母集団と標本・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(2022-06-20 Mon)

最終更新: Time-stamp: "2022-06-20 Mon 07:32 JST hig"

今日の目標

- 母集団, 標本, 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, 母比率を点推定できる

  岩薩林 確率・統計 §6.1,§7.1,§7.2,§7.3

L10-Q2

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 $\mathrm{N}(n\mu, n\sigma^2)$ すなわち $\mathrm{N}(40, 6^2)$ $Z = \frac{S-40}{6}$ は近似的に標準正規分布 $\mathrm{N}(0, 1^2)$ にしたがう. よって, 求める確率は,

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

L10-Q3

Quiz 解答: 独立同分布と中心極限定理

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

$n = 400$ 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 $\mathrm{N}(n\mu, n\sigma^2)$ すなわち $\mathrm{N}(1600, \frac{400}{3})$ 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 $\mathrm{N}(0, 1^2)$ にしたがう. よって, 求める確率は,

$P(S \leq -5\sqrt{3}) = F_{\mathrm{Z}}(-5\sqrt{3})$

L10-Q4

Quiz 解答: 二項分布と正規分布と中心極限定理

表の出る回数 X は, 二項分布 $\mathrm{B}(100, \frac{4}{5})$ にしたがう. $X_i = 0, 1$ を i 回目に表の出る回数とすると, $X = X_1 + \cdots + X_{100}$, X_i 独立同分布, $\mathrm{E}[X_i] = \frac{4}{5}, \mathrm{V}[X_i] = \frac{4}{5}\frac{1}{5}$.
よって, $\mathrm{E}[X] = 80$, $\mathrm{V}[X] = 4^2$ である.
$n = 100$ が大きいと考えると, 中心極限定理より, X は近似的に正規分布 $\mathrm{N}(80, 4^2)$ にしたがう.
標準化された $Z = \frac{X-80}{4}$ は近似的に標準正規分布 $\mathrm{N}(0, 1^2)$ にしたがう.
よって, 求める確率は,
$P(73 < X \leq 79) = P(-\frac{7}{4} < Z \leq -\frac{1}{7}) = F_{\mathrm{Z}}(-\frac{1}{4}) - F_{\mathrm{Z}}(\frac{7}{4}) = 0.4599 - 0.0987 = 0.3612$.

二項分布の正規近似 高校 数学 B

二項分布 $\mathrm{B}(n, p)$ は, 正規分布 $\mathrm{N}(np, p(1-p))$ で近似できる.

ここまで来たよ

10 中心極限定理と正規近似

11 母集団と標本・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) 有限母集団

岩薙林 確率・統計 §§5.1,5.2

某アイドルグループの身長ふたたび

- 某アイドルグループ全員 ($\to$ 有限母集団) の身長 x_i の平均値 $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ を求めたい!
  - メンバー 1 名を等確率で選んでくる, という試行を考えると, 確率変数 X の母平均値 $\mu = \mathrm{E}[X]$.
- メンバー全員分のデータがあれば定義の式使うだけ
- 握手会でメンバー 1 人ずつに質問しなければいけないとしたら?
- 握手会参加券 40 枚集めないで何とかすませたい.

$\rightsquigarrow$ 質問できたメンバー 5 人の身長 (=標本)(独立同分布にしたがう確率変数 $X_1, X_2, \cdots, X_5$) から推定したい.
5 人を '無作為に' 選ぶ (=標本抽出する)

母集団サイズ=46, 標本サイズ=5, 標本の個数=1.

母集団と標本 (2) 離散 or 連続型確率変数

岩薙林 確率・統計 §5.1,5.2

賞金額, 個数が謎のスピードくじ (引いて賞金額を見た後で箱に戻す).
賞金額 X は離散型確率変数 $\to$ 無限母集団 (何回でもひけるから).

- 賞金の母平均値 $\mu = \mathrm{E}[X] = \sum_x x \cdot p(x)$ を求めたい.
- くじの中を見れば ($p(x)$ の式を知れば) 定義の式使うだけ.
- しかし, 中を見ることはできない.
- $+\infty$ 回くじを買わず, 何とかすませたい.

$\rightsquigarrow$ 引いた 5 枚のくじの賞金額=標本)(独立同分布にしたがう確率変数 $X_1, X_2, \cdots, X_5$) から推定したい.
5 枚を '無作為に' 選ぶ (=標本抽出する).

母集団サイズ=$+\infty$, 標本サイズ=5, 標本の個数=1.

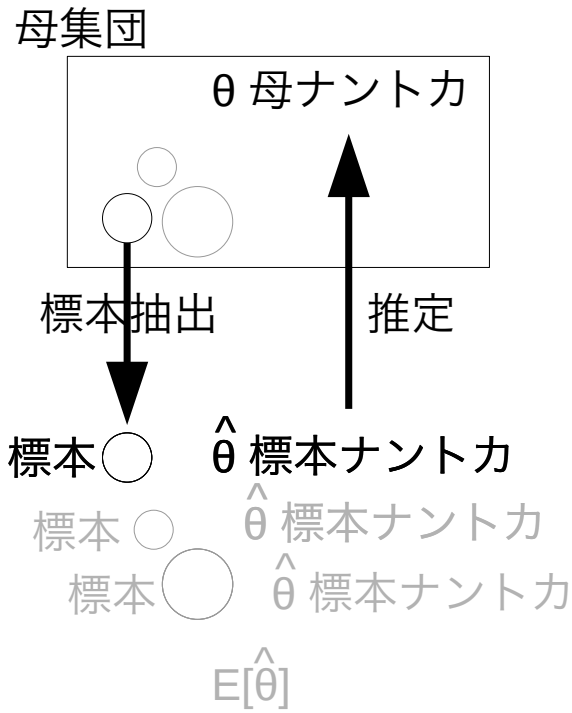
母集団・標本抽出・推定

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 ⇝ 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 ⇝ 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 を推定する量

推定には誤差ある. 標本の選び方ごとに答は違うし.

岩薩林 確率・統計 図 p.109,115,137,167



クラスから抽出した標本: 身長，滋賀県内高校

2 変量データ

- 身長 $X =$ 身長 (参加者)(cm)
- $Y = \mathrm{I}_{[\text{参加者の出身高校は滋賀県内}]}(\text{参加者}) = \begin{cases} 1 & (\text{Yes}) \\ 0 & (\text{No}) \end{cases}$.

母集団=クラスの回答者全体

1. 母集団サイズ 85 (クラス全体なら 115 だった)

今回の標本

1. 1 人に割り当てる標本の個数 1 個
2. 標本サイズ 10 から 16 くらい

ここまで来たよ

母平均値の (点) 推定

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

標本平均値

岩薙林 確率・統計 (5.4)p.114

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

が, 母平均値 $\mu = \mathrm{E}[X]$ の 'よい' 推定量になっている.

母平均値 μ はひとつに定まっているが, 標本平均値 $\overline{X}$ は確率変数であり, 試行=標本抽出のたびにかわる ($\overline{X}$ は確率分布をもつ)
Pandas/Python では .mean(), Excel では関数 average()

標本期待値

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

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

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 $\overline{X}$ 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

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

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

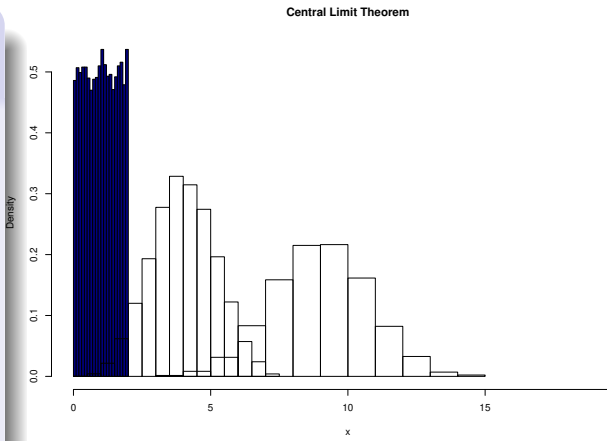
中心極限定理 岩薩林 確率・統計 定理 4.2(p.87)

中心極限定理 (いいかげんバージョン)

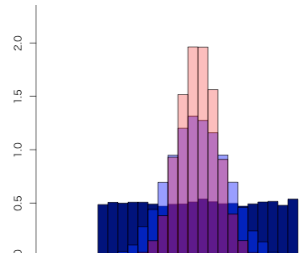
$X_1, \ldots, X_n$ が母平均値 μ, 母分散 σ^2 の独立同分布に従うとき, $n \to +\infty$ で

- $S_n = X_1 + \cdots + X_n$, の確率分布は,
  正規分布 $\mathrm{N}(n\mu, n\sigma^2)$ に似る
- $U_n = \frac{1}{n}(X_1 + \cdots + X_n)$ の確率分布は,
  正規分布 $\mathrm{N}(\mu, \sigma^2/n)$ 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 $\mathrm{N}(0, 1^2)$ に似る

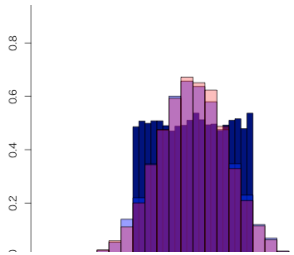
S



U



Z



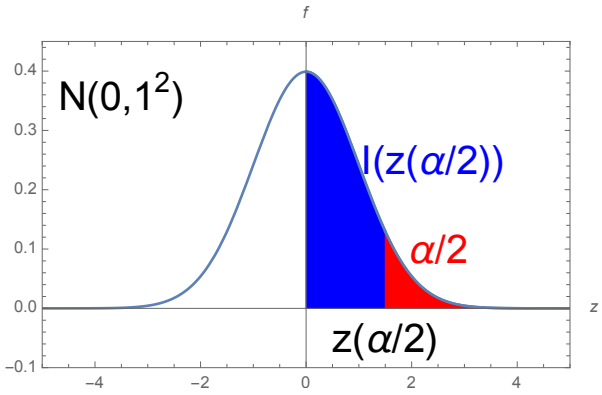
L11-Q1

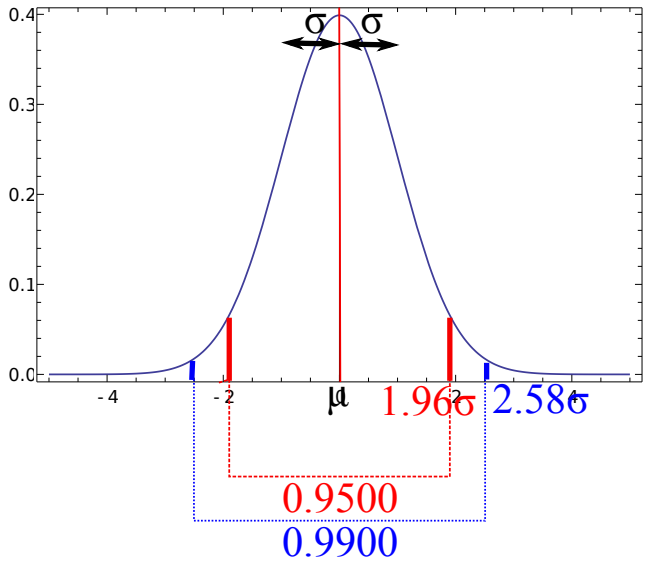
Quiz(確率変数としての標本平均値の分布)

正規分布 $\mathrm{N}(10, 6^2)$ にしたがう母集団 (正規母集団) から, サイズ $n = 4$ の標本 $X_1, \ldots, X_n$ 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.
4. Z は標準正規分布にしたがう. 確率 $P(7 < \overline{X} \leq 13)$ を, $P(Z$ 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.

標準正規分布 (ガウス分布) の確率 岩薩林 確率・統計 付表 1





標準正規分布の $z(\alpha)$

$Z \sim \mathrm{N}(0, 1^2)$ のとき, 上側確率 $P(Z \geq z(\alpha)) = \alpha$ で $z(\alpha)$ を定める.
$z(\alpha) = F_{\mathrm{Z}}^{-1}(1-\alpha)$.
$z(\alpha)$ =`scipy.stats.norm(loc=0,scale=1).ppf(`$1-\alpha$`)`, Excel では
$z(\alpha)$ =norm.s(1-α)

標準正規分布の確率密度関数は偶関数だから $z(1-\alpha) = -z(\alpha)$.

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.
$117\mathrm{g}, \quad 109\mathrm{g}, \quad 109\mathrm{g}, \quad 119\mathrm{g}, \quad 100\mathrm{g}, \quad 112\mathrm{g}.$

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 $= \mu = \mathrm{E}[X] \neq$ 標本平均値 $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$.
- 母分散 $= \sigma^2 = \mathrm{V}[X] \neq$ 不偏標本平均値 $S^2 = \frac{1}{n-1}(\cdots)$.
- 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 × 求めた○推定する」

上でタイプの間違いは厳しく弾圧します. × き

母分散の (点) 推定

岩薩林 確率・統計 (5.11) の V(p.122)

不偏標本分散

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

が, 母分散 σ^2 の 'よい' 推定値になっている.
ここで, $\overline{X}$ は母平均値でなく, 上の標本平均値.

Pandas/Python では .var(), Excel では関数 var.s()
$n-1$ の理由 こうするとちょうど不偏: $\mathrm{E}[S^2] = \sigma^2$.
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.

$n = 2$. $\mathrm{V}[X_i] = \sigma^2$. $\overline{X} = \frac{1}{2}(X_1 + X_2)$.
不偏標本分散の不偏性を確認.

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

ここまで来たよ

11 中心極限定理と正規近似

11 母集団と標本・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 $Y \sim \mathrm{B}(1,p)$ 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, 条件 $f(X)=$「X は…である」の成立不成立で 2 つに類別して作れる. カテゴリ変数

- $X \sim$ ある分布, $Y = \mathrm{I}_{[\ldots\text{である}]}(X)$, たとえば $X > 10$ 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 $Y = 1$.

母比率 岩薩林 確率・統計 p.107

$\mathrm{B}(1,p)$ の p. または母集団で条件 $f(x)$ から $\mathrm{B}(1,p)$ 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,

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

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

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,

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

が「…」の母比率 p のよい推定値になっている.

母平均値$\mathrm{E}[Y]$の推定値 =標本平均値$\overline{Y}$

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

岩薩林 確率・統計 問題 6(p.116)