



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!**

著
新谷 歩
Ayumi SHINTANI

講談社

はじめに

本書は、統計の専門家や医学分野の方だけでなく、すべての分野の皆さんに実践的なデータ解析をマスターしてほしいという願いを込めて、①統計コンセプトを数式を用いないで直観的にわかるように、②お金がなくとも質が担保された解析を行えるように、基本的な統計コンセプトから無料統計ソフト EZR を使った実践までをできるだけわかりやすくという点にフォーカスしてまとめた一冊です。

本書を手にとってくださった方の中には、「日常診療の結果を知りたい」「とにかくデータをとって何が起きているか知りたい」「学位のために研究結果を国際的な論文に発表しなければならない」「何をしたいかまだわからないけど、データ解析をマスターすると将来のためになりそう」など、いろいろな方がいらっしゃると思います。それらすべての人々ができるだけ簡単に、正しいデータ解析を行えるよう考えついたのが、無料ソフトにこだわった統計教育です。

本書で紹介している統計の基礎理論や EZR を用いた解析は、医療分野のみでなく、社会学やビジネスといったありとあらゆる分野で使うことができます。医学に携わっている方はもちろん、それ以外の方々でも、「みんなの統計学」として多くの方々を読んでいただければ、大変嬉しく思います。

本書の作成にあたり、みんなが読める入門書の執筆を進めて

くださり、細部にわたりいろいろなアドバイスをくださりご尽力いただいた、講談社サイエンティフィックの三浦洋一郎さん、おかげさまで大変満足の一冊ができました。大阪大学大学院臨床統計疫学寄附講座の山本紘司先生、修士学生の加葉田大志朗さん、基礎配属生のパニット・パトラプールさん、久野彩さん、最終内容のチェック大変助かりました。石橋恵子さん、EZRのスクリーンショットのチェックありがとうございました。そして、統計に関するさまざまな思い、コメント、質問をいただいた、日本中で出会った多くの研究者の方々、本当にありがとうございました。本書は多くの皆様からいただいた質問への答えができるだけ含まれるよう、書かせていただきました。最後に、お盆、お正月と休みを返上した本書の執筆でしたが、支えてくれた二人の娘と主人に心から感謝します。

2016年2月

新谷 歩

CONTENTS

はじめに.....	iii
EZRをインストールしよう.....	1
EZRのインストール 1つ目の方法.....	1
EZRのインストール 2つ目の方法.....	3
データセットを読み込む.....	8
R Console上で解析を行う.....	10
1日目 記述統計量.....	11
平均値と中央値.....	11
平均がデータの中心を表すとはかぎらない.....	12
標準偏差でばらつきがわかる.....	15
カテゴリ変数の記述の仕方.....	18
EZRを用いた記述統計量の計算.....	19
2日目 仮説検定.....	23
コイン投げゲーム——イカサマか偶然か？.....	23
5回連続で表はさすがにイカサマだろう.....	24
P値はまぐれあたりの確率.....	25
2つのダイエット法——P値の意味とは？.....	26
「統計的な差」と「臨床的に意味がある差」を混同しない.....	28
P値と信頼区間.....	29
真の値は絶対的.....	31
標準誤差.....	32
ハザード比の信頼区間.....	33
3日目 疫学研究のデザイン.....	35
統計的有意差≠因果関係.....	35
まずはPECOをおさえよ.....	38

疫学研究のデザイン分類	38
分類1 介入研究と観察研究	40
分類2 コホート研究とケースコントロール研究	41
コホート研究	41
ケースコントロール研究	43
必ずしも「ケース＝疾患あり」ではない	44
分類3 前向き研究, 後ろ向き研究, 横断研究	45
前向き研究	45
後ろ向き研究	45
前向きと後ろ向きのケースコントロール研究	47
横断研究	47
さあ分類してみよう	48
研究デザインの長所・短所	51

4日目 統計テストの選び方 53

ピアソンかスピアマンか, どちら?	53
正しい統計テストを選ぶための7つのQ	54
Q1 ランダム化研究か否か?	56
Q2 差を見るのか, 相関を見るのか?	57
Q3 比べたいデータ間の対応	58
Q4 連続変数アウトカムの種類	59
Q5 アウトカム(連続変数の場合)の正規性	60
Q6 比較群の数	62
Q7 症例数	63
Quiz	65

5日目 スチューデントのT検定, マンホイットニーのU検定 67

EZRを用いたスチューデントのT検定	69
スチューデントのT検定の仮定 (assumption) の確認	72
正規分布に従わなくても使えるマンホイットニーのU検定	74
EZRを用いたマンホイットニーのU検定	75

6日目	対応のあるT検定と ウィルコクソンの符号付順位和検定	77
	対応のあるT検定	77
	対応のあるT検定は差の平均を見る	78
	EZRで対応のあるT検定を行う	79
	ウィルコクソンの符号付順位和検定	81
	EZRでウィルコクソンの符号付順位和検定を行う	82
7日目	分散分析, クラスカルワリス検定, フリードマン検定	84
	多重性の問題と分散分析	84
	EZRを用いた分散分析	85
	多重性の補正: ボンフェローニ法	87
	分散分析の仮定(正規性)の確認	89
	EZRを用いた分散分析の等分散性の確認	90
	EZRを用いたクラスカルワリス検定	92
	EZRを用いたフリードマン検定	93
	相関が強いほど多重検定によるP値の補正をしなくてよい	97
	相関あり・なしのときの確率を計算する	98
	1回目と2回目のジャンプの成否がそれぞれ無関係(相関がない)な 場合	98
	1回目と2回目のジャンプの成否が無関係でない(相関がある) 場合	99
8日目	線形回帰と相関係数	102
	単回帰分析と重回帰分析	102
	線形回帰モデルによる最小二乗直線	103
	なぜ2乗するのか?	105
	最小二乗直線は残差の総和が最小	106
	最小二乗直線の意味	107
	傾きの大小が関連の強さを表すわけではない	108
	標準化の方法	109
	ピアソンの相関係数	110
	相関係数をグラフで見してみる	111

EZRを用いた線形回帰モデル	112
EZRを用いたピアソンの相関	113
EZRを用いたスピアマンの相関	115
線形回帰モデルの仮定の確認	116
Normal Q-Qプロットの作成	116
対数変換の方法	119
2乗変換の方法	121

9日目 リスク比, レート比, オッズ比とロジスティック回帰..... 122

リスク比の求め方	122
EZRを用いたリスク比の計算	124
P値の計算	124
期待値の計算	126
EZRを用いた期待値の計算	127
リスク計算では時間を考慮しよう	129
レートの計算法	131
リスク比とレート比ではどちらが正しい?	132
EZRを用いたレート比の計算	134
リスク比とリスク差	136
インフルエンザ予防接種の本当の効果は?	137
リスク差とNNT(治療必要数)	139
EZRを用いたリスク差とNNTの計算	140
NNTの信頼区間は, リスク差に有意差があるかどうか重要	141
リスク比とオッズ比	142
オッズとは?	142
単勝オッズの意味	144
オッズ比をリスク比として解釈しない	145
オッズ比はリスク比より1から遠ざかる	146
なぜオッズ比を使うのか?	148
リスクには時間の前後関係が必要	150
暴露割合を見る	152
コホート研究を行わなくてもリスク比が予想できる...?	152
オッズ比がリスク比より優れている点	154

EZRを用いたオッズ比の計算	156
補足：カイ2乗検定またはフィッシャーの正確検定のP値を データセットから直接計算する	158

10日目 感度・特異度・ROC図 159

一致割合とカッパの相関：あなたも気象予報士になれる？	160
カッパの相関係数	161
EZRを用いたカッパの相関係数の計算	163
感度・特異度	164
検査後有病率の計算	167
検査前の有病率をどうするか？	169
EZRを用いた各診断指標の計算	171
EZRを用いた検査後有病率の計算	173
ROCカーブ：検査の陽性・陰性を分けるカット値の決め方	173
EZRを用いたROCカーブの作成	177
曲線下面積(AUC)	179
多変量ROCカーブ	179
EZRを用いて多変量ROCカーブを描く	181
COLUMN 診断ツール開発に用いる統計	186

11日目 生存率解析： Kaplan-Meier図、ハザード比とコックス回帰 188

リスク評価では、イベントが起こるまでの時間も考慮する	189
生存率解析	190
累積イベント率	191
real worldでは抜け落ちがいっぱい！	192
累積イベント率の考え方	193
Kaplan-Meier法による累積イベント率の考え方	194
カレンダー年の表記は用いない	195
累積イベント率の計算方法	196
EZRを用いたKaplan-Meier図の作成	200
2つのKaplan-Meier図を比べる	202
Kaplan-Meier図の追跡期間を修正する	204
追跡時間を短くして図を作りかえる	205

カプランマイヤー図に対応するハザード比を計算する	209
コックスの比例ハザードモデル	209
カプランマイヤー図の本当の意味	211
患者によってイベント発生率が異なることに注意	212
競合リスクの解析	213
競合リスク法での累積イベント率の計算	214
カプランマイヤー法と競合リスク法, どちらを使うか?	217
EZRを用いた競合リスクの解析	218
くり返し起こるイベントや時間で変わる暴露の解析	222
データのクラスター	223
時間依存性コックス比例ハザードモデルをR Consoleで行う	223

12日目 研究に必要な症例数を計算しよう.....225

一度決めた症例数をむやみに変えてはならない	226
症例数を多くするのは有効だけど...	226
最低限必要な症例数	227
症例数の計算: サンプル問題を解いてみよう	228
2ステージアダプティブデザイン	240
索引	243

記述統計量

本章のキーワード

平均値

標準偏差と正規分布

分布と偏り

ばらつき

中央値

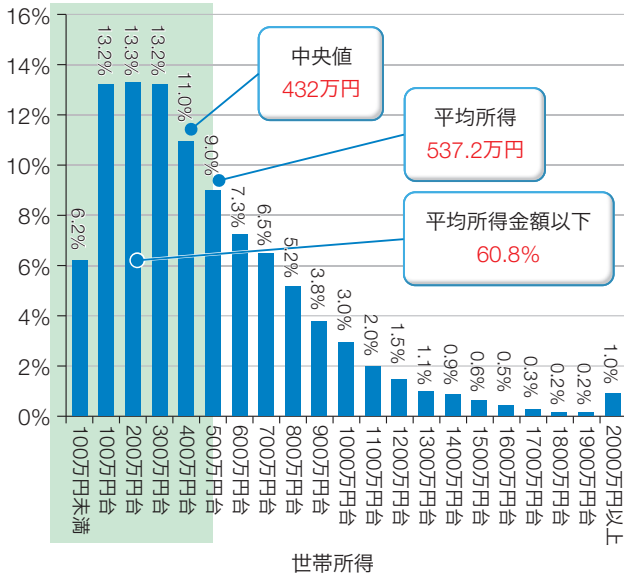
四分位点

平均値と中央値

次ページの図は、日本の1世帯あたりの所得分布を表しています。日本の世帯全体を100%とすると、所得金額が100万円未満の世帯が全体の6.2%、100万円台は13.2%、200万円台が13.3%…といった具合に、このグラフ1つで日本の全世帯の所得の様子がよくわかりますね。このグラフによると、1世帯あたりの所得金額の中央値は432万円で、平均は537.2万円だそうです。でもちょっと待ってください、平均以下の世帯が全体の60.8%って、どういうことでしょうか？

平均の計算法は、皆さんご存知の通り「すべての世帯所得を足して、世帯数で割る」という方法です。一方、**中央値432万円は所得が低い世帯から最も高い世帯まで並べてちょうど真ん中の世帯の所得、つまり半数（50%）の世帯は収入432万円以下ということがわかります。**日本の世帯年収の情報を1つの数字でまとめた場合、このグラフで用いられた①平均値、②中央値、③平均値以下の割合、の3つのうち、どれを使うのが最もよいのでしょうか？

所得金額階級別世帯数の相対度数分布
(2012年分・2013年調査, 政府統計より)



平均がデータの中心を表すとはかぎらない

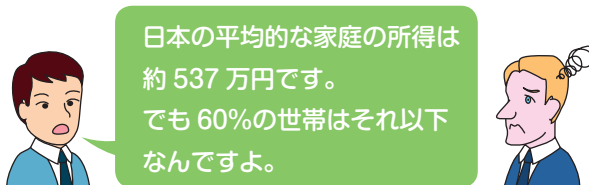
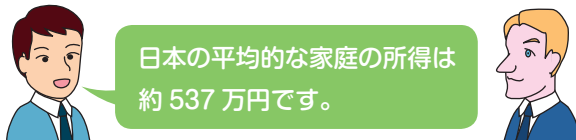
平均値は、非常に大きな値（ここでは高所得世帯）に引っ張られるという傾向があります。半数の世帯は所得が432万円以下ですが、残りの半数の世帯所得にはかなり差があり、所得が1億円を超えた人の数が約1万2千人（0.2%）もいたそうです。たとえば年間所得が100万円、200万円、300万円、400万円の世帯がそれぞれ1世帯ずつ、残りの1世帯の所得が1億円とすると、世帯所得の平均は

$$(100万 + 200万 + 300万 + 400万 + 1億) \div 5 = 2200万円$$

となります。中央値の300万円を1900万円も上回ってしまっています。所得の高い世帯にだいぶ引っ張られていることがわかりますね。



外国人の方にあなたが日本人の一般的な家庭の所得を説明するときに、平均の 537 万円か中央値の 432 万円か、どちらを選ぶでしょうか？



平均は、ここに挙げた例の日本の所得のように、片方に偏っていたり外れ値（他の数値より極端に離れた数値）がある場合は、データの中心を表す指標としてはあまり役に立たないどころか、かえって混乱を招いてしまいます。それでは、次のような説明はどうでしょうか。



日本の一般的な家庭の所得は
約 430 万円ですが、少なくとも
約 4 分の 1 の人は 200 万円以下
で、700 万円以上の世帯も
約 4 分の 1 いるのです。



ちょうど真ん中の値の中央値と、最初から 4 分の 1、最後から 4 分の 1 の**四分位点**を使ってデータを記述しました。どうでしょうか？ わかりやすくなったのではないのでしょうか。

次に、医学論文でデータを記述する典型的な方法を見てみましょう。

		コントロール群 (N = 60)	新薬群 (N = 120)
年齢	平均 (標準偏差)	50 (12)	53 (11)
性別	(男性 : 女性)	16 : 44	12 : 108
病歴 (年)	平均 (標準偏差)	9 (7)	9 (8)

この表は、ある薬剤開発時に行われたランダム化臨床試験の患者背景を示しています。**ランダム化 (無作為化) 臨床試験**とは、治療の効果などを調べるときに、コインの裏が出たら「治療あり」、表が出たら「治療なし」などという具合に、誰が治療を受けるか受けないかをまったくランダム (無作為) に決める (割り当てる) ことによって、**治療を受けるグループとそうでないグループの間に偏りがないようにする研究のこと**です。また、コントロール群とはこの場合「治療なし」に割り当てられたグループを指します。N は number の略で、グループの人数を表しています。

新薬の効果を調べるときに、新薬をとらなかった人がとった人に比べてかなり高齢で重篤な場合、新薬をとった人の生存率が結果的に良くても、それは薬が効いたからなのか最初から新薬をとった人の状態

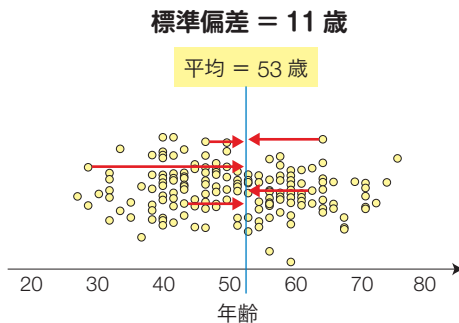
が良かったからなのか、わかりません。ランダム化は比較をしやすくするために、同じような人たちが治療あり群、治療なし群に入るように行われます。

標準偏差でばらつきがわかる

コントロール群に 60 人、新薬群に 120 人割り付けされ、平均年齢はコントロール群で 50 歳、新薬群では 53 歳です。でも、あれっ？ 53 の横の (11) って何でしょうか。これは標準偏差です。英語で Standard Deviation (SD) と表し、データのばらつきを示す指標です。それぞれの患者の年齢から平均年齢 53 歳までの差の平均で計算します。たとえば 36 歳から 53 歳までの差は 17 歳、53 歳から 70 歳までの差も 17 歳です。つまり、標準偏差が小さければ平均からの距離が小さいのでばらつきの小さいデータ、標準偏差が大きければ平均からの距離が大きいのでばらつきの大きいデータ、と解釈できます。

標準偏差 (SD : Standard Deviation)

個々の点から平均値までの平均的距離を表し、データのばらつきを示す。



データ（この場合は年齢）が正規分布に従う場合に、次のことがいえます。

- ① 「平均 \pm 2 \times 標準偏差」で示される範囲内に 95% のデータ値（年齢）が存在する。
- ② 「平均 \pm 1 \times 標準偏差」で示される範囲内に 67% のデータ値（年齢）が存在する。

標準偏差はデータの記述をするときに用いられ、よく使われるのは①の 95% の方です。この例では、「 $53 \pm 2 \times 11 = (31, 75)$ 」の範囲内に 95%（大多数）のデータ値（年齢）が存在すると解釈できます。

ここで正規分布とは、データをそれぞれの値の頻度で示したヒストグラムを見たときに、平均値の周辺の値をとる人の数が一番多く、平均値から離れるほど頻度が左右対象に減っていく「釣り鐘型」をとる分布のことをいいます。

標準偏差（SD）活用法

データが正規分布に従う場合：

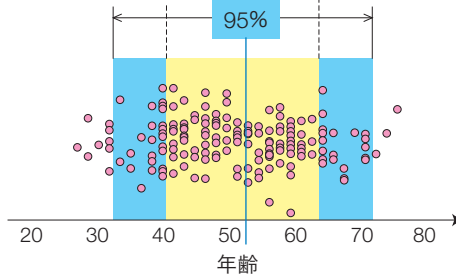
68% の患者の年齢が平均値 $\pm 1SD$ の間にある。

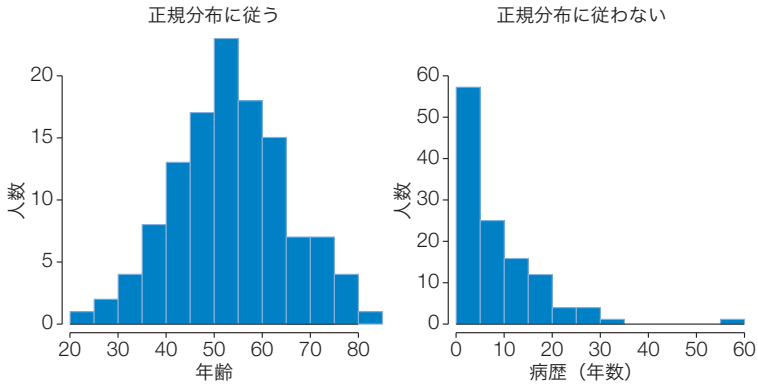
$$53 \pm 11 = (42, 64)$$



95% の患者の年齢が平均値 $\pm 2SD$ の間にある。

$$53 \pm 2 \times 11 = (31, 75)$$





「平均 \pm 2SD」の計算法を病歴に当てはめてみましょう。
 新薬群に割り付けられた人の平均病歴は9年、標準偏差が8年です。

$$9 \pm 2 \times 8 = (-7\text{年}, 25\text{年})$$

この研究に参加して新薬に割り付けられた120人の患者のうち、95%の病歴がマイナス7年からプラス25年？ どこかおかしいですね。

この研究に参加して新薬に割り付けられた
 120人の患者さんの95%の方の病歴は
 マイナス7年からプラス25年です。



どこで間違ったのでしょうか。左の病歴データをよく見てみましょう。
 病歴データはマイナスの値はとれないので、ゼロが最低値です。しかも、発症してすぐという人が少なくありません。ですから、ゼロ年に

人数が集中する一方、なんと病歴58年くらいの人もいるという、かなり歪んだ分布になっているのです。このデータも日本の世帯所得の例と同じく、注意が必要なデータです。

このように歪んだデータの場合は、平均や標準偏差といった指標を用いるのは好ましくありません。この場合も中央値と四分位点を使えば、かなりわかりやすくなるはずです。このデータでは、新薬に割り付けられた患者の病歴の、最初から4分の1、最後から4分の1の四分位点はそれぞれ2年と12年だったので、以下のように説明するとわかりやすいですね。

この研究で新薬に割り付けられた120人の患者さんの病歴は真ん中が6年、少なくとも4分の1は2年以下、または12年以上です。



カテゴリー変数の記述の仕方

		コントロール群 (N = 60)	新薬群 (N = 120)
年齢	平均(標準偏差)	50(12)	53(11)
性別	(男性:女性)	16:44	12:108
病歴(年)	平均(標準偏差)	9(7)	9(8)

最後に、性別のデータの表し方を見てみましょう。上の表を見てください。コントロール群の男性は16人、新薬群では12人でそれほど変わらないな、と思った人は要注意です。コントロール群では60人中で男性が16人ですから、割合は27%です。一方、新薬群では120

人のうちの 12 人ですから、男性の割合は 10%です。ランダムに割り付けたにも関わらず、27%と 10%ではかなり差が出てしまいました。

このように、研究に参加した人数に群間で差がある場合は、男性の人数を記載するだけでは比較が非常にしづらいので、注目すべき数字は割合 (%) の方なのです。多くの論文には人数を先に書いて割合は括弧の中 (%) で示しているものが多いのですが、%を前に出して人数を括弧の中に入れるようにしてみてもはどうでしょうか。私のおすすめの患者背景表は以下のようになります。

		コントロール群 (N = 60)	新薬群 (N = 120)
年齢	中央値 [四分位点]	49 [40, 60]	50 [38, 59]
男性	% (N)	27% (16)	10% (12)
病歴 (年)	中央値 [四分位点]	6 [3, 13]	6 [2, 12]

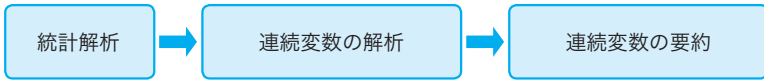
いかがでしょうか。ずっとわかりやすくなったと思いませんか？

次は EZR を用いた記述統計量の計算方法について説明します。

EZRを用いた記述統計量の計算

2 型糖尿病のランダム化臨床試験データ、【DMRCT180.csv】を使って、介入・コントロール群それぞれで年齢の平均、標準偏差、中央値、四分位点を計算してみましょう。まずは、p.8 の URL よりデータセットをダウンロードしてください。ダウンロードできたら、データセットを EZR に読み込んでください（読み込み方は、p.8 ~ 9 を参照）。

ここではアウトカム（調べたい項目のこと）が年齢、つまり連続変数なので、EZR のメニューバーの「統計解析」の下の「連続変数の解析」へカーソルを動かし、「連続変数の要約」を選択します。



① 年齢の平均を計算するので、年齢 (age) を選択

② 介入・コントロールを表す層別にする変数を選択し、OKをクリック

R コマンドーの「出力」欄に結果が表示される

拡大

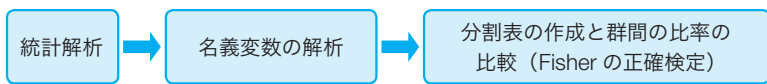
EZR の計算結果

	平均	標準偏差	0%	25%	50%	75%	100%	data:n
コントロール	57	11	21	50	57	64	79	83
介入	54	13	20	46	54	62	87	89

四分位点 (0%, 25%, 50%, 75%, 100%)

中央値 (50%)

次に男性の割合計算を行います。男性または女性の性別の変数はカテゴリーで表記され、これは名義変数と呼ばれています。解析は名義変数の解析、分割表の作成と群間の比率の比較を選択します。割り付け群を表す“arm”の変数と性別を表す“pat_gend”の変数は、どちらが行でも列でもかまいませんが、今回は“arm”を列に入れました。「パーセントの計算」は、群間を表す変数(“arm”)を列に入れたときは列を、行に入れるときは行を選択します。



分割表の作成と群間の比率の比較(Fisherの正確検定)

↓複数の選択はCtrlキーを押しながらクリック。

↓ 行の選択 (1つ以上選択)

↓ 列の変数 (1つ選択)

1 背景になる因子を表す変数を入れる

2 もう一方の群を表す変数を入れる

3 群を表す変数が列に入った時に「列のパーセント」を選択

4 症例数が 40 以上の場合選択。p.63 参照

5 症例数が 20 以上は「No」を選択

6 症例数が 20 以下は選択。20~40 の時は p.63 参照

パーセントの計算

- 行のパーセント
- 列のパーセント
- 総計のパーセント
- パーセント表示無し

仮説検定

- カイ2乗検定
- カイ2乗統計量の要素
- 期待度数の表示
- フィッシャーの正確検定

カイ2乗検定の連続性補正

- Yes
- No

↓2組ずつの比較(post-hoc検定)は比較する群が1つの場合のみ実施される。

- 2組ずつの比較(Bonferroniの多重比較)
- 2組ずつの比較(Holmの多重比較)

↓一部のサンプルだけを解析対象にする場合の条件式。例: age>50 & Sex==0 や age<50 | Sex==1

<全ての有効なケース>

ヘルプ リセット

EZR の計算結果

pat_gend	コントロール	介入
女性	43	37
男性	57	63
Total	100	100
Count	89	91

	arm = コントロール	arm = 介入
pat_gend = 女性	38	34
pat_gend = 男性	51	57

EZR での計算により、介入群には男性が 57 人で、それは介入群全体の 63% だったことが示されました。

本書において、EZ R の計算結果は上記のようにページの枠で示します。