母集団・標本・標本抽出・推定・点推定

樋口さぶろお https://hig3.net

龍谷大学 先端理工学部 数理・情報科学課程

確率統計 I L09(2025-06-16 Mon)

最終更新: Time-stamp: "2025-06-16 Mon 11:34 JST hig"

今日の目標

- 母集団, 標本抽出, 推定を説明できる 久保川 統計学入門 §7.1
- 母平均値, 母比率, 母分散を点推定できる 久保川 統計学入門 §7.2
- 不偏推定量, 一致推定量の意味を説明できる

L08-Q1

Quiz 解答: 独立同一分布と中心極限定理

$n = 400$ が大きいと考えると, 中心極限定理より, S は近似的に正規分布 $N(n\mu, n\sigma^2)$ すなわち $N(40, 6^2)$ にしたがう. $Z = \frac{S-40}{6}$ は近似的に標準正規分布 $N(0, 1^2)$ にしたがう. よって, 求める確率は, $P(S > 31) = P(Z > -\frac{9}{6}) = P(-\frac{9}{6} < Z < +\infty) = \Phi(\infty) - \Phi(-\frac{9}{6}) = 1 - \Phi(-\frac{9}{6}) = 0.9332$.

L08-Q2

Quiz 解答: 独立同一分布と中心極限定理

$\mu = \mathrm{E}[X_i] = \frac{3+5}{2}, \sigma^2 = \mathrm{Var}(X_i) = \frac{(5-3)^2}{12}$.

$n = 400$ が大きいと考えると, 中心極限定理より, S は近似的に正規分布 $N(n\mu, n\sigma^2)$ すなわち $N(1600, \frac{400}{3})$ にしたがう. よって, $Z = \frac{S-1600}{20/\sqrt{3}}$ は近似的に標準正規分布 $N(0, 1^2)$ にしたがう. よって, 求める確率は,

$P(-\infty < Z \leq -5\sqrt{3}) = \Phi(-5\sqrt{3}) - \Phi(-\infty) = \Phi(-5\sqrt{3}) - 0 = 2.35 \times 10^{-18}$.

とりうる最小の値は $400 \cdot 3$ だから 1500 はありうる値だが，確率はこんなに小さい.

L08-Q3

Quiz 解答: ベルヌーイ分布の独立同一分布の和と中心極限定理

100 回中表の出る回数は，$Y = X_1 + \cdots + X_{100}$，$X_i \sim Bin(1, \frac{4}{5})$，独立同一分布，$\mathrm{E}[X_i] = \frac{4}{5}, \mathrm{Var}(X_i) = \frac{4}{5}\frac{1}{5}$．よって，$\mathrm{E}[Y] = 80$，$\mathrm{Var}(Y) = 4^2$ である (これは，$Y \sim Bin(100, \frac{4}{5})$ からも求められる)

$n = 100$ が大きいと考えると，中心極限定理より，Y は近似的に正規分布 $N(80, 4^2)$ にしたがう．

標準化された $Z = \frac{Y-80}{4}$ は近似的に標準正規分布 $N(0, 1^2)$ にしたがう．

よって，求める確率は，

$P(73 < X \leq 79) = P(-\frac{7}{4} < Z \leq -\frac{1}{4}) = \Phi(-\frac{1}{4}) - \Phi(-\frac{7}{4}) = 0.4599 - 0.0987 = 0.3612.$

ここまで来たよ

統計モデル 久保川 統計学入門 §7.1 確率統計 I(2025)L08

統計モデル 久保川 統計学入門 §7.1

$$\boxed{\text{構造}} + \boxed{\text{ゆらぎ＝確率変数}} \;\; \begin{array}{c}\text{標本抽出}\\ \longrightarrow \\ \text{推定} \\ \longleftarrow\end{array} \;\; \boxed{\text{観測データ}}$$

記述統計の母集団と標本，有限母集団 久保川 統計学入門 §7.1

某アイドルグループ全員 (→ 有限母集団,$N = 46$)
身長 x_i の平均値 $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ を求めたい!
某アイドルグループの滋賀県出身率 $p_1 = m/N$ を求めたい!
握手券を 5 枚買って質問する． 標本 $X_1, \dots, X_5$, $Y_1, \dots, Y_5$
標本平均値 $\frac{1}{5}(X_1 + \cdots + X_5)$，標本比率 $\frac{1}{5}(Y_1 + \cdots + Y_5)$. データ分析

- 母集団サイズ=$\boxed{46}$ の母集団から，
- 標本サイズ=$\boxed{5}$ の標本 (sample) を
- 標本の個数=$\boxed{1}$ 個を標本抽出する

推定

- 身長の母平均値 $\mu = \mathrm{E}[X]$ を求めたい.
- 全員の身長を知れば定義の式使うだけ.
- しかし, いま，全員の身長 (母集団) を知ることはできない.
- 握手券 46 枚買わず (実際は重複するかも), 5 枚で何とかすませたい.

推測統計の母集団と標本，無限母集団 久保川 統計学入門 §7.1

某アイドルグループの各メンバーの身長は，何かの分布にしたがう
滋賀県出身の正否は，ベルヌーイ分布 $Ber(p_1)$ にしたがう．
確率変数と思えば，何回でも引くことができる (46 を超えて)．(→ 無限母集団)
「確率分布 $Ber(p_1)$ にしたがう試行を 5 回する」
＝「5 個の確率変数 $Y_1, \ldots, Y_n$, i.i.d. $\sim Ber(p_1)$ の実現値を得る」(p_1 は不明)
「平均値」$\overline{Y} = \frac{1}{5}(Y_1 + \cdots + Y_5)$ も確率変数．

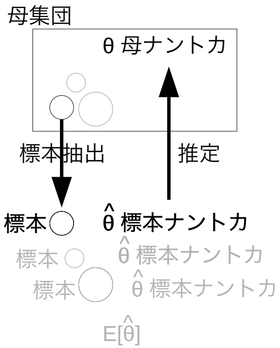
- 母集団サイズ＝$\boxed{+\infty}$ の母集団から，
- 標本サイズ＝$\boxed{5}$ の標本 (sample) を
- 標本の個数＝$\boxed{1}$ 個を標本抽出する

推定

- 分布の平均値を知りたい．
- 分布を知れば定義の式を使うだけ
- 分布を知ることはできない．
- 有限個の実現値から何とか「推測」

母集団・標本抽出・推定

- 母集団 population ＝考えたい集団．どんな分布，母平均値，母分散，などわかっていないことがあるが，全体を調べるわけにはいかない集団．パラメタ θ が備わる．
- 標本=sample (名詞) ＝母集団から ‘無作為に’ とってきた一部分
- 標本抽出する sample(動詞)=母集団から ‘無作為に’ とってくる ⇝ sampling (動名詞)
- 推定 する estimate(動詞) ＝標本を調べて母集団について，特に θ について，正しそうな事実を見つける ⇝ estimation (名詞)
- 確率変数 X, $\overline{X}$ 分布をもつ変数
- 実現値, 観測値 x, $\overline{x}$ 標本を 1 つとって確定した値
- 推定量 $\hat{\theta}(x)$ 母集団の量 θ を推定する式
- 点推定 1 個の値を提出する方式の推定



推定には誤差ある． 標本の選び方ごとに答は違う！

科目参加者全体から抽出した標本: 身長，滋賀県内高校

3 変量データ

- 身長 $X=$ 身長 (参加者)(cm) 量的データ 連続
- $Y=\begin{cases}1 & (\text{滋賀県高校 Yes})\\ 0 & (\text{No})\end{cases}$. 質的データ 名義尺度 久保川 統計学入門 p.34
- W A,B,C,D スクショする？ 質的データ 順位尺度 久保川 統計学入門 p.34

→ 導かれるデータ「身長高い」

- $W=\begin{cases}1 & (X\geq 170.7)\\ 0 & (\text{No})\end{cases}$

統計モデル

$X_i\sim$ 何かの独立同一分布.
$Y_i\sim Ber(p_1))$.
スクショ ∼ 多項分布
$W_i\sim Ber(p_2)$. X_i の言い換え.
こういう文脈での，ベルヌーイ分布のパラメタ，母平均値を，母比率 (ratio) という.

- $p_1=\mathrm{E}[Y]=P(Y=1)$.
- $p_2=\mathrm{E}[W]=P(X\geq 170.7)$.

母集団=回答者全体

1. 母集団サイズ 95 (クラス全体が回答すれば 118 だった)

サブチームに割り当られた標本

1. 1 サブチームに割り当てる標本の個数 1 個
2. 標本サイズ 様々

久保川 統計学入門 p.34

- よい標本抽出には実はいろいろなハイテクがある ⇝ 社会調査法, 社会統計学
- 本当は, 母集団が無限でないとき, 標本を作るのに重複選出を禁止するとき, 数学的にいろいろ複雑なことが起きるのだが, この授業では知らん顔する. ⇝ 確率・統計

ここまで来たよ

母平均値の (点) 推定 高校 数学 C 久保川 統計学入門 §7.2

組 $(X_1, X_2, \dots, X_n)$ はサイズ n の標本. 各 X_i は母平均値 $\mu = \mathrm{E}[X_i]$, 母分散 $\sigma^2 = \mathrm{Var}(X_i)$ の独立同一分布にしたがう確率変数.

定義 (標本平均値 久保川 統計学入門 公式 7.2)

$$\text{標本平均値 } \hat{\theta}(X_1, \dots, X_n) = \overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) = \text{先週の}U_n$$

母平均値 $\mu = \mathrm{E}[X_i]$ の 'よい' 推定量になっている. 'よい'≃ 不偏推定量

Pandas/Python では .mean(), Excel, Google スプレッドシートでは関数 average()

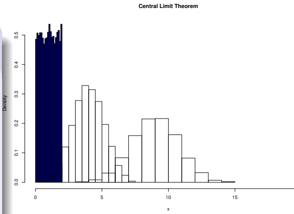
定義 (標本期待値)

$$g(X)\text{の標本期待値 } \overline{g(X)} = \frac{1}{n}(g(X_1) + \cdots + g(X_n))$$

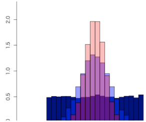
$\mathrm{E}[g(X_i)]$ の 'よい' 推定量になっている.

復習：中心極限定理 久保川 統計学入門 §7,4

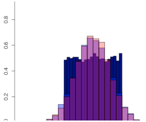
S



U



Z



定理 (中心極限定理 (いいかげんバージョン))

$X_1, \dots, X_n$ が母平均値 μ, 母分散 σ^2 の独立同分布に従うとき, $n \to +\infty$ で

- $S_n = X_1 + \cdots + X_n$, の確率分布は, 正規分布 $\boxed{N(n\mu, n\sigma^2)}$ に似る
- $U_n = \frac{1}{n}(X_1 + \cdots + X_n)$ の確率分布は, 正規分布 $\boxed{N(\mu, \sigma^2/n)}$ に似る
- 標準化した $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$ の確率分布は, $\boxed{\text{標準正規分布 } N(0, 1^2)}$ に似る

‘よい’ 推定量の資格 1

パラメタ θ を持つ分布の
母平均値 $\mathrm{E}[X]$，母期待値 $\mathrm{E}[g(X)]$ はひとつに定まっているが,
標本平均値 $\overline{X}$, $\overline{g(X)}$ は確率変数で, 試行=標本抽出のたびにかわる ($\overline{X}$ は確率分布をもつ)
一般に推定量 $\hat{\theta}(X_1, \dots, X_n)$ は確率分布を持つ.

定義 (推定量の不偏性 久保川 統計学入門 定義 7.1**)**

$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ が，パラメタ θ の不偏推定量であるとは，

$$\mathrm{E}[\hat{\theta}(X_1, \dots, X_n)] = \theta$$

が成り立つこと.

標本平均値 $\hat{\theta} = \overline{X}_n$ は，母平均値 $\theta = \mu$ の不偏推定量である: $\mathrm{E}[\overline{X}_n] = \mu$. 中心極限定理
また，$\mathrm{Var}(\overline{X}_n) = \sigma^2/n$. 中心極限定理

‘よい’ 推定量の資格 2

定義 (推定量の一致性 久保川 統計学入門 定義 7.5)

$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ が，パラメタ θ の一致推定量であるとは，

$$\forall \epsilon > 0 \qquad \lim_{n \to +\infty} P(|\hat{\theta}(X_1, \dots, X_n) - \theta| > \epsilon) = 0.$$

推定量と母ナントカに一定の差がある確率が, 標本サイズ n を大きくすると zero になること.

大数の法則 確率統計 I(2025)L08 より，標本平均値 $\hat{\theta} = \overline{X}_n$ は，母平均値 $\theta = \mu$ の一致推定量である.

L09-Q1

Quiz(確率変数としての標本平均値の分布)

母平均値 10, 母分散 6^2 の分布にしたがう母集団から, サイズ $n = 4$ の標本 $X_1, \dots, X_n$ を抽出し, 標本平均値 $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ を計算する.

1. $\overline{X}$ の母平均値を求めよう.
2. $\overline{X}$ の母分散を求めよう.
3. $\overline{X}$ を標準化する変換 $Z = \frac{\overline{X}-?}{?}$ を書こう.

L09-Q2

Quiz(母平均値, 母分散, 母比率の点推定)

フライドチキン屋さんのフライドチキンの大量の在庫 (=母集団) から, 無作為に 6 本のチキンを取り出したところ, 重さは次のようだった.

118g,　109g,　109g,　119g,　101g,　113g.

1. 重さの母平均値を点推定しよう.
2. 重さの二乗の母期待値を点推定しよう.
3. 重さの母分散を点推定しよう.
4. 105g 以上のものの母比率を点推定しよう.

記述上の注意　厳しく弾圧します. × き

- 母平均値 $= \mu = \mathrm{E}[X] \neq$ 標本平均値 $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$.
- 母分散 $= \sigma^2 = \mathrm{Var}(X) \neq$ 不偏標本分散 $V^2 = \frac{1}{n-1}(\cdots)$
- 母比率 $= p \neq$ 標本比率 $\hat{p} = \frac{k}{n}$.
- 「母ナントカを…と × 求めた○推定する」

比率=ratio 久保川 統計学入門 p.12

確率変数 $Y \sim Ber(p)$ ベルヌーイ分布, を考える.
こういう Y は, いろんな母集団を,「X は…である」という条件の成立不成立で 2 つに類別して作れる. 名義データ カテゴリ変数

- $X \sim$ ある分布, $Y = \begin{cases} 1 & (X \text{ の条件成立}) \\ 0 & (X \text{ の不成立}) \end{cases}$. 例 $X > 10$ なら $Y = 1$.
- 母集団=日本国民, その国民血液型が A であるなら $Y = 1$.

定義 (母比率)

$Ber(p)$ の p を母比率という. 母集団で X の条件「…」から $Y \sim Ber(p)$ を作ったとき, '母集団の「…である」ものの母比率 p', ともいう.

有限母集団なら,

$$\text{母集団の「…である」母比率 } p = \frac{\text{「…である」メンバー } x \text{ の個数}}{\text{すべてのメンバーの個数}} = \mathrm{E}[Y]$$

母比率の (点) 推定

定義 (標本比率)

ベルヌーイ分布にしたがう $Y_1, \dots, Y_n \sim Ber(p)$ に対しては，（推定量である）標本平均値を，特に標本比率 $\hat{p}$ という．$Y_i = 1$ のものが K 個とすると，

$$\hat{p}(Y_1, \dots, Y_n) = \frac{1}{n}(Y_1 + \cdots + Y_n) = \frac{K}{n}$$

標本比率は，母比率 p の不偏推定量，一致推定量

$$\frac{1}{n}[\underbrace{1 + \cdots + 1}_{k} + \underbrace{0 + \cdots + 0}_{n-k}] = \frac{k}{n}.$$

ここまで来たよ

8 中心極限定理・正規近似

9 母集団・標本・標本抽出・推定・点推定
- 母集団と標本
- 母平均値・母比率の (点) 推定
- 母分散の (点) 推定

母分散の (点) 推定

母分散 $\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2]$ もしょせん母期待値じゃん．データ分析で使った

$$\text{標本分散} S^2 = \frac{1}{n}[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2]$$

がよい推定値じゃないの？…ちょっと待った！
$\overline{X}$ の中には $X_1, \ldots, X_n$ がぜんぶ入ってるので，定数 μ であるかのような計算はできない．
分布の重心がちょっとずれるので修正が必要．

今後，その分散や分布を考えようと思っても，$(X_1 - \overline{X})^2, \ldots, (X_n - \overline{X})^2$ は互いに独立ではないから簡単ではない．

母分散の (点) 推定

定義 (不偏標本分散 久保川 統計学入門 公式 7.2)

$$\text{不偏標本分散 } \hat{\theta}(X_1, \dots, X_n) = V^2 = \frac{1}{n-1}[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2]$$
$$= \frac{n}{n-1}\left[\frac{1}{n}\sum_i X_i^2 - (\overline{X})^2\right]$$

不偏標本分散は母分散 σ^2 の ‘よい’ 推定値 不偏性 $\mathrm{E}[V^2] = \sigma^2$.
不偏標本分散は 久保川 統計学入門 p.155 の不偏分散のこと. 久保川 統計学入門 p.154 の標本分散は $n-1$ でなく n で割ったもの.
ここで, $\overline{X}$ は母平均値でなく, 上の標本平均値.
$n-1$ の理由 こうしないと不偏にならない
直観的理由 $\overline{X}$ は X_i の重心だから, μ より近くにある. $(X_i - \overline{X})^2$ は $(X_i - \mu)^2$ より小さくなりがち ($\frac{n-1}{n}$ 倍) なので修正.

不偏標本分散の不偏性の確認

$n = 2$. $\mathrm{Var}(X_i) = \sigma^2$. $\overline{X} = \frac{1}{2}(X_1 + X_2)$.

$$\begin{aligned}
\mathrm{E}[V^2] =& \mathrm{E}[\tfrac{1}{2-1}((X_1 - \overline{X})^2 + (X_2 - \overline{X})^2)] \\
=& \mathrm{E}[(X_1 - \tfrac{1}{2}(X_1 + X_2))^2 + (X_2 - \tfrac{1}{2}(X_1 + X_2))^2] \\
=& 2 \cdot \frac{1}{4}\mathrm{E}[(X_1 - X_2)^2] \\
=& 2 \cdot \frac{1}{4}\mathrm{E}[X_1^2 - 2X_1X_2 + X_2^2] \\
=& 2 \cdot \frac{1}{4}((\sigma^2 + \mu^2) - 2\mu\mu + (\sigma^2 + \mu^2)) \\
=& 2 \cdot \frac{1}{4}(2\sigma^2) = \sigma^2.
\end{aligned}$$

不偏標本分散は，Pandas/Python では .var(ddof=1), Excel では関数 var.s(), Google スプレッドシートでは, var(), var.s()s for sample(標本)
標本分散は，Pandas/Python では .var(), Excel では関数 var.p(), Google スプレッドシートでは, varp(), var.p()p for population(母集団)