母集団・標本抽出・推定

樋口さぶろお

龍谷大学理工学部数理情報学科

確率統計☆演習 I　L10(2015-12-04 Fri)

最終更新: Time-stamp: "2015-12-04 Fri 09:40 JST hig"

今日の目標

- 母集団, 標本, 推定の意味を説明できる
- 標本から母平均値を点推定できる
- 標本から母分散を点推定できる
- 標本から母比率を点推定できる

http://hig3.net

L09-Q1
Quiz 解答:多次元の確率変数の期待値

1. $\mathrm{E}[X+2Y] = 0\cdot(1+2\cdot 0) + \frac{2}{12}(2+2\cdot 0) + \frac{1}{12}(3+2\cdot 0) + \frac{4}{12}(1+2\cdot 2) + 0(2+2\cdot 2) + \frac{5}{12}(3+2\cdot 2) = \frac{62}{12}$.
2. $\mathrm{E}[\mathbf{1}_{[Y\geq 1]}(X,Y)] = 0\cdot 0 + \frac{2}{12}\cdot 0 + \frac{1}{12}\cdot 0 + \frac{4}{12}\cdot 1 + 0\cdot 1 + \frac{5}{12}\cdot 1 = \frac{9}{12}$.
3.

$$f_x^{\mathrm{X}} = \begin{cases} 4/12 & (x=1) \\ 2/12 & (x=2) \\ 6/12 & (x=3) \\ 0 & (\text{他}) \end{cases} \qquad f_y^{\mathrm{Y}} = \begin{cases} 3/12 & (y=0) \\ 9/12 & (y=2) \\ 0 & (\text{他}) \end{cases}$$

L09-Q2
Quiz 解答:離散型確率変数の独立性

1. $f_2^{\mathrm{X}} = f_4^{\mathrm{X}} = f_2^{\mathrm{Y}} = f_4^{\mathrm{Y}} = \frac{1}{2}$ であり，$f_x^{\mathrm{X}}\cdot f_y^{\mathrm{Y}} \neq f_{xy}^{\mathrm{XY}}$ であり独立でない.

2. $\mathrm{E}[X] = \frac{1}{2}\cdot 2 + \frac{1}{2}\cdot 4 = 3, \mathrm{E}[Y] = \frac{1}{2}\cdot 2 + \frac{1}{2}\cdot 4 = 3, \mathrm{E}[XY] = \frac{1}{2}\cdot 4 + 0\cdot 8 + 0\cdot 8 + 16\cdot\frac{1}{2} = 10, \mathrm{E}[X+Y] = 3+3 = 6, \mathrm{C_{XY}} = 10 - 3\times 3 = 1.$
$\mathrm{C_{XY}} \neq 0$ からも独立でないことがわかる.

L09-Q3

Quiz 解答:中心極限定理 T は母平均値が 10, 分散が $\frac{10}{3}$ の正規分布 $\mathrm{N}(10, \frac{10}{3})$ に従う.
U は母平均値が 1, 分散が $\frac{1}{30}$ の正規分布 $\mathrm{N}(1, \frac{1}{30})$ に従う.

ここまで来たよ

3 同時分布・確率変数の独立性・中心極限定理

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

母集団と標本 (1) 有限母集団

AKB48 の身長ふたたび

- AKB48 メンバー全員 (→ 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 $\mathrm{E}[X]$.
- 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 74 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) 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

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 $\mathrm{E}[X] = \sum_x f(x) \times x$ 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) 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

- 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

3. 同時分布・確率変数の独立性・中心極限定理

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

母平均値の (点) 推定

$X_1, X_2, \ldots, X_n$ 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

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

が, 母平均値 μ の ‘よい’ 推定値になっている.

母平均値は 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)

L10-Q1

Quiz(母平均値, 母分散の点推定)

フライドチキン屋さんのフライドチキンの在庫 (=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. 重さの母分散を点推定しよう.

よい推定値って?

標本平均値 $\overline{X}_{(n)}$ は不偏性を持つ

「標本平均値 $\overline{X}_{(n)}$」の母平均値$=X_i$ の母平均値

先週の $\mathrm{E}[U_n]=\mu$

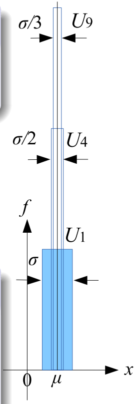
$$\forall n \qquad \mathrm{E}[\overline{X}_{(n)}]=\mu$$

標本平均値 $\overline{X}_{(n)}$ は一致性を持つ

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

大数の (弱) 法則

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



母分散の (点) 推定

(不偏) 標本分散

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

が, 母分散の 'よい' 推定値になっている.
ここで, $\overline{X}$ は母平均値でなく, 上のように計算した標本平均値.

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

おぼえ方 (不偏) 標本分散は $n = 1$ のとき, $\boxed{}$

$\mathrm{E}[S^2] = \sigma^2$ を $n = 2$ のときに確認

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

ここで,

$\sigma^2 = \mathrm{V}[X_1] = \mathrm{E}[X_1^2] - (\mathrm{E}[X_1])^2 = \mathrm{E}[X_1^2] - \mu^2$,

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

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

L10-Q2

Quiz(推定)

ある確率分布に従うスピードくじを 10 回ひいたところ, 賞金は,
0 円, 0 円, 0 円, 0 円, 0 円, 0 円, 10 円, 10 円, 30 円, 100 円
だった. 確率分布の母平均値と母分散と母標準偏差を推定しよう.

母標準偏差の (点) 推定

$$\text{母標準偏差の (点) 推定値} = \sqrt{\text{母分散の (点) 推定値 } S^2}$$

L10-Q3

Quiz(標本抽出と推定)

標本抽出と推定について, **正しい文の番号をすべて**答えよう.

1. 母平均値は, 標本平均値の推定値である.
2. 標本平均値は, 一般に, 標本抽出のたびに変化する
3. 不偏標本分散は, 母分散の推定値であり, 両者は必ずしも等しいわけではない
4. 標本平均値は, 一般に, 母平均値に等しい
5. 母分布 (母集団) が与えられたとき, 一般に, 標本のサイズは定まっている

ここまで来たよ

3. 同時分布・確率変数の独立性・中心極限定理

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

比率=ratio

データが「A 型である」「A 型でない」のような値を持つとき,

母比率

$$「\cdots」\text{の母比率 } p = \frac{「\cdots」\text{であるデータの個数}}{\text{母集団サイズ}} = \mathrm{E}[Y]$$

例 { 身長 165cm 未満, 身長 165cm 以上 }.

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

例 Y { サイコロの目が 1, サイコロの目が 1 以上 }.

母比率 $p =$ サイコロの目の 1 がでる確率

確率と言いそうになるけど, この文脈では比率という

ダミー変数と言われる確率変数 Y

A 型	Y	確率
「である」	1	p
「でない」	0	$1-p$

やりたいこと:母比率の推定

- 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.

ここまで来たよ

3. 同時分布・確率変数の独立性・中心極限定理

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

母比率の (点) 推定

標本比率

サンプルのデータ 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.

Y の母平均値 $\mathrm{E}[Y] = 1 \times p + 0 \times (1-p) = p$.
Y の母分散 $\mathrm{V}[Y] = (1-p)^2 \times p + (0-p)^2 \times (1-p) = p(1-p)$.
2 値 (ベルヌーイ試行) のときの特殊事情 母分散が母平均値 p の関数.

母比率 p を推定したい=母平均値 $\mathrm{E}[Y]$ を推定したい
母平均値の推定
サイズ n の標本中 k 個が「…である」とき,

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

L10-Q4

Quiz(母比率の点推定)

確率変数 Y の確率分布が次のように与えられる.

$$f_y = \begin{cases} p & (y = 1) \\ 1 - p & (y = 0) \\ 0 & (\text{他}) \end{cases}$$

ここで, p は未知のパラメタ.
Y のサイズ 20 の標本を得たところ, $y = 1$ が 12 個, $y = 0$ が 8 個だった.

1. $p = \mathrm{E}[Y]$ を推定しよう.
2. $\mathrm{V}[Y]$ を推定しよう.

L10-Q5

Quiz(母比率の点推定)

ある学部の学生 1200 人から, 20 人を無作為に選び出し, A 型であるかどうか質問したところ, 12 人が A 型であると答え, 残りが A 型でないと答えた.
学部の, 血液型 A である学生の母比率 p を (点) 推定しよう

連絡

- 配布資料は 1-503 向かいの引出, `http://hig3.net` で再配布.
- Quiz の略解は授業終了後に `http://hig3.net` で配布.
- 週のタイムラインで見たように, 非参照 Quiz 予習問題を RaMMoodle に金 17:00 ごろまでに公開. これで来週の Quiz に備えてね.
- Math ラウンジの配布資料訂正. 正しくは, 2015-12-23 水:授業なし, 2015-12-24 木:土曜授業.
- オフィスアワー月 4 木 6(1-502)

manaba / 出席カード
`https://manaba.ryukoku.ac.jp/s`