データ解析基礎

第7回 一つの母集団の平均に関する
推定と検定

区間推定と点推定

- 点推定
  - 「母平均の値はいくつ」と値そのものを推定する．
- 区間推定
  - 「母平均はこの区間に含まれる」ことを推定する．
- データ解析では，区間推定が用いられることが多い．
  - 実験するたびに推定値は変化するため．
  - 真の値が含まれる区間の信頼レベルを指定できる．
- 学習アルゴリズムなどは点推定の考え方に基づく．
  - 入出力関係を最もよく近似できる係数を求める．

区間推定 (interval estimation)

3

- 区間推定：
  - 母集団分布を定めるパラメータが含まれる区間を信頼係数(confidence coefficient)を定めて推定する．
- 母平均の推定区間の信頼度：
  - 区間が広いほど，そこに含まれる確率は高くなる．
  - データ数が十分あれば，その確率は中心極限定理で決まる．

    標本平均は平均μ，標準偏差σ/sqrt(n)の正規分布．

標本平均の分布

4

- 信頼係数を指定すれば，その確率で標本平均が入る区間が定まる ．

$$P\left(\left|\ (\bar{X} - \mu) / (\sigma/\mathrm{sqrt}(N))\right| < a\right) = \alpha$$

（a は α から決まる）

- つまり，$\bar{X}$ は確率は α で次の区間に含まれる．

$$\mu - a \times \sigma/\mathrm{sqrt}(N) < \bar{X} < \mu + a \times \sigma/\mathrm{sqrt}(N)$$

信頼区間(confidence interval)

5

- ここで，標本平均と母平均の相対関係を考慮すると，母平均が含まれる区間が求まる．

$$\bar{X} - a \times \sigma/\mathrm{sqrt}(N) < \mu < \bar{X} + a \times \sigma/\mathrm{sqrt}(N)$$

これを信頼区間という．

- 信頼区間は信頼係数が大きいほど広くなる．
  - 可能性がある区間を広めにとる．

信頼区間の解釈

6

- 信頼区間は式

$$P\left(\left|\bar{X}/(\sigma/\mathrm{sqrt}(N)) - \mu\right| < a\right) = \alpha$$

で決まるが，これは，「μ が区間

$$\left[\bar{X} - a \times \sigma/\mathrm{sqrt}(N),\ \bar{X} + a \times \sigma/\mathrm{sqrt}(N)\right]$$

に含まれる確率が α である」という意味ではない．

- 正しい意味は，「同じ実験を何度も繰り返すと，そのうち平均して α の割合で真のパラメータ μ が区間に入っている」ということ．

区間推定の例題：

7

- ある入学試験の点数の標準偏差が15点であることがわかっているとする．受験生100人のデータを取り出したところ，その平均点は65点であった．
- 信頼係数を95%としたときの母平均の信頼区間は？
- 信頼係数を99%としたときの母平均の信頼区間は？

例題の解法：

8

- step 1: 数表や計算機を用いて，標準正規分布の下での信頼区間の幅（つまり，a)を求める．
  - 数表によれば，
    95%の信頼区間の幅（の半分）は1.96.
    99%の場合は2.58.
- step 2:データ数と母集団の分散に基づいて，平均点が従う分布の標準偏差を求める．
  母集団の標準偏差は15点，データ数は100．よって，標本平均の標準偏差は15/sqrt(100)=1.5点である．

例題の解法

9

- step3：step1, step2の結果から，信頼区間は
  - 95%の場合

    [65 - 1.96 x 1.5, 65 + 1.96 x 1.5] = [62.1, 67.9]
  - 99%の場合

    [65 - 2.58 x 1.5, 65 + 2.58 x 1.5] = [61.1, 68.9]

  となる．

平均値に関する仮説検定

10

- 例題１：

  語学試験TOEFLは，平均500点，標準偏差100点になるように点数が決められる．平均が本当に500点に等しいかどうかを調べてみたい．

- 定式化
  - 帰無仮説H0: 母平均 $\mu = 500$
  - 対立仮説H1: 母平均 $\mu \neq 500$

    標準偏差は100であることがわかっているとする．

平均値がある値と異なることを検定したい

11

- 帰無仮説から予想される標本平均の分布

  中心極限定理から，標本平均は平均 μ，標準偏差 σ/sqrt(*n*)の正規分布に従う．

- したがって，確率の値を決めれば，標本平均がその区間に含まれる確率がわかる．

  → 採択域，棄却域が定まる．

- あとは，以前の例題と同様にして検定すればよい．

例題１の解法

12

- 受験生1600人の平均点を調べたところ507.2点だった．このとき，帰無仮説は棄却されるか？

- step 1: 平均点分布の標準偏差は，中心極限定理から

  100 / sqrt(1600) = 2.5

- step 2: 両側に棄却域を設け，5%の有意水準を設定．

  - 棄却域の境界の z 値は，およそ2 (1.96) である．

  - したがって，棄却域は 500 ± 2.5 x 2 の外側となる．

- step 3: 以上より，帰無仮説は棄却されることになる．

例題２

13

- A社のボールペンの寿命の平均は100時間，標準偏差は 9時間である．新しい製造工程で製品を400個作り，その平均寿命を計測したところ，101.5時間であった．新しい製品の寿命は延びているといえるだろうか？有意水準5%で検定せよ．

解法

14

- 仮説の設定：素直に考えれば，
  - 帰無仮説H0：$\mu \leqq 100$　（帰無仮説の設定に注意）
  - 対立仮説H1：$\mu > 100$
- これではいままでの手法では検定ができない．
  - 中心極限定理を使うには，帰無仮説の値を一つにする必要がある．
- 仮説の設定
  - 帰無仮説 H0： $\mu = 100$
  - 対立仮説 H1： $\mu > 100$

例題の続き

15

- 中心極限定理から，帰無仮説の下での標本平均値の分布は，平均100，標準偏差0.45の正規分布になる．
- 対立仮説と比較して帰無仮説を採択するには，棄却域を右側だけに設定するのが合理的である．
  - これを「片側検定」という．
- 以上の考察の下で，5%点のz値は1.645．
  - 境界は100+1.645x0.45=100.74
  - 試作品の平均寿命 101.5はこれより大きいので，帰無仮説は棄却される．

    「二つの製品の寿命の差は有意である」という．

仮説検定と区間推定の関係

16

- いま，データを64個とったところ平均が106であった．なお，標準偏差が16であることはわかっている．
- ここで，まず，
  - 帰無仮説 H0：μ =100，
  - 対立仮説 H1：$\mu \neq 100$

    として有意水準5%で両側検定することを考える．
- 標本平均の標準偏差は16 / sqrt(64) = 2．
  - したがって，z-scoreは (106 - 100) / 2 = 3 となる．

仮説検定と区間推定の関係

17

- 有意水準5%時の棄却域の限界点のz-scoreは1.96．
  - したがって，帰無仮説は棄却される．
- 一方，データ平均に基づいて母平均を区間推定することを考える．
- 95%の信頼区間は，

$$106 - 1.96 \times 2 < \mu < 106 + 1.96 \times 2$$

  - したがって，100は信頼区間外となる．
- 帰無仮説が棄却される ＝ 帰無仮説の値は信頼区間外

検出力（復習）

18

- 第1種過誤: 帰無仮説が正しいときに棄却してしまう
- 第2種過誤: 対立仮説が正しいときに棄却してしまう
- 両方を同時に小さくすることはできない．
  - そこで，通常は

    第1種過誤確率 α は，有意水準で統制する．

    第2種過誤確率 β を小さくする．

    → 検出力（$1-\beta$）を大きくする．
- しかし，検出力は一般に解析的に求められない．
  - 単純な例を以下で紹介する．

検出力を計算する例題

19

- **学位取得年齢の平均値が**30**歳以上であるかどうかを検定する．**
- **標準偏差**3**，データ数**100**，有意水準**5%**とする．**
  - **帰無仮説**H0：μ = 30
  - **対立仮説**H1：μ = 31
    （対立仮説下の標本分布を定めるため点にした）

検出力を計算する例題

20

- step1: **帰無仮説の棄却域を求める．**
  - **データ数**100**であるから，標本平均の標準偏差は**0.3**．**
  - **有意水準**5%**での片側検定の棄却域は**30.49**以上．**
- step2: **対立仮説の下での棄却域の部分の確率を求める．**
  - **この場合は，検出力**=0.9554**となる．**

検定の有意性と検定結果の重要性

21

- 検定の有意性と結果の重要性は別物である．
- 有意性：
  - 帰無仮説の下でそのデータが得られる確率が低いことを統計的に示している．
- 重要性：
  - 結果の重要性は利用する者が主観的に判断する．
  - 重要な結果が得られるように実験を設計した上で，統計的推測を下す．

分散未知の場合の推定と検定

22

- これまでの議論は，分散（標準偏差）既知という前提の下で進めてきた．
  - しかし，これは通常ありえないことである．
- 分散がわからければ，分散を推定しなければならない．
- 分散の不偏推定量は，前回の講義より

$$\hat{\sigma}^2 = \frac{1}{n-1} \Sigma \ (X - \overline{X})^2$$

- この不偏推定量を使えば，中心極限定理から，平均値の分布を定めることができそうに思える．しかし，

t 分布 (*t* - distribution)

23

- もとの確率変数が正規分布に従っていても，分散の不偏推定量を用いて標準化した平均値はもはや標準正規分布には従わない（中心極限定理が成立しない）
- では，平均値はどのような分布に従うのか？
- その分布のことを Studentの *t* 分布という．
  - *t* 分布の形はデータ数 *n* に依存して変化する．
    厳密には，データ数の関数である自由度 νというパラメータによって決まる．
  - 通常，自由度ν の *t* 分布を *t* (ν) と書く.

t 分布の自由度に対する依存性

24

- データ数 *n* のときの平均値は自由度 *n* - 1 のt分布．
- 自由度が 2より大きいとき，*t* 分布の分散は *n* / *n* -2
  -正規分布に比べて裾が重い．
- *n* が大きくなると，標準正規分布に近づく．
  - *n* が大きいほど分散は減少，*n* →∞で1に近づく．
- デモで実際にみてみよう．

自由度(degrees of freedom)とは何か．

25

- 自由度とはその統計量を定める独立なデータの数．
  - 当然ながら，データ数 n に依存している．
  - その統計量を求める際の制約条件数にも依存する．

平均値の分布の自由度

26

- n 個のデータから求めた平均値は，なぜ自由度n - 1の t 分布に従うのか？
- 自由度 ν の t 分布は自由度 ν のカイ2乗分布と関連．
- ν個のXiがそれぞれ独立に標準正規分布に従うとき，Xi^2の和は自由度νのカイ2乗分布に従う．
- Xが標準正規分布，Yが自由度νのカイ2乗分布に従うとき，$Z = X / \mathrm{sqrt}(Y / \nu)$ が従う分布が自由度ν の t 分布である．

偏差２乗和の自由度

27

- n 個のデータから計算した $S = \Sigma(X_i - \bar{X})^2$ の自由度はいくつか？
- 答えは n ではなく $n - 1$ である．
- 証明はやや tricky ．

step 1: 以下の Y_i (i は１から n-1まで）がすべて独立に平均 0, 分散 σ^2 の正規分布に従うことを示す．

$Y_1 = (X_1 - X_2) / \mathrm{sqrt}(2)$,

$Y_2 = (X_1 + X_2 - 2X_3)/\mathrm{sqrt}(6)$, …

偏差２乗和の自由度

28

step 2: $\bar{X}$ と Y_i が独立であることを示す．

step 3: $\Sigma_i Y_i^2 + n\bar{X}^2$ が $\Sigma_i X_i^2$ に等しいことを示す．

これと分散の展開公式から次式が導かれる．

$$S = \Sigma_i X_i^2 - n\bar{X}^2 = \Sigma_i Y_i^2$$

step 4: step 1での定義より，$\Sigma_i (Y_i/\sigma)^2$ は自由度 n-1のカイ2乗分布に従う (Y_i はn-1しかないことに注意）

step 5: step 3と4より，S/σ^2 が自由度 $n - 1$ のカイ2乗分布に従うことがわかる．

分散未知の場合の区間推定，検定

29

- 推定や検定の方法は，分散既知の場合と実質的に同じである．
  - 分散既知の場合は正規分布を使って作業をした．
  - 分散未知の場合は t 分布を使って作業をすればよい．

数値計算で行なうならば，関数を変える．
数表を使うならば，使う数表を変える．

例題：平均に関する仮説検定

30

- 小学校のある学年の生徒16人の身長を調べたところ，以下のデータが得られた．

  115, 117, 135, 128, 137, 127, 119, 130
  134, 126, 131, 123, 130, 123, 135, 122 (cm)

- 仮説「この学年の生徒の平均身長が130cmである」を有意水準1%で検定せよ．

例題の解法 1

31

- step1: **仮説の設定**
  - **帰無仮説**H0: $\mu = 130$
  - **対立仮説**H1: $\mu \neq 130$
- step 2: **各種数値の計算**
  - **データの平均：** $\overline{X} = 127$
  - **σの不偏推定量：** $\hat{\sigma} = 6.72$
  - **平均値の従う分布の標準偏差：**

$$\sigma_M = 6.72 / \mathrm{sqrt}(16) = 1.68$$

例題の解法 2

32

- step 3: **棄却域の決定**
  - **有意水準**1%**の両側検定のための棄却域を定める．**
  - **自由度**15**の** t **分布の数表をみると，**

$$t \leqq -2.947,\ 2.947 \leqq t$$

が棄却域であることがわかる．

- step 4: t **値の計算**

$$t = (\overline{X} - \mu) / \sigma_M = (127 - 130) / 1.68 = -1.79$$

- step 5: **結論**
  - **帰無仮説は棄却されない．**

平均値に関する検定方法の分類

33

- 正規分布に従っていて，かつ分散がわかっているとき
  - 正規分布を使って検定する．
- 正規分布に従っているが，分散がわからないとき，
  - データが少なければ(30以下)， t 検定を行なう．
  - データが多ければ(30以上)，正規分布を使って検定してもおよそ同じ結果が得られる．
- 正規分布に従っていない，あるいは，わからないとき
  - データが多ければ(30以上)，正規分布を使って検定しても実質的に問題がない．