統計情報処理入門

緒方裕光 編著
飯坂真司・清原康介
小西香苗・吉澤剛士 共著

建帛社
KENPAKUSHA

はじめに

現代では，ICT（電子通信技術）の進歩により，多くの人が数値，文章，画像，動画，記号など様々な情報を容易に収集し，電子媒体に保存できるようになっている。これらの情報は多種多様で，収集や発信の目的が不明確なものも多い。しかし，何らかの明確な目的を持って情報を利用しようとする場合には，情報の収集から分析，解釈までの全プロセスが合理的であることが求められる。本書では，この合理的プロセスのことを情報処理と呼ぶ。

情報処理の対象となるデータは，適切な分析を通じて有用な情報となる。すなわち，様々なデータを人間が使える情報とするために情報処理が必要となる。広い意味ではデータの収集から始まり，分析および分析結果の解釈までが情報処理に含まれるが，本書ではデータの分析方法に重点を置くことにする。

情報処理で扱うデータが数値の場合，その分析においては統計学が有力な方法論となる。特に数値データが観測や測定によって得られる場合には，統計的方法は必要不可欠である。本書では，統計的方法を用いて数値データを分析することを統計情報処理と呼ぶこととする。数値データの解析に統計的方法を用いる主な理由は，次の通りである。

まず，観測されたデータには必ず誤差が含まれるが，とりわけ人間集団を対象とした場合は非常に複雑な原因により誤差が発生する。誤差は偶然誤差と系統誤差に分かれ，統計学はこれらに対応する理論と方法を提供している。系統誤差の原因を探るには偶然誤差を分析に取り込む必要があり，大部分の統計的方法においてそのような偶然誤差を含めた分析モデルが使われている。また，観測データの多くは母集団の一部から取り出された標本データであり，統計学は，標本から母集団の状況を推測するための方法（推測統計，分析統計などと呼ばれる）が中心となっている。一般に，推測統計を行う前には，データの特徴を表現するための記述統計が有用であり，それはグラフ作成や指標の計算などを含んでいる。行政や一般事務では記述統計が多くの役割を持っており，その方法を学ぶことも重要である。さらに，観測データの多くは複数の変数から成り，そうした何種類ものデータの間に存在する複雑な関係を理解するには，多変量解析が有効である。統計学はこのような複雑な現象を取り扱う際にも，客観的かつ科学的な方法論となりうる。

本書は，特に人間集団の健康事象を観察して得られる数値データを取り扱うことを主題としている。人間集団の健康にかかわるデータを取り扱う統計学は，生物統計学，保健統計学，健康統計学などと呼ばれる。その分析結果は健康増進や疾病予防などのエビデンスとして用いられる。本書は，『疫学・健康統計学』（建帛社）の内容と関連しつつ，実際に読者がデータ解析を実行できるようになることを目標として書かれた。また，計算手段の一つとして，現時点では最も広く普及している「Microsoft EXCEL」（以下，Excel）を用いて基本的な統計計算ができるように，多くの例題を取り上げた。読者としては，人間集団のデータを扱う機会の多い栄養士・管理栄養士，保健学や栄養学を学ぶ学生・大学院生，保健に関わる技術者・行政職員などを想定している。数学的な表記は最小限にとどめ，わかりやすい表現になるように努めた。また，できる限り多くの例題を通じて計算ができるように心がけた。

本書は，６章により構成されており，第１章では，データ入力やデータの種類について，第２章では，Excel を使った基本的な計算方法について取り扱う。第３章と第４章では，Excel の基本的機能を使った記述統計の具体的方法について述べる。第５章では，主に分析ツールを使った分析統計として，区間推定，検定，回帰分析などについて述べる。第６章では，とくに健康にかかわる諸指標についてデータ分析の方法を述べる。また，付章では統計解析のフリーソフトである EZR についても少し紹介している。

統計学は，数学的な理論をもとにして成り立っているが，応用科学としての性格を持っている。すなわち，統計学は現実世界の様々な問題解決に適用されることで大きな意味を持つ。本書を通じて，多くの読者が統計的方法を実際のデータ分析に活用できるようになれば幸いである。

2024 年 8 月

著者を代表して　　緒方　裕光

目次

例題演習用 Excel データのダウンロードについて

本書に掲載の例題のサンプルデータ 付録 DATA を，建帛社ウェブサイトからダウンロードすることができます。あらかじめデータをダウンロードしておき，例題の解答を読みながら，実際にやってみましょう。

また，同じファイルには解答で作成したグラフなども収載しています。ご活用ください。

●例題サンプルデータのダウンロード方法

①建帛社ウェブサイト（https://www.kenpakusha.co.jp）の書籍検索から［統計情報処理入門］を検索します。

②本書の書籍詳細ページを開きます。

③書籍詳細ページにある「関連資料」より，ファイルをダウンロードしてください。

第1章　データを入力する

本章では，調査によって得られたデータの特性とその入力方法について学ぶ。近年では統計解析ソフトの操作性が良く統計手法も豊富であることから，統計解析に専用ソフトが用いられることも多いが，Microsoft Excel（以下 Excel と表記）にも統計解析のための豊富な関数が装備されている。ここでは，Excel を用いて調査データの入力を行う。Excel で入力されたデータは，主要な統計解析ソフトに取り込むことが可能なため，それらを用いることも念頭に入れたデータ入力を学ぶ。

1．データの種類

集めた調査データを取り扱う際には，いくつかの基本原則がある。この基本原則を理解するためには，まずデータの特性によって分類されたデータの種類を把握しておく必要がある。データは大きく量的データと質的データの２つに分けることができ，さらにそれぞれが２つに分けられ，合計４つの尺度に分類できる。

1）量的データと質的データ

量的データとは，その情報が数量的な意味を持つデータのことである。身長，体重，時間，気温などのように途切れることなく連続する連続データと，人数や個数のように飛び飛びの離散データに区別される。さらに，量的データは間隔尺度と比尺度の２つの尺度に分けることができる。

質的データとは，その情報が数量的な意味を持たないデータのことである。文字回答（テキストデータ）を除けば，上記の量的データ以外の全てのデータは質的データとなる。質的データは分類や種類を区別するためのデータで，何らかのカテゴリーを表すため**カテゴリーデータ**ともいわれる。質的データはさらに名義尺度と順序尺度の２つの尺度に分けることができる。

2）連続量と離散量

連続データは身長，体重，血圧のように，実験や観察から得られる多くのデータで，連続的な数値（**連続量**）として測定される。このように実測あるいは観測して得られる数値なので，**計量データ**ともいわれる。例えば，ヒトの体重は 52.35478…… kg と体重計の精度さえ許せば，原理的にはどこまでも細かく測定できる。このようなデータが連続データである。ただし，本来，連続量であっても，ある桁以下を四捨五入して，見かけ上離散量としてあらわされる場合もある。

離散データは飛び飛びの値（**離散量**）しかとらないデータのことで，測（量）るのではなく「数える」ので，**計数データ**とも呼ばれる。例えば，人の数，虫歯の数，歩数，心拍数などがある。これらは１，２，３，４のように整数の値をとり，0.5 人，1.5 本などと数えることはできない。

このように，量的データであっても離散データである場合もある。また，性別や摂取頻度などの質的データも，男性と女性の数，「食べない」「月に１-２回」「週に１回」「週に３-４回」と回答した数というように「数える」ので離散データとなる。

統計学ではデータが離散量であるか連続量であるかによって，データに適用される分布の種類や分布の特徴を示す表現方法などが異なる。

3）尺度

尺度とは，調査データがもつ数学的な特性に基づき評価する基準のことである。

(1) 名義尺度

質的データで，かつ順序に意味がなく，違いを識別することに意味がある尺度を，**名義尺度**という。名称自体には意味があるが，順序も大きさもないという特性をもっている。

図１-１は，生活習慣に関する調査票の例である。問１（性別）で得られる数字データは，男性を「０」，女性を「１」で表したとき，１＋２＝３あるいは１×２＝２という数式は意味をなさない。このように，名義尺度は加減乗除（足し算・引き算・掛け算・割り算）ができないデータである。問１（性別）や問２（所属学科）は，性別や所属学科の違いを識別する目的で割りつけられた数字である。

(2) 順序尺度

質的データで，かつ順序に意味がある尺度を**順序尺度**という。小さいものから大きいものへ，低いものから高いものへ，というように何らかの基準に従って順に並

生活習慣に関する調査　　ID_____

問１．性別　　名義尺度
　０．男性　　１．女性
問２．所属学科　　名義尺度
　１．看護学科　２．健康栄養学科　３．健康スポーツ学科
問３．出生の年月　　間隔尺度
　西暦(　　)年　(　　)月
問４．身長と体重　　比尺度
　身長(　　.　)cm　　体重(　　.　)kg
問５．朝食について，当てはまる番号を１つ選んでください。　　順序尺度
　１．ほとんど毎日食べる　　２．週 2-3 日食べない
　３．週 4-5 日食べない　　４．ほとんど食べない
問６．週に何日位お酒を飲みますか。当てはまる番号を１つ選んでください。　　順序尺度
　１．毎日　　２．週 5-6 日
　３．週 3-4 日　　４．週 1-2 日
　５．月に 1-3 日　　６．飲まない・飲めない

図１－１　生活習慣調査票における４つの尺度

べられているため，順序に意味をもつ特性がある。

図１－１の問５（朝食習慣）や問６（飲酒習慣）で得られる数字データは，頻度が多いもの「(ほとんど）毎日食べる（飲む)」から少ないもの「(ほとんど）食べない(飲まない)」へと順に並べられている。

順序尺度は，間隔（差）に意味を持たないデータであるため，足し算・引き算ができない。例えば，ほとんど毎日食べる「１」と週２－３日食べない「２」との間の間隔と，週４－５日食べない「３」とほとんど食べない「４」との間の間隔は，見かけの数値上は等しく「１」の差であるが，摂取頻度に関して等しい差とはいえない。

(3) 間隔尺度

量的データで，かつ絶対的ゼロ点を持たないデータを測定する尺度を，**間隔尺度**という。数値に絶対的ゼロ点を持たないとは，例えば年号はゼロ点があっても，それは歴史的な事情で決まっているので，絶対基準となる０が存在しないことになる。**図１－１**の問３（出生）の年月に回答された西暦 1950 年と 1955 年の差である「５」と 2000 年と 2005 年の差である「５」は等しいように，２つの差の数値が同じとき，差の元になっている２つの数値間の距離は等間隔である，というように間隔に意味をもつ特性がある。さらに，絶対的ゼロ点を持たない数値は，足し算・引き算はできるが，掛け算・割り算ができない特性をもっている。

（4）比尺度

量的データで，かつ絶対的ゼロ点をもつデータを測定する尺度を，**比尺度**という。通常は０以上の数値をとり，固有の単位を持ち，０は「何もない（無）」，つまり絶対的ゼロ点（原点）として特別な意味をもつ特性がある。**図１－１**の問４（身長と体重）は比尺度であり，このようなデータには血圧，年齢，血中マーカーの測定値，摂取栄養素量など，保健分野で扱うさまざまなものがある。

さらに，絶対的ゼロ点をもつ数値は，足し算・引き算だけでなく，掛け算・割り算もできるという特性がある。例えば，体重 50 kg と 10 kg の和の「60」や体重 90 kg と 20 kg の差の「70」には意味があり，さらに体重 30 kg と 60 kg の２つの数値の比は「２」（または１/２）であり，０という絶対的基準があるため，その比もおのずと意味をもつ。同様に，**図１－１**の問４（身長と体重）データから算出される BMI（体重（kg）/身長（m）^{2}）のように比尺度データ同士の比も意味をもつ。このように比に意味をもつのが比尺度の特性である。

2. データ入力の基本的形式

ここでは，Excel でデータ集計や解析を行うことを前提とした入力方法について解説する。統計解析においては，専用のソフトウェア（SAS，SPSS，STATA，JMP，R など）を利用することも多いが，ここで紹介する入力の仕方で作られた Excel のデータセットは，ほぼそのまま各ソフトウェアにインポート*することができる。

統計解析に利用する「データ」を作るために，まず調査票における質問文の回答形式と変数の作り方についての理解が必要である。**変数**とは，データ分析に用いるデータの項目のことを指し，質問文と変数の関係は質問文の回答形式によって異なる。

1）回答形式

質問に対する回答形式には，**選択回答形式**，**順位回答形式**，**自由回答形式**の３つがある（**表１－１**）。選択回答形式には，選択肢の中から一つだけ選ぶ**単一回答形式**と一つ以上選ぶ**複数回答形式**がある。選択回答形式における選択肢の一つひとつを**カテゴリー**と呼び，３つのカテゴリーの場合は３件法，４つのカテゴリーの場合は４件法などと呼ばれる。

＊ 各ソフトウェアで取り扱えるように形式を変更してデータを読み込むこと。

表1-1　回答形式

選択回答形式	単一回答形式	二項選択
		多項選択
	複数回答形式	制限付き
		制限なし（無制限）
順位回答形式		完全順位付け
		部分順位付け
自由回答形式		単語・数値・文章

また，個々の回答に対して，集計のために数値を与えることを**コーディング**と呼び，一般的には質問票の選択肢番号がコーディングされた数値となっている。**図1-2**に回答形式と変数名について示した。

(1) 単一回答形式

選択肢の中から一つだけ選ぶ**単一回答形式**には，２つの選択肢から一つ選ぶ**二項選択**と，３項以上の選択肢から一つを選ぶ**多項選択**がある。

図1-2の問１（性別）は，二つの選択肢から一つを選ぶ質問である（コラム p. 7）。疾病や症状の「有・無」，参加希望を「する・しない」，スクーリング検査の「陽性・陰性」，朝食欠食の「有り・無し」などのように，二つの選択肢に限定する回答形式が二項選択である。この二項選択の場合に，症状なし「０」・症状あり「１」，朝食欠食なし「０」・朝食欠食あり「１」のように「０・１」で数値をコーディングすることが多い。これは，解析において，二項ロジスティック回帰分析などに使用する際に便利であるからである。ちなみに，この０か１の値をとる変数のことを**ダミー変数**ともいう。

図1-2の問３（職業），問６（朝食頻度），問７（食生活満足度）は，三つ以上の選択肢から一つを選ぶ回答形式であるため多項選択と呼ばれている。同じ多項選択でも，問３（職業）は名義尺度であり，問６（朝食頻度）と問７（食生活満足度）には順序関係があり順序尺度となる。このように，選択肢に順序がある場合は**段階評価**とも呼ばれる。

段階評価の質問票では，一般的には質問票の選択肢番号がコーディングされた数値となっているが，場合によっては回答者の質問内容に対する意識的（時に無意識的）な回答バイアスを除くために，意図的なコーディングが行われていることもある。段階評価の項目において，他の項目とは評価の向きを逆にする**逆転項目**がそれである。

使用する解析ソフトによりそのデータ形式が異なる。近年，対象集団を追跡していくコホート研究のようなパネル調査（縦断調査）が行われるようになり，その実態や意識の変化などを時系列でとらえるパネルデータ分析も行われるようになってきた。このパネルデータ分析の際にはロング形式が用いられる。

ID	性別	身長 1回目	体重 1回目	BMI 1回目	身長 2回目	体重 2回目	BMI 2回目	身長 3回目	体重 3回目	BMI 3回目
1	1	175	79	25.8	175	75	24.5	175	70	22.9
2	2	155	61	25.4	155	58	24.1	155	57	23.7
3	1	182	86	26.0	182	79	23.8	182	77	23.2
4	1	171	73	25.0	171	71	24.3	171	67	22.9
5	2	162	76	29.0	162	70	26.7	162	66	25.1

ID	性別	測定時期	身長	体重	BMI
1	1	1	175	79	25.8
1	1	2	175	75	24.5
1	1	3	175	70	22.9
2	2	1	155	61	25.4
2	2	2	155	58	24.1
2	2	3	155	57	23.7
3	1	1	182	86	26.0
3	1	2	182	79	23.8
3	1	3	182	77	23.2
4	1	1	171	73	25.0
4	1	2	171	71	24.3
4	1	3	171	67	22.9
5	2	1	162	76	29.0
5	2	2	162	70	26.7
5	2	3	162	66	25.1

図１-９　ワイド形式（上図）
ロング形式（左図）

コラム：データ入力の前に決めておくこと

・入力作業の前に，同意記入欄に同意署名があるか，適切に回答されているかの確認作業が必要である。

・回答が無回答であった場合にどのように入力するかを決めておく。

　入力例：ブランク入力（セルに何も入力しない）

　入力例：特定の数字を入力（「９」，「999」など）

・必要に応じて，無回答と回答する必要がない設問（枝分かれ質問の回答）は区別して入力する。

　入力例：無回答は「99」入力，回答の必要がない場合はブランク入力など