母集団と標本

統計解析の目的は，関心のある対象の性質を，データから明らかにすることです．まず，最初にいくつかの例を考えてみましょう．

例 1　日本の雇用実態を知るために，国勢調査により全日本居住者の就業状態に関するデータを取った．

例 2　あるテレビ番組を，テレビ所有世帯のうち何パーセントが視聴したかを調べるために，モニター世帯に設置されているテレビに接続した専用の機器からデータを取った．

例 3　あるメーカーが，製造しているノート PC の耐久力を知るために，いくつかの製品を抜き取り，高さ何 cm から落としたら壊れるかを調べた．

これらの例で，関心のある対象とは，日本居住者全員（例 1），テレビ所有世帯（例 2），製造している全てのノート PC（例 3）ということになります．統計学では，データを取ることで性質を明らかにしたい関心のある対象のことを**母集団**といいます．

このとき，例 1 では母集団の全てのデータを取っていますが，例 2 では，母集団の全て（テレビを所有する全ての世帯）からデータを取るかわりに，一部のモニター世帯からのみデータを取っています．また，例 3 でも全ての製品で実験すると売るための製品がなくなってしまいますから，一部の製品で実験をするしかありません．すなわち，一般的に，母集団からデータを得るときには，母集団の全てのデータを取るのではなく，母集団の一部のデータのみを抽出することになります．このように，母集団からデータを抽出することを**サンプリング**，サンプリングによって得られたデータを**標本（サンプル）**といいます．

統計学の目的を統計学の言葉を使って言い直すと，「サンプリングによって得られた標本から母集団の性質を明らかにする」ということができます．既に述べたように，サンプリングでは母集団すべてのデータが得られるとは限りませんので，得られた標本から母集団の性質を推測する必要があります．統計学は，標本から母集団の性質を推測する手法に関する学問である，ということもできるでしょう．

しかし，標本から母集団の性質を推測する前に，まずは得られた標本自身の性質を明らかにする必要があります．具体的には，得られた標本を様々なグラフにより視覚化したり，**統計量**とよばれるものを算出することで，標本の要約をすることができます．統計学では，標本から母集団の性質を推測する手法については**推測統計**，標本自身を要約する手法については**記述統計**と分類されます．本書は，記述統計について扱っていきますが，そ

の内容は，推測統計を学習する際の基礎にもなります．また，記述統計・推測統計に関わらず，母集団と標本を明確に区別して違いを理解することは，統計学を通して非常に重要なことですので，その違いをしっかりと理解しておくようにしましょう．

データの分類と尺度

統計学はデータを扱う学問ですが，一口にデータと言っても，データは様々な種類に分類することができます．また，データの種類によっては，利用できる統計の手法が異なることもありますので，統計学を学習する上で，このデータの分類を知っておくことは重要になります．

まず，データは大きく分類して，**質的データ**と**量的データ**に分類することができます．更に，質的データは**名義尺度**と**順序尺度**に，量的データは**間隔尺度**と**比例尺度**に分類することができます．以下では，それぞれのデータ・尺度がどのように分類されているかを説明します．

質的データは，ものの属性や性質を示すカテゴリーで表されるデータのことです．例えば，血液型のデータを考えると，データはA型，B型，AB型，O型の4つのカテゴリーのいずれかの値を取るので質的データとなります．また，A・B・C・D・Fの5段階で評価される成績のデータも質的データです．上記の血液型のデータと成績のデータはいずれも質的データですが，両者には違いがあります．それは，成績のデータについては，「AはBよりも良い」というように，順序関係がありますが，血液型のデータについては順序関係がない，ということです．質的データの中でも，血液型のデータのように順序関係がないようなデータを**名義尺度**といい，成績のデータのように順序関係があるようなデータを**順序尺度**といいます．名義尺度の例としては，血液型のデータ以外に，テレビのチャンネル，学籍番号などが挙げられ，順序尺度の例としては，成績のデータ以外に，マラソンの順位，星の明るさの等級などが挙げられます．

順序尺度について一つ注意するべき点があります．順序尺度には順序関係がある，と書きましたが，その間隔には意味はありません．成績のデータを例に説明すると，AはBよりも良いと言えますし，DはFよりも良いと言えますが，AとBの間にある差と，DとFの間にある差は同じとはいえません．

量的データは，その名前が示す通り，物の量を表すデータです．例えば，気温（華氏，摂氏）や身長のデータなどがこれにあたります．ここでもやはり，気温のデータと身長のデータには違いがあります．気温は，20 °Cと10 °Cでは，「20 °Cの方が10 °Cよりも10 °C高い」というように，大きさの差には意味がありますが，「20 °Cは10 °Cの2倍暑い」とは言わず，その比率は通常意味を持ちません．一方で，身長は，200 cmと100 cmでは，「200 cmの方が100 cmより100 cm高い」というように，大きさの差に意味があるのと同時に，「200 cmは100 cmの2倍の高さ」というように，その比率も意味を持ちま

す．気温のように，大きさの大小と間隔に意味があるが，割合や比率には意味がないようなデータを**間隔尺度**といい，身長のように，間隔尺度の性質に加えて，割合や比率にも意味を持つようなデータを**比例尺度**といいます．間隔尺度の例としては，気温の他に偏差値や暦年などが挙げられ，比例尺度の例としては，体重や収入などが挙げられます．間隔尺度は和や差には意味がありますが，割合には意味が無いので乗除には意味がありません．一方，比例尺度は和や差だけでなく，乗除に意味があるので，四則演算をすることができます．

間隔尺度と比例尺度を見分けるのは少し難しいですが，「原点（多くの場合が 0)」が「何も無いこと」を表すなら比例尺度と言えます．例えば，長さで「0 cm」は「何も無いこと」を表しますが，「0 °C」は華氏に変換すると「32 °F」となることからもわかるように，何も無いことを表すわけではありません．

例題 1　以下のデータは，名義尺度，順序尺度，間隔尺度，比例尺度のどれに対応するか答えなさい．
(1) 生後間もない赤ちゃんの体重　(2) 野球選手の背番号　(3) 英語のテストのクラス内の順位　(4) 郵便番号　(5) 知能指数

(1)　体重なので，和や差に加え，割合や比率にも意味を持ちます．よって比例尺度です．
(2)　番号の大小に本質的な意味を持たないので，名義尺度です．
(3)　順番に意味を持ちますが，間隔には意味を持たないので，順序尺度です．
(4)　番号の大小に本質的な意味を持たないので，名義尺度です．
(5)　知能指数が 0 でも，知能がないわけではないので，間隔尺度です．