母集団と標本・点推定・区間推定

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(2021-06-16 Wed)

最終更新: Time-stamp: "2021-06-15 Tue 19:19 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
- 母比率を区間推定できる 岩薩林 確率・統計 §7.3

L09-Q1
Quiz 解答:指数分布

1. 間隔 X 分 は, パラメタ $\lambda = 0.05/$ 分の指数分布にしたがう (または間隔 X ゲーム は, パラメタ $\lambda' = 4.5/$ ゲームの指数分布にしたがう).
2. $\int_0^{+\infty} \lambda e^{-\lambda x}\, dx = \frac{1}{\lambda} = 20$ 分
3. $\int_0^{5} \lambda e^{-\lambda x}\, dx = [-e^{-\lambda x}]_0^5 = 1 - e^{-0.25} = 0.221.$
4. $\int_{15}^{25} \lambda e^{-\lambda x}\, dx = [-e^{-\lambda x}]_{15}^{25} = 0.186.$

L09-Q2
Quiz 解答:正規分布の応用

$X \sim \mathrm{N}(50, 10^2)$ なので, $Z = \frac{X-50}{10} \sim \mathrm{N}(0, 1^2)$.

$$P(60 \le X \le 65) = P(\tfrac{60-50}{10} \le Z \le \tfrac{65-50}{10}) = I(\tfrac{3}{2}) - I(1).$$

L10-Q3
Quiz 解答:独立同分布にしたがう変数の和

1. A は母平均値が $n\mu$, 母分散が $n\sigma^2$.
2. B は母平均値が μ, 母分散が $\frac{\sigma^2}{n}$.
3. $C = \frac{A-n\mu}{\sqrt{n}\sigma} = \frac{1}{\sigma/\sqrt{n}}(B-\mu) = \frac{1}{\sqrt{n}\sigma}(X_1 + X_2 + X_3 + \cdots + X_n - n\mu)$

L09-Q4

Quiz 解答:独立同分布と中心極限定理

1. 近似的に $\mathrm{N}(\frac{n}{\lambda}, \frac{n}{\lambda^2})$, 近似的に $\mathrm{N}(\frac{1}{\lambda}, \frac{1}{n\lambda^2})$.
2. 厳密に $\mathrm{B}(n,p)$ を近似して $\mathrm{N}(np, np(1-p))$, 近似的に $\mathrm{N}(p, \frac{1}{n}p(1-p))$.
3. (実は) 厳密にも $\mathrm{N}(n\mu, n\sigma^2)$, 厳密に $\mathrm{N}(\mu, \frac{1}{n}\sigma^2)$.
4. 厳密に $\chi(n)$ を近似して $\mathrm{N}(n, 2n)$, 近似的に $\mathrm{N}(1, \frac{2}{n})$.

L09-Q5

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, T は近似的に正規分布 $\mathrm{N}(n\mu, n\sigma^2)$ すなわち $\mathrm{N}(40, 6^2)$ $Z = \frac{T-40}{6}$ 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, $P(T > 31) = P(Z > -\frac{9}{6}) = Q(-\frac{3}{2}) - Q(\infty) = (1 - Q(\frac{3}{2})) - 0 = I(\infty) - I(-\frac{3}{2}) = \frac{1}{2} + I(\frac{3}{2}) = 0.9332.$

L09-Q6

Quiz 解答:二項分布と正規分布と中心極限定理

1. 表の出る回数 X は，二項分布 $\mathrm{B}(100, \frac{4}{5})$ にしたがう．よって，$\mathrm{E}[X] = 80$, $\mathrm{V}[X] = 4^2$ である．
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．
$Z = \frac{X-80}{4}$ 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，
$P(73 \leq X \leq 79) = P(-\frac{7}{4} < Z < -\frac{1}{7}) = I(-\frac{1}{4}) - I(-\frac{7}{4}) = -I(\frac{1}{4}) + I(\frac{7}{4})$.
2. $P(73 \leq X \leq 79) = I(1.75) - I(0.25) = 0.4599 - 0.0987 = 0.3612.$

ここまで来たよ

母集団と標本 (1) 有限母集団 岩薩林 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 岩薩林 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.

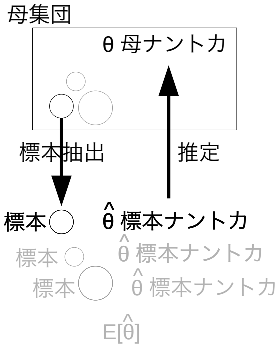
$\rightsquigarrow$ 引いた 5 枚のくじの賞金額=標本)(独立同分布にしたがう確率変数 $X_1, X_2, \cdots, X_5$) から推定したい.
5 枚を '無作為に' 選ぶ (=標本抽出する).
母集団サイズ= $+\infty$, 標本サイズ= 5, 標本の個数= 1.

母集団・標本抽出・推定

岩薩林 確率・統計 例 11(p.115)

岩薩林 確率・統計 図 p.109,115,137,167

- 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



推定には誤差あるかも．標本の選び方ごとに答は違うし．

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

母集団=クラスの回答者全体

1. 母集団サイズ 110 (クラス全体なら 126 だった)

標本 (Moodle が 1 人ずつ無作為抽出する)

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 $\mu = \mathrm{E}[X_i]$, 母期待値 $\mathrm{E}[g(X)]$ の独立同分布にしたがう確率変数.

標本平均値

岩薩林 確率・統計 (5.4)p.114

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

が, 母平均値 $\mu = \mathrm{E}[X]$ 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 $\overline{X}_{(n)}$ 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)
Excel では関数 average(), データ分析 > 基本統計量 > 平均

標本期待値

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

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 $\overline{X}_{(n)}$ 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

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

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

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

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

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

Excel では関数 var.s(), データ分析 > 基本統計量 > 分散 (を $\frac{n-1}{n}$ 倍しないそのまま)

$n-1$ 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.

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

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.

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

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. $110\mathrm{g}$ 以上のものの母比率を点推定しよう.

記述上の注意

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

L10-Q2

Quiz(母ナントカと標本ナントカ)

確率変数 X は二項分布 $\mathrm{B}(2, \frac{2}{3})$ にしたがう.
X のサイズ 6 の標本を抽出したところ,

$$0, 0, 0, 1, 2, 2$$

だった.

1. X の母平均値を求めよう.
2. X の母分散を求めよう.
3. この標本の標本平均値を求めよう.
4. この標本の不偏標本分散を求めよう.

ここまで来たよ

点推定 対 区間推定

点推定 岩薩林 確率・統計 §6.1

真の母平均値はわからないが, 標本平均値を使って,

「母平均値を A 円と推定する」

それどのくらい正確なの? 正確さは実は **母分散や標本サイズによる**

区間推定 岩薩林 確率・統計 §6.1

「母平均値が, B 円以上 C 円以下である '確率' は $1-\alpha=0.95$」

推定の精度・正確さまで表現

ここで '確率' というのは不誠実. 正しい言葉遣いは, 信頼係数=信頼度で

「母平均値の信頼係数 $1-\alpha=0.95$ の信頼区間は B 円以上 C 円以下」

動く (確率変数である) のは母平均値 μ でなく, B, C のほう.

母平均値の区間推定 (正規母集団, 母分散既知) 岩薩林 確率・統計 p.144

$\mathrm{N}(\mu, \sigma^2)$ にしたがう母集団 (正規母集団) の, サイズ n の標本を何回も取り出して, 毎回, 標本平均値 $\overline{X}_{(n)}$ を計算している. X_i:iid. よって,

$U_n = \overline{X}_{(n)} \sim \mathrm{N}(\mu, \sigma^2/n)$. $Z = \frac{\overline{X}_{(n)} - \mu}{\sqrt{\sigma^2/n}} \sim \mathrm{N}(0, 1^2)$.

> $n \to +\infty$ で正しいことは中心極限定理からわかる. 正規母集団でないときも, 標本サイズ n が大きい (30 くらい) なら, 近似的に成立することが多い.

標本平均値 $\overline{X}_{(n)}$ が母平均値 μ から大きく外れない確率は大きい (ここでは $1 - \alpha = 1 - 0.05$ に等しい) という式を書くと…

$$P(z(1 - \tfrac{\alpha}{2}) < \frac{\overline{X}_{(n)} - \mu}{\sqrt{\sigma^2/n}} < +z(\tfrac{\alpha}{2})) = 1 - \alpha.$$

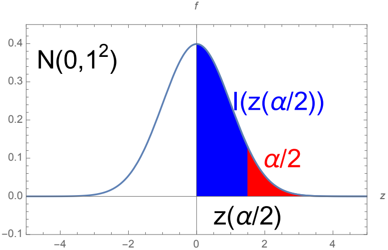
$$P(-1.96 < \frac{\overline{X}_{(n)} - \mu}{\sqrt{\sigma^2/n}} < +1.96) = 1 - 0.05.$$

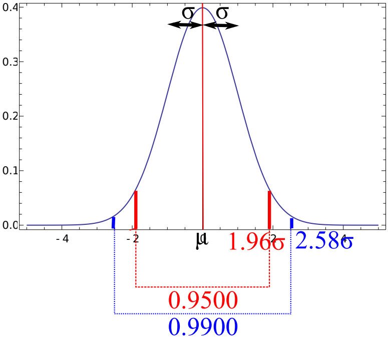
$$P(\mu - 1.96 \times \sqrt{\sigma^2/n} < \overline{X}_{(n)} < \mu + 1.96 \times \sqrt{\sigma^2/n}) = 1 - 0.05.$$

μ について不等式を解くと,

$$P(\overline{X}_{(n)} - 1.96 \times \sqrt{\sigma^2/n} < \mu < \overline{X}_{(n)} + 1.96 \times \sqrt{\sigma^2/n}) = 1 - 0.05.$$

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 $P(Z > z(\alpha)) = \alpha$ で $z(\alpha)$ を定める.
$I(z(\alpha)) = \frac{1}{2} - \alpha$.

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

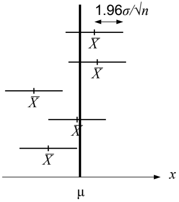
母平均値 (正規母集団, 母分散既知) の信頼区間 岩薩林 確率・統計 定理 6.1(p.18)

$\mathrm{N}(\mu, \sigma^2)$ にしたがう母集団の, σ^2 がわかっているとき, サイズ n の標本から区間推定すると, 母平均値 μ の 信頼係数 $1-\alpha$ の信頼区間 ($(1-\alpha)\times 100\%$信頼区間) は, $\overline{X}_{(n)}$ を標本平均値として,

$$\overline{X}_{(n)} - z(\tfrac{\alpha}{2}) \times \sqrt{\sigma^2/n} < \mu < \overline{X}_{(n)} + z(\tfrac{\alpha}{2}) \times \sqrt{\sigma^2/n}.$$

何回も標本抽出して何個も信頼区間を求めたとき, 信頼区間が μ を含む確率は, 信頼係数 $1-\alpha$. 推定がはずれる確率 α.

切りがいい α の $z(\alpha)$ は岩薩林 確率・統計 付表 1(p.227)の下.
$z(\frac{0.05}{2}) = 1.96$, $z(\frac{0.01}{2}) = 2.58$.



高校数学では, $z(\frac{0.05}{2}) = 1.96$ の場合のみ.
$a < \mu < b$ でなく, 閉区間の記号 $[a, b]$ で.
真の母分散 σ^2 の代わりに, (不偏 $\frac{1}{n-1}$ じゃない)
$S^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2$ の標本分散を使っていい.

L10-Q3

Quiz(母平均値の区間推定 (母分散既知))

あるドーナツ製造マシンが i 番目に製造するドーナツの重さ $X_i\mathrm{g}$ は, 独立で, 同じ正規分布にしたがう確率変数である. あらかじめ行った調査により, X_i の母分散は $\sigma^2 = 9\mathrm{g}^2$ であることがわかっている.
製造された 4 個のドーナツの重さを測定したところ, 次のようだった.
$51\mathrm{g}, 52\mathrm{g}, 47\mathrm{g}, 50\mathrm{g}.$

1. 母平均値 $\mu = \mathrm{E}[X_i]$ を, 信頼係数 $1 - \alpha = 0.95$ で区間推定しよう.
2. 母平均値 $\mu = \mathrm{E}[X_i]$ を, 信頼係数 $1 - \alpha = 0.99$ で区間推定しよう.

岩薩林 確率・統計 例題 6.4(p. 145) | 岩薩林 確率・統計 問題 3(p.146) | 岩薩林 確率・統計 第 6 章練習問題 1

ここまで来たよ

9 中心極限定理と正規近似

10 母集団と標本・点推定・区間推定
- 母集団と標本
- 母平均値・母分散の (点) 推定
- 母平均値の区間推定 (正規母集団, 母分散既知)
- 母比率の (点) 推定
- 母比率の区間推定

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, $Y = \mathrm{I}_{[\cdots である]}(X)$, たとえば $X > 10$ なら $Y = 1$.
- 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 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,

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

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.

母比率 p=母平均値 $\mathrm{E}[Y] = \mathrm{E}[\mathrm{I}_{[\text{条件}]}(X)]$ の推定

サイズ n の標本中 k 個が「…である」とき,

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

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

岩薩林 確率・統計 問題 6(p.116)

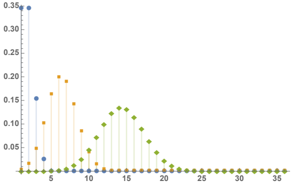
ここまで来たよ

母比率の信頼区間

母比率は母平均値の一種なので，さっきの区間推定の式で，$\sigma^2 = p(1-p)$ とおく．

別の見方

$K \sim \mathrm{B}(n, p)$． n が大きいとき近似的に $K \sim \mathrm{N}(np, np(1-p))$.



$p = 0.8, n = 4, 20, 40.$

信頼係数 $1-\alpha$.

$$P\left(p - z\left(\frac{\alpha}{2}\right) \times \sqrt{\frac{p(1-p)}{n}} < \hat{p} < p + z\left(\frac{\alpha}{2}\right) \times \sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha.$$

σ^2 の $p(1-p)$ は $\hat{p}(1-\hat{p})$ とする近似で，p について解く．

母比率の信頼区間 (母分散未知) 岩薩林 確率・統計 §7.3

X のサイズ n の標本で，標本比率 $\hat{p} = k/n$ のとき，母比率の 信頼係数 $1 - \alpha$ の信頼区間 ($(1-\alpha) \times 100\%$信頼区間) は，

$$\hat{p} - z(\tfrac{\alpha}{2}) \times \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z(\tfrac{\alpha}{2}) \times \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}}.$$

$z(\frac{0.05}{2}) = 1.96, z(\frac{0.01}{2}) = 2.58.$

L10-Q4

Quiz(母比率の区間推定)

選挙で出口調査をしたところ, 50 人中 35 人が A 候補に投票したと答えた. 母集団を投票した人全体とする. そのうち A 候補に投票した人の母比率 (得票率) を考える.

1. A 候補の得票率を, (点) 推定しよう
2. A 候補の得票率を, 信頼係数 $1-\alpha=0.95$ で区間推定しよう.
3. A 候補の得票率を, 信頼係数 $1-\alpha=0.99$ で区間推定しよう.

岩薩林 確率・統計 例題 7.6(p.170)	岩薩林 確率・統計 問題 7(p.171)	岩薩林 確率・統計 第 7 章練習問題 2(2)

注: 下限, 上限が 0,1 を越えるときは, 0,1 に直してしまっていい.