

中心極限定理・母集団と標本抽出・推定

樋口さぶろお

龍谷大学工学部数理情報学科

確率統計☆演習 I L10(2017-12-06 Wed)

最終更新: Time-stamp: "2018-12-03 Mon 21:29 JST hig"

今日の目標

- 中心極限定理の意味が説明でき、確率の近似計算に利用できる.

西川確率統計 §4.2



L09-Q1

Quiz 解答:標準正規分布の確率

標準正規分布の確率密度関数は偶関数 ($z = 0$ に関して対称) なので,

$$\begin{aligned}
 P(-\infty < Z < -2) &= \int_{-\infty}^{-2} f(z) dz \\
 &= Q(-\infty) - Q(-2) = 1 - (1 - Q(2)) = Q(2) = 0.0228.
 \end{aligned}$$

L09-Q2

Quiz 解答:標準正規分布の確率

確率密度関数が偶関数であることに注意する.

- ① $E[Z^2] = V[Z] + (E[Z])^2 = 1 + 0^2 = 1$.
- ② $P(-0.56 < Z < +1.23) = \int_{-0.56}^{1.23} f(z) dz = Q(-0.56) - Q(1.23) = (1 - Q(0.56)) - Q(1.23) = 1 - 0.1093 - 0.2877 = 0.6030$.

L09-Q3

Quiz 解答:正規分布の確率

定義にしたがって積分しても求まるが、正規分布の確率密度関数と比較すると、 $X \sim N(4, 3^2)$ なので、

- ① $x = 4$ を真ん中に幅 3 くらいの正規分布の確率密度関数のグラフ.
- ② $E[X] = 4$.
- ③ $V[X] = 3^2$.

L09-Q4

Quiz 解答:正規分布の確率

標準正規分布 $N(0, 1^2)$ にしたがう.

$Z = \frac{X-3}{2}$ とすると、 Z は標

- ① $P(X \geq 5) = P(\frac{5-3}{2} \leq Z < +\infty) = \int_1^{\infty} f(z) dz = Q(1) - Q(\infty) = Q(1.00) = 0.1587$.
- ② $P(1 \leq X \leq 7) = P(-1 \leq Z \leq 2) = \int_{-1}^2 f(z) dz = Q(-1) - Q(2) = 1 - Q(2.00) - Q(1.00) = 0.8186$.

ここまで来たよ

9 正規分布

10 中心極限定理・母集団と標本抽出・推定

- 中心極限定理と正規近似
- 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

独立同分布の復習

西川確率統計定理 4.1(p.84)

確率統計☆演習 I(2017)L07

X_1, \dots, X_n が独立同分布に従うとする. $E[X_i] = \mu, V[X_i] = \sigma^2$.

新しい確率変数: $U_n = X_1 + \dots + X_n$

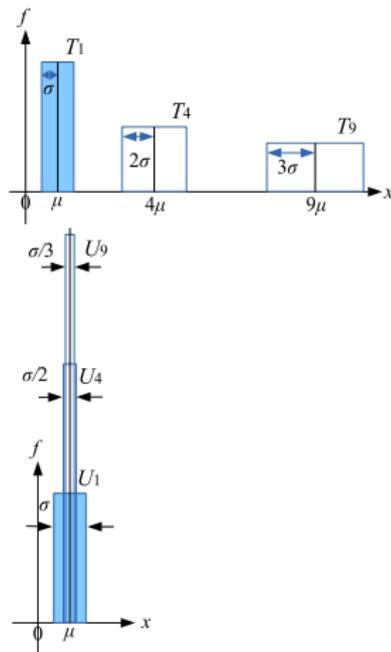
$$E[U_n] = \sum_{i=1}^n E[X_i] = n \times \mu.$$

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

新しい確率変数: $W_n = \frac{1}{n}U_n = \frac{1}{n}(X_1 + \dots + X_n)$

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

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



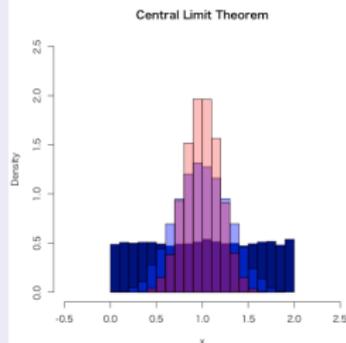
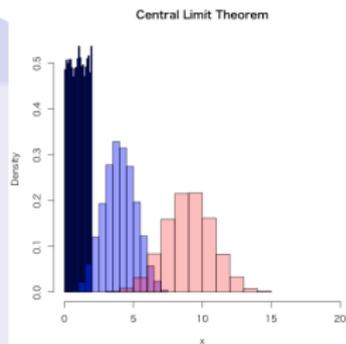
中心極限定理

西川確率統計 §4.2

中心極限定理 (いいかげんバージョン)

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

- $U_n = X_1 + \dots + X_n$ の確率分布は,
 の に似る
- $W_n = \frac{1}{n}(X_1 + \dots + X_n)$ の確率分布は,
 に似る
- $Z_n = \frac{W_n - \mu}{\sigma/\sqrt{n}}$ の確率分布は,
 に似る



中心極限定理 (厳密バージョン) 西川確率統計定理 4.3(p.87)

確率変数 X_1, X_2, \dots, X_n が, 母平均値 μ , 母分散 σ^2 の独立同分布に従うとする. **正規分布じゃない. どんな分布でも可**

$$Z_n = \frac{\frac{1}{n}(X_1 + \dots + X_n) - \mu}{\sigma} \times \sqrt{n} \text{ とすると,}$$

Z_n は, $n \rightarrow +\infty$ の極限で, $N(0, 1^2)$ に従う. すなわち

$$\lim_{n \rightarrow +\infty} P(a \leq Z_n < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

「 Z_n は $N(0, 1^2)$ にしたがう Z に**法則収束**する」

法則収束とは, 関数列がある関数に収束すること.

証明

$E[Z_n] = 0, V[Z_n] = 1$ はすぐわかるが...

モーメント母関数を使うと瞬殺 確率統計☆演習 II(L)

二項分布の正規近似

高校 数学 B

西川確率統計 §8.4



L10-Q1

Quiz(二項分布と正規分布と中心極限定理)

表が確率 $\frac{1}{10}$, 裏が確率 $\frac{9}{10}$ ででるコインを, 400 回投げるとき, 表がでる回数を確率変数 U とする.

- ① U はどのような二項分布にしたがうか. $B(?, ?)$ の形で答えよう.
- ② U は近似的にどのような正規分布にしたがうか. $N(?, ?)$ の形で答えよう.
- ③ 表が 31 回より多くでる確率を, 標準正規分布の上側確率 $Q(z)$ を用いて表し, さらに正規分布表を用いて小数値として近似的に求めよう.

実験 (あとでいう U_1, U_4, U_9 の標本抽出)

$$X_n \sim B(1, \frac{2}{3})$$

$$f(x) = \begin{cases} \frac{2}{3} & (x = 1) \text{ サイコロで } \boxed{3} \boxed{4} \boxed{5} \boxed{6} \\ \frac{1}{3} & (x = 0) \text{ サイコロで } \boxed{1} \boxed{2} \end{cases}$$

記入欄 $U_n = X_1 + \dots + X_n$.

n		1	2	3	4	5	6	7	8	9
目	(1-6)									
X_n	(0-1)									
U_n	(0-9)	*			*					*

<https://manaba.ryukoku.ac.jp> に送信.

ここまで来たよ

9 正規分布

10 中心極限定理・母集団と標本抽出・推定

- 中心極限定理と正規近似
- 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

母集団と標本 (1) 有限母集団

西川確率統計 §6.1

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 $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ 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 $E[X]$.
- 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.

⇨ 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 = , 標本サイズ = , 標本の個数 = .

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 \rightarrow 無限母集団 (何回でもひけるから).

- 賞金の母平均値 $E[X] = \sum_x f(x) \times x$ 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 = , 標本サイズ = , 標本の個数 = .

母集団・標本抽出・推定

西川確率統計 §6.1.6.2

- **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)

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 正規分布

10 中心極限定理・母集団と標本抽出・推定

- 中心極限定理と正規近似
- 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

X_1, X_2, \dots, X_n はサイズ n の標本.

各 X_i ($i = 1, \dots, n$) 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.

μ, σ^2 は母集団のパラメタ.

標本平均値 西川確率統計 p.132

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

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 $\bar{X}_{(n)}$ は確率変数であり, 試行=標本抽出のたびにかわる ($\bar{X}_{(n)}$ は確率分布をもつ)

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 重さの母分散を点推定しよう.
- ③ 110g 以上のものの母比率を点推定しよう.

よい推定値って? 西川確率統計定理 6.1(p.132)

標本平均値 $\bar{X}_{(n)}$ は不偏性を持つ

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

先週の $E[W_n] = \mu$

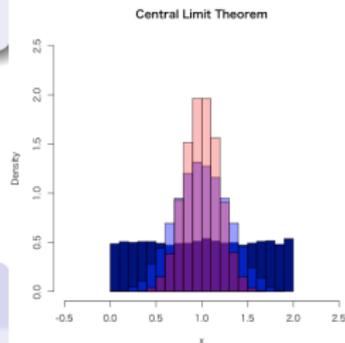
$$\forall n \quad E[\bar{X}_{(n)}] = \mu$$

標本平均値 $\bar{X}_{(n)}$ は一貫性を持つ

標本サイズ n が大きくなると, $\bar{X}_{(n)}$ 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

$$\forall \epsilon > 0 \quad \lim_{n \rightarrow +\infty} P(|\bar{X}_{(n)} - \mu| > \epsilon) = 0$$



母分散の(点)推定 高校 数学 B 西川確率統計 p.134

不偏標本分散 西川確率統計不偏分散 (p.134) であって標本分散 (p.134) と別

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

が、母分散の‘よい’推定値になっている。

ここで、 \bar{X} は母平均値でなく、上の標本平均値 ($\bar{X}_{(n)}$) の略記。

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

おぼえ方 (不偏) 標本分散は $n=1$ のとき、

$E[s^2] = \sigma^2$ を $n = 2$ のときに確認 (証明 西川確率統計定理 6.2,6.3)

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

ここで,

$$\begin{aligned}
 \sigma^2 &= V[X_1] = E[X_1^2] - (E[X_1])^2 = E[X_1^2] - \mu^2, \\
 \frac{\sigma^2}{n} &= V[\bar{X}] = E[\bar{X}^2] - (E[\bar{X}])^2 = E[\bar{X}^2] - \mu^2, \text{ より,}
 \end{aligned}$$

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

ここまで来たよ

9 正規分布

10 中心極限定理・母集団と標本抽出・推定

- 中心極限定理と正規近似
- 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

西川確率統計 §7.5.1,8.4

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, 条件「...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, $Y = \mathbf{1}_{[\dots\text{である}]}(X)$, たとえば $X > 10$ 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 $Y = 1$.

母比率

$B(1, p)$ の p . または母集団で条件 $f(x)$ 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,

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

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

標本比率

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。

母比率 p の推定=母平均値 $E[Y]$ の推定

サイズ n の標本中 k 個が「…である」とき、

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

連絡

- 来週は 7-002. 最初に紙の trial.
- 配布資料は 1-503 向かいの引出, <http://hig3.net> で再配布.
- 加減乗除と平方根(ルート)の使える電卓持ってきてね. 関数電卓でなくてもいいです. 携帯電話の機能・アプリでもかまいません.
- 樋口オフィスアワー月 3.5(1-539) 金 4(1-502), Math ラウンジ月-木昼 (1-614)
- 次回は区間推定 西川確率統計 §8.1-8.4 .