医療職のための統計シリーズ

医療職のための学び直し―研究デザインから論文報告までの生物統計学の道標―
第５回　データの分布と一変数の要約

カワハラ　タクヤ　　サカマキ　ケンタロウ
川原　拓也＊1　坂巻　顕太郎＊2

Ⅰ　はじめに

調査や実験を通して得られた各個人のデータから，あるグループ（集団）におけるデータの特徴を解釈するには，データの適切な要約が重要である。例えば，心筋梗塞患者の収集期血圧の解釈を考える。各個人の値は治療の選択などに必要と考えるかもしれないが，「心筋梗塞患者」という集団の特徴は個人ごとの値をみてもわからず，集団の特徴がわからなければ，各個人の値も解釈はできない。研究におけるデータの要約の第一歩は，データを得た集団の特徴をまとめることであり，論文における「表１」を作成することである。「表１」のような要約統計量（summary statistics）をデータ（変数の分布）の特徴を示す値として「表１」に記載することで，研究結果の一般化可能性や比較研究における群間の比較可能性を検討する。

医学系研究で扱うデータは，離散変数（discrete variable），連続変数（continuous variable），生存時間変数（time-to-event variable）のいずれかに分類できる。ここでいう「変数」は，簡単には「データ」のことであり，「尺度」とよばれる場合もある（正確にはこれらは区別される）。以下では，離散変数や連続変数の要約統計量，「表１」における要約統計量の記載，一つの変数に関する様々な情報を表現するグラフ，について説明する。なお，生存時間変数のまとめ方は後の連載に説明を任せる。

表１　仮想的な介入研究での「表1」

	介入群（N＝250）	非介入群（N＝250）	全体（N＝500）
年齢[a]	35.0（5.1）	34.4（5.1）	34.7（5.1）
≦30歳	40（16）	55（22）	95（19）
30-40歳	170（68）	155（62）	325（65）
≧40歳	40（16）	40（16）	80（16）
性別（女性）	221（88）	213（85）	434（87）
職種			
看護師	118（47）	129（52）	247（49）
保健師	73（29）	76（30）	149（30）
その他	59（24）	45（18）	104（21）
勤務年数[b]	3.1（1.2，6.9）	3.4（1.4，7.2）	3.2（1.3，7.0）

注　指定がない箇所は頻度（割合％）を表す。
a 平均値（標準偏差），b 中央値（第一四分位数，第三四分位数）

Ⅱ　データの型と図表による提示

（１）　離散変数

離散変数の中でも，「あり」または「なし」，「65歳以上」または「65歳未満」などのように，二つのカテゴリのいずれかの値をとる変数を二値変数（binary variable）とよぶ。例えば，心筋梗塞発症の「あり」「なし」といった結果（結果変数），「新治療」「標準治療」，「喫煙者」「非喫煙者」といった原因（説明変数）など，医学系研究では二値変数が扱われることが多い。

二値変数に対する要約統計量に，頻度（frequency）と割合（proportion）がある。頻度は，カテゴリ内の人数や，ある事象の発生回数などを表す。割合は，心筋梗塞発症の例を使うと，

$$\text{発症割合} = \left(\begin{array}{c}\text{発症あり}\\ \text{の人数}\end{array}\right) \Big/ \left(\begin{array}{c}\text{発症ありと発症なし}\\ \text{の合計の人数}\end{array}\right)$$

という計算から求まる。一般には，×100をして，パーセントで表記することが多い。説明変数のカテゴリごとに結果変数の割合を比較することで，二値変数である原因と結果の関連を検討するなどもよく行われる。

離散変数が三つ以上のカテゴリを取り得る場合，カテゴリに順序があるかないかが解釈するうえで重要となる。例えば，「あなたの健康状態は？」という質問に対する「1.良い」「2.やや良い」「3.普通」「4.あまり良くない」「5.良くない」という回答は，カテゴリを表す数字が

＊1 東京大学医学部附属病院臨床研究推進センター助教
＊2 横浜市立大学データサイエンス推進センター特任准教授