確率・統計の初歩 (5)

別所俊一郎

2006 年 4 月 28 日

いくつかの訂正

- 「t 統計量は使わない」わけではありません.「t 分布を使わない」が正しい表現でした.
- p 値は第 1 種の過誤に対応している概念ですが，第 2 種の過誤に対応する概念として**検定の検出力 power** というものがあります.

判定	H_0 が真	H_1 が真
H_0 を棄却	第 1 種の過誤	:-)
H_0 を受容	:-)	第 2 種の過誤

信頼区間 confidence interval

- 標本の抽出は確率的なので，母平均の正確な値を知ることはできない
- ある確率（**信頼水準 confidence level**）で真の値が含まれるような集合（**信頼集合 confidence set**）を見つけることはできる．1 変数のケースで，信頼集合が上限と下限を持つ実数の集合であるときには**信頼区間 confidence interval** ともいう
- 信頼集合を求めるには以下の手続きを繰り返せばよい
  1. 候補となる実数 $\mu_{Y,0}$ を取り上げ，$H_0 : \mu_Y = \mu_{Y,0}$ と $H_1 : \mu_Y \neq \mu_{Y,0}$ を設定
  2. H_0 に対応する p 値を求め，(1 − 信頼水準) よりも大きければ，この実数 $\mu_{Y,0}$ を書き留めておく
  3. また別の実数を候補として取り上げる

信頼区間の設定方法

- 候補となる実数をひとつひとつ取り上げていく方法では，真の値が信頼集合に含まれる確率は信頼水準に等しい
- ただし，この方法は候補を全て検討するという点であまりに非現実的
- 実際には,「95%両側検定では標本平均より $1.96 \times SE(\overline{Y})$ 以上遠いと信頼集合から外される」という性質を用いて,

$$\overline{Y} - 1.96 \times SE(\overline{Y}) \leq \overline{Y} \leq \overline{Y} + 1.96 \times SE(\overline{Y})$$

を 95%信頼区間とする.

異なる母集団の平均の検定 (1)

2 つの異なる母集団の分布の平均の差の検定を考える．2 つのグループをそれぞれ添え字 m, w で表す

- 2 つの母平均の差が d_0 であるという帰無仮説を立てる．母平均が等しいという仮説なら $d_0 = 0$

$$H_0 : \mu_m - \mu_w = d_0 \quad \text{v.s.} \quad H_1 : \mu_m - \mu_w \neq d_0$$

- 母平均の差 $\mu_m - \mu_w$ の推定量として $\overline{Y}_m - \overline{Y}_w$ を考える．これを用いて検定を行うにはこの分布を知る必要がある
- いま，中心極限定理を用いると，

$$\overline{Y}_m \xrightarrow{d} N\left(\mu_m, \frac{\sigma_m^2}{n_m}\right) \quad \overline{Y}_w \xrightarrow{d} N\left(\mu_w, \frac{\sigma_w^2}{n_w}\right)$$

異なる母集団の平均の検定 (2)

- $\overline{Y}_m$ と $\overline{Y}_w$ は異なる母集団からの標本から計算された標本平均であるから，互いに独立した確率変数と考えることができ，正規分布の性質を用いると，

$$\overline{Y}_m - \overline{Y}_w \xrightarrow{d} N\left(\mu_m - \mu_w, \frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}\right)$$

- 通常は σ_m^2 と σ_w^2 は未知だから，一致推定量としての標本分散 s_m^2，s_w^2 を用いると，標準誤差は

$$SE(\overline{Y}_m - \overline{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$$

であるから，標準化して検定統計量を求めると，H_0 のもとで

$$t = \frac{(\overline{Y}_m - \overline{Y}_w) - d_0}{SE(\overline{Y}_m - \overline{Y}_w)} \xrightarrow{d} N(0,1), \quad \text{p-value} = 2\Phi(-|t|)$$

異なる母集団の平均の差の信頼区間

- $d = \mu_m - \mu_w$ の信頼区間を形成する
- ほぼ先ほどと同様の手続きによって，95%信頼区間は

$$\left\{(\overline{Y}_m - \overline{Y}_w) \pm 1.96 \times \mathrm{SE}(\overline{Y}_m - \overline{Y}_w)\right\}$$

2 変数の関係の捉え方

散布図 平面上の点 (X_i, Y_i) で各観測値を代表させるグラフ

標本共分散 標本分散と同様に $n-1$ で割ると，一致推定量となる．

$$s_{XY} \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \xrightarrow{p} \sigma_{XY}$$

標本相関 n 個の観測値の線形のつながりの強さを表す

$$r_{XY} \equiv \frac{s_{XY}}{s_X s_Y} \xrightarrow{p} \mathrm{corr}(X, Y)$$

- $|r_{XY}| \leq 1$ が成り立つ
- $|r_{XY}| = 1$ であれば，散布図上の点は直線上に並ぶ．右上がりなら $r_{XY} = 1$，右下がりなら $r_{XY} = -1$．直線上に近いほど絶対値が 1 に近付く
- 2 変数に非線形な関係があるとき，相関係数がゼロに近くなるケースがある（Fig. 3.3）