

母集団と標本・母平均値/母分散/母比率の点推定

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 L11(2023-06-19 Mon)

最終更新: Time-stamp: "2023-06-19 Mon 12:06 JST hig"

今日の目標

- 母集団, 標本抽出, 推定を説明できる 岩薩林 確率・統計 §5.1,5.2
- 母平均値, 母期待値, 母分散, 母標準偏差, 母比率を点推定できる 岩薩林 確率・統計 §6.1,§7.1,§7.2,§7.3
- 標本平均値の分布を説明できる 岩薩林 確率・統計 §5.3



L10-Q1

Quiz 解答: 独立同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 $N(40, 6^2)$ $Z = \frac{S-40}{6}$ 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,

$$P(S > 31) = P(Z > -\frac{9}{6}) = P(-\frac{9}{6} < Z < +\infty) = F_Z(\infty) - F_Z(-\frac{9}{6}) = 1 - F_Z(-\frac{9}{6}) = 0.9332.$$

L10-Q2

Quiz 解答: 独立同分布と中心極限定理

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

$n = 400$ 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,

$$P(S \leq -5\sqrt{3}) = F_Z(-5\sqrt{3})$$

二項分布の正規近似 高校 数学 B I

L10-Q3

Quiz(ベルヌーイ分布の独立同分布の和と中心極限定理)

表が $\frac{4}{5}$, 裏が $\frac{1}{5}$ の確率で出る超いびつなコインを, 100 回投げる. 表が 73 回より多く 79 回以下で出る確率を求めたい.

標準正規分布の累積分布関数 $F_Z(z)$ を用いて表し, さらに `scipy.stats.norm.cdf()` (または正規分布表) を用いて小数値として近似的に求めよう.

二項分布にしたがう $Y \sim B(n, p)$ は, 独立同分布 (ベルヌーイ分布) にしたがう $X_i \sim B(1, p)$ ($i = 1, \dots, n$) の和とみなせるので, やはり正規近似できる.

岩薩林 確率・統計 第 5 章練習問題 1(3)

命題 (二項分布の正規近似 高校 数学 B)

二項分布 $B(n, p)$ は, $N(np, p(1 - p))$ で近似できる.

ここまで来たよ

- 10 中心極限定理・独立同分布の和の正規近似

- 11 母集団と標本・母平均値/母分散/母比率の点推定
 - 母集団と標本
 - 母平均値・母分散の(点)推定
 - 母比率の(点)推定

母集団と標本 (1) 無限母集団 特に 離散/連続型確率変数

岩薩林 確率・統計 §5.1.5.2

謎な仕組みのくじ (ひいたら戻す) の賞金額 (確率変数とも言う) の母平均値を, 結果から求めたい!

```
1 rvz=stats.norm(loc=0,scale=1) #N(0,1^2) 実は賞金は正規分布
2 rvz.rvs(size=5) # rvs=random variable sample
```

「確率変数 $Z \sim N(0, 1^2)$ にしたがう試行を 5 回して, 5 個の数値を得る」

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

無限母集団

数値は毎回生成されて何度でも引けるけど, 抽象的に, 無限個の数値のはいった袋があり, そこから数値に応じた確率で取り出していると考える. 全部を取り出すことはできない.

推定

- 値 (賞金) の母平均値 $\mu = E[X]$ を求めたい.
- 袋 (くじの仕組み = 正規分布) を知れば定義の式使うだけ.
- しかし, いま袋の中 ($f(x)$) を見ることはできない.
- $+\infty$ 回くじを買わず, 5 個で何とかすませたい.

母集団と標本 (2) 有限母集団

岩薩林 確率・統計 §5.1,5.2

某アイドルグループ全員 (→ 有限母集団) の身長 x_i の平均値

$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ を求めたい!

無作為にメンバーが1名出てきて質問できる握手券を5枚買って質問する
無作為にだれかの身長が返される試行

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

推定

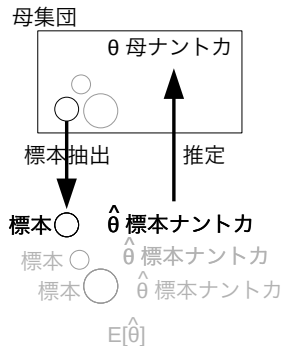
- 身長之母平均値 $\mu = E[X]$ を求めたい.
- 全員の身長を知れば定義の式使うだけ.
- しかし, いま, 全員の身長 (母集団) を知ることはできない.
- 46枚の握手券を買わず (実際は重複するかも), 5枚で何とかすませたい.

母集団・標本抽出・推定

岩薩林 確率・統計例 11(p.115)

- **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.
- **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 \rightsquigarrow sampling (動名詞)
- **推定**する estimate(動詞) =標本を調べて母集団について正しそうな事実を見つける \rightsquigarrow estimation (名詞)
- **確率変数** X , \bar{X} 分布をもつ変数
- **実現値, 観測値** x , \bar{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 を推定する量

岩薩林 確率・統計 図 p.109,115,137,167



推定には**誤差**ある. 標本の選び方ごとに答は違うし.

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
- W A,B,C,D スクショする? 質的データ (順序尺度)

母集団=回答者全体

- ① 母集団サイズ 99 (クラス全体なら 131 だった)

今回の標本

- ① 1 チームに割り当てる標本の個数 1 個
- ② 標本サイズ 10

ここまで来たよ

- 10 中心極限定理・独立同分布の和の正規近似

- 11 母集団と標本・母平均値/母分散/母比率の点推定
 - 母集団と標本
 - 母平均値・母分散の(点)推定
 - 母比率の(点)推定

母平均値の(点)推定

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

定義 (標本平均値) 岩薩林 確率・統計 (5.4)p.114

$$\text{標本平均値 } \bar{X} = \frac{1}{n}(X_1 + \dots + X_n) = \text{先週の } U_n$$

母平均値 $\mu = E[X]$ の 'よい' 推定量になっている.

Pandas/Python では `.mean()`, Excel では関数 `average()`

定義 (標本期待値)

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

$E[g(X)]$ の 'よい' 推定量になっている.

母期待値 $E[g(X)]$ はひとつに定まっているが, 標本平均値 \bar{X} , $\overline{g(X)}$ は確率変数で, 試行=標本抽出のたびにかわる (\bar{X} は確率分布をもつ)

よい(点)推定量がもつ性質

- 不偏性 (unbiased ナントカ)
推定量の母平均値は、推定したい母ナントカに等しい 岩薩林 確率・統計 p.141
- 一致性 (consistency)
推定量と母ナントカに一定の差がある確率は、標本サイズを大きくすると zero になる 岩薩林 確率・統計 p.143
- 最尤性 (maximum likelihood) 確率統計 II

標本平均値 \bar{X} の不偏性 岩薩林 確率・統計 p.113

母平均値 [母ナントカの推定量] = 母ナントカ

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

標本平均値 \bar{X} の一致性 大数の法則から 岩薩林 確率・統計 p.143

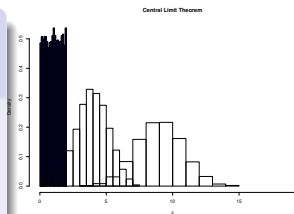
中心極限定理

定理 (中心極限定理 (いいかげんバージョン)) 岩薩林 確率・統計 定理 4.2(p.87)

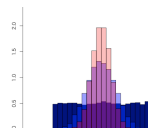
X_1, \dots, X_n が母平均値 μ , 母分散 σ^2 の独立同分布に従うとき, $n \rightarrow +\infty$ で

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

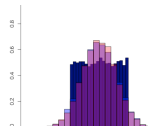
S



U



Z



L11-Q1

Quiz(確率変数としての標本平均値の分布)

正規分布 $N(10, 6^2)$ にしたがう母集団 (正規母集団) から, サイズ $n = 4$ の標本 X_1, \dots, X_n を抽出し, 標本平均値 $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ を計算する.

- ① \bar{X} の母平均値を求めよう.
- ② \bar{X} の母分散を求めよう.
- ③ \bar{X} を標準化する変換 $Z = \frac{\bar{X} - ?}{?}$ を書こう.
- ④ Z は標準正規分布にしたがう. 確率 $P(7 < \bar{X} \leq 13)$ を, $P(Z \text{ の不等式})$ に直して, 標準正規分布の累積分布関数 $F(z)$ で表そう.
- ⑤ 確率 $P(-c < Z \leq +c) = 1 - \alpha = 1 - \frac{1}{5}$ となるような c を, 標準正規分布の累積分布関数の逆関数 $F^{-1}(z)$ で表そう.

L11-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.

117g, 109g, 109g, 119g, 100g, 112g.

- 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 110g 以上のものの母比率を点推定しよう.

桁落ちに注意

数値計算法

記述上の注意

- 母平均値 $= \mu = E[X] \neq$ 標本平均値 $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$.
- 母分散 $= \sigma^2 = V[X] \neq$ 不偏標本分散 $S^2 = \frac{1}{n-1}(\dots)$.
- 母比率 $= p \neq$ 標本比率 $\hat{p} = \frac{k}{n}$.
- ここしばらくの問題で、「母ナントカを…と \times 求めた \bigcirc 推定する」

上でタイプの間違いは厳しく弾圧します. \times き

Moodle の練習問題スケジュール 今回から、練習問題の締切を月 09:15 に変更. それ以降が再練習問題.

統計検定自分で申し込んで好きなときに CBT 受験できる

<https://www.toukei-kentei.jp>

- 統計検定 3 級 = データ分析まで + この科目の一部
- 統計検定 2 級 = この科目まで + α . データサイエンスプログラム TM に勉強会記録あり.

母分散の(点)推定

定義 (不偏標本分散 岩薩林 確率・統計 (5.11) の $V(p.122)$)

$$\begin{aligned} \text{不偏標本分散 } S^2 &= \frac{1}{n-1} [(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2] \\ &= \frac{n}{n-1} \left[\frac{1}{n} \sum_i X_i^2 - (\bar{X})^2 \right] \end{aligned}$$

不偏標本分散は母分散 σ^2 の‘よい’推定値になっている。

不偏標本分散は 岩薩林 確率・統計 p.122 の不偏分散のこと。 岩薩林 確率・統計 p.113 の標本分散は $n-1$ でなく n で割ったもの。

ここで、 \bar{X} は母平均値でなく、上の標本平均値。

Pandas/Python では `.var(ddof=1)`, Excel では関数 `var.s()`

$n-1$ の理由 こうするとちょうど不偏: $E[S^2] = \sigma^2$ 。

直観的理由 \bar{X} は X_i の重心だから、 μ より近くにある。 $(X_i - \bar{X})^2$ 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

$$n = 2. \quad V[X_i] = \sigma^2. \quad \bar{X} = \frac{1}{2}(X_1 + X_2).$$

$$\begin{aligned} E[S^2] &= E\left[\frac{1}{2-1}((X_1 - \bar{X})^2 + (X_2 - \bar{X})^2)\right] \\ &= E\left[(X_1 - \frac{1}{2}(X_1 + X_2))^2 + (X_2 - \frac{1}{2}(X_1 + X_2))^2\right] \\ &= 2 \cdot \frac{1}{4} E[(X_1 - X_2)^2] \\ &= 2 \cdot \frac{1}{4} E[X_1^2 - 2X_1X_2 + X_2^2] \\ &= 2 \cdot \frac{1}{4} ((\sigma^2 + \mu^2) - 2\mu\mu + (\sigma^2 + \mu^2)) \\ &= 2 \cdot \frac{1}{4} (2\sigma^2) = \sigma^2. \end{aligned}$$

ここまで来たよ

- 10 中心極限定理・独立同分布の和の正規近似

- 11 母集団と標本・母平均値/母分散/母比率の点推定
 - 母集団と標本
 - 母平均値・母分散の (点) 推定
 - 母比率の (点) 推定

比率=ratio

岩薩林 確率・統計 p.107

確率変数 $Y \sim B(1, p)$ 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 岩薩林 確率・統計 p.107)

$B(1, p)$ の p . または母集団で X の条件「…」から $B(1, p)$ を作ったとき, 「母集団の「…である」ものの母比率」ともいう.

有限母集団なら,

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

やりたいこと: 母比率の推定

ベルヌーイ分布の p (母比率) を標本から推定したい!

- クラスの中で, 血液型 A 型の人々の比率は? n 人に質問しただけで推定したい.
- 候補者 A の得票率は何%? n 人に質問しただけで推定したい.
- 工場から出荷する製品のうち, 何% が不良品? n 個だけ抜き出して調査したい.
- このコインの表が出る確率は? n 回投げるだけで推定したい.

母比率の (点) 推定 岩薩林 確率・統計 p.115

定義 (標本比率 岩薩林 確率・統計 p.115)

標本のデータ n 個中 k 個が「…である」とき、

$$\text{標本比率 } \hat{p} = \frac{k}{n}$$

標本比率は「…」の母比率 p のよい推定値になっている。

母平均値 $E[Y]$ の推定値 = 標本平均値 \bar{Y}

$$= \frac{1}{n} \left[\underbrace{1 + \dots + 1}_k + \underbrace{0 + \dots + 0}_{n-k} \right] = \frac{k}{n} = \hat{p}.$$

岩薩林 確率・統計 問題 6(p.116)