かなり役立つ生涯学習

医療統計シリーズ③

「データの集め方・まとめ方：平均値・中央値、信頼区間、箱ひげ図」

大阪医科大学　研究支援センター　医療統計室　室長・准教授
伊藤　ゆり

研究に使用するデータの特徴を知る

今回は基本に立ち返って、臨床研究を行う上で、最初に知っておきたい「データの集め方・まとめ方」について紹介してみたいと思います。まず、「研究のために意図して集めたデータ」と「（診療など）他の目的で勝手に集まったデータ」とでは研究に使用する上での質が大きく異なります。前者は、最初に研究の計画を考え、必要な情報を決まった手順で収集します。後者は、「他の目的で勝手に集まったデータ」ですので、研究に利用する際に注意する必要があります。詳しくは今後、各種研究デザインの紹介をしていく際に述べますが、両者は大きく区別してください。

近年、後者のデータで臨床研究を行うことが増えてきました。2000年代以降、診療録や支払い情報が電子化され、いわゆる「医療ビッグデータ」と呼ばれる日常に集積していく膨大な医療情報が使用可能となりました。サイズが大きく、変数も多い「宝の山」のように感じるデータですが、研究に使用するには、そのデータの特徴を押さえておく必要があります。まず、留意しなければならないのは、そのデータの「母集団」です。例えば、皆さんがお勤めの病院や診療所で長年集積した医療情報を研究に使用するとします。そのデータの母集団はどのような集団ですか？ 多くの場合、単一の医療機関を受診した方のデータであるため、代表性に乏しく、「通院を続けた人（続けられた人）」をとらえている偏ったデータである可能性があります。皆さんの医療機関での治療に納得できず、離れてしまった患者さんのデータは残りません。来院せず、症状が悪化し、死亡してしまった患者さんのデータをとらえることも難しいです。そのような限界のあるデータであることを理解したうえで使用しましょう。

一方で、全国の患者さんのデータを悉皆的に収集したデータや、特定の保険加入者のデータなど、比較的代表性の高い既存データもあります。また、貴院独自で行っている新しい治療法などのデータは他施設にはない、貴重なデータとなりうるわけです。研究に使用するデータが他の目的で収集された場合には、そのデータの母集団の特定と、データの信頼性に留意して研究に利用しましょう。

データのまとめ方

データを扱う研究を実施する際には、基本的な統計の値をまとめていきます。皆さんよくご存じの平均値や中央値などの代表値とともに、データのばらつき具合や信頼度を示す標準偏差（標準誤差）・分散、信頼区間など様々な指標を表にまとめていくことになります。その際に、「どの指標をまとめたらいいのかな」と悩まれたことはないでしょうか？

データの代表値をまとめる前に、まず、そのデータの特性を確認します。データの種類が数値なのかカテゴリーなのかを確認します。数値データの場合、まずは平均値±標準偏差と考えがちですが、最初に確認してほしいのはその数値データの分布です。これはヒストグラムを描くことで確認できます。扱っている数値データが図1のような左右対称のデータの分布（正規分布）になっている場合には平均値±標準偏差を示すことに意味があります。

しかし、正規分布していない場合、通常、分布によらない解析方法（ノンパラメトリック手法）を取ります。これは数値データをそのまま扱うのではなく、数値の大きさを順位に置き換えて扱う手法です。正規分布（やその他の分布）を想定できない分布

になっているデータの場合には、代表値は中央値とデータのばらつきを四分位範囲（25パーセンタイル値と75パーセンタイル値）で示すことになります。図2のような箱ひげ図で表現するとわかりやすいでしょう。

また、2群の差を検定する場合にも、データの特徴に合わせた検定を使用しましょう。正規分布を仮定している場合、平均値の差の検定をt検定で行い、要約統計量は平均値と標準偏差（や95%信頼区間）を示します。一方、分布を仮定しない場合、Wilcoxon検定などノンパラメトリックな検定方法を用いますが、その場合の要約統計量は中央値と四分位範囲で示しましょう。どのようなデータの分布を想定し、何を比較しているのかを考えた上で、データをまとめるようにしてください。

略歴

大阪大学大学院医学系研究科博士前期・後期課程卒業後、大阪府立成人病センター（現大阪国際がんセンター）リサーチ・レジデント、研究員、主任研究員を経て、大阪医科大学研究支援センター医療統計室准教授（現職）。現在、がん疫学、健康格差、医療統計の研究に主に従事。

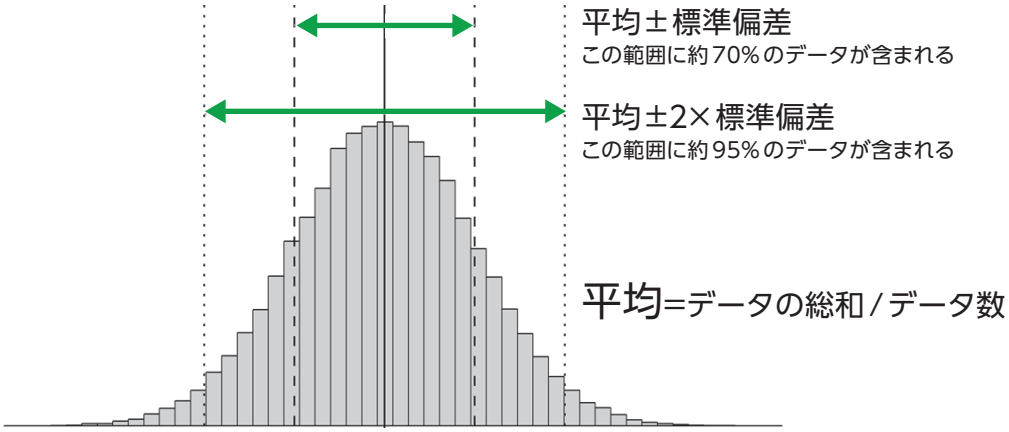

図1. 正規分布しているデータのヒストグラム

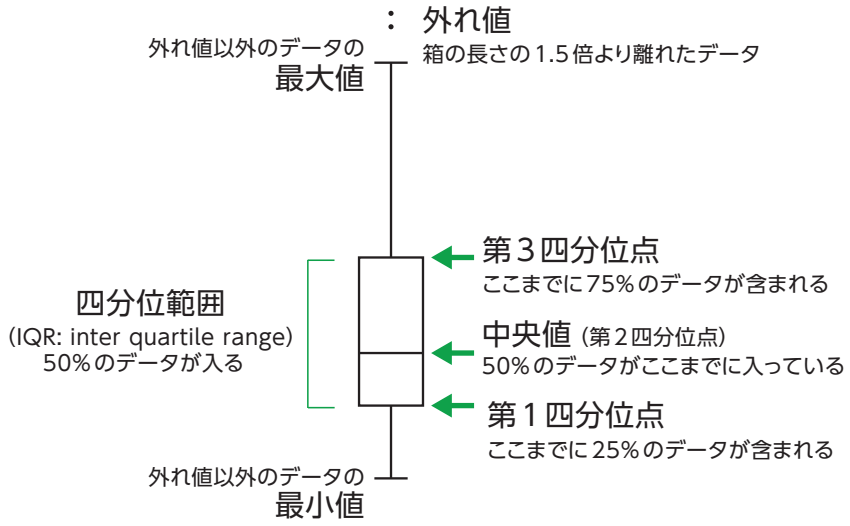

図2. 箱ひげ図