母平均値の区間推定・線形回帰モデル

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 L12(2025-07-07 Mon)

最終更新: Time-stamp: "2025-07-07 Mon 06:45 JST hig"

今日の目標

- t 分布を説明できる 久保川 統計学入門 §8.1
- 母平均値の区間推定ができる（正規母集団，小標本，母分散未知） 久保川 統計学入門 §8.2
- 線形回帰モデルを説明できる 久保川 統計学入門 §12.1

L11-Q1

Quiz 解答: カイ二乗分布

1. $F_{\mathrm{w}}(0.1)$. Python では `stats.chi2(df=2).cdf(0.1)`.
2. $1 - F_{\mathrm{w}}(0.9)$. Python では `1-stats.chi2(df=2).cdf(0.9)`.
3. $F_{\mathrm{w}}^{-1}(0.1)$. Python では `stats.chi2(df=2).ppf(0.1)`.
4. $F_{\mathrm{w}}^{-1}(1-0.1)$. Python では `stats.chi2(df=2).ppf(1-0.1)`.

L11-Q2

Quiz 解答: 母分散の区間推定

標本サイズは $n = 9$, 自由度は $9 - 1$.
不偏標本分散を V^2 とすると，$(9-1)\frac{V^2}{\sigma^2}$ が自由度 $9-1$ のカイ二乗分布に従うことから，母分散 σ^2 の信頼係数 $1-\alpha = 0.95$ の信頼区間は,

$$
\begin{aligned}
\frac{n-1}{F_{\mathrm{w8}}^{-1}(1-\frac{\alpha}{2})} \times V^2 <&\sigma^2 < \frac{n-1}{F_{\mathrm{w8}}^{-1}(\frac{\alpha}{2})} \times V^2 \\
\frac{n-1}{\chi^2_{n-1,\frac{\alpha}{2}}} \times V^2 <&\sigma^2 < \frac{n-1}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \times V^2 \\
\frac{8}{17.53} \times 72 <&\sigma^2 < \frac{8}{2.180} \times 72 \\
32.85 <&\sigma^2 < 264.2 \quad [\mathrm{g}^2]
\end{aligned}
$$

ここまで来たよ

11 母分散の区間推定

12 母平均値の区間推定・線形回帰モデル
- 母平均値の区間推定 (正規母集団, 母分散既知，小標本)
- t 分布
- 母平均値の区間推定 (正規母集団, 母分散未知，小標本)
- 線形回帰モデルとしての回帰分析

復習：大標本からの母平均値の近似的信頼区間 確率統計 I(2025)L10

未知（母集団）母平均値 μ, 母分散 σ^2
$\uparrow$ 推定
既知 標本 $X_1, \ldots, X_n \rightsquigarrow$ 標本平均値 $\overline{X}$, 不偏標本分散 V^2.

近似 1 サンプルサイズ n が大きいとき，中心極限定理により，近似的に.

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

$$\text{標準化} \quad \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

近似 2 σ のかわりに $\sqrt{V^2}$ を使っても近似的になりたつ.

$$\frac{\overline{X}-\mu}{\sqrt{V^2/n}} \sim N(0, 1^2)$$

命題 (母平均値の近似的信頼区間 久保川 統計学入門 公式 8.10)

母平均値 μ, 母分散 σ^2 のなにかの分布にしたがう母集団の, 母平均値 μ の 信頼係数 $1-\alpha$ の近似的信頼区間 ($(1-\alpha)$ 信頼区間) は, n が大きいとき，$\overline{X}$ を標本平均値, V^2 を不偏標本分散として,

$$\overline{X} - z_{\alpha/2} \times \sqrt{V^2/n} < \mu < \overline{X} + z_{\alpha/2} \times \sqrt{V^2/n}.$$

標本平均値のしたがう分布 (正規母集団, 母分散既知) 久保川 統計学入門 §8.2

$N(\mu,\sigma^2)$ 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 $\overline{X}$ を計算する.
標本平均値 $\overline{X}=\frac{1}{n}(X_1+\cdots+X_n)$ は確率変数

命題 (正規分布の再生性 久保川 統計学入門 公式 6.21)

$X\sim N(\mu_1,\sigma_1^2), Y\sim N(\mu_2,\sigma_2^2)$, X,Y 独立のとき,
$X+Y\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$.

$\to+\infty$ しなくても $n=\underline{2}$ から正規分布
$\overline{X}\sim N(\mu,\sigma^2/n)$. $Z=\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}\sim N(0,1^2)$.

超強い仮定として，正規母集団を仮定すれば，$\boxed{\text{近似 1}}$を取り除ける!

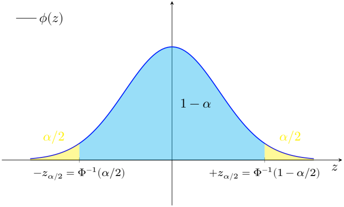
σ^2 はなぜかわかってるとして，$\boxed{\text{近似 2}}$ (σ^2 を V^2 で置き換えちゃう) なしで書いてみる

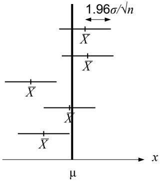
命題 (母平均値 (正規母集団, 母分散既知) の信頼区間 久保川 統計学入門 公式 8.6(1))

$N(\mu, \sigma^2)$ 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 ($(1-\alpha)$ 信頼区間) は, $\overline{X}$ を標本平均値として,

$$\overline{X} - \Phi^{-1}(1-\tfrac{\alpha}{2}) \times \sqrt{\sigma^2/n} < \mu < \overline{X} - \Phi^{-1}(\tfrac{\alpha}{2}) \times \sqrt{\sigma^2/n}$$
$$\overline{X} - z_{\alpha/2} \times \sqrt{\sigma^2/n} < \mu < \overline{X} + z_{\alpha/2} \times \sqrt{\sigma^2/n}.$$

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 $1-\alpha$. 推定が外れる確率 α.





ここまで来たよ

11 母分散の区間推定

12 母平均値の区間推定・線形回帰モデル

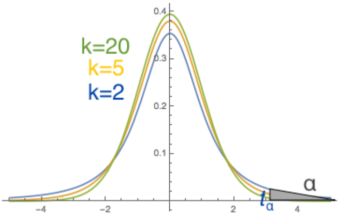
- 母平均値の区間推定 (正規母集団, 母分散既知，小標本)
- t 分布
- 母平均値の区間推定 (正規母集団, 母分散未知，小標本)
- 線形回帰モデルとしての回帰分析

t 分布

あとは，$\boxed{\text{近似 2}}$ をなんとかしたい．

定義 (t 分布 久保川 統計学入門 定義 8.3**)**

$Z \sim N(0,1^2)$, $W \sim \chi_k^2$, Z と W が独立,
のとき連続型確率変数 $T = \frac{Z}{\sqrt{W/k}}$ のしたがう分布を自由度 k の (スチューデントの, またはゴセットの)t 分布 t_k という.



自由度 k が小さいとき, $N(0,1^2)$ より低く広い.
自由度 $k \to +\infty$ で $N(0,1^2)$ に一致する.

確率密度関数 $f_k(x) = A_k \cdot \left(1 + \frac{1}{k}x^2\right)^{-\frac{k+1}{2}}$

t 分布の確率密度関数と累積分布関数 久保川 統計学入門 p.171

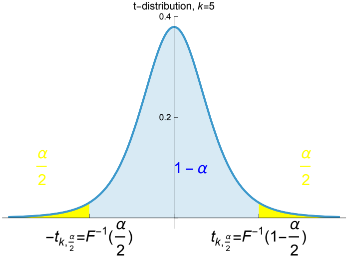
Python scipy.stats.t(df=k)

久保川 統計学入門 数値表 p.291 上側確率 $\alpha = P(T > t_{k,\alpha})$ となる $t_{k,\alpha}$ の値.

t 分布の確率密度関数は偶関数なので, $t_{k,1-\alpha} = -t_{k,\alpha}$.

久保川 統計学入門 数値 p.291 と p.290 はフォーマットが違う

$t_k \to N(0,1^2), t_{k,0.025} \to 1.960, t_{k,0.005} \to 2.576 \ (k \to +\infty)$.



久保川 統計学入門 図 8.5

L12-Q1

Quiz(t 分布の確率と $t_{k,\alpha}$)

標準正規分布にしたがう確率変数 $Z \sim N(0,1^2)$ と，自由度 $k = 40$ の t 分布にしたがう確率変数 $T \sim \mathrm{t}_{40}$ を考える．各分布の累積分布関数の逆関数 F_{t}^{-1} を使って答え，さらに数値にしよう．

1. 確率 $P(Z > z) = 0.025$ となる z を求めよう．
2. 確率 $P(Z > z) = 1 - 0.025$ となる z を求めよう．
3. 確率 $P(T > w) = 0.025$ となる w を求めよう．
4. 確率 $P(T > w) = 1 - 0.025$ となる w を求めよう．

ここまで来たよ

11 母分散の区間推定

12 母平均値の区間推定・線形回帰モデル
- 母平均値の区間推定 (正規母集団, 母分散既知，小標本)
- t 分布
- 母平均値の区間推定 (正規母集団, 母分散未知，小標本)
- 線形回帰モデルとしての回帰分析

母平均値の区間推定 (正規母集団, 母分散未知)

母分散既知の区間推定では, 標本平均値を標準化した. $Z = \frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$.
ふつう, μ がわからないときは σ^2 もわかってない.
σ^2 のかわりに不偏標本分散 V^2 (確率変数) を使った, 標準化もどき

$T = \frac{\overline{X}-\mu}{\sqrt{V^2/n}} = \frac{\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{(n-1)V^2}{\sigma^2}/(n-1)}}$ でやっちゃいたい. 分子 $\sim N(0,1^2)$, 分母 $= \sqrt{W/(n-1)}, W \sim \chi^2_{n-1}$，実は W と X は独立，より, $T \sim \mathrm{t}_{n-1}$.

命題 (T の分布 久保川 統計学入門 公式 8.4)

母集団 $N(\mu,\sigma^2)$ から, サイズ n の標本 $X_1,\cdots,X_n$ を取り出したとき,

$$T = \frac{\overline{X}-\mu}{\sqrt{V^2/n}}$$

は, 自由度 $n-1$ の Student の t 分布にしたがう.

標準化に σ^2 のかわりに V^2 を使うと，標準正規分布のかわりに t_k になる

母平均値の信頼区間 (母分散未知) 久保川 統計学入門 公式 8.6(2)

(母分散未知の) 正規分布 $N(\mu, ?^2)$ にしたがう母集団から, サイズ n の標本を得たとき, 母平均値 μ の 信頼係数 $1-\alpha$ の信頼区間は

$$\overline{X} - t_{n-1,\alpha/2} \times \sqrt{V^2/n} < \mu < \overline{X} + t_{n-1,\alpha/2} \times \sqrt{V^2/n}.$$

ただし, $\overline{X}$: 標本平均値, S^2: 不偏標本分散, n: 標本サイズ, $t_{n-1,\alpha/2}$: 自由度 $n-1$ の t 分布の上側確率が $\alpha/2$ となる点.

母分散既知 確率統計 I(2025)L10 と比べて,

$\sigma^2 \rightsquigarrow S^2$

$z_{\alpha/2} \rightsquigarrow t_{n-1,\alpha/2}$

L12-Q2

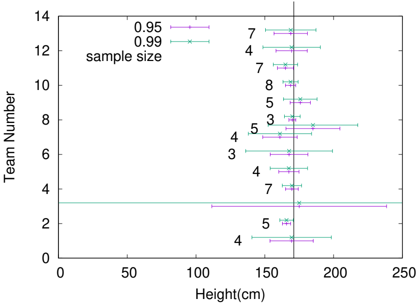
Quiz(母平均値の区間推定 (母分散未知))

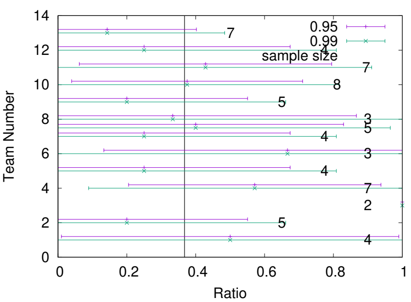
あるドーナツ製造マシンが製造するドーナツの重さ Xg は, 正規分布にしたがう確率変数である
製造された 4 個のドーナツの重さを測定したところ, 次のようだった.
$51\mathrm{g}, 52\mathrm{g}, 47\mathrm{g}, 50\mathrm{g}.$

1. 母平均値 $\mu = \mathrm{E}[X]$ を, 信頼係数 $1-\alpha = 0.95$ で区間推定しよう. まず, $t_{k,\alpha}$ または F_{t}^{-1} で書き, 小数に直そう.
2. 母平均値 $\mu = \mathrm{E}[X]$ を, 信頼係数 $1-\alpha = 0.99$ で区間推定しよう.

久保川 統計学入門 例題 8.7

クラス 2017 の身長の母平均値/「滋賀県」の母比率の区間推定





注: 母比率が 0.5 のとき, 信頼区間の幅は最大で $2 \times 1.96 \times 0.5/\sqrt{n}$.

ここまで来たよ

11 母分散の区間推定

12 母平均値の区間推定・線形回帰モデル

- 母平均値の区間推定 (正規母集団, 母分散既知，小標本)
- t 分布
- 母平均値の区間推定 (正規母集団, 母分散未知，小標本)
- 線形回帰モデルとしての回帰分析

線形モデル (統計モデルのある一族)

あるドーナツ製造機の作るドーナツの重さ Yは次のモデルに従う.
Y, u: 連続型確率変数, σ^2, β_0: 定数=パラメタ=母数.

$$Y = \beta_0 + u, \qquad u \sim N(0, \sigma^2)$$

u: 誤差, ノイズ. 小文字だけど確率変数.
(隠された) パラメタ β_0, σ^2 を, サイズ n 個の Yの標本から推定したい.

$$\begin{aligned} \mathrm{E}[Y] =& \beta_0 + \mathrm{E}[u] = \beta_0, \\ \mathrm{Var}(Y) =& \quad \mathrm{Var}(u) = \sigma^2. \end{aligned}$$

$u \sim N(0, \sigma^2) \rightsquigarrow Y = u + \beta_0 \sim N(\beta_0, \sigma^2)$. ここしばらくやってた, 正規母集団, 母平均値, 母分散の推定の言い換えに過ぎない.
だけど, 多数のパラメタを含む一般的なモデルにも使える考え方をする.

(確率変数でない) 変数 x に依存する確率変数 Y

ドーナツ製造機の作るドーナツの重さ Y は, 温度 x による場合，線形単回帰モデルを仮定できる．

定義 (線形単回帰モデル 久保川 統計学入門 (12.1))

Y, u: 連続型確率定数, β_0, β_1: 回帰係数, σ^2: 定数=パラメタ

$$Y = \beta_0 + \beta_1 \cdot x + u, \qquad u \sim N(0, \sigma^2)$$

あれ？ データ分析 久保川 統計学入門 §2.2 で見た？ ここでは (x, Y) は平等ではない．

x: 説明変数 (独立変数) ここでは確率変数でない

Y: 被説明変数目的変数 (従属変数) ここでは確率変数

$\beta_0, \beta_1, \sigma^2$ は，未知の母ナントカ．データ (x_i, Y_i) から推定したい．

β_1: 回帰係数 (x を 1 だけ変えたときの y の変化量)

ノイズ・誤差 $u = Y - (\beta_0 + \beta_1 \cdot x) \sim N(0, \sigma^2)$ は縦向きのみ.

x_i の値を変えて，Y_i の実現値を測定 $\rightsquigarrow$ 独立だけど n 個の同一分布でない！

最小 2 乗法

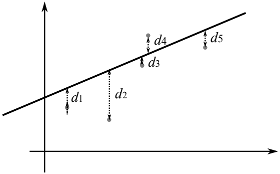
直線 $y = \beta_0 + \beta_1 x$ からの，実現値 (x_i, y_i) の縦のずれの 2 乗

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_1 x_i + \beta_0))^2.$$

これが最小になるという条件で決める $\hat{\beta}_0, \hat{\beta}_1$ を，最小二乗推定量という．

$\frac{\partial S}{\partial \alpha}(\hat{\beta}_0, \hat{\beta}_1) = \frac{\partial S}{\partial \beta}(\hat{\beta}_0, \hat{\beta}_1) = 0$ 微積分 II

u, Y が確率変数なので，$\hat{\beta}_0, \hat{\beta}_1$ も確率変数．

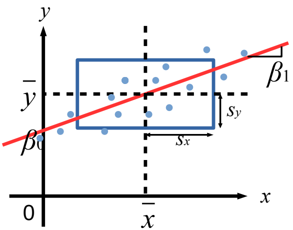


物理実験

回帰直線の公式再訪データ分析

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2}$$

$\hat{\beta}_0 =$(点 $(\overline{x}, \overline{Y})$ を通るような値)



命題 (回帰直線 久保川 統計学入門 公式 2.8)

x_i, Y_i $(i = 1, \ldots, n)$ の平均値を $\overline{x}, \overline{y}$, 標本標準偏差を S_x, S_y, 標本共分散を S_{xy} とする．このとき回帰直線は,

$$y = \frac{S_{xy}}{S_x^2} \times (x - \overline{x}) + \overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x.$$

定義

- 予測値 ($\neq$ 推定量) $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \overline{y} + \hat{\beta}_1(x_i - \overline{y})$.
- 残差 $e_i = y_i - \hat{y}_i =$ 実測値 − 予測値. これも確率変数.
- 残差平方和 $\mathrm{RSS} = \sum_i e_i^2$.

σ^2 の推定量 $\hat{\sigma}^2 = \frac{\sum_i e_i^2}{n-2}$.
自由度は $n-2$. β_0, β_1 の 2 個分ずらして e_i^2 を小さくできるから，その分を引いておく. $\hat{\beta}_1$ が最重要な推定量

$\hat{\beta}_1$ の母平均値と母分散

$$
\begin{aligned}
\hat{\beta}_1 =& \frac{1}{S_x^2}\frac{1}{n}\sum_i (x_i - \overline{x})(\beta_1(x_i - \overline{x}) + u_i), \\
\mathrm{E}[\hat{\beta}_1] =& \beta_1, \\
\mathrm{Var}(\hat{\beta}_1) =& \left(\frac{1}{nS_x^2}\right)^2 \mathrm{Var}(\sum_i (x_i - \overline{x}) u_i) = \cdots = \frac{\sigma^2}{nS_x^2}.
\end{aligned}
$$

命題 (久保川 統計学入門 公式 12.1)

- $\hat{\beta}_1$ は β_1 の不偏推定量: $N(\beta_1, \frac{\sigma^2}{nS_x^2})$.
- $\hat{\sigma}^2$ は何倍かするとカイ二乗分布にしたがう：$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$

$\hat{\beta}_1$ の精度を上げるには

標準偏差を小さくするには，n を大きくするか，S_x^2 を大きくするか．どちらも実験計画により実現できる．

さらなる問

久保川 統計学入門 p.232，公式 12.2

- β_1 の信頼区間は？ 確率統計 II,III
- $\beta_1 \neq 0$ なのか？ Y と x は「関係」あるのか？ 統計的仮説検定 確率統計 I(2025)L13