確率・統計　後期　第2回

母集団と標本、統計量と標本分布

稲積　泰宏（いなづみ　やすひろ）

今日の内容

- 母集団と標本
- 統計量
- 標本平均
- 標本分散と不偏分散
- 標本分布

母集団と標本

母集団

- 調査対象全体の集合
- 例：日本全国の大学生全員の身長

標本

- 母集団から取り出した一部
- 例：100人の大学生の身長

なぜ標本を使うのか

- 母集団全体を調べるのは困難
- 標本から母集団の性質を推測する

無作為抽出

無作為抽出

- すべての個体が等しい確率で選ばれる
- 偏りのない標本を得るために重要

乱数表の使い方

1. 母集団の各個体に番号を付ける
2. 乱数表から数字を読み取る
3. その番号の個体を選ぶ
4. 必要な標本数まで繰り返す

統計量

統計量

- 標本から計算される値
- 例：標本平均、標本分散

重要な違い

- 母平均 μ、母分散 σ^2 → 定数
- 標本平均 $\overline{X}$、標本分散 S^2 → 確率変数

統計量は標本の取り方で値が変わる

標本平均

n 個の標本 $X_1, X_2, \ldots, X_n$ から計算：

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

標本平均の性質

期待値：$E[\overline{X}] = \mu$

分散：$V[\overline{X}] = \frac{\sigma^2}{n}$

標本平均の性質を理解する

期待値が μ になる理由

各 X_i の期待値は μ なので

$$E[\overline{X}] = \frac{1}{n}(E[X_1] + \cdots + E[X_n]) = \frac{1}{n} \cdot n\mu = \mu$$

分散が $\frac{\sigma^2}{n}$ になる理由

標本が独立なので

$$V[\overline{X}] = \left(\frac{1}{n}\right)^2 (V[X_1] + \cdots + V[X_n]) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

大数の法則

標本サイズ n を大きくすると

標本平均 $\overline{X}$ は母平均 μ に近づく

理由

分散 $V[\overline{X}] = \frac{\sigma^2}{n}$ に注目

n が大きくなると $\frac{\sigma^2}{n} \to 0$

分散が小さい = ばらつきが小さい = μ に集中

標本分散

n 個の標本から計算される標本分散 S^2：

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

特徴

- 標本平均からの偏差の2乗平均
- 直感的で分かりやすい

問題点

- 母分散を系統的に過小評価する

なぜ過小評価するのか

標本平均を使うことの影響

標本平均 $\overline{X}$ 自体が標本から計算されている

X_i は $\overline{X}$ に近づく傾向がある

真の中心 μ からの散らばりより小さくなる

```
真の中心μ
       ↓
1      2      3      4      5
       ●             ●
   標本          標本
       ↓
標本平均 x̄
```

自由度の考え方

データ {3, 4, ?} が3個あるとする

標本平均を計算すると

2個の値が決まれば、残り1個は自動的に決まる

実質的に自由に動かせるのは $n-1$ 個だけ

だから $n-1$ で割る

具体例で確認

母集団：{1, 2, 3, 4, 5}

- 母平均 $\mu = 3$
- 母分散 $\sigma^2 = 2$

標本 {2, 4} の場合：

- 標本平均 $\overline{X} = 3$
- 偏差の2乗和 = $(2-3)^2 + (4-3)^2 = 2$

n **で割る**: $\frac{2}{2} = 1$ （過小評価）

$n-1$ **で割る**: $\frac{2}{1} = 2$ （正しい）

不偏分散

n 個の標本から計算される不偏分散 U^2：

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

$n-1$ で割る理由

- 標本分散の偏りを補正する
- 母分散の不偏推定量になる

不偏推定量

- 期待値が推定したいパラメータに等しい

シミュレーションで確認

母集団 {1, 2, 3, 4, 5}（$\sigma^2 = 2$）

標本サイズ $n = 5$ で何度も抽出

割る数	分散推定値の平均	結果
n	約 1.6	過小評価
$n-1$	約 2.0	正しい

> $n-1$ で割ることで母分散を正しく推定できる

問6）

正規母集団の標本平均

母集団が正規分布 $N(\mu, \sigma^2)$ のとき

標本平均は標本サイズに関わらず正規分布に従う：

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

母集団が正規分布でない場合は中心極限定理が必要

例4）

問7)

標本分布

標本分布

- 統計量の確率分布
- 標本平均の分布を「標本平均の分布」という

中心極限定理

母集団の分布に関わらず、n が大きいとき

標本平均は近似的に正規分布に従う

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

中心極限定理：サイコロの例

サイコロを n 個振って平均を計算

$n = 1$（1個）

```
確率
1/6 │  ■  ■  ■  ■  ■  ■
    +--+--+--+--+--+--+
     1  2  3  4  5  6
```

一様分布

中心極限定理：サイコロの例

$n = 2$（**2個の平均**）

```
確率
    │       ▲
    │     ▲▲▲▲▲
    │   ▲▲▲▲▲▲▲▲▲
    +--+--+--+--+--+--+
     1  2  3  4  5  6
```

真ん中が高くなる

$n = 30$（**30個の平均**）

ほぼ正規分布になる

なぜ正規分布になるのか

極端な値が出にくい

- 30個全部が6 → 確率 $(1/6)^{30}$ ≈ 0%
- 30個全部が1 → 確率 $(1/6)^{30}$ ≈ 0%
- 大小混在 → よく起こる

実用的な意義

- 元の母集団が正規分布でなくても使える
- 目安：$n \geq 30$ で正規分布に近い

標本平均の標準化

標本平均を標準化すると標準正規分布に従う：

$$Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

メリット

- 標準正規分布表が使える
- 確率計算が簡単になる

例題1）

問8）

まとめ

- 統計量は確率変数
- 標本平均：期待値 μ、分散 $\frac{\sigma^2}{n}$
- 標本分散（n で割る）は過小評価
- 不偏分散（$n-1$ で割る）が正しい推定量
- 中心極限定理：n が大きいと正規分布

次回までに：Basic154-156

感想を会議のチャット欄へ