2 標本ノンパラメトリックモデルにおける平均比の統計解析法

2016SS060 大竹 友香
指導教員：白石 高章

1 はじめに

2 標本のデータ解析において，平均の差の推測論を考えることが多い．しかしながら，(第 1 標本の平均)/(第 2 標本の平均) の推測を考えることにより，第 1 標本の平均が第 2 標本の平均の何倍であるかが分かる．本研究では，平均の比の推測論について考察する．はじめに信頼区間を求め，検定方式を与える．これを基に解析手法の C 言語プログラムを作成し, 名古屋市内にある飲食店の売上データを使って，着目する 2 つの月の，平均の比の信頼区間を求め，検定を行うことで 2 標本の平均の比を分析できる．

2 モデルの設定

$(X_1,..,X_{n_1})$, $(Y_1,...,Y_{n_2})$ をある 2 つの連続型分布に従う母集団からのそれぞれの大きさが n_1, n_2 の無作為標本とし，$E(X_i)=\mu_1$, $V(X_i)=\sigma_1^2$, $E(Y_j)=\mu_2$, $V(Y_j)=\sigma_2^2$ とし．分布関数はそれぞれ $P(X_i \leqq x)=F((x-\mu_1)/\sigma_1)$, $P(Y_j \leqq x)=F((x-\mu_2)/\sigma_2)$ とする．ただし，$F(x)$ は平均 0 分散 1 の未知の分布関数である．

このとき，平均の不偏推定量と分散の不偏推定量は,

$$\hat{\mu}_1 \equiv \bar{X}, \quad \hat{\mu}_2 \equiv \bar{Y}$$

$$\tilde{\sigma}_1^2 \equiv \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\bar{X})^2,$$

$$\tilde{\sigma}_2^2 \equiv \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_j-\bar{Y})^2$$

で与えられる．ただし,

$$\bar{X} \equiv \frac{1}{n_1}\sum_{i=1}^{n_1}X_i, \quad \bar{Y} \equiv \frac{1}{n_2}\sum_{j=1}^{n_2}Y_j,$$

とする．

3 平均比の推測論

平均の比 μ_1/μ_2 の推測法について考察する．次の条件 (c) を仮定する．

$$\text{条件 } (c): \qquad \lim_{n\to\infty}\frac{n_1}{n}=\lambda \ \ (0<\lambda<1).$$

ただし，$n \equiv n_1+n_2$ とする．
白石 [1] の定理 3．3 5を適用して次の補題 1 を得る．

補題 1 条件 (c) の下で，$n\to\infty$ として,

$$\begin{aligned}\sqrt{n}\left(\log\frac{\bar{X}}{\bar{Y}}\right)-\sqrt{n}\left(\log\frac{\mu_1}{\mu_2}\right) &\xrightarrow{\mathcal{L}} \frac{1}{\mu_1}\sigma_1\sqrt{\frac{1}{\lambda}}Z_1-\frac{1}{\mu_2}\sigma_2\sqrt{\frac{1}{1-\lambda}}Z_2 \\ &\sim N\left(0,\frac{\sigma_1^2}{\mu_1^2\lambda}+\frac{\sigma_2^2}{\mu_2^2(1-\lambda)}\right)\end{aligned}$$

が成り立つ．ただし，$\xrightarrow{\mathcal{L}}$ は法則収束を表し, Z_1，Z_2 は互いに独立で同一の $N(0,1)$ に従う確率変数とする．

定理 2 条件 (c) の下で，$n\to\infty$ として,

$$\frac{\sqrt{n}\left(\log\frac{\bar{X}}{\bar{Y}}\right)-\sqrt{n}\left(\log\frac{\mu_1}{\mu_2}\right)}{\tilde{\eta}_n} \xrightarrow{\mathcal{L}} Z\sim N(0,1)$$

が成り立つ．ただし,

$$\tilde{\eta}_n \equiv \sqrt{\frac{\tilde{\sigma}_1^2}{\hat{\mu}_1^2\frac{n_1}{n}}+\frac{\tilde{\sigma}_2^2}{\hat{\mu}_2^2\left(1-\frac{n_1}{n}\right)}}$$

とする．
定理 2 より, 次の定理 3 を得る．

定理 3 このとき, 標準正規分布の上側 $100(\alpha/2)\%$ 点を $z(\alpha/2)$ とする．

μ_1/μ_2 に対する $100(1-\alpha)\%$ の漸近的信頼区間として,

$$\frac{\bar{X}}{\bar{Y}}\exp\left\{\frac{-z\left(\alpha/2\right)\tilde{\eta}_n}{\sqrt{n}}\right\}<\frac{\mu_1}{\mu_2}<\frac{\bar{X}}{\bar{Y}}\exp\left\{\frac{z\left(\alpha/2\right)\tilde{\eta}_n}{\sqrt{n}}\right\}$$

が提案できる．

4 検定方式

ここで, 帰無仮説 $H_0^a: \mu_1/\mu_2=1$ vs. 対立仮説 $H_1^a: \mu_1/\mu_2\neq 1$ に対する水準 α の検定について考察する．
このとき，検定統計量を

$$S=\frac{\sqrt{n}\log\frac{\bar{X}}{\bar{Y}}}{\tilde{\eta}_n}$$

とおく．
検定統計量 S を用いた水準 α の検定方式を考える．帰無仮説 H_0^a の下で，$\log\frac{\mu_1}{\mu_2}=0$ であるので,
定理 2 により，H_0^a の下で,

$$S\xrightarrow{\mathcal{L}} Z\sim N(0,1) \tag{1}$$

を得る．ここで, 帰無仮説 $H_0^a: \mu_1/\mu_2=1$ vs. 対立仮説 $H_1^a: \mu_1/\mu_2\neq 1$ に対する水準 α の検定は次で与えられる．

$$\phi_1(\boldsymbol{X},\boldsymbol{Y})=\begin{cases}1 & (S<-z(\alpha/2) \text{ または } z(\alpha/2)<S)\\ 0 & (-z(\alpha/2)<S<z(\alpha/2))\end{cases}$$

$\Longleftrightarrow$

$$\begin{cases}H_0^a\text{を棄却する} & (S<-z(\alpha/2) \text{ または } z(\alpha/2)<S)\\ H_0^a\text{を棄却しない} & (-z(\alpha/2)<S<z(\alpha/2))\end{cases}$$

同様に，(1) より, 帰無仮説 $H_0^a: \mu_1/\mu_2=1$ vs. 対立仮説 $H_2^a: \mu_1/\mu_2>1$ に対する水準 α の検定は次で与えられる．

$$\phi_2(\boldsymbol{X},\boldsymbol{Y})=\begin{cases}1 & (S>z(\alpha))\\ 0 & (S<z(\alpha))\end{cases}$$

と表現される．同様に, (1) より, 帰無仮説 $H_0^a : \mu_1/\mu_2 = 1$ vs. 対立仮説 $H_3^a : \mu_1/\mu_2 < 1$ に対する水準 α の検定は次で与えられる．

$$\phi_3(\boldsymbol{X}, \boldsymbol{Y}) = \begin{cases} 1 & (S < -z(\alpha)) \\ 0 & (S > -z(\alpha)) \end{cases}$$

と表現される．

5　C 言語におけるプログラム解説

5.1　プログラムの流れ

1. 関数 average により標本平均を計算
2. 関数 muhat により平均の不偏推定量を計算
3. 関数 sigmatilde により分散の不偏推定量を計算
4. 関数 sinrai により信頼区間を計算
5. 関数 kentei により水準 α の検定を行い, 結果を出力する

5.2　main プログラム

```
int main(void)
{
    ave1 = average();
    ave2 = average();
    muH1 = muhat();
    muH2 = muhat();
    sigT1 = sigmatilde();
    sigT2 = sigmatilde();
    sinrai();
    kentei();

}
```

6　データ解析

3, 4節で提案した信頼区間と検定方式の C 言語プログラムによるプログラムを作成した. このプログラムを使って, 名古屋市内のカフェ＆バーの売上データを解析した.

ここでは水準 $\alpha = 0.05$ の検定とする.

両側検定, 片側検定, 両方の結果を出力させるプログラムを作成したため, 両方の結果を表示する.

検定結果：両側検定は常に行い, その場合は (I)

$\frac{\bar{X}}{\bar{Y}} > 1$ のとき右側検定を行い, その場合は (Ⅱ)

$\frac{\bar{X}}{\bar{Y}} < 1$ のとき左側検定を行い, その場合は (Ⅲ)

また棄却する場合は 1, 棄却しない場合は 0

両側検定を行い, 棄却された場合は (I) 1

表１は, 第１標本を２０１７年から２０１９年の８月の１日ごとの売上高 ９３個とし, 第２標本を２０１７年から２０１９年の１２月の１日ごとの売上高 ９０個とする.

売上, 客単価は棄却されたが客数は棄却されなかった.

表１　：８月と１２月の売上データで比較

項目	平均比	信頼区間	信頼区間の幅	検定結果
売上	0.816	(0.730,0.913)	0.183	(I) 1, (Ⅲ) 1
客数	0.930	(0.845,1.024)	0.179	(I) 0, (Ⅲ) 0
客単価	0.882	(0.831,0.937)	0.106	(I) 1, (Ⅲ) 1

つまり, 売上, 客単価の平均は, ８月より１２月の方が大きいといえる. 売上の信頼区間は客数より広い範囲で, 客単価の信頼区間は客数より狭い範囲で位置している.

表２は, 第１標本を２０１８年から２０１９年の８月の１日ごとのランチの売上高 ６１個とし, 第２標本を２０１８年から２０１９年の１２月の１日ごとのランチの売上高 ６０個とする.

表２　：８月と１２月のランチの売上データで比較

項目	平均比	信頼区間	信頼区間の幅	検定結果
売上	0.954	(0.833,1.092)	0.259	(I) 0, (Ⅲ) 0
客数	0.999	(0.876,1.138)	0.262	(I) 0, (Ⅲ) 0
客単価	0.959	(0.928,0.990)	0.062	(I) 1, (Ⅲ) 1

客単価の平均比は 0.959 と 1 に近い値だが棄却され, 売上, 客数は棄却されなかった. 客単価の平均は８月より１２月の方が大きいといえる. 客数の信頼区間は売上より広い範囲で, 客単価の信頼区間は売上より狭い範囲で位置している.

表３は, 第１標本を２０１８年から２０１９年の８月の１日ごとのディナーの売上高 ６１個とし, 第２標本は２０１８年から２０１９年の１２月の１日ごとのディナーの売上高 ６０個とする.

表３　：８月と１２月のディナーの売上データで比較

項目	平均比	信頼区間	信頼区間の幅	検定結果
売上	0.805	(0.674,0.962)	0.288	(I) 1, (Ⅲ) 1
客数	0.931	(0.797,1.090)	0.293	(I) 0, (Ⅲ) 0
客単価	0.961	(0.897,1.031)	0.134	(I) 0, (Ⅲ) 0

売上が棄却され, 客数, 客単価は棄却されなかった. つまり, 売上の平均は. ８月より１２月の方が大きいといえる. 客数の信頼区間は売上より広い範囲で, 客単価の信頼区間は売上より狭い範囲で位置している.

7　おわりに

本論文では,2 標本のノンパラメトリックモデルにおける平均比の統計解析法を提案した．信頼区間を求め, 検定を行うプログラムを C 言語で作成した. 実際のデータを用い, 結果の解析を行うことで理解を深めることができた.

参考文献

[1] 白石高章：『統計科学の基礎』．日本評論社, 東京,2012.