母集団と標本

数学と医療・第 6 回

数学・橋本貴宏

2024 年 10 月 9 日 1 限

本日の授業では

- 前回のおさらい
  - (1) 正規分布の性質
  - (2) 正規分布の確率計算
- 母集団と標本
  - (1) 標本抽出
  - (2) 標本平均の分布
  - (3) 標本比率の分布

前回のおさらい

正規分布の定義

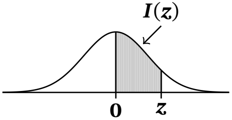
確率密度関数は $$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$N(\mu, \sigma^2)$ と表す．平均は μ, 分散は σ^2.

正規分布において $\mu = 0$, $\sigma = 1$ としたものを**標準正規分布**という．

確率の計算

標準正規分布の確率は標準正規分布表 (p.201) により求める．$I(z) = P(0 \leqq Z \leqq z)$ の値が記載されている．他の場合は，グラフの性質を用いて求める．



前回のおさらい (つづき)

一般の正規分布の確率を求めるには, $Z = \dfrac{X-\mu}{\sigma}$ と標準化して

$$P(a \leqq X \leqq b) = P\left(\frac{a-\mu}{\sigma} \leqq Z \leqq \frac{b-\mu}{\sigma}\right)$$

より求める.

- 確率が与えられていて, 横軸の値を求めたいときは, p.202 の「上側 100α % 点」の表を用いる.

二項分布の正規近似

二項分布 $B(n, p)$ は, $np \geqq 5$ かつ $n(1-p) \geqq 5$ のとき, 正規分布 $N(np, np(1-p))$ で近似することができる.

前回の演習問題

問題

赤ちゃん（女児）の出生時身長は, 平均 **48.5** (cm), 標準偏差 **2.5** (cm) の正規分布に従うことが知られている。このとき次を求めよ。ただし, 確率 (%) は整数値で, 身長は小数第 1 位までで答えよ。

(1) 身長が **47.0** (cm) 〜 **53.5** (cm) であるのは全体の何 % か。

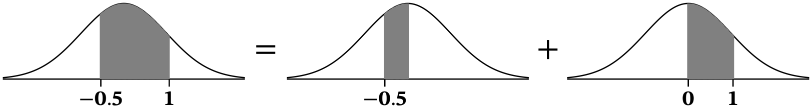
(2) 身長が **51.0** (cm) 以上であるのは全体の何 % か。

(3) 身長が高い方から **2.5** % の範囲内に入るのは, 何 (cm) 以上の赤ちゃんであるか。

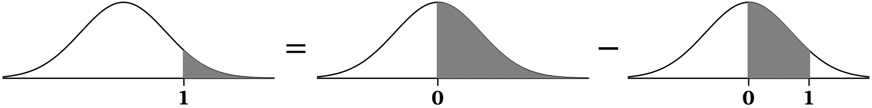
演習問題の解説 (その1)

(解) $\quad Z = \dfrac{X - 48.5}{2.5}$ と変換する.

(1) $P(47 \leqq X \leqq 53.5) = P\left(\dfrac{47 - 48.5}{2.5} \leqq Z \leqq \dfrac{53.5 - 48.5}{2.5}\right) = P(-0.6 \leqq Z \leqq 2) = I(0.6) + I(2) = 0.2257 + 0.4772 = 0.7029$. ゆえに 70% である。



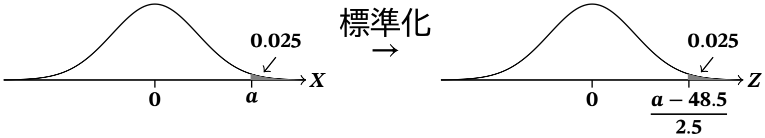
(2) $P(X \geqq 51) = P\left(Z \leqq \dfrac{51 - 48.5}{2.5}\right) = P(Z \geqq 1) = 0.5 - I(1) = 0.5 - 0.3413 = 0.1587$. ゆえに 16% である。



演習問題の解説 (その 2)

(3) 求めるのは $P(X \geqq a) = 0.025$ となる a の値である. 標準化すると

$$P(X \geqq a) = P(Z \geqq \frac{a - 48.5}{2.5}) = 0.025$$



$z_{0.025} = 1.960$ (上側 2.5 パーセント点; p.202) であるから $\dfrac{a - 48.5}{2.5} = 1.960$ を解くと, $a = 53.4$. 53.4 (cm) 以上.

みなさんからのコメント

- $(e^{f(x)})' = f'(x)e^{f(x)}$ はどんなところで使うのですか
  正規分布の密度関数の微分．大学入試レベルでは，ほとんど出てこないと思います．物理で e^r の微分は出てくるかも $(r^2 = x^2 + y^2 + z^2)$
- 半整数補正を考えた人，頭いいなと思いました
  角ばっているところを，半分の線を引いて補う．なかなか上手いですね
- 二項分布で，手計算では解けない問題を正規分布を用いてとくことができる点は数学の魅力だと思った
  上手く近似して，概算で求めていく．実生活でも活かせると思います
- 来週は、演習を解くプリント持ってきてほしいです！ あのプリントに毎週助けられてます。
  こういうコメントを書いてもらえると，絶対に忘れません

母集団 (p.79)

調査 (薬効など) 対象の全体を**母集団** (polulation) という．母集団に属する個々の要素を個体といい，母集団に属する個体の総数を**母集団の大きさ**という．

有限母集団と**無限母集団** (個数が非常に大きいとき)．

(I) 全数調査 (悉皆調査) ⋯ 対象の全部を調査する．

(例) 国勢調査は全数調査であるが，大変なコストがかかるので 5 年おきにやっている．

(II) 標本 (sample) 調査

母集団から一部を選び出して調査する．こちらの方が一般的である．

記述統計学と推測統計学

標本が選ばれたら，それらを要約し標本データの中身を分析する．「記述統計学」

「本当に知りたいのは母集団」

標本は一部分でしかない．そこから得られた情報から母集団のそれを推測する方法がある「推測統計学」

- **標本抽出** (sampling) $\Leftarrow$ 無作為抽出:　標本の大きさが $\boldsymbol{n}$ のとき，母集団のどの $\boldsymbol{n}$ 個の組み合わせも標本に選ばれる確率が等しく，どの標本も独立に取り出される．

標本抽出 (その 1) (p.81)

- **単純無作為抽出法**

最も基本的な抽出方法である. しかしながら, 母集団が大きい場合には非常に手間と時間がかかることがあるので, 母集団の特徴を反映させつつよりコストを軽減させた以下のような無作為抽出法がある.

(1) 系統抽出法
通し番号をつけた名簿を作成し、1 番目の調査対象を無作為に選び、2 番目以降の調査対象を一定の間隔で抽出する方法

(2) 層化無作為抽出法
母集団をあらかじめいくつかのグループ（層）に分けておき、各層の中から必要な数の調査対象を無作為に抽出する方法

標本抽出 (その 2)

(3) 多段抽出法

母集団をいくつかのグループに分け、そこから無作為抽出でいくつかグループを選び、さらにその中から無作為抽出でいくつかのグループを選び・・・という操作を繰り返して、最終的に選ばれたグループの中から調査対象を無作為抽出する方法

(4) クラスター（集落）抽出法

(a) 母集団を、小集団である「クラスター」に分ける
(b) 分けられたクラスターの中から、いくつかのクラスターを無作為抽出する
(c) それぞれのクラスターにおいて全数調査を行う

- 確率論の方法を用いる. ➡「単純無作為抽出法」でいく.

母集団分布と標本分布 (その1) (p.82)

母集団から大きさ $\mathbf{1}$ の標本を抽出するとき, その特性 $\boldsymbol{X}$ は確率変数と見なせる. この $\boldsymbol{X}$ が従う確率分布を**母集団分布**という.

有限母集団は, 大きさ $\boldsymbol{N}$ が非常に大きいときは**無限母集団**と見なせる.

母平均: $\boldsymbol{\mu} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_i$, **母分散**: $\boldsymbol{\sigma}^2 = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(x_i - \mu)^2$

これらをまとめて**母数** (parameter) と呼ぶ (記号 θ で表す).

$X_1, \cdots, X_n$: 母集団から無作為抽出された大きさ $\boldsymbol{n}$ の標本の確率変数.

X_i $(i = 1, \cdots, n)$ の関数 $T(X_1, \cdots, X_n)$ を標本**統計量** (statistic) という. 統計量は確率変数であり, その確率分布を**標本分布**という.

母集団分布と標本分布 (その 2)

(例 1)　大きさ $\boldsymbol{n}$ の標本 $\boldsymbol{X_1, \cdots, X_n}$ による統計量

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \quad S = \sqrt{S^2}$$

をそれぞれ**標本平均**, **標本分散**, **標本標準偏差**という.

(例 2)　標本分散に $\dfrac{n}{n-1}$ を掛けた統計量

$$U^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

を**不偏分散** (unbiased variance) という.

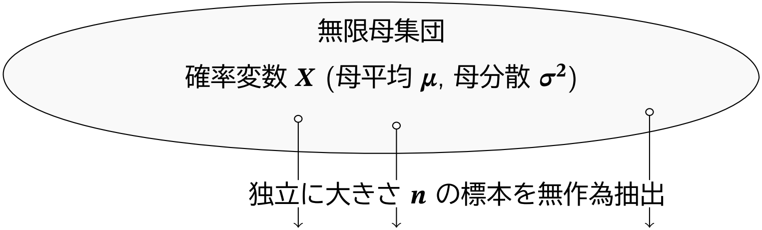
(注 1)　これらの呼び名は書籍によって異なることが多い.

(注 2)　Excel で標本分散 は VAR.P, 不偏分散は VAR.S である.

母集団分布と標本分布 (その3)

X_i の実際の標本の値を**実現値**といいい, x_i で表す．つまり, 標本平均の実現値は

$\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$. (確率変数は大文字, 実現値は小文字で表す)



確率変数: X_1 X_2 $\cdots\ \cdots$ X_n

実現値: x_1 x_2 x_n

統計量: $T(X_1, \cdots, X_n)$, 統計量の実現値: $T(x_1, \cdots, x_n)$

例題 4.1.1 (p.84)

大きさ $\boldsymbol{5}$ のデータ: $\mathbf{7.3,\ 8.3,\ 6.8,\ 8.6,\ 7.5}$ について, $\bar{\boldsymbol{x}}$, $\boldsymbol{s^2}$, $\boldsymbol{u^2}$ を求める.

(1) 手計算
(2) 電卓
(3) 表計算ソフト

のいずれかでやってみよう.

(1) $\boldsymbol{s^2 = \dfrac{1}{n}\sum_{i=1}^{5} x_i^2 - \bar{x}^{\,2}}$ で求めても良い.

(2) 関数電卓 fx-290A

取扱説明書の「13. 統計計算 (SD、REG)」に計算法が書いてある.

(3) Excel では, $\bar{\boldsymbol{x}}$ ⋯ AVERAGE, $\boldsymbol{s^2}$ ⋯ VAR.P, $\boldsymbol{u^2}$ ⋯ VAR.S で求める.

標本平均の分布 (その1) (p.84)

標本平均 $\overline{X}$ は統計量, すなわち確率変数である

定理 4.2.1　母平均 μ, 母分散 σ^2 の (無限) 母集団から抽出された大きさ n の標本 X_i $(i = 1, \cdots, n)$ の標本平均 $\overline{X}$ について, 次が成り立つ.

(1) $E(\overline{X}) = \mu, \quad V(\overline{X}) = \dfrac{\sigma^2}{n}$

(2) 母集団が正規母集団 $N(\mu, \sigma^2)$ ならば $\overline{X}$ は $N(\mu, \dfrac{\sigma^2}{n})$ に従う.

(証)　(1) 各 X_i は独立で, 母集団分布と同じ確率分布に従うので, $E(X_i) = \mu$, $V(X_i) = \sigma^2$ である. このとき

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu, \quad V(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{\sigma^2}{n}.$$

標本平均の分布 (その 2)

(2) を示すには正規分布に関する次の 2 つの性質を使う

- X が $N(\mu, \sigma^2)$ に従うとき, $aX + b$ は $N(a\mu + b, a^2\sigma^2)$ に従う
- **正規分布の再生性**: X_1, X_2 が互いに独立で, それぞれ $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ に従う $\Rightarrow X_1 + X_2$ は $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ に従う

これらにより, $\overline{X} = \dfrac{X_1 + \cdots X_n}{n}$ も正規分布に従うことがいえる.

この定理より, $\overline{X}$ の分散は n を大きくすると小さくなる (バラけていた分布が μ のまわりに集まってくる) ことがわかる.

中心極限定理 (p.72)

無限母集団からの標本について, 標本の大きさが十分大きいとき ($n \to \infty$ のとき) 標本平均 $\overline{X}$ は正規分布に近く.

定理 3.5.3 [中心極限定理; central limit theorem] 母平均 μ, 母分散 σ^2 の無限母集団から大きさ n の標本を独立に無作為抽出する. 標本平均 $\overline{X}$ を標準化した確率分布を

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

で定義すると Z は $n \to \infty$ のとき $N(0, 1)$ に従う.

- $\overline{X}$ は n が十分大きいとき近似的に $N(\mu, \frac{\sigma^2}{n})$ に従う, と考えることもできる.

例題 4.2.1 (p.85)

母集団は $N(\mu, \sigma^2) = N(240, 50^2)$ に従っている．定理 2 より，大きさ $n = 16$ の標本の標本平均 $\overline{X}$ は $N(240, \dfrac{50^2}{16})$ に従う．

$\overline{X}$ の標準化を Z とすると，$Z = \dfrac{\overline{X} - 240}{\sqrt{50^2/16}}$．よって求める確率は

$$P(\overline{X} \leqq 220) = P(Z \leqq \frac{220 - 240}{\sqrt{50^2/16}}) = P(Z \leqq \frac{-20 \cdot 4}{50})$$

$$= P(Z \leqq -1.6) = P(Z \geqq 1.6) = 0.5 - 0.4452 = 0.0548.$$

標本比率 (その 1) (p.86)

二項母集団: 母集団が２つのカテゴリーに分かれている

(例) ワクチン接種（接種済み, 未接種）, 内閣支持率（支持, 不支持） など

- 一方のカテゴリーを A とするとき, A に属する個体の割合 p を A の**母比率**という.
- 抽出された大きさ n の標本において, A に属する個体数が x であったとき, $\widehat{P}=\dfrac{x}{n}$ を**標本比率**という.

二項母集団の母集団分布は $B(1,p)$ である (属すか属さないかのとちらかだから).
大きさ n の標本 X_i について, A に属するとき $X_i=1$, 属さないとき $X_i=0$ とする. $Y=X_1+\cdots+X_n$ とおくと Y は $B(n,p)$ に従う.

- この後の議論は教科書のように中心極限定理で行っても良いが, 二項分布について成り立つ次の近似定理を使う

標本比率 (その 2) (p.86)

ド・モアブル-ラプラスの定理　確率変数 $\boldsymbol{X}$ が二項分布 $\boldsymbol{B}(\boldsymbol{n},\boldsymbol{p})$ に従っているとき, $\boldsymbol{X}$ を標準化変換した

$$\boldsymbol{Z} = \frac{\boldsymbol{X} - \boldsymbol{np}}{\sqrt{\boldsymbol{np}(\mathbf{1}-\boldsymbol{p})}}$$

の分布は, $\boldsymbol{n} \to \infty$ のとき標準正規分布 $\boldsymbol{N}(\mathbf{0},\mathbf{1})$ に従う. (証明は AIDLE-K に upload されている)

$\dfrac{\boldsymbol{Y}}{\boldsymbol{n}} = \dfrac{\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n}{\boldsymbol{n}}$ は標本比率 $\widehat{\boldsymbol{P}}$ であるから, この定理より,

$\boldsymbol{Z} = \dfrac{\boldsymbol{Y} - \boldsymbol{np}}{\sqrt{\boldsymbol{np}(\mathbf{1}-\boldsymbol{p})}} = \dfrac{\widehat{\boldsymbol{P}} - \boldsymbol{p}}{\sqrt{\frac{\boldsymbol{p}(\mathbf{1}-\boldsymbol{p})}{\boldsymbol{n}}}}$ は $\boldsymbol{n}$ が十分大きいとき近似的に $\boldsymbol{N}(\mathbf{0},\mathbf{1})$ に従う.

例題 4.3.1 (p.87)

無作為に選ばれた $\mathbf{150}$ 人のうち, 未接種である人の割合を $\widehat{\boldsymbol{P}}$ とする. $\boldsymbol{n} = \mathbf{150}$ は十分大きいと考えてよいから, $\widehat{\boldsymbol{P}}$ の標準化

$$\boldsymbol{Z} = \frac{\widehat{\boldsymbol{P}} - \mathbf{0.4}}{\sqrt{\frac{\mathbf{0.4 \cdot 0.6}}{\mathbf{150}}}} = \frac{\widehat{\boldsymbol{P}} - \mathbf{0.4}}{\mathbf{0.1} \cdot \sqrt{\mathbf{0.4 \cdot 0.4}}} = \frac{\widehat{\boldsymbol{P}} - \mathbf{0.4}}{\mathbf{0.04}}$$

は近似的に $\boldsymbol{N}(\mathbf{0}, \mathbf{1})$ に従う. よって求める確率は

$$\boldsymbol{P}(\widehat{\boldsymbol{P}} \geqq \mathbf{0.45}) = \boldsymbol{P}\left(\boldsymbol{Z} \geqq \frac{\mathbf{0.45} - \mathbf{0.4}}{\mathbf{0.04}}\right) = \boldsymbol{P}(\boldsymbol{Z} \geqq \mathbf{1.25})$$
$$= \mathbf{0.5} - \mathbf{0.3944} = \mathbf{0.1056}.$$

まとめ

- 母集団と標本

  標本のデータから, 母集団について (パラメータなど) の情報を得ることが「推測統計学」の目的である.

- 標本平均の分布

  正規母集団 $N(\mu, \sigma^2)$ からの標本の標本平均 $\overline{X}$ は $N(\mu, \dfrac{\sigma^2}{n})$ に従う.

- 標本比率の分布

  標本比率 $\widehat{P}$ は n が十分大きいとき, 近似的に $N(p, \dfrac{p(1-p)}{n})$ に従う.

10 月 9 日の演習問題

問題

令和 4 年度学校保健統計調査 (文部科学省) によると, ある県の 17 歳女子の身長は, 平均 $\mathbf{158.0}$ (cm), 標準偏差 $\mathbf{5.7}$ (cm) の正規分布に従うという。次の問いに答えよ。

(1) この中から無作為抽出で選んだ $\mathbf{15}$ 人の身長の平均を $\overline{\boldsymbol{X}}$ (cm) とするとき, $\overline{\boldsymbol{X}}$ はどのような確率分布に従うか。

(2) $\overline{\boldsymbol{X}}$ が $\mathbf{160.0}$ 以下である確率を求めよ。

次回

「標本分布」をやります．推定や検定で使う確率分布です

課題提出

AIDLE-K にアクセスし，「1009 課題提出」から提出する．提出期限は **10/12 23:59** とする．フィードバック (今日中) も忘れずに．

注意

- ファイル形式に注意し，拡張子をつける (jpg, pdf 等)
- 解答用紙にも名前を書くこと
- **過去のプリントで出していないものがある場合は，超過でもかまわないので出すこと**