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。

本書を手に取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こっているか知りたい」「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」として多くの方々に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クの三浦洋一郎さん，おかげさまで大変満足のいく一冊ができました。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

1日目

記述統計量

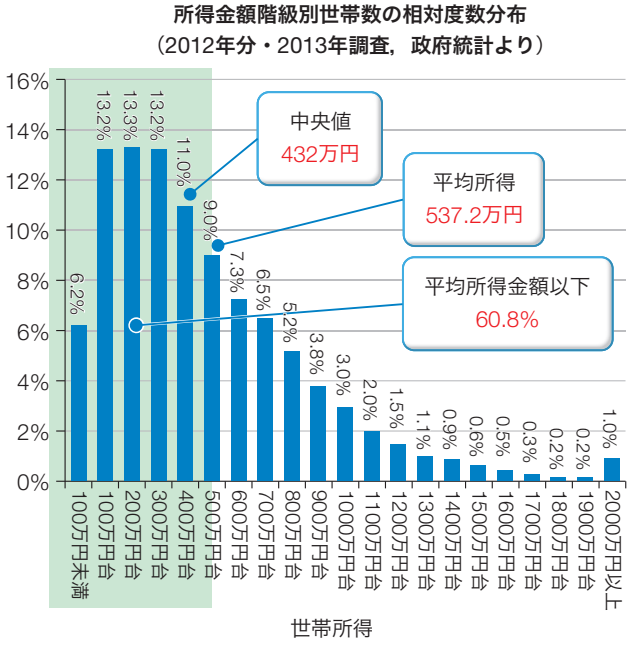
本章のキーワード

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でしょうか？



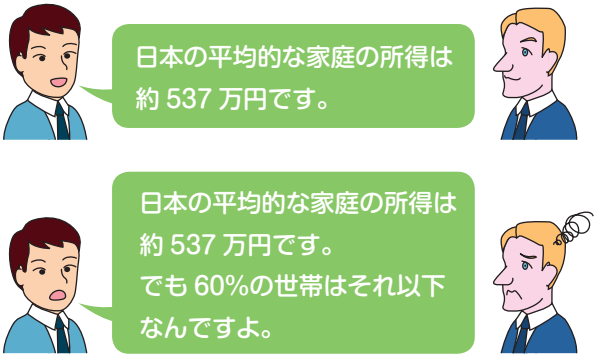
平均がデータの中心を表すとはかぎらない

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

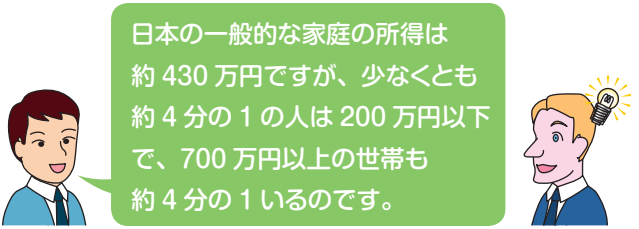
（100万 ＋ 200万 ＋ 300万 ＋ 400万 ＋ 1億）÷ 5 ＝ 2200万円

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。



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

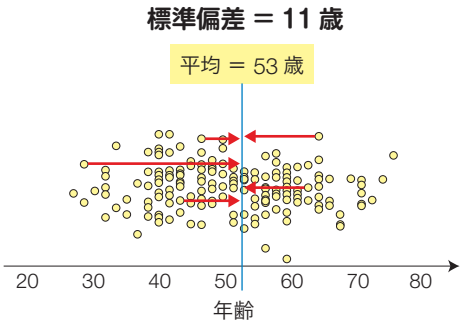
が良かったからなのか，わかりません。ランダム化は比較をしやすくするために，同じような人たちが治療あり群，治療なし群に入るように行われます。

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。

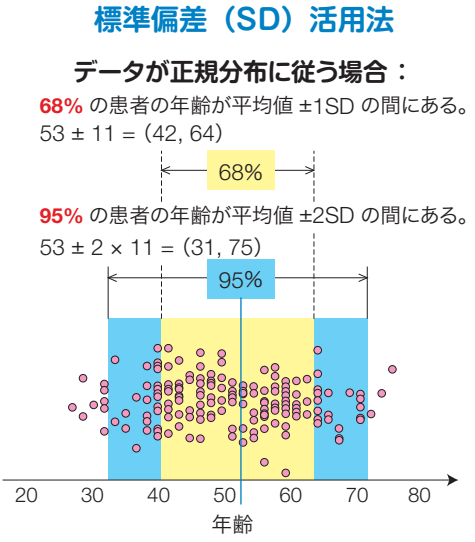


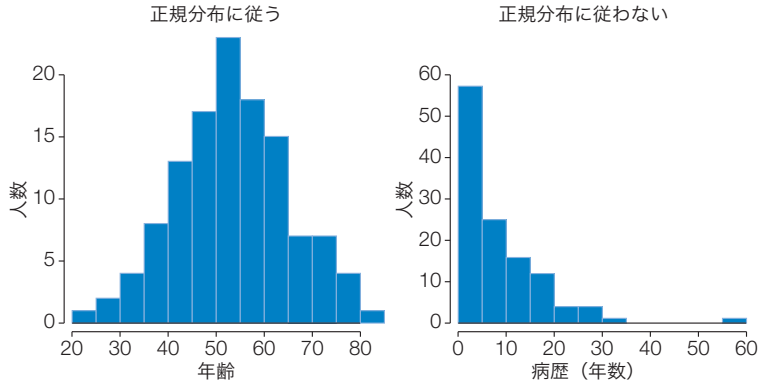
データ（この場合は年齢）が正規分布に従う場合に，次のことがいえます。

① 「平均± 2 ×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± 1 ×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 ± 2 × 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。



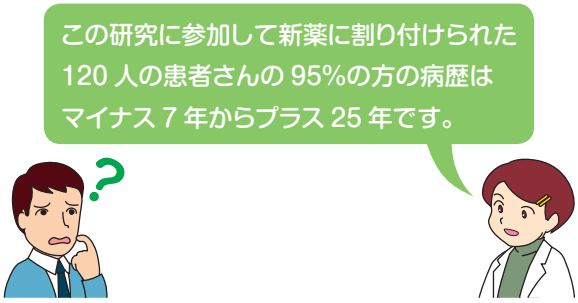


「平均 ± 2SD」の計算法を病歴に当てはめてみましょう。

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。

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

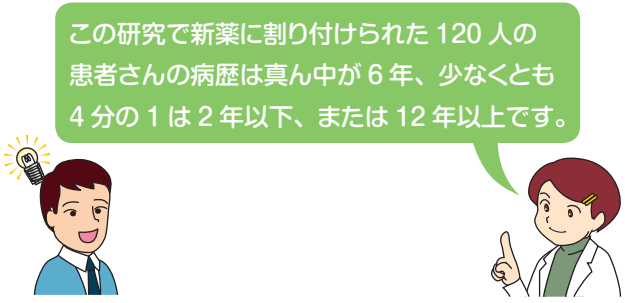
この研究に参加して新薬に割り付けられた 120 人の患者のうち，95% の病歴がマイナス 7 年からプラス 25 年？　どこかおかしいですよね。



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。



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？

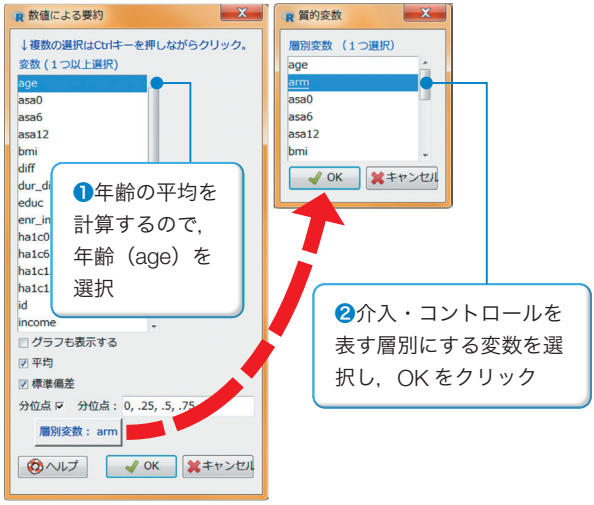
次は EZR を用いた記述統計量の計算方法について説明します。

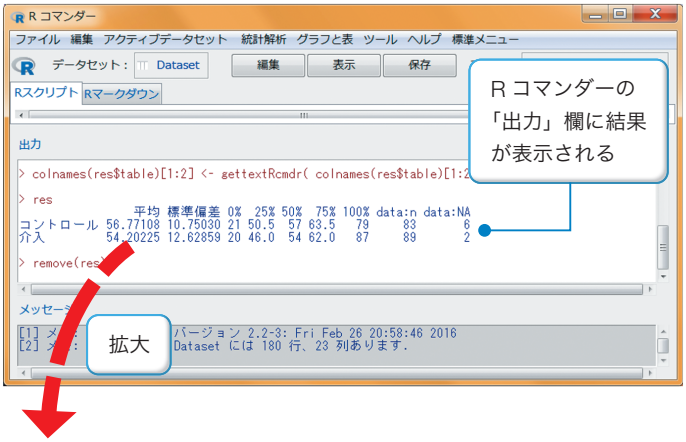
EZRを用いた記述統計量の計算

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。

統計解析 → 連続変数の解析 → 連続変数の要約





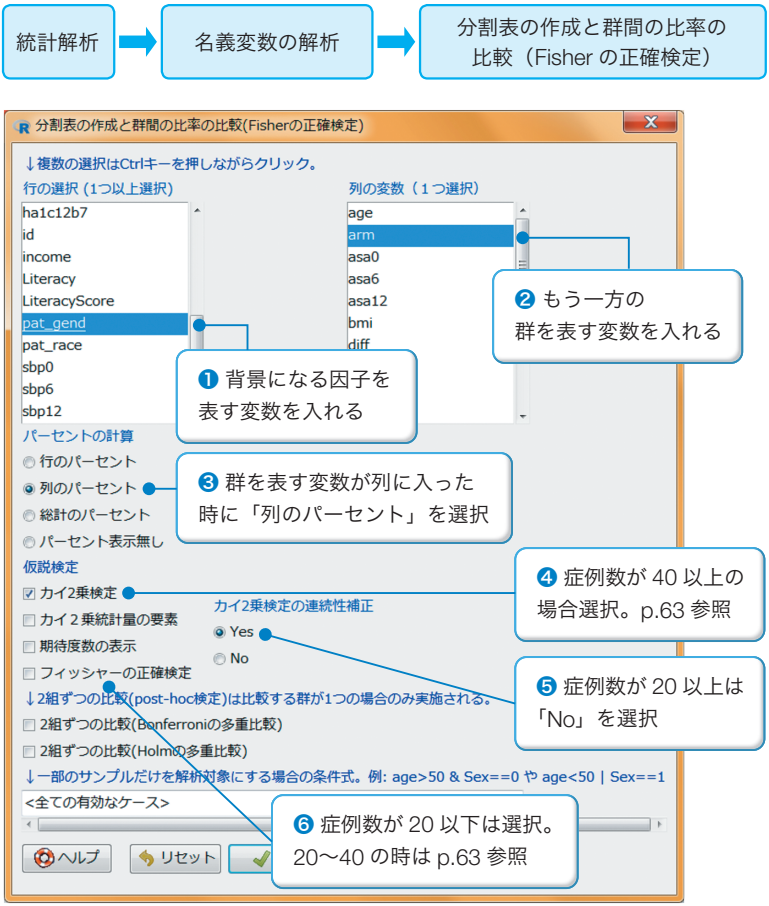
EZR の計算結果

四分位点

	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

中央値

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。



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，EZR の計算結果は上記のようにベージュの枠で示します。