確率・統計の初歩 (4)

別所俊一郎

2006 年 4 月 26 日

統計学とは

- データを使って世の中のことを知るための科学
- 問題にしている母集団の分布の特性（平均や分散など）を知る手がかり
  - 悉皆調査は困難：費用、集計の時間、非回答者の存在
  - 標本を用いた統計的推測 statistical inference
- よく使う統計手法
  - 推定 estimation：未知の特性値の best guess の計算
  - 仮説検定 hypothesis testing：特性値についての真偽の判定
  - 信頼区間 confidence interval：ありそうな区間の推測

推定量の分布

- ある母集団の平均 μ_Y を知りたいとき、サンプルが i.i.d. であればその標本平均 $\bar{Y}$ を計算するのは自然
  - $\bar{Y}$ の，μ_Y の推定量としての性質はどうか？
  - **推定量 estimator**：標本から実数への関数のこと
  - **推定値 estimate**：得られた数値そのもの
- 推定量の標本分布が持つべき望ましい性質とは何か？
  - 推定量は標本から計算されるから確率変数であり，分布を持つ
  - 母平均の推定量は標本平均だけではない．推定量自体は自由に定義できる
  - そのうちでも「何らかの意味で真の値に近づいている」「真の値の近くに分布している」などの「望ましい性質」を持つものを探す

推定量 $\hat{\mu}_Y$ の望ましい性質

不偏性 unbiasedness 推定量の標本分布の平均が母平均に等しい

$$E[\hat{\mu}_Y] = \mu_Y$$

一致性 consistency 標本の大きさが大きくなるほど，真の値の周りの区間に入る確率が 1 へ近づく．推定量が母平均に確率収束する

$$\hat{\mu}_Y \xrightarrow{p} \mu_Y$$

効率性 efficiency 2 つの推定量 $\hat{\mu}_Y$ と $\tilde{\mu}_Y$ がともに不偏推定量であるとき，$\hat{\mu}_Y$ の分散が $\tilde{\mu}_Y$ の分散よりも小さいとき，$\hat{\mu}_Y$ のほうが効率的である，とよぶ.

標本平均 $\overline{Y}$ の性質

- 不偏性，一致性を持つ.

$$E[\overline{Y}] = \mu_Y, \quad \overline{Y} \xrightarrow{p} \mu_Y$$

- 線形不偏推定量のなかでもっとも効率的
  - 母平均の推定量としての Y_1 と比べると，$n \geq 2$ で明らかに効率的
  - そのほかのものよりも（Ex. 3.7）.

μ_Y の最小二乗推定量としての $\overline{Y}$

- 標本平均 $\overline{Y}$ は，各データと $\overline{Y}$ との差の 2 乗の平均を最小化しているという意味でデータにフィットしている
- いま，次のような問題を考える．

$$\min_m \sum_{i=1}^{n} (Y_i - m)^2$$

$Y_i - m$ は予測誤差のようなものだから，予測誤差の 2 乗和の最小化問題

- この問題の解が**最小二乗推定量 least squares estimator**
- 標本平均 $\overline{Y}$ はこの問題の解

$$\frac{d}{dm} \sum_{i=1}^{n} (Y_i - m)^2 = -2 \sum_{i=1}^{n} Y_i + 2nm = 0, \quad m = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

無作為抽出の重要性

- ランダムでないサンプリングでは，一般に標本平均は不偏性を持たない
- ある特性を持った観測値を overrepresent や oversample している可能性がある
- たとえば「労働力調査」のばあい
  - 層化 2 段抽出法による標本調査
  - 調査区を第 1 次抽出単位，住戸を第 2 次抽出単位として系統抽出
  - `www.stat.go.jp/data/roudou/pdf/10.pdf`

仮説検定の基礎用語

仮説検定 hypothesis testing 母集団の分布の特性値について仮説を立て，その真偽を統計的（確率的）に判定すること

帰無仮説 null hypothesis 検定したい仮説.「平均値が〜に等しい」といった形で定式化されることが多い．たとえば

$$H_0 : E[Y] = \mu_{Y,0}$$

対立仮説 alternative hypothesis 帰無仮説が成り立たないときの仮説．たとえば

$$H_1 : E[Y] \neq \mu_{Y,0}$$（両側対立仮説）

受容 accept，棄却 reject 帰無仮説をさしあたって支持することを受容，支持しないことを棄却という．受容とは，帰無仮説が真であるという強く宣言ではない．それゆえ，もとの H_0 自体を棄却されるべきものとして設定することが多い

p値 significance probability, p-value

- 実際に計算された標本統計量が仮説と等しい（$\overline{Y} = \mu_{Y,0}$）ことはほとんどない
  - $E[Y] \neq \mu_{Y,0}$ である
  - サンプリングによる誤差がある
- **p 値**：帰無仮説が正しいとしたときに，実際に得られた値より「離れた」値が得られる確率

$$\text{p-value} = Pr_{H_0}\left[|\overline{Y} - \mu_{Y,0}| > |\overline{Y}^{act} - \mu_{Y,0}|\right]$$

  - 第 1 種の過誤 type I error の確率
  - p 値が大きい：H_0 のもとで「起こりやすい」こと
  - 分布の裾 tail の積分値に等しい

p 値の計算

- $\overline{Y}$ の標本分布がなければ p 値は計算できない
  - 小標本なら標本分布は複雑だが，大標本なら正規分布で近似可能（CLT）
  - たとえば，$H_0 : \overline{Y} = \mu_{Y,0}$ のもとで

$$\overline{Y} \sim N(\mu_{Y,0}, \sigma^2_{\overline{Y}}), \quad \sigma^2_{\overline{Y}} = \sigma^2_Y / n$$

  - 標準化すれば $N(0,1)$ の分布関数さえあれば p 値は求まり，母集団分布の情報は不要
- では，母集団の分散 σ^2_Y はわかっているのか？

母集団の分散 σ_Y^2 既知のときの p 値

- 大標本ならば，$H_0 : \overline{Y} = \mu_{Y,0}$ のもとで

$$\overline{Y} \sim N(\mu_{Y,0}, \sigma^2_{\overline{Y}}), \quad \sigma^2_{\overline{Y}} = \sigma^2_Y / n$$

だから，標準化すると，

$$\frac{\overline{Y} - \mu_{Y,0}}{\sigma_{\overline{Y}}} \sim N(0,1)$$

それゆえ，

$$\begin{aligned} \text{p-value} &= Pr\left(\left| \frac{\overline{Y} - \mu_{Y,0}}{\sigma_{\overline{Y}}} \right| > \left| \frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_{\overline{Y}}} \right| \right) \\ &= 2\Phi\left(- \left| \frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_{\overline{Y}}} \right| \right) \end{aligned}$$

- しかし一般に母分散は未知.

標本分散，標本標準偏差，標準誤差

標本分散 s_Y^2 **sample variance** 以下の式で定義され，母分散の一致推定量

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

n ではなく，$n-1$ で除することを**自由度修正 degree of freedom correction** と呼ぶ．一致性は LLN を用いて証明する

標本標準偏差 s_Y **sample standard deviation** $s_Y = \sqrt{s_Y^2}$

標準誤差 standard error 標本平均の標準誤差とは，標本平均の分布の標準偏差の推定量のこと．

$$\hat{\sigma}_{\overline{Y}} \equiv SE(\overline{Y}) = s_Y/\sqrt{n}$$

母集団の分散 σ_Y^2 未知のときの p 値，t 値

- s_Y^2 は σ_Y^2 の一致推定量だから，i.i.d. な大標本について

$$\text{p-value} \quad = \quad 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{SE(\overline{Y})}\right|\right)$$

  と定義するのが自然.

- しばしば用いられてきた**検定統計量 test statistics** として**t 値**がある.

$$\text{t-value} \quad = \quad \frac{\overline{Y}^{act} - \mu_{Y,0}}{SE(\overline{Y})} \xrightarrow{d} N(0,1)$$

  大標本では t 統計量は標準正規分布に従う.

t 統計量　t-statistics

- 母集団分布が正規分布であれば，t 値は自由度 $n-1$ の t 分布に従うから，任意の n に対して exact に p 値を計算可能
- ただし，この授業では t 分布を用いない
  - 母集団分布が正規分布であるときしか t 分布は使えない．実際の経済データには近似がよくないため，適用可能性が小さい
  - 正規分布と t 分布の違いは，n が大きい限りにおいて小さいか，無視できる

有意水準 significance level

- 事前に規定した水準より p 値が小さいときに帰無仮説を棄却する，という手続きがしばしば採られてきた
  - たとえば，$p < 0.05$ で棄却すると決めておけば，$|t^{act}| > 1.96$ で帰無仮説を棄却
  - このような水準を有意水準とよび,「有意水準 5%で μ_Y は $\mu_{Y,0}$ と統計的に有意に statistically significantly 異なる」という
  - ただし，このような言い方では情報が少なくなるので，p 値を報告することがおおい
  - 用いる有意水準は 0.1, 1, 5%など．低い有意水準では帰無仮説を棄却しやすくなる
  - 片側対立仮説でも同様に考えることができる.