

データ解析

サンプルサイズ設計1



創域理工学部

Faculty of Science and Technology

東京理科大学
創域理工学部情報計算科学科
安藤宗司

2023年12月14日

Contents

- サンプルサイズ設計とは？
 - 仮説検定の誤り
 - 検出力

- 母平均の差に関する2標本検定問題
 - 分散が既知の場合
 - 分散が未知の場合

- 推定精度を保証する方法
 - 母平均の推定

手元にあるコインはいかさまコインかどうか

- 表が出る確率 π は $1/2$ かどうか
- 実際にコインを N 回投げて、確かめる実験を考える

表 表 表 表 表 表 ... 表 n 回

裏 裏 裏 裏 裏 裏 ... 裏 $N - n$ 回

- 表が出た割合 $p = \frac{n}{N}$

コイン投げの実験を仮説検定で検証

□ 仮説の設定

- 表が出る確率は $1/2$ かどうか

□ 設定した仮説を評価するためのデータを収集

- 実際にコインを N 回投げる

↳ N を設定することをサンプルサイズ設計という

□ 事前に設定した判定基準に基づき判断

- 表が m 回以下（または以上）のとき、いかさまコインと判断

↳ この判定基準を確率論的に設定する

仮説の設定

- 「表が出る確率 π は $1/2$ かどうか」を検証するために2つの仮説を設定する
 - 帰無仮説 $H_0: \pi = 1/2$
 - 対立仮説 $H_1: \pi \neq 1/2$

- 帰無仮説が成り立つと仮定する
 - 手元にあるコインはいかさまコインではないと仮定する
 - 収集したデータに基づき帰無仮説が成り立つかどうかを判断する

設定した仮説を評価するためのデータを収集

- 「帰無仮説」と「対立仮説」のどちらが正しいかを判断するために、コインを N 回投げる
 - N （サンプルサイズ）はどう設定すればいいのだろうか？

- 検出力に基づいてサンプルサイズを設計する
 - 検出力は検定の精度を表す指標
 - 詳しくは後ほど紹介

事前に設定した判定基準に基づき判断

□ 判断基準の考え方

- いかさまコインではないとき
10回コインを投げれば常に表が5回出るとは限らない
- 表が出る回数は確率的に変動する
- 偶然に出る可能性のある「表の回数」の範囲を考える

□ この範囲を確率論的に設定する

偶然に出る可能性のある「表の回数」の範囲

□ いかさまコインではない（帰無仮説 $H_0: \pi = 1/2$ ）と仮定

表の回数	確率
0	0.1%
1	0.98%
2	4.39%
3	11.72%
4	20.51%
5	24.51%
6	20.51%
7	11.72%
8	4.39%
9	0.98%
10	0.1%

確率の計算式 ${}_{10}C_x \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{10-x}$

表の回数3から7である確率は85%以上

表の回数が1以下，または9回以上である確率は2.16%
表の回数が2以下，または8回以上である確率は10.94%

有意水準

- 帰無仮説のもとで、5%（または1%）未満でしか起きない事象は偶然ではないと考える

- 10回コインを投げた結果
 - 表の回数が1以下，または9回以上の場合 ➡ 偶然ではない
 - 表の回数が2以上，または8回以下の場合 ➡ 偶然である

検定結果の解釈

□ 10回コインを投げた結果

■ 表の回数が1以下，または9回以上の場合

- 統計学的に有意と判定
- 帰無仮説を棄却して，対立仮説を採択する
- 「表が出る確率 π は1/2ではない」と判断する

■ 表の回数が2以上，または8回以下の場合

- 統計学的に有意でないと判定
- 帰無仮説を採択する
- 「表が出る確率 π は1/2ではない」とはいえないと判断する

「表が出る確率 π は1/2である」とは判断できないことに注意！

仮説検定には誤りが存在する

- 帰無仮説のもとで5%未満の確率でしか起きない事象は偶然ではないと考えて有意水準を設定
- 裏を返せば，帰無仮説のもとでも，5%未満の確率で生じる事象ということになる
- 第1種の過誤
 - 帰無仮説が正しいときに，誤って帰無仮説を棄却する誤り
 - 第1種の誤りを起こす確率を第1種の過誤確率

		検定結果	
		帰無仮説が正しいと判断	対立仮説が正しいと判断
真実	帰無仮説が正しい	正しい	第1種の誤り
	対立仮説が正しい	第2種の誤り	正しい

2種類の誤り確率

- 仮説検定では、第1種の誤りと第2種の誤りが存在
- 有意水準を設定することで第1種の過誤確率を制御している
- 第2種の過誤確率はどのように制御するのか？

いかさまコインであると仮定

- これまでは、帰無仮説（いかさまコインではない）が成り立つと仮定して議論してきた
- いかさまコインである仮定して、表が出る回数の確率を求める
- 検出力
 - 対立仮説が正しいとき、対立が正しいと判断する確率
 - $1 - \text{第2種の過誤確率}$

表が出る確率が70%のいかさまコイン

いかさまコインではない場合

表の回数	確率
0	0.1%
1	0.98%
2	4.39%
3	11.72%
4	20.51%
5	24.51%
6	20.51%
7	11.72%
8	4.39%
9	0.98%
10	0.1%

第1種の
過誤確率

第1種の
過誤確率

いかさまコインの場合

表の回数	確率
0	0.0006%
1	0.01%
2	0.15%
3	0.90%
4	3.68%
5	10.29%
6	20.01%
7	26.68%
8	23.35%
9	12.11%
10	2.82%

検出力
(無視可能)

第2種の
過誤確率

検出力

表が出る確率が80%のいかさまコイン

いかさまコインではない場合

表の回数	確率
0	0.1%
1	0.98%
2	4.39%
3	11.72%
4	20.51%
5	24.51%
6	20.51%
7	11.72%
8	4.39%
9	0.98%
10	0.1%

第1種の
過誤確率

第1種の
過誤確率

いかさまコインの場合

表の回数	確率
0	1.024e-05%
1	0.0004%
2	0.007%
3	0.08%
4	0.55%
5	2.64%
6	8.81%
7	20.13%
8	30.20%
9	26.84%
10	10.74%

検出力
(無視可能)

第2種の
過誤確率

検出力

表が出る確率が90%のいかさまコイン

いかさまコインではない場合

表の回数	確率
0	0.1%
1	0.98%
2	4.39%
3	11.72%
4	20.51%
5	24.51%
6	20.51%
7	11.72%
8	4.39%
9	0.98%
10	0.1%

第1種の
過誤確率

第1種の
過誤確率

いかさまコインの場合

表の回数	確率
0	1.024e-05%
1	9e-07%
2	3.645e-05%
3	0.0009%
4	0.01%
5	0.15%
6	1.12%
7	5.74%
8	19.37%
9	38.74%
10	34.87%

検出力
(無視可能)

第2種の
過誤確率

検出力

12回コインを投げた結果

いかさまコインではない場合

表の回数	確率
0	0.02%
1	0.29%
2	1.61%
3	5.37%
4	12.08%
5	19.34%
6	22.56%
7	19.34%
8	12.08%
9	5.37%
10	1.61%
11	0.29%
12	0.02%

第1種の
過誤確率

第1種の
過誤確率

表が出る確率が80%のいかさまコインの場合

表の回数	確率
0	4.096e-07%
1	1.96608e-05%
2	0.0004%
3	0.006%
4	0.05%
5	0.33%
6	1.55%
7	5.32%
8	13.29%
9	23.62%
10	28.34%
11	20.62%
12	6.87%

検出力
(無視可能)

第2種の
過誤確率

検出力

検出力の特徴

□ 帰無仮説からの乖離の程度に依存する

■ コインのいかさまの程度（表の出る確率）に依存する

表の出る確率	検出力
70%	14.93%
80%	37.58%
90%	73.61%

□ サンプルサイズ N に依存する

コイン投げの回数	表の出る確率	検出力
10	80%	37.58%
12	80%	55.83%

諸前提

□ 試験治療 T と対照治療 C の応答変数の母平均の差に関する両側検定を考える

■ 試験治療 T の応答変数 $X_i \underset{\text{i.i.d}}{\sim} N(\mu_1, \sigma^2)$ ($i = 1, 2, \dots, n$)

■ 対照治療 C の応答変数 $Y_j \underset{\text{i.i.d}}{\sim} N(\mu_2, \sigma^2)$ ($j = 1, 2, \dots, m$)

□ 母平均が大きいことが臨床的に望ましい状態とする

■ $\varepsilon = \mu_1 - \mu_2 > 0$ が臨床的に望ましい

母平均の差に関する2標本検定

□ 試験治療の対照治療に対する優越性を検証

□ 仮説

■ 帰無仮説 $H_0: \varepsilon = 0$

■ 対立仮説 $H_1: \varepsilon \neq 0$

□ 応答変数の標本平均

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \bar{Y} = \frac{1}{m} \sum_{j=1}^m Y_j$$

□ 割付比

■ 各群の参加者数の比を $r = m/n$ とし, $\kappa = r/(r + 1)$ とする

相似検定

複合仮説	複合仮説
帰無仮説 $H_0: \theta \in \Theta_0$	対立仮説 $H_1: \theta \in \Theta_1 = \Theta \cap \Theta_0^c$
	$\Theta = \Theta_0 \cup \Theta_0^c$ かつ $\Theta_0 \cap \Theta_0^c = \phi$

すべての $\theta_0 \in \Theta_0$ に対して,

$$\beta_W(\theta_0) = P_{\theta_0}((X_1, X_2, \dots, X_n) \in W) = \alpha$$

を満たす棄却域 W を用いた検定を相似検定 (similar test) という

相似検定の例

□ 母集団1

- 平均 μ_1 （未知），分散 σ^2 （既知）の正規母集団
- 無作為標本 X_1, X_2, \dots, X_n

□ 母集団2


- 平均 μ_2 （未知），分散 σ^2 （既知）の正規母集団
- 無作為標本 Y_1, Y_2, \dots, Y_m

□ 仮説

- $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$

この統計的仮説検定に対する相似検定を構成する

両側検定の相似検定の構成 (1)


$$\begin{aligned} X_i &\underset{\text{i.i.d}}{\sim} N(\mu_1, \sigma^2) \quad (i = 1, \dots, n) \\ &\quad \text{分散}\sigma^2 \text{ (既知)} \\ Y_j &\underset{\text{i.i.d}}{\sim} N(\mu_2, \sigma^2) \quad (j = 1, \dots, m) \\ &\quad \text{分散}\sigma^2 \text{ (既知)} \end{aligned}$$

$$\begin{aligned} \bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu_1, \frac{\sigma^2}{n}\right) \\ \bar{Y} &= \frac{1}{m} \sum_{j=1}^m Y_j \sim N\left(\mu_2, \frac{\sigma^2}{m}\right) \end{aligned}$$

$\bar{X} \perp \bar{Y}$ であることから

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

両側検定の相似検定の構成 (2)

$H_0: \mu_1 = \mu_2$ のもとで

$$\bar{X} - \bar{Y} \sim N\left(0, \sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)\right), \quad Z \equiv \frac{\bar{X} - \bar{Y}}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1)$$

$H_1: \mu_1 \neq \mu_2$ のもとで

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)\right), \quad Z \equiv \frac{\bar{X} - \bar{Y}}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N\left(\frac{\mu_1 - \mu_2}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}, 1\right)$$

両側検定の相似検定の構成 (3)

棄却域 W を次のようにする。

$$W = \left\{ (x_1, x_2, \dots, x_n; y_1, y_2, \dots, y_m) \mid |Z| > z\left(\frac{\alpha}{2}\right) \right\}$$

$H_0: \mu_1 = \mu_2$ のもとで

$$\beta_W(\mu_1, \mu_2) = P_{\theta_0}((X_1, X_2, \dots, X_n; Y_1, Y_2, \dots, Y_m) \in W) = \alpha$$

第1種の過誤確率

$$\Leftrightarrow \int_{|Z| > z\left(\frac{\alpha}{2}\right)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{Z^2}{2}\right) dZ = \alpha$$

この棄却域 W に基づく検定は相似検定である

両側検定の相似検定の構成 (4)

$H_1: \mu_1 \neq \mu_2$ のもとで

$$\beta_W(\mu_1, \mu_2) = P_{\theta_1}((X_1, X_2, \dots, X_n; Y_1, Y_2, \dots, Y_m) \in W)$$

検出力

$$\Leftrightarrow \int_{|Z| > z(\frac{\alpha}{2})} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(Z - \frac{\mu_1 - \mu_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right)^2\right) dZ$$

分散が既知の場合

□ 次式が成立するときに帰無仮説 H_0 を有意水準 α で棄却

$$\left| \frac{\bar{X} - \bar{Y}}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \right| \geq z_{\alpha/2} \Leftrightarrow \left| \frac{\bar{X} - \bar{Y}}{\frac{\sigma}{\sqrt{nk}}} \right| \geq z_{\alpha/2}$$

$z_{\alpha/2}$: 標準正規分布の上側 $\alpha/2$ % 点

$$\begin{aligned} \sigma \sqrt{\frac{1}{n} + \frac{1}{m}} &= \sigma \sqrt{\frac{1}{n} + \frac{1}{nr}} \\ &= \sigma \sqrt{\frac{1}{n} \left(1 + \frac{1}{r} \right)} = \sigma \sqrt{\frac{1}{n} \frac{r+1}{r}} \\ &= \sigma \sqrt{\frac{1}{n} \frac{1}{\kappa}} = \frac{\sigma}{\sqrt{nk}} \end{aligned}$$

□ サンプルサイズ設計 (検出力の計算)

■ $\alpha, r, \varepsilon, \sigma^2$ はある値に固定する

検出力の計算

$\Phi(\cdot)$: 標準正規分布の累積分布関数

□ 対立仮説 $H_1: \varepsilon \neq 0$ のもとで、検出力は n の関数になる

$$\begin{aligned}\phi(n) &= P\left(\left|\frac{\bar{X} - \bar{Y}}{\sigma/\sqrt{n\kappa}}\right| \geq z_{\alpha/2}\right) = P\left(\frac{\bar{X} - \bar{Y}}{\sigma/\sqrt{n\kappa}} \geq z_{\alpha/2}\right) + P\left(\frac{\bar{X} - \bar{Y}}{\sigma/\sqrt{n\kappa}} \leq -z_{\alpha/2}\right) \\&= P\left(\frac{\bar{X} - \bar{Y} - \varepsilon}{\sigma/\sqrt{n\kappa}} \geq z_{\alpha/2} - \frac{\varepsilon}{\sigma/\sqrt{n\kappa}}\right) + P\left(\frac{\bar{X} - \bar{Y} - \varepsilon}{\sigma/\sqrt{n\kappa}} \leq -z_{\alpha/2} - \frac{\varepsilon}{\sigma/\sqrt{n\kappa}}\right) \\&= \left(1 - \Phi\left(z_{\alpha/2} - \frac{\varepsilon}{\sigma/\sqrt{n\kappa}}\right)\right) + \Phi\left(-z_{\alpha/2} - \frac{\varepsilon}{\sigma/\sqrt{n\kappa}}\right) \\&= \Phi\left(\frac{\varepsilon}{\sigma/\sqrt{n\kappa}} - z_{\alpha/2}\right) + \Phi\left(-z_{\alpha/2} - \frac{\varepsilon}{\sigma/\sqrt{n\kappa}}\right) \\&\approx \Phi\left(\frac{\varepsilon}{\sigma/\sqrt{n\kappa}} - z_{\alpha/2}\right) \quad \left(\because \frac{\varepsilon}{\sigma/\sqrt{n\kappa}} \gg 0\right)\end{aligned}$$

サンプルサイズ設計

□ 標準正規分布の上側 $\beta\%$ 点を z_β とする

■ $1 - \beta = \Phi(z_\beta)$

□ 検出力 $1 - \beta$ を満たす n は、次式の解として与えられる

$$\frac{\varepsilon}{\sigma/\sqrt{n\kappa}} - z_{\alpha/2} = z_\beta \Leftrightarrow n = \frac{(z_{\alpha/2} + z_\beta)^2}{\kappa \left(\frac{\varepsilon}{\sigma}\right)^2} \quad m = rn$$

n に小数点以下の単数が含まれるので、切り上げることで試験に必要な参加者数を得る。

検出力が $1 - \beta$ 以上となる最小の自然数として得られる。

分散が未知の場合

□ 次式が成立するときに帰無仮説 H_0 を有意水準 α で棄却

$$\left| \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n} + \frac{1}{m}}}\right| \geq t_{\alpha/2, \nu} \Leftrightarrow \left| \frac{\bar{X} - \bar{Y}}{\frac{s}{\sqrt{nk}}}\right| \geq t_{\alpha/2, \nu}$$

分散 σ^2 の不偏推定量

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2}{n + m - 2}$$

$t_{\alpha/2, \nu}$: 自由度 $\nu = n + m - 2 (= n + nr - 2)$ の t 分布の上側 $\alpha/2\%$ 点

$$\left| \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n} + \frac{1}{m}}}\right| \geq z_{\alpha/2} \Leftrightarrow \left| \frac{\bar{X} - \bar{Y}}{\frac{\sigma}{\sqrt{nk}}}\right| \geq z_{\alpha/2}$$

$$\frac{(\bar{X} - \bar{Y})}{\frac{\sigma}{\sqrt{nk}}} \sim N(0, 1)$$

$$\frac{s^2 \nu}{\sigma^2} \sim \chi^2_\nu$$

自由度 ν のカイ二乗分布

分散 σ^2 が未知のため使用 ✕

□ サンプルサイズ設計 (検出力の計算)

■ $\alpha, r, \varepsilon, s^2$ はある値に固定する

$$\frac{\frac{\bar{X} - \bar{Y}}{\sigma/\sqrt{nk}}}{\sqrt{\frac{s^2 \nu}{\sigma^2} \frac{1}{\nu}}} = \frac{\bar{X} - \bar{Y}}{\frac{s}{\sqrt{nk}}} \sim t_\nu$$

自由度 ν の t 分布 30

検出力の計算 (1)

□ 対立仮説 $H_1: \varepsilon \neq 0$ のもとで, 検出力は n の関数になる

$$\begin{aligned}\phi(n) &= P\left(\left|\frac{\bar{X} - \bar{Y}}{s/\sqrt{n\kappa}}\right| \geq t_{\alpha/2, \nu}\right) = P\left(\left|\frac{(\bar{X} - \bar{Y})\sqrt{n\kappa}/\sigma}{s/\sigma}\right| \geq t_{\alpha/2, \nu}\right) \\ &= P\left(\left|\frac{\frac{(\bar{X} - \bar{Y})}{\sigma}}{\frac{s/\sigma}{\sqrt{n\kappa}}}\right| \geq t_{\alpha/2, \nu}\right) = P(|T'| \geq t_{\alpha/2, \nu}) = P(T' \geq t_{\alpha/2, \nu}) + P(T' \leq -t_{\alpha/2, \nu})\end{aligned}$$

$$\frac{(\bar{X} - \bar{Y})}{\sigma} \sim N(\theta, 1)$$
$$\theta = \frac{\varepsilon}{\sigma/\sqrt{n\kappa}}$$

$$\frac{s^2 \nu}{\sigma^2} \sim \chi_\nu^2$$

自由度 ν のカイ二乗分布

$$T' \sim t(\nu; \theta)$$

自由度 ν , 非心度 θ の非心 t 分布

検出力の計算 (2)

$T_{\nu;\theta}(\cdot)$: $t(\nu; \theta)$ の累積分布関数

□ 対立仮説 $H_1: \varepsilon \neq 0$ のもとで、検出力は n の関数になる

$$\begin{aligned}\phi(n) &= P(T' \geq t_{\alpha/2, \nu}) + P(T' \leq -t_{\alpha/2, \nu}) \\ &= \left(1 - T_{\nu; \theta}(t_{\alpha/2, \nu})\right) + \left(T_{\nu; \theta}(-t_{\alpha/2, \nu})\right) \\ &\approx \left(1 - T_{\nu; \theta}(t_{\alpha/2, \nu})\right) \quad \left(\because \theta = \frac{\varepsilon}{\sigma/\sqrt{n\kappa}} \gg 0\right)\end{aligned}$$

サンプルサイズ設計

- 検出力 $1 - \beta$ を満たす n は、次式の解として与えられる

$$1 - \beta = 1 - T_{v;\theta}(t_{\alpha/2,v})$$

- 計算法（反復計算）

- 実際に試験に必要な参加者数を求めるには、 n を逐次的に変化させて、与えられた n のもとで検出力を計算
- 検出力が $1 - \beta$ 以上となる最小の自然数まで反復計算

推定精度を保証する方法 (1)

□ 母平均の推定

■ 応答変数 $X_i \underset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$ ($i = 1, 2, \dots, n$)

■ 応答変数の標本平均

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

□ 分散 σ^2 が既知の場合の母平均 μ の信頼区間

$$(L, U) = \left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

推定精度を保証する方法 (2)

□ 信頼区間幅が一定値 γ 以下になる n を求める

$$\begin{aligned}\left(\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) &= 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \gamma \\ \Leftrightarrow \sqrt{n} &\geq \frac{2z_{\alpha/2}\sigma}{\gamma} \\ \Leftrightarrow n &\geq \left(\frac{2z_{\alpha/2}\sigma}{\gamma}\right)^2\end{aligned}$$

この不等式を満たす最小の n が試験に必要な参加者数