2 標本パラメトリックモデルにおける分散の比の統計解析法

2016SS022 加藤志織
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，分散の比の推測論を考えることが多い．本稿でも，(第 1 標本の分散)/(第 2 標本の分散) の推測を考えることにより，第 1 標本の分散が第 2 標本の分散の何倍であるかが分かる．本研究では，分散の比の推測論について考察する．データの従っている分布は予め判定しておくものとする．はじめに信頼区間を求め，次に検定方式を与える．これらの解析手法の C 言語プログラムを作成し，飲食店の売上データについて解析する．

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 $P(X_i \leqq x)=F_0((x-\mu_1)/\sigma_1)$, $P(Y_j \leqq x)=F_0((x-\mu_2)/\sigma_2)$ とする．ただし，$F_0(x)$ は平均 0 分散 1 の既知の分布関数である．

命題 1　$f_0(x)=\frac{dF_0(x)}{dx}$ とし，$\ell \equiv \int_{-\infty}^{\infty} x^4 f_0(x)dx$ とおくと，

$$\ell = \frac{E\{(X_i-\mu_1)^4\}}{\sigma_1^4} = \frac{E\{(Y_j-\mu_2)^4\}}{\sigma_2^4}$$

が成り立つ．

ℓ の値を以下の表 1 にまとめる．

表 1　各分布における ℓ(尖度 +3) の値

正規分布	指数分布
3	9
ラプラス分布	ロジスティック分布
6	4.2

また，平均の不偏推定量と分散の不遍推定量は,

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

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

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

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　統計的推測法

分散の比 σ_1^2/σ_2^2 に対する推測法を得るために漸近的性質を導く．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，$n=n_1+n_2$ とする．

補題 2　条件 (c) の下で，$i=1,2$ に対して,

$$\sqrt{\frac{n}{\ell-1}}\{\log(\tilde{\sigma}_i^2)-\log(\sigma_i^2)\} \xrightarrow{\mathcal{L}} R_i \sim N\left(0, \frac{1}{\lambda_i}\right)$$

が成り立つ．ただし，$\lambda_1=\lambda$，$\lambda_2=1-\lambda$ とする．

定理 3 条件 (c) の下で,

$$\frac{\sqrt{\frac{n}{\ell-1}}\left\{\log\left(\frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_2^2}\right)-\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right)\right\}}{\tilde{\eta}_n} \xrightarrow{\mathcal{L}} Z \sim N(0,1)$$

が成り立つ．ただし，$\tilde{\eta}_n \equiv \sqrt{\frac{n^2}{n_1 n_2}}$ とする．

定理 3 より，次の定理 4 を得る．

定理 4　σ_1^2/σ_2^2に対する $100(1-\alpha)\%$ の漸近的信頼区間として,

$$\frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_2^2}\exp\left\{-\frac{\sqrt{\ell-1}\tilde{\eta}_n}{\sqrt{n}}z(\alpha/2)\right\} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_2^2}\exp\left\{\frac{\sqrt{\ell-1}\tilde{\eta}_n}{\sqrt{n}}z(\alpha/2)\right\}$$

が提案できる．

次に，帰無仮説 $H_0^b: \frac{\sigma_1^2}{\sigma_2^2}=1$ vs. 対立仮説 $H_1^b: \frac{\sigma_1^2}{\sigma_2^2}\neq 1$ に対する水準 α の検定を考える．

$$T \equiv \frac{\sqrt{\frac{n}{\ell-1}}\log\left(\frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_2^2}\right)}{\tilde{\eta}_n}$$

とおく．

ここで，帰無仮説 H_0^b vs. 対立仮説 H_1^b に対する水準 α の検定は次で与えられる．

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

$\Longleftrightarrow$

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

帰無仮説 H_0^b vs. $H_2^b: \frac{\sigma_1^2}{\sigma_2^2}>1$ に対する水準 α の検定は

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

帰無仮説 H_0^b vs. $H_3^b: \frac{\sigma_1^2}{\sigma_2^2}<1$ に対する水準 α の検定は

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

4　C 言語プログラム解説

4.1　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 sigma により分散を計算
3. 関数 sigmatilde により分散の不偏推定量を計算
4. 関数 shinraikukan により分散比の信頼区間を計算し出力
5. 関数 kentei により検定を行い結果を出力する

4.2　main プログラム

```
float main(){
    ave1 = average();
    ave2 = average();
    sigT1 = sigmatilde();
    sigT2 = sigmatilde();
    heikinhi =  average() / average();
    bunsanhi = sigma() / sigma();
    shinraikukan();
    kentei();
}
```

5　データ解析

前節で提案した信頼区間と検定方式の C 言語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．

解析結果を以下の表に示す．ただし，$\alpha = 0.05$ とし，分布について，正規分布のとき，1，指数分布のとき，2，ラプラス (両側指数) 分布のとき，3，ロジスティック分布のとき，4 で表す．これらの分布の判別は，外部のソフト [3] で予め判定しておくものとする．平均比は，$\hat{\mu}_1/\hat{\mu}_2$，分散比は，$\tilde{\sigma}_1^2/\tilde{\sigma}_2^2$ とする．

また，検定結果については，両側検定，片側検定を行うプログラム作成したため，両方の結果を表に示す．両側検定は常に行い，その結果を [Ⅰ]，$\frac{\sigma_1^2}{\sigma_2^2} > 1$ のとき，片側 (右側) 検定を行い，その結果を [Ⅱ]，$\frac{\sigma_1^2}{\sigma_2^2} < 1$ のとき，片側 (左側) 検定を行い，その結果を [Ⅲ] で表し，帰無仮説 H_0^b を棄却するとき 1，棄却しないとき 0 で表すものとする．例えば，両側検定を行い棄却された場合には，[Ⅰ]1 と表される．

表 2 は，第 1 標本を 2017 年から 2019 年の 10 月の 1 日の売上高 92 個，第 2 標本を 2017 年から 2019 年の 12 月の 1 日の売上高 90 個について解析した結果である．同様に，客数，客単価について解析した結果も載せている．

表 2　10 月と 12 月の 1 日での比較

項目	分布	平均比	分散比	信頼区間	検定結果
売上高	1	0.72	0.37	(0.25, 0.56)	[Ⅰ]1, [Ⅲ]1
客数	1	0.82	0.38	(0.25, 0.58)	[Ⅰ]1, [Ⅲ]1
客単価	1	0.88	0.69	(0.46, 1.04)	[Ⅰ]0, [Ⅲ]1

売上高と客数は棄却されたが，客単価の分散比は 0.69 であるが棄却されなかった．つまり，前者では 12 月の分散が大きくなっていると言える．また，売上高と客数の信頼区間は同じような範囲になっている．

表 3 は，第 1 標本を 2018 年から 2019 年の 10 月のランチの売上高 61 個，第 2 標本を 2018 年から 2019 年の 12 月のランチの売上高 60 個について解析した結果である．同様に，客数，客単価について解析した結果も載せている．

表 3　10 月と 12 月のランチでの比較

項目	分布	平均比	分散比	信頼区間	検定結果
売上高	1	0.80	0.57	(0.34, 0.94)	[Ⅰ]1, [Ⅲ]1
客数	1	0.84	0.61	(0.37, 1.01)	[Ⅰ]0, [Ⅲ]1
客単価	4	0.95	1.48	(0.78, 2.80)	[Ⅰ]0, [Ⅱ]0

売上高は棄却されたが，客数と客単価は棄却されなかった．つまり，売上高は 12 月の分散が大きくなっていると言える．また，売上高と客数の信頼区間は同じような範囲になっている．

表 4 は，第 1 標本を 2018 年から 2019 年の 10 月のディナーの売上高 61 個，第 2 標本を 2018 年から 2019 年の 12 月のディナーの売上高 60 個について解析した結果である．同様に，客数，客単価について解析した結果も載せている．

表 4　10 月と 12 月のディナーでの比較

項目	分布	平均比	分散比	信頼区間	検定結果
売上高	2	0.72	0.42	(0.15, 1.14)	[Ⅰ]0, [Ⅲ]1
客数	1	0.81	0.46	(0.28, 0.77)	[Ⅰ]1, [Ⅲ]1
客単価	4	0.89	0.99	(0.52, 1.87)	[Ⅰ]0, [Ⅲ]1

客数は棄却されたが，売上高と客単価は棄却されなかった．つまり，客数は 12 月の分散が大きくなっていると言える．

6　考察

比較した 2 標本について，分散比からは 2 標本の分散に差があると考えられるものでも，検定を行うことで棄却されることが確認できるものもあった．また，平均比が大きいと分散比も比較的大きくなる傾向があることがわかった．

7　おわりに

本論文では，2 標本パラメトリックモデルにおける分散の比の統計解析法を考察し，また，上記の手法を基にデータ解析を行うための C 言語プログラムを作成した．それを活用し，実際に飲食店の売上データを用いて，前期と後期，特徴的な月, そして，ランチとディナーなどの 2 標本の分散の違いについて解析し，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.

[2] 『Asoft』.
http://www.st.nanzan-u.ac.jp/info/marble/book/asoft.html