高額な
統計ソフトは
もういらない!?

フリーソフト

Rを使った らくらく

医療統計解析入門

すぐに使える事例データと実用Rスクリプト付き

著 | 大櫛陽一（大櫛医学情報研究所所長／東海大学名誉教授）

中山書店

はじめに—フリーソフトRの薦めと本書の特徴—

医学をはじめ，経済学，教育学，心理学などは人間を対象とした科学・学問です．人は複雑な生き物で，同じウイルスに感染しても発病する人もいれば，発病しない人もいます．病気の治療でも，ある薬剤で大きな効果のある人もいれば，効果のない人もいます．ある人の発症する可能性や，特定の人に対する薬剤の効果を予測することは簡単ではありません．また，がんや糖尿病などの生活習慣病には，原因物質への暴露状態，栄養状態，運動習慣，遺伝的要因など多くの因子が関係しています．これらの因子のうち何が重要かを特定するには，膨大な症例の蓄積が必要で，ひとりの医師の臨床経験だけでは不十分です．

統計学を使うと，上にあげた医療上の諸問題に対しての解答（エビデンス）を得ることができます．個人ごとの疾病の発症率の予測や，特定の薬剤の治療率や副作用の発生率を予測することも可能です．統計学では，過去のデータを集めて解析しますが，その結果は目の前の健康診断の受診者や患者の近未来の生活の質を最大にするために使われるのです．

統計処理には難しい計算が必要なので，苦手だと思う学生や医療関係者が多いでしょう．しかし，最近は質の良い統計ソフトウェアが開発されていますので，計算はコンピュータに任せればいいのです．皆さんは，統計的な判断のやり方を理解して，データの収集と統計手法の選択をするだけでいいのです．この本でも，Rと名付けられたソフトウェアを使って医学統計の実習ができるようになっています．

統計ソフトウェアには，有名なSASやSPSSなどの市販されているものもありますが，医学論文に使われているオプションを含むと50万円前後となり，学生個人や統計処理がときどき必要となるだけの医師にとってはあまりにも高価な商品となっています．Rはフリーソフトで，インターネットのWebサイトから無料でダウンロードでき，自由に使うことができます．Rの解説書は無料で提供されていますし，日本語の解説書もインターネット上に多くの研究者から提供されています．また，医学論文で使われる最新の統計処理も網羅されています．世界中の統計専門家が協力して開発し，利用者数も多いので，信頼性は市販ソフトウェアと同等です．Rを使って解析された医学論文は増えており，2015年の英語論文は307件ありました（PubMedで“R package”で検索）．市販ソフトウェアでは毎年のアップデートのたびに購入価格の半額程度も必要になりますから，これもかなりの経済的負担です．Rは随時アップデートされており，更新もフリーなので最新の統計ソフトウェアを使えます．

本書は2007年に刊行した「看護・福祉・医学統計学—SPSS入門から研究まで」（福村出版）の姉妹版です．Rの出力結果はSPSSと全く同じであることを確認しています．

Rで問題となるのはその操作性です．利用者としては，統計学やコンピュータの熟練者が想定されているために，簡単とはいえキーボードからスクリプトという文字列を入力して動かさなければなりません．しかも英語なので，統計やコンピュータが得意でない人では難解

かも知れません．また，利用説明書が公開されてはいますが，ものすごく多くの統計手法が掲載されており，膨大なので見るべきページを探すこと自体が簡単ではありません．さらに，統計学やコンピュータの用語で書かれていますから，やはり難解です．

本書では，医学領域で使われるほぼすべての統計手法を，簡単なものから最新の多変量解析や多重比較までを分類整理して，わかりやすく説明しています．数式は，結果を理解するために必要なものだけの最小限にとどめてあります．難解な統計処理は，PC 画面の画像を掲載して，R での操作手順を丁寧に示しました．

本書の最も大きな特徴は，実際のデータに基づいて実習しながら理解できるようにしたことです．本書には豊富な事例を掲載しています．これらのデータは，ほとんどが実際の研究に使われたもので，統計処理の具体的な理解や，統計処理用データのつくり方に役立つと思われます．本書にある【例】では，詳細な処理手順と結果を示しており，各自のデータをつくるときに参考になり，教師が授業で説明するときの事例になります．【問】は，巻末に略解を掲示しており，読者の復習や学習達成度の確認や，教師が学生に対する課題として使うこともできます．付録の「R スクリプト一覧」や「統計処理のガイダンス」も役立つはずです．

実習で必要となる R，事例データ，R スクリプトは，中山書店のサイトからダウンロードできます．なお，ダウンロード方法については序章をご参照ください．

ダウンロードした R スクリプトを使えば，ほとんどマウス操作だけで実習ができます．また，皆さんが収集したデータの統計処理を行いたい場合は，スクリプトに含まれるデータファイル名と変数名を置き換えれば，後はマウス操作で統計処理を行うことができます．スクリプトに慣れてくれば，本書の範囲を超えて R の大きな世界にチャレンジすることも夢ではありません．

2016 年 1 月

東海大学名誉教授
大櫛　陽一

フリーソフトRを使ったらくらく医療統計解析入門
—すぐに使える事例データと実用Rスクリプト付き

CONTENTS

付録

第2章 2群の比較

医療分野では，2群を比較して差があるかどうかを知りたい場合が多々ある．たとえば，

- Aという薬を服用すると，症状が改善された人のほうがされない人よりもかなり多いように思えるが，確信をもってそういえるのかどうか．
- B療法を受けたグループとC療法を受けたグループとでは改善の度合いが違っているようにみえるが，違うといえるのかどうか．
- ボランティア経験のある学生とない学生とでは，調査結果から障害者に対する意識が違っているように思えるが，そう言い切れるのかどうか．

といった疑問が生じてきた場合である．このような場合に，統計学的に意味のある違いなのかどうなのかを明らかにしようとするのが，統計学における差の検定である．

本章では，2群の比較について，Ⓐ母集団と標本との比較，Ⓑ対応のある2群の比較，Ⓒ独立した標本の比較，に分けて例題をあげながら解説する．

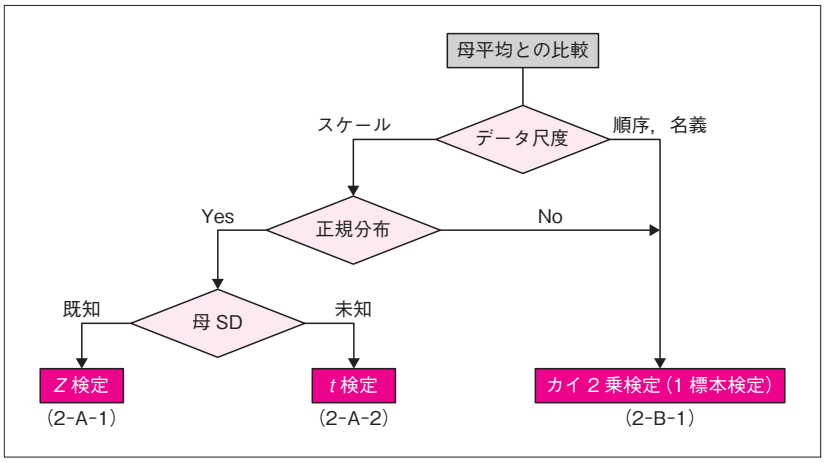
A 母集団と標本との比較

母集団と標本とを比較する場合，尺度がスケール（間隔・比率尺度）であれば母集団の平均と標本の平均とを比較することになる．また，順序尺度や名義尺度であれば母集団の比率と標本の比率とを比較する．

1. 母平均と標本平均の比較（スケールの場合）■

母集団の平均と標本の平均との差を比較することがある．たとえば，14歳児の身長の全国平均（母平均）とA中学校14歳児の身長の平均（標本平均）とを比べる場合などである．比較を行う場合，データが正規分布しているのかどうか，また，正規分布をしているとしても，母標準偏差（母SD）がわかっているのかどうか，などにより手法が異なる．この手法のフローチャートを次に示す．

- 母平均
- 標本平均
- 母標準偏差：母SD



1）正規分布するデータで母 SD が既知の場合：Z 検定

医療データでこの条件のそろった場合は少ない．R でも用意されていないので電卓で計算する．

母集団の平均（母平均），母 SD がすでにわかっている場合には，次の式を用いて差があるか否かを検定する．

μ：母平均，$\bar{x}$：標本平均，σ：母標準偏差，n：標本のデータ数．

$$Z=\frac{(\bar{x}-\mu)}{\frac{\sigma}{\sqrt{n}}} \qquad A=|Z|$$

統計仮説：標本平均（$\bar{x}$）は母平均（μ）と等しい．

統計仮説が成立すると $Z=0$（$A=0$）となる．データのばらつきにより Z が 0 でないこともあるが，Z は正規分布をするので，Z が 0 から離れるにしたがい，その確率は大きく下がってくる．正規分布の特徴については，1 章 B-4「正規分布」で述べた通りである．

ここで，統計的判断は次のようになる．

A の範囲	1.96 未満	1.96 以上～2.58 未満	2.58 以上～3.30 未満	3.30 以上
統計的判断	有意差なし	有意差あり	有意差あり	有意差あり
表　記	NS	$p<0.05$	$p<0.01$	$p<0.001$

2）正規分布するデータで母 SD が未知の場合：t 検定

母集団の平均のみが既知で，母 SD が未知の場合には，次の式を用いて検定を行う．ただし，μ：母平均，$\bar{x}$：標本平均，s^*：不偏標準偏差，ϕ：自由度，n：標本のデータ数である．

$$t=\frac{(\bar{x}-\mu)}{\frac{s^*}{\sqrt{n}}} \qquad \text{自由度：} \phi=n-1$$

- Z 検定
- t 検定
- 不偏標準偏差：SD の不偏推定値
- 自由度：データが自由に変化できる個数．t を計算するために $\bar{x}$ を使うので，$(n-1)$ 個のデータは自由に変化できるが，最後の 1 つは決められてしまう．自由度により t 分布の形が変化し，大きくなると正規分布に近づく．

統計仮説：標本平均は母平均と等しい（$t=0$）.

R で計算をする場合には，有意確率（p-value）をみる．有意確率の値が 0.05 未満であれば統計的に有意な差があると判断してよいことになる.

例 2.1　母平均と標本平均の比較（1 標本 t 検定）

データは「障害施設 .txt」に登録されています.

知的障害施設に入居している子ども 283 人のデータです．身長が低いように感じるのですが，文科省の全国データと比較してみましょう．身長は，年齢と性別の影響があるので，15 歳男子で全国の平均身長 168.4 cm と比較します.

〈解説〉

統計仮説：平均身長は全国と一致する.

①データの読み込み

```
> dat <- read.delim("障害施設.txt")
> dat1 <- subset(dat,(年齢==15)&(性別=="m"))
> attach(dat1)
```

② Shapiro-Wilk 検定で正規性の検定を行います．統計仮説：正規分布である.

```
> shapiro.test(身長)

        Shapiro-Wilk normality test

data:  身長
W = 0.97669, p-value = 0.2948
```

- $p \geqq 0.05$ なので正規性は否定されません．t 検定が適切です.

③全国平均値と比較するために t 検定を行います.

```
> t.test(身長,mu=168.4)

        One Sample t-test

data:  身長
t = -6.7925, df = 60, p-value = 5.645e-09
alternative hypothesis: true mean is not equal to 168.4
95 percent confidence interval:
 156.1682 161.7335
sample estimates:
mean of x
 158.9508
```

- t は -6.7925 と 0 より大きく離れた値でした．自由度（df）は 60，有意確率は 5.645×10^{-9} でした.
- $p<0.001$ なので，統計仮説は否定され，全国と有意差があります.
- 施設の子どもの平均身長が 158.9508 cm，全国が 168.4 cm なので，約 10 cm 低いことになります.

第3章 関係を調べる

医療分野において，データ A とデータ B との関係について調べることがしばしばある．たとえば，早起きと健康との関係，運動量と体力の関係，性別と介護に対する意識との関係など，関連性に注目が集まることはいろいろである．

データ尺度がスケールや順序で，変数 A が大きくなれば変数 B も大きくなる，その逆に変数 A が大きくなるにつれて変数 B が小さくなる，といった関係が認められる場合がある．他方，データ尺度が名義で，変数 A も B もいくつかのカテゴリーに別れ，そのクロス集計から関係が認められる場合もある．

ここでは，Ⓐ 2 変量の統計，Ⓑ順序およびスケール尺度データの統計図表と相関係数および回帰式，Ⓒ名義尺度データの統計表と検定，Ⓓ ROC 曲線，について事例を使って学習する．

A 2 変量の統計

1. 基本統計量

データ尺度がスケールの場合の基本統計量に関して，各変量の統計量に加えて，2 変量による統計量の数式による表し方を示す．

1）平方和と積和

- 平方和
- 積和

$S'_x = \sum x_i^2$　　x の平方和

$S'_y = \sum y_i^2$　　y の平方和

$S'_{xy} = \sum (x_i \times y_i)$　　x と y の積和

2）修正ずみ平方和と積和

平均値を引いて計算すると修正ずみとなる．

$S_x = \sum (x_i - \bar{x})^2$　　x の修正ずみ平方和

$S_y = \sum (y_i - \bar{y})^2$　　y の修正ずみ平方和

$S_{xy} = \sum \{(x_i - \bar{x}) \times (y_i - \bar{y})\}$　　x と y の修正ずみ積和

3）分散と共分散

- 分散
- 共分散

$V_x = S_x \div n$　　x の分散

$V_y = S_y \div n$　　y の分散

$V_{xy} = S_{xy} \div n$　　x と y の共分散

4）不偏分散と不偏共分散

- 不偏分散
- 不偏共分散

$V_x^* = S_x \div (n-1)$　　x の不偏分散

$V_y^* = S_y \div (n-1)$　　y の不偏分散

$V_{xy}^* = S_{xy} \div (n-1)$　　x と y の不偏共分散

B 順序およびスケール尺度データの統計図表と相関係数および回帰式

1. クロス集計表と箱ひげ図

- クロス集計表
- 箱ひげ図

順序尺度または名義尺度データの 2 変数の関係を示すためにクロス集計表が使われる．各変量を行と列にして，表内には各行と列に対応する人数が入る．1 つの変量が順序尺度または名義尺度データで，もう 1 つの変量がスケール尺度データのときには箱ひげ図が使われる．横軸に順序尺度または名義尺度データをとり，縦軸をスケール尺度データにする．

例 3.1　**関係を示す統計図表**

前にも使ったデータセット「意識 .txt」を例にします．

〈解説〉

「介護保険が導入される前の社会福祉に対する意識」を調査する試験の結果を，年代で比較してみましょう．変数は，福祉に対する意識クイズの得点：quiz，性別：sex（1：男，2：女），年齢 3 階級：age3，（1：20・30 歳代，2：40・50 歳代，3：60・70 歳代）です．

このクイズの得点が高いほど，福祉に対する理解があるという調査です．

①データを読み込みます．

```
> dat <- read.delim("意識.txt")
> attach(dat)
```

②クロス集計表：table（行，列）

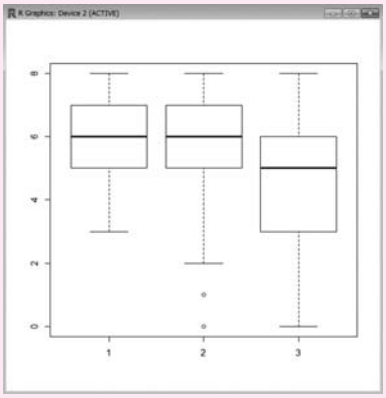
性別（sex）を行に，クイズの得点（quiz）を列にした表をつくります．

```
> table(sex,quiz)
   quiz
sex  0  1  2  3  4  5  6  7  8
  1  5  3  1  5  6 12 27 17  4
  2  8  5  3 10 14 29 32 21 11
```

③箱ひげ図：boxplot（量的変数～群分け変数）

クイズの得点（quiz）を 3 つの年齢階級で比較します．

```
> boxplot(quiz~age3)
```



- X 軸：1：20・30 歳代，2：40・50 歳代，3：60・70 歳代
- Y 軸：福祉クイズの得点（0〜8）
- 20・30 歳代と 40・50 歳代での 25%タイル値，中央値，75%タイル値は同じですが，60・70 歳代ではいずれも低くなっています．この調査は介護保険導入前に行われており，高齢者では「人のお世話になる」ことへの遠慮や「他人が家に入る」ことに抵抗があったと思われます．

2. 散布図と Pearson の相関係数

2 つの変数がスケール尺度データのときには，関係をみる図として散布図が使われる．原因と考えられる変数を横軸（X）に，結果と考えられる変数を縦軸（Y）にとる．この中に各データをプロットしたものが散布図で，この関係を数値で表現するのが Pearson（ピアソン）の相関係数（r）である．2 変量の共分散を 2 変量の標準偏差の積で除したものであり，次の式により算出する．

- Pearson の相関係数

$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\times(y_i-\bar{y})^2}}=\frac{S_{xy}}{\sqrt{S_x\cdot S_y}}$$

r が −1〜0 のときは負（逆）相関といい，r が 0〜1 のときは正（順）相関とよぶ．r が +1 に近いほど高い正の相関があるといい，−1 に近いほど高い負の相関があるという．r が 0 であれば，2 変量間に相関関係がまったくないことを意味している．

また，相関係数 r の有意差の検定は次の式で行う．

$$t=r\times\frac{\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \text{自由度：}n-2$$

統計仮説：相関係数（r）$=0$（$t=0$）

この仮説はデータ（X_i, Y_i）が 2 次元正規分布していることを前提と

中山書店の出版物に関する情報は，小社サポートページを御覧ください．
http://www.nakayamashoten.co.jp/bookss/define/support/support.html

大櫛　陽一（おおぐし　よういち）

大櫛医学情報研究所所長・東海大学名誉教授

S22. 01. 03 生.
昭和 46（1971）年大阪大学大学院工学研究科修了，大阪府に就職．府立成人病センター，羽曳野病院，母子センター，府立病院を歴任．昭和 63 年（1988）東海大学医学部教授．平成 24 年（2012）から東海大学名誉教授，大櫛医学情報研究所所長．医療統計学，医療情報学，脳卒中，高血圧，糖尿病，メタボリックシンドローム，脂質異常症，性差医療などに関する著書多数あり．

フリーソフト R を使った
らくらく医療統計解析入門

2016 年 3 月 7 日　初版第 1 刷発行
〔検印省略〕
著　者　大櫛陽一
発行者　平田　直
発行所　株式会社 中山書店
〒112-0006 東京都文京区小日向 4-2-6
TEL 03-3813-1100（代表）
振替 00130-5-196565
http://www.nakayamashoten.co.jp/
装丁　臼井弘志（公和図書デザイン室）
印刷・製本　株式会社 真興社

Published by Nakayama Shoten Co.,Ltd. Printed in Japan
ISBN 978-4-521-74364-6