母集団・標本・標本抽出と推定

樋口さぶろお　http://hig3.net

龍谷大学理工学部数理情報学科

確率統計☆演習 I　L10(2018-12-05 Wed)

最終更新: Time-stamp: "2018-12-05 Wed 14:51 JST hig"

今日の目標

- 母集団, 標本, 標本抽出, 推定の考えが説明できる
- 母平均値, 母期待値, 母分散, 母比率を点推定できる

L09-Q1
Quiz 解答:ベルヌーイ分布

1. ベルヌーイ分布 $\mathrm{B}(1, 0.05)$ に従う確率変数を X とすると,
   $Y = 1000X$.
2. $\mathrm{E}[Y] = \mathrm{E}[1000X] = 1000\mathrm{E}[X] = 1000p = 50.$(円)
   $\mathrm{V}[Y] = \mathrm{V}[1000X] = 1000^2\mathrm{V}[X] = 1000^2 p(1-p) = 47500.$ (円 ^{2})

L09-Q2
Quiz 解答:二項分布

1. 二項分布 $\mathrm{B}(100, \frac{2}{3})$ に従う確率変数を X とする. $P(X = 40)$ を求めればいいから, ${}_{100}\mathrm{C}_{40} p^{40}(1-p)^{100-40} = \frac{100!}{40!60!}(\frac{2}{3})^{40}(1-\frac{2}{3})^{60}$.
2. $\mathrm{E}[X] = n \times p = \frac{200}{3}$. $\mathrm{V}[X] = n \times p(1-p) = \frac{200}{9}$.

L09-Q3
Quiz 解答:二項分布の応用

1. $Y = 1000X + 50$. $\mathrm{E}[X] = 0.05$. $\mathrm{V}[X] = 0.05 \times 0.95$.
2. $\mathrm{E}[U_{20}] = 20\mathrm{E}[Y] = 20000\mathrm{E}[X] + 1000 = 2000$ (円).
   $\mathrm{V}[U_{20}] = 20\mathrm{V}[Y] = 20 \times 1000^2\mathrm{V}[X] = 950000$ (円 ^{2}).
3. w を 20 回中のあたりの回数とすると, W は二項分布 $\mathrm{B}(0.05, 20)$ にしたがう. $U_{20} = 1000W + 50 \times 20$.
4. $\mathrm{E}[U_{20}] = \mathrm{E}[1000W + 50 \times 20] = 1000\mathrm{E}[X] + 1000 = 1000 \times 20 \times 0.05 + 1000 = 2000$ (円). $\mathrm{V}[U_{20}] = \mathrm{V}[1000W + 1000] = 1000^2\mathrm{V}[W] = 1000^2 \times 20 \times 0.05(1 - 0.05) = 950000$ (円 ^{2}).
5. $P(U_{20} = 4000) = P(W = 3) = \frac{20!}{3!17!}0.05^3 0.95^{17}$.

ここまで来たよ

母集団と標本 (1) 有限母集団 前園確率統計 §5.1

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 (→ 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.

⇝ 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= $\boxed{}$, 標本サイズ= $\boxed{}$, 標本の個数= $\boxed{}$.

母集団と標本 (2) 離散 or 連続型確率変数 前園確率統計 §5.1

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 → 無限母集団 (何回でもひけるから).

- 賞金の母平均値 $\mu = \mathrm{E}[X] = \sum_x xf(x)$ を求めたい.
- くじの中を見れば ($f(x)$ 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= $\boxed{}$, 標本サイズ= $\boxed{}$, 標本の個数= $\boxed{}$.

母集団・標本抽出・推定 前園確率統計 §5.1

- 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)
- 推定量 母集団のパラメタをあてるために標本から作った量 (確率変数)

推定には誤差あるかも. 標本の選び方ごとに答は違うし.

ここまで来たよ

9 略解:二項分布, 独立同分布の和

10 母集団・標本・標本抽出と推定
- 母集団と標本
- 母平均値・母分散の (点) 推定
- 母比率とその (点) 推定

母平均値の (点) 推定高校 数学 B 前園確率統計 p.66,§5.1

$X_1, X_2, \ldots, X_n$ はサイズ n の標本.
各 X_i $(i = 1, \ldots, n)$ 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.
μ, σ^2 は母集団のパラメタ.

標本平均値 前園確率統計 p.71

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

が, 母平均値 μ の 'よい' 推定量になっている.

母平均値は μ はひとつに定まっているが, 推定量 (標本平均値) $\overline{X}_{(n)}$ は確率変数であり, 試行=標本抽出のたびにかわる ($\overline{X}_{(n)}$ は確率分布をもつ)

L10-Q1

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.

よい推定量がもつ性質 前園確率統計 §5.1

- 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
- 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
- 最尤性 (maximum likelihood) 確率統計 II

母平均値の推定量 $\overline{X}_{(n)}$ は不偏性を持つ

「標本平均値 $\overline{X}_{(n)}$」の不偏性

母ナントカの推定量の母平均値 ＝ 母ナントカ

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

先週の W_n の性質

独立同分布の性質 (復習) 前園確率統計 §3.4

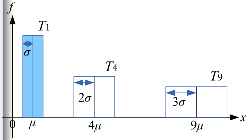
$X_1, \ldots, X_n$: 独立同分布. 母平均値 $\mathrm{E}[X_i] = \mu$, 母分散 $\mathrm{V}[X_i] = \sigma^2$.

和の確率変数 $U_n = X_1 + \cdots + X_n$

$$\mathrm{E}[U_n] = \sum_{i=1}^{n} \mathrm{E}[X_i] = n \times \mu.$$

$$\mathrm{V}[U_n] = \sum_{i=1}^{n} \mathrm{V}[X_i] = n \times \sigma^2$$

U_n の確率密度関数はこんな感じ?

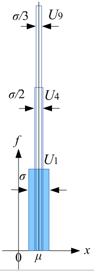


確率変数:
$W_n = \frac{1}{n} U_n = \frac{1}{n}(X_1 + \cdots + X_n)$ 前園確率統計 §3.2 定理 3.4

$$\mathrm{E}[W_n] = \mathrm{E}\left[\tfrac{1}{n} U_n\right] = \frac{1}{n} \times n \times \mu = \mu.$$

$$\mathrm{V}[W_n] = \mathrm{V}\left[\tfrac{1}{n} U_n\right] = \left(\frac{1}{n}\right)^2 \times n \times \sigma^2 = \frac{\sigma^2}{n}.$$

W_n の確率密度関数はこんな感じ?



L10-Q2

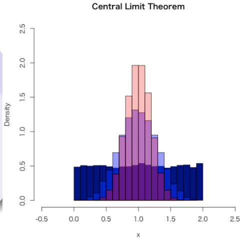
Quiz(独立同分布にしたがう確率変数の和)

ダーツで, 1 回投げるごとに点数が得られ, n 回投げた合計点で競うルールでプレイしている.
あるプレイヤーの i 回目の点数を確率変数 X_i とすると, X_i は互いに独立で同分布にしたがい, $\mathrm{E}[X_i] = 80, \mathrm{V}[X_i] = 200$ だという.
このプレイヤーが n 回投げた合計点を確率変数 U_n とする.

1. $\mathrm{E}[U_{10}]$, $\mathrm{V}[U_{10}]$ を求めよう.
2. $\mathrm{E}[\frac{1}{10}U_{10}]$, $\mathrm{V}[\frac{1}{10}U_{10}]$ を求めよう.
3. $\mathrm{E}[\frac{1}{100}U_{100}]$, $\mathrm{V}[\frac{1}{100}U_{100}]$ を求めよう.

母平均値の推定量 $\overline{X}_{(n)}$ は一致性を持つ

$\overline{X}_{(n)}$ と母平均値 μ が一定値離れている確率は，標本サイズ n が大きくなると 0 に近づく．



集合位相，確率統計 II

大数の (弱) 法則

$$\forall \epsilon > 0 \qquad \lim_{n\to+\infty} P(|\overline{X}_{(n)} - \mu| < \epsilon) = 1$$

確率的に収束

チェビシェフの不等式 (復習) と大数の (弱) 法則の証明

チェビシェフの不等式 前園確率統計 p.73

X を離散型または連続型確率変数とする. $\mu = \mathrm{E}[X]$: 母平均値,
$\sigma^2 = \mathrm{V}[X]$: 母分散
$a > 0$: 任意の正の実数.
このとき次が成立する.

$$P(|X - \mu| \geq a \times \sigma) \leq \frac{1}{a^2}$$

大数の弱法則の証明 前園確率統計 p.74 X として $\overline{X}_{(n)}$ (または W_n), $a = \frac{n\epsilon}{\sigma}$.

$$P(|\overline{X}_{(n)} - \mu| \geq \frac{n\epsilon}{\sigma} \times \frac{\sigma}{n}) \leq \frac{\sigma^2}{(n\epsilon)^2}$$

$$\lim_{n \to +\infty} P(|\overline{X}_{(n)} - \mu| \geq \epsilon) = 0$$

母期待値の推定

$Y = u(X)$ という新しい確率変数と思えば,

標本期待値

標本期待値

$$\overline{u(X)} = \frac{1}{n}(u(X_1) + \cdots + u(X_n))$$

が, 母期待値 $\mathrm{E}[u(X)]$ の ‘よい’ 推定量になっている.

母分散の (点) 推定 高校 数学 B 前園確率統計 p.75

不偏標本分散

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

が, 母分散の不偏な推定量になっている.

$n-1$ 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 $\overline{X}_{(n)}$ は X_i の重心だから, $(X_i - \overline{X}_{(n)})^2$ は $(X_i - \mu)^2$ より小さくなりがち ($\frac{n-1}{n}$ 倍) なので修正.

おぼえ方 (不偏) 標本分散は $n = 1$ のとき,

$\mathrm{E}[s^2] = \sigma^2$ を $n = 2$ のときに確認 (証明 前園確率統計定理 5.2)
標本平均値を $\overline{X} = \overline{X}_{(n)}$ と略記.

$$
\begin{aligned}
\text{左辺} =& \frac{1}{2-1}\mathrm{E}[(X_1 - \overline{X}_{(n)})^2 + (X_2 - \overline{X}_{(n)})^2] \\
=& \mathrm{E}[X_1^2 + X_2^2 - 2(X_1 + X_2)\overline{X}_{(n)} + 2\overline{X}_{(n)}^2] \\
=& \mathrm{E}[X_1^2 + X_2^2 - 2\overline{X}_{(n)}^2] \\
=& \mathrm{E}[X_1^2] + \mathrm{E}[X_2^2] - 2\mathrm{E}[\overline{X}_{(n)}^2]
\end{aligned}
$$

ここで,
$\sigma^2 = \mathrm{V}[X_1] = \mathrm{E}[X_1^2] - (\mathrm{E}[X_1])^2 = \mathrm{E}[X_1^2] - \mu^2$,
$\frac{\sigma^2}{n} = \mathrm{V}[\overline{X}_{(n)}] = \mathrm{E}[\overline{X}_{(n)}^2] - (\mathrm{E}[\overline{X}_{(n)}])^2 = \mathrm{E}[\overline{X}_{(n)}^2] - \mu^2$, より,

$$
\begin{aligned}
\text{左辺} =& (\mu^2 + \sigma^2) + (\mu^2 + \sigma^2) - 2(\mu^2 + \frac{\sigma^2}{2}) \\
=& \sigma^2 \\
=& \text{右辺}
\end{aligned}
$$

ここまで来たよ

9 略解:二項分布, 独立同分布の和

10 母集団・標本・標本抽出と推定
- 母集団と標本
- 母平均値・母分散の (点) 推定
- 母比率とその (点) 推定

比率=ratio 前園確率統計 §5.4

母集団 $\{X\}$ で確率 $p = P(X$ は$\cdots$である$)$ を考える.
$Y = \mathrm{I}_{[\cdots]}(X)$ とすると, $Y \sim \mathrm{B}(1, p)$ になる.

- $X \sim$ ある分布, $Y = \mathrm{I}_{[\cdots \text{である}]}(X)$, 例 $X > 10$ なら $Y = 1$ とか.
- 母集団=日本国民, 国民 X の血液型が A であるなら $Y = 1$.

母比率

確率 $P(X$ は$\cdots$である$)$ を, '母集団の「$\cdots$」の母比率', ともいう.
$Y \sim \mathrm{B}(1, p)$ の p に等しい

有限母集団なら,

$$\text{母集団の「}\cdots\text{」の母比率} = \frac{\text{「}\cdots\text{」である母集団のメンバー } x \text{ の個数}}{\text{母集団のメンバーの個数}} = \mathrm{E}[Y$$

例 母集団を, 条件「…である」の成立不成立で 2 つに類別すると,
{ 身長 165cm 未満, 身長 165cm 以上 }.

母比率 $p = \dfrac{\text{身長 165cm 未満の人の数}}{\text{母集団サイズ}}$.

例 Y{ サイコロの目が 1, サイコロの目が 1 以上 }. $Y \sim \mathrm{B}(1, p)$.
確率とも言えるけど, こういう状況では x の比率という習慣
ダミー変数と言われる確率変数 Y

やりたいこと:母比率の推定

母比率 p 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.

標本比率

標本比率 $\hat{p} = \overline{Y} = \overline{\mathrm{I}_{[\cdots]}(X)}$ が, 「…」の母比率の 'よい' 推定量になっている.
標本のデータ n 個中 k 個が「…」であるとき,

$$\text{標本比率 } \hat{p} = \overline{\mathrm{I}_{[\cdots]}(X)} = \frac{k}{n}$$

連絡

Moodle

https://learn.math.ryukoku.ac.jp/

Moodle モバイルアプリ

https://download.moodle.org/mobile

起動後, URL https://learn.math.ryukoku.ac.jp/moodle を登録

- 学期途中の振り返りのレポート. –2018-12-12 水. https://manaba.ryukoku.ac.jp
- trial に, 前回の問題と同種の問題を再出題してます (1/3 くらい)
- 予習復習問題を, 期限後も (再/初) 受験できます. 点数にはカウントしないけど, プチテスト準備に活用してね.
- Learn Math Moodle の予習復習問題は来週期限のものがあります. プチテストに備えてね.
- 教科書の中心極限定理 前園確率統計 §3.4, 区間推定 前園確率統計 §5.2,5.3,5.4 読んできてね.