中級統計学：後期定期試験

村澤 康友

2021 年 1 月 26 日

注意： 3 問とも解答すること．結果より思考過程を重視するので，途中計算等も必ず書くこと（部分点は大いに与えるが，結果のみの解答は 0 点とする）．教科書のみ参照してよい（他の講義資料・ノートは持込不可）．

1. （20 点）以下の用語の定義を式または言葉で書きなさい（各 20 字程度）．
   (a) 帰無仮説
   (b) 有意水準
   (c) 通常の最小 2 乗法（OLS）
   (d) ダミー変数
2. （30 点）ゴルトンは身長の遺伝を研究した．両親の平均身長と成人した子供の身長（女性の身長は 1.08 倍して男性に換算）の無作為標本を $((x_1, y_1), \ldots, (x_n, y_n))$ とする（単位はインチ）．$\ln y_i$ の $\ln x_i$ 上への古典的正規単回帰モデルは
$$\ln y_i | \ln x_i \sim \mathrm{N}\left(\alpha + \beta \ln x_i, \sigma^2\right)$$
   $\beta < 1$ となる現象を「平均への回帰」という．回帰分析の結果は次の通りであった．
$$\widehat{\text{l\_child}} = \underset{(0.17499)}{1.49881} + \underset{(0.041431)}{0.644298}\,\text{l\_parent}$$
$$T = 928 \quad \bar{R}^2 = 0.2062 \quad F(1, 926) = 241.84 \quad \hat{\sigma} = 0.033050$$
   (丸括弧内は標準誤差)

   (a)「子供の身長」の「両親の平均身長」に対する弾力性の OLS 推定値・標準誤差・t 値は幾らか？
   (b)「平均への回帰」の有無の検定問題を定式化しなさい．
   (c)「平均への回帰」の有無の検定統計量の値を求め，有意水準 5 %の検定を実行しなさい．
3. （50 点）Go To トラベル事業の 2020 年 8 月末までの利用者と非利用者で，9 月末までに発熱症状があった人の割合を p_X, p_Y とする．p_X と p_Y を比較したい．独立に抽出した大きさ n_X, n_Y の無作為標本で，発熱症状があった人の割合を $\hat{p}_X, \hat{p}_Y$ とする．
   (a) 検定問題を定式化しなさい（問題意識を踏まえること）．
   (b) 2 項母集団 $\mathrm{Bin}(1, p_X)$, $\mathrm{Bin}(1, p_Y)$ の平均と分散を求めなさい．
   (c) $\hat{p}_X, \hat{p}_Y, \hat{p}_X - \hat{p}_Y$ の漸近分布を求めなさい．
   (d) 検定統計量を定義し，その H_0 の下での分布から有意水準 5 %の検定の棄却域を定めなさい．
   (e) $n_X = 2500$, $n_Y = 6400$, $\hat{p}_X = .05$, $\hat{p}_Y = .04$ として検定統計量の値と漸近 p 値を求め，有意水準 5 %の検定を実行しなさい．

   ※数値例はフィクションです．この分析は Go To トラベル事業と発熱症状の相関関係の検証であり，結果を因果関係と解釈するのは誤りです．

解答例

1. 統計学の基本用語
   (a) とりあえず真と想定する仮説.
   (b) 許容する第 1 種の誤りの確率.
   (c) 残差 2 乗和を最小にするように回帰係数を定める方法.
   (d) あるカテゴリーに入るなら 1，入らないなら 0 とした変数.
      - 0 か 1 をとる変数に変換するのがポイントなので，「0 か 1 をとる変数」のみは 1 点減.
2. 単回帰分析
   (a) OLS 推定値は.644298，標準誤差は.0414309，t 値は.644298/.0414309=15.55.
      - OLS 推定値と標準誤差は各 3 点，t 値は 4 点.
      - t 値＝ OLS 推定値／標準誤差としていれば OK.
   (b)
   $$H_0 : \beta = 1 \text{（}\alpha, \sigma^2 \text{ は任意）} \quad \text{vs} \quad H_1 : \beta < 1 \text{（}\alpha, \sigma^2 \text{ は任意）}$$
   (c) 検定統計量は
   $$\begin{aligned} t &:= \frac{b-1}{s} \\ &= \frac{.644298 - 1}{.0414309} \\ &\approx -8.5854 \end{aligned}$$
   H_0 の下で $t \sim \mathrm{t}(926)$ より（近似的な）棄却域は $(-\infty, -1.645]$．検定統計量が棄却域に入るので，H_0 を棄却して H_1 を採択．すなわち「平均への回帰」は存在する．
      - 検定統計量で 5 点，棄却域で 5 点.
3. 母比率の差の検定
   (a)
   $$H_0 : p_X = p_Y \quad \text{vs} \quad H_1 : p_X > p_Y$$
      - 両側検定は 2 点.
   (b) $\mathrm{Bin}(1, p_X)$ の平均は
   $$1 \cdot p_X + 0 \cdot (1 - p_X) = p_X$$
   分散は
   $$\begin{aligned} (1-p_X)^2 \cdot p_X + (0 - p_X)^2 \cdot (1 - p_X) &= (1-p_X)^2 p_X + p_X^2 (1 - p_X) \\ &= p_X(1-p_X) \end{aligned}$$
   $\mathrm{Bin}(1, p_Y)$ についても同様.
   (c)
   $$\begin{aligned} \hat{p}_X &\overset{a}{\sim} \mathrm{N}\left(p_X, \frac{p_X(1-p_X)}{n_X}\right) \\ \hat{p}_Y &\overset{a}{\sim} \mathrm{N}\left(p_Y, \frac{p_Y(1-p_Y)}{n_Y}\right) \\ \hat{p}_X - \hat{p}_Y &\overset{a}{\sim} \mathrm{N}\left(p_X - p_Y, \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}\right) \end{aligned}$$

- $\hat{p}_X, \hat{p}_Y$ は各 3 点，$\hat{p}_X - \hat{p}_Y$ は 4 点.

(d) 標準化すると

$$\frac{\hat{p}_X - \hat{p}_Y - (p_X - p_Y)}{\sqrt{p_X(1-p_X)/n_X + p_Y(1-p_Y)/n_Y}} \overset{a}{\sim} \mathrm{N}(0,1)$$

検定統計量は

$$Z := \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p}_X(1-\hat{p}_X)/n_X + \hat{p}_Y(1-\hat{p}_Y)/n_Y}}$$

棄却域は $[1.645, \infty]$.

- 検定統計量で 5 点，棄却域で 5 点.

(e)

$$\begin{aligned}
\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} &= \frac{(1/20)(1-1/20)}{2500} \\
&= \frac{(1/20)(19/20)}{50^2} \\
&= \frac{19}{1000^2} \\
\frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y} &= \frac{(1/25)(1-1/25)}{6400} \\
&= \frac{(1/25)(24/25)}{80^2} \\
&= \frac{24}{25^2 80^2} \\
&= \frac{6}{25^2 40^2} \\
&= \frac{6}{1000^2}
\end{aligned}$$

したがって

$$\begin{aligned}
\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y} &= \frac{19}{1000^2} + \frac{6}{1000^2} \\
&= \frac{25}{1000^2}
\end{aligned}$$

すなわち

$$\begin{aligned}
\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} &= \frac{5}{1000} \\
&= \frac{1}{200}
\end{aligned}$$

検定統計量は

$$\begin{aligned}
Z &:= \frac{.05 - .04}{1/200} \\
&= 2
\end{aligned}$$

漸近 p 値は.02275．したがって有意水準 5 %で H_0 を棄却する.

- 検定統計量と p 値は各 4 点，検定は 2 点.