データ解析基礎

第９回分散に関する検定

分散に関する推定と検定

- 分散に関する議論が必要になるのはどういうときか？
- 二つの母集団の分散が等しいことを確認したい．
  - t 検定では，母集団の分散が等しいことが前提．
- データの平均よりばらつきの方に関心がある．
  - 学生の履修度に差がないかどうかを知りたければ，成績の平均よりもばらつきの方が重要である．

本講義で扱う問題

3

- 一つの母集団の分散
  - ある母集団の分散が特定の値であるかどうか？
- 二つの無関係な母集団の分散
  - 二つの母集団の分散が等しいかどうか？
- 二つの関連した母集団の分散
  - 分散が等しいかどうか？

一つの母集団の分散に関する検定

4

- 仮説：
  - 帰無仮説 H0：$\sigma^2 = 25$
  - 対立仮説 H1：$\sigma^2 \neq 25$
- 検定のステップ
  - 標本（データ）に基づいて分散の推定量を求める.
  - 帰無仮説の下での標本分布（確率モデル）を用い，得られた分散推定値が棄却域に入っているかどうかを調べる.
- 標本分布として何を使えばよいか？

χ^2分布（カイ2乗分布）

5

- χ^2分布の定義
  - 標準正規分布に従う独立な確率変数 $X1, X2, \ldots, Xn$ があるとき，それらの2乗の和は自由度 n のχ^2 分布 (χ^2 (n)と表す) に従う.
- 例：自由度1の場合
  - 正規分布$N(\mu, \sigma^2)$に従う母集団から標本を取りだし標準化．標準化したデータを2乗した値が従う分布が自由度1の χ^2 分布である.

→ つまり，z-scoreの2乗は $\chi^2(1)$分布

χ^2分布

6

- 自由度2の場合
  - 正規分布 $\mathrm{N}(\mu, \sigma^2)$ に従う母集団から，独立に二つ標本を取り出す．それらを標準化して2乗して和をとる.

→　その和が従う分布が $\chi^2(2)$ 分布である.

- 自由度 n の場合
  - 同じことを n 個標本を取り出して計算する.

→　その和が従う分布が $\chi^2(n)$ 分布である.

χ^2分布

7

- χ^2分布の期待値と分散
  - 期待値：自由度と等しい．
  - 分散：2×自由度.
- χ^2分布の形をみてみよう．
  - MATLABのデモンストレーション

χ^2分布をどうやって使うのか？

8

- 分散の推定値とχ^2分布の関係は？
  - 真の分散をσ^2，n 個の標本から求めた不偏推定量を $\hat{\sigma}^2$とすると，

$$\frac{(n-1)\,\hat{\sigma}^2}{\sigma^2} = \frac{\Sigma i\,(Xi - \bar{X})^2}{\sigma^2}$$

は，自由度 n -1 のχ^2 分布に従う．

（その理由は，t 分布の自由度のところで説明済み）

χ^2分布に基づく検定

9

- 数表の見方：
  - 自由度と有意水準が決まれば，数表から限界点が求まるので，そこから棄却域を決めればよい．
- 検定の手続き：検定の進め方はこれまでと全く同じ．
  - 仮説を決定する（帰無仮説の分散の値を定める）．
  - 標本を n 個集めて，分散の不偏推定量を求める．
  - 不偏推定量と仮説の分散から検定統計量を求める．
  - 自由度 n - 1 の χ^2 分布の数表を用いて限界点を求め，棄却域を決定する．そして，判断を下す．

例題１：両側検定

10

- ある試験の点数分布の分散が25であることがわかっている．いま，ある新しい教育方法を受けた学生に同じ試験を課して，22人からデータをとったところ，分散の推定値が20であることがわかった．新しい教育方法が学力のばらつきを変化させる効果があるかどうかを有意水準1%で検定せよ．
- 仮説の設定
  - 帰無仮説H0: $\sigma^2 = 25$
  - 対立仮説H1: $\sigma^2 \neq 25$

例題１（続き）

11

- 検定統計量の計算

$$\chi 2 = \frac{(n - 1)\,\hat{\sigma}^2}{\sigma^2} = \frac{(22 - 1)\,20}{25} = 16.8$$

- 自由度21のχ2分布の限界点は，有意水準1%のときは，左側が8.03， 右側が41.401である．
- 検定統計量は棄却域に含まれていない．
- したがって，帰無仮説は棄却されない．

二つの無関係な母集団の分散の比較

12

- 仮説：
  - 帰無仮説 H0：$\sigma_1^2 = \sigma_2^2$
  - 対立仮説 H1：$\sigma_1^2 \neq \sigma_2^2$
- 検定のステップ
  - いままでと全く同じ．
  - 問題は標本分布として何を使えばよいかだけ．
- 分散を比較する場合はF分布を使う．

F分布

- F分布の定義：
  - Xが自由度 ν_1 の $\chi 2$ 分布，Yが自由度 ν_2 の$\chi 2$分布に従うとき，

$$F = \frac{X / \nu_1}{Y / \nu_2}$$

は，自由度 (ν_1, ν_2) のF分布に従う．

F分布

- F分布の確率密度関数の形状
  - MATLABのデモ
- 確率密度関数や累積分布関数の値は，簡単には計算できないので，通常，数表を使う．
- 期待値と分散

期待値 $\dfrac{\nu_2}{\nu_2 - 1}$　　分散 $\dfrac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$

分散の比較とF分布

15

- 分散の違いを検定するには，グループ1の分散の分布とグループ2の分散の分布を比較すればよい．
- グループ1, 2それぞれのついて，分散の不偏推定量を真の分散で割り，自由度をかけた値はχ^2分布に従うことがわかっている．
- 両者の真の分散が等しいと仮定すれば（つまり，帰無仮説の下では），両者の推定量の比は，F分布に従うことになる（分母分子で真の分散が相殺される）．

F検定に使う統計量

16

- 以下の量は各々自由度 n_1-1 と n_2-1 のχ^2分布に従う．

$$\frac{(n_1 - 1)\hat{\sigma}_1^2}{\sigma_1^2} \sim \chi^2(n_1-1)$$

$$\frac{(n_2 - 1)\hat{\sigma}_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$$

- それぞれを自由度で割った結果の比は，自由度n_1-1, n_2-1 の F 分布に従う．

$$\frac{\hat{\sigma}_1^2}{\sigma_1^2} \Big/ \frac{\hat{\sigma}_2^2}{\sigma_2^2} \sim F(n_1-1, n_2-1)$$

F検定に使う統計量

17

- 帰無仮説の下で $\sigma_1^2 = \sigma_2^2$ であるから，分散推定量の比 $\hat{\sigma}_1^2 / \hat{\sigma}_2^2$ が検定統計量になる．

片側検定の例

18

- 個別指導制度を導入することで，学力のばらつきが減少したかどうかを検定したい．通常の制度と個別指導制度の学生をそれぞれ75人ずつ集め，それぞれの分散の不偏推定量を求めたところ，125と100であった．有意水準1%で検定せよ．
- 仮説の設定：
  - 帰無仮説 H0：$\sigma_1^2 = \sigma_2^2$
  - 対立仮説 H1：$\sigma_1^2 > \sigma_2^2$

    したがって，片側検定となる．

例題（続き）

19

- 検定統計量の計算：

$$F = \hat{\sigma}_1^2 / \hat{\sigma}_2^2 = 125 / 100 = 1.25$$

- 有意水準1%であることから，自由度 (74, 74) の F分布の限界点を読みとる．
  - 自由度74は表にないので60で代用すると，F=1.836．
  - したがって，棄却域は F >1.836．
- 求めた分散比は棄却域に含まれていない．
- したがって，帰無仮説は棄却されない．

両側検定の例

20

- 以下の二群の分散が等しいかどうかを有意水準1%で検定せよ．
  - グループ1：データ数61，分散の不偏推定量10
  - グループ2：データ数41，分散の不偏推定量 8
- これは両側検定である．しかし，F分布の数表には左側の限界点が掲載されていない．どうするか？
- 分散比を検定しているので，分散が大きい方を分子にして検定すれば，右側だけを考えればよいことになる．

例題（続き）

21

- 検定統計量の計算
  - グループ1の方が分散が大きいので，

$$F = \hat{\sigma}_1^2 / \hat{\sigma}_2^2 = 10/8 = 1.25$$

- 有意水準を1%にしたので，自由度60, 40のF分布の右側限界点は2.184.
- F値は棄却域に含まれていないので，帰無仮説は棄却されない．
  - あるいは，F(60, 40) = 1.25 (p = 0.11) とかく．

二つの標本に関連のある場合

22

- 天下り的であるが，以下の統計量が帰無仮説の下で自由度 n - 2の t 分布に従うことを使う．

$$t = \frac{\hat{\sigma}_1^2 - \hat{\sigma}_2^2}{\mathrm{sqrt}\left(\frac{4\hat{\sigma}_1^2\hat{\sigma}_2^2}{n - 2}(1 - r_{12}^2)\right)}$$

  - r_{12} は二つのデータの相関係数．
    （$1 - r_{12}^2$はデータの2乗誤差に相当する）
  - その他の手続きは今までと全く同じ．

例題

- 学生の創造力のばらつきが brain storming によって拡がるかどうかを検定したい．いま，brain stormingの前後で創造力テストの成績を調べたところ，以下のデータが得られた．有意水準5%で検定せよ．ただし，相関係数は0.5である．
  - 学生数：62名
  - セッション前の分散の推定量：$\hat{\sigma}_1^2 = 25$
  - セッション後の分散の推定量：$\hat{\sigma}_2^2 = 36$

例題（続き）

- 仮説の設定
  - 帰無仮説H0：$\sigma_1^2 = \sigma_2^2$　対立仮説H1：$\sigma_1^2 < \sigma_2^2$
- 検定に使う統計量の計算

$$t = \frac{25 - 36}{\mathrm{sqrt}\left(\frac{4 \times 25 \times 36}{60} (1 - 0.5^2)\right)} = -1.640$$

- 棄却域の設定と結論
  - 自由度60のt 分布の左側の5%の限界点は-1.671
  - 統計量は限界点内なので, 帰無仮説は棄却されない．