Sq Mean Sq F value Pr(>F)
factor(薬)   2  99.52   49.76   6.271 0.0053 **
Residuals   30 238.04    7.93
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output.1 は, 分散分析表と言われるものである. ここで, Df は自由度を表している. 自由度は, 因子(factor)の場合には水準数-1, 誤差(Residuals)の場合には症例数-水準数である. また, Sum は平方和を表している. 因子, 誤差それぞれの平方和は, 次のように定義される.

・因子:(3 群(水準)のそれぞれの平均値)ー(観測値全体の平均値)を 2 乗したときの総和

・誤差:(観測値)ー(観測値が属する群(水準)での平均値)を 2 乗したときの総和

また, 平均平方和(Mean Sq)は, 平方和(Sum)／自由度(Df)で計算される. したがって, 因子の平均平方和は平均の分散を表しており, 誤差の分散は観測値の分散を表している. つまり,「因子の平均平方和＞誤差の平均平方和」であれば, 有意であると結論付けられる.

これらの分散の違いを表しているのが F 値(F value)である. F 値は(因子の平均平方和)／(誤差の平均平方和)で計算され, 検定統計量(帰無仮説 H_0 が正しいと判断できる確率である p 値を計算するための測度)として用いられる. F 値から計算される p 値(Pr(>F))は, 上記の帰無仮説 H_0 に対して求められ, これまでの検定と同様に評価される.

その結果, p 値は, 0.0053 であることから, 有意水準 0.05 のもとで有意である. したがって, 3 種類の薬剤で痛みの程度の平均値に違いが認められている.

Output.2

```
          平均 標準偏差     P 値
薬=A  6.978889 1.590072 0.0053
薬=B  9.822857 4.030230
薬=C 10.888750 3.733996
```

Output.2 は, 各薬剤(水準)での平均値および標準偏差を表しており, P 値は一元配置の分散分析によって計算されたものであり, Output.1 の「Pr(>F)」と同じ数値になっている.

Output.3

```
          Pairwise comparisons using t tests with pooled SD

data:  Dataset$痛みの程度 and Dataset$薬

  A      B
B 0.0924 -
C 0.0082 1.0000

P value adjustment method: bonferroni
```

Output.3 は，Bonferroni の多重比較の結果である(太字の部分に多重比較の結果が表示されている)．ここで，対比較には 2 標本 t 検定が用いられている．薬剤 A vs. C のあいだで有意差が認められている．

Output.4

```
        Pairwise comparisons using t tests with pooled SD

data:  Dataset$痛みの程度 and Dataset$薬

  A      B
B 0.0616 -
C 0.0082 0.4704

P value adjustment method: holm
```

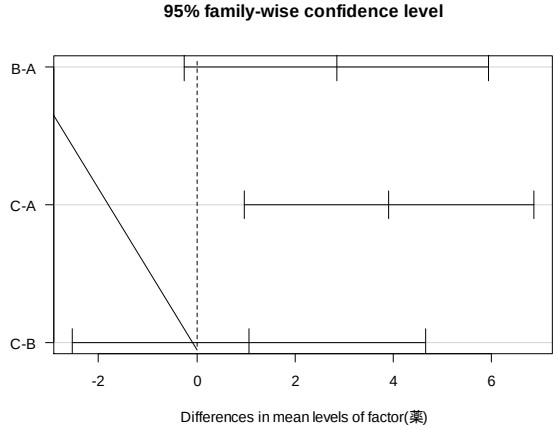
Output.4 は，Holm の多重比較の結果である(太字の部分に多重比較の結果が表示されている)．ここで，対比較には 2 標本 t 検定が用いられている．薬剤 A vs. C のあいだで有意差が認められている．Holm の多重比較は，Bonferroni の多重比較を修正したものであり，A vs. B 及び B vs. C の p 値が小さくなっていることが分かる[8]．

Output.5

```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = 痛みの程度 ~ factor(薬), data = Dataset, na.action = na.omit)

$`factor(薬)`
        diff        lwr      upr     p adj
B-A 2.843968 -0.2492548 5.937191 0.0763250
C-A 3.909861  0.9591144 6.860608 0.0074620
C-B 1.065893 -2.5281076 4.659893 0.7471241
```

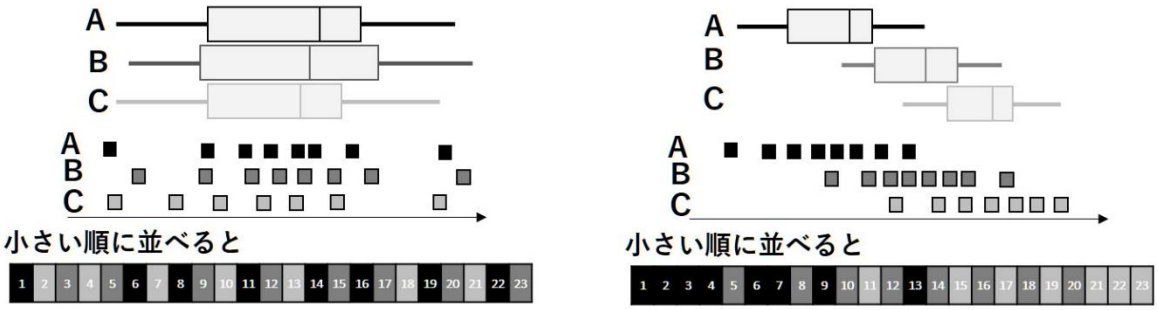
Output.5 は，Tukey の多重比較の結果である．ここで，diff とは群間の平均値の差(例えば，B-A であれば，薬剤 B の平均-薬剤 A の平均を表している)である．また，lwr 及び upr は，それぞれ，多重比較調整をおこなったときの平均値の差の 95%信頼区間である．したがって，この信頼区間が 0 を含まなければ有意になる．p adj は，Tukey の多重比較における p 値である．

また，Tukey の多重比較では，平均の差および 95%信頼区間をグラフ化したものが表示される．



ここで，エラーバーの中央の縦線は平均の差を表しており，信頼幅は，平均の差に対する 95%信頼区間を表している．

今回の場合には，すべての多重比較で薬剤 A vs. 薬剤 C のみ有意差が認められた．一方で，一元配置の分散分析では有意であるものの，Bonferroni の多重比較，および Holm の多重比較では有意差は認められない場合がある(つまり，一元配置の分散分析と多重比較で結果の不一致が起こる)．そのため，一元配置の分散分析の結果に基づ



(a) 有意な場合　　(b) 有意でない場合

図 1.8：Kruskal–Wallis 検定の概念図

いて，多重比較を実施する場合には，Tukey の多重比較を用い，群間比較のみを実施する場合には，Bonferroni の方法あるいは，Holm の方法を用いることが推奨される．

1.5.2 3 群以上でのノンパラメトリック検定：Kraskal-Wallis 検定

(1) Kraskal–Wallis 検定の概要

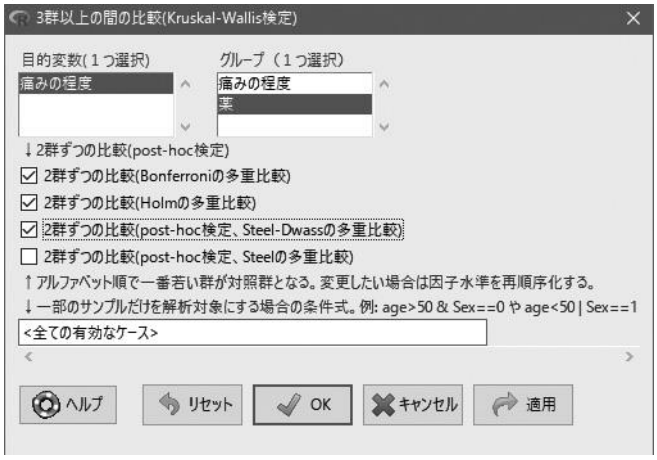
分散分析では，観測値が正規分布に従うことが仮定される．他方，医学系研究では観測値が正規分布に従っていない場合も少なくない．このような場合に用いることができるのが，ノンパラメトリック検定である．一元配置の分散分析に対するノンパラメトリック検定は，Kruskal–Wallis 検定である．Kruskal–Wallis 検定は，2 標本で用いられる Wilcoxon 検定と同様に観測値を小さい順に並べ替えたときの順位に基づいて検定する．

図 1.8 は，3 群を比較する場合の Kruskal–Wallis 検定の概念図である．ボックスプロットの下側の帯は，観測値を小さい順に並べ替え，群毎に色分けしたものである．有意でない場合(図 1.8(a))，それぞれの色がおおよそ交互に並んでいる．一方で，有意である場合(図 1.8(b))，左側に A 群，右側に C 群が集中している．Kruskal–Wallis 検定では，この偏りを評価しており，Wilcoxon 検定の拡張型と考えることができる．

(2) EZR による Kraskal–Wallis 検定の実行

Kraskal–Wallis 検定の関心は，「3 種類の薬剤(A,B,C)の除痛効果に違いがあるか」にある．因みに，Kraskal–Wallis 検定には，両側対立仮説，片側対立仮説はない．Kraskal–Wallis 検定の手順を以下に示す．

Kraskal-Wallis 検定の実行
1:　「統計解析」→「ノンパラメトリック検定」→「3 群以上の間の比較(Kraskal-Wallis 検定)」を選択する．
2:　次のようなメニューが表示される．



このとき，
- 「目的変数(1 つ選択)」で「痛みの程度」を選択する．
- 「グループ(1 つ選択)」で「薬」を選択する．
- 「2 群づつの比較(Bonferroni の多重比較)」，「2 群づつの比較(Holm の多重比較)」，「2 群づつの比較(post-hoc 検定、Steel-Dwass の多重比較)」にチェックを入れる．

3: 「OK」ボタンを押す

多重比較では，一元配置の分散分析の場合には表示されなかった，Steel-Dwass の多重比較と Steel の多重比較が表示される．Steel-Dwass の多重比較は，Tukey の多重比較のノンパラメトリック版であり，Steel の多重比較は，Dunnett の多重比較のノンパラメトリック版である．

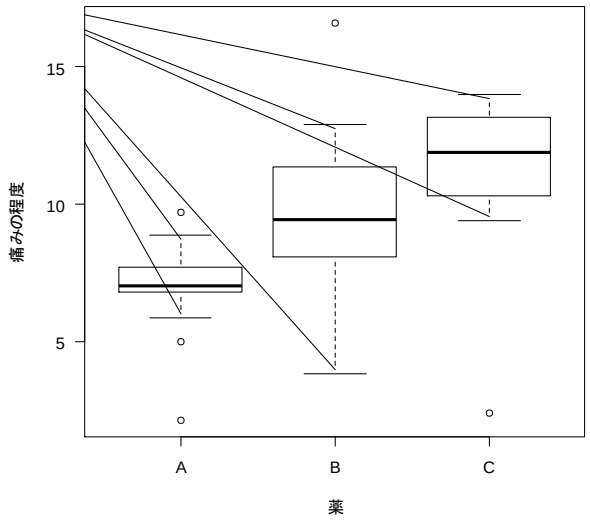
上記では，複数の多重比較を選択しているが，いずれもペアワイズ比較であり，その傾向を評価するのに複数を選択している．これに対して，Steel の多重比較では，コントロール群と複数の治療群の比較を実施するのに用いる．EZR では，変数名のアルファベットの頭文字が一番若いものをコントロール群と認識する．

EZR の出力では，様々な出力が表示される．表示された青色の箇所毎に説明する．

Output.1

```
      A       B       C
 7.035   9.420 11.900
```

Output.1 は，各群の中央値を表している．また，同時に次のような箱ひげ図



が表示される．薬 A の痛みの程度が最も低く，中央値が最小である．

Output.2

```
3 群以上の間の比較(Kruskal-Wallis 検定) P 値 = 0.00291
```

Output.2 は，Kraskal-Wallis 検定の結果である．p 値が 0.00291 であることから，有意水準 α=0.05 のもとで有意である．したがって，3 種類の薬剤における除痛効果に違いが認められる．なお，上側の青色のアウトプット(Kraskal-Wallis rank sum test)は，この出力と同じ意味なので無視してよい．

Output.3

```
            Pairwise comparisons using Mann-Whitney U test

data:  Dataset

  A      B
B 0.0594 -
C 0.0088 1.0000

P value adjustment method: bonferroni
```

Output.3 は，Bonferroni の多重比較の結果である(太字の部分に多重比較の結果が表示されている)．ここで，対比較には Mann-Whitney U(Wilcoxon)検定が用いられている．薬剤 A vs. C のあいだで有意差が認められている．

Output.4

```
           Pairwise comparisons using Mann-Whitney U test

data:  Dataset

  A      B
B 0.0396 -
C 0.0088 0.4630

P value adjustment method: holm
```

Output.4 は，Holm の多重比較の結果である(太字の部分に多重比較の結果が表示されている)．ここで，対比較には Mann-Whitney U 検定が用いられている．薬剤 A vs. C だけでなく，薬剤 A vs. C のあいだで有意差が認められている．

Output.5

```
             t           p
A:B 2.3608414 0.047870161
A:C 3.0005130 0.007595555
B:C 0.8100926 0.696761645
```

Output.5 は，Steel-Dwass の多重比較の結果である(太字の部分に多重比較の結果が表示されている)．薬剤 A vs. C だけでなく，薬剤 A vs. C のあいだで有意差が認められている．

1.5.3 繰り返し測定の分散分析

(1) データの概要:脳下垂体と翼突上顎裂の距離のデータ

ここでは，11 人の女の子の脳下垂体と翼突上顎裂の距離を 8 歳，10 歳，12 歳，14 歳の時点で比較する研究のデータを用いる(新谷，2016[9])．

ID	gt8	gt10	gt12	gt14
1	21.0	20.0	21.5	23.0
2	21.0	21.5	24.0	25.5
3	20.5	24.0	24.5	26.0
4	23.5	24.5	25.0	26.5
5	21.5	23.0	22.5	23.5
6	20.0	21.0	21.0	22.5
7	21.5	22.5	23.0	25.0
8	23.0	23.0	23.5	24.0
9	20.0	21.0	22.0	21.5
10	16.5	19.0	19.0	19.5
11	24.5	25.0	28.0	28.0

ここで，ID は被験者番号，gt8 は 8 歳のときの距離，gt10 は 10 歳のときの距離，gt12 は 12 歳のときの距離，gt14 は 14 歳のときの距離である．年齢によって脳下垂体と翼突上顎裂の距離に違いがあるだろうか．このデータは，dental_growth.csv に含まれている．

(2) 繰り返し測定の分散分析

いま，2 群比較(薬剤 1，薬剤 2)において，介入後の 3 時点(時点 1，時点 2，時点 3)でアウトカムがとられた状況を考える(アウトカムに変化量を用いるなど，介入前の値で調査委されていることとする)．このとき，次の分散分析モデル．

(アウトカム) = (薬剤や時間に依存しない効果)＋(薬剤による効果)＋(時間による効果)＋(誤差)

で得られる．

---

9 新谷歩: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，2016.

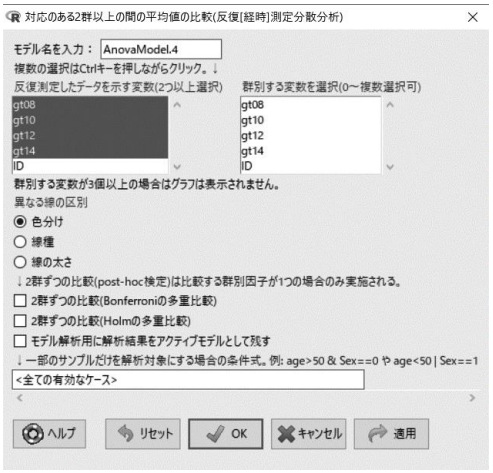
ただし，経時繰り返し測定データにおいて，このようなことは稀である．例えば，時点 1 と時点 2 のアウトカムには相関関係があることは，平易に理解できる．このような相関のことを系列相関(serial correlation)という．このような，系列相関のなかでも，すべての時点間の相関係数(すなわち分散)が等しいと仮定できる場合を球面性(sphericity)という．アウトカムが球面性の仮定を満さない場合には，Greenhouse & Geisser 法あるいは，Huynh-Feldt 法などを用いて分散分析の自由度を調整することができる．EZR では，これらの調整方法が利用されている．医学系研究において，繰り返し測定の分散分析(repeated measured ANOVA)と記載されているものの多くが，方法を採用している．

(3) EZR による繰り返し測定の分散分析の実行

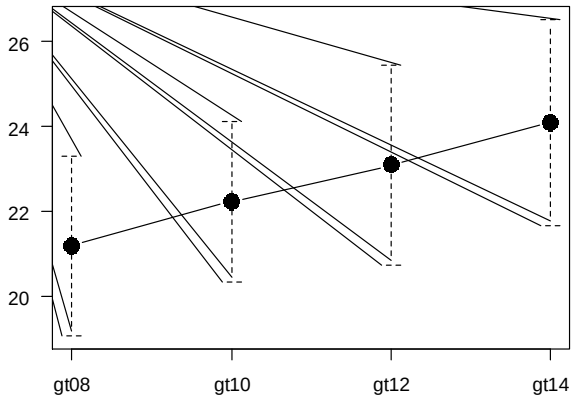
ここでは，脳下垂体と翼突上顎裂の距離のデータを用いて繰り返し測定の分散分析の適用方法を示す．ここでの関心は，「年齢によって脳下垂体と翼突上顎裂の距離に違いがあるか」にある．このとき，そのままの形式では，「g10」→「gt10」→「gt12」→「gt8」になってしまう(頭文字から若い順序で解釈されるためである)．そのため，「gt8」の変数名を「gt08」に変更する．

変数名の変更
1: 「アクティブデータセット」→「変数の操作」→「変数名を変更する」を選択する． 2: ウィンドウ「変数名を変更する」が表示されるので，「gt8」を選択して，「OK」ボタンを押す． 3: ウィンドウ「変数名」が表示されるので，新しい変数名「gt08」を選択して，「OK」ボタンを押す

次いで，繰り返し測定の分散分析を実行する．

繰り返し測定の分散分析の実行
1: 「統計解析」→「連続変数の解析」→「対応のある 2 群以上の間の平均値の比較(反復(継時)測定分散分析)」を選択する． 2: 次のようなメニューが表示される．    このとき， ・「反復測定したデータを示す変数(2 つ以上選択)」で「gt08」，「gt10」，「gt12」，「gt14」を選択する． なお，図 1.5(c)のように，治療・薬剤の経時的変化を比較する場合には，「群別する変数を選択(0〜複数選択可)」において選択する． 3: 「OK」ボタンを押す

EZR の出力では，様々な出力が表示される．また，各時点での平均値±標準偏差のウィスカー・プロット



が表示される. また, EZR の「出力」において, 表示された青色の箇所毎に説明する.

Output.1

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

                  SS num Df Error SS den Df        F    Pr(>F)
(Intercept) 22568.5      1  177.227     10 1273.419 7.090e-12 ***
Time           50.7      3   19.409     30   26.098 1.673e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output.1 は, 繰り返し測定の分散分析の結果である(ちなみに, EZR では Time を質的変数として扱われるので, 変数名の変更に依らず結果は同じになる). 「Time」が経時的変化に対する結果であり, 「Pr(>F)」が p 値を表している. その結果, p 値は 1.673×10^{-8}(e-08 とは 10^{-8} を表す)であることから, 成長に対して, 有意な変化が認められる

Output.2

```
Mauchly Tests for Sphericity

     Test statistic p-value
Time        0.69474  0.6745
```

Output.2 は, 球面性の検定(Machly の検定)の結果である. 球面性の検定では, 帰無仮説 H_0「球面性を満たす」に対して, 対立仮説 H_1「球面性を満たさない」を検定する. したがって, 球面性の検定の結果, 有意であるならば, 下側の Greenhouse-Geisser の方法あるいは Huynh-Felt の方法のいずれかの結果を用いる. このデータでは, 球面性の検定が有意でないため, Output.1 の結果を用いても差し支えない.

Output.3

```
Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

      GG eps    Pr(>F[GG])
Time 0.83516 0.0000002039 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      HF eps       Pr(>F[HF])
Time 1.13685 0.00000001673366
```

Output.3 は, Greenhouse-Geisser の方法及び Huynh-Felt の方法の結果である. いずれも, p 値が 0.05 を下回っていることから, 調整を行った場合でも経時的変化に対して有意差が認められた.

1.5.4 ノンパラメトリック検定による繰り返し測定データの解析：Friedman 検定

(1) 繰り返しのある 3 群以上でのノンパラメトリック検定:Friedman 検定の概要

Kruskal-Wallis 検定は, 3 群以上の独立なグループに対する検定である. 一方で, 3 期 3 剤以上でのクロスオーバー試験(チェンジオーバー試験)あるいは, 経時的にとられたデータでは, 対応があるデータになる(これを繰り返しのある

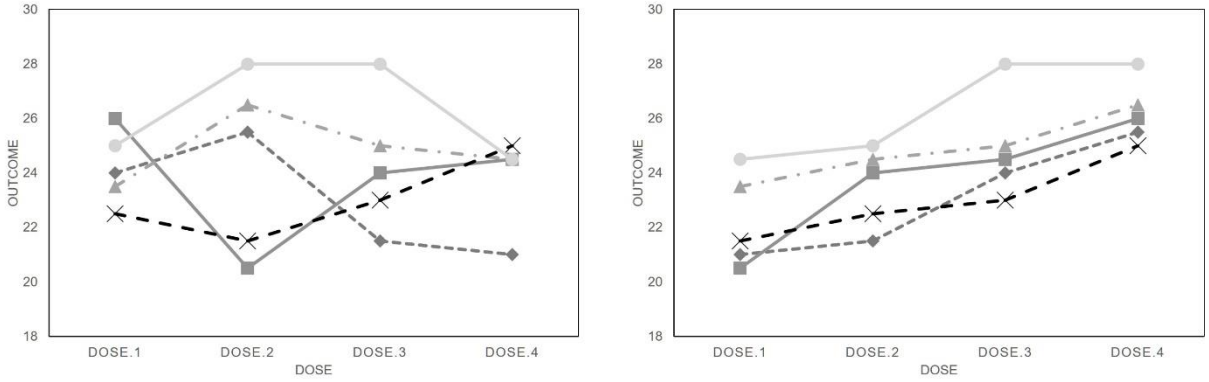
有意でない場合

ID	Dose.1	Dose.2	Dose.3	Dose.4
1	24.0	25.5	21.5	21.0
2	26.0	20.5	24.0	24.5
3	23.5	26.5	25.0	24.5
4	22.5	21.5	23.0	25.0
5	25.0	28.0	28.0	24.5

有意な場合

ID	Dose.1	Dose.2	Dose.3	Dose.4
1	21.0	21.5	24.0	25.5
2	20.5	24.0	24.5	26.0
3	23.5	24.5	25.0	26.5
4	21.5	22.5	23.0	25.0
5	24.5	25.0	28.0	28.0

(a) 仮想データ



(b) 仮想データの折れ線グラフ

ID	Dose.1	Dose.2	Dose.3	Dose.4
1	2	1	3	4
2	1	4	3	2
3	4	1	2	3
4	3	4	2	1
5	3	1	1	4
計	**13**	**11**	**11**	**14**

ID	Dose.1	Dose.2	Dose.3	Dose.4
1	4	3	2	1
2	4	3	2	1
3	4	3	2	1
4	4	3	2	1
5	4	3	1	1
計	**20**	**15**	**9**	**5**

(c) 順位和の計算

図 1.9：Friedman 検定の概要

データという). 例えば, 脳下垂体と翼突上顎裂の距離のデータでは, 個々の被験者から, 8,10,12,14 歳での距離を測定し, 年齢による違いを評価している.

ここでは, 5 名の被験者に対する 4 種類の用量(DOSE.1 < DOSE.2 < DOSE.3 <DOSE.4)を投与したときの反応(OUTCOME)の仮想例に基づいて Friemdman 検定を説明する. このデータでは, 個々の被験者に対して, 4 種類の用量の薬剤を投与したときのアウトカムを比較している. このとき, 前に投与した薬剤影響が消失するのに十分な期間をおいているとする(このような期間をウォッシュアウト期間という). 図 1.9 は, Friedman 検定が有意な場合と有意でない場合を表している. 上側の表はこのときのデータを表している. そして, 中央の折れ線グラフは, 用量(DOSE)に対する被験者毎の反応(OUTCOME)の変化を表している. 有意でない場合には, 用量による反応(OUTCOME)が被験者によって異なっている. 一方で, 有意な場合には, いずれの被験者も用量が増加するにつれて反応が上昇している(すなわち, 因子(用量)によって OUTCOME に違いがある).

下側の表は, 被験者毎に OUTCOME に昇順に順序付けたものであり, 一番下側は, 各 DOSE での順位和を表している. 有意でない場合には, 各 DOSE での順位和が類似しており(バラツキが小さい), 有意な場合には, 各 DOSE での順位和が異なっている(バラツキが大きい). すなわち, Friedman 検定では, 順位和のバラツキを評価することで検定している.

(2) EZR による Friedman 検定の実行

ここでは，1.5.3 節の脳下垂体と翼突上顎裂の距離のデータを用いて Friedman 検定の適用方法を示す．Friedman 検定の関心は，「年齢によって脳下垂体と翼突上顎裂の距離に違いがあるか[10]」にある．因みに，Friedman 検定には，両側対立仮説，片側対立仮説はない．Friedman 検定の手順を以下に示す．

Friedman 検定の実行
1: 「統計解析」→「ノンパラメトリック検定」→「対応のある 3 群以上の間の比較(Friedman 検定)」を選択する． 2: 次のようなメニューが表示される．    このとき， ・「繰り返しのある変量(2 つ以上選択)」で「gt8」，「gt10」，「gt12」，「gt14」を選択する． ・「2 群づつの比較(Bonferroni の多重比較)」，「2 群づつの比較(Holm の多重比較)」にチェックを入れる． 3: 「OK」ボタンを押す

多重比較では，Bonferroni の多重比較及び Holm の多重比較が存在するが，これらは，時点間でのすべての組み合わせでの評価を行う．

EZR の出力では，様々な出力が表示される．表示された青色の箇所毎に説明する．

Output.1	対応のある 3 群以上の間の比較(Friedman 検定) P 値 = 0.00000314

Output.1 は，Friedman 検定の結果である．p 値が 0.00000314 であることから，有意水準 α=0.05 のもとで有意である．したがって，年齢によって脳下垂体と翼突上顎裂の距離に違いが認められる．なお，上側の青色のアウトプット(Friedman rank sum test)は，この出力と同じ意味なので無視してよい．

Output.2

```
        Pairwise comparisons using Wilcoxon signed rank test

data:  Dataset

     gt8   gt10  gt12
gt10 0.138 -     -
gt12 0.023 0.131 -
gt14 0.023 0.022 0.058

P value adjustment method: bonferroni
```

[10] この事例の場合には，成長によって脳下垂体と翼突上顎裂に上昇傾向(成長とともに距離が大きくなるか)に南進があるかもしれない．そのような傾向変化を評価する場合には，Jonckeere-Terpstra 検定を用いる．Jonckeere-Terpstra 検定では，帰無仮説 H_0「傾向変化がない」に対して，両側対立仮説 H_1 では，「傾向変化がある」，片側対立仮説 H_1 では，「上昇傾向がある」あるいは「減少傾向がある」が評価される．
EZR における Jonckeere-Terpstra 検定の実行は，
「統計」→「ノンパラメトリック検定」→「連続変数の傾向の検定(Jonckeere-Terpstra 検定)」
を選択すればよい．なお，EZR においても，両側対立仮説，片側対立仮説を選択することができる．

Output.2 は，Bonferroni の多重比較の結果である(太字の部分に多重比較の結果が表示されている)．ここで，対比較には Wilcoxon 符号付き順位検定が用いられている．8 歳 vs 12 歳，8 歳 vs 14 歳，10 歳 vs.14 歳のあいだで有意差が認められている．

Output.3

```
           Pairwise comparisons using Wilcoxon signed rank test

data:  Dataset

     gt8   gt10  gt12
gt10 0.044 -     -
gt12 0.022 0.044 -
gt14 0.022 0.022 0.029

P value adjustment method: holm
```

これは，Holm の多重比較の結果である(太字の部分に多重比較の結果が表示されている)．ここで，対比較には Wilcoxon 符号付き順位検定が用いられている．すべての年齢のペアで有意差が認めらる．

1.5.4 多元配置の分散分析

(1) データの概要:疼痛薬・性別のデータ

ここでは，1.5.1 節の 3 種類の疼痛薬(A,B,C)による痛みの程度の比較のデータを一部変更するとともに，性別の情報を追加したデータを用いる．

薬	性別	観測値						
薬 A	男性	8.18	7.24	7.00	7.00	8.00		
	女性	7.69	9.69	8.89	6.94	2.13	7.26	5.87
		7.20	6.81	6.67	6.98	7.07	5.00	
薬 B	男性	12.90	16.60	9.81	7.84	9.42	10.67	10.83
	女性	8.35	3.84	4.62	9.29	6.43		
薬 C	男性	12.40	14.00	11.60	13.90	11.20	13.21	
	女性	12.20	9.41	2.40	9.78	7.04		

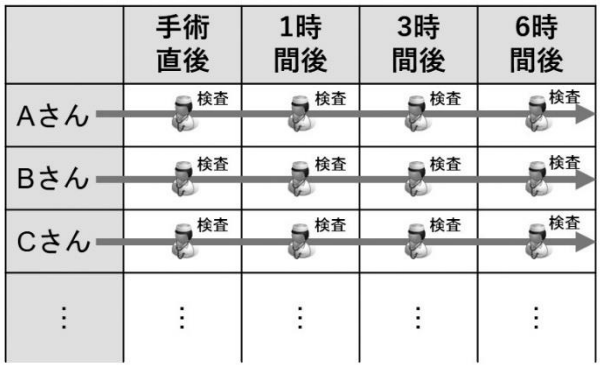
このデータは，Analgesics2.csv で与えられる．

(2) 多元配置の分散分析の概要

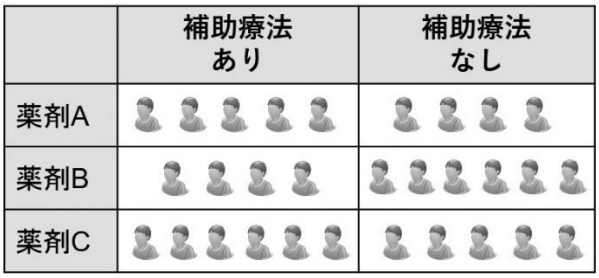
複数の因子が存在するときの分散分析として，2 元配置の分散分析を検討する．なお，2 元配置の分散分析では，因子が 2 個になったものの，分散分析表の作成方法は，一元配置の分散分析と同様である．ただし，適用場面には幾つかのパターンが存在する．ここでは，2 つの場面を考える:

［場面 1］新たな手術法を開発したときの，術後の検査値の推移を手術直後，1 時間後，3 時間後，6 時間後に測定した研究．

［場面 2］3 種類の薬剤(薬剤 A，薬剤 B，薬剤 C)と補助療法(あり，なし)の投与前後での検査値の変化を評価する研究．



(a) 同一被験者から複数時点でアウトカムがとられた場合



(b) 2 種類の介入が存在する場合

図 1.10:2 元配置の分散分析が適用される場面

それぞれの場面での観測値のイメージを図 1.10 に示す. 場面 1 では, 各被験者から 4 回(手術直後, 1 時間後, 3 時間後, 6 時間後)の検査値(アウトカム)を取得する. また, 被験者と手術時間の組み合わせでは, 1 個の観測値のみが与えられる. そして, アウトカムに影響を及ぼす要因として被験者と術後時間が存在するものの, 研究の関心は術後の検査値の経時的変化である(図 1.10(a)). このような場合では, 繰り返し測定の分散分析を用いる. そして, 術後時間の因子が有意であるならば, 被験者の個人差に依らず術後の検査値に経時的変化があると解釈される. このとき, 術後時間を傾向変化として扱う場合には数値情報になることから量的因子と呼ばれ, 被験者の因子は質的因子と呼ばれる.

一方で, 場面 2 では, 薬剤の種類と補助療法の有無の 2 因子が存在するため(図 1.10(b)), 2 元配置の分散分析を用いることになる. ただし, 場面 1 と異なるのは, 薬剤と補助療法の組み合わせによる効果の吟味である. 統計学では, この組み合わせ効果を交互作用(interaction)といい, それぞれの因子(薬剤, 補助療法)の効果を主効果(main effect)という.

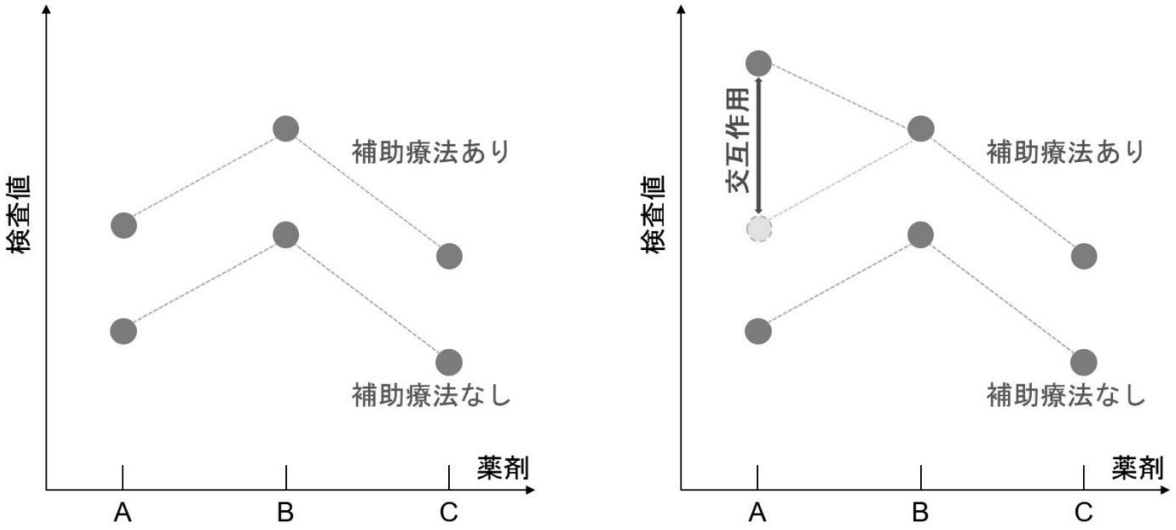
図 1.11 は, 場面 2 における薬剤と補助療法の組み合わせでの平均を表している. 交互作用が存在しない場合(図 1.11(a)), 薬剤 B の検査値が最も高く, 補助療法を追加することで, いずれの薬剤でも検査値が同じように増加している. 一方で, 交互作用が存在する場合には(図 1.11(b)), 薬剤 A において, 補助療法が加えられたことで, 他の 2 剤に比べて大幅に検査値が増加している. なお, 交互作用(薬剤×補助療法)を評価するには, 主効果(薬剤, 補助療法)に加えて, 交互作用を要因に加えたうえで分散分析を行う必要がある.

本稿では, 2 元配置の分散分析での解説だったが, 分散分析では, 3 因子以上の主効果あるいは, 複雑な交互作用を含むことができる. 一方で, 複雑な交互作用は, 解釈を困難にさせる恐れがあるため, 注意が必要である.

(3) EZR による多元配置の分散分析の実行

1.5.1 節では, 3 種類の薬(A,B,C) の効果のみを因子とした一元配置の分散分析を用いて解析した. 本節では性別も因子に加えた, 二元配置の分散分析を考える. ここでは, 交互作用を含めた検討を行う. この事例における交互作用は,「薬剤 A を男性に投与すると女性に投与するよりも有効である」というような相乗効果が認められる状況などが検討できる. 従って, 分散分析のモデルは,

(痛みの程度) = (平均) + (疼痛薬の影響) + (性別の影響) + (疼痛薬×薬剤の影響) + (誤差)

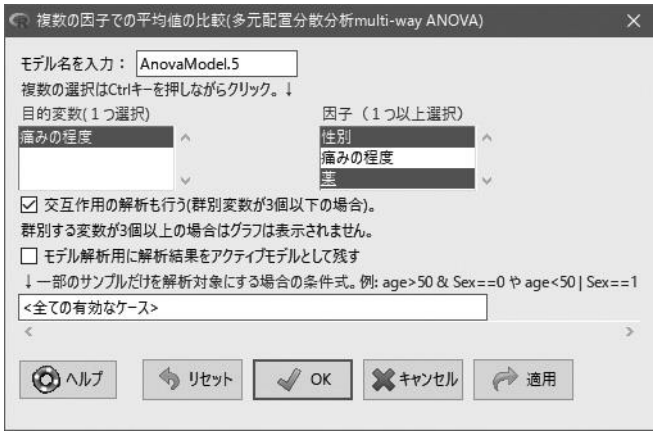


(a) 交互作用が存在しない場合　　(b) 交互作用が存在する場合

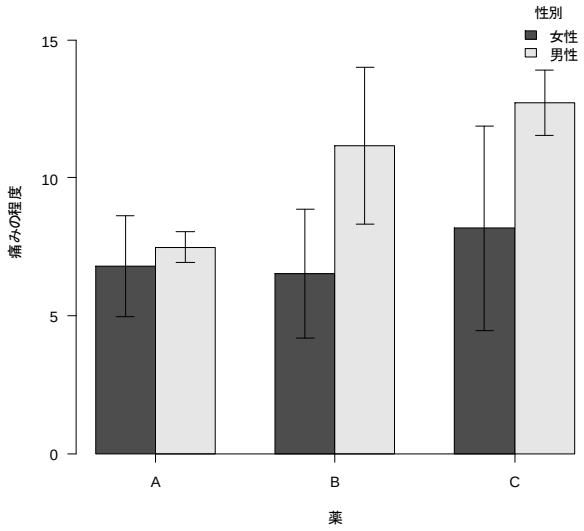
図 1.10: 場面 2 における薬剤と補助療法の交互作用の有無による傾向

で与えられる. ここでの平均とは, 疼痛薬や性別の影響がない, 全体での平均的な痛みの程度を表しており, 具体的には, 全ての被験者における平均値を意味する.

このときの, EZR による解析方法を以下に示す.

2 元配置の分散分析の実行
1:　「統計解析」→「連続変数の解析」→「複数の因子での平均値の比較(多元配置分散分析 multi-way ANOVA)」を選択する. 2:　次のようなメニューが表示される.  このとき, ・「目的変数(1 つ選択)」で「痛みの程度」を選択する. ・「因子」で「性別」, 「薬」を選択する(このとき, CTRL キーを押しながらクリックする). ・「交互作用の解析も行う(群別変数が 3 個以下の場合)」にチェックを入れる. 3:　「OK」ボタンを押す

このとき, 薬剤と性別の組み合わせ毎の平均値と標準偏差のグラフが次のように表示される.



その結果，薬 A の痛みの程度が低かった．また，男性よりも女性のほうがいずれの薬剤でも痛みの程度が低く，薬 A に比べて，薬 B，薬 C のほうが男女差が顕著だった．

EZR の出力では，様々な出力が表示される．表示された青色の箇所毎に説明する．

Output.1

```
      薬
性別          A        B        C
  女性 6.784615  6.50600  8.16600
  男性 7.484000 11.15286 12.71833
```

Output.1 は，各因子の組み合わせにおける平均値を表している(すなわち，上図の棒グラフと同様である)．因みに，出力では，意味が記載されていないが，このアウトプットの上側の R のコマンド

```
> tapply(TempDF$痛みの程度, list(性別=TempDF$性別, 薬=TempDF$薬), mean, na.rm=TRUE) # means
```

の右側に means(平均)と記載されているので，それを参考にすればよい．

Output.2

```
      薬
性別          A        B        C
  女性 1.8242926 2.335665 3.706087
  男性 0.5654025 2.853137 1.177971
```

Output.2 は，各因子の組み合わせにおける標準偏差を表している(すなわち，上図のエラーバーと同様である)．因みに，出力では，意味が記載されていないが，このアウトプットの上側の R のコマンド

```
> tapply(TempDF$痛みの程度, list(性別=TempDF$性別, 薬=TempDF$薬), sd, na.rm=TRUE) # std. deviations
```

の右側に std. deviation(標準偏差)と記載されているので，それを参考にすればよい．

Output.3

```
      薬
性別   A B C
  女性 13 5 5
  男性  5 7 6
```

Output.3 は，各因子の組み合わせにおける被験者数を表している．因みに，出力では，意味が記載されていないが，このアウトプットの上側の R のコマンド

```
> tapply(TempDF$痛みの程度, list(性別=TempDF$性別, 薬=TempDF$薬), function(x) sum(!is.na(x))) # counts
```

の右側に counts(個数)と記載されているので，それを参考にすればよい．

Output.4

```
Anova Table (Type III tests)

Response: 痛みの程度
                          Sum Sq Df  F value     Pr(>F)
(Intercept)             2827.41  1 569.5257  < 2.2e-16 ***
Factor1.性別              99.33  1  20.0077 0.00007791 ***
Factor2.薬                68.60  2   6.9087   0.002959 **
Factor1.性別:Factor2.薬   33.54  2   3.3778   0.045571 *
Residuals               173.76 35
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output.4 が二元配置の分散分析の結果である. ここで,「Factor1.性別」は, 性別の主効果,「Factor2.薬」は薬剤の主効果,「Factor1.性別:Factor2.薬」は, 性別×薬剤の交互作用を表している. そして,「Pr(>F)」がそれぞれの効果に対する p 値を表している. いずれも, 有意水準 0.05 のもとで有意であり, 有意な効果が認められた. 棒グラフの解釈から,

・薬剤による痛みの程度に違いがあり, 薬剤 A の痛みの程度が最も低い,

・性差が認められ, 男性よりも女性のほうが痛みの程度が低い,

・薬剤×性別の交互作用が認められ, 薬剤 A に比べて薬剤 B,薬剤 C における性差が顕著であり, 男性の痛みの程度が高い,

ことがわかった. ちなみに,「Smu Sq」は, 平均平方和,「Df」は自由度,「F value」は F 値を表しているが, これらは, p 値「Pr(>F)」を計算するのに用いるものであり, 結果の解釈には用いない場合が多い.

1.6 相関分析

1.6.1 Pearson の相関係数

(1) データの概要:健康診断における血圧とコレステロール値のデータ

健康診断を受診した 20 名の被験者のコレステロール値と収縮期血圧が観測されている.

被験者 ID	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
コレステロール	216	261	169	287	181	187	159	194	187	244	241	295	215	206	245	259	238	271	134	231
血圧	139	153	124	168	116	145	128	114	116	169	130	156	127	116	149	147	128	133	103	147

相関関係があるだろうか. このデータは, Col.csv で与えられる.

(2) Pearson の相関係数の概要

相関関係とは, 2 変数間の関連性を表す用語である. 図 1.12 は相関関係を表す 3 種類の散布図である. 図 1.12(a)は X(横軸)が増加するほど Y(縦軸)が増加している. このような状態を正の相関関係があるという. そして, 図 1.12 (b)は X(横軸)が増加するほど Y(縦軸)が減少している. このような状態を負の相関関係あるという. さらに, 図 1.12 (c)は X(横軸)が増加しても Y(縦軸)に変化はない. このような状態を無相関関係という. これらの相関関係を数値化したものが相関係数である.

相関係数には, 次のような特徴がある.

- 相関係数は, −1 から 1 までの範囲をとる.
- 相関係数が正値の場合に正の相関関係があり, 1 に近づくほど散布図のデータ点が比例(右肩上がり)の直線状に布置する(正の相関関係が強いと判断される).
- 相関係数が負値の場合に負の相関関係があり, −1 に近づくほど散布図のデータ点が反比例(右肩下がり)の直線状に布置する(負の相関関係が強いと判断される).
- 相関係数が 0 に近づくほど無相関関係であることが示され, 散布図のデータ点が一様に散らばる.

EZR では，3 種類の相関係数(相関係数(Pearson の相関係数)，Spearman の順位相関係数，Kendall の順位相関係数)が存在する． Pearson の相関係数は，最も一般的に用いられている相関係数であり，2 変数が正規分布に従っていることが仮定される．単に相関係数と呼ぶ場合には，Pearson の相関係数を表す．

(3) 無相関性の検定

観察研究などでは，複数の検査項目間の相関関係を評価する場合がある．このとき，2 個の検査項目に相関関係があるか否かを統計学的に評価するために無相関性の検定を用いることが多い．無相関性の検定では，帰無仮説 H_0「相関係数が 0 である」に対して，3 種類の対立仮説は

両側対立仮説 H_{1a}：相関係数は 0 でない．

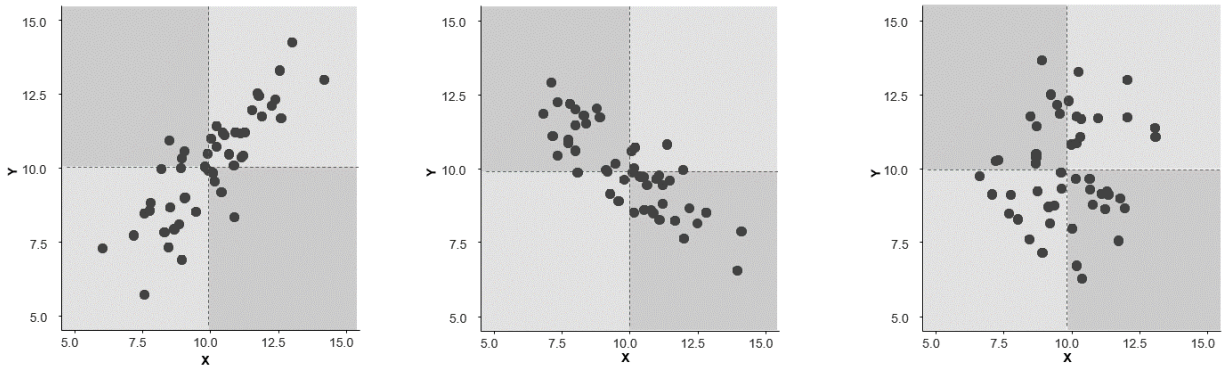
片側対立仮説 H_{1b}：相関係数は 0 よりも大きい(正の相関関係がある)．

片側対立仮説 H_{1c}：相関係数は 0 よりも小さい(負の相関関係がある)．

である．

図 1.13 は，無相関性の検定に対する 2 つの例示である．図 1.13(a)は，相関係数=0.713 のデータに対する散布図である(標本サイズ=15)．データ点が右肩上がりの傾向を示すことから，正の相関関係が認められる．そして，無相関性の検定における p 値は 0.003 であることから，有意水準 0.05 のもとで有意である．図 1.13 (b)は，相関係数=0.051 のデータに対する散布図である(標本サイズ=2,500)．無相関性の検定における p 値は 0.010 で有意であるものの，散布図のデータ点の布置からは，相関関係が殆ど認められない．

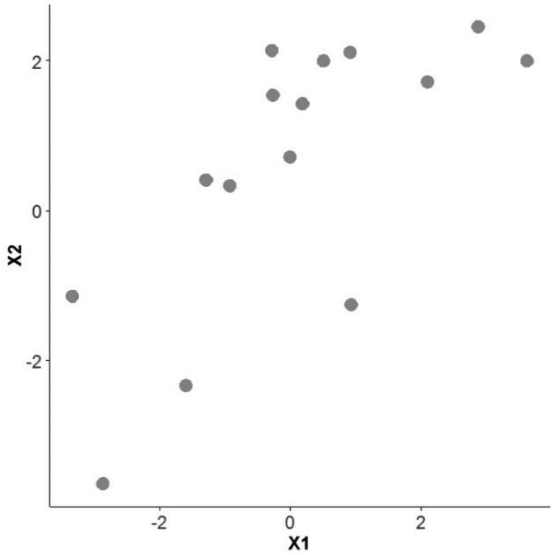
相関分析において，相関係数の解釈で重要なのは「相関係数が 0 であるか否かではなく，どの程度の相関関係の強さがあるか」を知ることにある．一方で，無相関性の検定では，「相関係数が 0 である」ことを統計学的に判断する手段であり，相関関係の強さを示すものではない．SAMPL(Statistical Analysis and Methods in the Published Literature)ガイドライン[11]では，相関係数を表す場合には，p 値とともに散布図および信頼区間を表記することが推奨されている．その理由は，散布図を描写することで観測値の正規性，外れ値，相関関係を視覚的に捉えることができ，相関係数の 95%信頼区間を記載することで，相関関係の信頼性(標本サイズが小さい場合には，「偶然」に得られた相関関係であるかもしれない)を表すことができるためである．



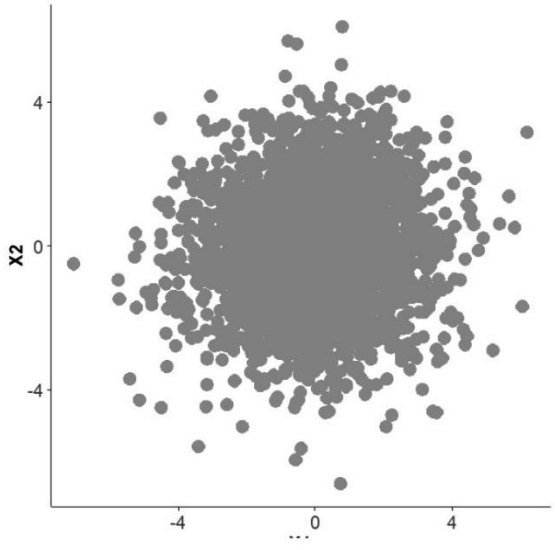
(a) 正の相関関係　(b) 負の相関関係　(c) 無相関関係

図 1.12：相関関係の図示

[11] Lang, T.A. and Altman, D.G.：Reporting Basic Statistical Analyses and Methods in the Published Literature: The SAMPL Guidelines for Biomedical Journals, http://www.equator-network.org/wp-content/uploads/2013/07/SAMPL-Guidelines-6-27-13.pdf.



(a) 標本サイズが 15 の場合の散布図
相関係数=0.713 (p 値=0.003)



(a) 標本サイズが 2,500 の場合の散布図
相関係数=0.051 (p 値=0.010)

図 1.13：無相関性の検定と相関係数の関係を表す 2 種類の散布図

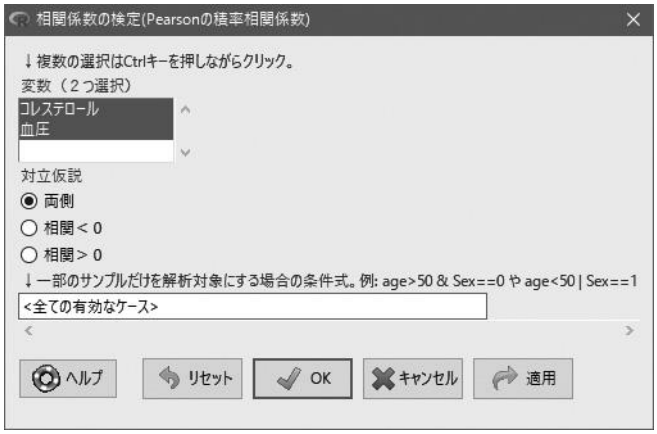
図 1.13 (a)の相関係数及び 95%信頼区間は 0.713 [0.317, 0.898]であり，図 1.13 (b)では 0.051[0.012, 0.090]である．図 3(a)では，比較的高い正の相関関係が示されているものの，標本サイズが小さいため，その 95%信頼区間の信頼幅は大きく，図 1.13 (b)では，(無相関性の検定では有意だったものの)殆ど相関関係が認められないことを散布図及び 95%信頼区間を用いて評価できる．

(4) EZR による Pearson の相関係数の計算

ここでは，EZR による Pearson の相関係数の計算を行う．

Pearson の相関係数の実行

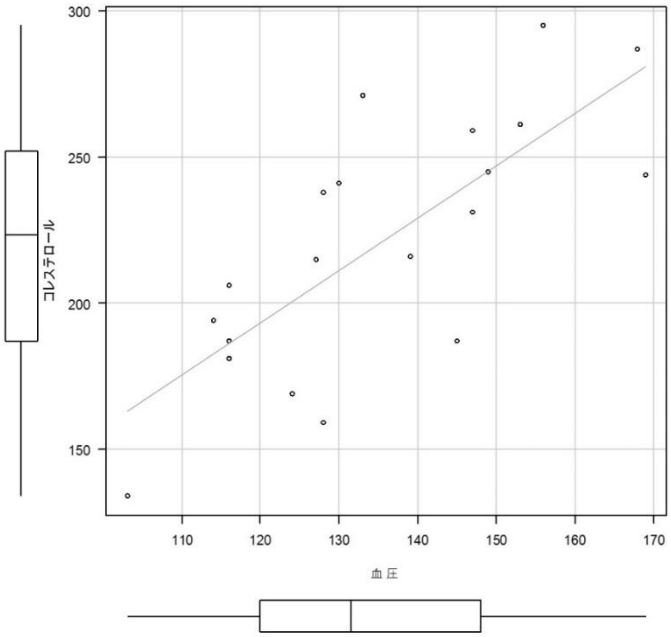
1: 「統計解析」→「連続変数の解析」→「相関係数の検定 (Pearson の積率相関係数)」を選択する．
2: 次のようなメニューが表示される．



このとき，
- 「変数(2 つ選択)」で「コレステロール」，「血圧」を選択する．
- 「対立仮説」で「両側」を選択する．

3: 「OK」ボタンを押す

このとき，散布図が次のように表示される．



ここで，直線は回帰直線を表しており，相関関係の目安として表示される．また，座標軸の外側の箱ひげ図は，それぞれの変数に対応しており，ヒゲは最小値，最大値を表している．直線が右斜め上になっていることから，正の相関関係が示唆される．

このときの出力を以下に示す．

```
相関係数 = 0.755，95%信頼区間 0.468-0.897，P 値 = 0.000121
```

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される．．また，上側の青色の出力部分(Pearson's product-moment correlation のなかで記載されている部分)は，この出力と同じ意味なので，無視してかまわない．相関係数は 0.755 なので，高い正の相関関係が認められた．このときの 95%信頼区間は[0.468,0.897]であった．さらに，無相関性の検定の p 値が 0.000121 なので，有意水準 0.05 のもとで有意である．よって，コレステロールと収縮期血圧には，有意な正の相関が認められた．

1.6.2 Spearman の順位相関係数

(1) Spearman の順位相関係数の概要

図 1.14 は，胃癌患者 63 名の AST と ALT の散布図である．このとき，Pearson の相関係数は 0.819 であり，高い正の相関関係が認められる．しかしながら，散布図のデータ点の布置(とくに青色の範囲)を眺めると，正の相関関係は認められるものの，「高い」相関関係であるとは言えない．この事例では，2 名の被験者の AST，ALT が高い数値を示しており(緑色の範囲)，これらを除外して Pearson の相関係数を計算すると，0.615 であり，0.204 減少する．したがって，これらの値が Pearson の相関係数に影響を及ぼしていると考えられる．

Pearson の相関係数では，2 変数が正規分布に従うことが仮定されている．そのため，正規分布に従わない場合(例えば，データが歪んでいる場合)や外れ値が存在する場合に Pearson の相関係数を利用すると，誤った解釈を導く惧れがある．図 1.14 の場合には，2 個の外れ値が Pearson の相関係数の結果に影響を及ぼし，「高い」相関関係が示された．正規分布に従わない場合や外れ値が存在しない場合，あるいは計数データや順序カテゴリカル・データなどの相関関係を評価する方法が，ノンパラメトリック相関係数である．ノンパラメトリック相関係数には，Spearman の順位相関係数や Kendall の順位相関係数などがあるが，本節では前者の Spearman の順位相関係数をとり上げる．

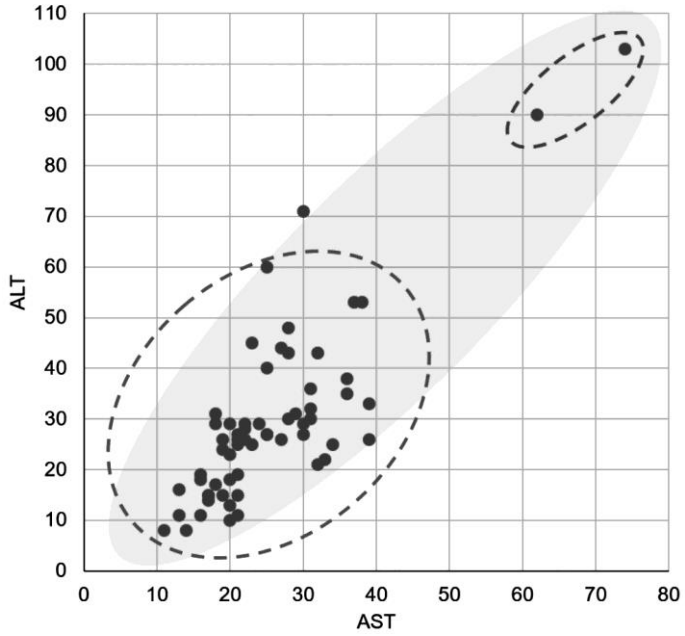


図 1.14：胃癌患者 63 名の ALT と AST に関する散布図

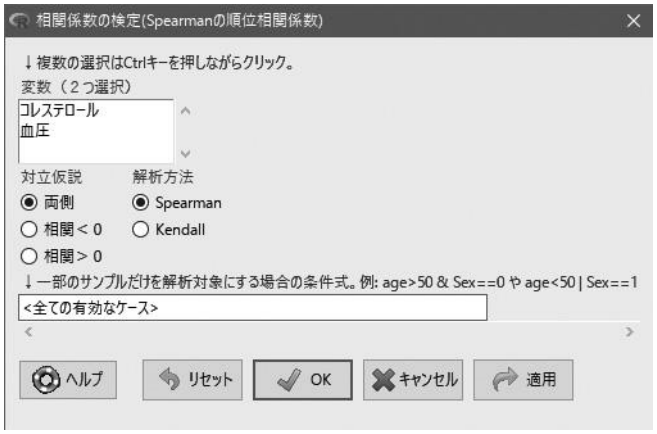
Spearman の順位相関係数とは，2 変数のそれぞれを順位付けしたもとで，順位を用いて相関係数を計算する方法である(2 変数を順位付けしたもとで Pearson の相関係数を計算すると Spearman の順位相関係数に一致する)．図 1.14 の観測値において，Spearman の順位相関係数は 0.727 であることから，Pearson の相関係数(0.819)に比べて減少したものの，図 1.14 の相関関係を反映しているように思われる．

なお，SAMPL(Statistical Analysis and Methods in the Published Literature)ガイドラインでは，適切な相関係数を選択することとともに，利用した相関係数の名称(例えば，Pearson の相関係数，Spearman の順位相関係数など)を論文に記載することが明記されている．

(2) EZR による Spearman の順位相関係数の計算

ここでは，1.6.1 節のデータを用いて Spearman の順位相関係数を計算する．

Spearman の順位相関係数の実行
1: 「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)」を選択する．
2: 次のようなメニューが表示される．



このとき，

・「変数(2 つ選択)」で「コレステロール」，「血圧」を選択する.
・「対立仮説」で「両側」を選択する.
・「解析方法」で「Spearman」を選択する.

3: 「OK」ボタンを押す

このとき，Pearson の相関係数と同様に，散布図が表示される(記載は割愛する). このとき注意しないといけないのは，順位相関係数は，「順位」の関係性を評価しているため，直線との直接的な関連性がない点にある.

このときの出力を以下に示す.

```
Spearman の順位相関係数 0.786 P 値 = 0.0000406
```

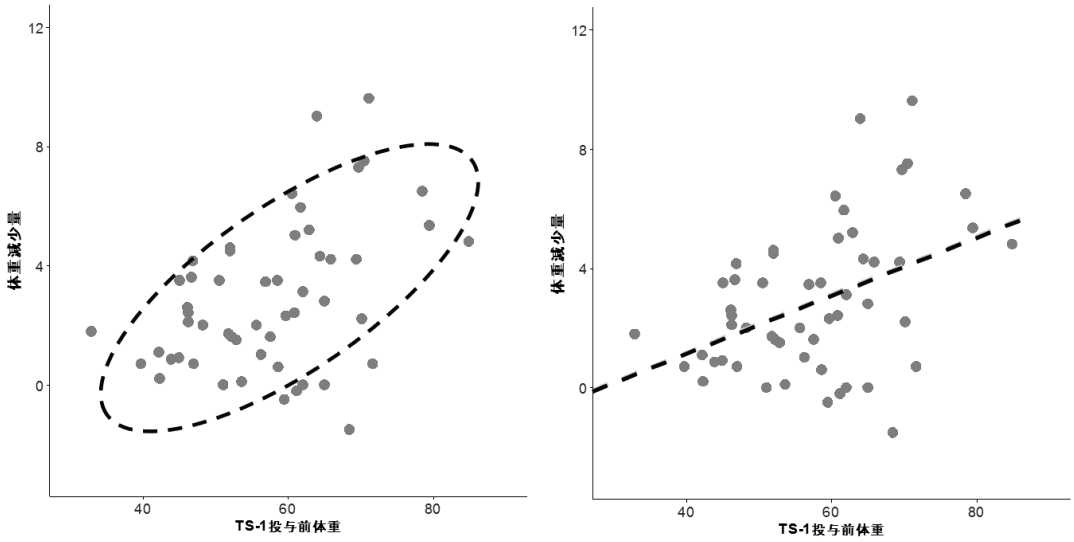
この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される. また，上側の青色の出力部分(Spearman's rank correlation rho のなかで記載されている部分)は，この出力と同じ意味なので，無視してかまわない. 相関係数は 0.786 なので，高い正の相関関係が認められた. さらに，Spearman の順位相関係数に対する無相関性の検定の p 値が 0.000121 なので，有意水準 0.05 のもとで有意である. よって，コレステロールと収縮期血圧には，有意な正の相関が認められた.

1.7 回帰分析

1.7.1 単回帰分析

(1) 単回帰分析の概要

図 1.15 は，TS-1 による補助化学療法が施行された 100 名の胃癌患者に対する投与前と投与後 6 カ月での体重減少量を表している. 相関係数が 0.427 であり，正の相関が認められることから，TS-1 投与前の体重が重いほど体重減少量が大きいと解釈される. そのため，投与前の体重から，投与後 6 カ月での体重減少量を予測することも可能かもしれない. このように，一方の変数(複数の場合もある)からもう一方の変数を予測する統計的方法を回帰分析という. とくに，予測する側の変数は 1 個の場合は単回帰分析と呼ばれ，複数の場合は重回帰分析と呼ばれる. このとき，予測する側の変数のことを説明変数，独立変数，入力変数と呼び，予測される側の変数を応答変数，従属変数，出力変数という. 本稿では，説明変数及び応答変数の名称を用いる.



(a) 相関分析の図示　　(b) 回帰分析の図示

図 1.15：相関分析と回帰分析の違い

EZR の出力では，相関分析の結果を表す散布図に回帰直線(図 1.15(a))が描写される．しかしながら，相関分析と回帰分析には明確な違いがある．相関分析とは，2 変数の関連性(相関関係)を分析する方法であり，正の相関が高いとは，片方の変数の値が上がれば，もう一方の変数の値が上がる(負の相関関係の場合には下がる)ことを表す．一方で，回帰分析は，説明変数から応答変数を予測するための統計モデル(回帰直線)を推定する方法である(図 1.15(b))．

単回帰分析では，1 個の説明変数から応答変数を予測するための統計モデルを推定する．単回帰分析における統計モデルを単回帰直線あるいは単回帰モデルという．単回帰直線は，

$$(\text{応答変数}) = \beta_0 + \beta_1 \times (\text{説明変数}) + (\text{誤差})$$

で与えられる．ここで，単回帰直線の切片 β_0 および傾き β_1 は回帰係数(回帰パラメータ)と呼ばれる．また，誤差は単回帰直線で説明できなかった応答変数の予測値に対する乖離(誤差)である．説明変数(投与前の体重)の任意の値 x に対する単回帰直線に基づく応答変数(体重変化量)の予測値 $\hat{y}$ は，回帰係数の推定値 $\hat{\beta}_0, \hat{\beta}_1$ を用いて

$$\hat{y}=\hat{\beta}_0 + \hat{\beta}_1 x$$

である．予測値 $\hat{y}$ と応答変数の値 y の差 $y - \hat{y}$ (推定された回帰直線では説明できなかった値)は，残差と呼ばる(統計学ではモデルで記述する場合には誤差，実際の予測値と応答変数の値の差を残差という)．因みに，回帰係数の推定値 $\hat{\beta}_0, \hat{\beta}_1$ は，残差の平方和(残差平方和)が最小になるように求められる．

因みに，図 1.15(b)の TS-1 による補助化学療法施行の胃癌患者に対する体重減少量のデータにおいて，推定された単回帰直線は

$$\hat{y}= -2.682 + 0.097 \cdot x$$

である．すなわち，投与前の体重が 1kg 増加することで，TS-1 投与による体重減少量は 0.097kg であることが予測される．

(2) 寄与率

応答変数の各観測値と応答変数の平均値の差の 2 乗値を求め，それらを総計したものを「総変動の平方和 SS_T 」という．また，予測値の各観測値と応答変数の平均値の差の 2 乗値を求め，それらを総計したものを「回帰変動の平方和 SS_R 」という．予測値 $\hat{y}$ と応答変数の値 y の差 $y - \hat{y}$ (残差)の平方和を残差平方和を SS_E とするとき，それぞれの平方和には

$$SS_T = SS_R + SS_E$$

の関係がある．このような関係式のことを回帰分析の変動分解という．回帰変動の平方和 SS_R は推定された回帰直線が当てはまっている度合いを表しており，残差平方和 SS_E は推定された回帰直線が当てはまっていない度合いを表す．

回帰変動が総変動に占める割合を計算することで，推定された(単)回帰直線が応答変数のどのぐらいの割合を説明しているかを要約することができる．この指標は寄与率(決定係数)と呼ばれ，0 から 1 の範囲で表すことができる．

(3) 適合度評価:F 検定

先ほどは，推定された回帰モデルの適合度を数値化する方法として寄与率について説明した．本項では，推定された回帰モデルには統計学的な意味があるか否を検定する方法について説明する．このときの検定は，F 検定と呼ばれる．F 検定では，

帰無仮説 H_0:「回帰モデルに意味がある」

対立仮説 H_1:「回帰モデルに意味がない」

が検定される. F 検定は, 分散分析表を用いるが, このときの分散分析表を「回帰の分散分析」と呼ぶことがある.

(4) 回帰係数に対する検定

推定された回帰直線の適合度が高くても, 回帰係数 β_1 の推定値 $\hat{\beta}_1$ が小さければ, 説明変数が応答変数の値に影響を殆ど与えないことを意味する. したがって,

帰無仮説 H_0:「回帰係数 β_1 は 0 である」

対立仮説 H_1:「回帰係数 β_1 は 0 でない」

を検定することは, 応答変数を予測するのに説明変数が必要であるか否かを判断することになる. このような検定を回帰係数に対する検定(回帰係数に対する t 検定)と呼ぶ.

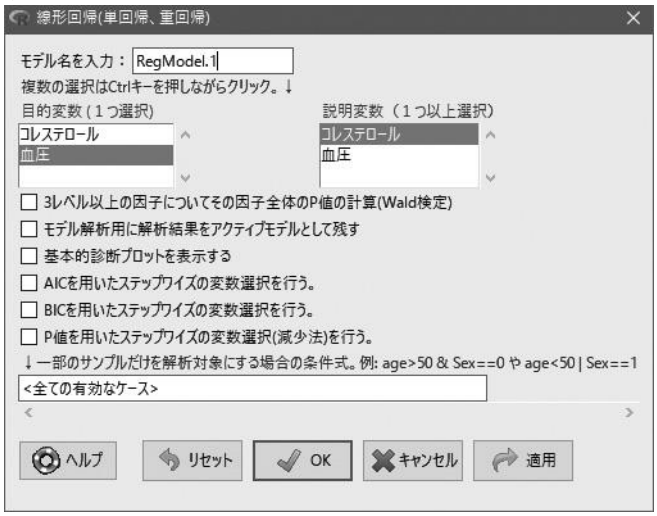
(5) EZR による単回帰分析の実行

ここでは, 1.6.1 節のデータを用いて単回帰分析を行う. その関心は,「コレステロール値」から「収縮期血圧」を予測するための単回帰モデルを推定することにある. したがって, 目的変数(応答変数)は「血圧」であり, 説明変数は,「コレステロール」である.

単回帰分析の実行

1: 「統計解析」→「連続変数の解析」→「線形回帰(単回帰、重回帰)」 を選択する.

2: 次のようなメニューが表示される.



このとき,

- 「目的変数(1 つ選択)」で「血圧」を選択する.
- 「説明変数(1 つ以上選択)」で「コレステロール」を選択する.
- 「解析方法」で「Spearman」を選択する.

3: 「OK」ボタンを押す

このときの出力を以下に示す.

```
                回帰係数推定値 95%信頼区間下限 95%信頼区間上限     標準誤差  t 統計量
(Intercept)         65.0298122       34.1650303       95.894594 14.69106484 4.426487
コレステロール       0.3184171        0.1812833        0.455551  0.06527319 4.878223
                        P 値
(Intercept)     0.0003258135
コレステロール 0.0001209234
```

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される. その結果, 推定された回帰モデルは,

$$\hat{y} = 65.03 + 0.318 \cdot (\text{コレステロール})$$

であった. また, コレステロールに対する回帰係数の 95%信頼区間は, [0.181, 9.456]であり, 0 を含まなかった. そのため, 回帰係数に対する検定の p 値も 0.00012 であり, 有意水準 α=.0.05 のもとで有意だった.

上側の R の出力

```
Call:
lm(formula = 血圧 ~ コレステロール, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-18.321  -9.372  -1.731   6.572  26.276

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    65.02981   14.69106   4.426 0.000326 ***
コレステロール  0.31842    0.06527   4.878 0.000121 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.44 on 18 degrees of freedom
Multiple R-squared:  0.5693,    Adjusted R-squared:  0.5454
F-statistic:  23.8 on 1 and 18 DF,  p-value: 0.0001209
```

を用いることで, 推定された回帰モデルの適合度を評価できる. 寄与率(Multiple R-squared の部分)は, 0.5693 であることから, 推定された回帰モデルは, 血圧(応答変数)に対して 56.93%の説明能力をもつことがわかる. また, F 検定の p 値(p-value の部分)は, 0.001 未満であり, 有意水準 α=0.05 のもとで有意である. つまり, 推定された回帰モデルには意味があることがわかった.

1.7.2 重回帰分析

(1) データの概要:糖尿病データ

ここでは, 糖尿病患者 442 名のデータを用いる(Efron et al., 2004)[12]. 応答は, 糖尿病患者のベースライン時点から 1 年後における症状進行状況のスコアであり, 説明変数は, 年齢, 性別, BMI, 血圧, 6 種類の血清検査(総コレステロール, LDL, LDL, TCH, LTG, グルコース)である. このデータは, Diabetes.csv で与えられる.

(2) 重回帰分析の概要

1.7.1 節では, 説明変数が 1 個の場合の回帰分析法として, 単回帰分析について説明した. 一方で, 医学系研究では, アウトカムに対する複数の要因(説明変数)を評価するために, 重回帰分析(いわゆる多変量解析)を用いる場面も少なくない.

説明変数が 3 個のときの重回帰分析のモデル(重回帰直線と呼ぶことも多いが, 重回帰分析の場合は, 重回帰モデルが一般的であるので, 本稿では重回帰モデルと呼ぶことにする)は,

$$(\text{応答変数}) = \beta_0 + \beta_1 \times (\text{説明変数 } 1) + \beta_2 \times (\text{説明変数 } 2) + \beta_3 \times (\text{説明変数 } 3) + (\text{誤差})$$

で表される. (説明変数 1)以外を左辺に移動すると,

$$(\text{応答変数}) - \beta_2 \times (\text{説明変数 } 2) - \beta_3 \times (\text{説明変数 } 3) = \beta_0 + \beta_1 \times (\text{説明変数 } 1) + (\text{誤差})$$

になる. 上式の左辺は応答変数に対して(説明変数 2)と(説明変数 3)の影響を調整してしている(排除している)ことを意味しており, 右辺は調整された応答変数のもとで(説明変数 1)の影響を評価していることを意味している. つまり, 重

[12] Efron, B., Hastie, T., Johnstone, J., and Tibshirani, R.:Least Angle Regression. Annals of Statistics (with discussion), 32, 407-499, 2004.

回帰分析は，他の要因(説明変数)の影響を考慮(調整)したうえで当該要因(説明変数)の影響を評価できることを意味する.

回帰係数 β_p の推定は，単回帰分析と同様であり，応答変数と予測値の差の平方和(残差平方和)を最小にするように求めることになる. このような推定の方法は，最小 2 乗法と呼ばれる.

いま，例示として，嚢胞性肺線維症の患者 25 名に対するデータを用いる[13]. このデータは，背景情報に関連する 5 項目(年齢，性別，身長，体重，BMI)及び，肺機能に関連する 5 項目(最大呼吸圧，努力肺活量，残気量，機能的残気量，総肺気量)により構成されている. ここでの目標は，最大呼吸圧を予測するための重回帰モデルを推定することにある.

性別は名義尺度なので，そのままの形式では利用できない. そのため，女性を 1，男性 0 と置き換えたもとで連続変数と同様に利用する. このように，連続変数に置き換えられた変数のことをダミー変数という. 性別のダミー変数に対する推定回帰係数 は，「女性のほうが男性に比べて最大呼吸圧が $\beta_{性別}$ ほど大きい」ことを意味する.

このときの重回帰分析の結果を図 1.16 に示す. 回帰係数に対する検定の結果，有意である(回帰係数が 0 でない)ことを示す説明変数が一つもないことがわかる.

(3) 自由度調整済み寄与率

1.7.1 節で説明した寄与率の問題点は，説明変数の数が増加するにつれて寄与率が高くなることにある. 図 1.17(a)は，説明変数の増加に伴う寄与率の変化をシミュレーションによって表している(シミュレート回数=100). ここで，X 軸はアウトカム(応答変数)に影響しない変数の数，Y 軸は寄与率を表している. データ点は，個々のシミュレーションの結果であり，点線は寄与率の平均値の推移を表している. 寄与率はアウトカムに影響しない変数の増加に伴い上昇していることがわかる.

重回帰分析では，寄与率の代わりに，自由度調整済み寄与率を用いることが殆どである. 自由度調整済み寄与率は

(自由度調整済み寄与率) = (回帰変動の不偏分散)／(総変動の不偏分散)

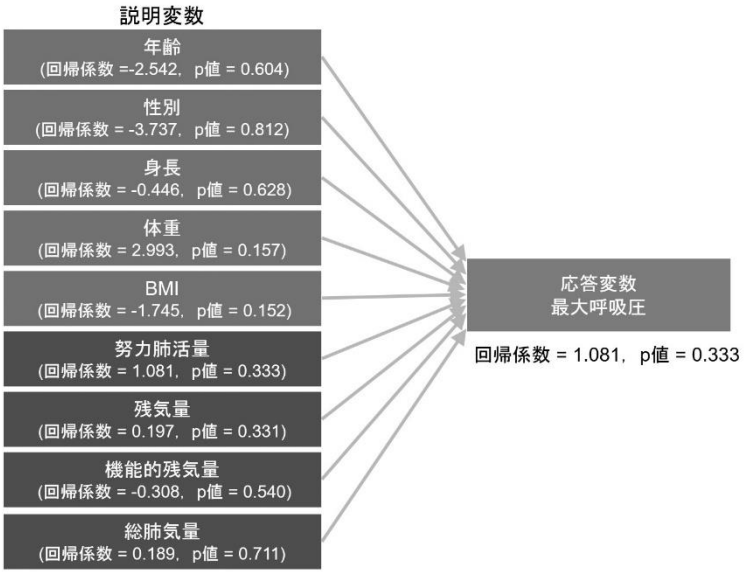
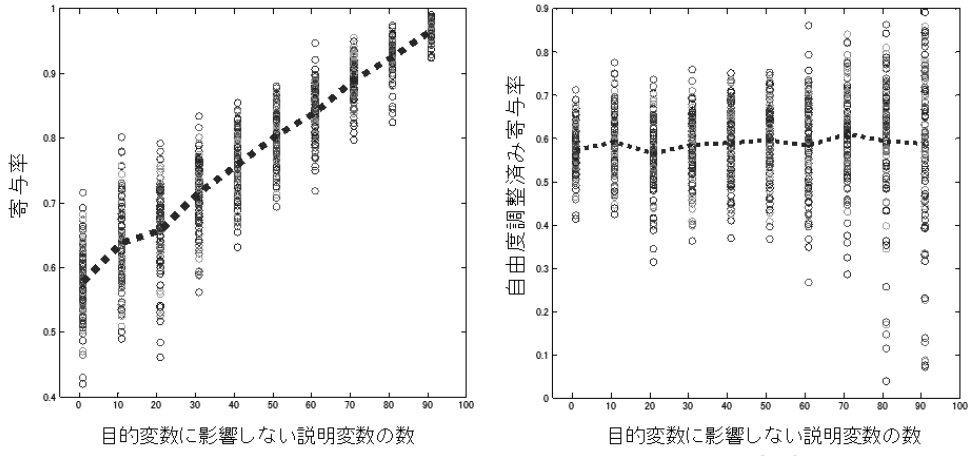


図 1.16：嚢胞性肺線維症のデータに対する重回帰分析の結果

[13] Altman, D.G.: Practical Statistics for Medical Research, Chapman & Hall, 1991.



(a) 寄与率　　(b) 自由度調整済み寄与率

図 1.17：アウトカム(応答変数)に影響しない説明変数の数を増加したときの寄与率及び自由度調整済み寄与率の推移(各説明変数の数に対して 100 回のシミュレートを実施している．点線は平均値)

で与えられる．なお，総変動の不偏分散は，(総変動の平方和)／(総変動の重度)で計算できる．

図 1.17(b)は説明変数の増加に伴う自由度調整済み寄与率の変化を表している．説明変数が増加しても自由度調整済み寄与率が変化しないことがわかる．

因みに，嚢胞性肺線維症のデータに対する寄与率が 0.637 であり，自由度調整済み寄与率は 0.420 である．結果の解釈には，自由度調整済み寄与率の 0.420 を重回帰モデルの適合度の評価に用いるべきであり，数値が高いという理由で寄与率 0.637 を用いてはならない．

(4) 変数選択

重回帰分析では，複数の説明変数を評価することができる．一方で，少しでも多くの説明変数を重回帰モデルに含めたほうが良い結果を導くかというと，そうではない．なぜなら，不要な説明変数は「ノイズ」として重回帰モデルに含まれるため，「不要な説明変数は含めるべきではない」．

嚢胞性肺線維症のデータでは，9 個の説明変数があるが，全ての説明変数が必要であるとは限らない．すなわち，不要な説明変数を削除しても重回帰モデルの予測結果に影響がないかもしれない(むしろ，良くなるかもしれない)．

不要な説明変数を削除することは，応答を予測するうえでの「ノイズ」を除去することにも繋がり，より安定的な回帰モデルの推定に繋がる．そのため，重回帰分析を実施する場合には，変数選択を併せて実施することが多い．このとき，応答を適切に予測するための説明変数を選択することは変数選択と呼ばれる．

応答を予測するための最適な説明変数を選択するには，全ての説明変数のパターンを計算しなければならない．嚢胞性肺線維症のデータの場合には，9 個の説明変数があることから，$2^9-1=511$ パターンの重回帰モデルを推定し，最適な説明変数の組み合わせを選択することになる．511 パターンであれば，現在のコンピュータの能力であれば実行可能かもしれない．しかしながら，20 個の説明変数がある場合には，$2^{20}-1=1{,}048{,}575$ パターンでの評価を行わなければならず，計算が困難になる．

そのため，変数選択では，説明変数の組み合わせの全パターンを評価するのではなく，ステップワイズ法というアルゴリズムを用いることが多い．ステップワイズ法には次の 3 種類がある：

(a) 変数増加法：切片のみのモデルから出発し，1 個ずつ説明変数をモデルに加える方法．

(b) 変数減少法：全ての説明変数を含むモデルから出発し，1 個ずつ説明変数をモデルから除外する方法．

(c) 変数増減法：切片のみのモデルから出発し，1 個づつ説明変数を加えるのか除外するのかを逐次選択する方法．

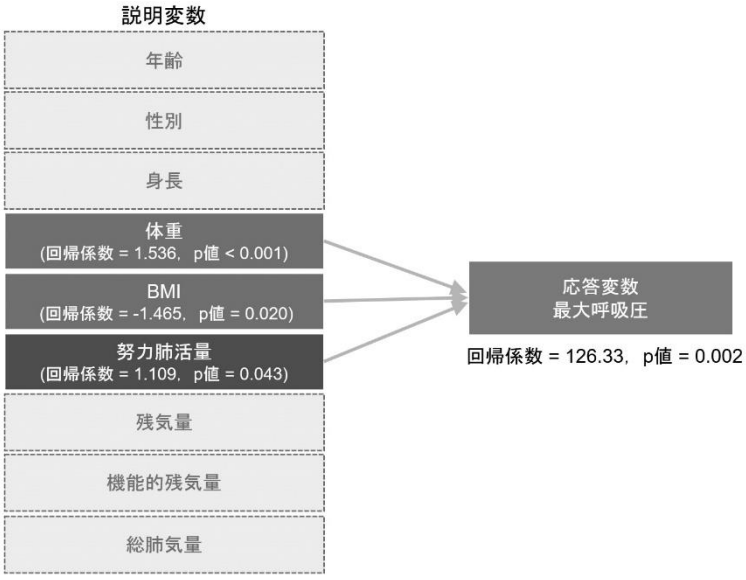


図 1.18：嚢胞性肺線維症のデータに対する後退ステップワイズ法を用いた重回帰分析の結果（説明変数の点線の括弧は，ステップワイズ法により削除された説明変数を表している）

ステップワイズ法のアルゴリズムに対するゴールド・スタンダードは存在しない．浜田(2013)[14]は，変数選択に対する経験則をまとめている．ここでは，それを参考に以下に示す．

(a) 評価したい要因は変数選択に強制的に加える

無作為化比較試験の結果を評価する場合，治療群を表す共変量を含まなければ意味をもたない．つまり，このような場合には，背景因子などの他の共変量を調整したうえで治療群(評価変数)を調べることに意義がある．

(b) 変数増加法の落とし穴

標本サイズが小さい場合に，変数増加法を用いて変数選択を行う場合，結果の解釈が困難なモデルを選択することがしばしばある．また，本当は必要な共変量が取り込まれる前に変数選択が終了する場合がある． 回帰モデルでの変数選択において，変数減少法のほうが変数増加法よりも取り込まれる変数が多いため，医学系研究では変数減少法を選ぶことがある．これは，本当は必要な共変量の「取りこぼし」が変数減少法のほうが少ないことを意味する．

(c) 多数の共変量(項目)がある場合の留意点

医学系研究では，多数の調査項目(共変量)を評価に用いることは少なくない．このような場合には，全ての共変量を用いて変数選択を行うのではなく，事前スクリーニングを行うことが推奨される[15]．事前スクリーニングでは，共変量毎に単変量解析(1 個の共変量による回帰モデルを推定する)を実施し，その回帰係数に対する検定(回帰係数が 0 であるか否かを評価する検定)の p 値，あるいはハザード比によって評価するが，p 値を用いることのほうが多いようである．

p 値に基づいて評価する場合には有意水準 α未満の変数を多変量解析に用いる．このとき，有意水準 αを 0.05 でなければいけないわけではなく，例えば，0.10 を用いる場合もある．例えば，p 値が 0.06 であったとしても，多変量解析を用いたときの調整ハザード比のもとでは，p 値が 0.05 を下回る可能性があるためである．

---

[14] 浜田知久馬：学会・論文発表のための統計学(新版)，新興交易(株)医療出版部，2012.

[15] 多数の共変量がある場合，多変量解析（重回帰分析，多重ロジスティック回帰分析，比例ハザードモデル）を行う場合，多重共線性（相関が高い共変量が不適切な結果を与える），あるいは解釈が困難な結果を得る恐れがある．

(d) 欠測が多い共変量(項目)には注意が必要である

多変量解析では，共変量のなかで 1 個でも欠測があれば，その被験者を削除しなければならない．そのため，欠測が多い共変量をモデルに含めると，多くの被験者を削除することになる．また，観測方法が煩雑な場合には，欠測が多くなる傾向にある．そのため，このような共変量は，予め変数選択の候補から覗いておくことが望ましい．

(e) 可能であれば総当たり法を用いる

変数増加法や変数減少法が必ずしも最適なモデルに到達するとは限らない．最適なモデルを選択できる唯一の方法は，すべての候補モデルを評価する総当たり法のみである．共変量の数が 10 個の場合，候補となるモデルの数は 1,023 個である．最近のコンピュータであれば実現不可能な数ではない(共変量の数が 20 個の場合には 1,048,575 個となり，不可能に近い数値となる)．そのため，臨床的知見あるいは，事前スクリーニングなどを用いて変数選択に用いる共変量を可能な限り少なくし，そのもとで，総当たり法によって変数選択を実施することが考えられる．

また，変数を増加(減少)させるか否かを評価する指標には，検定を用いる方法と情報量規準を用いる方法がある．情報量規準とは，推定された回帰モデルの適切性を評価する測度(ものさし)であり，赤池の情報量規準(AIC；Akaike's Information Criteria)などの方法が提案されている．検定を用いる方法では有意水準 α を事前に設定したもとで評価しなければならず，恣意的に説明変数の数が制御されてしまう恐れがある．そのため，情報量規準を用いることが多くなっている．

図 1.18 は，嚢胞性肺線維症のデータにおいて，AIC を用いた変数減少法で変数選択を実施した結果である．体重，BMI，努力肺活量のみの重回帰モデルが選択された．これらの説明変数の回帰係数に対する検定では，すべて有意な結果(回帰係数は 0 でない)ことを示すことができた．また，このときの自由度調整済み寄与率は 0.509 なので，全変数を用いた重回帰モデルよりも，適切な適合結果を示した．

SAMPL(Statistical Analysis and Methods in the Published Literature)ガイドライン[16]では，重回帰分析を用いた場合には，単回帰分析での結果，重回帰分析での結果，そして，変数選択を実施したときの結果について，変数選択の方法とともに記載することが指摘されている．表 1.1 は，嚢胞性肺線維症の回帰分析の結果を SAMPL ガイドラインにあわせて記載した表である．年齢，機能的残気量は単回帰分析において有意だったにも関わらず，重回帰分析では有意でなく，かつ変数選択後には削除されている．単回帰分析での結果では，他の説明変数の影響が考慮されていな

表 1.1：嚢胞性肺線維症のデータを要約するための SAMPL ガイドラインを遵守した表記例
(論文などでは，回帰係数に対する 95%信頼区間を併記する場合もある)

	単変量解析	多変量解析	
		変数選択なし	変数選択あり
年齢	4.055 (0.001)	-2.542 (0.604)	—
性別	-19.045 (0.162)	-3.737 (0.812)	—
身長	0.932 (0.002)	-0.446 (0.628)	—
体重	1.187 (0.001)	2.993 (0.157)	1.536 (p<0.001)
BMI	0.639 (0.270)	-1.745 (0.152)	-1.465 (p=0.020)
努力肺活量	1.354 (0.023)	1.081 (0.333)	1.109 (p=0.020)
残気量	-0.123 (0.124)	0.197 (0.331)	—
機能的残気量	-0.319 (0.038)	-0.308 (0.540)	—
総排気量	-0.358 (0.385)	0.189 (0.711)	—

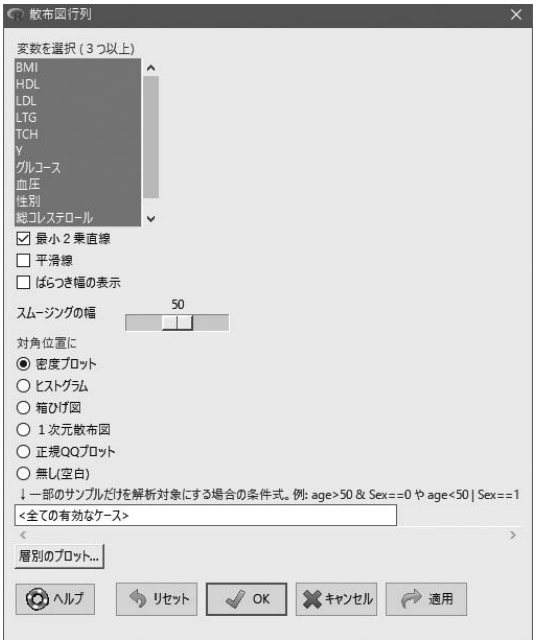
[16] Lang, T.A. and Altman, D.G.：Reporting Basic Statistical Analyses and Methods in the Published Literature: The SAMPL Guidelines for Biomedical Journals, http://www.equator-network.org/wp-content/uploads/2013/07/SAMPL-Guidelines-6-27-13.pdf.

い. 一方で, その他の説明変数(要因)を評価した場合, これらの説明変数は必ずしも必要でなかったことが伺える. なお, 論文等では, 回帰係数とともに, 回帰係数に対する 95%信頼区間を併記する場合もある.

(5) EZR による重回帰分析の実行

ここでは, 糖尿病データを用いる. 応答変数(従属変数)は, ベースラインから 1 年後の糖尿病の進行を表すスコア(「Y」: 以降, 糖尿病スコアと呼ぶ)であり, 説明変数は,「年齢」,「性別」,「BMI」,「血圧」,「総コレステロール」,「HDL」,「TCH」,「LTG」,「グルコース」である. ここで,「性別」は, 男性が 1, 女性が 0 のダミー変数で与えられている.

重回帰分析を実行する前に, 散布図行列を作成し, 応答変数と各説明変数の相関関係を省察する. ここで, 散布図行列とは, 全ての変数のペアの散布図を描写したグラフのことである.

散布図行列の描写
1: 「グラフと表」→「散布図行列」を選択する. 2: 次のようなメニューが表示される.    このとき, ・「変数(3 つ以上)」ですべての変数を選択する. 3: 「OK」ボタンを押す

このときの結果を, 図 1.19 に示す. ここで, 右側と上側の変数名および四角の枠は, 追記したものである. また, 縦と横が同じ変数を示す部分の曲線は, その変数の分布(密度関数)を表している. これが著しく歪んでいる場合には, 変数変換を実施するほうが良い.

散布図行列より, 糖尿病スコア(Y)と BMI, LTG には高い正の相関関係が認められた. 一方で, 糖尿病スコアと年齢の相関関係は低かった. また, 総コレステロールと LTG には, 非常に強い正の相関関係が示唆された. このような変数間では, 多重共線性[17]が生じる可能性があるので, 以降の解析で注意しなければならない.

---

[17] 重回帰分析では, 説明変数間の相関関係が高い場合には, 個々の説明変数の応答変数への影響を相殺することがある. これを, 多重共線性という.
多重共線性の存在を評価するには, (1)説明変数毎での相関係数を計算する, (2) 分散拡大係数(VIF; Variance Inflation Factor)を計算する, が考えられる. いずれの場合にも, 説明変数間の相関関係を評価することを目的とするが, 解釈の方法に違いがある. 相関係数の場合には, 説明変数の相関関係をペアワイズに評価しなければならないのに対して, 分散拡大係数の場合には, 任意の説明変数とその他の説明変数の相関関係が評価できる. 分散拡大係数が 2 未満の場合には問題がないと判断される(10 以上の場合には非常に問題があると判断される).

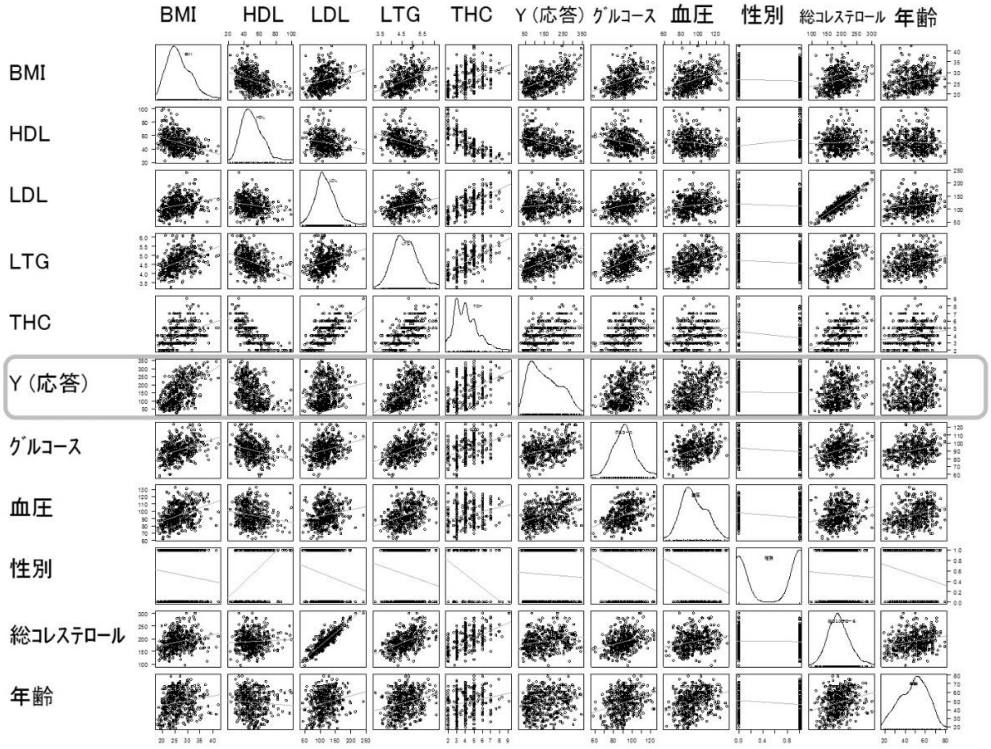
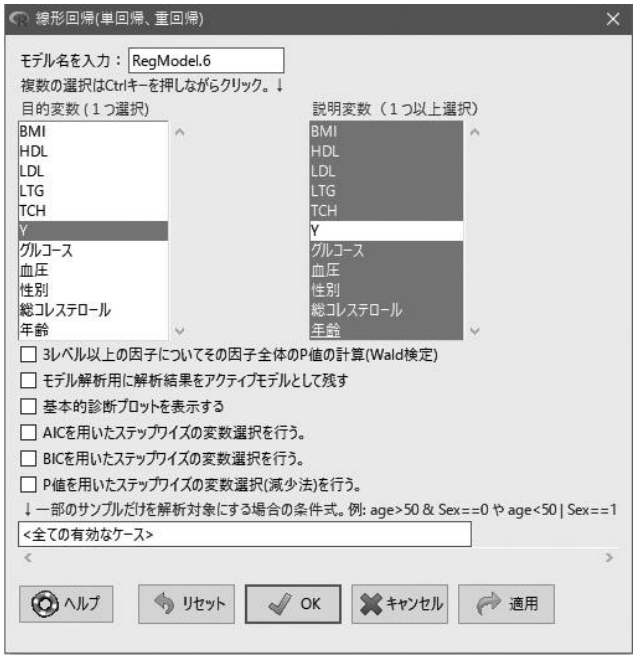


図 1.19：糖尿病データに対する散布図行列

次いで，重回帰分析を用いて解析する．このとき，Bayes 流情報量規準(BIC)を用いて変数選択を実施する．EZR での変数選択には，変数増減法(EZR の元になっている R では，変数増加法，変数減少法も用意されているが，EZR では，変数増減法のみが採用されている)が用いられている．また，情報量規準には，BIC の他に，赤池の情報量規準(AIC)を用いる方法，および p 値を用いる方法が用意されている．

重回帰分析の実行
1:　「統計解析」→「連続変数の解析」→「線形回帰(単回帰、重回帰)」を選択する． 2:　次のようなメニューが表示される．



このとき,

・「目的変数(１つ選択)」で「Y」を選択する.

・「説明変数(1 つ以上選択)」で「Y」以外の変数を選択する.

・「BIC を用いたステップワイズの変数選択を行う」にチェックを入れる.

3: 「OK」ボタンを押す

EZR の出力では, 様々な出力が表示される. 表示された青色の箇所毎に説明する.

Output.1

```
Call:
lm(formula = Y ~ BMI + HDL + LDL + LTG + TCH + グルコース + 血圧 +
    性別 + 総コレステロール + 年齢, data = Dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-155.827  -38.536   -0.228   37.806  151.353

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -380.28643   67.16813  -5.662 2.74e-08 ***
BMI                5.60296    0.71711   7.813 4.30e-14 ***
HDL                0.37200    0.78246   0.475 0.634723
LDL                0.74645    0.53083   1.406 0.160390
LTG               68.48312   15.66972   4.370 1.56e-05 ***
TCH                6.53383    5.95864   1.097 0.273459
グルコース         0.28012    0.27331   1.025 0.305990
血圧               1.11681    0.22524   4.958 1.02e-06 ***
性別              22.85965    5.83582   3.917 0.000104 ***
総コレステロール  -1.09000    0.57333  -1.901 0.057948 .
年齢              -0.03636    0.21704  -0.168 0.867031
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.15 on 431 degrees of freedom
Multiple R-squared:  0.5177,    Adjusted R-squared:  0.5066
F-statistic: 46.27 on 10 and 431 DF,  p-value: < 2.2e-16
```

Output.1 は, 変数選択前の回帰係数の推定結果と適合度を表している. 有意だった説明変数は,「BMI」,「LTG」,「血圧」,「性別」の 4 変数(10 変数中)のみだった. F 検定の p 値は, 0.001 未満で有意であり, 寄与率は, 0.5066 だった.

Output.2

BMI	HDL	LDL	LTG	TCH	グルコース
1.509437	15.402156	39.193370	10.075967	8.890986	1.484623
血圧	性別	総コレステロール	年齢		
1.459428	1.278071	59.202510	1.217307		

Output.2 は, 分散拡大係数(VIF; Variance Inflation Factor)である. 出力に説明がないが, 上側の R のコマンドが vif(モデル名)になっているので, これを参考にされたい. その結果,「HDL」,「LDL」,「LTG」,「総コレステロール」の VIF が 10.0 を上回っており, 多重共線性が示唆された.

Output.3

	回帰係数推定値	95%信頼区間下限	95%信頼区間上限	標準誤差	t 統計量	P 値
(Intercept)	-380.28643470	-512.3042757	-248.26859369	67.1681309	-5.6617093	2.740208e-08
BMI	5.60296209	4.1935032	7.01242099	0.7171055	7.8133023	4.296391e-14
HDL	0.37200472	-1.1659149	1.90992435	0.7824638	0.4754274	6.347233e-01
LDL	0.74645046	-0.2968957	1.78979659	0.5308344	1.4061833	1.603902e-01
LTG	68.48312496	37.6845532	99.28169676	15.6697192	4.3704117	1.555899e-05
TCH	6.53383194	-5.1777713	18.24543522	5.9586378	1.0965311	2.734587e-01
グルコース	0.28011699	-0.2570770	0.81731100	0.2733140	1.0248909	3.059895e-01
血圧	1.11680799	0.6741061	1.55950986	0.2252382	4.9583425	1.024278e-06
性別	22.85964809	11.3894387	34.32985749	5.8358213	3.9171261	1.041671e-04
総コレステロール	-1.08999633	-2.2168705	0.03687787	0.5733319	-1.9011613	5.794761e-02
年齢	-0.03636122	-0.4629525	0.39023010	0.2170414	-0.1675313	8.670306e-01

Output.3 は, Output.1 をわかりやすく表したものであり, 内容が重複するため, 割愛する.

R のコマンド「res <- stepwise(RegModel.1, direction="backward/forward", criterion="BIC")」以降では, 長い出力があるが, これは, ステップワイズ法の実行過程であるため, 解釈は不要である.

Output.4

```
Call:
lm(formula = Y ~ BMI + LDL + LTG + 血圧 + 性別 + 総コレステロール,
    data = TempDF)

Residuals:
     Min       1Q   Median       3Q      Max
-158.275  -39.476   -2.065   37.219  148.690

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -356.9486    26.5192 -13.460  < 2e-16 ***
BMI                 5.7111     0.7073   8.075 6.69e-15 ***
LDL                 0.8433     0.2298   3.670 0.000272 ***
LTG                73.3065     7.3083  10.031  < 2e-16 ***
血圧                1.1266     0.2158   5.219 2.79e-07 ***
性別               21.5910     5.7056   3.784 0.000176 ***
総コレステロール   -1.0429     0.2208  -4.724 3.12e-06 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.06 on 435 degrees of freedom
Multiple R-squared:  0.5149,    Adjusted R-squared:  0.5082
F-statistic: 76.95 on 6 and 435 DF,  p-value: < 2.2e-16
```

Output.4 は，変数選択後の回帰分析の結果を表している．その結果，「BMI」，「LDL」，「LTG」，「血圧」，「性別」，「総コレステロール」の 6 変数が選択され，回帰係数に対する検定の p 値はいずれも有意だった．また，このときの F 検定は有意であり，自由度調整済み寄与率は，0.5082 なので，全変数を用いるよりも僅かに上昇した．

因みに，選択された変数のみを用いて VIF を計算するために，新ためて重回帰分析を実施すると，

BMI	LDL	LTG	血圧	性別	総コレステロール
1.473160	7.366460	2.199033	1.344669	1.225743	8.805871

となり，10 を超える変数がなくなっている．すなわち，高い多重共線性の存在もなくなっている．

1.8 共分散分析

1.8.1 データの概要：降圧剤データ

これは，10 名づつランダムに割り付けたうえで，2 種類の降圧剤(A,B)のいずれかを投与したときの，投与前後での収縮期血圧の値である．

薬剤 A	投与前	159	127	142	146	157	183	149	141	189	167
	投与後	158	126	137	134	148	176	136	131	177	151
薬剤 B	投与前	181	162	188	130	127	186	137	173	143	150
	投与後	162	149	173	122	110	159	129	141	124	144

このデータは，Pressure.csv で与えられる．

1.8.2 共分散分析の概要

胃癌患者に対する TS-1 補助化学療法による体重減少の抑制を意図した成分栄養剤服用の有効性を検討するために，成分栄養剤服用群(53 名)と非服用群(47 名)をランダムに割り付け，補助化学療法投与前と投与後 6 カ月の体重減少を検討している．図 1.20 は，X 軸を補助化学療法開始前の体重，縦軸を体重減少量としたときの散布図である(黒色：服用群，灰色：非服用群)．ここで，赤色の直線は服用群の観測値に当てはめた回帰直線であり，緑色の直線は非服用群の観測値に当てはめた回帰直線である．

いずれの群も補助化学療法投与前の体重が重いほど体重減少量が大きいことがわかる．アウトカムに影響を及ぼす要因が存在するとき，服用群と非服用群の体重減少量を 2 標本 t 検定によって単純に比較することはできず， TS-1 投与前の体重の影響を調整したもとで，成分栄養剤服用群と非服用群を比較しなければならない．アウトカムに影

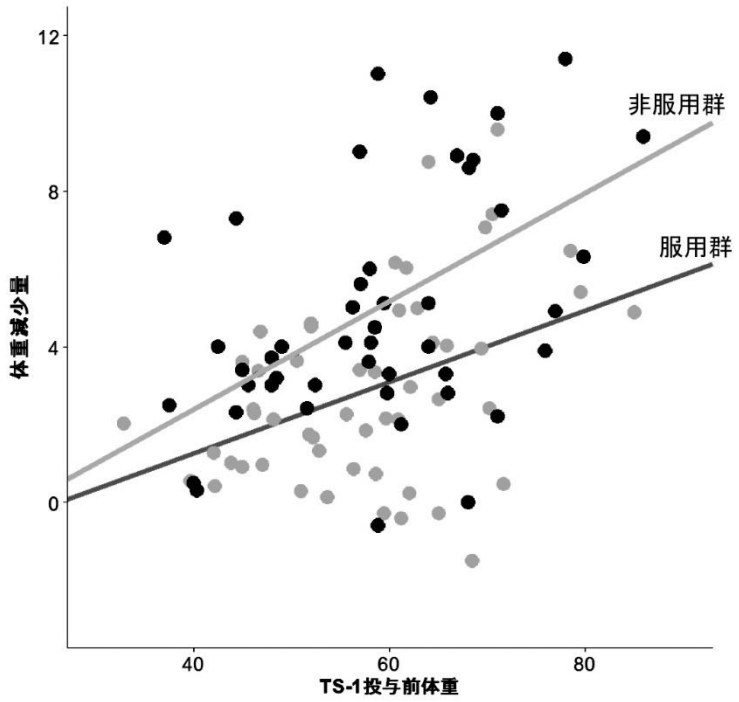


図 1.20：胃癌患者に対する TS-1 補助化学療法による体重減少の抑制を意図した成分栄養剤服用に関する無作為化比較第 III 相試験の結果（灰色：服用群，黒：非服用群）

響を及ぼす連続変数(共変量)の影響を調整したもとで群間を比較する統計的方法が共分散分析(ANCOVA; ANalysis of COVAriance)である.

いま，TS-1 投与前の体重を x とするとき，成分栄養剤服用群の回帰直線を $\beta_{0A}+\beta_{1A}x$，成分栄養剤非服用群の回帰直線を $\beta_{0C}+\beta_{1C}x$ とするとき，もし，$\beta_{0A}=\beta_{0C}$（TS-1 投与前の体重による体重減少量の大きさは成分栄養剤服用の有無に関わらず同じである)であることが仮定できれば(2 つの回帰直線が並行である)，成分栄養剤投与の有無による影響は β_{0A} と β_{0C} を比較すればよいことになる.

共分散分析では，成分栄養剤服用の有無による体重減少量の比較に先立って，2 つの回帰直線が並行であることを検定する. すなわち，帰無仮説 H_0「2 つの回帰直線が並行である」に対して，対立仮説 H_1「2 つの回帰直線が並行でない」が評価される. 有意であれば，2 つの回帰直線(TS-1 投与前体重による体重減少量の変化)の並行性の仮定を満たさないため，共分散分析による群間比較(成分栄養剤服用の有無による体重減少量の比較)を行うことができない. 群間で共変量の影響が異なることは，「交互作用がある」と呼ばれことから，この検定を交互作用検定と呼ぶことがある. 有意でなければ，2 つの回帰直線の並行性の仮定が否定できないとして，β_{0A} と β_{0C} の比較(共変量 x を調整した群間比較)を行う. すなわち，帰無仮説 H0「β_{0A} と β_{0C} が等しい」に対して，対立仮説 H1「β_{0A} と β_{0C} が等しくない」が検定される[18].

TS-1 補助化学療法胃癌患者に対する成分栄養剤投与に関する無作為化比較第 III 相試験のデータの場合には，回帰直線の並行性に対する検定の結果，p 値は 0.288 であることから，帰無仮説が受容されるため，並行性は否定できない(交互作用効果があるとは言えない). 次いで，回帰直線の切片が等しいか否かの検定(群間比較)を行う. その

[18] ここでは，2 群比較のデータを用いて共分散分析を説明しているが，3 群以上の場合にも用いることができる. その場合には，3 個以上の切片を比較することになり，解釈は 1 元配置の分散分析と同様に評価できる.

結果，p 値は 0.0002 であった．成分栄養剤服用群の回帰直線が成分栄養剤非服用群の回帰直線の下側に布置していることから，成分栄養剤を服用することで，TS-1 投与による体重減少を有意に抑制することが認められた．

1.8.3 EZR による共分散分析の実行

ここでは，投与前の収縮期血圧(投与前)を共変量としたうえで，降圧剤(降圧剤)による投与後の収縮期血圧(投与後)の違いを比較する．つまり，投与前の収縮期血圧によって調整した投与後の収縮期血圧に降圧剤が影響するか否か(降圧剤によって違いがあるか)を共分散分析により評価する．降圧剤データは，Pressure.csv で与えられる．

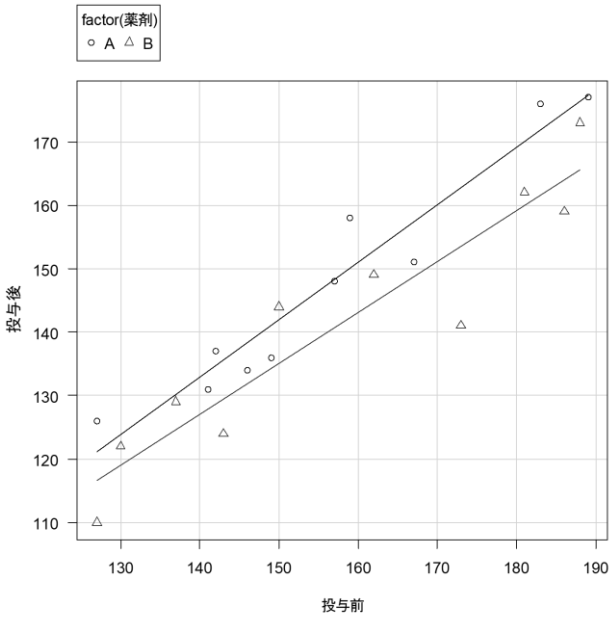
このとき，共分散分析は，次の手順で実行できる．

共分散分析(ANCOVA)の実行
1: 「統計解析」→「連続変数の解析」→「連続変数で補正した 2 群以上の間の平均値の比較」を選択する(共分散分析 ANCOVA).
2: 次のようなメニューが表示される.

このとき，
- 「目的変数(１つ選択)」で「投与後」を選択する．
- 「比較する群(1 つ選択)」で「薬剤」を選択する．
- 「補正に用いる連続変数(1 つ選択)」で「投与前」を選択する．

3: 「OK」ボタンを押す

このとき，次のような散布図が構成される．



このグラフにおいて，直線は，それぞれの群(A，B)に対して単回帰直線をあてはめたものである．いずれの降圧剤も，投与前の収縮期血圧が高いほど，投与後の収縮期血圧が高くなる傾向にある．また，この直線がおおよそ並行でなければ，共分散分析を実行することはできない．上図では，おおよそ並行になっていることから，並行性の仮定は，おおよそ満たしそうである．また，薬剤 A の直線に比べて，薬剤 B の直線のほうが下側に布置していることから，薬剤 B のほうが薬剤 A よりも降圧効果が期待できる．

EZR の出力では，様々な出力が表示される．表示された青色の箇所毎に説明する．

Output.1	群別変数と共変数の交互作用の P 値は 0.473

Output.1 これは，降圧剤×投与前の収縮期血圧の交互作用を検定した結果であり，有意な場合には，共分散分析による評価ができないことを意味する(共分散分析の仮定を満たさなくなるため)．その結果，p 値は 0.473 であり，有意水準 α=0.05 のもとで有意でないことから，このデータでは，共分散分析による評価が可能であることが示唆される．

Output.2

```
Anova Table (Type III tests)

Response: 投与後
              Sum Sq Df  F value    Pr(>F)
(Intercept)    80.1  1   2.0510   0.17024
factor(薬剤)  283.5  1   7.2634   0.01533 *
投与前       5917.0  1 151.6012 6.782e-10 ***
Residuals     663.5 17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output.2 は，共分散分析の結果である．降圧剤による影響は，factor(薬剤)を見ればよい．その結果，p 値は 0.01533 であることから，有意水準 α=0.05 のもとで有意である．したがって，降圧剤によって投与後の収縮期血圧に違いがあることが認められた．

ちなみに，もし，交互作用検定において，有意差が認められる場合(Output.1 が有意である場合)，EZR では，Output.2 が表示されない．なぜなら，共分散分析の仮定を満たさないからである．

2 章：質的データにおける統計解析

2.1 2 値変数に対する 1 標本データの解析：母比率に対する推測

（1）母比率の検定：臨床試験における動機とその意味

単アーム第 II 相試験では，ヒストリカル・コントロールに対する試験治療の有効性・安全性を検討する．ヒストリカル・コントロールの選定方法には，(1)これまでの文献等で得られた結果に基づいて選定する，(2)試験実施機関における臨床成績に基づいて選定する，などが考えられる．(1)によるヒストリカル・コントロールは，公表された情報に基づくことから，実施機関以外の研究者が確認することが可能である．一方で，(2)によるヒストリカル・コントロールは，実施機関以外の研究者が確認することはできない．したがって，(1)による設定のほうがエビデンスのある情報であるといえる．いずれにしても，臨床的な妥当性に基づいてヒストリカル・コントロールを設定することが重要である．

単アーム試験の評価には，母比率の検定(binomial test)を用いることができる．ここでは，胃癌患者に対する単アーム第 II 相試験の結果を用いる．Kurokawa et al.(2014)[19]は，HER2 陽性の進行・再発胃癌患者に対する S1+CDDP+Tmab の 3 剤併用療法(新規レジメン)の有効性・安全性を検討している．この試験では，フッ化ピリミジン系抗癌剤と CDDP を併用した既存レジメンでの奏効率 35%(ヒストリカル・コントロール)を閾値奏効率としたもとで，53 例の被験者に対して新規レジメンを実施し，奏効率を主要評価項目(primary endpoint)として評価している．

上記の臨床試験を例に母比率の検定を説明する．

帰無仮説 H_0「母比率は，ヒストリカル・コントロールでの比率と同じである(新規レジメンでの奏効率は閾値奏効率 35%と同じである)」

両側帰無仮説 H_{1a}「母比率は，ヒストリカル・コントロールでの比率と異なる(新規レジメンでの奏効率は閾値奏効率 35%と異なる)」

片側帰無仮説 H_{1b}「母比率は，ヒストリカル・コントロールでの比率を上回る(新規レジメンでの奏効率は閾値奏効率 35%よりも大きい)」

片側帰無仮説 H_{1c}「母比率は，ヒストリカル・コントロールでの比率を下回る(新規レジメンでの奏効率は閾値奏効率 35%よりも小さい)」

母比率の検定では， 3 種類の p 値の計算方法が提案されている：

---

[19] Kurokawa Y. et al.: Phase II study of trastuzumab in combination with S-1 plus cisplatin in HER2-positive gastric cancer (HERBIS-1), Br. J. Cancer, 110(5), 1163-1168, 2014.

(1)近似を用いる方法

(2)近似に補正(連続性の補正)を行う方法

(3)正確な p 値(exact p-value)をコンピュータを用いて計算する方法.

近年, コンピュータの性能の向上などに伴い, 正確な p 値を利用することが増えている. とくに, 単アーム試験では, 症例数(標本サイズ)が比較的小さいため, 殆どの論文で採用されている. 一方で, 症例数が 100 例以上の規模の大きな試験では, 正確な p 値の計算に膨大な時間を要するため, (1)あるいは(2)のいずれかを用いるほうが良い.

また, ヒストリカル・コントロールとの比較を意図して実施される単アーム試験の場合には, 片側対立仮説(片側 p 値)を用いることが多い. 単アーム試験が実施された状況は, ヒストリカル・コントロールが得られた状況(試験が実施された時期, 実施施設(国, 地域を含む), 被験者背景)と異なるため,「新規治療レジメンとヒストリカル・コントロールの差 ＝ 新規治療レジメンと既存治療レジメンの差」には必ずしもなり得ない. 既存治療レジメンに対する新規治療レジメンの有効性の差は無作為(ランダム)化比較第 III 相試験で検証する必要がある. ただし, 新規治療レジメンが閾値アウトカム(ヒストリカル・コントロールでの有効性)を上回らなければ, 資金と時間を要する無作為化比較試験を実施する意味がない. このことが, 片側対立仮説を用いる理由である.

(2) データの要約

あるクリニックで有効率が 0.6 といわれている薬剤を 10 人の患者に投薬したところ, 2 人しか有効でなかった. そこでこのクリニックでは, この薬剤の有効率が 0.6 より低いことが疑われた. 有効率が 0.6 より低いかどうかの検定を有意水準 0.05 で行いなさい(柳川・荒木, 2010)[20]. このデータは, Drug_efficacy.csv で与えられる. 因みに, Drug_efficacy.csv では, 有効症例を 1, 無効症例を 0 としている.

(3) EZR による母比率の検定の実行

ここでは,「Drug_efficacy.csv」のデータを用いて, 帰無仮説 H_0「あるクリニックにおける薬剤の有効率は 0.6 である」に対して, 片側対立仮説 H_1「あるクリニックにおける薬剤の有効率は 0.6 を下回る」を母比率の検定により評価する.

母比率の検定の実行
1: 「統計解析」→「名義変数の解析」→「1 標本の比率の検定」 を選択する. 2: 次のようなメニューが表示される.

[20] 柳川 堯・荒木由布子：バイオ統計の基礎―医薬統計入門 (バイオ統計シリーズ), 近代科学社, 2010.

このとき，
・「二値変数(1 つ選択)」で「結果」を選択する．
・「正確検定」にチェックを入れる(カイ 2 乗検定は不要)．
・「対立仮説」で「母比率 p<p0」を選択する．
・「帰無仮説」の右側の白枠に「0.6」と入力する．
3: 「OK」ボタンを押す

EZR では，正確検定(正確 p 値)とカイ 2 乗検定(母比率の検定と同じ意味である)を選択することができる．標本サイズが小さい場合には，正確検定を選択したほうが良いが，標本サイズが大きい(例えば，200 例以上を基準にしている文献が多い)場合には，カイ 2 乗検定で十分である．一方で，症例数が多くなると，いずれの検定でも p 値に大きな違いがない．また，カイ 2 乗検定では，近似計算を用いて p 値を算出する．「カイ 2 乗検定の連続性補正」は，近似計算に補正を行うことで，より正確性の高い p 値を計算している．いずれにしても，医学系研究では，正確検定を選択するほうが賢明である．

このとき，次のような出力が表示される．

```
１標本の比率の検定（母不良率の検定） P 値 = 0.0123
```

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される．赤色が R のコマンドであるが，無視してかまわない(EZR では，出力情報は，すべて青色で表示される)．また，上側の青色の出力は，R での解析結果を表しているが，下側の結果と内容が重複しているので割愛する．

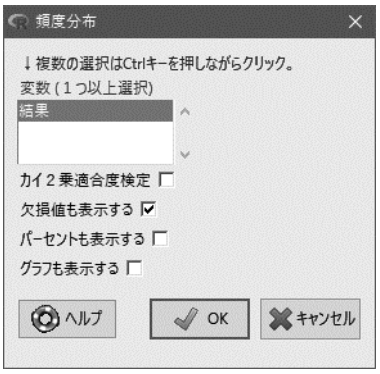
p 値は 0.0123 なので，有意水準 α=0.05 を下回るため，有意である．したがって，あるクリニックにおける薬剤の有効率は 0.6 を下回ることが分かった．

(4) EZR による母比率の信頼区間の実行

ここでは，母比率に対する 95%信頼区間の計算方法について述べる．データは，先ほどと同様に「Drug_efficacy.csv」を用いる．ただし，EZR による母比率に対する 95%信頼区間の計算では，ファイルではなく，標本サイズ(総サンプル数)とイベント数を手入力しなければならない．そのため，「頻度分布」を用いて，度数を計算する．

度数分布(頻度分布)計算の実行

1: 「統計解析」→「名義変数の解析」→「頻度分布」を選択する．
2: 次のようなメニューが表示される．

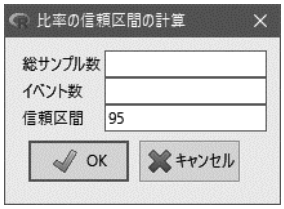


このとき，
・「変数(1 つ選択)」で「結果」を選択する．
3: 「OK」ボタンを押す

このときの結果を以下に示す．

```
0 1
8 2
```

すなわち，有効症例(1)が 2 例，無効症例(0)が 8 症例の合計 10 症例である．次いで，母比率の信頼区間を計算する．

母比率の信頼区間の実行
1:　「統計解析」→「名義変数の解析」→「比率の信頼区間の計算」を選択する. 2:　次のようなメニューが表示される.    このとき, ・「総サンプル数」に「10」と入力する. ・「イベント数」に「2」と入力する. ・「信頼区間」に「95」と入力する. 3:　「OK」ボタンを押す

このときの結果を以下に示す. このとき, 赤色が R のコマンドであるが, 無視してかまわない(EZR では, 出力情報は, すべて青色で表示される).

```
[1] 比率          : 0.2
[1] 95% 信頼区間                 : 0.025 - 0.556
```

その結果, 母比率の点推定値は 0.2 であり, 95%信頼区間は, [0.025, 0.556]であった.

2.2 クロス集計表による統計的推測

2.2.1 クロス集計表の概要

表 2.1 は, 上部消化管または下部消化管の開腹外科手術が施行された患者 558 名(上部消化管 413 例, 下部消化管 101 例)に対して, 真皮縫合群とステープラー群の 2 群に割り付けて, 手術後 30 日以内の創合併発現率を比較したランダム化比較第 III 相試験の結果である(Tsujinaka et al., 2013)[21]. これは, クロス集計表と呼ばれるものであり, 創合併発現の有無と介入群(真皮縫合群, ステープラー群)の関係を表しており, 例えば, 左上の 47 例は, 真皮縫合群でかつ創合併が発現した人数を表している. また, 下側の括弧は, 割合(行パーセント)と呼ばれる. その結果, 真皮縫合群のなかで創合併が発現した割合は 8.4%であり, ステープラーでは 11.5%であることから, 真皮縫合術のほうがステープラーに比べて創合併の発現割合が 3.1%低いことがわかる. 因みに, クロス集計表では, 列方向に介入(あるいは要因), 行方向に結果(アウトカム)を記載するのが一般的である.

2.2.2 オッズ比とリスク比

(1) オッズ比とリスク比の概要

リスク比(相対リスクと呼ぶこともある)とオッズ比は, それぞれリスク, オッズの 2 群間の比によって計算される. いま, 関心のあるイベントに対するリスクおよびオッズの定義は,

リスク = (関心のイベントが起きた被験者数) ÷ (被験者数)

オッズ = (関心のあるイベントが起きた割合) ÷ (関心のあるイベントが起きなかった割合)

で定義される. リスク比および, オッズ比はそれぞれの比率で表されることから, 表 2.1 の事例の場合には,

リスク比 = (真皮縫合術でのリスク) ÷ (ステープラーでのリスク) = 0.084÷0.115 = 0.730

オッズ比 = (真皮縫合術でのオッズ) ÷ (ステープラーでのオッズ) = 0.092÷0.130 = 0.710[22]

である. いずれの測度も「真皮縫合術はステープラーに比べて〇〇倍ほどイベント(創感染症)が起こる」と解釈される.

[21] Tsujinaka, T. et al. : Subcuticular sutures versus staples for skin closure after open gastrointestinal surgery: a phase 3, multicentre, open-label, randomised controlled trial. Lancet, 382(9898), 1105-1112, 2013.

[22] 真皮縫合術のオッズは 0.084/(1-0.084)=0.092 であり, ステープラーのオッズは 0.115/(1-0.115)=0.130 である.

表 2.1：開腹外科手術の縫合術に対するランダム化比較第 III 相試験の結果

	創合併あり	創合併なし	計
真皮縫合術	47 (8.4%)	511 (91.6%)	558
ステープラー	59 (11.5%)	455 (85.5%)	514
計	106 (9.9%)	966 (90.1%)	1072

リスク比のほうが解釈しやすそうだが，その利用はコホート研究に限定される．疫学研究の縦断研究には，コホート研究とケース・コントロール研究がある．いま，肺癌と喫煙習慣の関係を調査したいと考える．コホート研究の場合には，喫煙習慣のある被験者と喫煙習慣のない被験者に分けて，その後の経過を追跡し，肺癌に罹患したかどうかを調査する(原因で群分けを行い，その後の経過(結果の有無)を調査する)．ケース・コントロール研究では，肺癌に罹患した被験者とそうでない被験者のデータを集め，喫煙習慣がなかったかを調査する(結果で群分けを行い，原因の有無を調査する)．リスクの定義をみればわかるように，リスクの計算には被験者数が必要になる．コホート研究では，原因をもとに被験者を集めるが，ケース・コントロール研究では，結果をもとに被験者を集める．そのため，被験者数の適切な集計を行うことができない．

また，リスク比には，数学的な問題もある．一つは，リスク比の場合には，0 倍～有限倍(分母のリスクで上限が決まる)までしか定義域がないのに対して，オッズ比の場合には，0 倍～∞倍まで定義可能である．もう一つは，ラベル付けの問題である．関心のあるイベントではなく，関心のあるイベントの非発現を考えると，「〇〇倍ほど皮膚かぶれになる」といっていたものが「1/〇〇倍ほど関心のあるイベントが起きない」と逆数での解釈になるはずだが，リスク比ではこのような数値にはならない．

さらに，今回は解説しないがロジスティック回帰分析では，オッズ比による解釈をおこなうため，最近では，研究の形式に依らず，オッズ比を用いることが多くなってきている．

2.2.3 クロス集計表の形式と手法の取捨選択

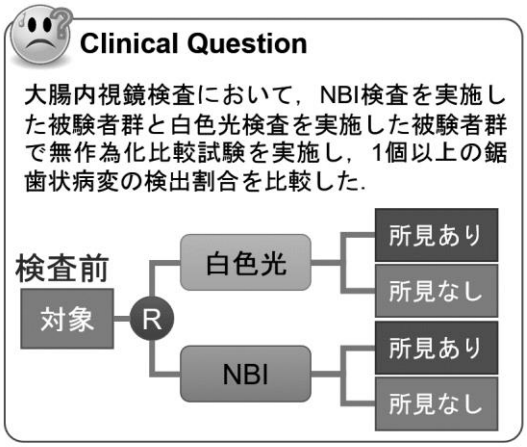
図 2.1 は，様々なシチュエーションと検定の関係をあらわしている．図 2.1(a)は，異なる内視鏡検査を実施した 2 群(白色光，NBI)のアウトカム(所見の有無)を評価している．このような状況には，無作為化比較試験などがある．この場合には，2.2.4 節のカイ 2 乗検定，あるいは 2.2.3 節の Fisher の正確検定を用いることができる．

図 2.1(b)は，同一の被験者に対して 2 種類の介入を実施した場合である．このような状況には，治療前後でのアウトカムの比較，あるいはクロスオーバー試験などがある．このように，同一被験者から複数のアウトカムを取得する場合(対応があるデータ)の解析には，2.4.1 節の McNemar 検定を用いることができる．

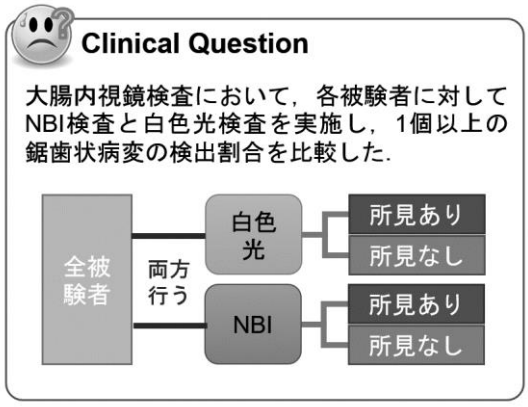
図 2.1(c)は，図 2.1(b)に類似しているが，介入が 3 種類(3 群)以上存在する場合である．このようなデータの解析には，2.4.2 節の Cochran の Q 検定を用いることができる．

図 2.1(d)は，アウトカムが 2 値以上のカテゴリで構成されている場合である．このようなデータの解析には，2.4.1 節の McNemar 検定を用いることができる．

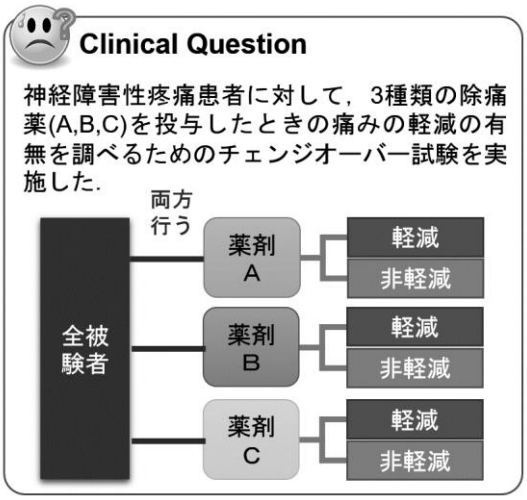
図 2.1(e)は，介入されるカテゴリに順序関係が存在する場合である．例えば，この事例の場合には，薬剤の投与量が 5 カテゴリに分けられ，それぞれの投与群において，2 値アウトカム(治療の有効・無効)がとられている．このときの



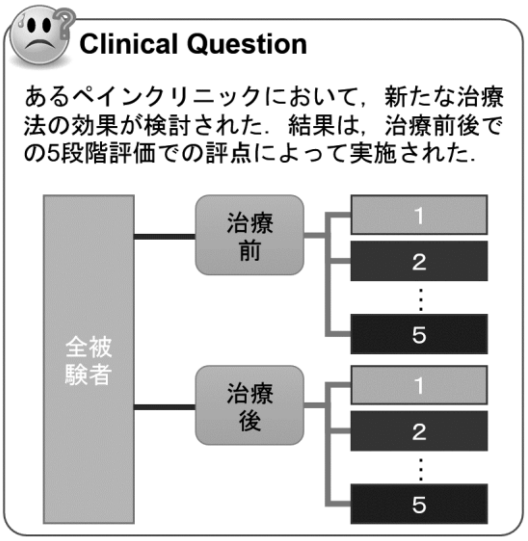
(a) カイ 2 乗検定(Fisher の正確検定)の適用場面



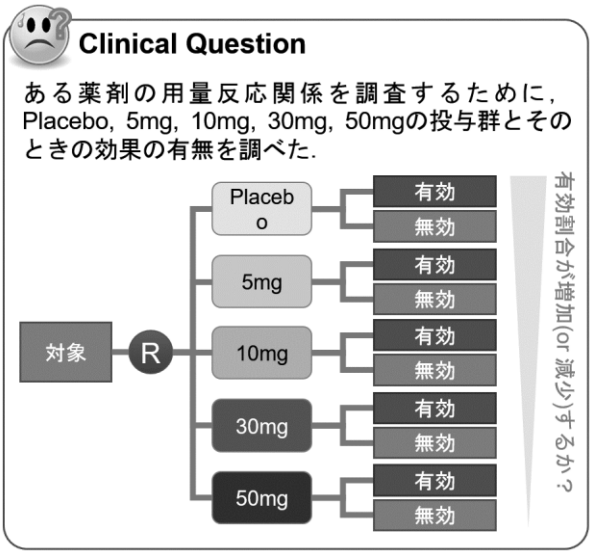
(b) McNemar 検定の適用場面(1)



(c) Cochran の Q 検定の適用場面



(d) McNemar 検定の適用場面(2)



(e) Cochran–Armitage 検定の適用場面

図 2.1：クロス集計表におけるデータの形式と検定の取捨選択の関係

関心は，薬剤投与量に対して，有効割合が増加するかどうかにある．このような傾向変化を評価するには，2.3 節の Cochran-Armitage 検定を用いる．

2.2.4 カイ 2 乗検定

表 2.2 は，大腸内視鏡検査における 2 種類の検査方法(NBI 検査，白色光検査)の診断能を比較するためのランダム化比較第 III 相試験 [2)]の結果である．ここでの評価項目は，近位結腸の鋸歯状病変を 1 個以上検出した被験者割合(病変検出割合)である．NBI の病変検出割合は 51.1%であり，白色光では 39.4%であることから，NBI のほうが白色光よりも診断能が優れている．要因(大腸内視鏡検査の方法)の質的アウトカム(病変検出の有無)への影響を評価する統計的方法がカイ 2 乗検定である．カイ 2 乗検定では，帰無仮説 H_0「要因(NBI，白色光)によるアウトカム(病変の有無)に違いがない」に対して対立仮説 H_1「要因(NBI，白色光)によってアウトカム(病変の有無)に違いがある」を検定する[23]．

カイ 2 乗検定では，帰無仮説 H_0 が正しいと仮定したもとでのクロス集計表の度数と試験で得られた度数の違いを用いて評価する．ここで，帰無仮説を仮定したときの度数は期待度数と呼ばれる．表 2.3 は期待度数の計算方法を表している．期待度数は，周辺度数[24]の積を症例数(総度数)で割ったものである．そして，クロス集計表の全てのセルに対して(観測度数-期待度数)^2/(期待度数)を計算し，それらの総和を用いることで p 値を求めることができる．すなわち，カイ 2 乗検定とは，「観測されたクロス集計表が帰無仮説からどの程度離れているか」を評価する検定方法である．

大腸内視鏡検査に対するランダム化比較第 III 相試験において，カイ 2 乗検定の p 値は 0.001 であることから，有意である．NBI のほうが白色光よりも病変検出割合が高いことから，NBI の診断能が優れていることが認められる．

カイ 2 乗検定でも p 値の計算には近似を用いる．そのため，近似による p 値の計算方法には，(母比率の検定と同様に)連続性の補正を行う場合と行わない場合の 2 種類が存在する．このとき，カイ 2 乗検定における補正の方法を

表 2.2：大腸内視鏡検査における 2 種類の検査方法に対する近位結腸の鋸歯状病変の発現に関するクロス集計表

	病変あり	病変なし	計
NBI	204 **(51.1%)**	195 **(48.9%)**	399
白色光	158 **(39.4%)**	243 **(60.6%)**	401
計	362 **(45.2%)**	438 **(54.2%)**	800

表 2.3：大腸内視鏡検査に対するランダム化比較第 III 相試験における期待度数

	病変あり	病変なし	計
NBI	$\frac{399\times362}{800}=180.5$	$\frac{399\times438}{800}=218.5$	399
白色光	$\frac{401\times362}{800}=181.5$	$\frac{401\times438}{800}=219.5$	401
計	362	438	800

行周辺度数

列周辺度数　　総度数

[23] 統計学の教科書では，カイ 2 乗検定を独立性の検定と呼ぶことがある．独立性とは「要因にアウトカムが影響を受けない」ことを意味する．したがって，帰無仮説 H_0「アウトカム(病変の有無)は要因(大腸内視鏡検査の方法)に対して独立である」に対して対立仮説 H_1「アウトカム(病変の有無)はアウトカム(病変の有無)は要因(大腸内視鏡検査の方法)に対して独立でない」と書くことができる．

[24] クロス集計表では，行方向の合計値を行周辺度数(NBI：399，白色光：401)，列方向の合計値を列周辺度数(病変あり：362，病変なし：438)という．

表 2.4:開腹外科手術の縫合術に対するランダム化比較第 III 相試験の結果

	奏効	非奏効	計
新規抗癌剤	**12 (54.5%)**	10 (45.5%)	22
既存抗癌剤	3 (20.0%)	12 (80.0%)	15
計	**15 (40.5%)**	**22 (59.5%)**	**37**

Yates の補正という. 上記の p 値は Yates の補正を行った場合の p 値であり, 統計パッケージでは補正後の値を採用することが多い. なお, Yates の補正を行わない場合の p 値は 0.0009 であることから, Yates の補正後の p 値のほうが若干大きくなる.

因みに, 母比率の検定に対して, 2 標本での母比率を比較する検定(母比率の差の検定)があるが, その結果はカイ 2 乗検定に一致する

2.2.5 Fisher の正確検定

表 2.4 は, ある病院において実施された胃癌患者に対する新規抗癌剤(22 例)と既存抗癌剤 (15 例)のパイロット試験(仮想例)の結果である. 前項で述べたように, カイ 2 乗検定では p 値の計算に近似を用いる. この近似の精度は症例数が小さくなるほど悪くなるため, 本試験のような少数例(37 例)の場合には適切でない. このような場合に用いられる方法が Fisher の正確検定(Fisher's exact test)である. つまり, Fisher の正確検定での仮説はカイ 2 乗検定と同である. 本事例の場合には, 帰無仮説 H_0「抗癌剤(新規, 既存)によるアウトカム(奏効割合)に違いがない」に対して対立仮説 H_1「抗癌剤(新規, 既存)によるアウトカム(奏効割合)に違いがある」を検定する[25].

2×2 クロス集計表[26]では, 周辺度数(太線で囲んだ部分)を固定すると, 1 個のセルが決まれば, その他のセルは全て決まる. 例えば, 表 2.4 の場合には,「新規」かつ「奏効」の被験者数(緑色の数値)が 12 例であることが決まれば,「新規」かつ「非奏効」は 22−12=10,「既存」かつ「奏効」は 15−12=3,「既存」かつ「非奏効」は 37−(12+10+3)=12 である. このことを利用すると, 考えられ得る全てのクロス集計表のパターンは 1 個のセルの数値を用いて表すことができる. Fisher の正確検定では, 全てのクロス集計表のパターンに対して, 帰無仮説 H_0 が正しいと仮定したもとで, それぞれのクロス集計表が得られる確率を計算する. そして, 当該試験で得られたクロス集計表とそれよりも小さな確率をもつ(極端な)クロス集計表が得られる確率を合計することで p 値を求めることができる. 図 2.2 は, 本事例の考え得る全てのクロス集計表及び, 帰無仮説が正しいと仮定したときの確率である. 本事例のクロス集計表(黒色の枠で囲まれたクロス集計表)が得られる確率は 0.0314 である. 薄色の背景で囲まれたクロス集計表は, 本事例の結果よりも極端なものを表している(確率が 0.0314 よりも小さなクロス集計表). p 値は, これらの確率の総和であることから,

p 値=0.0000 + 0.0000 + 0.0000 + 0.0001+ 0.0011 + 0.0084 + 0.0314 + 0.0056 + 0.0005 + 0.0000 = 0.0471

[25] 脚注 a と同様に独立性のもとで仮説を考えると, 帰無仮説 H_0「アウトカム(奏効割合)は抗癌剤(新規, 既存)に対して独立である」に対して両側対立仮説 H_1「アウトカム(奏効割合)は抗癌剤(新規, 既存)に対して独立である」と書くことができる.

[26] クロス集計表では, (行のセル数)×(列のセル数)クロス集計表と呼ぶことが多い. 表 2.4 では, 行のセル数, 縦のセル数ともに 2 個のなので 2×2 クロス集計表になる. 2×2 クロス集計表のみを用いて Fisher の正確検定を説明している文献が多いものの, それ以上のセル数が存在する場合にも計算できる.

図 2.2：開腹外科手術の縫合術に対するランダム化比較第 III 相試験の結果

である．有意水準 α=0.05 よりも小さいことから，抗癌剤の種類による奏効割合に違いが認められる．なお，このときのカイ 2 乗検定の p 値は 0.0783 であることから有意でない．本事例がパイロット試験で実施された少数例の臨床試験であることを考えると，カイ 2 乗検定による統計解析は適切でない．カイ 2 乗検定を誤って用いた場合，本来は抗癌剤の種類によって奏効割合に違いが認められるにも関わらず，違いがないと解釈することになる．他方，症例数が大きい場合には，カイ 2 乗検定と Fisher の正確検定の結果はほぼ一致する．

また，Fisher の正確検定について，Cochran[27]は，(1)期待度数が 1 未満のセルが 1 個以上存在する場合，(2)期待度数が 5 未満のセルが全体のセル数の 20%以上存在する場合，には Fisher の正確検定を用いるほうが良いことを指摘している．例えば，ランダム化比較第 III 相試験のような規模の大きな試験であっても，数パーセント程度の感染症の発現割合を群間で比較する場合には Fisher の正確検定のほうが適切である．

2.2.6 EZR によるクロス集計表及び検定の実行

(1) データの概要

テープ剥離に用いるベンジンが皮膚かぶれの原因と考え，剥離材にオリーブ油を利用した．このとき，剥離剤(ベンジン・オリーブ油)と皮膚かぶれの発現の有無には違いがあるだろうか．このデータは，Skin_rash.csv に保存されている．変数は，テープ離脱(ベンジン，オリーブ油)，皮膚かぶれ(あり，なし)である．

[27] Cochran W.G. : Some methods for strengthening the common χ^2 tests, Biometrics, 10, 417-451, 1954.

		皮膚かぶれ		
		あり	なし	合計
オリーブ油	例数	24	100	124
	全体パーセント	10.0	41.8	51.9
	列パーセント	32.4	60.6	51.9
	行パーセント	19.4	80.6	100.0
ベンジン	例数	50	65	115
	全体パーセント	20.9	27.2	48.1
	列パーセント	67.6	39.4	48.1
	行パーセント	43.5	56.5	100.0
合計	例数	74	165	239
	全体パーセント	31.0	69.0	100.0
	列パーセント	100.0	100.0	100.0
	行パーセント	31.0	69.0	100.0

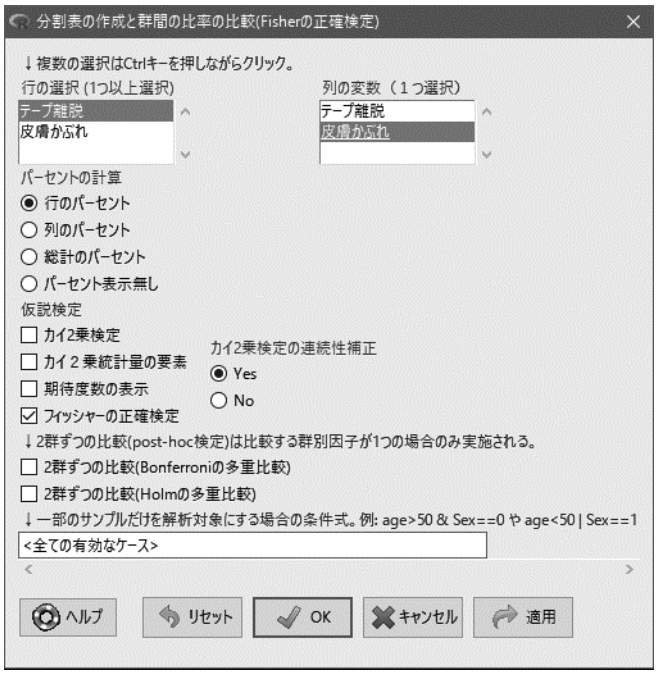
(2) EZR による計算

ここでは，EZR によるクロス集計表の作成，及び検定(Fisher の正確検定，カイ 2 乗検定)の方法について述べる．

クロス集計表の作成及び検定の方法

1: 「統計解析」→「名義変数の解析」→「分割表の作成と群間の比率の比較(Fisher の正確検定)」を選択する．

2: 次のようなメニューが表示される．



このとき，

- 「行の選択(1 つ以上選択)」で「テープ離脱」を選択する．
- 「列の選択(1 つ選択)」で「皮膚かぶれ」を選択する．
- 「パーセントの計算」で「行のパーセント」を選択する．
- 「仮説検定」で「カイ 2 乗検定」と「フィッシャーの正確検定」にチェックを入れる．
- 「カイ 2 乗検定の連続性補正」で「Yes」を選択する．

3: 「OK」ボタンを押す

ここで，「行」は群(説明変数)を表し，「列」はアウトカムを表す．また，パーセントの意味は，以下のとおりである．

・全体パーセント(EZR では総計のパーセント):被験者全体のなかで，何パーセントの被験者が各セルに属しているかを表す．例えば，(オリーブ油，あり)のセルの場合には，「オリーブ油を剥離剤に用いて，かつ皮膚かぶれになった割合は，10.0%である」と解釈される．

・列パーセント(EZR では列のパーセント):皮膚かぶれの有無で分けたときの，それぞれの剥離剤の割合を表している(縦方向に 100%になるように計算している)．例えば，(オリーブ油，あり)のセルの場合には，「皮膚かぶれになった被験者のうち，32.4%がオリーブ油を用いた」と解釈される．

・行パーセント(EZR では行のパーセント):剥離剤の種類で分けたときの，皮膚かぶれの有無の割合を表している(横方向に 100%になるように計算している)．例えば，(オリーブ油，あり)のセルの場合には，「オリーブ油を剥離剤に利用した被験者のうち，19.4%が皮膚かぶれになった」と解釈される．

仮説検定では，カイ 2 乗検定と Fisher の正確検定の二つを選択したが，実際のデータ解析では，いずれか一方のみを用いればよい．このとき，カイ 2 乗検定の連続性補正とは，2.2.4 節での Yates の補正を表す．

「分割表の作成と群間の比率の比較(Fisher の正確検定)」では，複数の出力(青色の部分が複数存在する)が表示される．ここでは，R 及び EZR での計算結果(青色の部分)のみを解釈する．

Output.1

```
              皮膚かぶれ
テープ離脱    あり なし
  オリーブ油    24  100
  ベンジン      50   65
```

Output.1 は，クロス集計表による要約の結果である．つまり，

	皮膚かぶれあり	皮膚かぶれなし
オリーブ油	24	100
ベンジン	50	65

である．

Output.2

```
              皮膚かぶれ
テープ離脱    あり なし Total Count
  オリーブ油 19.4 80.6   100   124
  ベンジン   43.5 56.5   100   115
```

Output.2 は，クロス集計表の列パーセント(EZR では行のパーセント)を表している．つまり，

	皮膚かぶれあり	皮膚かぶれなし
オリーブ油	19.4%	80.6%
ベンジン	43.5%	56.5%

である．オリーブ油のほうが，ベンジンよりも皮膚かぶれの割合が低いことが示唆される．因みに，Count は各群の症例数を表している．

Output.3

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

data:  .Table
X-squared = 15.135, df = 1, p-value = 0.0001001
```

Output.3 は，Yates の補正を伴うカイ 2 乗検定の結果である．帰無仮説 H_0 は「剥離剤の種類と皮膚かぶれの有無には関連性がない(剥離剤の種類によって皮膚かぶれの有無に違いがない)」に対して，対立仮説 H_1 は「剥離剤の種類と皮膚かぶれの有無には関連性がある(剥離剤の種類によって皮膚かぶれの有無に違いがある)」である．「p-value」が p 値を表している(p値=0.0001001 である)．有意水準 α=0.05 を下回ることから，剥離剤によって皮膚かぶれの有無に違いが認められる．

Output.4

```
        Fisher's Exact Test for Count Data

data:  .Table
p-value = 0.00007754
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1669801 0.5767046
sample estimates:
odds ratio
 0.3135849
```

Output.4 は，Fisher の正確検定の結果である．カイ 2 乗検定と同様に，帰無仮説 H_0「剥離剤の種類と皮膚かぶれの有無には関連性がない(剥離剤の種類によって皮膚かぶれの有無に違いがない)」に対して，対立仮説 H_1「剥離剤の種類と皮膚かぶれの有無には関連性がある(剥離剤の種類によって皮膚かぶれの有無に違いがある)」を検定している．「p-value」が p 値を表している(p値=0.00007754 である)．有意水準 α=0.05 を下回ることから，剥離剤によって皮膚かぶれの有無に違いが認められる．また，「odds ratio」下側の 0.3135849 が(オリーブ油)/(ベンジン)のオッズ比であり，「95 percent confidence interval」下側の 0.1669801 0.5767046 がオッズ比に対する 95%信頼区間である．すなわち，オッズ比[95%信頼区間]を小数点以下 3 桁で四捨五入すると，0.314 [0.167, 0.577]である．また，95%信頼区間が 1.00(オリーブ油とベンジンで皮膚かぶれの罹患が同じ)を含んでいないことから，オリーブ油の皮膚かぶれに対する罹患リスクはベンジンに比べて有意に小さいと言える．

Output.5

	皮膚かぶれ=あり	皮膚かぶれ=なし	Fisher 検定の P 値
テープ離脱=オリーブ油	24	100	0.0000775
テープ離脱=ベンジン	50	65	

Output.5 は，Output.1 のクロス集計表と Ouput.4 の Fisher の正確検定の結果を表示した EZR での出力である．説明が重複するため，内容の解釈は割愛する．

(3) EZR によるクロス集計表の直接入力による計算

EZR では，クロス集計表を直接入力して計算することができる．先ほどの皮膚かぶれのデータのクロス集計表は，

	皮膚かぶれあり	皮膚かぶれなし
オリーブ油	24	100
ベンジン	50	65

であった．これを EZR に直接入力した場合でも，同様の解析が実行できる．

直接入力による解析では，先ず，「統計解析」→「名義変数の解析」→「分割表の直接入力と解析」を選択する．すると，メニューが表示されるので，次のように入力する．

STEP1:「数を入力:」の下側のセルに次のように入力する．

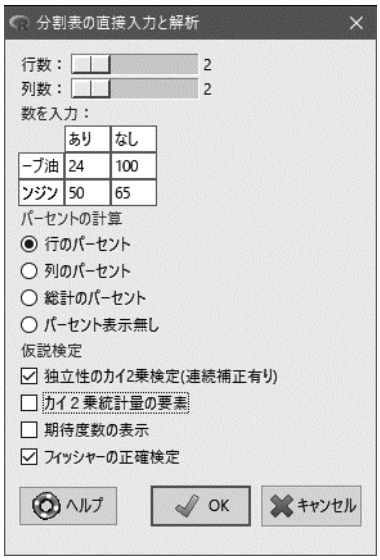
	あり	なし
オリーブ油	24	100
ベンジン	50	65

なお，セルの大きさの関係で，「オリーブ油」と「ベンジン」と入力すると，最初の文字が見えなくなるが，計算には問題はない．また，今回は，2×2 分割表なので，「行数」と「列数」はデフォルトになるが，3 群以上(行数が 3 以上)の場合には，「行数」のスライダーを右側に動かすと「数を入力:」の行数が増加し，アウトカムが 3 カテゴリ以上の場合には，「列数」のスライダーを右側に動かすと「数を入力:」の列数が増加する．

STEP2:「パーセントの計算」で「行のパーセント」を選択する(列パーセントが表示される)．

STEP3:「仮説検定」で「独立性のカイ 2 乗検定(連続補正有り)」及び「フィッシャーの正確検定」にチェックする．

これらの作業を行った場合，メニューは，次のようになる．



「OK」ボタンを押すと，先ほどの場合と同様の出力が得られる．

2.3 傾向変化の検定：Cochran-Armitage 検定

2.2.1 Cochran-Armitage 検定の概要

表 2.5 は，ある薬剤と有害事象の発現の有無を調査した研究の結果(仮想例)である．薬剤の用量が増加するにつれて有害事象の発現割合が増加していることがわかる．このように，薬剤の用量に対する反応(有害事象の発現)を評価する研究は，用量−反応試験と呼ばれ，医薬品開発における初期の第 II 相試験などに用いられる．このとき，要因に対する 2 値で得られた事象の発現割合の傾向性を評価するための統計的検定が Cochran−Armitage 検定(Cochran−Armitage 傾向性検定)である ．

Cochran−Armitage 検定では，帰無仮説 H_0「要因(薬剤の投与量)の変化に対して事象(有害事象)の発現割合に変化はない」に対して対立仮説 H_1「要因(薬剤の投与量)の変化に対して事象(有害事象)の発現割合に変化がある 」を検討する．ここで，「変化」とは，例示の場合には，薬剤の用量が増加するほど副作用の発現割合が増加(あるいは減少)することを意味する．Cochran−Armitage 検定による評価は，要因を説明変数，各要因のカテゴリ毎(Placebo，5mg，10mg,. .を 1,2,3,. .)での事象の発現割合を応答変数としたもとで単回帰分析を行い，その単回帰分析の傾き(つまり要因に対して増加傾向・減少傾向があるか)を検討することと同様である．

表 2.5 の例の場合には，p 値が<0.001 であることから(有意である)，薬剤の用量が増加するほど有害事象の発現頻度が傾向変化することが認められる．そして，その傾向変化は各要因のカテゴリ毎の発現割合から増加傾向にあることがわかる．

表 2.5：ある薬剤と有害事象の発現の有無を調査した研究の結果

	Placebo	5mg	10mg	30mg	50mg
有効	19	81	169	318	379
無効	23,531	13,234	24,332	30,514	20,432
発現割合	0.08%	0.61%	0.69%	1.04%	1.85%

2.2.2 EZR による Cochran-Armitage 検定の実行

(1) データの概要

女性の肺癌患者 108 名，対照群 108 名を選出し，喫煙量(0，1～4 本，5～14 本，15 本～)について調査したケース・コントロール研究の結果がある(丹後，2013)．肺癌と喫煙量に関連性があるかどうかを検討しなさい．なお，このデータは，breast_smok.e.csv で与えられる．このデータの変数は，群(肺癌，対照)，喫煙量(0:0 本，1:1～4 本，2:5～14 本，3:15 本)である．

(2) EZR による計算

ここでは，先ず，喫煙量のダミー変数にカテゴリ化を行う．

変数のカテゴリ化

1: 「アクティブデータセット」→「変数の操作」→「連続変数を因子に変換」を選択する．

2: 次のようなメニューが表示される．

このとき，

- 「変数(1 つ以上選択)」で「喫煙」を選択する．
- 「因子水準」で「水準名」を選択する．
- 「新しい変数名または複数の変数に対する接頭文字列」で「喫煙カテゴリ」と入力する(任意)．

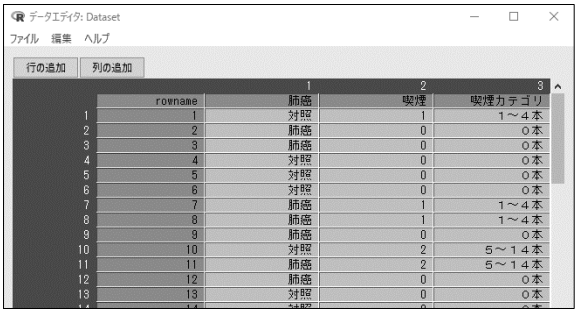
3: 「OK」ボタンを押すと，次のような新たなメニューが表示させる．

このとき，

- 「数値」の「0」で「０本」と入力する．
- 「数値」の「1」で「１～４本」と入力する．
- 「数値」の「2」で「５～１４本」と入力する．
- 「数値」の「3」で「１５本以上」と入力する．

3: 「OK」ボタンを押す

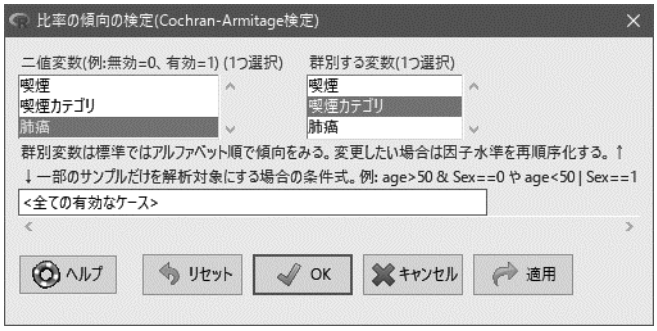
これにより，新たに，「喫煙カテゴリ」という新たな変数が追記される．確認する場合には，メニュー下の「編集」あるいは「編集」を押せばよい．表示を押した場合には，



	1	2	3
	rowname	肺癌	喫煙	喫煙カテゴリ
1	1	対照	1	１～４本
2	2	肺癌	0	０本
3	3	肺癌	0	０本
4	4	対照	0	０本
5	5	対照	0	０本
6	6	対照	0	０本
7	7	肺癌	1	１～４本
8	8	肺癌	1	１～４本
9	9	肺癌	0	０本
10	10	対照	2	５～１４本
11	11	肺癌	2	５～１４本
12	12	肺癌	0	０本
13	13	対照	0	０本

が表示される(上記は一部である).

次いで, Cochran-Armitage 検定を実行する.

Cochran-Armitage 検定の実行
1: 「統計解析」→「名義変数の解析」→「比率の傾向の検定」を選択する. 2: 次のようなメニューが表示される.    このとき, ・「二値変数(例：無効＝0, 有効=1)(一つ選択)」で「肺癌」選択する. ・「群別する変数(1 つ選択)」で「喫煙カテゴリ」を選択する. 3: 「OK」ボタンを押す

「分割表の作成と群間の比率の比較(Fisher の正確検定)」では, 複数の出力(青色の部分が複数存在する)が表示される. ここでは, R 及び EZR での計算結果(青色の部分)のみを解釈する.

Output.1

```
                肺癌
喫煙カテゴリ 対照 肺癌
  ０本        59   40
  １～４本    25   16
  ５～１４本  18   24
  １５本以上   6   28
```

Output.1 は, クロス集計表による要約の結果である. つまり,

	0 本	1～4本	5～14 本	15 本以上
対照	59	25	18	6
肺癌	40	16	24	28

である.

この出力の下側の R の出力(「Chi-squared Test for Trend in Proportions」と記載された部分)は, EZR による最後の出力と同じ内容なので割愛する.

Cochran-Armitage 検定では, 帰無仮説 H_0「喫煙量に対して肺癌の罹患率に変化はない」に対して対立仮説 H_1「喫煙量に対して肺癌の罹患率に変化がある 」を検討する. その結果,

Output.2 比率の傾向の検定(Cochran-Armitage 検定) P 値 = 0.0000332

より, 有意な結果が得られた. したがって, 喫煙量の変化に対して, 肺癌の罹患率が変化することが分かった.

2.4 カテゴリカル変数に対する対応があるクロス集計表の解析

2.4.1 対応のあるクロス集計表・対応のある 2 値アウトカムの 2 群比較

2.4.1.1 対応のあるクロス集計表

通常のクロス集計表では, 例えば, 被験者を 2 群以上に分けられたもとで(あるいはランダムに 2 種類以上の介入を割付けたもとで), それぞれの群に異なる介入を行い, アウトカムを評価している. そのため, 対応のないクロス集計表では, 列(縦方向)に要因, 行(横方向)にアウトカムを配置したうえで作成される. つまり, 対応のないクロス集計表によってまとめられる研究では, 2 種類の介入のいずれかのみが被験者に実施される.

表 2.6：大腸内視鏡検査の臨床研究結果に対する対応のあるクロス集計表(括弧内は総パーセント)

		白色光		合計
		あり	なし	
NBI	あり	(a) **85** **[48.9%]**	(b) **18** **[10.3%]**	**103**
	なし	(c) **10** **[5.7%]**	(d) **61** **[35.1%]**	**71**
合計		**95**	**79**	**174**

これに対して，対応がある場合には，介入前後のアウトカムを比較する場合や，あるいは同一被験者に異なる治療・検査法が施される．そのため，すべての被験者に対して両方の介入が行われる．したがって，対応がない場合とクロス集計表の構成が異なる．

表 2.6 は，大腸内視鏡検査の NBI 検査と白色光検査で鋸歯状病変の検出の有無を比較した臨床研究の結果に対して，対応のあるクロス集計表を作成したものである(仮想例)．この研究は，定期検診受診者のなかで要精密検査と診断された 50 歳以上の被験者に対して，NBI による大腸検査と白色光による 2 種類の大腸検査の両方を実施している．この対応のあるクロス集計表では，列(縦方向)に NBI 検査における鋸歯状病変の検出の有無，行(横方向)に白色光検査における鋸歯状病変の検出の有無を配置している．すなわち，それぞれのセルの解釈は以下のとおりである．

(a) NBI 検査，白色光検査のいずれでも鋸歯状病変ありと診断された被験者数は 85 例

(b) NBI 検査では鋸歯状病変ありと診断されたが，白色光では鋸歯状病変なしと診断された被験者数は 18 例

(c) NBI 検査では鋸歯状病変なしと診断されたが，白色光では鋸歯状病変ありと診断された被験者数は 10 例

(d)) NBI 検査，白色光検査のいずれでも鋸歯状病変なしと診断された被験者数は 61 例

対応のあるクロス集計表では，列パーセント点あるいは行パーセント点を用いることはなく，総パーセント点のみが利用される．例えば，NBI 検査，白色光検査のいずれでも鋸歯状病変ありと診断された被験者(表 1 のセル(a))の割合は，48.9%であると解釈される．

2.4.1.2 対応のある 2 値アウトカムに対する 2 群の比較：McNemar 検定

対応のある 2×2 クロス集計表において，2 種類の介入によるアウトカムにおける事象の発現率を比較する方法が McNemar 検定である．すなわち，McNemar 検定では，帰無仮説 H_0「介入によるアウトカムの事象の発現率に違いがない」に対して対立仮説 H_1「介入によるアウトカムの事象の発現率に違いがある」を検定する．表 2.6 の事例において，NBI 検査で「病変あり」と診断される(真の)確率をp_{NBI}，白色光検査で「病変あり」と診断される(真の)確率をp_{WH}とする．このとき，上記の仮説は，帰無仮説 H_0「$p_{\mathrm{NBI}} - p_{\mathrm{WH}} = 0$」に対して，対立仮説 H_1「$p_{\mathrm{NBI}} - p_{\mathrm{WH}} \neq 0$」を検定することを意味する．実際に得られた試験結果で診断能を考える．試験結果において，NBI 検査で「病変あり」と診断された割合を$\hat{p}_{\mathrm{NBI}}$，白色光検査で「病変あり」と診断される割合を$\hat{p}_{\mathrm{WH}}$とするとき，これらの割合は

$$\hat{p}_{\mathrm{NBI}} = \frac{\text{セル(a)} + \text{セル(b)}}{N},\ \hat{p}_{\mathrm{WH}} = \frac{\text{セル(a)} + \text{セル(c)}}{N}$$

である．ここに，Nは被験者数を表す．McNemar 検定では，これらの割合の差$\hat{\Delta}$を検討することになるので，

$$\hat{\Delta} = \hat{p}_{\mathrm{NBI}} - \hat{p}_{\mathrm{WH}} = \frac{\text{セル(b)} - \text{セル(c)}}{N}$$

になり，セル(b)とセル(c)を比較すればよいことになる[28].

表 2.6 の事例における McNemar 検定での p 値は 0.185 なので，有意でなかった．つまり，内視鏡検査(NBI 検査，白色光検査)によって診断能(病変の検出率)に差異があるとはいえなかった．また，NBI 検査における所見ありの割合は 59.2%(103/174)であり，白色光による所見ありの割合は 54.6%であることから，NBI 検査と白色光検査では，5%程度の差異であった．

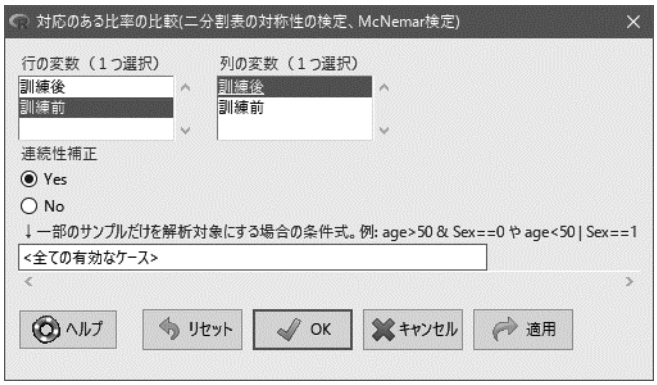
2.4.1.2 EZR による McNemar 検定の実行

(1) データの概要

65 歳以上の高齢者を対象に，転倒予防訓練と運動機能の低下の有無に関する研究が実施された．この研究では，200 人の被験者に対して，転倒予防訓練前に運動機能検査を行い，3 カ月の転倒予防訓練後に同様の検査を実施している．ここでの目標は，転倒予防訓練後に運動機能の低下が改善していることを確認することにある．のデータは，「fall_risk.csv」に保存されている．ここで変数「訓練前」は訓練前の運動機能の低下の有無(低下あり，低下なし)であり，「訓練後」は訓練後の運動機能の低下の有無(低下あり，低下なし)である．

(2) EZR による計算

ここでは，EZR を用いて McNemar 検定を実行する．McNemar 検定は，帰無仮説 H_0「歩行訓練を行っても運動機能の低下割合に変化がない」に対して対立仮説 H_1「歩行訓練を行うことで運動機能の低下割合に変化がある」を検定する．

McNemar 検定の実行
1: 「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)」を選択する． 2: 次のようなメニューが表示される．    このとき， ・「行の変数(1 つ選択)」で「訓練前」選択する． ・「列の選択(1 つ選択)」で「訓練後」を選択する． ・「連続性補正」で「Yes」を選択する． 3: 「OK」ボタンを押す

「連続性補正」とは，McNemar 検定もカイ 2 乗検定と同様に p 値の計算に近似を用いるため，それを補正するものである．ここでは，EZR での計算結果(青色の部分)のみを解釈する

Output.1

	訓練後	
訓練前	低下あり	低下なし
低下あり	40	60
低下なし	30	70

[28] 対応のあるクロス集計表を 2×2 の行列であると考えると，McNemar 検定とは，その行列の対称性を評価していると考えることができる．

Output.1 は，クロス集計表による要約の結果である．因みに，全体パーセントは，「統計解析」→「名義変数の解析」→「分割表の作成と群間の比率の比較(Fisher の正確検定)」において，「総計パーセント」を選べばよい(2.2.5 節を参照)．

この出力の下側の R の出力(「McNemar's Chi-squared test with continuity correction」と記載された部分)は，EZR による最後の出力と同じ内容なので割愛する．

McNemar 検定は，一番下の出力

Output.2	McNemar 検定 P 値 = 0.00224

である．p 値が 0.05 未満なので，有意な結果が得られた．したがって，歩行訓練を行うことで運動機能の低下割合に変化が認められた．

2.4.2 対応のある 2 値アウトカムの 3 群以上の比較

2.4.2.1 Cochran の Q 検定

表 2.8 は，癌性疼痛患者に対して，3 種類の除痛薬を投与したときの神経障害性疼痛の改善の有無を評価したクロスオーバー試験[29]の結果である．ここで，1 は改善を表しており，0 は非改善を表している．このように，3 種類以上の介入によるアウトカムでの事象の発現率を評価する場合には，McNemar 検定を用いることができず，Cochran の Q 検定を用いることになる[30]．

Cochran の Q 検定では，帰無仮説 H_0「介入(要因)によるアウトカムの事象の発現率はすべて同じである」に対して対立仮説 H_1「帰無仮説 H_0 ではない」を検定する[31]．表 2.8 の事例において，薬剤 A での(真の)改善率をp_{A}，薬剤 B での(真の)改善率をp_{B}，薬剤 C での(真の)改善率をp_{C}，とする．このとき，上記の仮説は，帰無仮説 H_0「$p_{\mathrm{A}} = p_{\mathrm{B}} = p_{\mathrm{C}}$」に対する検定を行うことを意味する．

表 2,8 の事例における値は 0.034 なので，有意であった．つまり，除痛薬によって神経障害性疼痛の改善率に違いが認められた．一方で，Cochran の Q 検定では，3 種類以上の介入(要因)によるアウトカムの事象の発現率の違いを評価できるものの，「どこに違いがあるか」を検討することはできない．そのため，薬剤 A vs. 薬剤 B(比較 AB)，薬剤 B vs. 薬剤 C(比較 AC)，薬剤 B vs. 薬剤 C(比較 BC)のすべての組み合わせ(ペアワイズ)での比較を McNemar 検定のもとで評価する必要がある．このとき，検定を 3 回繰り返すことから，多重比較が必要になる．

表 2.8：癌性疼痛患者に対する 3 種類の除痛薬のクロスオーバー試験の結果(0：非改善，1：改善)

患者	制吐剤		
	A	B	C
1	**0**	**1**	**0**
2	**1**	**1**	**0**
3	**1**	**1**	**1**
4	**0**	**0**	**0**
5	**1**	**0**	**0**
6	**0**	**1**	**1**
7	**0**	**0**	**0**
8	**1**	**1**	**0**
9	**0**	**1**	**0**
10	**1**	**1**	**1**
11	**0**	**1**	**0**
12	**1**	**1**	**0**
計	**6**	**9**	**3**
改善率(%)	**60.0%**	**75.0%**	**25.0%**

[29] チェンジオーバー・デザインと呼ばれることもある．

[30] Fless 愛好会 訳：計数データの統計学，株式会社アーム,2009 [原著：Fless J. L., Levin B., Paik MC.: Statistical Methods for Rate and Proportions (3rd edition), Wiley, 2003].

[31] Cochran の Q 検定は，2 値アウトカムに対する検定である．一方で，量的アウトカムに対する検定には，正規分布が仮定できる場合には繰り返し測定の分散分析(Repeated Measured ANOVA)，仮定できない場合には Freadman 検定がある．

表 2.9:癌性疼痛患者に対する 3 種類の除痛薬のクロスオーバー試験に対するペアワイズでの対応のあるクロス集計表

(a) 薬剤Aと薬剤Bの比較

		薬剤B		合計
		改善	非改善	
薬剤A	改善	5 [41.7%]	1 [8.3%]	6
	非改善	4 [33.3%]	2 [16.7%]	6
合計		9	3	12

(b) 薬剤Aと薬剤Cの比較

		薬剤C		合計
		改善	非改善	
薬剤A	改善	2 [16.7%]	4 [33.3%]	6
	非改善	1 [8.3%]	5 [41.7%]	6
合計		3	9	12

(c) 薬剤Bと薬剤Cの比較

		薬剤C		合計
		改善	非改善	
薬剤B	改善	3 [25.0%]	6 [50.0%]	9
	非改善	0 [0.0%]	3 [25.0%]	3
合計		3	9	174

表 2.9 は，3 剤(薬剤 A,薬剤 B,薬剤 C)のすべての組み合わせでの対応のあるクロス集計表である．2 値アウトカムにおいて一般的に用いられる多重比較には，Bonferroni 法や Holm 法のように，p 値を調整する方法である．McNemar 検定の p 値の Bonferroni の方法による調整 p 値は

・比較 AB:$0.180 \times 3 = 0.540$

・比較 AC:$0.180 \times 3 = 0.540$

・比較 BC:$0.014 \times 3 = 0.042$

である．したがって，薬剤 B と薬剤 C において有意であった．薬剤 B で改善したにも関わらず，薬剤 C で改善しなかった割合が 50.0%(6 例)であるのに対して，薬剤 C で改善したにも関わらず，薬剤 B で改善しなかった割合が 0.0%(0 例)であることから，薬剤 B による除痛効果は薬剤 C よりも優れているといえる(表 2.9(c))．

2.4.2.2 EZR による Cochran の Q 検定の実行

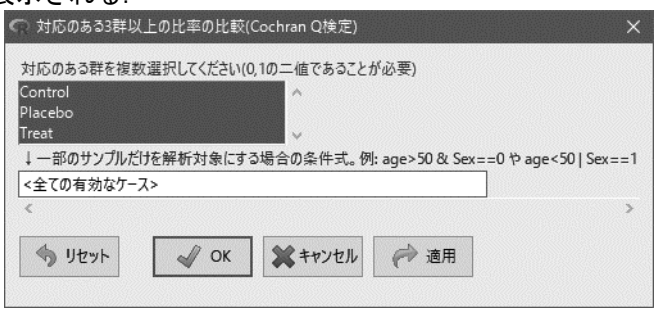
(1) データの概要

ここでは，ある疾患患者 17 名に 3 種類の薬剤(Treat:新薬，Control:既存薬，Placebo:プラセボ)のチェンジオーバー試験(すなわち，それぞれの被験者は，ウォッシュアウト期間を通じて，3 種類の薬剤の全てが投与・評価されている)を実施したときの結果(仮想例)である．このとき，アウトカムは，2 値 (0:無効，1:有効)がとられている．このデータは，「changeover.csv」で得られる．

(2) EZR による計算

EZR を用いて Cochran の Q 検定を実行する．Cochran の Q 検定は，帰無仮説 H_0「3 種類の薬剤間の有効割合に違いがない」に対して対立仮説 H_1「3 種類の薬剤間の有効割合に違いがある」を検定する．EZR における Cochran の Q 検定の注意点は，アウトカムが 0 あるいは 1 のダミー変数で与えられなければいけない点にある(その他の手法ではこのようなことはない)．そのため，カテゴリデータで与えられている場合には，「アクティブデータセット」→「変数の操作」→「ダミー変数を作成する」を用いて，2 値化しなければならない．

EZR による解析方法を以下に示す．

Cochran の Q 検定の実行
1: 「統計解析」→「名義変数の解析」→「対応のある 3 群以上の比率の比較(Cochran の Q 検定)」を選択する. 2: 次のようなメニューが表示される.    このとき, ・「対応のある群を複数選択してください(0,1 の二値であることが必要)」で「Control」,「Placebo」,「Treat」を選択する. 3: 「OK」ボタンを押す

ここでは, EZR での計算結果(下側の青色の部分)のみを解釈する. 上側の R の出力(「Cochran's Q test」の下に出力された部分) は, EZR での出力と同じのためである.

```
Cochran Q 検定 P 値 = 0.0013
```

その結果, p 値は 0.013 であることから, 有意である. したがって, 薬剤によって, 有効割合に違いが認められる.

どの薬剤間に違いが認められるかを検定する場合には, 2.4.1 節の McNemar 検定を実施し, p 値を 3 倍すればよい. そのときの結果のみ, 以下に示す.

	p 値	Bonferroni による多重比較[32]
Treat vs. Control	0.0455	0.1365
Treat vs. Placebo	0.0026	0.0078
Control vs. Placebo	0.2890	0.8670

すなわち, 新薬(Treat)とプラセボ(Placebo)のあいだに有意差が認められる.

2.5 ロジスティック回帰分析

2.5.1 ロジスティック回帰の概要

2.5.1.1 ロジスティック回帰分析の基礎

医学系研究では, 疾患の有無, 治療の成功／非成功など 2 値アウトカムで得られる状況は少なくない. 図 2.3 は, 胃がん患者に対して, ある術後補助化学療法を実施したときの被験者の年齢と奏効の有無をプロットした散布図である(仮想例). ここでは, 奏効例を 0, 非奏効例を 1 としており, 黒丸のデータ点は, 被験者(n=30)を表している. つまり, ここでの関心のあるイベントは, 非奏効例である[33].

このデータに対して, 2 値アウトカムを計量データと見做して単回帰直線をあてはめたものが青色の点線である. 腫瘍縮小率は, 0 から 1 までの範囲であるにも関わらず, マイナスの値や 1.0 を超える値をとる可能性がある. 例えば, 単回帰直線における 52 歳の被験者の非奏効率の予測値は, −0.25 となるが, そのような確率は存在しない.

ロジスティック回帰分析のモデルは,

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 \times (\text{年齢})$$

[32] Bonferroni による多重比較は, p 値×3 で計算できる.
[33] 奏効例を 1, 非奏効例を 0 とする場合が多いと思われるが, ロジスティック曲線の方向が逆になるので, 便宜上, このように定義している.

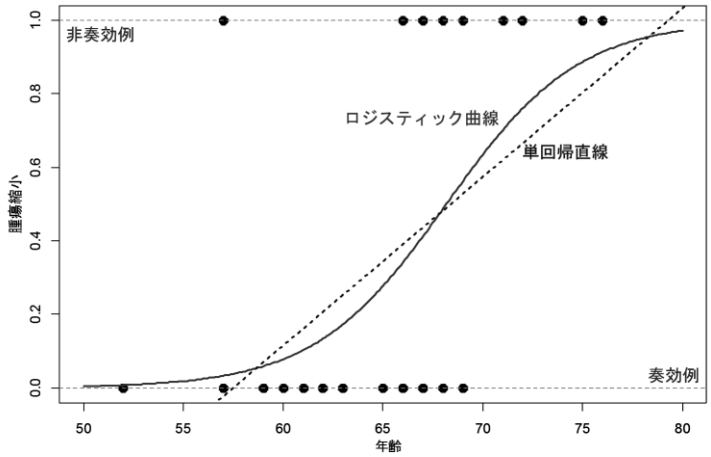


図 2.3: 胃癌患者に対する術後補助化学療法データに対するロジスティック曲線の図示

である. ここで, β_0 と β_1 は, それぞれ回帰係数であり, 左側(左辺)は確率 p に対するオッズ比の対数値(log は自然対数)であり, 対数オッズ比と呼ばれる. 通常の回帰分析では、応答変数(アウトカム)を予測するためのモデルであるのに対して, ロジスティック回帰分析では, 説明変数(年齢)に対する関心のあるイベントが発生する確率 p を予測する. 図 2.3 において, ロジスティック回帰分析の結果を表す赤色の実線は, 確率 p を表すことから, 0 から 1 までの範囲しかとらないことがわかる. 因みに, このような曲線のことをロジスティック曲線あるいはシグモイド曲線という.

図 2.3 におけるロジスティック回帰の結果は,

$$\log\frac{p}{1-p} = -20.618 + 0.302 \times (\text{年齢})$$

である. 確率 p の計算には, 上式を利用して,

$$p = \frac{\exp(-20.618 + 0.302 \times (\text{年齢}))}{1 + \exp(-20.618 + 0.302 \times (\text{年齢}))}$$

で計算できる. ここに exp は指数関数を表している[34]. 図 2.3 のシグモイド関数は, この式をプロットしたものである.

2.5.1.2 ロジスティック回帰分析とオッズ比の関係

表 2.10 は, 65 歳をカットオフ値にした場合のクロス集計表である. このクロス集計表より, 65 歳未満に対する 65 歳以上の奏効に対するオッズ比は,

$$\text{オッズ比} = \frac{11\times10}{11\times1} = 10.0$$

表 2.10: 胃癌患者に対する術後補助化学療法データに対して, 65 歳をカットオフ値(65 歳未満, 65 歳以上)としたときのクロス集計表

	非奏効(1)	奏効(0)	計
65 歳以上 (1)	11 (50.0%)	11 (50.0%)	22
65 歳未満 (0)	1 (9.1%)	10 (90.9%)	11
計	12 (36.4%)	21 (63.6%)	33

[34] Excel で計算する場合には, exp 関数を用いる.

表 2.11：開腹手術における縫合術に対する無作為化比較第 III 相試験

部位	縫合の方法	創合併症		計
		あり(1)	なし(0)	
上部(1)	真皮縫合術(1)	29 **(7.6%)**	353 **(92.4%)**	382
	ステープラー(0)	39 **(9.4%)**	374 **(90.6%)**	413
	計	68 **(8.6%)**	727 **(91.4%)**	795
下部(0)	真皮縫合術(1)	18 **(10.2%)**	158 **(89.8%)**	176
	ステープラー(0)	20 **(19.8%)**	81 **(80.2%)**	101
	計	38 **(13.7%)**	239 **(86.3%)**	277

表 2.12：開腹手術の縫合術に関するデータのロジスティック回帰分析の結果

	説明変数毎のロジスティック回帰分析 (simple logistic regression)			多重ロジスティック回帰分析 (multiple logistic regression)		
	回帰係数	オッズ比	p 値	回帰係数	オッズ比	p 値
縫合術	-0.343	0.709	0.094	-0.419	0.658	0.412
部位	-0.531	0.588	0.016	-0.600	0.549	0.008

である. したがって, 65 歳以上の被験者は, 65 歳未満に比べて非奏効となる割合が 10.0 倍であることがわかる.

次いで, 65 歳以上を 1, 65 歳未満を 0 としたときのロジスティック回帰分析は,

$$\log\left(\frac{p}{1-p}\right) = -2.303 + 2.303 \times (2\text{値の年齢})$$

である. このとき, β_1=2.393 の指数値 exp(2.393)=10.0 になる. すなわち, 説明変数に対する回帰係数 β_1 の指数値は, オッズ比に一致する.

2.5.1.3 ロジスティック回帰分析とオッズ比の関係

表 2.11 は, 消化器癌患者の開腹手術における縫合術に対する無作為化比較第 III 相試験の結果である(Tsujinaka et al., 2013). 説明変数は, 縫合術の種類のダミー変数(真皮縫合術：1, ステープラー：0), および, 手術部位(上部：1, 下部：0)であり, 応答変数は, 創合併症の有無(創合併症有：1, 創合併症無：0)である.

このように, 2 個以上の説明変数がある場合のロジスティック回帰分析を多重ロジスティック回帰分析(multiple logistic regression)という. 因みに, 多変量解析(multivariate analysis)あるいは多変量ロジスティック回帰分析(multivariate logistic regression)と記載された文献等を散見するが, 統計学での「多変量(multivariate)」とは, 応答が多変数で構成される場合を指す.

このときの多重ロジスティック回帰分析の回帰係数とオッズ比, 及び単一変量でのロジスティック回帰分析の回帰係数とオッズ比を表 2.12 に示す. 説明変数毎にロジスティック回帰分析を実施した場合と, 多重ロジスティック回帰分析を実施した場合で回帰係数及びオッズ比が異なることがわかる. これは, 多重ロジスティック回帰分析では, 説明変数間で調整が行われているためである. 例えば, 縫合術のオッズ比 0.658 とは, 部位による影響を調整したうえでオッズ

比を計算している．このようなオッズ比のことを調整オッズ比という．無作為解比較試験の群間比較において，割付調整因子を共変量とした調整オッズ比を用いるのは，(無作為割り付けで調整しきれなかった)割付調整因子の影響を調整したうえで，群間のオッズ比を評価するためである．

表 2.12 の p 値は，帰無仮説「回帰係数は 0 である」に対して，対立仮説「回帰係数は 0 でない」を検定したときの検定の結果である．手術部位(部位)はロジスティック回帰分析と多重ロジスティック回帰分析のいずれでも有意な結果が得られる．一方で，縫合術は，ロジスティック回帰分析では有意でないものの，多重ロジスティック回帰分析では有意な結果が得られた．本試験では，創合併症割合が低く，かつ縫合術間の差が小さい上部の割合が高いため(795/1072)，手術部位で調整しないロジスティック回帰では縫合術で有意な結果が得られなかったと推察される．

2.5.1.4 変数選択の方法

多重ロジスティック回帰を利用する場合，多くの論文で変数選択が実施される．変数選択を実施するとき，(1) 変数選択の評価基準，(2) 変数選択のアルゴリズム，を予め選ばなければならない．

変数選択の評価基準には，検定方法を用いる方法と情報量規準を用いる場合の 2 種類が存在する．検定方法を用いる場合とは，増加あるいは減少する変数に対して，回帰係数に対する検定あるいは適合度検定(モデルの適切性を表す検定)の p 値を用いて評価する方法である．一方で，情報量規準を用いる方法とは，赤池の情報量規準(AIC; Akaike’s Information Criteria)あるいは Bayes 流情報量規準(BIC; Bayesian Information Criateria)といったモデル適合度を表す統計量を用いる方法である．最近では，情報量規準を用いる方法が主流となっている．情報量規準の選択については，ゴールドスタンダードが存在するわけではないが，AIC を用いるよりも BIC を用いるほうが選択される変数の数が少なくなる傾向にある．

変数選択のアルゴリズムとして一般的に用いられる方法がステップワイズ法である．ステップワイズ法には，変数増加法(前進ステップワイズ法)，変数減少法(後退ステップワイズ法)，そして変数増減法がある．

(a) 変数増加法：切片のみのモデルから出発し，1 個ずつ説明変数をモデルに加える方法．

(b) 変数減少法：全ての説明変数を含むモデルから出発し，1個ずつ説明変数をモデルから除外する方法．

(c) 変数増減法：全ての説明変数を含むモデルから出発し，1個ずつ説明変数を加えるのか除外するのかを評価・実施する方法．

ステップワイズ法のアルゴリズムに対するゴールド・スタンダードは存在しない．変数選択に関する議論は， 1.7.2 節を参照されたい．

2.5.2 EZR によるロジスティック回帰の実行

(1) データの概要

ここでは，頭部外傷データを用いる．このデータは，カナダの 3121 名の軽度頭部外傷患者に対する CT による脳所見の有無に対して，10 個の共変量がとられている[35]．このデータは，headInjury.csv で与えられる．

各変数の名称と意味は，以下のとおりである．

- age.65：年齢(65 歳未満(0)／60 歳以上(1))
- amnesia.before：衝撃前の記憶喪失(30 分未満(0)／30 分以上(1))
- basal.skull.fracture：頭蓋底骨折(無(0)／有(1))
- GCS.decrease：グラスゴー・コーマ・スケール低下の有無(意識障害の評価)(低下なし(0)／低下(1))

[35] Stiell IG, et al.; The Canadian CT head rule for patients with miner head insury, The Lancet, 357, 1391-1396.

- GCS.13：初期のグラスゴー・コーマ・スケール(13 未満(0)，13 点以上(1))
- GCS.15.2hours：2 時間後のグラスゴー・コーマ・スケール(15 点未満(0)，15 点(1))
- high.risk：臨床医が脳神経学的介入のリスクが高いと判断したか否か(いいえ(0)，はい(1))
- loss.of.consciousness：気絶(無(0)，有(1))
- open.skull.fracture：蓋開放骨折(無(0)，有(1))
- vomiting：嘔吐(無(0)，有(1))
- clinically.important.brain.injury：CT による脳所見(無(0)，有(1))

ここでの目的は，脳所見の有無に影響を及ぼす共変量を探索することにある．

(2) EZR による実行

先ず，脳所見の有無による共変量の要約(背景表)を作成する．いずれの共変量も 2 値化されているので，すべて，カテゴリ変数として扱う．

背景表の作成

1: 「グラフと表」→「サンプルの背景データのサマリー表の出力」を選択する．
2: 次のようなメニューが表示される．

このとき，

- 「群別する変数(0〜1 つ選択)」で「clinically.important.brain.injury」を選択する．
- 「カテゴリ変数(名義変数，順序変数)」でその他の変数を選択する．

3: 「OK」ボタンを押す

なお，「自動選択」をクリップボードにすると，クリップボードに結果が保存され，WORD などに結果を貼り付けることができ，CSV ファイルを選択した場合には，結果をファイルに保存することができる．

また，カテゴリカル変数の場合には，カイ 2 乗検定と Fisher の正確検定を選択することができ，「連続変数(正規分布)」の場合には，2 標本 t 検定のp値，「連続変数(非正規分布)」の場合には，Wilcoxon 検定の p 値が選択される．

このときの結果を以下に示す(紙面の都合上，縦書きで描写している)．

Factor	Group	clinically.important.brain.injury		p.value
		0	1	
n		2871	250	
age.65 (%)	0	2590 (90.2)	169 (67.6)	<0.001
	1	281 (9.8)	81 (32.4)	
amnesia.before (%)	0	2318 (80.7)	164 (65.6)	<0.001
	1	553 (19.3)	86 (34.4)	
basal.skull.fracture (%)	0	2723 (94.8)	180 (72.0)	<0.001
	1	148 (5.2)	70 (28.0)	
GCS.13 (%)	0	2788 (97.1)	216 (86.4)	<0.001
	1	83 (2.9)	34 (13.6)	
GCS.15.2hours (%)	0	2595 (90.4)	131 (52.4)	<0.001
	1	276 (9.6)	119 (47.6)	
GCS.decrease (%)	0	2820 (98.2)	230 (92.0)	<0.001
	1	51 (1.8)	20 (8.0)	
high.risk (%)	0	2244 (78.2)	119 (47.6)	<0.001
	1	627 (21.8)	131 (52.4)	
loss.of.consciousness (%)	0	2581 (89.9)	191 (76.4)	<0.001
	1	290 (10.1)	59 (23.6)	
open.skull.fracture (%)	0	2777 (96.7)	229 (91.6)	<0.001
	1	94 (3.3)	21 (8.4)	
vomiting (%)	0	2631 (91.6)	182 (72.8)	<0.001
	1	240 (8.4)	68 (27.2)	

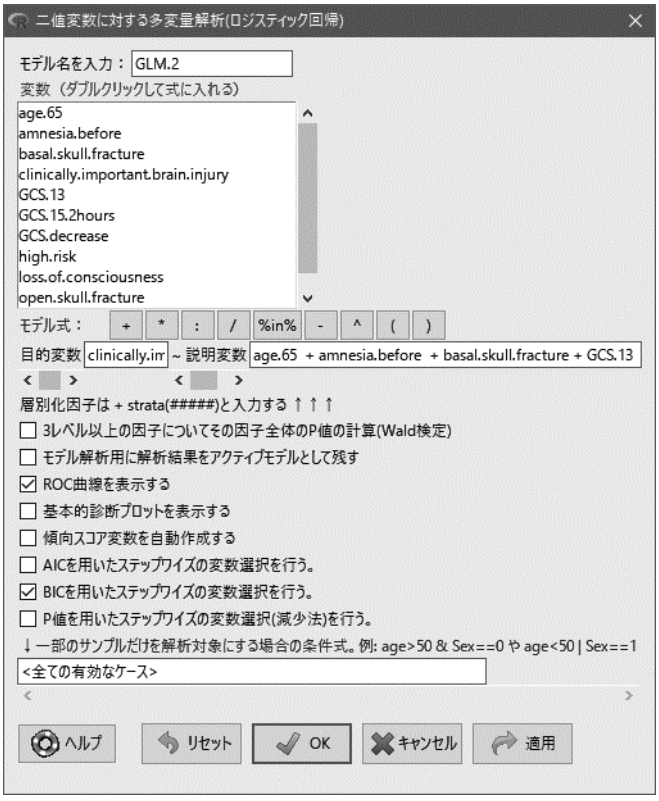
すべての共変量で有意差が認められた．一方で，GCS.decrease，GCS.13，及び，GCS.15.2hours は，いずれもグラスゴー･コーマ･スケールを扱っていることから，いずれかが不要であるかもしれない．

そのため，変数選択を伴うロジスティック回帰分析を用いて統計解析を行う．なお，EZR では，情報量規準(AIC,BIC)を用いる場合には，変数増減法による変数選択法が用いられ，検定を用いる方法(p 値を用いたステップワイズの変数選択)では，変数減少法が用いられる．ここでは，BIC による変数選択法を採用する．

ロジスティック回帰分析の実行

1: 「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)」を選択する．

2: 次のようなメニューが表示される．



このとき，

- 「モデル式：」において，
  目的変数 clinically.important.brain.injury ～説明変数 (共変量[36])
  と入力する．なお，CTRL キーを押しながら共変量を選択し，「＋」ボタンを押せば自動的に和として表示される．
- 「ROC 曲線を表示する」にチェックを入れる．
- 「BIC を用いたステップワイズ法の変数選択を行う」にチェックを入れる．

3: 「OK」ボタンを押す

ここで，ROC 曲線とは，受信者動作特性曲線(Receiver Operating Characteristic Curve)の略称であり，4.2 節の ROC 曲線と同じである．ロジスティック回帰分析における ROC 曲線は，推定されたロジスティック回帰モデルの予測値によって，2 値応答を適切に分けることができるか否かを評価しており，予測確度を確認するのに用いられる．

このときに，重要なのは ROC 曲線の曲線下面積 AUC(Area Under Curve)である．AUC とは，ROC 曲線の曲線下の面積であり，0.5～1.0 までの範囲をとる．曲線下面積は，1.0 に近づくほど予測確度が高いと解釈される．

その結果，多くの出力が表示される．ここでは，必要な結果のみ解釈する．

[36]「age.65 + amnesia.before + basal.skull.fracture + GCS.13 + GCS.15.2hours + GCS.decrease + high.risk + loss.of.consciousness + open.skull.fracture + vomiting」になる．

Output.1

```
Call:
glm(formula = clinically.important.brain.injury ~ age.65 + amnesia.before +
    basal.skull.fracture + GCS.13 + GCS.15.2hours + GCS.decrease +
    high.risk + loss.of.consciousness + open.skull.fracture +
    vomiting, family = binomial(logit), data = Dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2774  -0.3511  -0.2095  -0.1489   3.0028

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -4.4972     0.1629 -27.611  < 2e-16 ***
age.65                  1.3734     0.1827   7.518 5.56e-14 ***
amnesia.before          0.6893     0.1725   3.996 6.45e-05 ***
basal.skull.fracture    1.9620     0.2064   9.504  < 2e-16 ***
GCS.13                  1.0613     0.2820   3.764 0.000168 ***
GCS.15.2hours           1.9408     0.1663  11.669  < 2e-16 ***
GCS.decrease           -0.2688     0.3680  -0.730 0.465152
high.risk               1.1115     0.1591   6.984 2.86e-12 ***
loss.of.consciousness   0.9554     0.1959   4.877 1.08e-06 ***
open.skull.fracture     0.6304     0.3151   2.001 0.045424 *
vomiting                1.2334     0.1961   6.290 3.17e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1741.6  on 3120  degrees of freedom
Residual deviance: 1201.3  on 3110  degrees of freedom
AIC: 1223.3

Number of Fisher Scoring iterations: 6
```

Output.1 は，変数選択前のロジスティック回帰の結果である．GCS.decrease(グラスゴー・コーマ・スケール低下の有無)は，有意でなかった．また，ロジスティック回帰モデルの適合結果を表す AIC(赤池の情報量規準)は，1223.3 であった．

Output.2

```
Analysis of Deviance Table

Model 1: clinically.important.brain.injury ~ age.65 + amnesia.before +
    basal.skull.fracture + GCS.13 + GCS.15.2hours + GCS.decrease +
    high.risk + loss.of.consciousness + open.skull.fracture +
    vomiting
Model 2: clinically.important.brain.injury ~ 1
  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
1      3110     1201.3
2      3120     1741.6 -10   -540.3 < 2.2e-16 ***
```

Output.2 は，モデル適合度に対する尤度比検定の結果である．この検定は，null モデル(共変量がない場合のロジスティック回帰の結果)と適合度を比較することで，帰無仮説 H_0「回帰モデルに意味がない」に対して，対立仮説 H_1「回帰モデルに意味がある」を評価する．その結果，p 値が 0.001 未満(<2.2e-16(2.2×10^{-16}))と非常に小さいことから，回帰モデルに意味があることが伺える．

Output.3

age.65	amnesia.before	basal.skull.fracture	GCS.13	GCS.15.2hours
1.027664	1.014121	1.080201	1.015580	1.031707
GCS.decrease	high.risk	loss.of.consciousness	open.skull.fracture	vomiting
1.074796	1.021963	1.010002	1.010030	1.030185

Output.3 は，各共変量に対する VIF(Variance Inflation Factor，分散拡大係数(分散拡大要因))である．VIF が 10 を超える場合には多重共線性の程度が大きいと解釈される場合が多い．今回の事例では，そのような共変量は認められなかった．

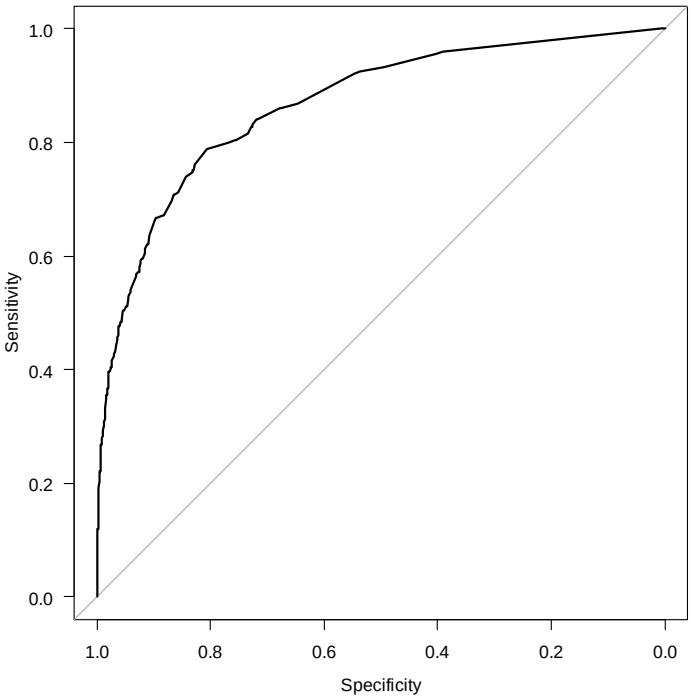


図 2.4:頭部外傷データに対する ROC 曲線

Output.4

	オッズ比	95%信頼区間下限	95%信頼区間上限	P 値
(Intercept)	0.0111	0.0081	0.0153	8.11e-168
age.65	3.9500	2.7600	5.6500	5.56e-14
amnesia.before	1.9900	1.4200	2.7900	6.45e-05
basal.skull.fracture	7.1100	4.7500	10.7000	2.02e-21
GCS.13	2.8900	1.6600	5.0200	1.68e-04
GCS.15.2hours	6.9600	5.0300	9.6500	1.83e-31
GCS.decrease	0.7640	0.3720	1.5700	4.65e-01
high.risk	3.0400	2.2200	4.1500	2.86e-12
loss.of.consciousness	2.6000	1.7700	3.8200	1.08e-06
open.skull.fracture	1.8800	1.0100	3.4800	4.54e-02
vomiting	3.4300	2.3400	5.0400	3.17e-10

Output..4 は，各共変量に対する調整オッズ比(回帰パラメータに指数をとったもの)及び，95%信頼区間である．basal.skull.fracture(頭蓋底骨折の有無)のオッズ比が最も高く， 次いで， GCS.15.2hours(2 時間後のグラスゴー・コーマ・スケール)が高かった．いずれも，有のほうが無に比べて，7 倍程度の脳所見の発現が認められた．

Output.5 曲線下面積 0.867 95%信頼区間 0.842 - 0.892

Output.5 は，図 2.4 の ROC 曲線における曲線下面積及び 95%信頼区間である．信頼区間の下限値が 0.5 を含まないことから，予測の点からも推定されたロジスティック回帰モデルが良好であることが示された．なお，この結果は，変数選択前のものであり，変数選択後の ROC 曲線は，以降の変数選択で選ばれた共変量を用いて，再度，ロジスティック回帰を実行しなければならない．

以降の部分，すなわち，以下の R コマンド(赤色の部分)

res <- stepwise(GLM.1, direction="backward/forward", criterion="BIC")[37]

は，変数選択の過程を表しているので，解釈は不要である．ここで，GLM.1 は，R でのオブジェクト，direction は，変数選択のアルゴリズム(EZR では変数増減法のみだが，R では変数増加法，変数減少法を選ぶことができるため)，criterion は，選択基準である(つまり，AIC で変数選択を行う場合には，criterion="AIC"になる)．

[37] このコマンドにおいて，GLM.1 は，R での GLM の保存したオブジェクトなので，名称が変わる可能性がある．

表 2.13:頭部外傷データに対する調整オッズ比

共変量	変数選択前		変数選択後	
	OR (95%C.I.)	p 値	OR (95%C.I.)	p 値
age.65(年齢)	3.95[2.76, 5.65]	<0.001	3.96[2.77, 5.65]	<0.001
amnesia.before(衝撃前の記憶喪失)	1.99[1.42, 2.79]	<0.001	2.01[1.43, 2.81]	<0.001
basal.skull.fracture(頭蓋底骨折)	7.11[4.75, 10.70]	<0.001	6.90[4.65, 10.20]	<0.001
GCS.13(初期の GCS)	2.89[1.66, 5.02]	<0.001	2.88[1.65, 5.03]	<0.001
GCS.15.2hours(2 時間後の GCS)	6.96[5.03, 9.65]	<0.001	6.94[5.02, 9.58]	<0.001
GCS.decrease(GCS 低下)	0.76[0.37, 1.57]	0.465	―	―
high.risk(脳神経学的介入リスク)	3.04[2.22, 4.15]	<0.001	3.04[2.23, 4.15]	<0.001
loss.of.consciousness(気絶)	2.60[1.77, 3.82]	<0.001	2.58[1.76, 3.78]	<0.001
open.skull.fracture(蓋開放骨折)	1.88[1.01, 3.48]	0.045	―	―
vomiting(嘔吐)	3.43[2.34, 5.04]	<0.001	3.48[2.37, 5.09]	<0.001

変数選択を実行した後の結果を以下に示す.

Output.6

```
Call:
glm(formula = clinically.important.brain.injury ~ age.65 + amnesia.before +
    basal.skull.fracture + GCS.13 + GCS.15.2hours + high.risk +
    loss.of.consciousness + vomiting, family = binomial(logit),
    data = TempDF)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3348  -0.3392  -0.2132  -0.1508   2.9940

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.4707     0.1616 -27.659  < 2e-16 ***
age.65                 1.3760     0.1813   7.590 3.20e-14 ***
amnesia.before         0.6976     0.1719   4.057 4.97e-05 ***
basal.skull.fracture   1.9318     0.2016   9.581  < 2e-16 ***
GCS.13                 1.0595     0.2838   3.734 0.000189 ***
GCS.15.2hours          1.9366     0.1650  11.735  < 2e-16 ***
high.risk              1.1123     0.1582   7.031 2.05e-12 ***
loss.of.consciousness  0.9466     0.1957   4.836 1.33e-06 ***
vomiting               1.2464     0.1947   6.400 1.55e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1741.6  on 3120  degrees of freedom
Residual deviance: 1205.5  on 3112  degrees of freedom
AIC: 1223.5

Number of Fisher Scoring iterations: 6
```

Output.6 は, 変数選択後のロジスティック回帰の結果である. GCS.decrease(GCS 低下)及び, open.skull.fracture(蓋開放骨折)が削除されている. そして, 全ての共変量の回帰パラメータに対する検定の p 値が 0.001 未満で高度に有意だった.

変数選択後の AIC は 1223.5 であった, 全変数の場合の AIC が 1223.3 なので僅かに上昇した(AIC は小さいほど良い). これは, 変数選択の基準(BIC)と評価基準(AIC)が異なるためである. 実際に, 全変数でのロジスティック回帰モデルの BIC が 1289.84 であるのに対して, 変数選択後は 1277.88 であった.

Output.7

	odds ratio	lower .95	upper .95	p.value
(Intercept)	0.0114	0.00833	0.0157	2.16e-168
age.65	3.9600	2.77000	5.6500	3.20e-14
amnesia.before	2.0100	1.43000	2.8100	4.97e-05
basal.skull.fracture	6.9000	4.65000	10.2000	9.57e-22
GCS.13	2.8800	1.65000	5.0300	1.89e-04
GCS.15.2hours	6.9400	5.02000	9.5800	8.46e-32
high.risk	3.0400	2.23000	4.1500	2.05e-12
loss.of.consciousness	2.5800	1.76000	3.7800	1.33e-06
vomiting	3.4800	2.37000	5.0900	1.55e-10

これは，変数選択後のロジスティック回帰モデルでの調整オッズ比である．変数選択前後での調整オッズ比を表 2.13 に示す．変数選択前後で，オッズ比に大きな違いは認められなかった．

2.6 共変量調整を伴うクロス集計表の解析：Mantel-Haentzel 検定

2.6.1 Mantel-Haentzel 検定

表 2.14 は，上部消化管または下部消化管の開腹外科手術が施行された患者 558 名(上部消化管 413 例，下部消化管 101 例)に対して，真皮縫合群とステープラー群の 2 群に割り付けて，手術後 30 日以内の創合併発現割合を比較したランダム化比較第 III 相試験の結果である(Tsujinaka et al.5))．表 2.14 (a) のクロス集計表は，縫合術(真皮縫合術，ステープラー)と創合併症の有無の関係を表している．ステープラーに対する真皮縫合術のオッズ比は 0.709 であるものの，カイ 2 乗検定の p 値は 0.116 であることから，縫合術による創合併症発現の有無に統計的な違いは認められなかった．表 2.14 (b)のクロス集計表は，部位(上部消化管，下部消化管)と創合併症の有無の関係を表している．下部消化管のほうが上部消化管に比べて創合併症発現割合が高く(オッズ比=1.701 倍)，カイ 2 乗検定の p 値は 0.018 であることから，統計的な違いが認められた．

本試験では，真皮縫合群のなかで上部消化管は 382 例，下部消化管は 176 例であるのに対して，ステープラー群では上部消化管は 413 例，下部消化管は 101 例であった．つまり，真皮縫合群には創合併症の発現が多い下部消化管の被験者の割合がステープラー群に比べて著しく多い(真皮縫合群：31.5%，ステープラー群：19.6%)．つまり，手術部位が縫合術の違いによる創合併症発現割合の比較に影響を及ぼしている可能性がある．

手術部位(上部・下部)で層化 したもとでクロス集計表を作成したものが表 2.14(c)である．このように，層化したもとで構成されるクロス集計表のことを多重クラス集計表といい，層化に用いた変数のことを共変量という．その結果，上部消化管のサブグループにおけるステープラーに対する真皮縫合術のオッズ比は 0.788 であり，カイ 2 乗検定の p 値は 0.380 であるのに対して，下部消化管のサブグループでのオッズ比は 0.461 であり，カイ 2 乗検定の p 値は 0.030 であった．いずれの部位でも真皮縫合群のほうがステープラー群に比べて創合併症の発現割合は低かったものの，有意だったのは下部消化器のサブグループのみであった．

創合併症発現割合に違いがある手術部位の偏りが，縫合術での比較に影響を与えている可能性は否定できない．共変量の影響を考慮したうえで群間比較を行う統計的方法が Mantel-Haenszel 検定(Cochran-Mantel-Haenszel 検定)である．Mantel-Haenszel 検定は，共変量の影響を考慮したもとで計算される調整オッズ比(Mantel-Haenszel 推定量)を用いて，帰無仮説「調整オッズ比は 1.0 である」に対して，対立仮説「調整オッズ比は 1.0 でない」を検定する．本事例における調整オッズ比は 0.658 (調整なしでのオッズ比=0.709)である．このときの Mantel-Haenszel 検定での p 値は 0.044 であることから，手術部位による調整オッズ比において，真皮縫合群のほうがステープラー群に比べて創合併症の発現割合を有意に減少させることが分かった．

ランダム化比較試験では，無作為割り付けを実施する際に，被験者の均一化を意図して割付調整因子を設定することが多い．ただし，割付調整因子が群間で完全に均一化されることは殆どない．Mantel-Haenszel 検定は，アウトカムを割付調整因子で調整(割り付け調整因子の均一性の補完)したもとで評価するのに用いられる．

表 2.14：開腹手術における縫合術に対する無作為化比較第 III 相試験の結果

(a)縫合術(真皮縫合術，ステープラー)と創合併症の有無のクロス集計表

	創合併症あり	創合併症なし	計
真皮縫合術	47 (8.4%)	511 (91.6%)	558
ステープラー	59 (11.5%)	455 (85.5%)	514
計	106 (9.9%)	966 (90.1%)	1,072

(b)手術部位(上部・下部)と創合併症の有無のクロス集計用

	創合併症あり	創合併症なし	計
上部	68 (8.6%)	727 (91.4%)	795
下部	38 (13.7%)	239 (86.3%)	277
計	106 (9.9%)	966 (90.1%)	1,072

(c) 手術部位を共変量としたときの縫合術(真皮縫合術，ステープラー)と創合併症の有無の多重クロス集計表

部位	縫合の方法	創合併症		計
		あり	なし	
上部	真皮縫合術	29 (7.6%)	353 (92.4%)	382
	ステープラー	39 (9.4%)	374 (90.6%)	413
	計	68 (8.6%)	727 (91.4%)	795
下部	真皮縫合術	18 (10.2%)	158 (89.8%)	176
	ステープラー	20 (19.8%)	81 (80.2%)	101
	計	38 (13.7%)	239 (86.3%)	277

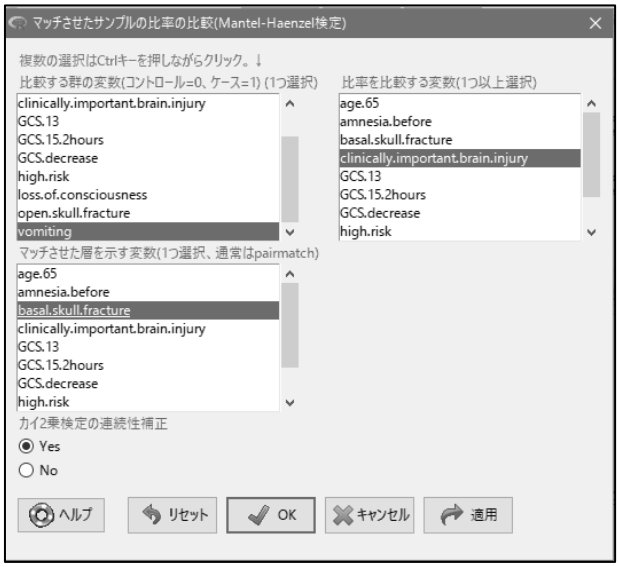
2.6.2 EZR による Mantel-Haentzel の実行

ここでは，2.5.2 節のデータを用いる．ここでは，嘔吐の有無(vomiting)が脳所見(clinically.important.brain.injury)と関連するか否かを評価する．このとき，頭蓋底骨折の有無(basal.skull.fracture)が影響を及ぼすことが想定されるため，調整を行う．

Mantel-Haentzel 検定

1: 「統計解析」→「マッチドペア解析」→「マッチさせたサンプルの比率の比較(Mantel-Haenzel 検定)」を選択する．

2: 次のようなメニューが表示される．



このとき，

- 「比較する群の変数(コントロール=0，ケース=1)(1 つ選択)」において「vomiting」を選択する．このとき，ケースは 1，コントロールは 0 にダミー変数化されていなければならない．ダミー変数化されていない場合には，「アクティブデータセット」→「ダミー変数を作成する」から，ダミー変数化することができる．
- 「比率を比較する変数(1 つ以上選択)」において，「clinically.important.brain.injury」を選択する．
- 「マッチさせた層を示す変数(1 つ選択選択、通常は pairmatch)」において，「basal.skull.fracture」を選択する．

3: 「OK」ボタンを押す

EZR の出力では，様々な出力が表示される．表示された青色の箇所毎に説明する．

Output.1

```
        Mantel-Haenszel chi-squared test with continuity correction
data:  Dataset$vomiting and Dataset$clinically.important.brain.injury and Dataset$basal.skull.fracture
Mantel-Haenszel X-squared = 77.034, df = 1, p-value < 2.2e-16
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 2.909690 5.590633
sample estimates:
common odds ratio
         4.033238
```

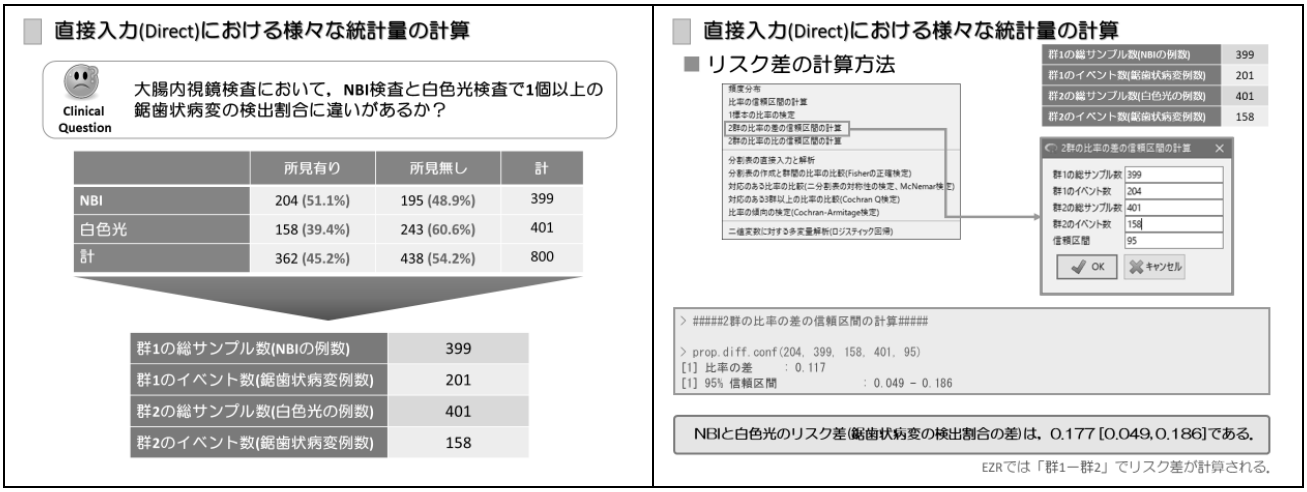
Output.1 は，Mantel＝Haentzel 検定に対する R の出力結果である．「p-value」が p 値を表している．$<2.2\times10^{-16}$(e-16 は，10^{-16}を表している)よりも小さいことから，高度に有意である．したがって，嘔吐の有無が脳所見に影響を与えることがわかる．また，「common odds ratio」とは，「頭蓋底骨折の有無(basal.skull.fracture)」で調整したオッズ比であり，「95 percent confidence interval」は，このときの 95%信頼区間である．つまり，嘔吐がある場合は，嘔吐がない場合に比べて，4.03 倍(95%信頼区間[2.91, 5.59])の脳所見が認められた．

Output.2

```
vomiting=0 vomiting=1 MH.p.value
clinically.important.brain.injury=0       2631         240    1.68e-18
clinically.important.brain.injury=1        182          68
```

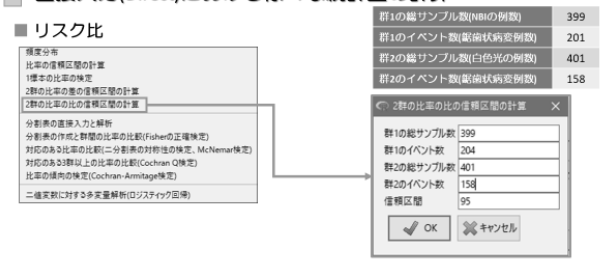
Output.2 は，EZR の結果である．クロス集計表及び p 値で表される．p 値は，Output.1 での「p-value」と同じである．

2.7 質的データの解析における補足的資料



直接入力(Direct)における様々な統計量の計算

■ リスク比




```
> #####2群の比率の比の信頼区間の計算#####

> prop.ratio.conf(204, 399, 158, 401, 95)
[1] 比率の比      : 1.298
[1] 95% 信頼区間          : 1.112 - 1.515
```

NBIと白色光のリスク比(NBIの検出割合／白色光の検出割合)は，1.298[1.112, 1.515]である．つまり，NBIは白色光に比べて1.298倍ほど鋸歯状病変の検出能がある．

EZRでは「群1/群2」でリスク比が計算される．

特殊な状況でのロジスティック回帰分析

■ 条件付ロジスティック回帰

EZR：「マッチドペア解析」→「マッチさせたサンプルの比率の多変量(条件付ロジスティック回帰)」

■ 皮膚科等での無作為化ハーフサイド試験
■ 1対1マッチングされたケースコントロール研究

では，共変量が同じで，アウトカムに対応がある．

このような場合に利用できるロジスティック回帰モデルが**条件付きロジスティック回帰分析**である．

3 章：生存時間データにおける統計解析

3.1 生存曲線に対する統計的推測

3.1.1 生存時間データの特徴

癌臨床試験における真のエンドポイントの一つは全生存期間(OS; Overall Survival)である．全生存期間とは，被験者が登録(手術・手技の場合は施行日の場合もある)された日を起算日として，死亡(イベント)までの期間を指す．一方で，追跡期間中に転院等で死亡日が不明になることは少なくない．このような被験者のデータでは，本来の全生存期間を得ることができないため，打ち切り(censoring)データと呼ばれる．

図 3.1 は，全生存期間を主要エンドポイントとしたときの臨床試験の例である．ここで，左図の直線は観測できた期間を表しており，点線は，観測できなかった期間を表している．転院が起こった場合には，その後の追跡が不能になることから，最終全生存確認日で打ち切られる．また，生存期間をエンドポイントとした臨床試験では，被験者を登録する期間(登録期間)及び，(最終症例登録日からの)追跡期間を設定するが[38]，追跡期間内に死亡が観測されなかった症例も打ち切りになる．

生存時間解析では，個々の被験者に対して，生存期間(survival time，あるいは time to event)と打ち切りの有無がペアでとられる．このときの統計的関心は，個々の被験者の生存期間(個別評価)ではなく，被験者全体から得られる生存曲線(集団評価)にある．

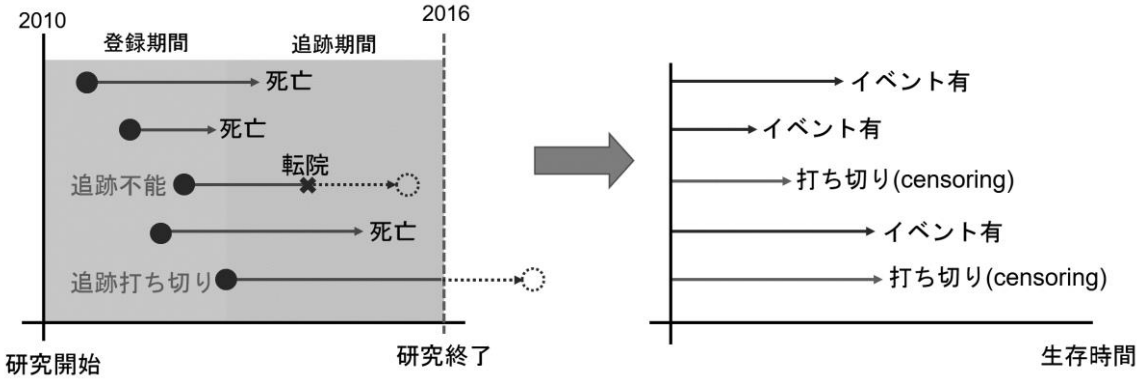


図 3.1：臨床支援と打ち切りのメカニズム

[38] 登録期間と追跡期間は，研究計画において検討しなければならない．生存期間を主要エンドポイントとする臨床試験では，統計学的な例数の設定は，必要イベント数(イベントが観測された被験者の人数)で与えられ，必要症例数では与えられない．そのため，登録期間＋追跡期間が短いと必要イベント数を観測するための症例数(必要症例数)が多くなり，一方で，長いと必要症例数は必要イベント数とほぼ同じになる．

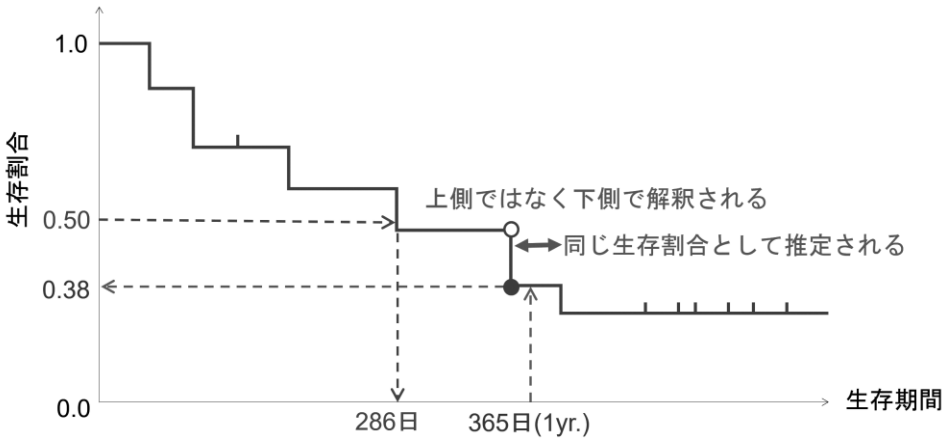


図 3.2：仮想例に対する Kaplan-Meier プロットとその解釈

3.1.2 生存曲線の推定：Kaplan-Meier 法

生存曲線とは，X 軸が生存期間，そして Y 軸が生存割合で描かれた曲線であり，例えば，「期間 T まで生存した割合は P である」ことが解釈できる．生存曲線の推定に一般的に用いられている方法が Kaplan-Meier 法である．

図 3.2 は，仮想データ(全生存期間)に対する Kaplan-Meier 法による生存曲線の推定結果(Kaplan-Meier 曲線)である．Kaplan-Meier 曲線は，生存期間が 0，そして生存割合が 1.0(生存期間が 0 の時点では被験者全員が生存していることを意味する)からの階段状プロット(stairs plot)によって表される．このとき，各階段は死亡(イベント)が観測された時点を表している(死亡が観測された時点で生存割合が減少することを意味する)．また，打ち切り時点には，目印(図 2 の場合には，「｜」を目印としている)が付与される．

Kaplan-Meier 曲線の解釈には，(1) X 軸(生存期間)から Y 軸(生存割合)を評価する場合，(2) Y 軸(生存割合)から X 軸(生存期間)を評価する場合，の 2 種類が存在する．

評価(1)の一般的な用途は，年次生存割合の推定である．図 3.2 では 1 年生存割合の推定の過程を表している．Kaplan-Meier 曲線では，階段の下側が生存割合を表しており，次の階段までの期間は同じ生存割合として解釈される．したがって， 1 年生存割合は，1 年(365 日[39])までの期間において，最後に死亡(イベント)が観測されたときの生存割合である．図 3.2 における，1 年生存割合は，0.38(38%)である．

評価(2)の一般的な用途は，中央生存期間(MST：Median Survival Time)である．中央生存期間とは，被験者の 50%が死亡するまでの期間(50%にイベントが発現するまでの期間)である．図 3.2 における中央生存期間は，286 日である．

3.1.3 EZR による生存曲線の推定

(1) データの概要

ここでは，North Central Cancer Treatment Group によって実施された進行肺癌患者に対するデータ[40]を用いる．このデータは，228 名の進行肺癌患者の全生存期間(日)がとられている．このデータは，Lung.csv で与えられる．変数は，time が生存期間，status(1：死亡，0：打ち切り)である．

[39] 年次生存率の推定では，うるう年を調整するために，1 年を 365.25 日とすることも多い．

[40] Loprinzi CL., et al.: Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. Journal of Clinical Oncology. 12(3):601-7, 1994.

(2) EZR による Kaplan–Meier 推定の方法

EZR を用いて Kaplan–Meier 推定を実行する. ここでは，日数で記載された生存期間を年に変換し，リスク集合のサイズ(任意の時点で死亡リスクに曝された被験者数)を X 軸の下に記載する.

生存曲線の Kaplan-Meier 推定の方法

1: 「統計解析」→「生存時間の解析」→「生存曲線の記述と群間の比較」を選択する.

2: 次のようなメニューが表示される.

このとき,

- 「観察期間の変数(1 つ選択)」で「time」を選択する.
- 「イベント(1)，打ち切り(0)の変数(1 つ選択)」で「status」を選択する.
- 「X 軸の単位」で「日を年に変換」を選択する.
- 「At risk のサンプル数を表示する」にチェックを入れる.

3: 「OK」ボタンを押す

ここで注意しなければいけないのは，イベント・打ち切りを表す変数のコードが決まっており，イベントは 1 で表し，打ち切りは 0 で表さなければならない.

このときの結果(青色の部分)の説明を以下に示す.

Output.1

```
Call: survfit(formula = Surv((time/365.25), status == 1) ~ 1, data = Dataset,
    na.action = na.omit, conf.type = "log-log")

  time n.risk n.event survival std.err lower 95% CI upper 95% CI
0.0137    228       1   0.9956 0.00438        0.9693        0.999
0.0301    227       3   0.9825 0.00869        0.9539        0.993
0.0329    224       1   0.9781 0.00970        0.9481        0.991
                          (省略)
0.9938     69       2   0.4154 0.03583        0.3448        0.484
0.9966     67       1   0.4092 0.03582        0.3387        0.478
1.0157     65       2   0.3966 0.03581        0.3264        0.466
                          (省略)
1.9357     15       1   0.1246 0.02904        0.0748        0.188
1.9932     14       1   0.1157 0.02830        0.0676        0.178
2.0014     13       1   0.1068 0.02749        0.0606        0.168
                          (省略) c
```

これは，生存表と呼ばれるものであり，年次生存割合の推定値を得るために用いる. ここで，「time」は生存期間，「n.risk」は time においてリスクに曝されている被験者数，「n.event」は time においてイベントがあった被験者数，

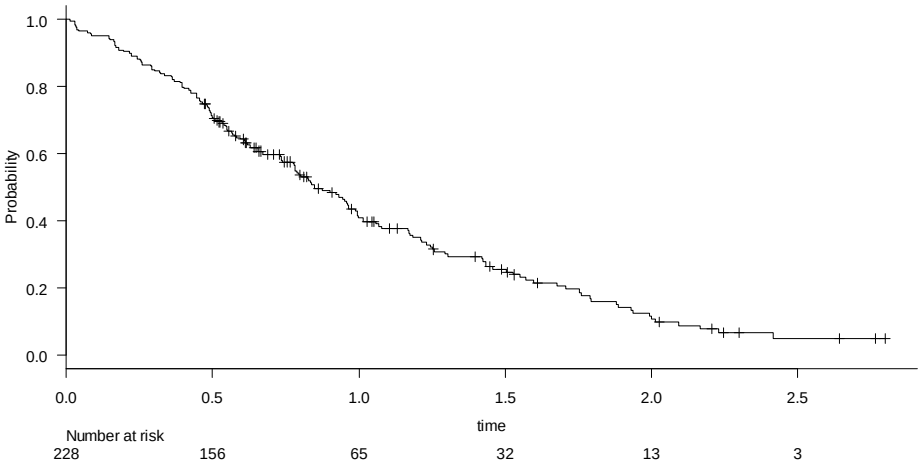


図 3.3：肺癌データに対する Kaplan-Meier プロット

「survival」,「std.err」,「lower 95% CI」,「upper 95% CI」は，それぞれ time における生存割合，標準誤差，95%信頼区間の下限値，上限値である．

例えば， time が 0.9966 での survival の 0.4092(40.92%)が 1 年生存割合であり，1.9932 での survival の 0.1157(11.57%)が 2 年生存割合である(太字の部分)．すなわち，年次生存割合は，当該生存期間以下の time のなかの最大値をとる(例えば，2 年生存率では 2.0014 のほうが 2 年に近いが 1.9932 の行の情報を用いる)．

Output.2	サンプル数	生存期間中央値	95%信頼区間
	1 228	0.848733744010951	0.777549623545517-0.988364134154689

すなわち，中央生存期間は 0.849(年)，95%信頼区間は[0.778, 0.988](年)であることがわかる．また，このときの Kaplan-Meier 曲線を図 3.3 に示す．今回は，95%信頼区間を描写していないが，表示したい場合には，メニューの「95%信頼区間を表示する」にチェックを入れればよい．

3.2 生存曲線の比較

3.2.1 生存曲線を比較するための基本的知識

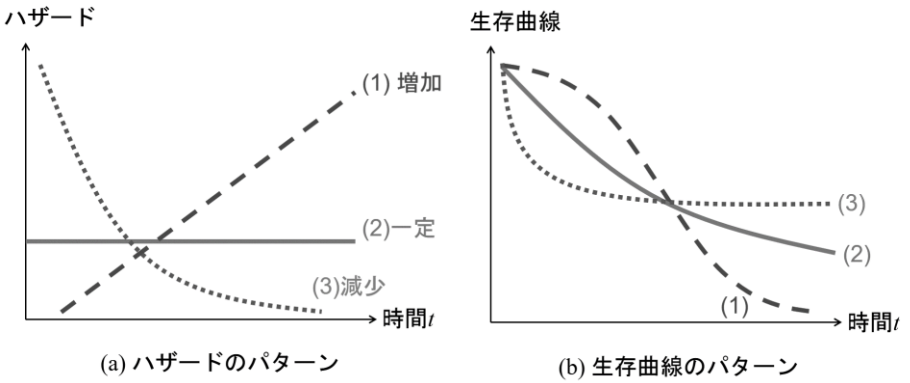
生存時間解析のなかで重要な要素の一つがハザード(瞬間死亡率)の考え方である．ハザードとは，時間 t まで生存している症例が，時間 t[41]において死亡(イベントが発生)する確率である．言いかえれば，ハザード比は時間 t における死亡リスクを表す．

図 3.4 は，生存期間(時間 t)に対するハザード及び生存曲線のパターンを表している．(1)は生存期間(時間 t)とともにハザード(死亡リスク)が増加している場合である，次いで，(2)は生存期間(時間 t)に対してハザード(死亡リスク)が一定の場合である．因みに，臨床試験において必要症例数を計算する場合には，この仮定のもとで計算する場合が多い．(3)は生存期間(時間 t)に対してハザード(死亡リスク)が減少する場合である．

2種類の治療法(新規治療，既存治療)が存在するとき，治療効果の違いをハザードの比で表したものがハザード比である．新規治療の既存治療に対するハザード比は，

$$\text{ハザード比} HR = \frac{\text{新規治療のハザード}}{\text{既存治療のハザード}}$$

[41] 厳密には，時間 t まで生存しているという条件で，微小期間 $t+\Delta t$ に死亡する確率を表す．



(a) ハザードのパターン　　(b) 生存曲線のパターン

図 3.4: ハザードと生存曲線のパターン

で与えられる. 上式のハザード比 HR は, 新規治療の死亡リスクは既存治療に対して, HR 倍であることを意味する. すなわち, ハザード比 HR が 1.0 を下回るとき, 新規治療が既存治療に比べて良好である(死亡リスクを軽減する)と判断できる.

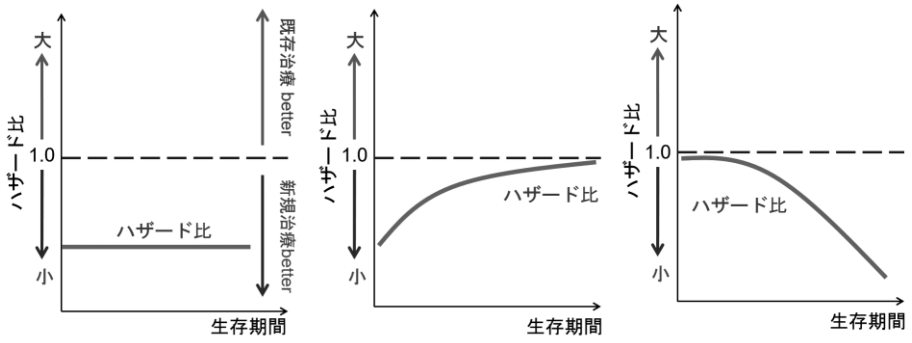
ハザードが時間 t に対して変化することから, ハザード比 HR も変化する. 図 3.5 はハザード比 HR のパターン例を表している. 図 3.5(a)は, 時間 t に対してハザード比が同じである. また, ハザード比 HR が 1.0 を下回ることから, 新規治療は既存治療に比べて, 時間 t に依らず有効性が高い(ハザード(死亡リスク)が低い). 図 3.5(a)のように, 時間 t に対して一定のハザード比を示すことを比例ハザード性という. 比例ハザード性は, 3.2.2 節で述べるログランク検定, 及び 3.2.3 節で述べる比例ハザード・モデルにおいて仮定される. また, 多くの論文・学会発表において,「ハザード比が〇〇である」と記載されているが, このような解釈も比例ハザード性が仮定されている.

図 3.5(b)は, ハザード比 HR が時間 t とともに上昇している. これは, 観察期間前期では, 新規治療のハザード(死亡リスク)が既存治療に比べて低いものの, 観察期間後期になるにつれて同程度になることを意味する. 図 3.5(c)は, ハザード比 HR が時間 t とともに減少している. これは, 観察期間前期に死亡(イベント)があった症例では, 新規治療と既存治療のハザード(死亡リスク)が同程度であったものの, 観察期間後期になるにつれて, 新規治療のほうが既存治療に比べてハザードが低くなることを意味する.

3.2.2 生存曲線の比較

3.2.2.1 ログランク検定

生存曲線を比較するための方法として広範に利用されている統計的検定の方法は, ログランク検定である. ログランク検定では, 帰無仮説 H_0「ハザード比は 1.0 である」に対して, 対立仮説 H_1「ハザード比は 1.0 でない」を検定する.



(a) ハザード比が一定　　(b) ハザード比が上昇　　(c) ハザード比が減少

図 3.5: ハザード比のパターン

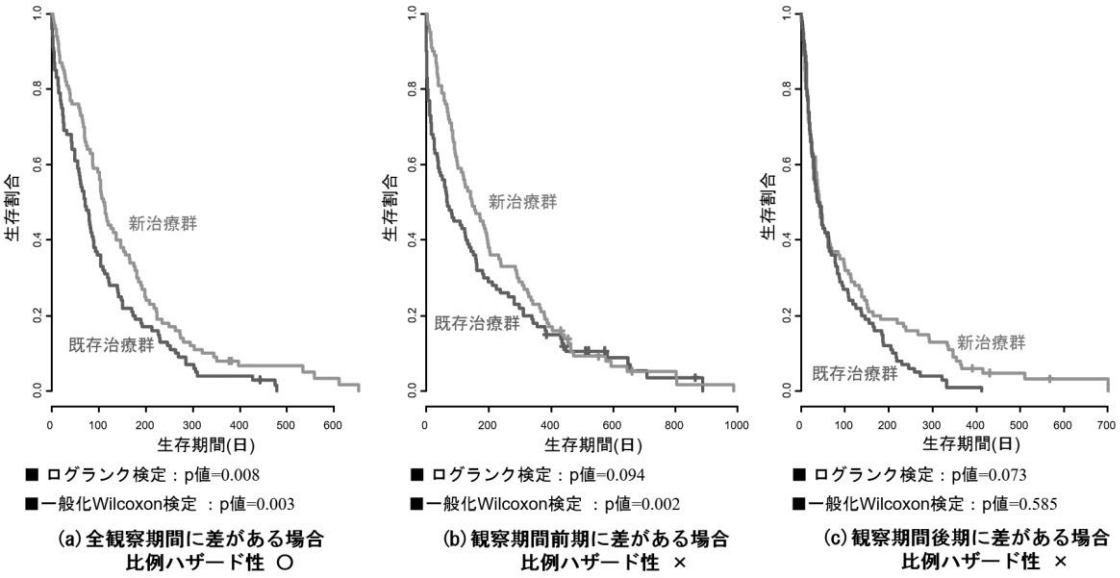


図 3.6：シミュレート・データに対する 3 種類の検定の結果(いずれの標本サイズも 100 である)

ハザード比は時間 t に対して変化するにも関わらず[42]，「ハザード比が〇〇だからポジティブ・スタディだった(あるいはネガティブ・スタディだった)」という解釈を行うことがしばしばある．これは，多くの医学系研究において，比例ハザード性（ハザード比 HR が時間 t に対して一定である)が暗黙裡に仮定されるためである．ログランク検定においても比例ハザード性が仮定されるため，上記のような仮説になる．そのため，比例ハザード性の仮定を満たさない，あるいは，ハザードが交差する状況において有意になりにくい傾向にある．

図 3.6 は 3 種類のシミュレート・データに対する検定結果を表している．ログランク検定は，比例ハザード性を満たす状況では有意であるものの(図 3.6(a))，比例ハザード性を満たさない状況(図 3.6(b)(c))では有意ではなかった．

3.2.2.2 一般化 Wilcoxon 検定

抗癌剤の 1 次治療の臨床試験などでは，全生存期間による評価の問題が指摘されることがしばしばある．なぜなら，このような臨床試験では，観察期間前期には全生存期間に差があっても，後続の治療法の影響によって，全生存期間の差が次第に小さくなるためである．とくに，後続治療が重複する可能性が高い投与レジメンの違い(例えば，4 週投与 2 週休薬 vs. 2 週投与 1 週休薬のレジメンの比較)を比較する臨床試験，あるいは後続治療において治療法がクロスオーバーする臨床試験では，その傾向が顕著である．

観察期間前期には差が認められても，次第に差がなくなる(ハザード比が 1.0 に近づく)ような場合，比例ハザード性の仮定は満たされず，図 3.5(b)のような形状を示す．このような状況に対する対処法としては，(1) 主要エンドポイントをサロゲート・エンドポイント(例えば，無増悪生存期間)に変更する，(2) 比例ハザード性を仮定するログランク検定以外の検定方法を採用する，ことが考えられる．

対処(2)の候補となる一つの検定が，一般化 Wilcoxon 検定である．一般化 Wilcoxon 検定の特徴は，観察期間前期の生存期間の差に敏感(有意になりやすい)なものの，観察期間後期には鈍感(有意になりにくい)ことにある．図 3.6(b)における，ログランク検定の p 値は 0.094 で有意でないものの，一般化 Wilcoxon 検定では有意差が認められた

[42] ハザード HR は時間 t の関数である．

(p=0.002). 図 3.6(c)は観察期間後期に差があるものの，観察期間前期に差が認められない場合である. この場合の一般化 Wilcoxon 検定の p 値は 0.585 であり，他の 2 手法に比べて極端に高かった.

3.2.3 EZR による生存曲線の比較

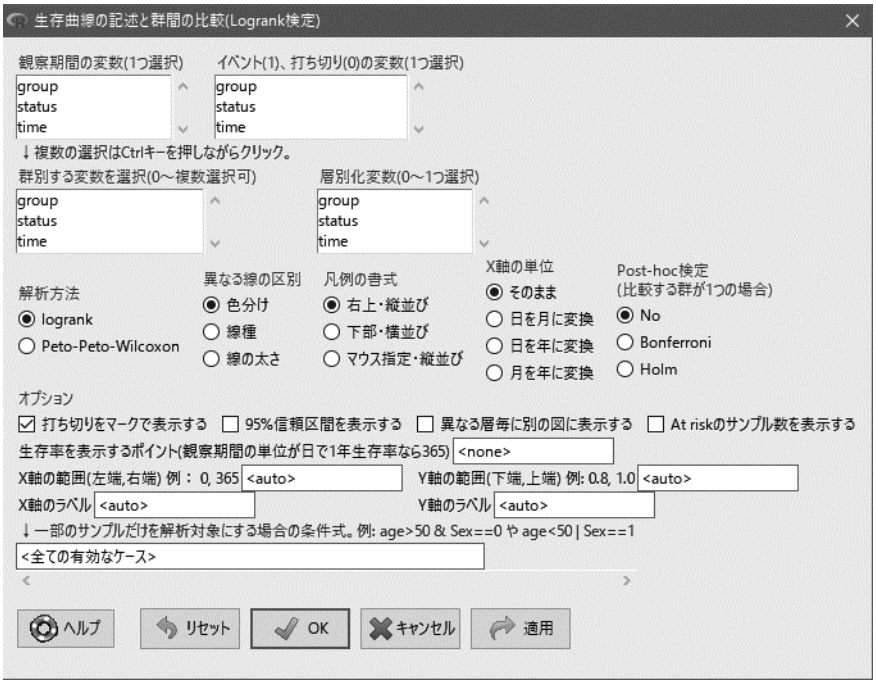
(1) データの概要

ここでは，卵巣癌データ[43]を用いて生存曲線を比較する. このデータは，26 名の卵巣癌患者に対する 2 種類の抗癌剤(既存薬，新薬)における全生存期間(日)がとられている. このデータは，Ovarian.csv で与えられる. 変数は，time が生存期間，status(1：死亡，0：打ち切り)，及び gorup(0：既存薬，1：新薬)である.

(2) EZR による生存曲線の比較

EZR を用いて治療群(group)による生存曲線を比較する. ここでは，日数で記載された生存期間を年に変換し，リスク集合のサイズ(任意の時点で死亡リスクに曝された被験者数)を X 軸の下に記載する. また，生存曲線の比較には，ログランク検定を用いる.

Logrank 検定による生存曲線の比較
1:　「統計解析」→「生存時間の解析」→「生存曲線の記述と群間の比較」を選択する. 2:　次のようなメニューが表示される.



このとき，

- 「観察期間の変数(1 つ選択)」で「time」を選択する.
- 「イベント(1)，打ち切り(0)の変数(1 つ選択)」で「status」を選択する.
- 「群別する変数を選択(0～複数選択可)」で「group」を選択する.
- 「X 軸の単位」で「日を年に変換」を選択する.
- 「At risk のサンプル数を表示する」にチェックを入れる.

3:　「OK」ボタンを押す

[43] Schumacher M, et al. G. : Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. Journal of Clinical Oncology, 12, 2086-2093, 1994.

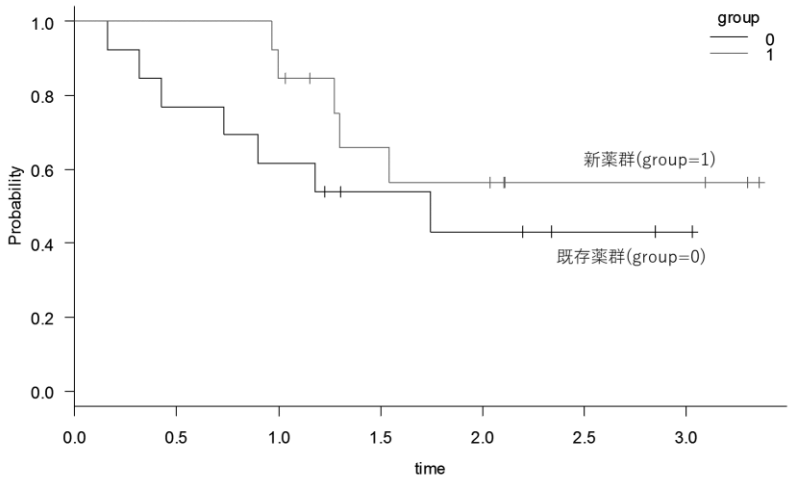


図 3.7:卵巣癌データに対する Kaplan-Meier プロット(日本語の説明は出力に追記している)

ここで注意しなければいけないのは，イベント・打ち切りを表す変数のコードが決まっており，イベントは 1 で表し，打ち切りは 0 で表さなければならない．また，群数についても，0～1,2,3,…のようなダミー変数で与える．さらに，一般化 Wilocoxon 検定は，解析方法の「Peto-Peto-Wilcoxon」を選択すればよい．

生命表は，群毎に次のように与えられる．

Output.1

```
Call: survfit(formula = Surv((time/365.25), status == 1) ~ group, data = Dataset,
    na.action = na.omit, conf.type = "log-log")

                group=0
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
0.162     13       1    0.923  0.0739        0.566        0.989
0.315     12       1    0.846  0.1001        0.512        0.959
0.427     11       1    0.769  0.1169        0.442        0.919
0.734     10       1    0.692  0.1280        0.373        0.872
0.901      9       1    0.615  0.1349        0.308        0.818
1.180      8       1    0.538  0.1383        0.248        0.760
1.747      5       1    0.431  0.1467        0.156        0.683

                group=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
0.966     13       1    0.923  0.0739        0.566        0.989
0.999     12       1    0.846  0.1001        0.512        0.959
1.270      9       1    0.752  0.1256        0.407        0.914
1.300      8       1    0.658  0.1407        0.320        0.858
1.541      7       1    0.564  0.1488        0.244        0.793
```

したがって，既存薬群(group=0)の 1 年生存割合は 61.5%[95%信頼区間：30.8%-81.8%]であり，新薬群の 1 年生存割合は，84.6%[95%信頼区間：51.2%-95.9%]だった．したがって，新薬の 1 生存割合のほうが既存薬群に比べて，20%以上高かった．

このときの Kpalan-Merier プロットを図 3.7 に示す．新薬群(group=1)の生存曲線が，既存薬群(group=0)の上側に布置した．したがって，新薬での有効性が示唆される．

Output.1 の下側の出力，すなわち，次の R コマンド

(res <- survdiff(Surv(time,status==1)~group, data=Dataset, rho=0, na.action = na.omit))

の下側は，ログランク検定の結果を表しているが，下側の EZR の出力と同じ内容であることから，割愛する．

Output.2

```
          サンプル数 生存期間中央値          95%信頼区間    P 値
group=0         13       1.746749 0.427104722792608-NA 0.303
group=1         13             NA   1.2703627652293-NA
```

Output.2 は，各群の標本サイズ(サンプル数)，中央生存期間，95%信頼区間及び，ログランク検定の p 値である．ここで，新薬群(group=1)の中央生存期間が NA(欠測)になっているのは，生存曲線が中央生存期間まで下がっていないためである．95%信頼区間の上限値が NA(欠測)になっているのも同様である．

また，ログランク検定では，帰無仮説 H_0「ハザード比は 1.0 である」に対して，対立仮説 H_1「ハザード比は 1.0 でない」を評価するが，その p 値が 0.303 であることから，生存曲線に対する有意な違いは認められなかった．

3.3 比例ハザードモデル

3.3.1 比例ハザードモデルの基本

生存時間データに対する回帰分析の方法として一般的に用いられているのが，比例ハザードモデルである(Cox の比例ハザードモデル)．比例ハザードモデルでは，任意の共変量の値 $\boldsymbol{x}$ に対する時間 t におけるハザードλを推定することができる．

ここでは，3.2.3 節で用いた，ECOG(Eastern Cooperative Oncology Group)が実施した卵巣癌に対する無作為化比較試験のデータを用いて比例ハザードモデルについて説明する(Edmunson et al., 1979)[44]．

いま，既存治療群を 0，新治療群を 1 で表した共変量(ダミー変数と呼ばれる)を「治療」とするとき，卵巣癌に対する無作為化比較試験のデータに対する比例ハザードモデル(ハザード λ を推定するための回帰モデル)は，

$$\lambda = \lambda_0(t) \cdot \exp\{\beta \times (\text{治療})\}$$

で表すことができる．ここで，$\lambda_0(t)$は，共変量(治療)に依らないハザードであり，ベースライン・ハザード(基線ハザード)と呼ばれる．比例ハザードモデルの特徴は，ベースライン・ハザード$\lambda_0(t)$には共変量が入っておらず，また，共変量による影響を表す$\exp\{\beta \times (\text{治療})\}$には時間 t が入っていないことにある．つまり，共変量(治療)による影響は，時間 t に依らず一定(比例ハザード性)が仮定される．

したがって，既存治療群(治療群=0)におけるハザード$\lambda_{\text{既存}}$は，

$$\lambda_{\text{既存}} = \lambda_0(t) \cdot \exp\{\beta \times 0\} = \lambda_0(t)$$

であり，新規治療群(治療群=1)におけるハザード$\lambda_{\text{新規}}$は，

$$\lambda_{\text{新規}} = \lambda_0(t) \cdot \exp\{\beta \times 1\} = \lambda_0(t)\exp\{\beta\}$$

である．これらを用いて既存治療群に対する新規治療群のハザード比 HR で表すと

$$HR = \frac{\lambda_{\text{新規}}}{\lambda_{\text{既存}}} = \frac{\lambda_0(t) \cdot \exp\{\beta\}}{\lambda_0(t)} = \exp\{\beta\} \qquad (1)$$

となる[45]．つまり，ハザード比は回帰係数の指数値$\exp\{\beta\}$である．

卵巣癌に対する無作為化比較試験のデータにおける回帰係数βの推定値$\hat{\beta}$は，$\hat{\beta} = -0.595$であることから，ハザード比 HR は，

$$HR = \exp\{\hat{\beta}\} = \exp\{-0.595\} = 0.551$$

である．ハザード比 HR が 1.0 を下回ることから，新規治療のほうが，既存治療に比べて死亡リスクが減少することがわかる．

---

[44] Edmunson, J.H. et al. : Different Chemotherapeutic Sensitivities and Host Factors Affecting Prognosis in Advanced Ovarian Carcinoma vs. Minimal Residual Disease. Cancer Treatment Reports, 63:241-47, 1979.

[45] 前回述べたように，比例ハザード性の仮定のもとでは，ハザード比は時間 t に依らず一定である．Cox 比例ハザードモデルにおけるハザード比 HR は，ベースライン・ハザード$\lambda_0(t)$が削除されることから，比例ハザード性の仮定のもとで構成されることがわかる．

3.3.2 比例ハザードモデルと調整ハザード比

卵巣癌に対する無作為化比較試験のデータでは，残像病変を有する被験者の割合が，対照群 61.5%(8/15)に対して処理群 53.8%(7/13)であり，若干の相違が認められている．また，残像病変の有無が被験者の予後に影響を与える可能性がある．そのため，残像病変の有無の影響を排除(調整)したもとでハザード比を評価することを考える．いま，既存治療群を 0，新規治療群を 1 で表した共変量を「治療」，残存病変無を 0，残存病変有を 1 で表した共変量を「残存病変」するとき，比例ハザードモデルは，

$$\lambda = \lambda_0(t) \cdot \exp\{\beta_1 \times (\text{治療}) + \beta_2 \times (\text{残存病変})\}$$

で与えられる．このとき，残存病変の有無が同じであるときの既存治療群(治療=0)に対する新規治療群(治療=1)のハザード比 HR は，

$$HR = \frac{\lambda_0(t) \cdot \exp\{\beta_1\} \cdot \exp\{\beta_2 \times (\text{残存病変})\}}{\lambda_0(t) \cdot \exp\{\beta_2 \times (\text{残存病変})\}} = \exp\{\beta_1\}$$

である．すなわち，「残存病変」を共変量に加えた場合においても，「治療」に対するハザード比は，回帰係数β_1の指数値$\exp\{\beta_1\}$によって計算できる．このときのハザード比は，調整ハザード比と呼ばれる．

卵巣癌に対する無作為化比較試験のデータでは，「治療」に対する回帰係数$\hat{\beta}_1 = -0.763$であり，「残存病変」に対する回帰係数$\hat{\beta}_2 = 1.320$であった．したがって，「治療」に対する調整ハザード比は，

$$HR = \exp\{\hat{\beta}_1\} = \exp\{-0.763\} = 0.466$$

である．残存病変の有無による影響を調整しない場合のハザード比が 0.551 であったことから，調整ハザード比のほうが僅かに小さくなることがわかった．

3.3.3 比例ハザードモデルにおける変数選択

比例ハザードモデルにおいても，これまでに説明した重回帰分析，多重ロジスティック回帰分析と同様に変数選択を実施することが多い．変数選択の方法についても，これまでと同様であり，(1) 変数選択のアルゴリズム，(2) 変数選択の評価基準，を予め選ばなければならないが，いずれもこれまでと同様である．

3.3.4 EZR による比例ハザードモデルの実行

(1) データの概要

ここでは，乳癌データを用いる．このデータは，ホルモン療法の効果を検討するために，ドイツ乳癌研究グループ(GBSG; German Breast Cancer Study Group)が実施した無作為化比較第 III 相試験の結果である．このデータは，GBSG2.csv で与えられる．変数は，生存時間(time)，イベントの有無(1:イベント(死亡)，0:打ち切り)とともに，以下の 8 個の予後因子がとられている．

- ・年齢(age)　・閉経の有無(menostat)　・腫瘍径(size)　・腫瘍のグレイド(grade)
- ・リンパ節転移個数(pnodes)　・ホルモン療法の有無(horth)
- ・プロゲステロン・レセプタ個数(progrec)　・エストロゲン・レセプタ個数(estrec)

ここで，年齢，腫瘍径，リンパ節転移個数，プロゲステロン・レセプタ個数，エストロゲン・レセプタ個数は連続変数であり，閉経の有無(Post, Pre)，ホルモン療法の有無(Yes, No)は 2 値変数，腫瘍のグレイドは順序変数である．

(2) EZR による実行

ここでは，4 個の連続データ(年齢(age)，腫瘍径(size)，リンパ節転移個数(pnodes)，プロゲステロン・レセプタ個数(progrec)，エストロゲン・レセプタ個数(estrec))を中央値で 2 値化したもとで評価を行う．

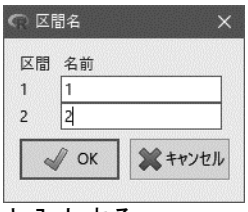
連続データの 2 値化 (age を 2 値化して 2 値変数 age.bin を作成する)
1: 「アクティブデータセット」→「変数の操作」→「数値変数を区分に分ける」を選択する. 次のようなメニューが表示される.

このとき,

- 「区分に分ける変数(0～1 つ選択)」で「age」を選択する.
- 「新しい変数」に「age.bin」と入力する.
- 「区間の数」を「2」に設定する.
- 「区分の方法」で「同データ数の区分」を設定する.

これらの作業後に「OK」ボタンを押す.

3: 次のようなメニューが表示される.



ここで, 区間 1 に「0」, 区間 2 に「1」と入力する.

4: 「OK」ボタンを押す

これにより, 同データ数(中央値)で 2 群に分けたデータ集合が作成される. この作業を腫瘍径(size), リンパ節転移個数(pnodes), プロゲステロン・レセプタ個数(progrec), エストロゲン・レセプタ個数(estrec)に実行し, size.bin, pnodes.bin, prog.bin, est.bin を作成する.

ここでは, 変数選択を伴う比例ハザードモデルを用いる. このとき, 連続変数の共変量(年齢, 腫瘍径, リンパ節転移個数, プロゲステロン・レセプタ個数, エストロゲン・レセプタ個数)には, 前述の 2 値化したものを用いる.

また, 腫瘍のグレイド(I, II, III)は,「グレイド II か否か」,「グレイド III か否か」の 2 個のダミー変数で表現される. そのため, グレイド自体の評価には, 共変量全体(ここでは, 腫瘍グレイドに対する)での検定が必要になる. EZR では, Wald 検定を用いて検定することができる.

EZR における比例ハザードモデルの変数選択は, ロジスティック回帰モデルと同様である. すなわち, 情報量規準(AIC,BIC)を用いる場合には, 変数増減法による変数選択法が用いられ, 検定を用いる方法(p 値を用いたステップワイズの変数選択)では, 変数減少法が用いられる. ここでは, BIC による変数選択法を採用する.

比例ハザードモデルの実行
1: 「統計解析」→「生存時間の分析」→「生存時間に対する多変量解析(Cox 比例ハザード回帰)」を選択する.
2: 次のようなメニューが表示される.

このとき，

・「モデル式：」において，

時間 | time | ，イベント | cens | ～説明変数 | (共変量)

と入力する．ここで，(共変量)には，

age.bin + est.bin + horTh + menostat + pnodes.bin + prog.bin + tgrade + tsize.bin

と入力する．なお，共変量をダブルクリックすれば，「＋」が自動的に付与される．

・「3 レベル以上の因子についてその因子全体の P 値の計算(Wald 検定)」にチェックを入れる．

・「BIC を用いたステップワイズ法の変数選択を行う」にチェックを入れる．

3: 「OK」ボタンを押す

その結果，多くの出力が表示される．ここでは，必要な結果のみ解釈する．

Output.1

```
Call:
coxph(formula = Surv(time, cens == 1) ~ age.bin + est.bin + horTh +
    menostat + pnodes.bin + prog.bin + tgrade + tsize.bin, data = Dataset,
    method = "breslow")

  n= 686, number of events= 299

                    coef exp(coef) se(coef)      z Pr(>|z|)
age.bin[T.1]     0.29968   1.34942  0.19407  1.544 0.122553
est.bin[T.1]     0.06331   1.06535  0.13334  0.475 0.634938
horTh[T.yes]    -0.42676   0.65262  0.12856 -3.320 0.000901 ***
menostat[T.Pre]  0.16086   1.17451  0.19997  0.804 0.421174
pnodes.bin[T.1]  0.92849   2.53069  0.12240  7.586 3.30e-14 ***
prog.bin[T.1]   -0.69046   0.50135  0.13934 -4.955 7.22e-07 ***
tgrade           0.16906   1.18419  0.11085  1.525 0.127248
tsize.bin[T.1]   0.09406   1.09862  0.12029  0.782 0.434271
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
age.bin[T.1]       1.3494     0.7411    0.9225    1.9740
est.bin[T.1]       1.0654     0.9387    0.8203    1.3835
horTh[T.yes]       0.6526     1.5323    0.5073    0.8396
menostat[T.Pre]    1.1745     0.8514    0.7937    1.7381
pnodes.bin[T.1]    2.5307     0.3951    1.9909    3.2168
prog.bin[T.1]      0.5013     1.9946    0.3815    0.6588
tgrade             1.1842     0.8445    0.9529    1.4716
tsize.bin[T.1]     1.0986     0.9102    0.8679    1.3907

Concordance= 0.699  (se = 0.018 )
Rsquare= 0.164   (max possible= 0.995 )
Likelihood ratio test= 123  on 8 df,   p=0
Wald test            = 120.4  on 8 df,   p=0
Score (logrank) test = 127.4  on 8 df,   p=0
```

Output.1 は，変数選択前の比例ハザードモデルの結果である．ホルモン療法の有無(horth)，リンパ節転移個数のダミー変数(pnodes.bin)，プロゲステロン・レセプタ個数のダミー変数(prog.bin)において，有意だった．このとき，変数名[〇〇.1]あるいは horTH[T.yes となっているのは，カテゴリカル変数において，カテゴリ 1，あるいはカテゴリ yes のときに 1，それ以外の場合に 0 のダミー変数によって推定された回帰パラメータであることを意味する．

exp(coef)は，ダミー変数において 1/0 のハザード比を表している．一方で，exp(-coef)は，ダミー変数において 0/1 のハザード比である．なお，95%信頼区間[lower .95, upper.95]は，1/0 のハザード比に対するものなので，0/1 の場合には，その逆数を計算すればよい．その結果，pnodes.bin(リンパ節転移個数のダミー変数)の影響が高く，転移個数が多い場合(1)のほうが，少ない場合(0)に比べて，死亡リスクを 2.53 倍に上昇させることがわかった．また，horTh(ホルモン療法の有無)は，ホルモン療法を実施したほうが(yes)，しない場合(no)に比べて死亡リスクを 0.65 倍に減少させるようである．

モデルの予測確度の指標一つである C 指標(Concordance index)は，0.699 であった．C 指標は，0～1 までの範囲をとり，寄与率と同様の解釈を行うことができる．その下側に，Rsquare(寄与率)が存在するが，比例ハザードモデルで用いることは少ないので，割愛する．

適合度検定を表す，尤度比検定(Likelihood ratio test)，Wald 検定(Wald test)，スコア検定(Sqore (logrank) test)は，いずれも有意だった．

Output.2

	ハザード比	95%信頼区間下限	95%信頼区間上限	P 値
age.bin[T.1]	1.3490	0.9225	1.9740	1.226e-01
est.bin[T.1]	1.0650	0.8203	1.3840	6.349e-01
horTh[T.yes]	0.6526	0.5073	0.8396	9.014e-04
menostat[T.Pre]	1.1750	0.7937	1.7380	4.212e-01
pnodes.bin[T.1]	2.5310	1.9910	3.2170	3.297e-14
prog.bin[T.1]	0.5013	0.3815	0.6588	7.221e-07
tgrade	1.1840	0.9529	1.4720	1.272e-01
tsize.bin[T.1]	1.0990	0.8679	1.3910	4.343e-01

Output.2 は，ハザード比に対する R のアウトプットを EZR のなかで日本語に翻訳したものなので割愛する．

以降の部分，すなわち，以下の R コマンド(赤色の部分)

res <- stepwise(CoxModel.1, direction="backward/forward", criterion="BIC")[46]

は，変数選択の過程を表しているので，解釈は不要である．ここで，CoxModel.1 は，R でのオブジェクト，direction は，変数選択のアルゴリズム(EZR では変数増減法のみだが，R では変数増加法，変数減少法を選ぶことができるため)，criterion は，選択基準である(つまり，AIC で変数選択を行う場合には，criterion="AIC"になる)．

変数選択を実行した後の結果を以下に示す．

Output.3

```
Call:
coxph(formula = Surv(time, cens == 1) ~ horTh + pnodes.bin +
    prog.bin, data = TempDF, method = "breslow")

  n= 686, number of events= 299

                   coef exp(coef) se(coef)      z Pr(>|z|)
horTh[T.yes]    -0.4132    0.6615   0.1252 -3.299 0.000969 ***
pnodes.bin[T.1]  0.9512    2.5888   0.1193  7.975 1.55e-15 ***
prog.bin[T.1]   -0.7348    0.4796   0.1193 -6.159 7.32e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
horTh[T.yes]       0.6615     1.5116    0.5176    0.8456
pnodes.bin[T.1]    2.5888     0.3863    2.0491    3.2706
prog.bin[T.1]      0.4796     2.0850    0.3796    0.6060

Concordance= 0.693  (se = 0.018 )
Rsquare= 0.156   (max possible= 0.995 )
Likelihood ratio test= 116.5  on 3 df,   p=0
Wald test            = 113.3  on 3 df,   p=0
Score (logrank) test = 120.1  on 3 df,   p=0
```

[46] このコマンドにおいて，CoxModel.1 は，R での GLM の保存したオブジェクトなので，名称が変わる可能性がある．

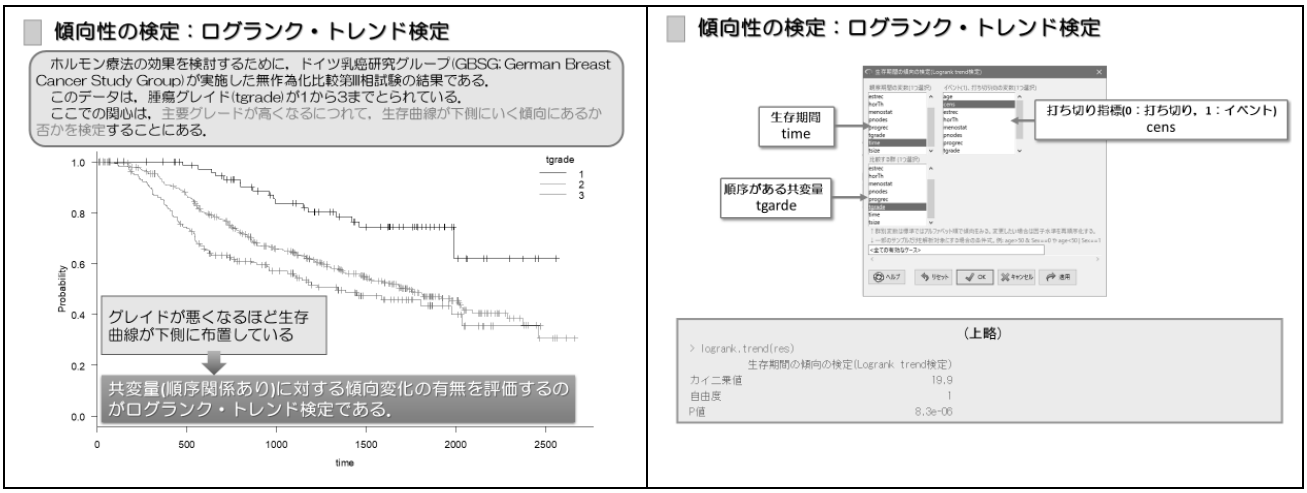
表 3.1：変数選択前後の調整ハザード比

	変数選択前		変数選択後	
	HR [95%C.I.]	p 値	HR [95%C.I.]	p 値
年齢(age.bin)	1.349[0.923, 1.974]	0.123	―	―
エストロゲン・レセプタ個数(est.bin)	1.065[0.820, 1.384]	0.635	―	―
ホルモン療法の有無(horTh)	0.653[0.507, 0.840]	0.001	0.662[0.518, 0.846]	0.001
閉経の有無(menostat)	1.175[0.794, 1.738]	0.421	―	―
リンパ節転移個数(pnodes.bin)	2.531[1.991, 3.217]	<0.001	2.589[2.049, 3.271]	<0.001
プロゲステロン・レセプタ個数(prog.bin)	0.501[0.382, 0.659]	<0.001	0.480[0.380, 0.606]	<0.001
腫瘍のグレイド(tgrade)	1.184[0.953, 1.472]	0.127	―	―
腫瘍径(tsize.bin)	1.099[0.868, 1.391]	0.434	―	―

Output.3 は，変数選択後の比例ハザードモデルの結果である．ホルモン療法の有無(horth)，リンパ節転移個数のダミー変数(pnodes.bin)，プロゲステロン・レセプタ個数のダミー変数(prog.bin)のみがモデルに含まれた．変数選択後の C 指標は 0.693 であった，全変数の場合の C 指標が 0.699 なので僅かに減少したものの，変数を大幅に減少することができた．

変数選択前後での調整ハザード比を表 3.1 に示す．変数選択前後で，調整ハザード比比に大きな違いは認められなかった．

2.8 生存時間データの解析における補足的資料



4 章：臨床検査データにおける統計解析

4.1 定性検査値の評価

4.1.1 定性検査値の要約

(1) データの概要：マンモグラフィ検査のデータ

ここでは，乳癌に対するマンモグラフィ検査の予測確度を評価する仮想例を用いる(新谷，2015[47])．このデータは，病理診断の結果，乳癌ありと診断された 12 名と乳癌なしと診断された 9,988 名に対するマンモグラフィ検査の結果(陽性，陰性)を用いて，マンモグラフィ検査の診断能を評価している．このときのクロス集計表を以下に示す．

	乳癌あり	乳癌なし	合計
検査陽性	10	799	809
検査陰性	2	9,189	9,191
合計	12	9,988	10,000

(2) 定性検査値を要約するための統計的方法

ここでは，定性検査を解析するのに用いる用語について整理する．下表は，定性検査における呼び方を表している．

	疾患有	疾患無
検査陽性	真陽性 (TP: True Positive)	偽陽性 (FP: False Positive)
検査陰性	偽陰性 (FN: False Negative)	真陰性 (TN: True Negative)

以下では，定性検査のための評価指標について略説する．

感度・特異度

疾患有の被験者を陽性と正しく診断する確率(疾患患者を陽性と判断する確率)を感度(sensitivity)といい，また，疾患無の被験者を陰性と正しく診断する確率(非疾患患者を陰性と判断する確率)を特異度(specificity)という．感度と特異度は，次のように定義される．

$$\text{感度} = \frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}} + n_{\mathrm{FN}}} \quad , \quad \text{特異度} = \frac{n_{\mathrm{TN}}}{n_{\mathrm{TN}} + n_{\mathrm{FP}}}$$

（n_{TP}：真陽性の被験者数，n_{FP}：偽陽性の被験者数，n_{TN}：真陰性の被験者数，n_{FN}：偽陰性の被験者数）

[47] 新谷歩：今日から使える医療統計，医学書院，2015.

陽性的中率・陰性的中率

感度と特異度の利点は，当該疾患の有病率(prevalence)に影響されずに診断性能を評価できる点にある．一方で，「感度≠臨床的有用性」でることに注意しなければならない．感度は「疾患患者を陽性と判断する確率」であり，臨床検査の診断能を評価しているのに対して(医師・研究者の立場)，実際の臨床検査では「陽性と判断された被験者が実際に疾患である確率」という臨床的有効性が重要である(患者の立場)．

このようなときに用いられるのが陽性的中率 (positive predictive value)及び陰性的中率(negative predictive value)である．陽性的中率とは，陽性と診断された被験者が疾患である確率を意味する．また，陰性的中率とは，陰性と診断された被験者が疾患である確率を意味する．陽性的中率と陰性的中率は，次のように定義される．

$$\text{陽性的中率} = \frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}} + n_{\mathrm{FP}}} \quad , \quad \text{陰性的中率} = \frac{n_{\mathrm{TN}}}{n_{\mathrm{TN}} + n_{\mathrm{FN}}}$$

一方で，上式による陽性的中率及び陰性的中率の計算は，真の有症率が(疾患者数)÷(総被験者数)とした場合であり，実際の有症率が異なる場合には，陽性的中率及び陰性的中率の値が変化する．すなわち，陽性的中率及び陰性的中率は，有病率の影響を受ける．

因みに，陽性的中率は陽性予測度，陰性的中率は陰性予測度と呼ばれることもある．

陽性尤度比・陰性尤度比

疾患有が疾患無よりも何倍陽性になりやすいかを表す指標に陽性尤度比(likelihood ratio of a positive result)がある．

$$\text{陽性尤度比} = \frac{\text{感度}}{1 - \text{特異度}} = \frac{n_{\mathrm{TP}} / (n_{\mathrm{TP}} + n_{\mathrm{FN}})}{1 - n_{\mathrm{TN}} / (n_{\mathrm{TN}} + n_{\mathrm{FP}})}$$

で定義される．すなわち，陽性尤度比は，疾患有を陽性と診断した場合と疾患無を陽性と診断した場合の比で表されており，大きいほど確定診断に優れるといえる(一般に尤度比といった場合には陽性尤度比を表す）．

また，疾患有が疾患無よりも何倍陰性になりやすいかを表す指標が陰性尤度比(likelihood ratio of a negative result)であり，

$$\text{陰性尤度比} = \frac{1 - \text{感度}}{\text{特異度}} = \frac{1 - n_{\mathrm{TP}} / (n_{\mathrm{TP}} + n_{\mathrm{FN}})}{n_{\mathrm{TN}} / (n_{\mathrm{TN}} + n_{\mathrm{FP}})}$$

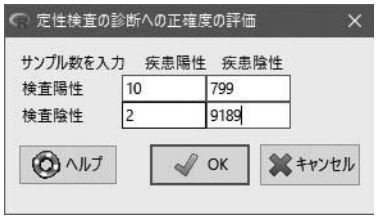
で定義される．すなわち，陰性尤度比は，疾患有を陰性と診断した場合と疾患無を陰性と診断した場合の比で表される指標である．

(3) EZR による定性検査値の評価

ここでは，マンモグラフィ検査のデータを用いて EZR による解析方法を解説する．

定性検査の診断への正確度の評価

ここでは，感度，特異度，陽性的中率，陰性的中率，陽性尤度比，陰性尤度比などの指標を計算する方法について述べる．EZR での計算では，マンモグラフィ検査のデータのように，クロス集計表を予め用意したうえで，それを直接入力することで実行できる．

定性検査の診断への正確度の評価
1:　「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」を選択する.
2:　クロス集計表のデータを次のように入力する.

3:　「OK」ボタンを押す

このときのアウトプットは，以下のとおりである.

	疾患陽性	疾患陰性	計
検査陽性	10	799	809
検査陰性	2	9189	9191
計	12	9988	10000

点推定と 95 % 信頼区間

	推定値	信頼区間下限	信頼区間上限
検査の陽性率	0.081	0.076	0.086
真の有病率	0.001	0.001	0.002
感度	0.833	0.516	0.979
特異度	0.920	0.915	0.925
陽性的中率	0.012	0.006	0.023
陰性的中率	1.000	0.999	1.000
診断精度	0.920	0.914	0.925
陽性尤度比	10.417	8.019	13.532
陰性尤度比	0.181	0.051	0.642

ここで，検査の陽性度とは，陽性と診断された被験者の割合である. また，真の有病率とは，このデータから計算された有病率であり，

$$\text{真の有病率} = \frac{10+2}{10,000} = 0.0012 \approx 0.001$$

である.　さらに，診断精度とは，正しく診断された被験者の割合であり，

$$\text{診断精度} = \frac{10+9189}{10,000} = 0.9199 \approx 0.920$$

で計算される. また，信頼区間下限，信頼区間上限は，各指標に対する 95%信頼区間を表している.

事例の結果より，

・感度(乳癌患者をマンモグラフィ検査で陽性とする確率)は，83.3% (0.830)，

・特異度(非乳癌患者をマンモグラフィ検査で陰性とする確率)は，92.0% (0.920)，

・陽性的中率(陽性の被験者が乳癌である確率)は，1.2% (0.012)，

・陰性的中率(陰性の被験者が乳癌でない確率)は，100.0% (1.000)，

・乳癌患者は非乳癌患者よりも 10.417 倍陽性になる(陽性尤度比)，

・乳癌患者は非乳癌患者よりも 0.181 倍陰性になる(陰性尤度比)

と解釈される.

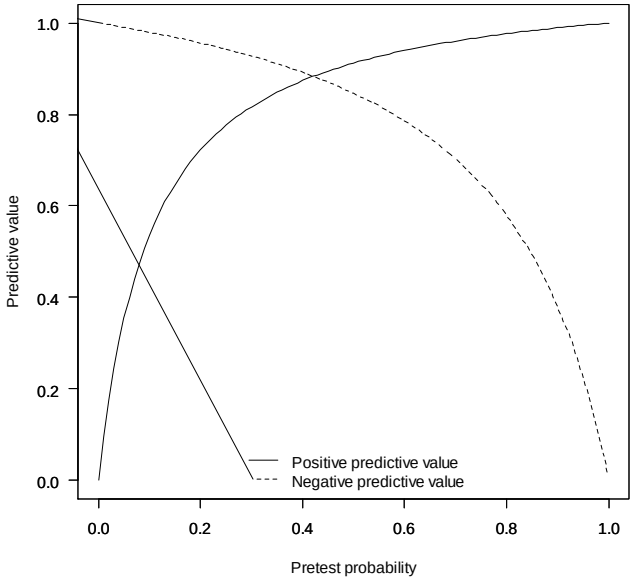
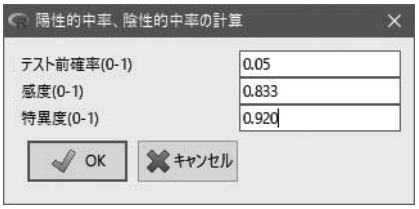


図 4.1：陽性的中率および陰性的中率の推移

陽性的中率，陰性的中率の計算

「定性検査の診断への正確度の評価」での陽性的中率及び陰性的中率は，真の有症率(疾患患者数÷被験者数)に基づいて計算されたものである．一方で，疫学研究等により，当該疾患の有症率が分かる場合には，それを用いて陽性的中率，陰性的中率を計算したほうが適切である．

ここでは，疫学調査で報告された乳癌の有症率を 5%(0.05)としたときの陽性的中率，陰性的中率を計算する．

陽性的中率・陰性的中率の計算(有病率が存在する場合)
1:　「統計解析」→「検査の正確度の評価」→「陽性的中度、陰性的中度の計算」を選択する． 2:　データを次のように入力する．  3:　「OK」ボタンを押す

このときのアウトプットは，以下のとおりである．

```
                    仮定
テスト前確率(0-1)    0.05
感度                0.833
特異度              0.92

                  計算結果
陽性的中率          0.354
陰性的中率          0.991
```

先ほどの事例(有症率=0.001)に比べて，陽性的中率が上昇する一方で，陰性的中率が減少していることがわかる．すなわち，有症率が陽性的中率及び陰性的中率に影響を及ぼすことがわかる．このことは，同時に表示されるグラフ(図 4.1)からも明らかである．ここで，実線は陽性的中率であり，破線は陰性的中率である．横軸は有病率を表している．したがって，有病率が上昇するほど陽性的中率が高くなるのに対して，陰性的中率は低くなる．

4.1.2 二つの定性検査の一致性の評価：Kappa 係数

(1) データの概要：2 人の病理医による非小細胞肺癌の診断データ

ここでは，2 人の病理医による非小細胞肺癌の組織標本 75 枚の組織学的分類結果のデータを用いる(Gardis, 2009[48]).

		病理医 A		
		Grade II	Grade III	合計
病理医 B	Grade II	41	3	3
	Grade III	4	27	27
	合計	45	0	75

(2) Kappa 係数の説明

ここでは，二つの定性検査の一致性を表す指標として Kappa 係数について説明する．病理医 A と病理医Bの診断結果が一致した割合(測定者の全一致率)は，

$$\text{測定者の全一致率} = \frac{41+27}{75} = 0.907$$

によって計算できる．しかしながら，測定者の全一致率は，病理医Aと病理医Bの結果が偶然にも一致する割合を考慮していない．病理医Aと病理医Bは組織標本を見ずに適当に診断したとしても Grade がある程度一致する．このような偶然による一致率を調整したうえで測定者間の一致率を計算したものが Kappa 係数である

偶然による一致率を組織学的分類のデータを用いて説明する．先ず，Grade2 に対する偶然の一致数，Grade3 に対する偶然の一致数は，

$$\text{Grade2に対する偶然の一致数} = \frac{(\text{病理医AがGrade2と診断した例数}) \times (\text{病理医BがGrade2と診断した例数})}{(\text{全組織標本数})} = \frac{45 \times 44}{75} = 26.4$$

$$\text{Grade3に対する偶然の一致数} = \frac{(\text{病理医AがGrade3と診断した例数}) \times (\text{病理医BがGrade3と診断した例数})}{(\text{全組織標本数})} = \frac{30 \times 31}{75} = 12.4$$

なので，偶然による一致率は，

$$\text{偶然による一致率} = \frac{(\text{Grade2に対する偶然の一致数}) + (\text{Grade2に対する偶然の一致数})}{(\text{全組織標本数})} = \frac{26.4 \times 12.4}{75} = 0.517$$

である．このとき，Kappa 係数は

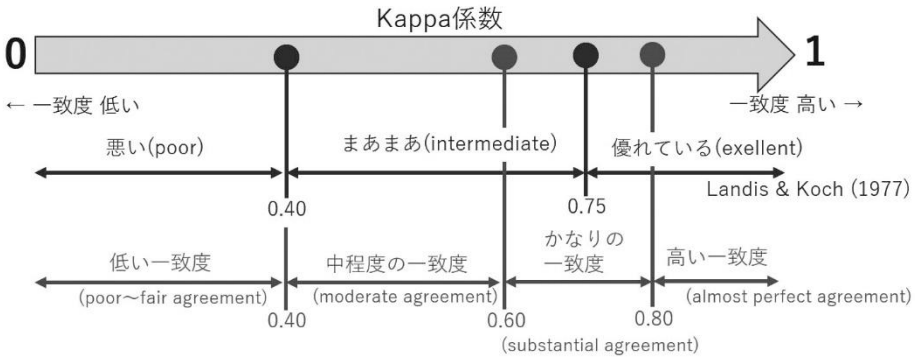
$$\text{Kappa係数} = \frac{(\text{測定者の全一致率}) - (\text{偶然による一致率})}{1 - (\text{偶然による一致率})} = \frac{0.907 - 0.517}{1 - 0.517} = 0.810$$

である．Kappa 係数の定義を説明すると，分子は「測定者間の一致率が偶然によって期待される一致率よりどれぐらい大きいか」を表しており，分母は「2 人の測定者の一致率が偶然による一致を含めずに最大限取り得る値はいくらか」を表している．したがって，Kappa 係数は，偶然の一致を調整したときに取り得る最大値を 1 としたときに，測定者の偶然による一致率を除いた時の全一致率がどの程度の割合になるかを表している．

そのため，Kappa 係数は，0 から 1 の範囲をとる．このとき，Kappa 係数の解釈には，次のようなものがある．

---

48 Gardis, L: Epidemiology (forth edition), Elsevier [木原正博・木原雅子・加治正行：疫学 医学的研究と実践のサイエンス，メディカル・サイエンス・インターナショナル，2010].

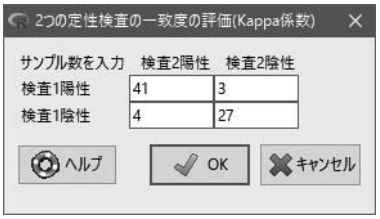


本事例では，Kappa 係数が 0.810 なので，Landis & Koch(1977)[49]の基準では優れている(excelent)であり，一般的な解釈では，高い一致度(almost perfect agreement)と解釈される．

(3) EZR による Kappa 係数の計算

ここでは，組織学的分類のデータを用いて，EZR での計算方法について述べる．

Kappa 係数の計算
1: 「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)」を選択する．
2: 対応のあるクロス集計表のデータを次のように入力する．
3: 「OK」ボタンを押す



このときのアウトプットは，以下のとおりである．

```
$kappa
    点推定値 信頼区間下限 信頼区間上限
1 0.8066298      0.6702305      0.9430292
```

ここで，信頼区間下限，信頼区間上限は，Kappa 係数に対する 95%信頼区間を表している．

(4) 余禄：Kappa 係数と McNemar 検定

対応があるデータには，2 種類(一致性を見る場合，変化(違い)を見る場合)の考え方がある．ここでは，それぞれの見方と評価方法の取捨選択について概説する．

<u>一致性を見る場合</u>

いま，検査 A と検査 B があったとする．検査 A は精度の高い検査方法であるが費用がかかり，検査 B は新しい検査方法で簡便に行えるとする．この 2 つの検査を同じ被験者に行った場合，次のような対応のあるクロス集計表が構成される．

		検査 B	
		陽性	陰性
検査 A	陽性	(a)	(b)
	陰性	(c)	(d)

[49] Landis J.R., and Koch G.G: The measurement of observer agreement for categorical data, Biometrics, 33(1), 159-174, 1977.

この例の場合には，(a) 検査 A および検査 B ともに陽性，(d) 検査 A および検査 B ともに陰性の度数(被験者数)が診断結果が検査 A，検査 B で診断結果が異なる(b)(c)よりも大きくなることが期待される．このように一致性を評価する場合には，Kappa 係数およびその検定が用いられる．

変化（違い）を見る場合

例えば，手術による不安感に関する調査を実施したとする．この調査では，医師からの手術に関する説明前に，手術に対する不安感(あり，なし)を質問したうえで，説明後に同じ質問を行い，医師の説明の適切性を検討している．この場合には，次のような対応のあるクロス集計表が作成される．

		説明後	
		あり	なし
説明前	あり	(a)	(b)
	なし	(c)	(d)

この例の場合には，説明前に不安ありだった患者が説明後に不安なしに変化することが期待される．このように，要因による影響の違いは McNemar 検定を用いて評価する．つまり，帰無仮説「説明前に不安ありの割合と説明後に不安ありの割合」に対して，対立仮説「説明前に不安ありの割合」と「説明後に不安ありの割合は異なる」を検定している．したがって，上記の手術前説明の例の場合には，有意であれば説明前後で被験者の手術に対する不安意識に変化がみられると解釈できる．

4.2 定量検査値の評価

4.2.1 ROC 曲線

(1) データの概要：頭部外傷症データ

頭部外傷症の重篤度を識別するために，CK-BB(クレアチン・キナーゼ BB)が有効か否かを判定している(Zhou et al., 2011[50])．ここに，重篤度は，重度および非重度の 2 値とする．

重症群					非重症群		
140	740	543	490	523	136	60	46
1087	126	913	156	76	286	17	
230	153	230	356	303	281	27	
183	283	463	350	353	23	126	
1256	90	60	323	206	200	100	
700	303	509	1560		146	253	
16	193	576	120		220	70	
800	76	671	216		96	40	
253	1370	80	443		100	6	

このデータは，「ROC_example.csv」で保存されている．

(2) ROC 曲線の概要

ROC 曲線の構成

ROC 曲線は，受信者動作特性曲線(Receiver Operating Characteristic Curve)の略称であり，定量検査値の診断能の評価，最適カットオフ値の選定などに用いられる．

いま，疾患群(D=1)の検査値を X，健常群(D=0)の検査値を Yとする．このとき，疾患の有無を予測するための任意のカットオフ値を u とするとき，診断結果は次のように表すことができる．

[50] Zhou X.H, et al.: Statistical Methods in Diagnostic Medicine (2nd edition), Wiley, 2011.

	疾患群 (D=1)	健常群(D=1)
検査陽性 (検査値≧u)	真陽性 (TP: True Positive)	偽陽性 (FP: False Positive)
検査陰性 (検査値＞u)	偽陰性 (FN: False Negative)	真陰性 (TN: True Negative)

すなわち，定量検査値であっても，カットオフ値 u が決定すれば，定性予測値と同様に，感度及び特異度を定義できる．すなわち，

$$\text{カットオフ値}u\text{での感度}\Pr(X \geq u) = \frac{\text{検査値}X\text{が}u\text{以上の疾患患者数}}{\text{疾患患者数}}$$

$$\text{カットオフ値}u\text{での特異度}\Pr(Y \geq u) = \frac{\text{検査値}X\text{が}u\text{以上の健常者数}}{\text{健常者数}}$$

である．このとき，ROC 曲線は，カットオフ値 u を逐次に変化したときの(1-特異度)をX軸，感度をY軸にプロットした階段グラフで構成される．

いま，簡単な数値例を示す．以下のデータは，ある疾患の患者(7 名)と健常者(7 名)の仮想の検査値を表している．

健常者	50.0	40.5	42.9	52.7	40.8	37.4	32.6
疾患患者	55.7	61.7	37.3	50.3	68.4	48.3	65.1

このデータに対して，任意のカットオフ値 u を逐次に変化したときのクロス集計表は，次のように与えられる．

u=68.4

	疾患群 (D=1)	健常群 (D=0)
陽性(≧u)	1	0
陰性(＜u)	6	7

感度 =0.143
1−特異度 =0.000

・・・

u=52.7

	疾患群 (D=1)	健常群 (D=0)
陽性(≧u)	4	1
陰性(＜u)	3	6

感度 =0.571
1−特異度 =0.143

・・・

u=40.5

	疾患群 (D=1)	健常群 (D=0)
陽性(≧u)	6	5
陰性(＜u)	1	2

感度 =0.857
1−特異度 =0.714

・・・

このように計算した結果，次のような表を得ることができる．

カットオフ値	∞	68.4	65.1	61.7	55.7	52.7	50.3	50.0
真陽性	0	1	2	3	4	4	5	5
真陰性	7	7	7	7	7	6	6	5
偽陽性	0	0	0	0	0	1	1	2
偽陰性	7	6	5	4	3	3	2	2
1−特異度	0.000	0.000	0.000	0.000	0.000	0.143	0.143	0.286
感度	0.000	0.143	0.286	0.429	0.571	0.571	0.714	0.714

下に続く

カットオフ値	48.3	42.9	40.8	40.5	37.4	37.3	32.6	32.6
真陽性	6	6	6	6	6	7	7	7
真陰性	5	4	3	2	1	1	0	0
偽陽性	2	3	4	5	6	6	7	7
偽陰性	1	1	1	1	1	0	0	0
1−特異度	0.286	0.429	0.571	0.714	0.857	0.857	1.000	1.000
感度	0.857	0.857	0.857	0.857	0.857	1.000	1.000	1.000

上表において，X軸に 1−特定度，Y軸に感度をプロットしたものが図 4.2 である．これが ROC 曲線である(図 4.2 において，データ点を描写しているが，一般には描写しないことに注意されたい)． ROC 曲線は，座標(0,0)から座標(1,1)までの階段状にプロットされる[51]．

[51] 正規分布を仮定した場合，あるいは曲線を当てはめた場合には，曲線で描写することもできる．これを平滑化 ROC 曲線と呼ぶが，臨床研究においては，平滑化 ROC 曲線を用いるのは稀である．

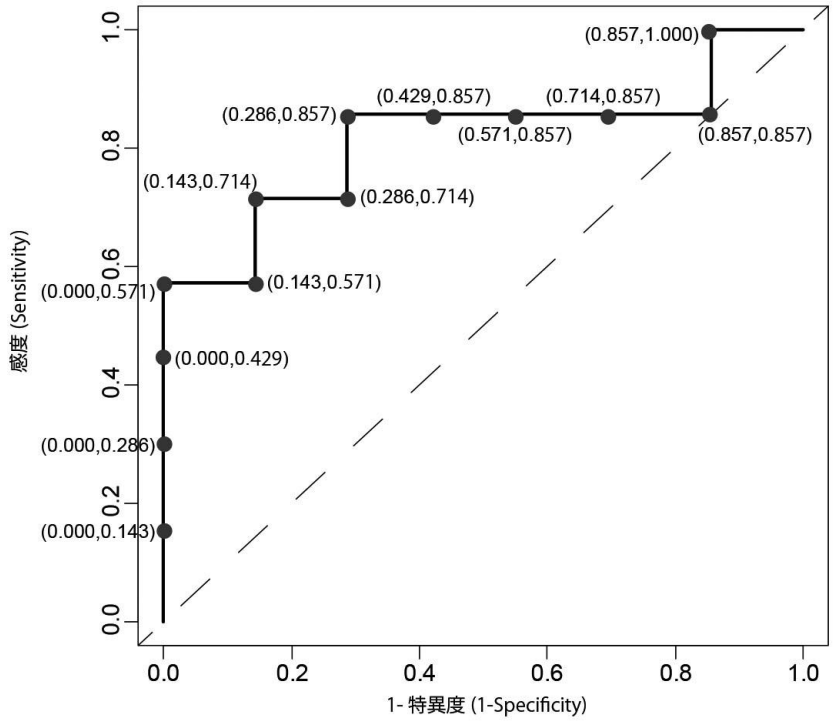
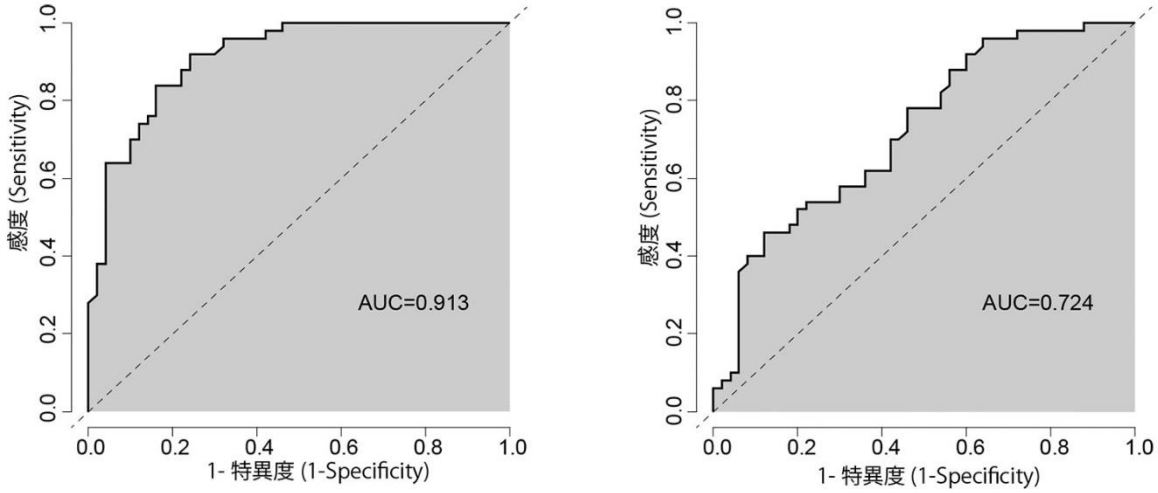


図 4.2：ROC 曲線の例

ROC 曲線の解釈

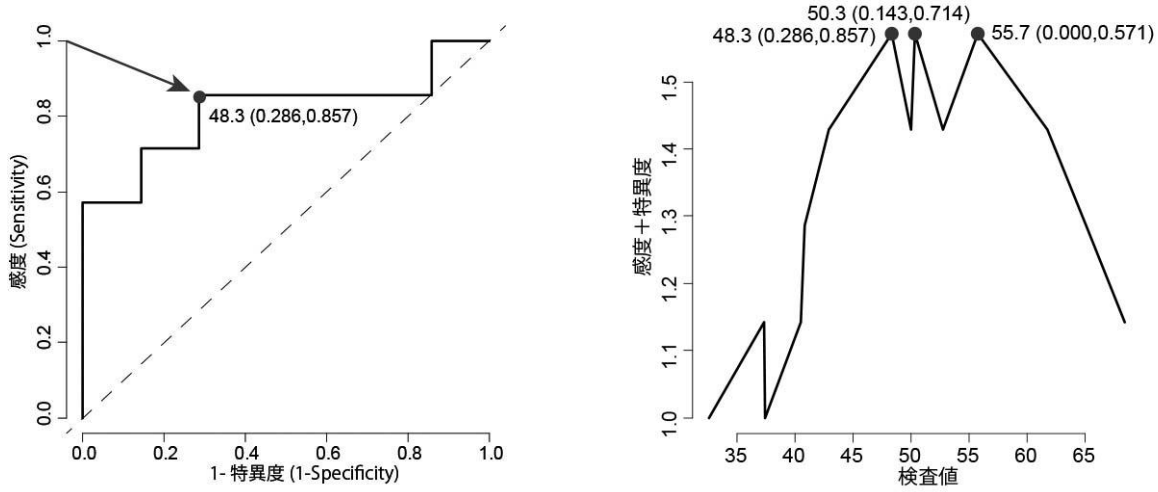
図 4.3 は，2 種類の ROC 曲線を表している．検査値Aの ROC 曲線のほうが(図 4.3(a))，検査値 B の ROC 曲線(図 4.3(b))よりも 45 度の直線(点線)から離れており，座標(0,1)に近くなっている．このような場合に，検査値Aのほうが，検査値 B よりも診断能に優れていると解釈される．

また，ROC 曲線に基づく，診断能を評価する指標に曲線下面積(AUC; Area Under Curve)がある．ROC 曲線の曲線下面積とは，ROC 曲線より下部分の面積(図 4.3 の灰色の部分の面積)であり，0.5～1.0 までの範囲をとる．このとき，



(a) 検査値 A の ROC 曲線　　(b) 検査値 B の ROC 曲線

図 4.3：ROC 曲線の解釈



(a) 座標(0,1)に最も近いカットオフ値　　(b) 感度+特異度が最大になるときのカットオフ値

図 4.4：ROC 曲線における最適カットオフ値の選定

1.0 に近づくほど，検査値に診断能があると解釈できる．図 4.3 の二つの ROC 曲線では，検査値Aの ROC 曲線の曲線下面積 AUC=0.913 に対して，検査値 B の ROC 曲線の曲線下面積 AUC=0.724 であることから，検査値 A の曲線下面積のほうが検査値 B よりも高く，診断能に優れていることがわかる．

ROC 曲線に基づく最適カットオフ値の選定

ROC 曲線の用途の一つが，最適なカットオフ値の選定である．ROC 曲線に基づく最適カットオフ値の選定には，様々な方法が提案されているが，EZR では，(a) 座標(0,1)に最も近いカットオフ値を選定する，(b) 感度+特異度が最大になるときのカットオフ値を選定する，の 2 種類が提案されている．

先ほどの ROC 曲線の数値例において 2 種類のカットオフ値を選定したときの例示を図 4.4 に示す．座標(0,1)に最も近いカットオフ値を選定する場合には，48.3 が最適カットオフ値に選定される(図 4.4(a))．一方で，感度+特異度を最適カットオフ値に選定する場合には，48.3, 50.3, 55.7 の 3 個が選定される．このとき，EZR では最小値が選定されるため，48.3 が最適カットオフ値として選定される．

最適カットオフ値の選定には，ゴールド・スタンダードが存在しないが，座標(0,1)に最も近いカットオフ値を選定することが多いように思われる．

(3) EZR による ROC 曲線の計算

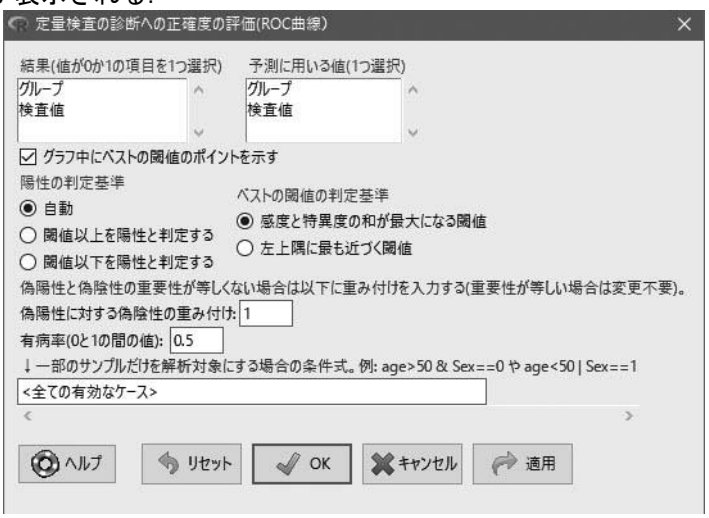
ここでは，頭部外傷症のデータ(ROC_example.csv)を用いて，EZR での計算方法について述べる．このとき，最適カットオフ値の選定には，座標(0,1)に最も近い検査値を用いることにする．

なお，頭部外傷症データは，以下の手順で読み込むことができる．

「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」を選定し，ファイル(ROC_example.csv)を選択する．ここでは，「グループ」にグループ変数(重症，非重症)，「検査値」に検査値が入力されている．

このとき，ROC 曲線の描写は，以下の手順で行うことができる．

ROC 曲線の描写
1:　「統計解析」→「検査の正確度の評価」→「定量検査診断への正確度の評価(ROC 曲線)」を選択する.
2:　次のようなメニューが表示される.

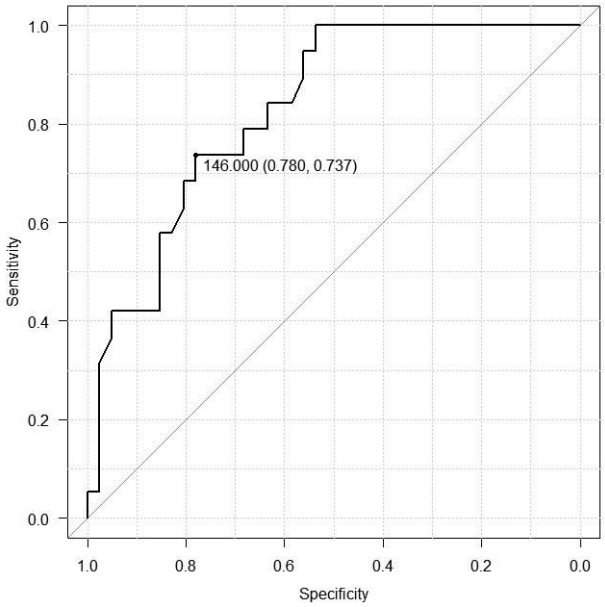


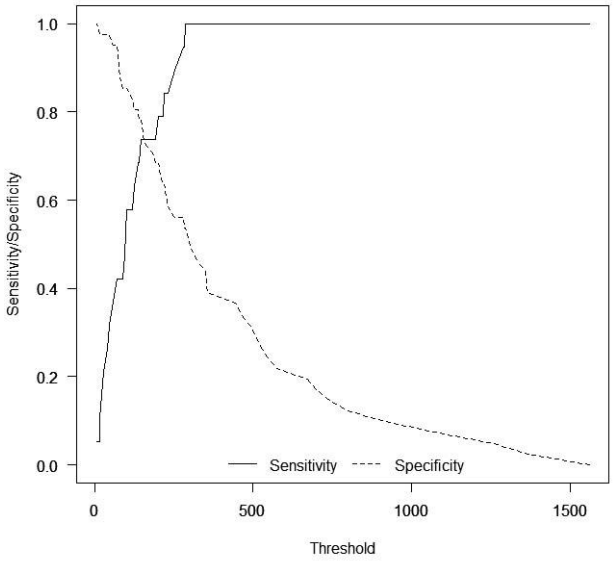
このとき,

- 「結果(値が 0 か 1 の項目を 1 つ選択)」で「グループ」を選択する.
- 「予測に用いる値(1 つ選択)」で「検査値」を選択する.
- 「ベストの閾値の判定基準」で「左上隅に最も近づく閾値」を選択する.

3:　「OK」ボタンを押す

このとき, 次のような 2 種類のグラフが描写される.





ここで, 左側のグラフは, ROC 曲線であり, 最適なカットオフ値が黒丸で表され, カットオフ値(特異度, 感度)が表示される. 頭部外傷症データでの最適なカットオフ値は, 146.00 であり, このときの特異度は 0.780(78.0%)であり, 感度は 0.737(73.7%)であった. また, 右側は ROC 曲線の感度(実線), 特異度(点線)をグラフで表したものである. ここで X 軸はカットオフ値を表している.

さらに, ROC 曲線では, 曲線下面積(AUC)に関する情報についても「出力」画面に表示される.

```
曲線下面積 0.829 95%信頼区間 0.726 - 0.932
```

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される. 赤色が R のコマンド, 青色が出力であるものの曲線下面積以外の情報は, ROC 曲線に描写されていることから, 改めて見る必要がない(数値情報が知りたい場合には参照されたい).

曲線下面積における 95%信頼区間であるが, このとき, 信頼区間のなかに 0.50(診断能が全くない)を含んで切る場合には(信頼区間の下限値が 0.5 を下回る), 当該検査値の診断能は不十分であると判断される. 今回の場合には, 信頼区間の中に 0.5 を含まないことから, 十分な診断能があると判断される.

4.2.2 二つの ROC 曲線の曲線下面積の比較

(1) データの概要：仮想の 2 種類の診断データ

ここでは, 仮想データを用いて二つの ROC 曲線を比較する. このデータは, 被験者毎に 2 種類の検査方法で検査値を取得したときの診断能を比較している.

疾患の有無	検査方法 1	検査方法 2	疾患の有無	検査方法 1	検査方法 2
あり	59	95	なし	63	93
あり	61	123	なし	64	99
あり	55	74	なし	65	119
あり	66	145	なし	64	92
あり	52	64	なし	68	112
なし	60	84	なし	64	99
なし	61	128	なし	69	113
なし	51	79	あり	62	92
あり	60	112	あり	64	112
あり	61	107	なし	67	128
あり	56	67	なし	65	111
なし	65	98	なし	66	105
なし	63	105	なし	62	104
なし	58	95	なし	66	106
なし	59	79	あり	65	112
あり	61	81	あり	60	115
あり	62	91	なし	68	128
あり	65	142	あり	62	116
あり	63	84	なし	68	134
あり	62	85	なし	70	172

このデータは,「ROC_comp.csv」で保存されている.

(2) ROC 曲線における比較

ROC 曲線の比較には, (1)最適なカットオフ値での正診率を比較する, (2) ROC 曲線の曲線下面積を比較することが考えられる. (1)の場合には, McNemar 検定を用いることで, 定性検査値と同様の手順で実行できる. 一方で, 最適なカットオフ値は選定方法によって様々であり, 定量検査値の診断能を評価しているわけではない. したがって, (2)を用いて評価することが推奨される.

ROC 曲線の曲線下面積の比較では, 対応がない場合と対応がある場合が存在する. 対応がない場合とは, 例えば, 性別で 2 群に分けられたグループに対して, 2 つの ROC 曲線を構成する. そして, 検査値の診断能に性差があることを評価する場合が該当する. 一方で, 対応がある場合とは, 被験者から 2 種類の検査値を採取し, それらの検査値の診断能を比較する場合が該当する. EZR では対応がない場合の ROC 曲線の曲線下面積の検定方法は実装されておらず, 対応がある場合のみが実装されている. そのため, ここでは, 対応がある場合を想定して議論する.

帰無仮説 H_0「2 つの検査値の曲線下面積は等しい」に対して，対立仮説 H_1「2 つの検査値の曲線下面積は異なる」を検定する．このような検定方法には，様々な方法が提案されているが，EZR では，Delong の検定[52]が採用されている．

(3) EZR による ROC 曲線の計算

ここでは，仮想データ(ROC_comp.csv)を用いて，EZR での計算方法について述べる．

なお，仮想データは，以下の手順で読み込むことができる．

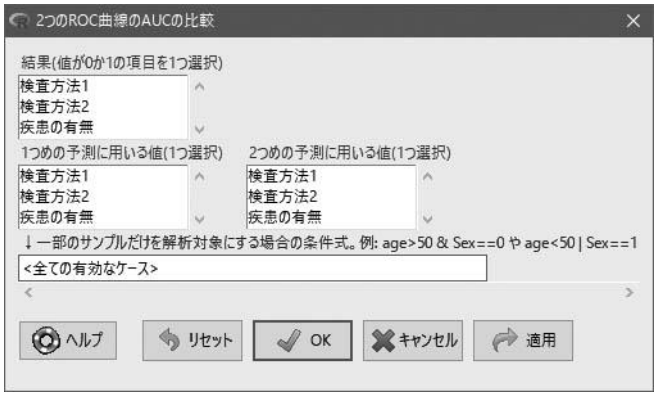
「ファイル」→「データのインポート」→「ファイルまたはクリップボード、URL からテキストデータを読み込む」を選定し，ファイル(ROC_comp.csv)を選択する．ここでは，「疾患の有無」にグループ変数(あり，なし)，「検査方法 1」「検査方法 2」にそれぞれ検査値が入力されている．

このとき，ROC 曲線の曲線下面積の比較は，以下の手順で行うことができる．

ROC 曲線の曲線下面積の比較

1: 「統計解析」→「検査の正確度の評価」→「2 つの ROC 曲線の AUC の比較」を選択する．

2: 次のようなメニューが表示される．



このとき，

- 「結果(値が 0 か 1 の項目を 1 つ選択)」で「疾患の有無」を選択する．
- 「1 つめの予測に用いる値(1 つ選択)」で「検査方法 1」を選択する．
- 「2 つめの予測に用いる値(1 つ選択)」で「検査方法 2」を選択する．

3: 「OK」ボタンを押す

このとき，次のような出力が表示される．

```
            曲線下面積     P 値
検査方法 1      0.736  0.0542
検査方法 2      0.578
```

この出力の上側には R のスクリプト(赤色)及び出力結果(青色)が表示される．赤色が R のコマンド， 青色が出力であるものの同様の情報が重複して表示されているだけであることから，改めて見る必要がない．

その結果，検査方法1の曲線下面積が 0.736，検査方法2の曲線下面積が 0.578 であり，検査方法 1 のほうが診断能は高かったものの，p 値は 0.542 であることから，有意な違いは認められなかった．

[52] Delong E.R., Delong D.M., Clarke-Peason, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics, 44(3), 837-845, 1988.

5 章：傾向スコアによる解析

5.1 傾向スコアの概要

5.1.1 共変量の種類と傾向スコアの関係

ここでは，説明のために薬剤(新薬，既存薬)投与によるアウトカムへの影響を考える．無作為化比較臨床試験では，被験者の背景因子等が同一になるようにそれぞれの薬剤をランダムに割り付ける．一方で，観察研究では，被験者に対する治療の選択が研究者に委ねられていない(介入がない)ため，薬剤群間で背景因子等に偏りが生じ，その結果としてアウトカムに影響を及ぼす可能性がある．このような，治療(要因)とアウトカム(結果)の因果関係に影響を及ぼす第 3 の変数のことを共変量(covariate)という．

図 5.1 は，因果関係に対する共変量の影響を表している(Leite, 2017)[53]．共変量には，4 種類のパターンが存在する．治療予測子(treatment predictor)は，治療選択のみに関連する因子である．例えば，新薬とジェネリック医薬品が存在するときに，患者は，薬価によって薬剤を選択するかもしれない．このとき，薬価が治療予測子になる．治療予測子はアウトカムに影響を及ぼさないことから，傾向スコアの計算には不要である．

媒介変数(mediator)とは，治療とアウトカムを媒介する因子である．例えば，抗癌剤の支持療法(実薬群，プラセボ群)に対する無作為化比較臨床試験において，抗癌剤の治療完遂割合がアウトカムの一つとする．この試験では，支持療法の有無が(任意の)有害事象の発現に影響を与え，その有害事象が抗癌剤の治療完遂割合に影響を与える．

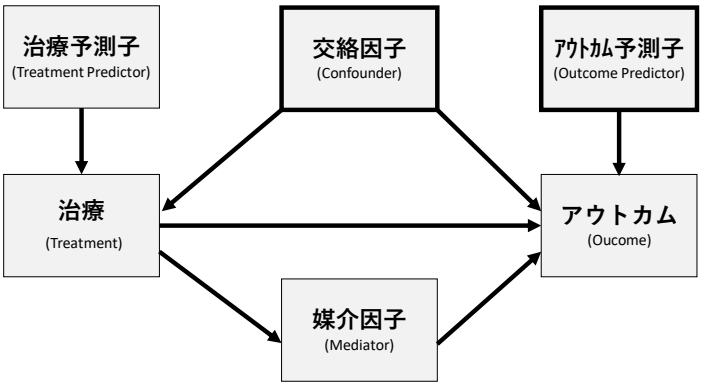


図 5.1：因果関係に関する模式図

[53] Leite, W.：Practical Propensity Score Methods using R, SAGE, 2017.

このような場合には，(任意の)有害事象の発現が媒介変数になる．媒介変数は，原因(例：薬剤投与・治療の選択)の影響を受けることから，傾向スコアの計算には不向きである．

交絡因子(confounder)とは，治療とアウトカムの双方に影響を及ぼす因子である．例えば，2 種類の治療方法(治療 A，治療 B)の有効割合を比較する観察研究を考える．研究の結果，若年者では治療 A が選択される傾向にあり，高齢者では治療 B が選択される傾向が認められ，また，高齢者に比べて若年者のほうが有効割合が高い傾向が認められたとする．このような場合には，年齢層が原因(治療法)と結果(有効割合)に影響していることから，交絡因子になる．傾向スコアとは，主として交絡因子を調整することで，原因への影響を排除することを意図している．

アウトカム予測子(outcome predictor)とは，アウトカムのみに影響を及ぼす因子である．例えば，2 種類の治療方法(治療 A，治療 B)の有効割合を比較する臨床試験を考える．試験の結果，軽症患者に比べて重症患者のほうが有効割合が高い傾向が認められたとする．このような場合には，進行程度が結果(有効割合)に影響を与えることからアウトカム予測子になる．アウトカム予測子によるアウトカムへの影響は，共分散分析あるいは多変量解析手法を用いることで，統計学的に排除することができる．

5.1.2 医学系研究のデザインと因果推論

臨床研究において，研究対象の最小の単位(統計学では個体と呼ぶ)は，被験者である．ある疾患に対する治療(新薬，既存薬)の効果を比較するとき，個体 i に新薬を投与したときの結果を Y_i^{T}，個体 i に既存薬を投与したときの結果を Y_i^{C} とするとき，個体 i に対する潜在的な治療の差(個体治療効果) τ_i は，

$$\tau_i = Y_i^{\mathrm{T}} - Y_i^{\mathrm{C}}$$

で与えらえる．個体毎での潜在的な治療効果の差がわかれば，研究対象での平均的な潜在的な治療効果(平均治療効果)を求めることができる．

しかしながら，新薬が投与された被験者(個体)は既存薬が投与されることはなく，既存薬が投与された被験者(個体)は新薬が投与されることはない[54]．

図 5.2 は，薬剤投与群(新薬投与群：z_i=1，既存薬投与群：z_i=0)と実際に投与された薬剤での組み合わせを表している．ここで，

- ・新薬投与群に新薬を投与した場合の結果：$Y_{i|z=1}^{\mathrm{T}}$
- ・新薬投与群に既存薬を投与した場合の結果：$Y_{i|z=1}^{\mathrm{C}}$
- ・既存薬投与群に新薬を投与した場合の結果：$Y_{i|z=0}^{\mathrm{T}}$
- ・既存薬投与群に既存薬を投与した場合の結果：$Y_{i|z=0}^{\mathrm{C}}$

である．また，$Y_{i|z=1}^{\mathrm{T}}$，$Y_{i|z=1}^{\mathrm{C}}$，$Y_{i|z=0}^{\mathrm{T}}$，$Y_{i|z=0}^{\mathrm{C}}$ の期待値(平均)をそれぞれ $\mathrm{E}\left[Y_{z=1}^{\mathrm{T}}\right]$，$\mathrm{E}\left[Y_{z=1}^{\mathrm{C}}\right]$，$\mathrm{E}\left[Y_{z=0}^{\mathrm{T}}\right]$，$\mathrm{E}\left[Y_{z=0}^{\mathrm{C}}\right]$とする．このとき，

個体 i が新薬投与群 z_i=1 の場合には，既存薬を投与した場合の結果 $Y_{i|z=1}^{\mathrm{C}}$ は不明(欠測)であり(平均 $\mathrm{E}\left[Y_{z=1}^{\mathrm{C}}\right]$ も不明)，

個体 i が既存薬投与群 z_i=0 の場合には，新薬を投与した場合の結果 $Y_{i|z=0}^{\mathrm{T}}$ は不明(欠測)である($\mathrm{E}\left[Y_{z=0}^{\mathrm{T}}\right]$も不明)．した

[54] クロスオーバー試験では，新薬および既存薬が投与される．しかしながら，新薬が投与されたときの被験者の状況(背景因子等)と既存薬が投与されたときの被験者の状況が完全に一致することはない．

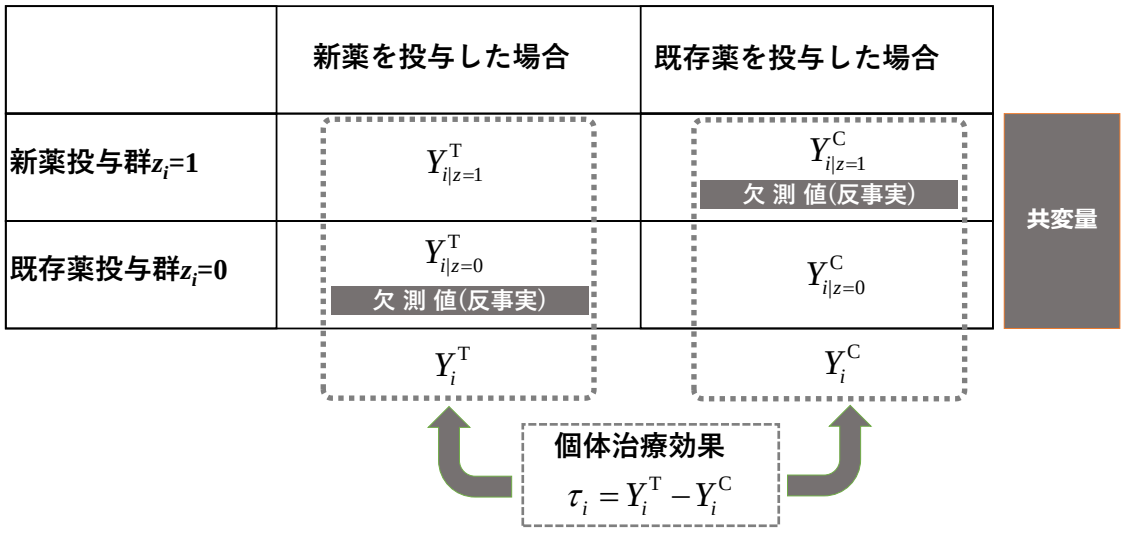


図 5.2：Neyman−Rubin の反事実モデル

がって，潜在的な個体治療効果を得ることは不可能である．そのため，潜在的な平均治療効果も知ることはできない．これを，Neyman−Rubin の反事実モデルという[55]．

無作為化比較試験では，ランダム割付を行うことで，投与群(新薬投与群，既存薬投与群)のあいだの被験者層をそろえることができる．いいかえれば，個々の被験者では違いがあるものの，その平均的な結果には違いがないことが仮定される．つまり，潜在的な個体治療効果を知ることはできないものの，$\mathrm{E}\left[Y^{T}_{z=0}\right]$ に $\mathrm{E}\left[Y^{T}_{z=1}\right]$ を代用し，$\mathrm{E}\left[Y^{C}_{z=1}\right]$ に $\mathrm{E}\left[Y^{C}_{z=0}\right]$ を代用することで，潜在的な平均治療効果を推定できる．

観察研究では，ランダム割付を行うことができないため，上述のような代用を行うことができない．一方で，観察研究の多くでは原因(説明変数)と結果(応答変数)のみを測定するのではなく，それらに影響を与えることが想定される共変量も測定し，それらを考慮した解析が行われる．星野・岡田(2006)[56]は，観察研究における共変量を考慮した研究の方法を以下の 3 つに分類するとともにその問題点を指摘している．

(1) 均衡化

共変量の値が同じになるペアをつくることで 2 つの群の被験者をサンプリングする方法である．均衡化を行うことでペアの被験者がほぼ同一の共変量になり，2 つの群を構成する被験者集団が均一になることが期待できる．しかしながら，完全に一致するペアを作ることはほぼ不可能である．また，連続量の共変量を用いることはできず(幾つかのカテゴリに分けるしかない)，また，多数の共変量を考慮することは困難である．さらに，共変量の選定には，研究者の主観に委ねられるため，恣意性を排除することはできない．

[55] 本来の Neyman-Rubin の反事実モデルでは，平均治療効果で記載される場合が多いものの，記法が統計学的になるため，ここでは個体治療効果で記載している．
[56] 星野崇宏・岡田謙介：傾向スコアを用いた共変量調整による因果効果の推定と臨床医学・疫学・薬学・公衆衛生分野での応用について，保健医療科学，55(3), 230-243, 2006.

(2) 恒常化・限定

同じ共変量をもつ被験者のみに限定してサンプリングする方法である. この方法では, 被験者集団全体の共変量が均一になるが, 一部の被験者に限定するため, 研究結果の一般可能性が低くなる. また, 均等化と同様に共変量選択の恣意性, 多数の共変量の考慮は困難である.

(3) 統計的な調整

多変量解析などの統計的手法を用いて調整を行う方法である. 後ろ向き研究の多くが, 統計的な調整に基づいて評価されている. 一方で, 統計的な調整では,「応答変数と共変量・説明変数をモデル化」しなければならない. そのため, 誤ったモデルを選択した場合には, 誤った結果を導く惧れがある. また, 統計的調整では, 共変量とアウトカムのあいだの関係性をモデル化してるため, 共変量が交絡因子の場合には, 交絡因子と説明変数のあいだの関係性を調整していない.

これらの問題点を解決するために, Rosenbaum & Rubin[57]が提案した統計学的な概念が傾向スコア(propensity score)である. 傾向スコアとは, 複数の共変量を一つの変数に集約することで, マッチングや層別化などを行う方法である.

5.1.3 傾向スコア・マッチング

傾向スコア解析の手順は, (1)傾向スコアを推定する, (2)傾向スコアを用いて群間の均衡化を行う, (3)傾向スコアにより均衡化された結果を用いて平均治療効果を推定する, の 3 段階で行われる.

傾向スコアの推定は, 治療群を 2 値(1:処理群, 0:対照群)で表した応答変数に対する回帰分析(説明変数は共変量である)を用いる. そして, 回帰モデルによって推定される予測確率(個体 i が処理群に属する確率)が傾向スコアの推定値として用いられる.

傾向スコアを推定するための回帰モデルとして一般的に用いられているのがロジスティック回帰分析である. (1)適切にモデルが当てはまっているかを検討する, (2) 傾向スコアによる均衡化後に共変量の分布が群間で同じになっていることを確認する, ことが重要である. モデル適合度の評価には, 疑似決定係数あるいは C 統計量(C-index)を用いることができる. 最近の多くの研究では, C 指標を用いており, 0.80 以上であることが一つの判断基準になっている.

傾向スコアを用いて群間の均衡化を行う方法には, (1) マッチング, (2) 層別化, (3) 逆確率重み付け, (4) 共分散分析, がある. ここでは, 最も用いられているマッチングについて触れる.

マッチングとは, 傾向スコアの一致した(あるいは極めて近い)個体同士を選択する方法であり, 傾向スコアによる均衡化のなかで最も応用されている.

図 5.3 はマッチングのアルゴリズムを表している . マッチングでは, 処理群の任意の個体に対して, 傾向スコアが最も近い対照群の個体を対応させる作業をすべての処理群の個体に対して実行する. その利点は, (a) 均衡化の実行過程が理解しやすい, (b) マッチング後の共変量の分布を点検することが容易である, (c) マッチングされたデータは通常のデータ解析と同様に取り扱うことができる, がある. 一方で, 群間で傾向スコアの重なりが少ない場合, あるいは, 処理群の標本サイズが対照群に比べて著しく少ない場合には, マッチング

[57]Rosenbaum PR, Rubin DB: The central role of the propensity score in observational studies for causal effect, Biometrika, 70, 41-55, 1983.

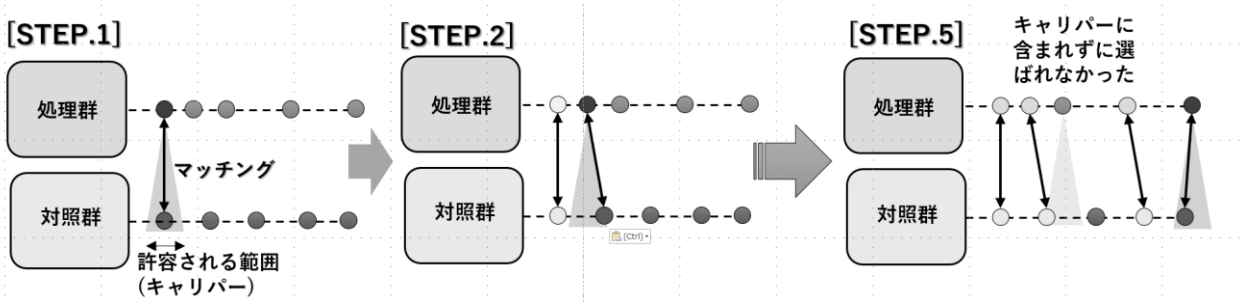


図 5.3:マッチングのアルゴリズム例(キャリパーを伴う 1:1 強欲アルゴリズム)

後のデータが大幅に削減されるため，効率が悪くなる(標本サイズの削減に伴い，検定の検出力が下がる)ことがある.

表 5.1 は，マッチングにおける留意点を整理したものである．マッチングは，(1) マッチングのアルゴリズム，(2) マッチング後の対照群の個体の取扱い，(3) マッチングの比率，(4) キャリパーの設定，を行わなければならない.

マッチングのアリゴリズムとして広範に用いられているが最近傍マッチング(nearest neibhor matching)及び最適マッチング(optimal matching)である．最近傍マッチングは，処理群の任意の個体に対して，傾向スコアが最も近い対照群の個体を逐次に探索する方法である(図 5.4 の説明は最近傍マッチングを用いている)．一方で，最適マッチングは，傾向スコアの距離の合計が最小になるように個体間をマッチングする方法である．最適マッチングは，処理群と対照群の標本サイズに違いが大きいとき，あるいは標本サイズが小さい場合に最近傍マッチングに比べて優れている.

マッチングのアルゴリズムには，多くの方法が提案されており，例えば，遺伝的マッチング(genetic mathing)，フルマッチング(full matching)などがある.

表 5.1:マッチングにおける留意点

項目	説明
マッチングのアルゴリズム	・ 最近傍マッチング(nearest neighbor matching) 処理群の任意の個体に対して，傾向スコアが最も近い対照群の個体を逐次に探索する方法(マッチグの順番の影響を受ける) ・ 最適マッチング(optimal matching) マッチング後の傾向スコアの距離の合計値が最小になるようにマッチングを行う.
マッチング後の対照群の個体の取扱い	・ 復元マッチング 処理群の異なる個体に対して同じ対照群の個体を対応させることを許容する. ・ 非復元マッチング 処理群の異なる個体に対して同じ対照群の個体を対応させることを許容しない.
マッチングの比率	・ 1:1 マッチング 1 名の治療群と 1 名の対照群をマッチングを行う. ・ 固定比マッチング(1:k マッチング) 1 名の治療群と k 名の対照群をマッチングを行う. ・ 変動比マッチング 1 名の治療群と複数(個体毎に変動，上限のみ設定)の非暴露群でマッチングを行う.
キャリパー(マッチングさせる許容領域)の設定	キャリパーとは，マッチングさせる許容領域を表しており，マッチングされたペアの傾向スコアの距離がキャリパー以上であればマッチングしない.

マッチング後の個体の取扱いには，復元マッチング(with replacement matching)と非復元マッチング(without replacement matching)がある．復元マッチングは，処理群の異なる個体に対して同じ対照群の個体を対応させることを許容し，非復元マッチングでは許容しない．そのため，傾向スコアのバイアス低減の点では，非復元マッチングのほうが優れている．一方で，非復元マッチングでは，群間の症例数のインバランスが起きる可能性がある．症例数のインバランスは，検出力を低下させる可能性がある．また，対照群の 1 名の個体に複数の治療群をマッチングさせる可能性があるため，復元マッチングは殆ど用いられていない．

マッチングの比率の設定には，1:1 マッチング(one-to-one matching)，固定比(1:k)マッチング(fixed rate matching, one-to-k matching)，変動比マッチング(variable rate mathing, one-to-many matching)がある，1:1 マッチングは，1 名の処理群と 1 名の対照群でマッチングする方法である．1:1 マッチングは，例数の減少が最も顕著であるが，群間の例数の不均衡が起こらない．したがって，マッチングによる症例数の減少が少なければ，アウトカムの比較における検出力の低下が最も少ない(Cohen, 1988)[58].

固定比マッチング(1:k マッチング)とは，1 名の処理群と k 名の対照群でマッチングする方法である．固定比マッチングは，選択される対照群の個体数が固定されるため，推奨されない(Leite, 2017)[59].

変動比マッチングとは，1 名の処理群と複数(個体ごとに変動，上限のみ設定)の対照群でマッチングする方法である．変動比マッチングは，処理群の例数が対照群の例数よりもかなり少ない場合には， 1:1 マッチングに比べて有効である(Leite, 2017)[60].

キャリパーとは，マッチングさせる許容領域を表しており，マッチングされたペアの傾向スコアの距離がキャリパー以上であればマッチングしない．キャリパーの設定は，(定数)×(傾向スコアの標準偏差 SD)で設定される．定数が大きくなるほどマッチングの許容領域が広くなる(マッチングの制限が緩くなる)．定数は任意に設定することができるが，Rosenbaum & Rubin(1983)[61]は 0.25×SD をキャリパーに設定することを推奨している．近年では，0.2×SD を採用する論文が多くなっている．

5.2 傾向スコア・マッチングによる統計解析

5.2.1 データの概要

これは，新薬(A)を投与した 71 例と既存薬 B(C)を投与した 101 例の背景因子(性別，年齢，喫煙の有無，BMI，重症度スコア)と治療効果を調査した後ろ向き研究のデータである．このデータは，PSexample.csv で与えられる．このファイルにおいて，変数の名称と説明を以下に示す．

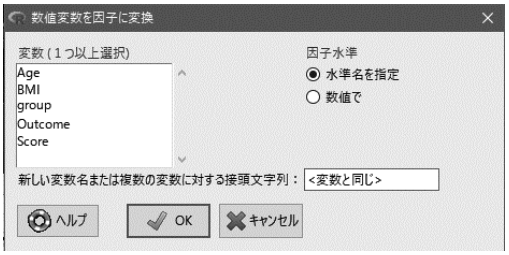
Sex：性別(M：男性，F：女性)，　Age：年齢，　Smoke：喫煙歴(1：有，0：無)，
BMI：Body Mass Index，　Score：重症度スコア，　group：薬剤(A:新薬，C：既存薬)
Outcome：アウトカム(1：改善，0：非改善)

5.2.2 EZR による傾向スコア・マッチング

共変量の要約:ここでは，先ず，喫煙歴をカテゴリカル変数に変換する方法について説明する(必須ではない)．

カテゴリ変数への変換
1:　「アクティブデータセット」→「変数の操作」→「サンプルの背景データのサマリー表の出力」を選択する．
2:　次のようなメニューが表示される．

[58] Cohen J: Statistical Power Analysis for the Behavioral Science (2nd edition), Routledge, 1988.
[59] Leite, W.：Practical Propensity Score Methods using R, SAGE, 2017.
[60] Leite, W.：Practical Propensity Score Methods using R, SAGE, 2017.
[61] Rosenbaum PR, Rubin DB: The central role of the propensity score in observational studies for causal effect, Biometrika, 70, 41-55, 1983.

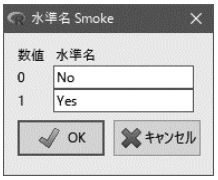


このとき,

- 「変数(1 つ以上選択)」で「Smoke」を選択する.
- 「因子水準」で「水準名を指定」を選択する.
- 「新しい変数名または複数の変数に対する接頭文字列」で「<変数名と同じ>(デフォルト)」を選択する.

3: 「OK」ボタンを押す

4: 次の画面が表示される.



このとき,

- 「0」で「No」を入力する.
- 「1」で「Yes」を入力する.

5: 「OK」ボタンを押す

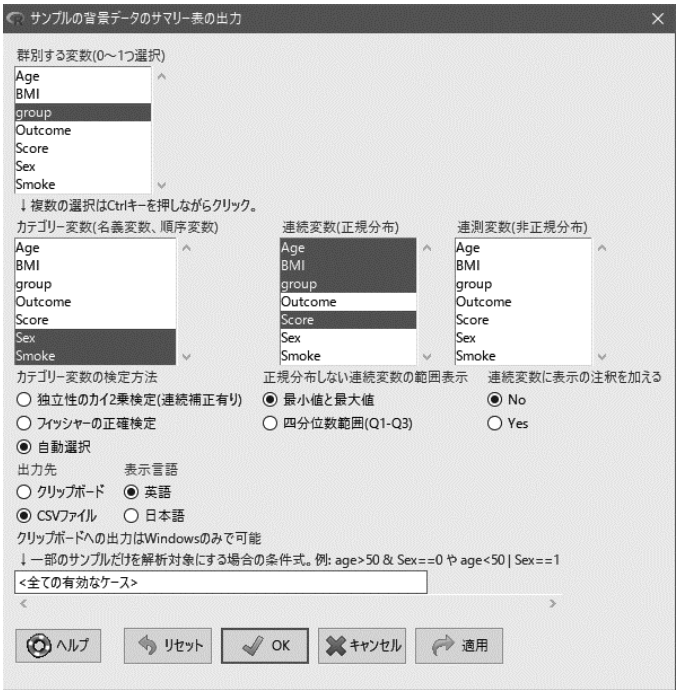
これにより, 1 が Yes, 0 が No に置き換えられる. 確認する場合には, メニュー下に「編集」「表示」「保存」と並んでいるボタンのなかで「表示」を選択すると, データの内容を閲覧できる.

傾向スコアの推定: 次いで, 新薬群と既存薬群の共変量について要約(背景表の作成)する. このとき, 「Sex」(性別), 「Smoke」(喫煙歴)はカテゴリカルデータであり, 「Age」(年齢), 「BMI」(Body Mass Index), 「Score」(重症度スコア)は連続データなので, 「群」(group)別に要約すると次のような手順で実行できる.

背景表の作成

1: 「グラフと表」→「検査の正確度の評価」→「サンプルの背景データのサマリー表の出力」を選択.

2: 次のようなメニューが表示される.



このとき，
・「群別する変数(0～1 つ選択)」で「group」を選択する．
・「カテゴリ変数(名義変数，順序変数)」で「Sex」「Smoke」を選択する．
・「連続変数(正規分布)」で「Age」，「BMI」，「Score」を選択する．
3: 「OK」ボタンを押す

なお，「自動選択」をクリップボードにすると，クリップボードに結果が保存され，WORD などに結果を貼り付けることができ，CSV ファイルを選択した場合には，結果をファイルに保存することができる．

さらに，カテゴリカル変数の場合には，カイ 2 乗検定と Fisher の正確検定を選択することができ，「連続変数(正規分布)」の場合には，2 標本 t 検定のp値，「連続変数(非正規分布)」の場合には，Wilcoxon 検定の p 値が選択される．

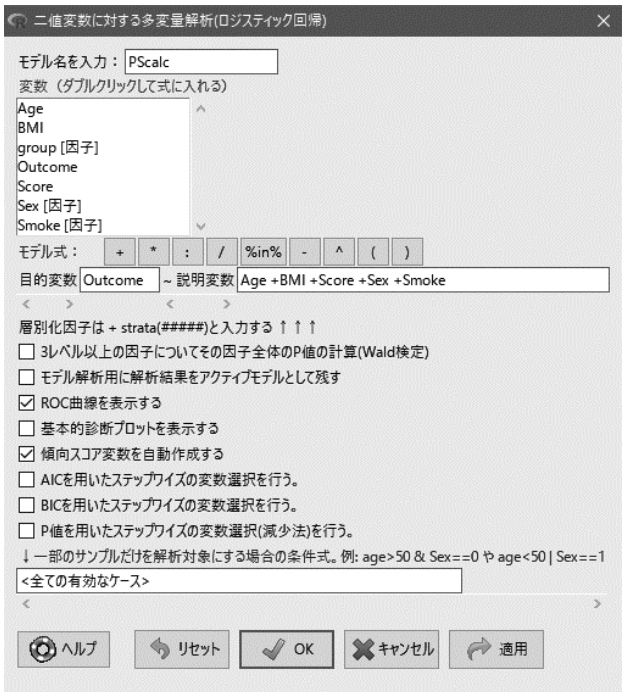
このときの結果を以下に示す．

Factor	Group	group A	group C	p.value
n		71	110	
Sex (%)	F	19 (26.8)	41 (37.3)	0.150
	M	52 (73.2)	69 (62.7)	
Smoke (%)	No	37 (52.1)	80 (72.7)	0.007
	Yes	34 (47.9)	30 (27.3)	
Age		54.11 (8.62)	53.45 (8.32)	0.604
BMI		25.03 (3.77)	24.10 (2.92)	0.063
Score		8.85 (1.68)	7.30 (1.94)	<0.001

喫煙歴(Smoke)，BMI(BMI)，および重症度スコア(Score)に違いが認められている．言いかえれば，治療群間で，これらの共変量に偏りが認められる(患者背景が異なる)．このとき，ロジスティック回帰分析による傾向スコアの計算は，

傾向スコアの推定

1: 「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)」を選択する．
2: 次のようなメニューが表示される．



このとき，
・「モデル名を入力」で「PScalc」と入力(任意で設定しても構わない)．
・「モデル式：」において，

目的変数 group ～説明変数 Age+BMI+Score+Sex+Smoke

と入力する.

・「ROC 曲線を表示する」にチェックを入れる.

・「傾向スコア変数を自動作成する」にチェックを入れる.

3: 「OK」ボタンを押す

で実行できる. 上記のロジスティック回帰は, 群(group)を応答変数, 年齢(Age), BMI(BMI), 重症度スコア(Score), 性別(Sex), 喫煙歴(Smoke)を説明変数としたうえで計算している.

このときの結果(青色の部分)を以下に示す.

Output.1

```
Call:
glm(formula = group ~ Age + BMI + Score + Sex + Smoke, family = binomial(logit),
    data = Dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4224  -0.8836   0.4825   0.8979   1.8517

Coefficients:
             Estimate Std. Error z value    Pr(>|z|)
(Intercept)   8.56483    2.19862   3.896 0.000097976 ***
Age          -0.01321    0.02062  -0.640     0.52187
BMI          -0.12633    0.06327  -1.997     0.04587 *
Score        -0.50981    0.10091  -5.052 0.000000436 ***
Sex[T.M]      0.38266    0.43739   0.875     0.38165
Smoke[T.Yes] -1.20326    0.39155  -3.073     0.00212 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 242.45  on 180  degrees of freedom
Residual deviance: 198.92  on 175  degrees of freedom
AIC: 210.92

Number of Fisher Scoring iterations: 4
```

Output.1 は, ロジスティック回帰の適合結果を表している. 年齢(Age)および性別(Sex)の回帰パラメータに対する p 値は有意でない. 傾向スコアの推定に変数選択を用いるほうが良いとの意見があるものの, 一方で, モデル自体を解釈するわけではないため, 多重共線性が認められなければよいとの意見もある(有意でなくても, 僅かでも各共変量を調整したほうが良いという意見があるためである).

Output.2

```
AAnalysis of Deviance Table

Model 1: group ~ Age + BMI + Score + Sex + Smoke
Model 2: group ~ 1
  Resid. Df Resid. Dev Df Deviance      Pr(>Chi)
1       175     198.92
2       180     242.45 -5  -43.532 0.00000002882 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output.2 は, 推定されたロジスティック回帰モデルの適合度を検定したものである. p 値が 0.05 を下回ることから, null モデル(共変量が存在しない切片のみのモデル)に対して有意に適合していることが分かる.

Output.3

```
Age      BMI    Score      Sex    Smoke
1.033550 1.203480 1.065016 1.349430 1.218888
```

Output.3 は, 各共変量に対する VIF(Variance Inflation Factor, 分散拡大係数(分散拡大要因))である. VIF が 10 を超える場合には多重共線性の程度が大きいと解釈される場合が多い. 今回の事例では, そのような共変量は認められなかった.

Output.4

	オッズ比	95%信頼区間下限	95%信頼区間上限	P 値
(Intercept)	5240.000	70.500	390000.000	0.000098000
Age	0.987	0.948	1.030	0.522000000
BMI	0.881	0.779	0.998	0.045900000
Score	0.601	0.493	0.732	0.000000436
Sex[T.M]	1.470	0.622	3.460	0.382000000
Smoke[T.Yes]	0.300	0.139	0.647	0.002120000

Output.4 は，各共変量のオッズ比，95%信頼区間および，回帰パラメータに対する有意性検定の p 値である．通常のロジスティック回帰分析の場合には，解釈の中心になるが，傾向スコアの推定では，個々の共変量に対する回帰パラメータを解釈することはないため，無視してかまわない．

Output.5 曲線下面積 0.779 95%信頼区間 0.708 - 0.85

Output.5 は，推定された傾向スコアに対する ROC 曲線の曲線下面積及び 95%信頼区間である．曲線下面積は，ロジスティック回帰モデルの予測確度の指標の一つである．C 指標(C-index)に一致する．傾向スコアでは，C 指標が 0.8 以上を一つの基準にしている．本事例では，0779 なので，僅かに下回るが，そのまま解析する[62]．

なお，このときの傾向スコアがデータに追加される(「PropensityScore」が頭文字になっている．通常は「PropensityScore.GLM.1」である)．

傾向スコア・マッチング：ここでは，傾向スコア・マッチングを実施する．EZR では，「統計解析」→「マッチドペア解析」→「マッチさせたコントロールの抽出」を用いてマッチングを実施できる．一方で，このマッチングでは，キャリパーを設定できないことから，実用的ではない．そのため，R のパッケージ Match を用いる方法を説明する．

R のパッケージ Match(EZR の場合も同じ)では，処理群を 1，コントロール群を 0 としたダミー変数を設定しなければならない．手順を以下に示す．

ダミー変数への変換

1: 「アクティブデータセット」→「変数の操作」→「ダミー変数を作成する」を選択する．
2: 次のようなメニューが表示される．

このとき，
・「ダミー変数を作成する変数を選択」で「group」を選択する．
・「ダミー変数であることを示す文字列」で「Dummy.Group」(任意)と入力する．
3: 「OK」ボタンを押す

上記の処理を実施すると，新たに，groupDummyGroup という変数が作成される．ここで，groupDummyGroupA は，処理群(A)を 1，対照群(C)を 0 とした場合であり，groupDummy.GroupC は，処理群(A)を 0，対照群(C)を 1 とした場合である．

傾向スコア・マッチングは，R のスクリプトを用いる．R のスクリプトは，EZR の画面の R スクリプト内で実行する．

[Step.1] パッケージ Match をインストールする(EZR の操作画面については，0.2.1 節を参照)．

[62] 実際の解析の場合には，2 次交互作用を含めたり，あるいは，高度な非線形回帰モデルを用いる．

```
install.packages("Matching")
```

と入力したうえで，ドラックする．そして，R スクリプト下側の「実行」を押す．

[Step.2] パッケージ Match を読み込む

```
library(Matching)
```

と入力したうえで，ドラックする．そして，R スクリプト下側の「実行」を押す．

[Step.3] マッチングを行う．マッチングの関数 Match()の書式は，

Matching <- Match(Y=*応答* Tr=*群(1:処理群, 0:対照群)*, X=*傾向スコア*, caliper=*キャリパー*, ties=F, replace=F)

である．ここでは，応答が「Dataset$Outcome」，群が「Dataset$groupDummyGroupA」，傾向スコアが「Dataset$PropensityScore」である(メニュー下の「表示」ボタンを押せば，変数名がわかる)．変数の入力は，「Dataset$変数名」で行う．また，キャリーパーが 0.25 とすると，次のようなコマンドになる．

```
Matching <- Match(Y=Dataset$Outcome, Tr=Dataset$groupDummyGroupA,
   X=Dataset$PropensityScore.GLM.1, caliper=0.25,ties=F, replace=F)
summary(Matching)
```

と入力したうえで，ドラックする．そして，R スクリプト下側の「実行」を押す．すると，次のように表示される．

```
Estimate. .  0.2381
SE. . . . .  0.11071
T-stat. . .  2.1507
p.val. . .   0.031499

Original number of observations. . . . . .   181
Original number of treated obs. . . . . .  71
Matched number of observations. . . . . .  42
Matched number of observations  (unweighted).  42

Caliper (SDs). . . . . .    0.25
Number of obs dropped by 'exact' or 'caliper'  29
```

ここで，「Matched number of observations」は，マッチング後の各群の標本サイズを表している．

[Step.4] マッチング後のデータを抽出する．

```
treat <- Dataset[Matching$index.treated,]
control <- Dataset[Matching$index.control,]
Match.data <- rbind(treat,control)
```

と入力したうえで，ドラックする．そして，R スクリプト下側の「実行」を押す．

これにより，新たなデータ集合「Match.data」が生成される．生成後にメニュー下のデータセット右側にある「Dataset」を左クリックして，「Match.data」を選択する．先ほどと同様に，背景表を作成する．

Factor	Group	group A	group C	p.value
n		42	42	
Sex (%)	F	14 (33.3)	16 (38.1)	0.820
	M	28 (66.7)	26 (61.9)	
Smoke (%)	No	27 (64.3)	26 (61.9)	1.000
	Yes	15 (35.7)	16 (38.1)	
Age		54.83 (9.38)	53.52 (7.84)	0.490
BMI		24.61 (2.96)	24.14 (3.16)	0.482
Score		8.26 (1.65)	8.31 (1.94)	0.904

背景表より，すべての共変量が有意でないことが分かる．

そのため，2 群間を Fisher の正確検定により評価する(詳しくは，2.2 節を参照)．このとき，

・「行の選択(1 つ以上選択)」で「group」を選択する．

・「列の選択(1 つ以上選択)」で「outcome」を選択する．

・「パーセントの計算」で「行のパーセント」を選択する．

すると，次のように表示される．

Output.1

```
Outcome
group    0    1 Total Count
    A 50.0 50.0   100    42
    C 73.8 26.2   100    42
```

上記のアウトプットより，処理群の有効割合は 50.0%であり，対照群の有効割合は，26.2%であった．

Output.2

```
           Fisher's Exact Test for Count Data

data:  .Table
p-value = 0.04238
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1272557 0.9693956
sample estimates:
odds ratio
 0.3594044
```

上記のアウトプットより，オッズ比は 0.359 であり，95%信頼区間は[0.128,0.969]であった．オッズ比が 1 をまたいでいないことから，有意であることが伺える．

Output.3

	Outcome=0	Outcome=1	Fisher 検定の P 値
group=A	21	21	0.0424
group=C	31	11	

Fiseher の正確検定の p 値が 0.0424 なので，有意水準 0.05 のもとで有意だった．すなわち，傾向スコア・マッチングにおいて治療群間で違いが認められた

6 章：臨床試験における必要症例数の計算

6.1 症例数設計の基本

症例数設計を行ううえで重要な考え方が第 1 種の過誤(α エラー)と第 2 種の過誤(β エラー)である. 第 1 種の過誤とは, 帰無仮説 H_0 が正しいにも関わらず, 対立仮説 H_1 が正しいと判定する誤りである. そして, 第 2 種の過誤とは, 対立仮説 H_1 が正しいにも関わらず, 帰無仮説 H_0 が正しいと判定する誤りである.

仮説検定の p 値とは, 帰無仮説 H_0 が正しいと仮定した場合に, 臨床試験の結果が得られる確率を表している. すなわち, 仮説検定とは, 第 1 種の過誤が一定水準未満(有意水準 α 未満)であるか否かを確認することを意味する. したがって, 第 1 種の過誤に関する評価は, 臨床試験の結果から得ることができる. 一方で, 第 2 種の過誤が一定水準未満(β 未満)になることを確保するには, 任意の症例数以上にしなければならない. すなわち, 必要症例数の設計とは, 予め規定した第 2 種の過誤 β 未満にすることを目的としている.

ちなみに, 第 1 種の過誤 α は, 論文等では,「α エラー(alpha error)」あるいは「有意水準(significance level)」で表されることが多い. 一方で, 第 2 種の過誤 β は,「β エラー(beta error)」あるいは「検出力(power)」1-β で表されることが多い. すなわち, 検出力とは, 対立仮説 H_1 が正しいときに, 対立仮説 H_1 が正しいと判定する確率を表している.

図 6.1 は, 仮説(帰無仮説 H_0, 対立仮説 H_1)と第 1 種の過誤(α エラー)及び第 2 種の過誤(β エラー)の関係を表している. 実線の曲線が帰無仮説 H_0 のもとでの検定統計量の分布(帰無分布)を表しており, 点線の曲線が対立仮説 H_1 のもとでの検定統計量の分布を表している. そして, これらの分布が重なる部分での検定統計量を u とするとき, 帰無仮説 H_0 のもとで検定統計量が u よりも大きな値をとる確率(濃い灰色の領域の面積)が第 1 種の過誤(α エラー)であり, 対立仮説 H_1 のもとで検定統計量が u よりも小さな値をとる確率(うすい灰色の領域の面積)が第 2 種の過誤である.

このとき, 二つの過誤を小さくするには, (1) 対立仮説 H_1 で想定される治療効果(エフェクトサイズ)を大きくすることで, 対立仮説 H_1 のもとでの検定統計量の分布を右側に移動させる, (2) 標本サイズを大きくすることで分布のバラツキ(青色の矢印)を小さくする, ことが考えられる.

対立仮説 H_1 のもとでの検定統計量の分布は,「臨床的に有効(あるいは安全)であると判断される治療効果の大きさ(エフェクトサイズ)」によって設定される. このとき, エフェクトサイズの設定は, 統計学的な観点ではなく, 臨床的な観点から設定しなければならない. 例えば,「対照治療での中央全生存期間 24.5 カ月に対して, 試験治療では 24.6 ヵ月であるため, 試験治療は有効である」とは言い切れないはずである.

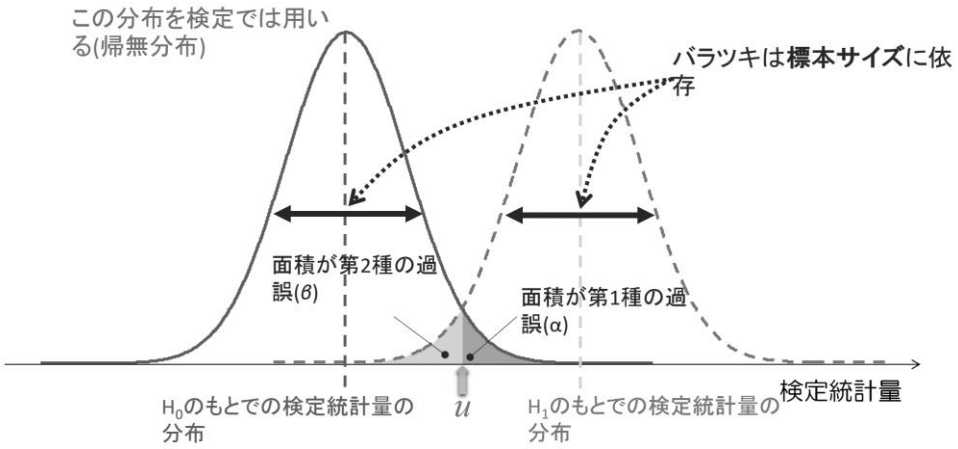


図 6.1：仮説(帰無仮説 H_0, 対立仮説 H_1)と第 1 種の過誤(α エラー)及び第 2 種の過誤(β エラー)の関係

表 6.1：試験デザイン及びアウトカム毎の症例数設計に必要な情報と対応する検定

	1標本（単アーム試験）		2標本（無作為化比較試験）	
	必要な情報	対応する検定	必要な情報	対応する検定
比率	閾値(比率) 期待値(比率)	母比率の検定	対照治療での比率 試験治療での比率 (2群の比率 → オッズ比でも可能)	カイ2乗検定 OR Fisherの正確検定
平均値	閾値(平均値) 期待値(平均値)	母平均の検定	対照治療での平均値 試験治療での平均値 共通の標準偏差	2標本t検定
生存曲線	閾値(中央生存期間 OR 年次生存率) 期待値(中央生存期間 OR 年次生存率) (登録期間及び追跡期間)	1標本ログランク検定 OR 信頼区間	対照治療での中央生存期間 OR 年次生存率 試験治療での中央生存期間 OR 年次生存率 (2群の生存率 → ハザード比でも可能) (登録期間及び追跡期間)	ログランク検定

	対応のある場合	
	必要な情報	対応する検定
比率	試験治療が有効で対象治療が無効の比率 試験治療が無効で対象治療が有効の比率	McNemar検定
平均値	被験者毎のアウトカムの差の平均値 被験者毎のアウトカムの差の標準偏差	対応のあるt検定

表 6.1 は，治療効果の大きさ(エフェクトサイズ)を設定するうえで一般的に必要な情報を表している．ここで，比率とは奏効割合，根治切除割合など，被験者毎に 2 値(奏効の有無，根治切除の有無)で与えられる主要評価項目を表している．また，平均値とは，手術における出血量，定量的な検査値など，被験者毎に量的データで与えられる主要評価項目を表している．さらに，生存曲線は，全生存期間，無増悪生存期間，治療成功期間など，被験者毎に生存期間とイベントの有無で与えられる主要評価項目を表している．なお，生存曲線における症例数設計では，生存期間 t に対して，ハザードが一定であることを仮定することが多い．

1 標本(単アーム試験)における閾値とは，試験治療が上回りたい(否定したい) 主要評価項目の値である．また，期待値とは，試験治療によって期待される主要評価項目の値である．言いかえれば，期待値以上の試験結果が与えられたとき，帰無仮説 H_0「試験治療による真の治療効果が閾値である」を棄却し，対立仮説 H_1「試験治療による真の治療効果は閾値を上回る」を支持できる．

主要評価項目が比率の 2 標本(無作為化比較試験)の症例数設計は，オッズ比に基づいて行われる．そのため，対照治療・試験治療での期待される比率あるいは期待されるオッズ比の情報が必要である．なお，比率の場合には，帰無仮説 H_0「試験治療と対照治療の真のオッズ比は 1 である」に対して症例数設計が行われる．

表 6.2:EZR で計算可能な標本サイズの計算

	タイトル	必要な情報	備考
1	**閾値奏効率，期待奏効率からのサンプルサイズの計算**	臨：閾値奏効率，期待奏効率 統：有意水準，検出力	3とほぼ同じ （オプションが異なる）
2	**1群の比率の信頼区間をある幅におさめるためのサンプルサイズの計算**	臨：想定する比率，信頼区間の幅 統：信頼係数(confidence level)	
3	**1群の比率を既知の比率と比較するためのサンプルサイズの計算**	臨：既知の比率，想定する比率 統：有意水準，検出力	1とほぼ同じ （オプションが異なる）
4	**1群の比率を既知の比率と比較するための検出率の計算**	臨：既知の比率，想定する比率 統：有意水準，標本サイズ	3の検出力計算版
5	**2群の比率の比較のためのサンプルサイズの計算**	臨：グループ1の比率，グループ2の比率 統：有意水準，検出力，サンプルサイズの比	
6	**2群の比率の比較のための検出力の計算**	臨：グループ1の比率，グループ2の比率 統：有意水準，各群の標本サイズ	5の検出力計算版
7	**2群の比率の比較(非劣性)のためのサンプルサイズの計算**	臨：各群の比率，臨床的に意味のある差*1 統：有意水準，検出力	
8	**1群の平均値の信頼区間をある幅におさめるためのサンプルサイズの計算**	臨：想定する標準偏差，信頼区間の幅 統：信頼係数(confidence level)	信頼区間の幅に平均値は関係ない
9	**2群の平均値の比較のためのサンプルサイズの計算**	臨：2群間の平均値の差，2群共通の標準偏差 統：有意水準，検出力，サンプルサイズの比	
10	**2群の平均値の比較のための検出力の計算**	臨：2群間の平均値の差，2群共通の標準偏差 統：有意水準，各群の標本サイズ	9の検出力計算版
11	**2群の平均の比較(非劣性)のためのサンプルサイズの計算**	臨：平均の差，標準偏差，臨床的に意味のある差*1 統：有意水準，検出力	
12	**対応のある2群の平均値の比較のためのサンプルサイズの計算**	臨：2群間の平均値の差*2，2群共通の標準偏差*2 統：有意水準，検出力	
13	**対応のある平均値の比較のための検出力の計算**	臨：2群間の平均値の差*2，2群共通の標準偏差*2 統：有意水準，標本サイズ	12の検出力計算版
14	**2群の生存曲線の比較のためのサンプルサイズの計算**	臨：登録期間，試験期間*3，年次，各群の生存率 統：有意水準，検出力，サンプルサイズの比	
15	**2群の生存曲線の比較のための検出力の計算**	臨：登録期間，試験期間*3，年次，各群の生存率 統：有意水準，各群の標本サイズ	14の検出力計算版
16	**2群の生存曲線の比較(非劣性)のためのサンプルサイズの計算**	臨：登録期間，試験期間*3，年次，各群の生存率，臨床的に意味のある差*1 統：有意水準，検出力，サンプルサイズの比	

*1:非劣性マージンと呼ばれる．非劣性試験において許容されるアウトカムの範囲を表す．
*2:2 群間の平均値の差，2 群共通の標準偏差とあるが，対応のあるデータなので，正しくは，個々の被験者における差の平均値，差の標準偏差を意味する．
*3:試験期間とは，登録期間＋フォローアップ期間を表している．

主要評価項目が平均値の 2 標本(無作為化比較試験)の症例数設計は，(平均値の差)／(共通の標準偏差)に基づいて行われる[63]．そのため，対照治療・試験治療での期待される平均値および共通の標準偏差の情報が必要である．なお，平均値の場合には，帰無仮説 H_0「試験治療と対照治療の真の平均値の差は 0 である(試験治療と対照治療)」に対して症例数設計が行われる．

主要評価項目が生存曲線の 2 標本(無作為化比較試験)の症例数設計は，ハザード比に基づいて行われる．そのため，期待される試験治療／対照治療のハザード比の情報が必要である．あるいは，各治療の年次生存割合(1 年生存割合，3 年生存割合など)または中央生存期間(MST; Median Survival Time)からハザード比を計算することができる．生存曲線による症例数設計では，必要症例数ではなく，必要イベント数で与えられる．一方で，必要症例数は，必要イベント数に打ち切り(censoring)症例数を加えたものであるため，登録期間および追跡期間に基づいて必要症例数を計算する場合がある(1 標本の場合も同様である)．なお，生存曲線の場合には，帰無仮説 H_0「試験治療と対照治療の真のオッズ比は 1 である」に対して症例数設計が行われる．

[63] 量的データは，平均値の差が測度に依存するため，標準偏差で割っている．

6.2 EZR による症例数設計

表 6.2 は，EZR で実行可能な標本サイズの設計を表している．ここでは，幾つかのシチュエーションのもとで，症例設計の方法について述べる．

6.2.1 2 値アウトカムにおける必要症例数の計算

Senario1. 2 値アウトカムに対する単群試験での症例設計

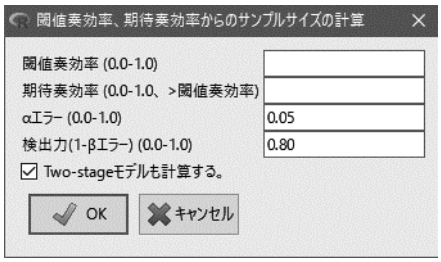
いま，ある癌における標準薬での奏効割合が 30%であることが，論文で報告されている．製薬企業 A が新たな抗癌剤を開発している．新薬では奏効割合が 50%になることを期待している．このことを，確認するための単群第 II 相試験を有意水準 α=0.05，検出力 1−β=80%での必要症例数を計算しなさい．

このとき，EZR での計算は以下のようになる．

2 値アウトカムに対する単群試験での症例設計(1)：Simon の方法

1: 「統計解析」→「必要サンプルサイズの計算」→「閾値奏効率，期待奏効率からのサンプルサイズの計算」を選択する．

2: 次のようなメニューが表示される．



このとき，

- 「閾値奏効率(0.0 – 1.0)」に「0.3」と入力する．
- 「期待奏効率(0.0 – 1.0)、>閾値奏効率」に「0.5」と入力する．
- 「α エラー(0.0 – 1.0)」に「0.05」と入力する．
- 「検出力(1-β エラー) (0.0 – 1.0)」に「0.8」と入力する．
- 「Two-stage モデルも計算する」にチェックを入れる

3: 「OK」ボタンを押す

このとき，2 つの出力が表示される．先ず，上側の青色の部分(ph2single(0.3, 0.5, 0.05, (1-0.80), nsoln=1 の下側)は，

```
   n  r Type I error Type II error
1 39 16   0.04998419      0.1683918
```

である．これは，必要症例数が 39 例であり，奏効例数が 16 例以下だった場合には，有効性が認められないことを意味する．一方で，抗癌剤の試験では，無効である治療を引き続いて実施することは，倫理的・医学的に認められないという観点から，2 段階デザインで実施されることがある．2 段階デザインでは，当該臨床試験において，中間解析を実施し，試験を継続しても，有効性が望めない場合には，早期無効中止を行うデザインである．下側の青の部分(ph2simon(0.3, 0.5, 0.05, (1-0.80), nmax=200)の下側)の出力

```
Simon 2-stage Phase II design

Unacceptable response rate:  0.3
Desirable response rate:  0.5
Error rates: alpha =  0.05 ; beta =  0.2

        r1 n1  r  n EN(p0) PET(p0)
Optimal  5 15 18 46  23.63  0.7216
Minimax  6 19 16 39  25.69  0.6655
```

は，Simon の 2 段階デザインでの結果である(抗癌剤の第 II 相試験のデザインとして良く用いられる)．2 段階デザインでは，n1 の例数が集積された時点で評価が行われ，奏効例数が r1 以下であれば，早期無効中止と判断される．

それ以外の場合には，n まで症例が集積される(すなわち，n が必要症例数である)．そして，奏効例数が r 以下だった場合には，有効性が認められないと判断される．

なお，Simon の 2 段階デザインには，Optimal デザインと Mini-Max デザインの 2 種類がある[64]． Optimal デザインに比べて，Mini-Max デザインのほうが，第 1 段階での症例数が多く，全体での症例数が少なくなる傾向にある．

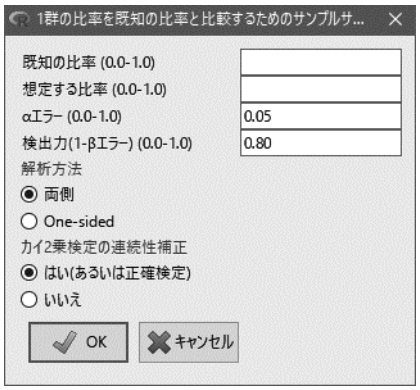
EZR における，「閾値奏効率，期待奏効率からのサンプルサイズの計算」は，Simon の単群デザインに基づいて計算されている．その考え方は，決定論的[65]に決められており，何らかの検定方法の裏付けがあるわけではない．そのため，両側対立仮説の設定が存在しない．また，中間解析におけるオーバーシュートが認められないため，それらの問題を緩和する方法として，SWOG の 2 段階デザインを用いることも多い[66]．

EZR では，Simon の方法とは別に，母比率の検定(1 群でのカイ 2 乗検定)に基づく標本サイズの決定方法がある．その場合の設定方法を以下に示す．

2 値アウトカムに対する単群試験での症例設計(2)：母比率の検定に基づく方法

1: 「統計解析」→「必要サンプルサイズの計算」→「1 群の比率を既知の比率と比較するためのサンプルサイズの計算」を選択する．

2: 次のようなメニューが表示される．



このとき，

- 「既知の比率(0.0 – 1.0)」に「0.3」と入力する．
- 「想定する比率(0.0 – 1.0)」に「0.5」と入力する．
- 「α エラー(0.0 – 1.0)」に「0.05」と入力する．
- 「検出力(1-β エラー) (0.0 – 1.0)」に「0.8」と入力する．
- 「解析方法」に「One-sided」を選択する．
- 「カイ 2 乗検定の連続性補正」において「はい(あるいは正確検定)」を選択する．

3: 「OK」ボタンを押す

ここで，「カイ 2 乗検定の連続性補正」とは，母比率の検定において，連続性の補正を行うか否かを選択するものであり，連続性を補正したほうが必要症例数が多くなる．

このときの結果を以下に示す．

	仮定
想定する比率	0.3
比較する比率	0.5
α エラー	0.05
	片側検定
検出力	0.8
	計算結果
必要サンプルサイズ	45

[64] 帰無仮説のもとでの期待症例数が最小になるように計算するのが Optimal デザインであり，最大の症例数を最小にするように計算するのが Mini-Max 法である．

[65] 「pick the winner rule」という．

[66] https://stattools.crab.org/Calculators/twoStage.htm

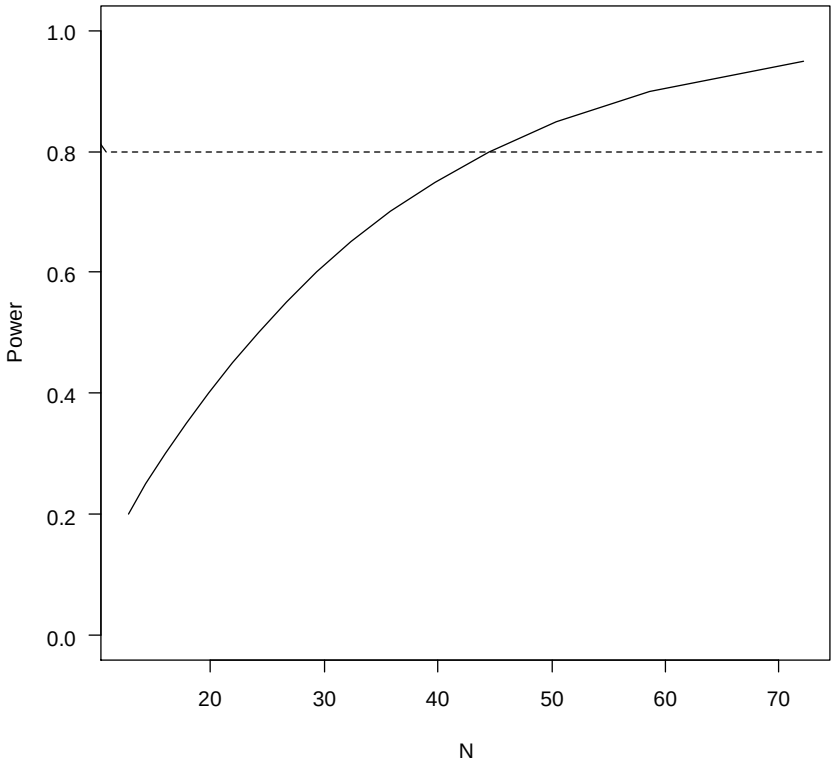


図 6.2：母比率の検定における検出力曲線

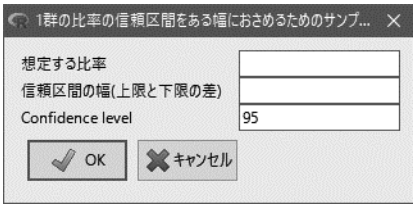
したがって，必要症例数は，45 例である．このとき，検出力と症例数のグラフ(図 6.2)が表示される．このグラフでは，X 軸に症例数 Y 軸に検出力が表示されている．そして，点線の平行線は，今回のデザインにおける検出力(0.8)を表している．Simon の Mini-Max デザインおよび 1 段階デザインでの必要症例数は 39 例であるが，この場合，母比率の検定における検出力は，80%未満になることがわかる．また，Optimal デザインでの必要最小例数は，46 例なので，検出力は 46 例なので検出力は 80%を上回るものの，19 例以上が positive study となる．一方で，この決定を母比率の検定に当てはめた場合には，連続性の補正を行った場合には，有意でない．実際には，不適格例を見込んだ症例数になるため，これらの方法の違いは少なくなるが，注意が必要である．

Senario2. 2 値アウトカムに対する観察研究での信頼区間に基づく症例設計

いま，ある難治疾患に対する治療成績に関する前向き観察研究を検討している．ここでのアウトカムには，治療成功／非成功の 2 値でとることを考えている．当該医療機関での治療成績から，60%の治療成功割合であることが分かっている．今回の前向き研究では，多施設で実施したいと考えており，信頼区間の幅(上側信頼限界－下側信頼限界)は，10%程度を想定している．必要症例数を計算しなさい．

このとき，EZR での計算は以下のようになる．

2 値アウトカムに対する信頼区間での症例設計
1:　「統計解析」→「必要サンプルサイズの計算」→「1 群の比率の信頼区間をある幅におさめるためのサンプルサイズの計算」を選択する．
2:　次のようなメニューが表示される．



このとき，
- 「想定する比率」に「0.6」と入力する．
- 「信頼区間の幅(上限と下限の差)」に「0.1」と入力する．
- 「Confidence level」に「95」と入力する．

3: 「OK」ボタンを押す

このとき，次のような出力が表示される．

```
                    仮定
想定する比率          0.6
信頼区間              0.1
Confidence level     0.95

                  計算結果
必要サンプルサイズ     369
```

この結果は，母比率に対する 95%信頼区間に基づいて計算したものである． したがって，必要症例数は 369 例である．このとき，信頼区間の幅に対する必要症例数のグラフが表示される(図 6.3)．信頼区間の幅を 0.1 未満にすると，非常に多くの症例数が必要になることが分かる．

Senario3. 2 値アウトカムに対する比較試験での症例設計

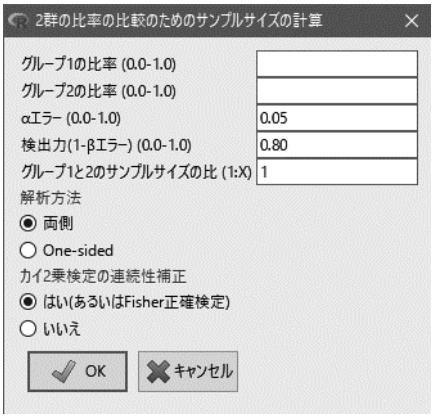
いま，ある疾患に対する治療法の無作為化比較第 II 相試験を検討している．これまでの治療法における治療成功割合は，50%であることが論文調査から明らかになっている．これを新規治療法では，60%の治療成功割合まで上昇できることを期待している．今回は，無作為化比較第 II 相試験なので，片側対立仮説(優越性)での検討を考えている．このとき，有意水準 α=0.10，検出力 1−β =80%での必要症例数を検討しなさい．

このとき，EZR での計算は以下のようになる．

2 値アウトカムに対する比較試験での症例設計

1: 「統計解析」→「必要サンプルサイズの計算」→「2 群の比率の比較のためのサンプルサイズの計算」を選択する．

2: 次のようなメニューが表示される．



このとき，
- 「グループ 1 の比率(0.0 – 1.0)」に「0.5」と入力する．
- 「グループ 2 の比率(0.0 – 1.0)」に「0.6」と入力する．

・「α エラー(0.0 – 1.0)」に「0.05」と入力する.
・「検出力(1-β エラー(0.0 – 1.0))」に「0.80」と入力する.
・「グループ 1 と 2 のサンプルサイズの比(1:X)」に「1」と入力する.
・「解析方法」で「One-sided」を選択する.
・「カイ 2 乗検定の連続性補正」で「はい(あるいは Fisher 正確検定)」を選択する.

3: 「OK」ボタンを押す

このとき, 次のような出力が表示される.

```
                                  仮定
P1                                 0.5
P2                                 0.6
α エラー                           0.05
                                片側検定
検出力                             0.8
N2 と N1 のサンプルサイズの比        1

必要サンプルサイズ                 計算結果
N1                                 325
N2                                 325
```

である. したがって, 必要症例数は 1 群あたり 325 例(全体で 650 例)である. これは, 連続補正を伴うカイ 2 乗検定(母比率の差の検定)に基づいて計算されている. このとき, 検出力に対する必要症例数のグラフも表示される(図 6.2 と同様のグラフであるが, 解釈は行わないため省略する).

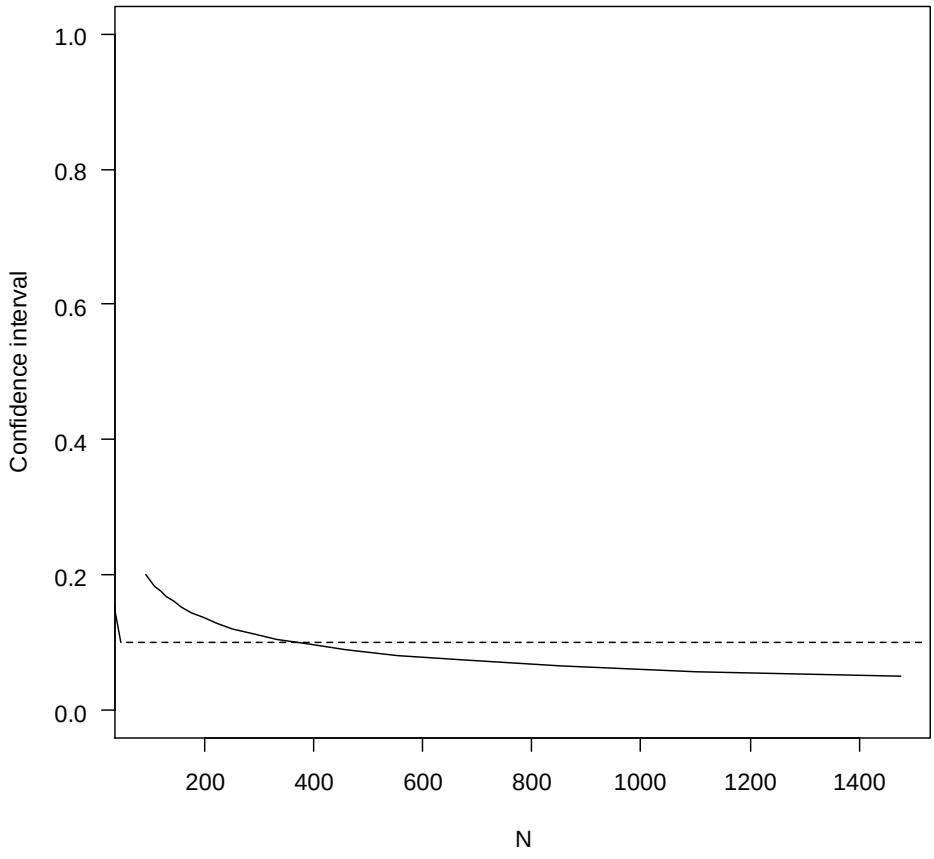


図 6.3:母比率の信頼区間における標本サイズと信頼区間の関係

Senario4. 2 値アウトカムに対する非劣性試験での症例設計

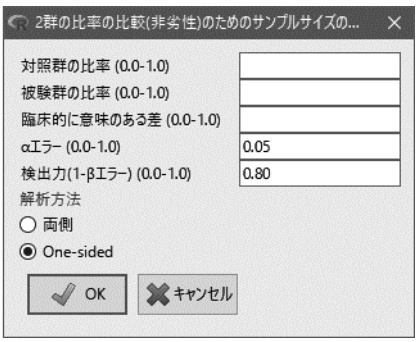
いま，新しい手術法による創感染症の発現割合に対する無作為化比較第 III 相試験を検討している．これまでの手術での創感染症発現割合は，7%であることがわかっている．新しい手術においても同程度の 7%であると期待しているものの，13%までであれば臨床的に創感染症発現割合が上昇していないと判断したいと考えている(非劣性試験)このとき，有意水準 α=0.025，検出力 1−β =80%での必要症例数を検討しなさい．

このとき，EZR での計算は以下のようになる．

2 値アウトカムに対する非劣性試験での症例設計

1: 「統計解析」→「必要サンプルサイズの計算」→「2 群の比率の比較(非劣性)のためのサンプルサイズの計算」を選択する．

2: 次のようなメニューが表示される．



このとき，

- 「対照の比率(0.0 – 1.0)」に「0.07」と入力する．
- 「被験者群の比率(0.0 – 1.0)」に「0.07」と入力する．
- 「臨床的に意味のある差(0.0 – 1.0)」に「0.05」と入力する．
- 「α エラー(0.0 – 1.0)」に「0.025」と入力する．
- 「検出力(1-β エラー(0.0 – 1.0))」に「0.80」と入力する．
- 「解析方法」で「One-sided」を選択する．

3: 「OK」ボタンを押す

このとき，次のような出力が表示される．

```
                        仮定
P1                      0.07
P2                      0.07
意味のある差            0.05
αエラー                 0.025
                    片側検定
検出力                   0.8

必要サンプルサイズ 計算結果
N1                       409
N2                       409
```

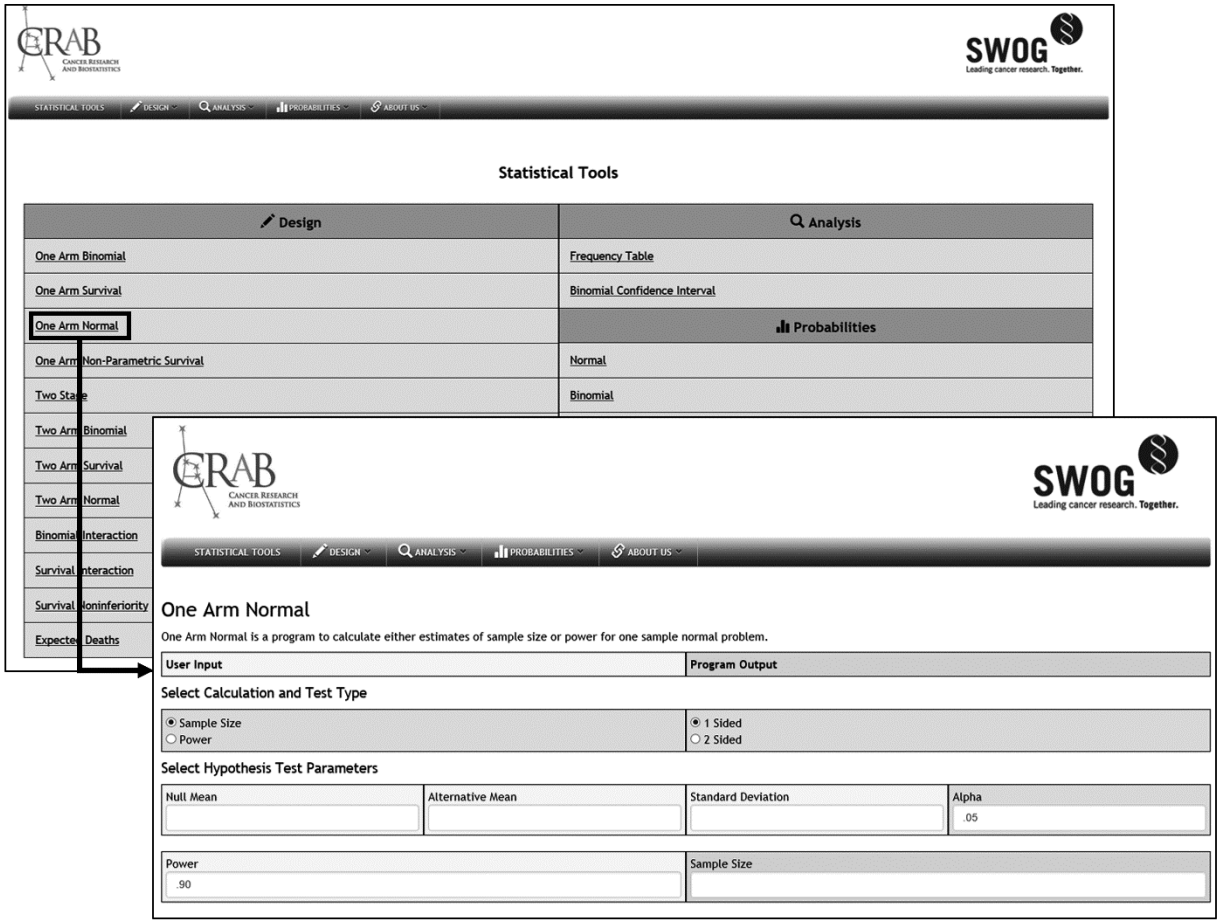
である． したがって，必要症例数は 1 群あたり 409 例(全体で 818 例)である．これは，母比率の差の検定に基づくハンディキャップ検定を用いて計算されている．このとき，検出力に対する必要症例数のグラフも表示される(図 6.2 と同様のグラフであるが，解釈は行わないため省略する)．

6.2.2 連続アウトカムにおける必要症例数の計算

Senario5. 連続アウトカムに対する単群試験での症例設計

いま，手術による心理的不安を軽減するためのカウンセリングを実施したいと考えている．これまでの調査では，平均 100，標準偏差 50 であることがわかっている．本カウンセリングによって平均 80 まで軽減することを期待している．片側対立仮説のもとで，有意水準 α=0.05，検出力 1−β =80%での必要症例数を計算しなさい．

EZR には，連続アウトカムに対する単群試験での症例設計を行うことができない．ここでは，SWOG(SouthWest Oncology Group)の Web サイト(https://stattools.crab.org/)の CRAB(Cancer Reseach And Bistatistics)のツール(One Arm Normal)を用いる．このときの Web の画面を以下に示す．



この Web サイトによる症例設計の方法を以下に示す．

連続アウトカムに対する単群試験での症例設計
1:　Web サイトの画面において ・「Select Calculation and Test Type」で「Sample Size」を選択する． ・「Select Calculation and Test Type」で「1 Sided」を選択する． ・「Select Hypothesis Test Parameters」の「Null Mean」に「100」と入力する． ・「Select Hypothesis Test Parameters」の「Alternativel Mean」に「80」と入力する． ・「Select Hypothesis Test Parameters」の「Standard Deviation」に「50」と入力する． ・「Power」に「0.80」と入力する． 2:　「Calculate」ボタンを押す

すると，「Sample Size」に「39」が表示される．すなわち，必要症例数は 39 症例である．なお，この検定は 1 標本 t 検定に基づいて計算されている．

Senario6. 連続アウトカムに対する観察研究での信頼区間に基づく症例設計

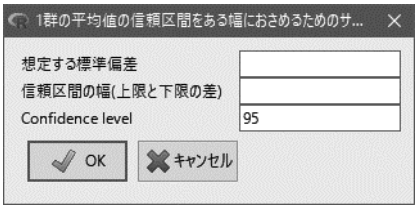
いま，ある地域における心臓病疾患に対する治療成績に関する前向き観察研究を検討している．ここでのアウトカムには，収縮期血圧を用いることにしている．当該地域の医療機関での治療成績から，集種期血圧の標準偏差が 50 であることが報告されている．今回の前向き研究では，多施設で実施したいと考えており，信頼区間の幅(上側信頼限界－下側信頼限界)は，20 程度を想定している．必要症例数を計算しなさい．

このとき，EZR での計算は以下のようになる．

連続アウトカムに対する単群試験での症例設計

1: 「統計解析」→「必要サンプルサイズの計算」→「1 群の平均値の信頼区間をある幅におさめるためのサンプルサイズの計算」を選択する．

2: 次のようなメニューが表示される．



このとき，

- 「想定する標準偏差」に「50」と入力する．
- 「信頼区間の幅」に「20」と入力する．
- 「Confidence level」に「95」と入力する．

3: 「OK」ボタンを押す

このとき，次のような出力が表示される．

```
                    仮定
標準偏差              50
信頼区間              20
Confidence level    0.95

                  計算結果
必要サンプルサイズ    97
```

この結果は，母平均に対する 95%信頼区間に基づいて計算したものである． したがって，必要症例数は 97 例である．このとき，信頼区間の幅に対する必要症例数のグラフが表示される(図 6.3 と同様の解釈になるので，ここでは割愛する)．

Senario7. 連続アウトカムに対する比較試験での症例設計

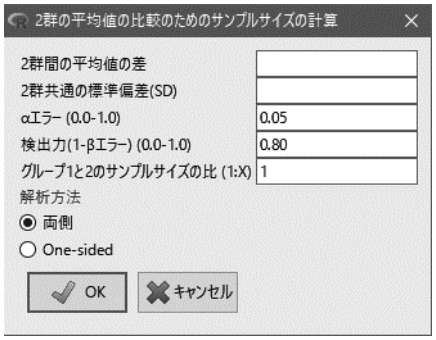
いま，薬剤による臨床検査値の軽減に対する比較試験を検討している．既存薬では，平均 30，標準偏差 40 の軽減効果が報告されている．新薬では，45 の軽減を期待している．両側対立仮説のもとで，有意水準 α=0.05，検出力 $1-\beta$=80%での必要症例数を計算しなさい．

本試験では，既存薬の平均は 30 であり，新薬では 45 なので，2 群間の平均値の差(新薬 – 既存薬)は，15 である．このとき，EZR での計算は以下のようになる．

連続アウトカムに対する比較試験での症例設計

1: 「統計解析」→「必要サンプルサイズの計算」→「2 群の平均値の比較のためのサンプルサイズの計算」を選択する．

2: 次のようなメニューが表示される．



このとき，

- 「2 群間の平均値の差」に「15」と入力する．
- 「2 群共通の標準偏差(SD)」に「0.07」と入力する．
- 「α エラー(0.0 – 1.0)」に「0.025」と入力する．
- 「検出力(1-β エラー(0.0 – 1.0))」に「0.80」と入力する．
- 「グループ 1 と 2 のサンプルサイズの比(1:X)」に「1」と入力する．
- 「解析方法」で「両側」を選択する．

3: 「OK」ボタンを押す

このとき，次のような出力が表示される．

```
仮定
2 群間の平均値の差                    15
標準偏差                              40
α エラー                            0.05
                                 両側検定
検出力                               0.8
N2 と N1 のサンプルサイズの比          1

必要サンプルサイズ                計算結果
N1                                   112
N2                                   112
```

である． したがって，必要症例数は 1 群あたり 112 例(全体で 224 例)である．これは，母平均の差の検定(2 標本 t 検定)に基づいて計算されている．このとき，検出力に対する必要症例数のグラフも表示される(図 6.2 と同様のグラフであるが，解釈は行わないため省略する)．

Senario8. 連続アウトカムに対する非劣性試験での症例設計

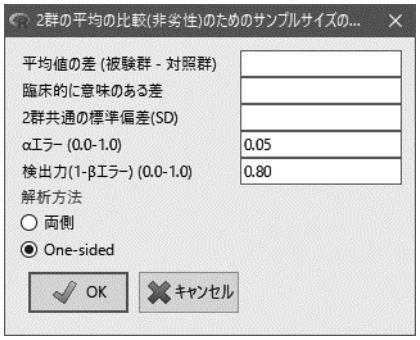
いま，薬剤による臨床検査値の軽減に対する非劣性試験を検討している．既存薬は，平均 30，標準偏差 10 の軽減効果が報告されている．一方で，副作用が少ないと考えられる新薬の効果は同程度であると期待されるが，90%程度の 27 までは許容されると考える．したがって，非劣性マージンは 3 である．このとき，有意水準 α=0.05，検出力 1−β =80%での必要症例数を計算しなさい．

ここで，既存薬の効果は 30 であり，新薬で期待される効果は同程度なので，平均の差は，0 である．このとき，EZR での計算は以下のようになる．

連続アウトカムに対する非劣性試験での症例設計

1: 「統計解析」→「必要サンプルサイズの計算」→「2 群の平均の比較(非劣性)のためのサンプルサイズの計算」を選択する．

2: 次のようなメニューが表示される．



このとき，

- 「平均値の差(被験群 － 対照群)」に「0」と入力する．
- 「臨床的に意味のある差」に「3」と入力する．
- 「共通の標準偏差(SD)」に「10」と入力する．
- 「αエラー(0.0 – 1.0)」に「0.05」と入力する．
- 「検出力(1-βエラー(0.0 – 1.0))」に「0.80」と入力する．
- 「解析方法」で「One-sided」を選択する．

3: 「OK」ボタンを押す

このとき，次のような出力が表示される．

```
                          仮定
2 群間の平均値の差          0
意味のある差                3
標準偏差                   10
αエラー                  0.05
                       片側検定
検出力                    0.8

必要サンプルサイズ 計算結果
N1                       138
N2                       138
```

である． したがって，必要症例数は 1 群あたり 138 例(全体で 276 例)である．これは，母平均の差の検定(2 標本 t 検定)に基づくハンディキャップ検定を用いて計算されている．このとき，検出力に対する必要症例数のグラフも表示される(図 6.2 と同様のグラフであるが，解釈は行わないため省略する)．

6.2.3 対応のある連続データに対する必要症例数の計算

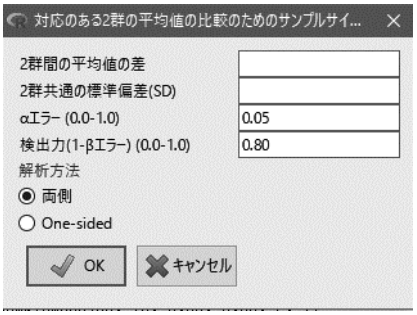
Senario9. 対応のある連続アウトカムでの症例設計

いま，手術による心理的不安を軽減するためのカウンセリングを実施したいと考えている．これまでの調査では，平均 100，標準偏差 30 であることがわかっている．そのため，カウンセリング前のストレス指標の平均を 100，カウンセリング後のストレス指標の平均 70，前後での標準偏差を 80 とするとき．両側対立仮説のもとで，有意水準 α =0.05，検出力 1−β =80%での必要症例数を計算しなさい．

本試験では，カウンセリング前の平均は 100 であり，カウンセリング後では 70 なので，2 群間の平均値の差(新薬 – 既存薬)は，30 である．このとき，EZR での計算は以下のようになる．

対応のある連続アウトカムでの症例設計

1: 「統計解析」→「必要サンプルサイズの計算」→「対応のある 2 群の平均値の比較のためのサンプルサイズの計算」を選択する．
2: 次のようなメニューが表示される．



このとき，

- 「2 群間の平均値の差」に「30」と入力する．
- 「2 群共通の標準偏差(SD)」に「80」と入力する．
- 「α エラー(0.0 – 1.0)」に「0.025」と入力する．
- 「検出力(1-β エラー(0.0 – 1.0))」に「0.80」と入力する．
- 「解析方法」で「両側」を選択する．

3:　「OK」ボタンを押す

このとき，次のような出力が表示される．

```
仮定
2 群間の平均値の差          30
標準偏差                    80
αエラー                   0.05
                     two.sided
検出力                     0.8

必要サンプルサイズ　計算結果
N                          58
```

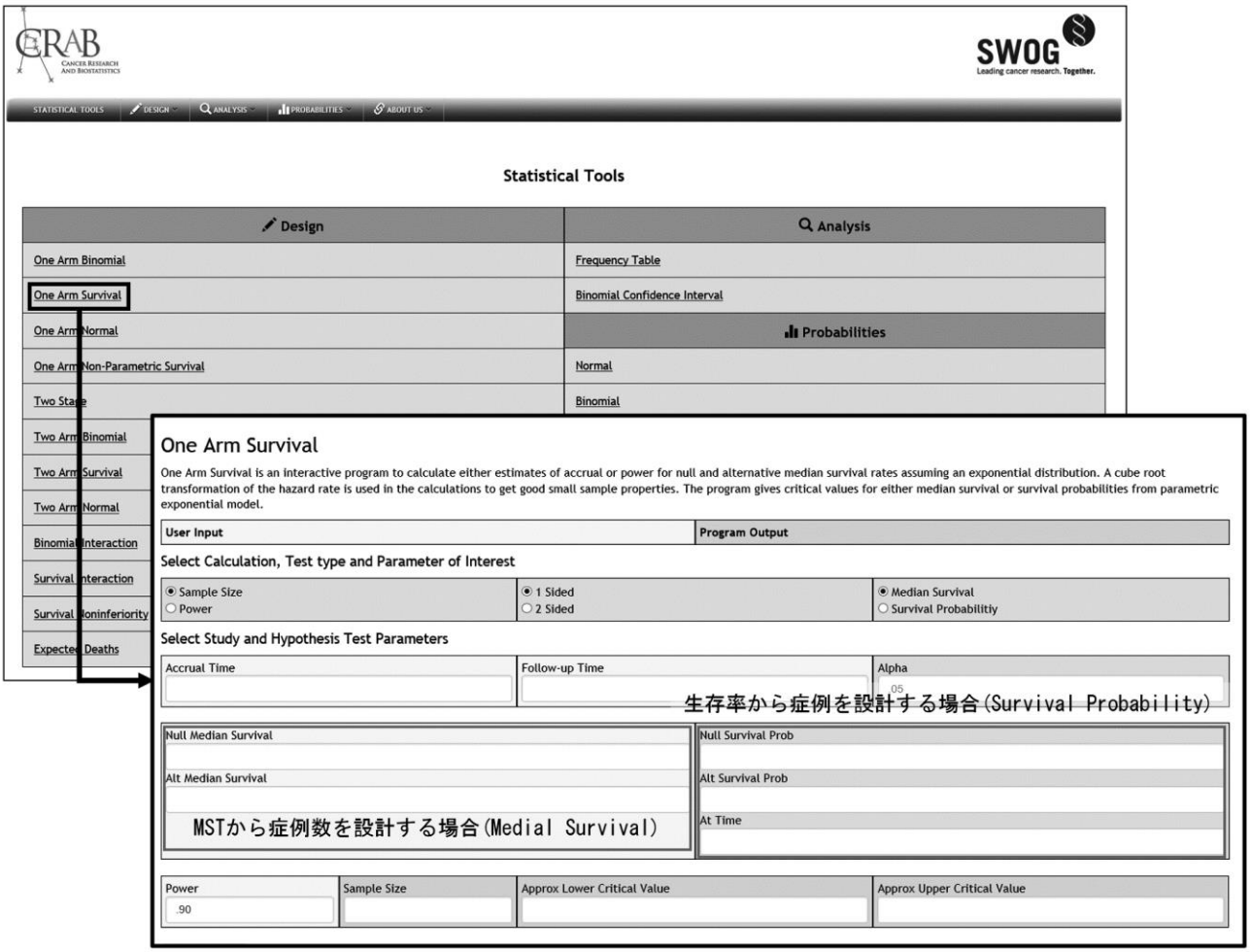
である．したがって，必要症例数は 58 例である．これは，対応のある t の検定に基づいて計算されている．このとき，検出力に対する必要症例数のグラフも表示される(図 6.2 と同様のグラフであるが，解釈は行わないため省略する)．

6.2.4 生存時間アウトカムにおける必要症例数の計算

Senario10. 生存曲線に対する単群試験での症例設計

いま，多発性骨髄腫に対する中央全生存期間(MST)が 46 ヵ月であることが報告されている．今回，新たな治療薬が開発され，MST が 60 ヵ月まで延長することが期待されている．登録期間 3 年，フォローアップ期間 5 年の片側対立仮説のもとで，有意水準 α=0.05，検出力 1−β=80%での必要症例数を計算しなさい．

EZR では，単群試験での症例設計を行うことができない．ここでは，SWOG(SouthWest Oncology Group)の Web サイト(https://stattools.crab.org/)の CRAB(Cancer Reseach And Bistatistics)のツール(One Arm Normal)を用いる．このときの Web の画面を以下に示す．



この Web サイトによる症例設計の方法を以下に示す.

生存曲線に対する単群試験での症例設計
1: Web サイトの画面において ・「Select Calculation, Test type and Parameter of Interest」で「Sample Size」を選択する. ・「Select Calculation, Test type and Parameter of Interest」で「1 Sided」を選択する. ・「Select Calculation, Test type and Parameter of Interest」で「Median Survival」を選択する. ・「Select Hypothesis Test Parameters」の「Accrual Time」に「36」(3 年)と入力する. ・「Select Hypothesis Test Parameters」の「Follow-up Time」に「60」(5 年)と入力する. ・「Null Median Survival」に「46」と入力する. ・「Alt Median Survival」に「60」と入力する. ・「Power」に「0.80」と入力する. 2: 「Calculate」ボタンを押す

すると,「Sample Size」に「138」が表示される. すなわち, 必要症例数は 138 症例である. なお, この検定は生存曲線に指数分布を想定したときの信頼区間に基づいて計算されている. なお, このときの信頼区間は,「Approx Upper Critical Value」に表示される. 今回の場合には,「54, 73」が想定される信頼区間幅になる.

Senario11. 生存曲線に対する比較試験での症例設計

いま，切除不能局所進行・再発胃癌患者における標準治療での MST が 10 ヵ月であることが報告されている．新たな抗癌剤＋標準療法の上乗せ効果によって，(標準療法)／(標準療法＋新規抗癌剤)のハザード比が 1.3 になることを期待している．登録期間 3 年，フォローアップ期間 2 年の両側対立仮説のもとで，有意水準 α=0.05，検出力 1−β=80%での必要症例数を計算しなさい．

EZR では，MST ではなく，年次生存割合(survival)で計算しなければならない．生存曲線が指数分布に従うとき，次の関係がある．

$$\text{ハザード} = \frac{-\log(\text{生存割合})}{\text{生存期間}}$$

ここで，log は自然対数である．MST は生存割合が 0.5 の生存期間なので，標準治療でのハザード比は

$$\text{標準治療のハザード} = \frac{-\log(\text{標準治療の生存割合})}{\text{標準治療の生存期間}} = \frac{-\log(0.5)}{10} = 0.06931$$

である[67]．したがって，標準治療の 1 年生存割合は，

$$\text{標準治療の1年生存割合} = \exp\left(-\text{標準治療のハザード} \times \text{標準治療の生存期間}\right) = \exp(-0.06931 \times 12) = 0.435$$

で与えられる[68]．すなわち，標準治療での 1 年生存割合は，43.5%である．

次いで，標準治療＋新規抗癌剤群(新規治療群)の 1 年生存率を計算する．(標準治療)／(新規治療群)のハザード比が 1.3 なので，新規治療のハザードは，

$$\text{新規治療のハザード} = \frac{\text{既存治療のハザード}}{\text{ハザード比}} = \frac{0.06931}{1.3} = 0.05332$$

なので，新規治療法の 1 年生存割合は，

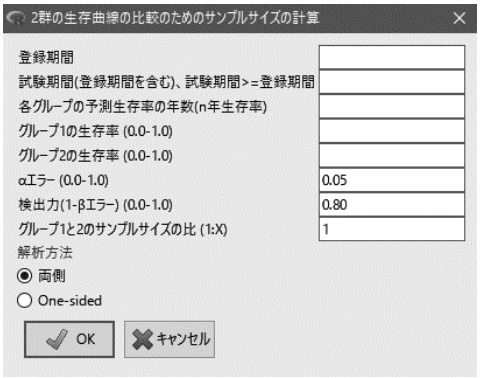
$$\text{新規治療の1年生存割合} = \exp\left(-\text{新規治療のハザード} \times \text{標準治療の生存期間}\right) = \exp(-0.05332 \times 12) = 0.527$$

である．すなわち，新規治療での 1 年生存割合は，52.7%である．

対応のある連続アウトカムでの症例設計

1: 「統計解析」→「必要サンプルサイズの計算」→「対応のある 2 群の平均値の比較のためのサンプルサイズの計算」を選択する．

2: 次のようなメニューが表示される．



[67] Excel では，「= −LN(0.5)/10」で計算できる．

[68] Excel では，「= exp(−0.06931*12)」で計算できる．

このとき，
- 「登録期間」に「3」と入力する.
- 「試験期間(登録期間を含む)，試験期間＞＝登録機関」に「5」と入力する.
- 「各グループの予測生存率の年数(n 年生存率)」に「1」と入力する.
- 「グループ 1 の生存率 (0.0 – 1.0)」に「0.435」と入力する.
- 「グループ 2 の生存率 (0.0 – 1.0)」に「0.527」と入力する.
- 「グループ 1 と 2 のサンプルサイズの比(1:X)」に「1」と入力する.
- 「解析方法」で「両側」を選択する.

3: 「OK」ボタンを押す

このとき，次のような出力が表示される.

	仮定
P1	0.435
P2	0.527
P1、P2 の観察期間	1
登録期間	3
全研究期間	5
αエラー	0.05
	両側検定
検出力	0.8
N2 と N1 のサンプルサイズの比	1
必要サンプルサイズ	計算結果
N1	256
N2	256

である． したがって，必要症例数は1群あたり 256 例である(合計 512 例). これは，ログランク検定に基づいて計算されている. このとき，検出力に対する必要症例数のグラフも表示される(図 6.2 と同様のグラフであるが，解釈は行わないため省略する).

Senario12. 生存曲線に対する非劣性試験での症例設計

いま，切除不能進行・再発膵癌に対する既存治療に対して，レジメンを変更することで，有害事象の発現を抑制できることが報告されている. 既存治療での 1 年生存率は 30%である. 非劣性試験を検討するとき，(新規レジメン)／(既存治療)の非劣性マージンは 1.2 とする. 登録期間 3 年，フォローアップ期間 2 年とするとき，有意水準 α=0.05，検出力 1−β=80%での必要症例数を計算しなさい.

非劣性マージンが，ハザード比で与えられていることから，生存期間に変更しなければならない. 先ほどの比較試験の場合と同様に計算する. 対照群でのハザードは，

$$\text{対照群のハザード} = \frac{-\log(\text{対照群の生存割合})}{\text{対照群の生存期間}} = \frac{-\log(0.3)}{1} = 1.20397$$

なので，非劣性マージン 1.2 でのハザードは，$1.20397 \times 1.2 = 1.444764$ である. つまり，生存割合は

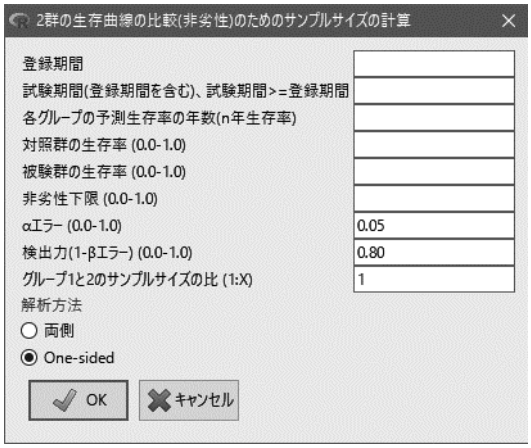
$$\text{非劣性マージンでの生存割合} = \exp(-1.4447 \times 1) = 0.236$$

である. よって，非劣性下限は，0.236 である.

生存曲線に対する非劣性試験での症例設計

1: 「統計解析」→「必要サンプルサイズの計算」→「2 群の平均の比較(非劣性)のためのサンプルサイズの計算」を選択する.

2: 次のようなメニューが表示される.



このとき,

- 「登録期間」に「3」と入力する.
- 「試験期間(登録期間を含む), 試験期間＞＝登録機関」に「5」と入力する.
- 「各グループの予測生存率の年数(n 年生存率)」に「1」と入力する.
- 「対照群の生存率 (0.0 – 1.0)」に「0.3」と入力する.
- 「試験群の生存率 (0.0 – 1.0)」に「0.3」と入力する.
- 「非劣性下限」に, 「0.236」と入力する.
- 「α エラー(0.0 – 1.0)」に「0.05」と入力する.
- 「検出力(1-β エラー(0.0 – 1.0))」に「0.80」と入力する.
- 「グループ 1 と 2 のサンプルサイズの比(1:X)」に「1」と入力する.
- 「解析方法」で「One-sided」を選択する.

3: 「OK」ボタンを押す

このとき, 次のような出力が表示される.

仮定	
P1	0.3
P2	0.3
非劣性下限 (0.0-1.0)	0.236
P1、P2 の観察期間	1
登録期間	3
全研究期間	5
α エラー	0.05
	片側検定
検出力	0.8
N2 と N1 のサンプルサイズの比	1
必要サンプルサイズ	計算結果
N1	384
N2	384

である. したがって, 必要症例数は 1 群あたり 386 例(全体で 768 例)である. これは, ログランク検定に基づくハンディキャップ検定を用いて計算されている. このとき, 検出力に対する必要症例数のグラフも表示される(図 6.2 と同様のグラフであるが, 解釈は行わないため省略する).