データ解析基礎

第８回 二つの母集団の平均の比較

二つの母集団に関する平均の比較

- 二つの母集団の平均が等しいかどうか？
  - 調布市民の年収と府中市民の年収はどちらが高いか？
  - 東京の物価と大阪の物価はどちらが高いか？
  - 雨の日の試験点数と晴れの日の試験点数はどちらが高いか？
  - 薬を飲むと熱が下がるか？
- これらの問題は，いずれも二つの条件での値を比較しているので，平均値に関する検定になる．

関連のある標本と関連のない標本

3

- 二つの条件間で違いがあるかどうかを調べる心理実験
  - 同じ被験者を二つの条件に参加させる場合
    within-subject design
  - 条件ごとに異なる被験者を使う場合
    between-subject design
- これら二つの条件において，データ処理は同じように行なってよいのか？

関連のある標本と関連のない標本

4

- 関連のある標本：dependent (related) sample
  - 二つの母集団を構成する要素に関係がある（あるいは同一である）場合
  - within-subject の条件がこれにあたる.
- 関連のない標本：independent (unrelated) sample
  - 二つの母集団を構成する要素が無関係な場合
  - between-subjectの条件がこれにあたる.
- これら二つのあいだで統計的な処理の仕方は異なる.

関連のない標本の分布

5

- 二つの標本に関連がない場合

  平均値がどのような分布をするかを，二つの母集団それぞれ独立に計算できるため，議論が簡単になる．

- 母集団が一つのときと同様に，

  -分散既知の場合

  -分散未知の場合

  に分けて議論する．

分散が既知の場合

6

- 二種類のデータ

  - 正規分布 $N(\mu_1, \sigma^2)$ に従う母集団 X_1 からの N_1 個

  - 正規分布 $N(\mu_2, \sigma^2)$ に従う母集団 X_2 からの N_2 個

  （二つの母集団の分散は等しく σ^2 であるとする）

- データの平均値の差はどのように分布するか？

  → 以下の平均と分散をもつ正規分布に従う．

  平均 $\mu_0 = \mu_1 - \mu_2$ 分散 $\sigma_0^2 = (1 / N_1 + 1 / N_2)\ \sigma^2$

  あとは母集団一つの場合と同じ．例題は省略．

分散が未知の場合

7

- 少々ややこしい．分散は等しいとする．

1. 二つの集団の分散の不偏推定量をそれぞれ求める．
2. 二つの母集団をあわせた分散の推定値を求める．
3. 平均値の差が従う分布の分散を推定する．
4. t 分布に従う統計量を定める．
5. 有意水準から棄却域を求め，結論を出す．

Steps 1 and 2:

8

1. 二つの集団の分散の不偏推定量をそれぞれ求める．

$$\hat{\sigma}_1^2 = \Sigma\ (X_1 - \overline{X_1})^2\ /\ (N_1 - 1)$$

$$\hat{\sigma}_2^2 = \Sigma\ (X_2 - \overline{X_2})^2\ /\ (N_2 - 1)$$

2. 二つの母集団をあわせた分散の推定値を求める．

$$\hat{\sigma}^2 = \frac{N_1 - 1}{N_1 + N_2 - 2}\hat{\sigma}_1^2 + \frac{N_2 - 1}{N_1 + N_2 - 2}\hat{\sigma}_2^2$$

この式を直接求める方がわかりやすいかもしれない．

Steps 1 and 2:

9

3. **平均値の差が従う分布の分散を推定する．**

$$\hat{\sigma}_D^2 = \left(\frac{1}{N_1} + \frac{1}{N_2}\right)\hat{\sigma}^2$$

4. t **分布に従う統計量を求める．**

帰無仮説 $\mu_1 = \mu_2$ **の下で，二つの平均値の差から得られる量**

$$\frac{\overline{X_1} - \overline{X_2}}{\hat{\sigma}_D}$$

はt **分布に従う．**

Steps 4 and 5

10

4'. t**分布の自由度は自由度は**

母集団1**に関する自由度が**$N_1 - 1$

母集団2**に関する自由度が**$N_2 - 1$

であるから，それらの和になって，$\nu = N_1 + N_2 - 2$

5. **有意水準から棄却域を求め，結論を出す．**

これまでと同様に数表から棄却域を定めればよい．

例題：

11

- 野球部とテニス部の部員の体重に差があるかどうかを調べたい．アンケートの結果，以下のデータが得られた．有意水準1%で差があるかどうかを検定せよ．
  - 野球部
    回答数：62
    平均値：80Kg，分散の不偏推定量：15kg^2
  - 軟式テニス部
    回答数：62
    平均値：83Kg，分散の不偏推定量：20kg^2

例題の解法：

12

- Step 1: 分散の不偏推定量を求める．
  - すでに与えられている．
- Step 2: 全体の分散の推定量を求める．

$$\sigma^2$$
$$= (62 - 1) / (62 + 62 -2) \ * 15 + 61/122 * 20$$
$$= 17.5$$

- Step 3: 平均値の差の分散を求める．

$$\sigma_D^2 = (1 / 62 + 1/ 62) * 17.5 = 0.5645...$$

例題の解法（続き）

13

- Step 4: **検定に使う統計量を求める．**

  $(\bar{X}_1 - \bar{X}_2) / \sigma_D = (80 - 83) / 0.7513 = -3.99$

- Step 5: **有意水準**1%**，自由度**122**に対する** t **分布の限界点を求め，棄却域を定める．**
  - **自由度**122**の数表はないので，**120**で代用すると，棄却域は，** $t < -2.617$, $2.617 < t$ **であることがわかる．**
- **結論**
  - **「平均値に差がない」という帰無仮説は棄却される．**

p値を用いた判断

14

- **検定統計量は以下のように求めた．**

  $(\bar{X}_1 - \bar{X}_2) / \sigma_D = (80 - 83) / 0.7513 = -3.99$

- **これまでは有意水準を先に定めて結論を下していたが，実際の論文では，次のような表記をすることが多い．**

  $t(122) = -3.99$, $p < 0.001$

  - **これは，以下の二つのことを表している．**
    1. **自由度**122**の** t **分布に従う統計量の値**
    2. **その値での（逆向きの）累積確率（**p**値）が基準値** (0.001)**より小さいこと．**

p値を用いた判断

15

- p値は，数表か統計パッケージにより求める．
- 基準には，0.05 (5%), 0.01, 0.001 がよく使われる．
- 不等号で表さず，p値のそのまま記載する場合もある．
  - 上の例では，$p = 0.000113$
  - p値の大きさから有意性を直接判断できる．
    （値が小さいほど，有意の程度が高い）

平均値の差の区間推定

16

- 検定と区間推定は表裏一体であることから，以上の議論から平均の差の区間推定がすぐに導ける．
  - 以上の議論から，

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_D} \quad \text{（または} \quad \frac{\bar{D} - \mu_D}{\hat{\sigma}_D} \text{）}$$

  は，自由度 $N_1 + N_2 - 2$ の t 分布に従う．
- $\bar{X}_1 - \bar{X}_2$ と $\hat{\sigma}_D$はデータから求められるので，信頼係数に応じた区間を数表から読みとれば，平均値の差 $\mu_1 - \mu_2$ （つまり，μ_D）の区間推定ができる．

区間推定の例題：

17

- 以下のケースについて，信頼係数99%としたときの平均の差の信頼区間を求めよ．
  - グループ１：
    標本数：40
    平均値：62，分散の不偏推定量：81
  - グループ２：
    標本数：38
    平均値：50，分散の不偏推定量：78

例題の解法：

18

- 二つのグループをあわせた分散の推定量は，

$$\hat{\sigma}^2 = (40 - 1) / (40 + 38 - 2) * 81 + 37/76 * 78$$

$$= 6045 / 76$$

- 「平均の差」の分散の推定量は，

$$\hat{\sigma}_D^2 = (1 / 40 + 1 / 38) * 6045 / 76$$

$$= 4.082\ldots$$

  - したがって，標準偏差はsqrt(4.082...) = 2.020

例題（続き）

19

- 自由度76のt分布の信頼係数99%に対応する区間幅は
  - 数表より2.660である．
- 一方，データの平均値の差は12である．
  - したがって，求めるべき区間は，

    (12 - 2.660 * 2.020, 12 + 2.660 * 2.020)

  - すなわち，(6.63 , 17.37)　である．

仮定が満たされない場合

20

- 上で用いた統計量がt分布に従うための条件は，
  1. 帰無仮説が成立していること
  2. 母集団がいずれも正規分布に従っていること
  3. 二つの母集団の分散が等しいこと
- これらが満たされないときはどうなるか？

  1.は検定すべき仮説なので考える必要なし．

  問題となりそうなのは，2.と3.

t 検定の頑健性

21

- 母集団が正規分布に従っていない場合
  - データ数が30より多ければ通常は問題にならない．
- 二つの母集団の分散が等しくない場合．
  - 二つの母集団からのデータ数がほぼ等しければ，本質的な問題にはならない．
- 一般に，t検定は母集団の分布には左右されにくい．
  - 頑健であるといえる．

関連のある標本の場合

22

- 二つの母集団を構成する要素が対応している場合
  - 同じ被験者を使って二つの条件での結果を比べる．
  - 同じ装置を使って二つの条件での結果を比べる．
- 検定の方法：
  - 条件1，2でのデータ：X_i and Y_i
  - それらの差 $D_i = X_i - Y_i$

    → Dを一つの確率変数と考え，一つの母集団に関する t 検定を用いればよい．

例題：

23

- 安全運転のビデオを視聴することで，運転免許試験の成績がよくなるかどうかを検定したい．
- 10人の被験者にビデオを見る前と見た後でそれぞれ試験を受けてもらった結果，以下のような結果が得られた．ビデオの効果について有意水準1%で検定せよ．

No.	1	2	3	4	5	6	7	8	9	10
視聴前：	4	5	8	7	9	6	5	4	8	7
視聴後：	6	4	7	8	9	7	5	3	9	6

解答：

24

- 仮説
  - 帰無仮説 $\mu_D = 0$　　対立仮説 $\mu_D > 0$
- 被験者ごとに差を計算してみると，

No.	1	2	3	4	5	6	7	8	9	10
視聴前：	4	5	8	7	9	6	5	4	8	7
視聴後：	6	4	7	8	9	7	5	3	9	6
差：	2	-1	-1	1	0	1	0	-1	1	-1

- 差の平均および分散の不偏推定量は，それぞれ

$\bar{D} = 0.1,\qquad \hat{\sigma}^2 = 1.21$

例題の解法 2

25

- 「差の平均値」の分散は，

$$\sigma_D = \hat{\sigma} / \mathrm{sqrt}(n) = \mathrm{sqrt}(1.21 / 10) = 0.348$$

- したがって，統計検定量は

$$t(9) = (\bar{D} - \mu_D) / \sigma_D = (0.1 - 0) / 0.348 = 0.287$$

- 有意水準1%の片側検定のための棄却域を定める．
  - 自由度9の t 分布の数表より，限界点は2.821．
- 0.287 < 2.821であるから，帰無仮説は棄却されない．