1 多変量解析に向けて

1.1 母集団と標本

実験・調査を行うときの興味ある調査対象の集団を**母集団**（population）と呼ぶ．母集団が小規模であるならば，全数調査も可能である．例えば，ある小学校の 4 年 5 組 36 名における学力に関して調べたい場合である．しかし，日本の小学 4 年生の学力に関して調査する場合は，集団の規模が大きくなり全数調査には時間とお金が必要となるから実際に行うことは難しくなってくる．また，ある工場で製造される製品について良品か不良品かを調べたい場合なども全数調査は実質不可能であろう．全数調査が難しい場合においては，母集団から**標本**（sample）を**無作為抽出**（random sampling）によって取り出して調査する標本調査が行われる（図 1）．無作為抽出するのは，母集団の縮図となるように標本をとりたいからである．作為的に標本を選んだのでは，そこから得られる推論結果は母集団の特性と異なるものになることは想像に易い．無作為抽出法を用いる考え方は，昔から豚汁の味見に例えられる．鍋いっぱいに豚汁を作りその味を確認するとき，必ず鍋の豚汁を良くかき混ぜてから味見をするだろう．かき混ぜる前の豚汁は，上の方は味噌味が薄く，底の方は味噌味が濃い．つまり，確認したい味が偏っているのである．したがって，正しく味を推定するために良くかき混ぜるのである．この行為が無作為化（ランダム化）であり，その重要性が直感的に理解できるだろう．

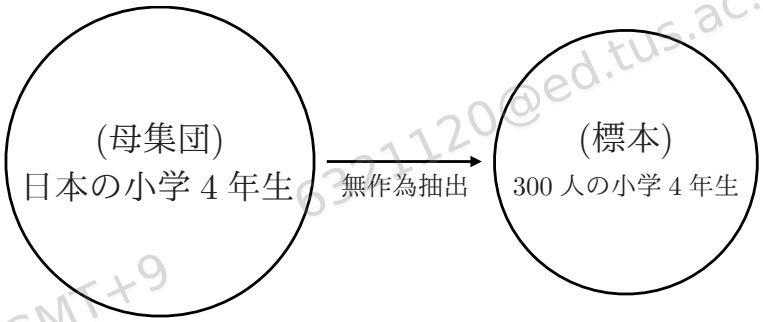

図 1: 標本調査

有限母集団と無限母集団

個体の個数が有限の母集団を有限母集団，個体の個数が無限の母集団を無限母集団という．有限母集団の場合は，復元抽出と非復元抽出とでは，一般には異なる結果になるため，有限母集団特有の議論が必要となる[a]．一方，無限母集団では，復元抽出のように，次の抽出に影響を及ぼさないと考える．また，有限母集団の場合でも，大規模母集団を対象にした標本調査では，有限母集団修正が近似的に1に等しく，多くの場合にその影響を無視できる．

[a]国沢清典, 確率統計演習 2 統計, 培風館, 1996

統計的な意味での興味の対象は，母集団それ自身ではなく，母集団の各要素の特徴を数量化した特性値 X である．例えば，小学4年生の学力試験の点数，工場で生産される製品の寿命，などである．特性値 X の値は母集団の各要素ごとに変動するので，X のことを母集団確率変数といい，その確率分布を**母集団分布**（population distribution）という．母集団分布を定める関数を $f(x)$ とする．ここでは，関数 $f(x)$ として確率密度関数や確率関数を想定する．母集団分布のもつ特性値を**母数**（population parameter）といい，**母平均**（population mean）や**母分散**（population variance）がその例である．ここに，母平均（μ と記す）は母集団分布の平均 $E(X)$ であり，母分散（σ^2 と記す）は母集団分布の分散 $V(X)$ である．

確率変数 $X_1, X_2, \ldots, X_n$ が互いに独立に同一の母集団分布に従うとき，**無作為標本**（random sample）という．ここに，n を**標本の大きさ**（sample size）という．**推測統計**（inferential statistics）の場合，母集団から無作為に取り出した母集団の一部であるサイズ n の標本の特性値を $x_1, x_2, \ldots, x_n$ とすると，各 x_i は観測するまで値がわからないから，母集団分布と同一の確率分布に従う確率変数 X_i の数値として x_i を捉える．数値 $x_1, x_2, \ldots, x_n$ を（確率変数 $X_1, X_2, \ldots, X_n$ の）**実現値**という．このことから，確率論を用いて推測に関する信頼性を見積もることが可能になる．一方，**記述統計**（descriptive statistics）の場合，数値 $x_1, x_2, \ldots, x_n$ を確率変数の実現値として捉えずに，数値 $x_1, x_2, \ldots, x_n$ からなるサイズ n のデータとして捉えてヒストグラムを描いたり，平均値や分散などを計算して全体の傾向や状況を調べる．

確率論と無作為標本

母集団 Ω の要素の数を N とする．母集団 Ω の個体 ω の特性値を $X(\omega)$ で表す．この母集団に対して，大きさ n の標本の復元無作為抽出という試行 T を考える．大きさ n の標本 $(\omega_1, \ldots, \omega_n)$ に対して，i 番目の個体 ω_i の特性値を $X(\omega_i) = X_i$ で表すと，標本から特性値の組 $(X_1, \ldots, X_n)$ が定まる．各 X_i は，試行 T の結果に対して値をとる変数であるから確率変数である[a]．この確率変数の組 $(X_1, \ldots, X_n)$ を，大きさ n の標本変量と呼ぶ．試行 T の標本空間は，Ω の n 個の直積 $\Omega \times \cdots \times \Omega$ であり，試行 T の標本点は，この直積の要素に対応する．

大きさ n の標本変量 $(X_1, \ldots, X_n)$ の性質について考える．復元無作為抽出は独立試行であるから，確率変数 $X_1, \ldots, X_n$ は互いに独立である．また，X_i $(i = 1, \ldots, n)$ の確率分布は明らかに X の分布（母集団分布）に一致する．本文中では，抽象的に大きさ n の無作為標本を定めた．ここで述べた標本変量 $(X_1, \ldots, X_n)$ から，無作為標本に要請された性質や確率論との関わりが見て取れる．

[a] 1 つの大きさ n の標本 $(\omega_1^*, \ldots, \omega_n^*)$ が与えられれば，サイズ n のデータ $(x_1, \ldots, x_n) = (X(\omega_1^*), \ldots, X(\omega_n^*))$ を得る

母集団と標本

教科書によっては，母集団は，個体の集まりではなく，個体の特性値の集まりとするものもある．例えば，日本の小学 4 年生の学力に関して調査する場合，日本の小学 4 年生ではなく，日本の小学 4 年生の学力試験の点数の集まりを母集団とする．同様にして，無作為抽出された 300 人の小学 4 年生ではなく，無作為抽出された 300 人の学力試験の点数を標本とする．これは，学力試験の点数という特性値を決めれば，関心があるのは学力試験の点数データのみだからである．このような考え方では，標本は，母集団から抽出された特性値 x_i の集まり $(x_1, \ldots, x_n)$ を意味する．

無作為標本 $X_1, X_2, \ldots, X_n$ を用いて推論を行う場合，無作為標本の実数値関数（母数を含んでいないもの）

$$T = t(X_1, X_2, \ldots, X_n)$$

を用いる．これを**統計量**（statistic）といい，その確率分布を**標本分布**（sampling distribution）という．

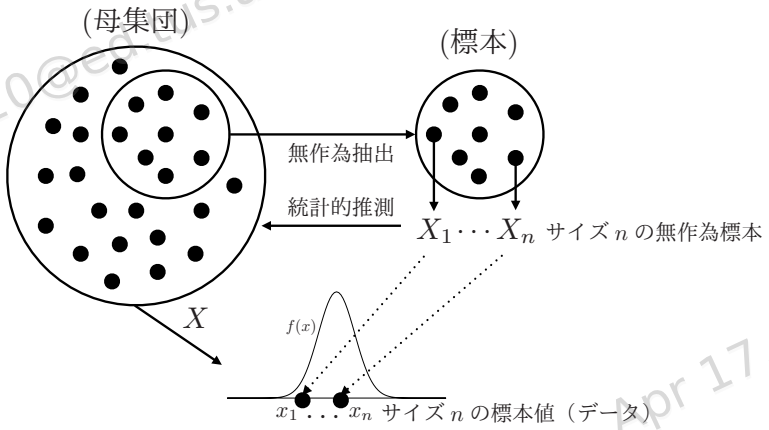

図 2: 無作為標本とデータ

1.2 1変量から多変量へ

統計学1，統計学2では，母集団分布が正規分布である場合を中心に学んだ．つまり，図2における X が正規分布に従う場合である．例えば，母集団を情報科学科の1年生とし，そこから無作為に選ばれた50名の期末試験の点数を調べることにする．このとき，情報数学1Bの期末試験の点数 X が正規分布に従うと仮定すれば，母平均に関する区間推定や仮説検定を行うことができる．そして，正規分布が統計的推測において重要な役割を果たすことも学んだ（中心極限定理など）．

多変量解析では，その名の通り，1変量だけでなく多変量を扱う．上の例を用いれば，情報数学1Bの期末試験の点数だけを考えるのではなく，その他の科目についても同時に扱うことを考える．つまり，情報数学1Bの期末試験の点数を X_1，解析学2の期末試験の点数を X_2，線形代数2の期末試験の点数を X_3 とし，確率変数ベクトル $\boldsymbol{X} = (X_1, X_2, X_3)^t$ が正規分布を拡張した確率分布に従うことを仮定する．そして，上で述べた枠組みを用いて区間推定，仮説検定を考える．講義の前半は，正規分布を多変量正規分布へ拡張することを考え，多変量正規分布が持ついくつかの性質を確認する．講義の後半は，多変量正規分布のパラメータ推定，線形重回帰分析などを学習する．

1.3 演習問題

問1 正規母集団 $N(\mu, \sigma^2)$ からの無作為標本を $X_1, X_2, \ldots, X_n$ とする．標本平均 $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ は，平均 μ，分散（ア）の正規分布に従う．

文中の（ア）に当てはまるものとして，次の ① ～ ④ のうちから適切なものを一つ選べ． **1**

① σ　② $\dfrac{\sigma}{n}$　③ σ^2　④ $\dfrac{\sigma^2}{n}$

問2 母集団 $\Omega = \{A, B, C, D\}$ から，大きさ2の無作為標本を取り出すことを考える．母集団 Ω から無作為に2個の個体を同時に取り出すとき，大きさ2の無作為標本が選ばれる確率はそれぞれ $1/6$ である．また，母集団 Ω から非復元無作為抽出で2個の個体を取り出すとき，大きさ2の無作為標本が選ばれる確率はそれぞれ（イ）である．さらに，母集団 Ω から復元無作為抽出で2個の個体を取り出すとき，大きさ2の無作為標本が選ばれる確率はそれぞれ （ウ）である．

1. 文中の（イ）に当てはまるものとして，次の ① ～ ④ のうちから適切なものを一つ選べ． **2**

   ① $1/4$　② $1/8$　③ $1/12$　④ $1/16$

2. 文中の（ウ）に当てはまるものとして，次の ① ～ ④ のうちから適切なものを一つ選べ． **3**

   ① $1/4$　② $1/8$　③ $1/12$　④ $1/16$