母集団と標本・母平均値/母比率/母分散の点推定

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　L10(2024-06-24 Mon)

最終更新: Time-stamp: "2024-06-24 Mon 11:19 JST hig"

今日の目標

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

独立同分布にしたがう確率変数の和の正規近似 I

L09-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 $\mathrm{N}(n\mu, n\sigma^2)$ 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, $P(S > 31) = P(Z > -\frac{9}{6}) = P(-\frac{9}{6} < Z < +\infty) = \Phi(\infty) - \Phi(-\frac{9}{6}) = 1 - \Phi(-\frac{9}{6}) = 0.9332.$

L09-Q2

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

L09-Q3

独立同分布にしたがう確率変数の和の正規近似 II

Quiz 解答: ベルヌーイ分布の独立同分布の和と中心極限定理

100 回中表の出る回数は，$Y = X_1 + \cdots + X_{100}$, $X_i \sim \mathrm{B}(1, \frac{4}{5})$, 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, $\mathrm{E}[Y] = 80$, $\mathrm{V}[Y] = 4^2$ である (これは, $Y \sim \mathrm{B}(100, \frac{4}{5})$ 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 $\mathrm{N}(80, 4^2)$ 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 $\mathrm{N}(0, 1^2)$ にしたがう.

よって, 求める確率は,

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

チェビシェフの不等式の証明 (離散型)

岩薩林 確率・統計 §4.3(連続型での証明)

$$P(|X-\mu| \geq a\sigma) \leq \frac{1}{a^2} \quad \forall a > 0$$

「離れた x」特徴関数 $g(x) = \begin{cases} 1 & (|x-\mu| \geq a\sigma) \\ 0 & (|x-\mu| < a\sigma) \end{cases}$

とおく．意味を考えず $\mathrm{E}[g(X)\cdot(X-\mu)^2]$ を定義に戻って書くと，

$$\begin{aligned}
(a\sigma)^2 \times P(|X-\mu| \geq a\sigma) &= (a\sigma)^2 \sum_{|x-\mu| \geq a\sigma} p(x) \\
&= \sum_{x=-\infty}^{+\infty} g(x)\cdot(a\sigma)^2\, p(x) \\
&\leq \sum_{x=-\infty}^{+\infty} g(x)\cdot(x-\mu)^2\, p(x) \\
&\leq \sum_{x=-\infty}^{+\infty} 1\cdot(x-\mu)^2\, p(x) = \mathrm{V}[X] = \sigma^2
\end{aligned}$$

ここまで来たよ

9 中心極限定理・独立同分布の和の正規近似

10 母集団と標本・母平均値/母比率/母分散の点推定
- 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 $Z \sim \mathrm{N}(0, 1^2)$ 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 $\mu = \mathrm{E}[X]$ 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 岩薙林 確率・統計 §§5.1,5.2

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!
握手券で無作為にメンバーが 1 名出てきて教えてくれる身長を確率変数 X とする.
握手券を 5 枚買って質問する. $X_1, \ldots, X_5$ 独立同分布.

- 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

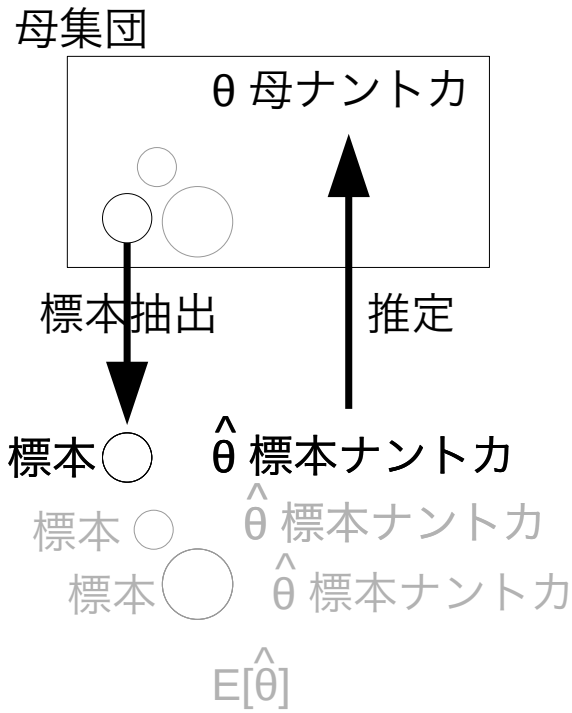
- 身長の母平均値 $\mu = \mathrm{E}[X]$ を求めたい.
- 全員の身長を知れば定義の式使うだけ.
- しかし, いま, 全員の身長 (母集団) を知ることはできない.
- 握手券 46 枚買わず (実際は重複するかも), 5 枚で何とかすませたい.

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, $\overline{X}$ 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

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 94 (クラス全体なら 120 だった)

今回の標本

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

ここまで来たよ

9 中心極限定理・独立同分布の和の正規近似

10 母集団と標本・母平均値/母比率/母分散の点推定
- 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.

定義 (標本平均値 岩薩林 確率・統計 (5.4)p.114)

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

母平均値 $\mu = \mathrm{E}[X_i]$ 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)

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

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

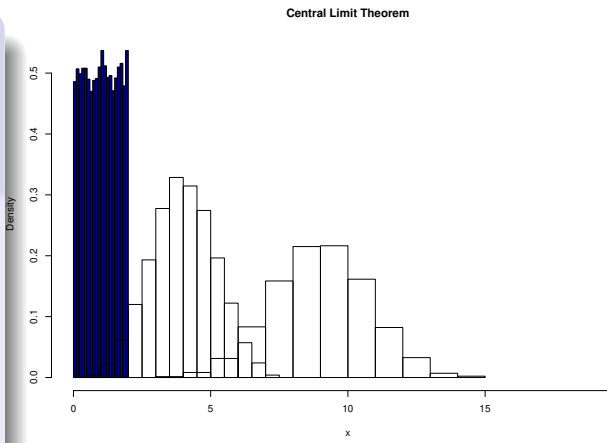
中心極限定理

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

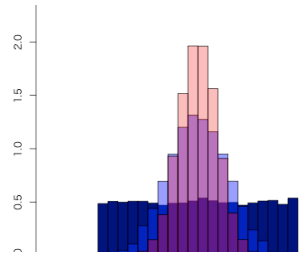
$X_1, \ldots, X_n$ が母平均値 μ, 母分散 σ^2 の独立同分布に従うとき, $n \to +\infty$ で

- $S_n = X_1 + \cdots + X_n$, 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

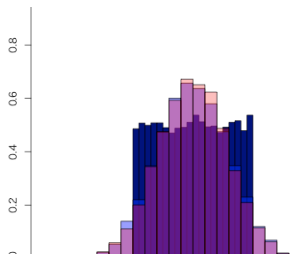
S



U



Z



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)

標本平均値 $\overline{X}$ の不偏性 岩薩林 確率・統計 p.113
標本平均値の分布の重心は母平均値

母平均値 [母ナントカの推定量] =母ナントカ

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

- 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

標本平均値 $\overline{X}$ の分布 $n \to +\infty$ では標本平均値の分布は正規分布に近づく．
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
標本平均値の分布の幅は σ^2/n くらい．

- 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 n を大きくすると zero になる 岩薩林 確率・統計 p.143

最尤性

- 最尤性 (maximum likelihood)
-

確率統計 II

L10-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 $X_1, \ldots, X_n$ 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.

L10-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 $= \mu = \mathrm{E}[X] \neq$ 標本平均値 $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$.
- 母分散 $= \sigma^2 = \mathrm{V}[X] \neq$ 不偏標本分散 $S^2 = \frac{1}{n-1}(\cdots)$.
- 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

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,「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)

$\mathrm{B}(1,p)$ の p を母比率という. 母集団で X の条件「…」から $Y \sim \mathrm{B}(1,p)$ 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 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.
$\mathrm{E}[Y] = p$ だから，母平均値のときのように標本平均値でよい.

$$\text{母平均値}\mathrm{E}[Y]\text{の推定値} = \text{標本平均値}\overline{Y}$$
$$= \frac{1}{n}[\underbrace{1 + \cdots + 1}_{k} + \underbrace{0 + \cdots + 0}_{n-k}] = \frac{k}{n} = \hat{p}.$$

岩薩林 確率・統計 問題 6(p.116)

ここまで来たよ

9 中心極限定理・独立同分布の和の正規近似

10 母集団と標本・母平均値/母比率/母分散の点推定
- 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 $\mathrm{V}[X] = \mathrm{E}[(X-\mu)^2]$ もしょせん母期待値じゃん．データ分析で使った

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

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 岩薩林 確率・統計 (5.11) の V(p.122))

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

不偏標本分散は母分散 σ^2 の 'よい' 推定値 一致性 $\mathrm{E}[S^2]=\sigma^2$.
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, $\overline{X}$ は母平均値でなく, 上の標本平均値.
$n-1$ 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.

不偏標本分散の不偏性の確認

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

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

不偏標本分散は，Pandas/Python では .var(ddof=1), Excel では関数 var.s()