

第1回 たかが記述統計，されど記述統計

横浜薬科大学医療統計学

奥田千恵子 Chieko OKUDA

はじめに

「客観的な証拠に基づく医療 (evidence-based medicine: EBM)」の考え方が世界的に普及してきたことに伴って、医療研究者は統計リテラシー(統計的な考え方)を要求されるようになった。多くの医療施設で最新の高度な医療が行われるようになった現在、直接研究に携わらない医療従事者であっても、専門誌に掲載された科学論文を日常的に読まなければならない。研究が正しく行われたかどうかを判断し、発表された結果を批判的に評価する能力が求められるようになったのである。

現状では、医療研究者の統計リテラシーは十分とはいえない。加えて、医療研究の特殊性を理解できる統計学者が少ないことも原因となって、ほかの分野に比べて統計学の誤用が多いことが以前から指摘されてきた。相互理解が不十分なまま、論点がかみ合わない議論の末、さらに誤用を重ねる医療研究者も少なくない。このような状態が続けば医療研究の発展を妨げることになりかねない。

統計学は「情報の不確かさの度合い」を数値化する道具である。統計ソフトが普及していなかった頃は、統計学の主な内容は解析理論や実際の数値計算だったが、現在は、そのような仕事はコンピューターがやってくれる。研究者に求められている統計リテラシーとは、

- ・自分の目的に合った研究デザインで、適正にデータを収集できる
- ・データに合った解析手法を選択し、統計ソフトが出力した結果を正しく解釈できる
- ・解析結果から結論を導き出して、論文などでわかりやすく報告できる

ということである。

本稿では、誰でも統計リテラシーを身に着けて、医療研究を行ったり論文を読んだりできるようになることを目指して、できるだけ難解な統計用語や数式を用いず、医療従事者の日常感覚に近い言葉で説明し、具体例を用いて、気を付けなければならないポイントを示す。

記述統計は脇役？

収集したデータを統計学的に扱う方法には、大きく分けて、記述統計 (descriptive statistics) と推測統計 (inferential statistics) がある。

記述統計とは、得られたデータを集計し、グラフなどを用いてわかりやすく表示することである。国勢調査に代表されるような大規模な調査であれば、記述統計だけでもデータがもつ情報を十分に引き出すことができる。ところが、医療分野の研究では、サンプル数がせいぜい数十から数百程度の中小規模の研究が多いので、研究ターゲット集団に関して普遍性のある情報を引き出すには推測統計を用いる必要がある。

論文の冒頭で、解析に用いた対象の背景因子などのデータを表にまとめる際に、ちょっとした脇役を演じるのが記述統計、というイメージだろうか。

推測統計では再現性が求められる

では、主役となる推測統計とはどんなものだろうか？

例えば、1組の(複数の)対象から何らかの数値(連続量データ)を得て、その平均値を求めたとしよう。記述統計では、その平均値は唯一無二のものとして扱われる。

一方、推測統計では、母集団から偶然選ばれた1組の標本の平均値に過ぎないと考える。別の1組が選ばれていたなら異なった平均値が得られたはずである。さらに、また別の1組を選んだら……。推測統計では、同じ調査や実験を何回も繰り返すという仮定がなされる。同じ調査や実験を、誰が、いつ、どこで行っても、同様の結果が得られるという暗黙の再現性が求められるのである。

ところで、「特定の疾患をもつ患者全体」という母集団を想定する場合、現在、世界中に存在する患者だけでなく、過去にも、未来にも同様の患者がいるはずである。そんな集団を実体としてとらえることができるだろうか。そこから標本抽出を行うということはほとんど不可能である。

現実には、特定の期間に、特定の医療機関に入院あるいは受診した患者など、ごく小規模な集団を対象とし、原則、全数調査を行うのが一般的である。しかし、病院の窓口を通して訪れる患者というのは、病院の規模や専門性、評判などを判断材料として自由意志で集まった、かなり偏った集団である。標本の偏りを統計学的に修正することはできない。

従って、このような研究から普遍的な情報を引き出すのは容易なことではない。研究結果が当てはまると考えられる集団を常識で判断し、これを研究ターゲット集団（架空の母集団）とし、推測統計の手法を用いて結論を導かなければならない。だがその前に、まず、十分なデータ探索を行い、研究仮説を明確にする必要がある。

探索的研究と検証的研究

研究仮説は日常業務のなかから生まれる。最初は個人的な印象、主観に過ぎなかった考えが、意識的にデータ収集を行うことで、次第に客観性をもった研究仮説となっていく。研究のスタート時点では、試行錯誤しながら小規模な研究を行い、結果が蓄積するにつれて最初に立てた研究仮説を少しずつ修正していくことが多い。このような研究を探索的研究 (exploratory study) と呼ぶ。

探索的研究を積み重ねて最終的な研究仮説が決まったら、規模を拡大して、仮説の統計学的な証明をしなければならぬ。そのような研究を検証的研究 (confirmatory study) と呼ぶ。検証的研究では、事前にしっかりとした計画を立て、できるだけ途中で変更を加えずに研究を進める必要がある。

研究デザインの序列

比較の対照を設けて、推測統計の手法を用いて解析を行った研究は分析的研究 (analytical study) と呼ばれる。医療研究における分析的研究の多くは因果関係を求めることを目的にしており、実験的研究 (experimental study) と観察的研究 (observation study) に分けられる。

実験的研究では、研究者が対象にどのような処置 (原因) を行うかを制御できるので、アウトカム (結果) は明瞭である。一方、そのような制御が行われない観察的研究では、因果関係の曖昧さが少ない (原因が先、結果が後という関係が判別できる) デザインかどうかによってエビデンスレベルが異なる (表1)。

表1 医療研究における研究デザインの分類
(表の下のほうほどエビデンスが強い)

記述的研究 (主として探索的研究として行われる、対照がない) 例) 症例報告、症例集積研究、特定地域の健康調査など
分析的研究 (主として検証的研究として行われる、対照がある)
観察的研究 (対象を制御せず、聞き取り調査や健康診断のみを行う)
横断的研究 (時間の要素がない) 例) 有病率や検査値の群間比較、相関関係など
縦断的研究 (時間の要素がある)
後向き研究 (スタート時点で「結果」が得られている) 例) ケース・コントロール研究、後向きコホート研究など
前向き研究 (スタート時点で「結果」が得られていない) 例) 前向きコホート研究など
実験的研究 (対象を制御し、薬剤の投与や処置などの介入を行う) 例) 臨床試験など

記述統計を正しく使う

冒頭に述べたように、医療研究では記述統計は脇役というイメージは否めないが、だからといって、推測統計の手法が使われていればエビデンスのレベルが高いというわけではない。恣意的な群分けをして無意味な群間比較を行ったり、多項目間で手当たり次第に相関分析をしても、再現性のある有意差や相関係数は得られない。

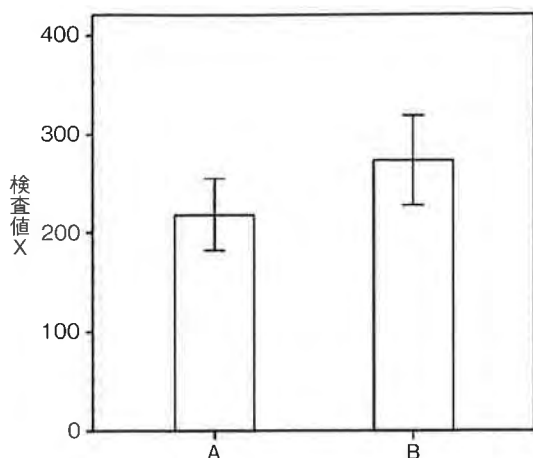
研究仮説が曖昧で未だ検証段階に至らない研究では、データを適切に要約して、表やグラフで表したほうが、研究仮説や研究デザインを構築し検証的研究へと発展させるための有用な情報を提供できる。探索的研究では、記述統計を正しく使うことにより、独断的な結論を出さずに最終判断を論文の読者に委ねることができる。

データの属性

どんな属性のデータであっても統計学的に解析することができるが、属性に合わせて統計手法を選ぶ必要がある。おおよその解析計画を立ててからデータ収集を行わなければならない。無計画なアンケート調査や評価方法で得られたデータを解析しても意味のある情報は得られない (表2)。

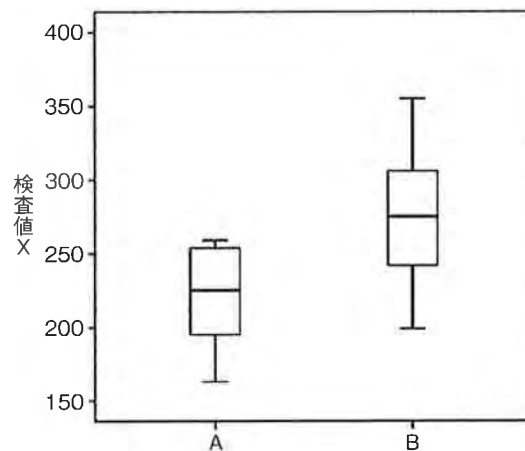
表2 データの属性

<p>連続量 (continuous data)</p> <p>数値データ (numerical data) であり、以下のような尺度で測定されている。</p> <ul style="list-style-type: none"> ・比尺度 (ratio scale) : 重さ, 長さ, 濃度など, 2つの数値の比および差をとることができ, 数字の0が絶対的なゼロ点を示す。 ・間隔尺度 (interval scale) : 温度や知能指数など, 2つの数値の差をとることができるが, 比は意味をもたず, 数字の0が絶対的なゼロ点ではない。
<p>離散量 (discrete data)</p> <p>個体数や回数などの数値データ (整数) や, 数量的には測れないカテゴリデータ (categorical data) がある。カテゴリデータには順序の有無により以下のような尺度がある。</p> <ul style="list-style-type: none"> ・順序尺度 (ordinal scale) : 健康状態 (非常に悪い, 悪い, 普通, 良い, 非常に良い), 尿糖 (++, +, ±, -) など, データの順序には意味がある。順序カテゴリデータ (ordered categorical data) と呼ぶ。 ・名義尺度 (nominal scale) : 性別 (男, 女), 生死 (生存, 死亡), 疾患名 (循環器疾患, 呼吸器疾患, 消化器疾患) など, データには順序がない。2つのカテゴリで表されている場合は2値データ (binary data) と呼ぶ。



平均値 (SD)

図1 エラーバー付棒グラフの例



箱の中央：中央値
箱の上下：25および75パーセンタイル値
ひげの上下：10および90パーセンタイル値

図2 箱ひげ図の例

記述統計の報告の仕方 To Do & Not To Do

1. 連続量データ

データの分布が(1)正規分布 (normal distribution) に従うか, (2)正規分布に従わないかによって, データの要約方法を変える必要がある。ただし, 医療研究で厳密な正規性が要求されることは少ない。通常は, ヒストグラムによる視覚的な対称性の確認だけで十分である。データの範囲の区切り方によって印象が大幅に変わってしまうことがあるので, 区間を変えたものも描いてみて比較するとよい。

簡単なデータ変換で正規分布に改変できる場合もある。例えば, 血液中の物質濃度は高値側に裾を長く引く傾向があるが, 対数変換すると正規分布に近似できる例が多いことが経験的に知られている。

データの中心性の指標となる「代表値」, および, その周りにどの程度散らばっているかを表す「ばらつきの値」を求める。通常, (1)の場合は, 平均値 (mean) と標準偏差 (standard deviation : 以下, SD), (2)の場合は, 中央値 (median) と四分位範囲 (interquartile range) が用いられる。

「代表値」として中央値を使っているにもかかわらず, 「ばらつきの値」としてSDを使うという, 組み合わせの間違いがよくみられる。

「平均値±SD」ではなく, 「平均値 (SD)」と報告することが推奨されている。

例) 平均値 (SD) は58.4歳 (9.0歳)

中央値と四分位範囲で表す場合は,

例) 中央値 (四分位範囲) は59.5歳 (51.3~66.5歳)

これらをグラフで表す場合, (1)では, エラーバーを付けた棒グラフ (図1) などが一般的であるが, (2)では, 箱ひげ図 (図2) を用いる。箱ひげ図は(1)の場合にも使

うことができるが、エラーバー付きの棒グラフは、(2)の場合には使えない。エラーバーや箱ひげ図には様々な統計量を描くことができるので、何が描かれているか必ず図中に表記する。

従来、エラーバーは短い方が見栄えがよいという理由で、SDではなく、標準誤差 (standard error: 以下、SE) が用いられてきたが、誤解を引き起こしやすいので推奨できない。データのばらつきを示す場合は、SDを、平均値の信頼性を示す場合は、SEではなく、信頼区間 (第2回で詳しく述べる) を描く。

推測統計で用いる手法との間に齟齬を来さないようにしなければならない。t検定や分散分析などはパラメトリック検定法 (parametric test) と呼ばれる。パラメトリックとは、正規分布を表す数式に含まれる、母平均値、および、母標準偏差という母数 (parameter) を利用することを指す。従って、記述統計においても平均値やSDを用いる。

一方、マン・ホイットニー検定やクラスカル・ウォリス検定など、ノンパラメトリック検定法 (nonparametric test) と呼ばれる手法を用いる場合、そのような母数を利用しないので、記述統計で平均値やSDを用いるのは

不適切である。実際にデータが正規分布に従うか従わないかにかかわらず、中央値と四分位範囲を用いる。

2. 離散量データ

離散量のうち、評価スケールなどのフルスケールが10点程度以上の整数値で表される順序カテゴリデータは、連続量データと同じ扱いをすることができる。それ以下の順序カテゴリデータや2値データ、順序のないカテゴリデータは、各カテゴリの度数を実数で示す。必要に応じて、各カテゴリの度数や割合を帯グラフや円グラフなどで表す。

割合を示す場合、必ず分母に当る値が何を表すか明記すべきである。割合は、あくまでもデータを読みやすくするために補助的に用いる。

例) 女性は全被験者の60% (60人中36人) である。

参考文献

- 1) B.H. Stephen *et al.*: "医学的研究のデザイン 第2版", 木原雅子, 木原正博訳, メディカル・サイエンス・インターナショナル, 東京, 2004.
- 2) T.A. Lang, M. Secic: "わかりやすい医学統計の報告", 大橋靖雄, 林 健一監訳, 中山書店, 東京, 2011.
- 3) 奥田千恵子: "親切的な医療統計学", 金芳堂, 京都, 2014.