

統計学II

早稲田大学政治経済学術院

西郷 浩

本日の目標

- 2標本問題
- 母平均の差の検定
- 成功確率(母比率)の差の検定
- 無作為割付

利用データについて

- JGSS2010、JGSS2000
 - 抽出実験に用いたデータは、東京大学社会科学研究所付属社会調査・データアーカイブ研究センターSSJデータアーカイブのリモート集計システムを利用し、同データアーカイブが所蔵する[JGSS2010, JGSS2000]の個票をデータを集計したものである。

2標本問題の例

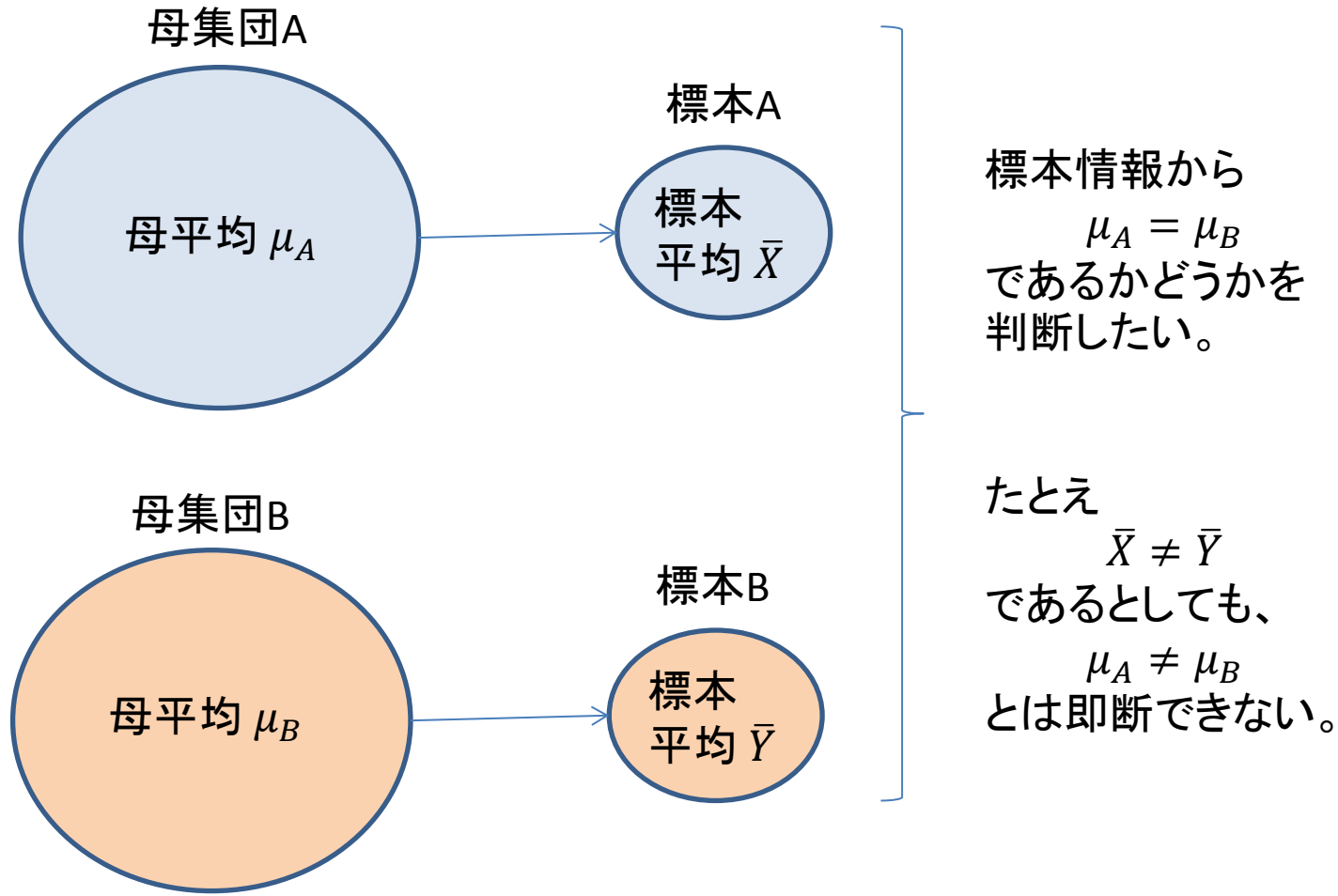


図1: 2標本問題の例

母平均の差の検定(1)

- 2標本問題

- 2つの独立標本

- 標本A $X_i \sim iid(\mu_A, \sigma_A^2) \ i = 1, 2, \dots, n.$

- 標本B $Y_j \sim iid(\mu_B, \sigma_B^2) \ j = 1, 2, \dots, m.$

- 仮説の設定

- $$\begin{cases} H_0: \mu_A = \mu_B \Leftrightarrow \mu_A - \mu_B = 0 \\ H_1: \mu_A \neq \mu_B \Leftrightarrow \mu_A - \mu_B \neq 0 \end{cases}$$

母平均の差の検定(2)

– 検定統計量と検定手続き(標本サイズが両方とも大きいとき)

$$\bullet T = \frac{\bar{X} - \bar{Y} - (\mu_A - \mu_B)}{\sqrt{\frac{S_A^2}{n} + \frac{S_B^2}{m}}} \rightarrow_d N(0, 1) \quad (\text{正規近似})$$

$$\begin{aligned} - \bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i, \bar{Y} = \frac{1}{m} \sum_{j=1}^m Y_j, S_A^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \\ S_B^2 &= \frac{1}{m-1} \sum_{j=1}^m (Y_j - \bar{Y})^2 \end{aligned}$$

$$\bullet \text{もし、} H_0 \text{ が正しければ、} T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_A^2}{n} + \frac{S_B^2}{m}}} \rightarrow_d N(0, 1)$$

• 両側検定(有意水準0.05)

– もし、 $|T| > 1.96$ であれば、 H_0 を棄却する。

母平均の差の検定(3)

- 例：平均就労年数の差 (p. 175)
 - 標本A: JGSS2010から、 $n = 50$ の無作為標本
 - $\bar{X}_{obs} = 17.1, S_{A obs}^2 = 211.9$
 - 標本B: JGSS2000から、 $m = 50$ の無作為標本
 - $\bar{Y}_{obs} = 13.3, S_{B obs}^2 = 107.4$
 - 検定統計量の値
 - $|T|_{obs} = \frac{|17.1 - 13.3|}{\sqrt{\frac{211.9}{50} + \frac{107.4}{50}}} = 1.50 < 1.96$
 - 検定の結果
 - H_0 は棄却されない。
 - 平均就労年数に差があるとはいえない。

母平均の差の検定(4)

- 2標本問題(分散が等しい2つの正規母集団を想定した場合)
 - 2つの正規独立標本
 - 標本A $X_i \sim iid N(\mu_A, \sigma^2) \ i = 1, 2, \dots, n.$
 - 標本B $Y_j \sim iid N(\mu_B, \sigma^2) \ j = 1, 2, \dots, m.$
 - 仮説の設定
 - $\begin{cases} H_0: \mu_A = \mu_B \Leftrightarrow \mu_A - \mu_B = 0 \\ H_1: \mu_A \neq \mu_B \Leftrightarrow \mu_A - \mu_B \neq 0 \end{cases}$

母平均の差の検定(5)

- 検定統計量と検定手続き(標本サイズが両方とも大きいとき)

- $T = \frac{\bar{X} - \bar{Y} - (\mu_A - \mu_B)}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{1}{m} \right)}} \sim t(n + m - 2) \quad (\text{厳密に})$

- $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \bar{Y} = \frac{1}{m} \sum_{j=1}^m Y_j, \hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2}{n + m - 2} = \frac{(n-1)S_A^2 + (m-1)S_B^2}{(n-1) + (m-1)}$

- もし、 H_0 が正しければ、 $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{1}{m} \right)}} \sim t(n + m - 2)$

- 両側検定(有意水準0.05)

- もし、 $|T| > t_{0.025}(n + m - 2)$ であれば、 H_0 を棄却する。

母平均の差の検定(6)

- 例 : Karl Pearson の親子の身長 (教科書と違う数値例)
 - 標本A: 10人の父親の無作為標本
 - 63, 65, 66, 68, 70, 66, 64, 66, 71, 69.
 - $\bar{X}_{A\text{ obs}} = 66.8, S_{A\text{ obs}}^2 = 6.84.$
 - 標本B: 10人の息子の無作為標本
 - 71, 69, 67, 72, 64, 72, 70, 69, 74, 72.
 - $\bar{Y}_{B\text{ obs}} = 70.0, S_{B\text{ obs}}^2 = 8.44.$
 - 検定統計量の値
 - $|T|_{\text{obs}} = \frac{|66.8 - 70.0|}{\sqrt{\frac{(10-1)6.84 + (10-1)8.44}{10+10-2} \times \left(\frac{1}{10} + \frac{1}{10}\right)}} = 2.588 >$
 $t_{0.025}(10 + 10 - 2) = 2.101$
 - H_0 は棄却される。

成功確率(母比率)の差の検定(1)

- 2標本問題(Berunoulli分布の場合)
 - 2つの独立標本
 - 標本A $X_i \sim_{iid} \text{Bernoulli}(p_A)$ $i = 1, 2, \dots, n$.
 - 標本B $Y_j \sim_{iid} \text{Bernoulli}(p_B)$ $j = 1, 2, \dots, m$.
 - 仮説の設定
 - $\begin{cases} H_0: p_A = p_B \Leftrightarrow p_A - p_B = 0 \\ H_1: p_A \neq p_B \Leftrightarrow p_A - p_B \neq 0 \end{cases}$

成功確率(母比率)の差の検定(2)

－ 検定統計量と検定手続き(標本サイズが両方とも大きいとき)

$$\bullet T = \frac{\hat{p}_A - \hat{p}_B - (p_A - p_B)}{\sqrt{\frac{p_A(1-p_A)}{n} + \frac{p_B(1-p_B)}{m}}} \sim N(0, 1) \quad (\text{正規近似})$$

$$- \hat{p}_A = \frac{1}{n} \sum_{i=1}^n X_i, \hat{p}_B = \frac{1}{m} \sum_{j=1}^m Y_j$$

• もし、 H_0 が正しければ、 $p_A = p_B = p$ と記して

$$T = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim N(0, 1)$$

成功確率(母比率)の差の検定(3)

- p の推定 (H_0 が正しいとしたときの成功確率の推定)

$$- \hat{p} = \frac{\sum_{i=1}^n X_i + \sum_{j=1}^m Y_j}{n+m} = \frac{n\hat{p}_A + m\hat{p}_B}{n+m}$$

» 標本Aと標本Bを合併して求めた経験的成功確率

- 検定統計量と検定手続き(有意水準0.05)

$$- |T| = \frac{|\hat{p}_A - \hat{p}_B|}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

– もし、 $|T| > 1.96$ であれば、 H_0 を棄却する。

成功確率(母比率)の差の検定(4)

- 例: 2つの地域の罹患率 (p. 180)
 - 2つの独立標本
 - 標本A: 地域Aから無作為標本200人(罹患者50人)
 - 標本B: 地域Bから無作為標本200人(罹患者65人)
 - 両地域で罹患率に差があるといえるか?
 - 検定統計量の値
 - $\hat{p}_{A\text{ obs}} = \frac{50}{200}, \hat{p}_{B\text{ obs}} = \frac{65}{200}, \hat{p}_{\text{ obs}} = \frac{50+65}{200+200} = \frac{115}{400}$

成功確率(母比率)の差の検定(5)

$$\bullet |T|_{obs} = \frac{\left| \frac{50}{200} - \frac{65}{200} \right|}{\sqrt{\frac{115}{400} \left(1 - \frac{115}{400} \right) \left(\frac{1}{200} + \frac{1}{200} \right)}} = 1.66 < 1.96$$

－ 検定の結果

- 両地域の罹患率に差があるとはいえない。

無作為割付(1)

- 例：インフルエンザと投薬
 - インフルエンザの薬
 - 患者が投薬を選べるとする。
 - 服用・非服用ごとの治癒の状況

表1a: 重症の患者

	治癒	非治癒	合計
服用	40	40	80
非服用	0	20	20
合計	40	60	100

表1b: 軽症の患者

	治癒	非治癒	合計
服用	20	0	20
非服用	60	20	80
合計	80	20	100

出所: 狩野裕(2002)「構造方程式モデリング、因果推論、そして非正規性」(甘利俊一他(2002)『多変量解析の展開』岩波書店 91ページ。)

無作為割付(2)

- 全患者(重症+軽症)における服用・非服用ごとの治癒の状況

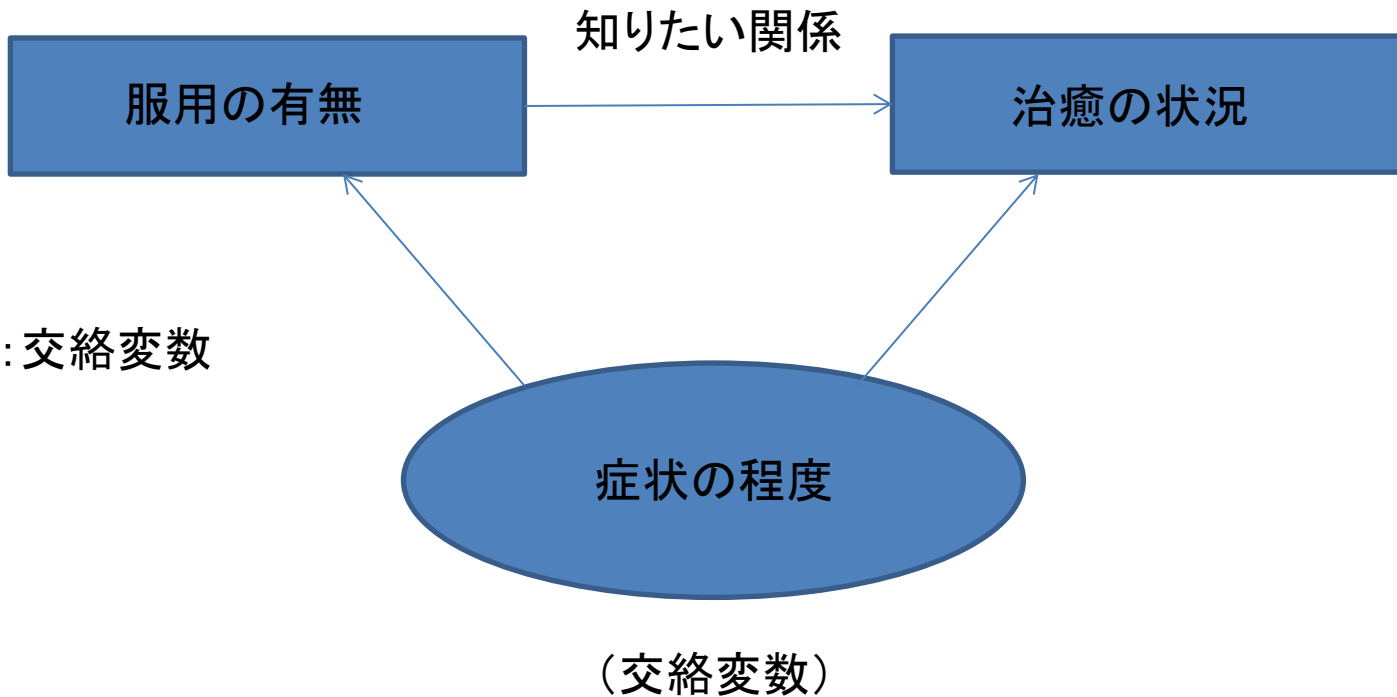
表2: 全患者

	治癒	非治癒	合計
服用	60	40	100
非服用	60	40	100
合計	120	80	200

- 服用・非服用は治癒の状況と独立(薬に効き目がない)ように見える。
 - » Simpson の逆説と呼ばれる。

無作為割付(3)

－ 交絡変数



無作為割付(4)

- 交絡変数の制御(調整)
 - 交絡変数が観察可能な場合
 - 直接的な調整
 - 例:グループ分け
 - 交絡変数が観察不可能な場合
 - 間接的な調整
 - 無作為割付
- 無作為割付
 - どの患者にも一定の確率で薬を投与する。
 - 例:五分五分の確率
 - 重症の患者にも五分五分の確率で投与される。
 - 軽症の患者にも五分五分の確率で投与される。
 - 投与するグループ(処置群)と投与しないグループ(制御群)とで、症状の分布がほぼ同じになる。

無作為割付(5)

－ 無作為割付の結果

- どの患者にも確率50%で薬を投与する。
- 表1a, b の確率で治癒・非治癒が発生するとする。

表3a: 重症の患者

	治癒	非治癒	合計
服用	25	25	50
非服用	0	50	50
合計	25	75	100

表3b: 軽症の患者

	治癒	非治癒	合計
服用	50	0	50
非服用	37.5	12.5	50
合計	87.5	12.5	100

表4: 全患者

	治癒	非治癒	合計
服用	75	25	100
非服用	37.5	62.5	100
合計	112.5	87.5	200

無作為割付(6)

－ 無作為割付の結果の解釈

- 服用者の治癒率： 75/100
- 非服用者の治癒率： 37.5/100
 - － 症状の程度が観察不可能な場合にも、薬効の有無を検定できる。

－ 無作為割付の効果

- 服用者グループと非服用者グループの症状の程度（および他の要因）を両者間で平均的に均質化できる。
 - － 処置群：服用者のグループ
 - － 制御群：非服用者のグループ