

母平均値の区間推定・線形モデル

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 L13(2024-07-15 Mon)

最終更新: Time-stamp: "2024-07-15 Mon 11:41 JST hig"

今日の目標

- 岩薩林 確率・統計 p.127 t 分布を説明できる
- 岩薩林 確率・統計 §7.1 母平均値を区間推定できる
- 線形モデルの最尤推定を説明できる



L12-Q1

Quiz 解答: カイ二乗分布

- ① $F_W(0.1)$.
- ② $1 - F_W(0.9)$.
- ③ $F_W^{-1}(0.1)$.
- ④ $F_W^{-1}(1 - 0.1)$.

L12-Q2

Quiz 解答: カイ二乗分布の確率と $\chi_k^2(\alpha)$

- ① $\Phi^{-1}(1 - 0.025) = z(0.025) = 1.960$.
- ② 標準正規分布の確率密度関数は偶関数なので,
 $F_Z^{-1}(0.025) = z(1 - 0.025) = -z(0.025) = -1.960$.
- ③ $F^{-1}(1 - 0.05) = \chi_1^2(0.05) = 3.841$. 別解. $0.05 = P(W > w_0) = P(Z^2 > w_0) = P(Z < -\sqrt{w_0} \text{ or } Z > +\sqrt{w_0}) = 2 \times P(Z > \sqrt{w_0})$.
 よって, $\sqrt{w_0} = 1.960$.

$$\textcircled{4} F^{-1}(0.05) = \chi_1^2(1 - 0.05) = 0.00393.$$

L12-Q3

Quiz 解答: 母分散の区間推定

標本サイズは $n = 9$, 自由度は $9 - 1$, 母分散 σ^2 の信頼係数 $1 - \alpha = 0.95$ の信頼区間は,

$$\begin{aligned} \frac{n-1}{F_W^{-1}(1 - \frac{\alpha}{2})} \times S^2 < \sigma^2 < \frac{n-1}{F_W^{-1}(\frac{\alpha}{2})} \times S^2 \\ \frac{n-1}{\chi_{n-1}^2(\frac{\alpha}{2})} \times S^2 < \sigma^2 < \frac{n-1}{\chi_{n-1}^2(1 - \frac{\alpha}{2})} \times S^2 \\ \frac{8}{17.53} \times 72 < \sigma^2 < \frac{8}{2.180} \times 72 \\ 32.85 < \sigma^2 < 264.2 \quad [g^2] \end{aligned}$$

ここまで来たよ

12 母分散の区間推定

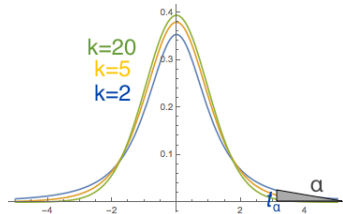
13 母平均値の区間推定・線形モデル

- t 分布
- 母平均値の区間推定 (正規母集団, 母分散未知)
- 最尤推定・線形モデルとしての回帰分析

t 分布

t 分布

$Z \sim N(0, 1^2)$, $W \sim \chi_k^2$, 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 $k \rightarrow +\infty$ で $N(0, 1^2)$ に一致する.

$$\text{確率密度関数 } f_{tk}(t) = A_k \cdot \left(1 + \frac{1}{k}t^2\right)^{-\frac{k+1}{2}}.$$

t 分布の確率密度関数と累積分布関数

岩薩林 確率・統計 付表 2

確率密度関数 $f_{t_k}(t)$, 累積分布関数 $F_{t_k}(t)$ は自由度 k ごとに異なる

Python `scipy.stats.t(df=k)`

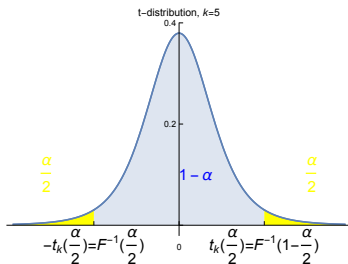
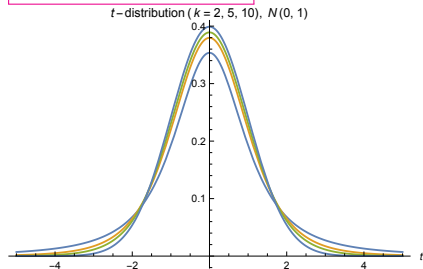
上側確率が $\alpha = P(T > t)$ となる $t = F_{t_k}^{-1}(1 - \alpha)$ を $t = t_k(\alpha)$ とかく.

岩薩林 確率・統計 付表 2 は付表 1 とフォーマットが違う

t 分布の確率密度関数 f は偶関数なので, $t_k(1 - \alpha) = -t_k(\alpha)$.

$t_k \rightarrow N(0, 1^2)$, $t_k(0.025) \rightarrow 1.960$, $t_k(0.005) \rightarrow 2.576$ ($k \rightarrow +\infty$).

岩薩林 確率・統計 図 5.8(p.128)



L13-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 $T \sim t_{40}$ を考える. 各分布の累積分布関数の逆関数 Φ^{-1}, F^{-1} を使って答え, さらに Python を使って数値にしよう.

- 1 確率 $P(|Z| > z) = 0.05$ となる z を求めよう.
- 2 確率 $P(|T| > w) = 0.05$ となる w を求めよう.

母平均値の区間推定 (正規母集団, 母分散未知)

岩薩林 確率・統計 §7.1

正規母集団 $N(\mu, \sigma^2)$ からサイズ n の標本を抽出したとする。

母分散 σ^2 既知なら, 標本平均値 \bar{X} を標準化 $Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$.

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 S^2 (確率変数) を使った, 標準化もどき

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} = \frac{\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{S^2}{\sigma^2}}} \text{ でやっちゃいたい.}$$

分子 $Z \sim N(0, 1^2)$, 分母 $= \sqrt{W/(n-1)}$, $W \sim \chi_{n-1}^2$, 実 Z, W は独立 より,
 $T \sim t_{n-1}$.

T の分布

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

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

は, 自由度 $n - 1$ の Student の t 分布にしたがう。

母集団が厳密に正規分布にしたがわなくても近似的に正しいことが多い。

ここまで来たよ

12 母分散の区間推定

13 母平均値の区間推定・線形モデル

- t 分布
- 母平均値の区間推定 (正規母集団, 母分散未知)
- 最尤推定・線形モデルとしての回帰分析

標本平均値 \bar{X} 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 $1 - \alpha = 1 - 0.05$ とか) という式を書くと…

$$P\left(F_{tn-1}^{-1}\left(\frac{\alpha}{2}\right) < \frac{\bar{X} - \mu}{\sqrt{S^2/n}} < F_{tn-1}^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha.$$

$$P\left(\mu + F_{tn-1}^{-1}\left(\frac{\alpha}{2}\right) \times \sqrt{S^2/n} < \bar{X} < \mu + F_{tn-1}^{-1}\left(1 - \frac{\alpha}{2}\right) \times \sqrt{S^2/n}\right) = 1 - \alpha.$$

μ について不等式を解くと,

$$P\left(\bar{X} - F_{tn-1}^{-1}\left(1 - \frac{\alpha}{2}\right) \times \sqrt{S^2/n} < \mu < \bar{X} - F_{tn-1}^{-1}\left(\frac{\alpha}{2}\right) \times \sqrt{S^2/n}\right) = 1 - \alpha.$$

$$P\left(\bar{X} - t_{n-1}\left(\frac{\alpha}{2}\right) \times \sqrt{S^2/n} < \mu < \bar{X} - (-t_{n-1}\left(\frac{\alpha}{2}\right)) \times \sqrt{S^2/n}\right) = 1 - \alpha.$$

母平均値の信頼区間 (母分散未知) 岩薩林 確率・統計定理 7.1(7.3)

(母分散未知の) 正規分布 $N(\mu, \sigma^2)$ 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

$$\bar{X} - F_{tn-1}^{-1}\left(1 - \frac{\alpha}{2}\right) \times \sqrt{S^2/n} < \mu < \bar{X} - F_{tn-1}^{-1}\left(\frac{\alpha}{2}\right) \times \sqrt{S^2/n}.$$

$$\bar{X} - t_{n-1}\left(1 - \frac{\alpha}{2}\right) \times \sqrt{S^2/n} < \mu < \bar{X} + t_{n-1}\left(\frac{\alpha}{2}\right) \times \sqrt{S^2/n}.$$

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

業界の習慣で、しばしば $\alpha = 0.05, 0.01$.

L13-Q2

Quiz(母平均値の区間推定 (母分散未知))

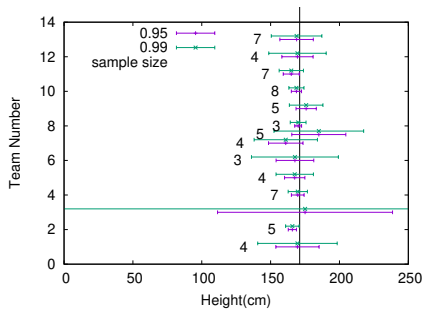
あるドーナツ製造マシンが製造するドーナツの重さ X g は, 正規分布にしたがう確率変数である

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.
51g, 52g, 47g, 50g.

- ① 母平均値 $\mu = E[X]$ 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 $\mu = E[X]$ 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 7.1(p.158), 問題 1(p.158) 第 7 章練習問題 1(2)

クラス 2017 の身長之母平均値の区間推定



ここまで来たよ

12 母分散の区間推定

13 母平均値の区間推定・線形モデル

- t 分布
- 母平均値の区間推定 (正規母集団, 母分散未知)
- 最尤推定・線形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.

$$Y = \beta_0 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Y, ϵ : 連続型確率変数, $\sigma > 0, \beta_0$: 定数=パラメタ=母数

ϵ : 誤差, ノイズ. 小文字だけど確率変数.

(未知の) パラメタ β_0, σ を, n 個のドーナツの重さのデータから推定したい.

$$E[Y] = \beta_0 + E[\epsilon] = \beta_0,$$

$$V[Y] = V[\epsilon] = \sigma^2.$$

$Y \sim N(\beta_0, \sigma^2)$ じゃん!

ここしばらくやってた, 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.

尤度 likelihood

$\epsilon = Y - \beta_0 \sim N(0, \sigma^2)$ より, ドーナツの重さ y を得る確率密度は,

$$f(y|\beta_0, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\beta_0)^2}{2\sigma^2}}$$

サイズ n の標本が y_1, \dots, y_n である確率密度は, 独立同分布なので積で,

$$\begin{aligned} f(y_1, y_2, \dots, y_n | \beta_0, \sigma) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\beta_0)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{-n/2} e^{-\sum_{i=1}^n \frac{(y_i-\beta_0)^2}{2\sigma^2}} \end{aligned}$$

n 次元正規分布

多変量解析及び演習

この f を, y_i の確率密度関数と思わず, が測定済データ y_1, \dots, y_n が定数, β_0, σ が変数と思ったとき, 尤度 (ゆうど) 関数 $L(\beta_0, \sigma)$ という。

$$L(\beta_0, \sigma) = f(y_1, y_2, \dots, y_n | \beta_0, \sigma)$$

最尤推定

岩薩林 確率・統計なし

最尤推定

β_0, σ の推定値として $L(\beta_0, \sigma)$ が最大になる値を選ぶ

2変数関数の最大値 \rightsquigarrow 偏微分

微積分 II

$$0 = \frac{\partial L}{\partial \beta_0}(\beta_0, \sigma) = \frac{\partial L}{\partial \sigma}(\beta_0, \sigma)$$

ここでは最初の等式だけ解く．合成微分．

$$0 = (\text{定数})^{-n/2} \times \sum_{i=1}^n \frac{1}{\sigma^2} (y_i - \beta_0) \times e^{\text{同じ}}$$

$$0 = \sum_{i=1}^n (y_i - \beta_0)$$

$$\beta_0 = \frac{1}{n} \sum_{i=1}^n y_i$$

ここでは最初の式だけ．
合成微分．

β_0 (母平均値) の最尤推定
(MLE=maximum likelihood
estimation) 値 $\hat{\beta}_0$ は標本平均値
じゃん

L13-Q3

Quiz(正規分布の母数の最尤推定)

$$Y = \beta_0 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Y, ϵ : 連続型確率変数, $\sigma > 0, \beta_0$: 定数=パラメタ=母数である.

- ① サイズ 2 の標本 y_1, y_2 が得られたとき, 尤度関数 $L(\beta_0, \sigma)$ を具体的に書こう.
- ② 連立方程式 $\frac{\partial L}{\partial \beta_0}(\beta_0, \sigma) = 0, \frac{\partial L}{\partial (\sigma^2)}(\beta_0, \sigma) = 0$ を解いて, σ^2 を最尤推定しよう.

線形モデルとしての回帰分析

このドーナツ製造機で作るドーナツの重さ Y は、温度 x によるらしい。
次の線形回帰モデルを仮定する。

岩薩林 確率・統計 §9.1(9.2)

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

Y, ϵ : 連続型確率変数, β_0, β_1 : 回帰係数, $\sigma > 0$: 定数=パラメタ=母数,

Y : 目的変数 (従属変数) ここでは確率変数

x : 説明変数 (独立変数) ここでは確率変数でない

ノイズ・誤差 $\epsilon = Y - \beta_0 + \beta_1 \cdot x \sim N(0, \sigma^2)$.

$\epsilon = Y - \beta_0 - \beta_1 x \sim N(0, \sigma^2)$ より、ドーナツの重さ y を得る確率密度は、

$$f(y|x, \beta_0, \beta_1, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\beta_0-\beta_1 \cdot x)^2}{2\sigma^2}}$$

(正確に) 指定した温度 x_i ($i = 1, \dots, n$) で製造したときの重さが y_i ($i = 1, \dots, n$) である確率密度は, 独立分布なので積.

$$\begin{aligned} f(y_1, y_2, \dots, y_n | x_1, \dots, x_n, \beta_0, \beta_1, \sigma) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 - \beta_1 \cdot x_i)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{-n/2} e^{-\sum_{i=1}^n \frac{(y_i - \beta_0 - \beta_1 \cdot x_i)^2}{2\sigma^2}} \\ &= L(\beta_0, \beta_1, \sigma) \end{aligned}$$

最尤推定

測定済のデータ x_i, y_i $i = 1, \dots, n$ を定数と思ったときの、3変数関数

$L(\beta_0, \beta_1, \sigma) = f(y_1, y_2, \dots, y_n | x_1, \dots, x_n, \beta_0, \beta_1, \sigma)$ の最大値は? \rightsquigarrow 偏微分 微積分 II

$$0 = \frac{\partial L}{\partial \beta_0}(\beta_0, \beta_1, \sigma) = \frac{\partial L}{\partial \beta_1}(\beta_0, \beta_1, \sigma) = \frac{\partial L}{\partial \sigma}(\beta_0, \beta_1, \sigma)$$

ここでは最初の2つの等式だけ解く。

$$0 = \frac{\partial L}{\partial \beta_0} = (\text{定数}) \sum_i \frac{1}{\sigma^2} (y_i - \beta_0 - \beta_1 x_i) \times e^{\text{同じ}}$$

$$0 = \frac{\partial L}{\partial \beta_1} = (\text{定数}) \sum_i \frac{1}{\sigma^2} x_i (y_i - \beta_0 - \beta_1 x_i) \times e^{\text{同じ}}$$

しよせん、 β_0, β_1 の連立1次方程式

正規方程式 岩薩林 確率・統計 §9.2

$$n\beta_0 + \left(\sum_i x_i\right)\beta_1 = \sum_i y_i$$

$$\left(\sum_i x_i\right)\beta_0 + \left(\sum_i x_i^2\right)\beta_1 = \sum_i x_i y_i$$

加減法 \rightsquigarrow

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}, \quad y = \frac{s_{xy}}{s_{xx}}x + \bar{y} - \frac{s_{xy}}{s_{xx}}\bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \frac{s_{xy}}{s_{xx}}\bar{x} \quad y - \bar{y} = \frac{s_{xy}}{s_{xx}}(x - \bar{x})$$

ここで,

$$\bar{x} = \frac{1}{n} \sum_i x_i \quad \text{平均値っぽい形}$$

$$\bar{y} = \frac{1}{n} \sum_i y_i \quad \text{平均値っぽい形}$$

$$s_{xy} = \frac{1}{n} \sum_i x_i y_i - \bar{x} \cdot \bar{y} \stackrel{\text{岩薩林 確率・統計定理 1.5}}{=} \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \quad \text{共分散っぽい形}$$

$$s_{xx} = \frac{1}{n} \sum_i x_i^2 - \bar{x}^2 \stackrel{\text{岩薩林 確率・統計定理 1.2}}{=} \frac{1}{n} \sum_i (x_i - \bar{x})^2 \quad \text{分散っぽい形}$$

$Y = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \epsilon$ ($\hat{\beta}_0, \hat{\beta}_1$: 回帰係数) に代入.

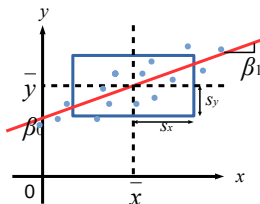
回帰直線

岩薩林 確率・統計 §9.1

推定結果 $\hat{\beta}_0, \hat{\beta}_1$ を係数とする xy 平面の直線

データ分析

$$y - \bar{y} = \frac{s_{xy}}{s_{xx}}(x - \bar{x})$$



回帰係数, 予測値の信頼区間

確率統計 II,III