統計学基礎

第５回

今回のテーマ

- テキスト：99ページから110ページ
- 母集団と標本
- 統計的な研究の種類
- 標本調査と抽出方法
- 推定について
- 点推定

最初に補足：確率！

母集団と標本

- 母集団：分析の結果をあてはめる対象となる全体の集合、研究対象となっているすべての集合、対象の特性値（たとえば、身長や体重など）全体の集合

- 標本：母集団の部分集合

- 母集団のサイズがあまり大きくない場合は、全部を調べることも出来る。これを全数調査という。
- 一般的には費用や時間的な問題により全数調査は出来ない。その場合は、母集団から標本を得て、標本について調査・分析を行う。このような調査方法を標本調査という。

板書あり

例

1. ある大学において、学科の学生120人の履修科目数を調査する。全数調査可能であろう。

2. ある大学において、学生の通学時間を把握するために100人の学生を選んで通学時間を聞き取り調査する。母集団はその大学の学生全体の通学時間であるが、交通事情により日によって異なる。この場合、母集団を特定するのは難しく、全数調査は可能でない。標本は実際に得られた100人分のデータである。

3. 新しく開発されたある病気の治療薬の有効性を評価するため、その病気の患者50人にその薬剤を投与して病気が治癒するかどうかを調査する。母集団はその病気の患者全体であるが、未来における患者も含まれる。つまり全数調査は不可能である。標本は実際に薬を投与した50人である。

統計的な研究の種類（実験研究）

- 条件の設定が研究者自らの手でできるものを実験という。
- 新薬の効果を評価するための臨床試験やある種の肥料の効果を知るための農事試験なども実験である。
- 新薬の効果の評価では、患者を新薬を投与する群（実験群）と新薬ではない薬を投与する群（対照群）に分けて、その差を測定する。

板書あり

統計的な研究の種類（実験研究）

- 実験研究では以下に述べるフィッシャーの3原則が重要とされる。
  - 無作為化
  - 繰り返し
  - 局所管理
- 無作為化１：複数の銘柄のビールについて比較したい。この場合、被験者に飲んでもらうビールの順番など被験者によってランダムに変える。
- 無作為化２：新薬と既存薬について比較したい。この場合、どの患者にどちらの薬剤を投与するかランダムに決める。

統計的な研究の種類（実験研究）

- 繰り返し：データにはばらつきがある。ばらつきの大きさを見積もるために、その実験を繰り返さないといけない。人間を対象とした臨床試験では、個体差があるため多くの被験者に対するデータを取る必要があり、これも繰り返しとみなされる。

- 局所管理：実験の場を条件がなるべく均一ないくつかのブロックに分けることがある。例えば新薬の臨床試験で新薬を60代の男性に既存薬を20代の女性に投与して比較することは意味がない。このブロック化することを局所管理という。ブロックの設定では同じブロック内ではなるべく均一に、異なるブロック間での違いは大きめにというのが原則である。

統計的な研究の種類（観測研究）

- 実験が倫理的あるいは現実的に不可能なこともある。喫煙の調査をするために喫煙を強要できない。
- このような場合には観測実験を行う。
- 観測実験は、フィッシャーの3原則のうち、無作為化が成立しない。

例4

- ある健康飲料が健康に与える影響を調べるため、その飲料を通常飲むグループと飲まないグループとを比較した結果、飲料を飲まないグループの方が健康度が高かったとする。

- これは、その健康飲料に効果がないのではなく、健康に自信のない人がその飲料を飲んでいるためかもしれない。

標本調査と抽出方法

- 単純無作為抽出法：標本調査を行う最も基本的な方法である。n個の標本を得る時に、各個体が標本となる確率が等しく、またどのn個の個体の組も選択される確率が等しい。

- 系統抽出法：母集団から個体すべてに番号をつける。次に第1番目の個体を無作為に抽出する。第2番目以降は番号について同じ間隔で抽出する。

単純無作為抽出法と同じく 偏りなく 調査対象を 選択.
〃 に比べて 標本抽出の 手間が 少ない.

標本調査と抽出方法

- 層化無作為抽出法：いくつかの層（性別、年代、職業など）に分けて、層ごとにランダムに抽出する方法。 各層内は等質なので誤差分散小.
- 多段抽出法：全国学校調査では、いくつかの県を抽出し、各県からいくつかの学校を抽出する。そして最後にその学校から生徒を抽出する。このような手順で抽出する方法を多段抽出法という。 調査が大規模で、調査対象を直接抽出することが困難.（調査実施上の制約やコストの問題）
  - 段数が多くなると平均など推定の精度が悪くなる。その欠点を補うために層化多段抽出法がある。大都市、中規模都市などの層に分ける。
- クラスター（集落）抽出法：母集団を網羅した小集団を構成する。これをクラスターという。次にクラスターを抽出する方法である。そのクラスターに属する個体すべてが標本となる。

練習2と3に取り組む.

例5

- 男女5人ずつの計10人からなる母集団から4人を選ぶとき、男性から2人、女性から2人をランダムに選ぶとすると、各人の選ばれる確率はそれぞれ2/5であるが、「男性3人、女性1人」といった組が選ばれることはないので、この抽出法は単純無作為抽出ではない。単純無作為抽出であるためには10人全体から4人をランダムに抽出し、どの4人の組も選ばれる確率が等しくある必要がある。
  - このときの確率は$1/\binom{10}{4} = 1/210 = 0.0048$である。

推定について

地域住民（総数N=100,000）のうち、ある政策に賛成の人の母比率（割合）pを知りたい。そのため、大きさn=100の標本を無作為に抽出し賛否調査を行った。このうち、x人が賛成であったとする。標本比率x/nで母比率を推定する。

- 点推定：母比率の推定値として標本比率で推定する方法
- 区間推定：母比率の推定値として標本比率を含むある範囲で推定する方法
- ここでは、母比率、母平均、母分散についての推定について行う

板書あり.

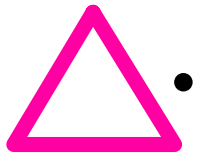
点推定

- 母平均μの推定量に$\hat{\mu}$とすることが多い。また標本平均として$\bar{X}$とすることが多い。
- 標本から求められる1つの値（たとえば、標本平均）で示す推定法を点推定とよぶ。

- 標本$x_1, x_2, \ldots, x_n$を大きさの順に並べ替えて$x_{n_1} \leq x_{n_2} \leq \cdots \leq x_{n_n}$としたものを順序統計量という。 x_{n_1}は最小値、 x_{n_n}は最大値である。

点推定

- 中央値m：母平均の推定値の一つであり、$n = 2k + 1$のとき$m = x_{n_{k+1}}$とし、$n = 2k$のとき$m = {\left(x_{n_k} + x_{n_{k+1}}\right)}/{2}$とする。



- 刈込み平均$\bar{x}_\alpha$：観測値を大きさの順に並べ替えて上下同じ数だけ観測値を取り除いて（刈込み）、残りの観測値の平均として定義されるもの。αは取り除く比率を表している。
  - たとえば、観測値が$1, \dots, 10$の10個だとし、$\bar{x}_{0.1}$は上下一つずつ取り除いた$2, \dots, 9$の平均である。
  - 正確には両側から$[n\alpha]$個ずつの観測値を取り除いて、残りの$n - 2[n\alpha]$個の観測値から平均を計算する。ここで実数aに対して$[a]$はaの整数部分を表している。
  - フィギュアスケートなどの採点で、9人の審判のうち最高点と最低点を除いて7人の平均点を求める方法は、刈込み平均の例である。

点推定

- 外れ値：観測値には、ときとしてほかの観測値と大きく異なる外れ値が発生する。誤記入、異質な観測値の混入など原因はさまざまである。小数個の外れ値によって母数の推定値が大きく変化することは好ましくない。

- 例6：$\{0,1,2,3,4,5,6,7,8,9,10,50\}$からなる標本がある。$(n = 12)$最後の数値が外れ値となる。この標本に対して平均$\bar{x}$、中央値$m$、刈込み平均$\bar{x}_{0.1}$を求めると次の通りである。

$$\bar{x} = 8.75, \qquad m = 5.5, \qquad \bar{x}_{0.1} = 5.5$$

点推定

- 前例のように平均$\bar{x}$は外れ値の影響を強く受ける。よって丁寧にデータを見ることがその利用の前提となる。多数の候補のうち、母平均の推定量として何を用いたらよいのか？

- 推定量として一致性や不偏性をみたすことが大事になる。

点推定

- 一致性：nが大きくなると母数の推定量が母数の値に収束することを一致性とよび、その推定量を一致推定量とよぶ。
  - 標本平均は一致推定量の一つで母平均に収束する。
  - 標本分散s^2や不偏標本分散（単に不偏分散）$\hat{\sigma}^2$も一致推定量で母分散σ^2に収束する。

- 標本ではnが有限である限り推定値$\hat{\theta}$のばらつきも無視できない。よって推定値のばらつきも一緒に表示することがよい。ただし、$\hat{\theta}$の 分散は未知なので、その推定量 $\hat{V}[\hat{\theta}]$ の平方根が用いられる。これを標準誤差とよぶ。$se(\hat{\theta})$と表す。

点推定

- 標本平均の場合、標本平均の分散は$V[\bar{x}] = \sigma^2/n$なので、σ^2をその推定値$\hat{\sigma}^2$で置き換え、その平方根$se(\bar{x}) = \hat{\sigma}/\sqrt{n}$を表示する。

  - $n = 100$の無作為標本を抽出し、標本平均$\bar{x} = 38.5$、標本分散$\hat{\sigma}^2 = 12.5^2$を得たとする。結果の報告には、母平均μの推定値$\bar{x} = 38.5$だけでなく、その標準誤差$se = 12.5/\sqrt{100} = 1.25$を合わせて記載することが望ましい。

点推定

- 不偏性：推定量$\hat{\theta}$の期待値が常に母数のその値θと等しくなる性質を不偏性とよぶ。その性質をもつ推定量を不偏推定量とよぶ。つまり$E\left[\hat{\theta}\right] = \theta$をみたす。また、不偏推定量でなければ、推定には偏りがある。その偏りを$E\left[\hat{\theta}\right] - \theta$で定義する。

- 標本平均$\bar{x}$は不偏性をみたす。
  - $E[\bar{x}] = \mu, \quad V[\bar{x}] = \frac{\sigma^2}{n}$が成立する。
- 不偏分散$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$も不偏性をみたす。

第6回の最初のトピックまでやる。

付録・標本平均の一致性・不偏性の証明

- 定理（チェビシェフの不等式）平均μ、分散σ^2とする確率変数Xに対して次が成立する。
  任意の$k > 0$に対して$P(|X - \mu| \geq k) \leq \sigma^2/k^2$をみたす。
  - この定理の証明は省略する。

- この定理より一致性を示すことができる。
- 標本平均$\bar{x}$に対して平均μ、分散σ^2/nなのでチェビシェフの不等式より次をえる。

$$0 \leq P(|\bar{x} - \mu| \geq k) \leq (\sigma^2/n)/k^2$$

そこで$n \to \infty$とすると、$P(|\bar{x} - \mu| \geq k) = 0$となる。どんなに小さな$k$についても成立するので$\lim_{n\to\infty} \bar{x} = \mu$となる。

- 不偏性については、「標本平均$\bar{x}$に対して平均μ」から明らかである。

付録・不偏分散の不偏性の証明

- 偏差平方和を$T_{xx} \coloneqq \sum_{i=1}^{n}(x_i - \bar{x})^2$とおく。
- $T_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}[(x_i - \mu) - (\bar{x} - \mu)]^2$
- $= \sum_{i=1}^{n}(x_i - \bar{x})^2 - 2(\bar{x} - \mu)\sum_{i=1}^{n}(x_i - \mu) + n(\bar{x} - \mu)^2$
- $= \sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{x} - \mu)^2$
- ここでT_{xx}の期待値をとる。
- $E[T_{xx}] = E[\sum_{i=1}^{n}(x_i - \mu)^2] - nE[(\bar{x} - \mu)^2]$
- $= \sum_{i=1}^{n} E[(x_i - \mu)^2] - nE[(\bar{x} - \mu)^2] = \sum_{i=1}^{n} \sigma^2 - n\frac{\sigma^2}{n}$
- $= n\sigma^2 - \sigma^2 = (n-1)\sigma^2$
- 両辺を$n-1$で割ると得られる。