データ解析基礎

第１回

本日の予定

- 講義の方針
- 講義内容
- データの表し方
- データの代表値
- 平均値，分散，パーセンタイルなど

統計処理はなぜとっつきにくいか？

3

- 難解な確率理論
  - 測度論などの難しい数学や積分の多い複雑な計算．
- 処理が使える前提条件や仮定
  - どういう時にどの方法を使えばよいかわからない．
- 具体的な作業の欠如
  - 講義を聴いているだけなので，実感が湧かない．
  - 自分の手で使ってみないので，身につかない．

講義の方針

4

- 使えるようになることに主眼をおく．
- 理論的な計算には極力踏み込まない．
- 内容を限定する（大学院レベルの内容はない）．
- 例題を多くする（できるだけ実例を出す）
- 毎回宿題を出す．

講義予定（その１）

5

- 準備的事項（変数，ヒストグラム，平均，分散）
- 相関，線形回帰分析（その１）
- 確率，分布，大数の法則，中心極限定理
- 確率分布（２項分布，ポアソン分布，正規分布）
- 仮説検定
- 母集団と標本

講義予定（その２）

6

- 平均値の推定と仮説検定
- 分散の仮説検定
- 回帰分析（その２）
- 分散分析
- 多変量解析，主成分分析

（ただし，進度に応じて柔軟に対応）

教科書と参考書

7

- 教科書：
  - 特定の教科書は指定しない．
- 参考書：
  - 蓑谷千凰彦：統計学入門１，２，東京図書．
  - 東大統計学教室編：統計学入門，東大出版会．

その他

資料の開示

8

- 講義のホームページの開設

  http://www.hi.is.uec.ac.jp/lecture/stat/

  - スライドの写しと演習問題の配布
    いずれも講義当日の午後にアップロードの予定．
  - その他，関連資料や連絡事項の掲示
- 学外からのアクセスにはパスワードが必要
  - user: data
  - passwd: basic

演習と成績

9

- 演習問題の配布と答案の提出
  - 配布はホームページでのみ行なう．
  - 講義の翌々週月曜日までにIS事務室横のポストへ．
  - 電子媒体での提出は受け付けない．
- 成績
  - 演習問題への提出状況に応じて判定．
  - 最終回のレポート提出が単位取得の必要条件．

統計的データ解析の対象と目的

10

- データに隠れている法則性を見つけだす．
- 例：
  - 自然現象
  - アンケート調査
  - 経済現象
  - ゲーム・プログラム
  - 学習システム

統計的推測

11

- 限られたデータに基づき，現象の法則性を推論する．
  (cf. 記述統計学)
- 「標本」に基づいて「母集団」の性質を推論する．
  - 「標本」とは計測や実験によって得られたデータ．
  - 「母集団」とは，背後にある現象そのもの．
    我々は母集団全体を調べることができないので，観測した標本に基づいて母集団の性質を推測するしかない．これが，統計的推測の考え方である．
- 前ページのほとんどの例が統計的推測の枠組みに入る．

変数（変量）

12

- 定数と変数
  - 定数：集合に属する要素にすべて共通するもの．
  - 変数：要素ごとに異なる値をとりうるもの．
    例：調布市の小学校5年生の男女の算数の成績
- 離散値と連続値
  - 変数が連続的な値をとるかどうか．
    数学的な意味ではなく，数値的な意味での分類
    例：5段階評価の成績，試験の点数

データの表し方（尺度）

13

- 尺度の種類
  - 名義尺度(nominal)：値を区別するだけの意味しかもたない．
  - 順序尺度(ordinal)：値に順序関係がある．
  - 間隔尺度(interval)：値の差に意味がある．
  - 比尺度(ratio)：値の比に意味がある．

これらは階層的な関係にある．

度数（頻度）とその分布

14

- 例題：学生アンケート
  - 質問１：出身学科
  - 質問２：自分自身の成績の自己評価
  - 質問３：年齢
  - 質問４：試験の点数

（データファイルを参照）

離散値の場合

15

- 階級(class)と度数(frequency)：
  - データをグループに分類したものが階級．
  - 各階級に属するデータの数が度数．
- 相対度数(relative frequency)：
  - 度数を全体のデータ数で割ったもの．
  - すべての階級について和をとると１になる．
- ヒストグラム：
  - 度数分布をグラフに表したもの．
  - 通常棒グラフが使われる．

連続値の場合

16

- 階級への分割
  - 連続的な量をいくつかの階級に排他的に分類する．
  - データの種類を減らすと同時に，情報を失わないように階級の数を保つ．
  - 多くの場合，階級の幅は等しく設定する．
- 度数など
  - 度数，ヒストグラムの定義は離散値の場合と同じ．
- 累積度数，累積相対度数：
  - 度数を下の階級から順につみあげたもの．

percentile rankとpercentile point

17

- percentile rank:
  - **その値は全データの中のどこに位置するか？**

    [**データが属する階級の一つ下の階級までの累積相対度数**]
    ＋[**属する階級の相対度数**] × (**データ** - **階級下端**) / [**階級幅**]

- percentile point:
  - **全体の◯%の位置にあるデータは何か？**

    [**当該階級の下端**] ＋ [**当該階級の階級幅**] ×
    (**順位** - **一つ下の階級までの累積度数**) / [**当該階級の度数**]

度数分布の形状

18

- **歪度**(skewness)：**分布の形が左右対称であるか？**
  - **歪度が負である：裾野が左側に伸びている．**
  - **歪度が正である：裾野が右側に伸びている．**
- **尖度**(kurtosis)：**分布の形がとがっているかどうか？**
  - **値が大きいと，尖り具合いがより甚だしい．**
- **単峰**(unimodal)，**双峰**(bimodal)
  - **分布の山が一つであるかどうか**

総和記号を使った計算の復習

19

- 次の値を計算せよ．

1) $\Sigma_i X_i$

2) $\Sigma_i Y_i$

3) $\Sigma_i X_i^2$

4) $(\Sigma_i X_i)^2$

5) $\Sigma_i X_i Y_i$

6) $(\Sigma_i X_i)(\Sigma_i Y_i)$

$X_1=4,\ X_2=2,\ X_3=1,\ X_4=1,\ X_5=3$

$Y_1=1,\ Y_2=2,\ Y_3=3,\ Y_4=4,\ Y_5=5$

分布を中央を表す値（代表値）

20

- 最頻値（mode）：度数が最も大きな値．
  - 連続量の評価では要注意．
- 中位数（median）：分布の中央の値．
  - データがすべてわかっているときは，大きなデータから順に数えればよい．

    データ数Nが奇数：(N+1) / 2番目のデータ値

    偶数：N/2番目と N/2+1 番目の中央値

  - 度数分布を基にするときは，50%パーセンタイル値を求める．

四分位数

21

- **４分位数**（quartile）

  分布を4**分する値**

  （25%, 50%, 75%**のパーセンタイル値）．**

平均値

22

- **算術平均** (arithmetic mean)**：いわゆる平均値．**
  - **データの総和をとり，データ数で割ったもの．**

$$\overline{X} = \frac{1}{N} \Sigma_i X_i$$

  - **度数分布で与えられた場合は，階級値** X_i**に度数** f_i **の重みをかけて計算する．**

$$\overline{X} = \frac{1}{N} \Sigma_i f_i X_i$$

$$N = \Sigma_i f_i$$

算術平均のもつ性質

23

- データと平均値の差をすべて足すと0になる．

$$0 = \Sigma_i (X_i - \overline{X})$$

- 全データが A だけ増加すれば，平均も A だけ増える．
- 全データが A 倍になれば，平均も A 倍になる．
- 算術平均と中位数(median)の関係
  - 分布が正の歪度をもつ場合は，中位数 ＜ 平均値
  - 分布が負の歪度をもつ場合は，中位数 ＞ 平均値

幾何平均 (geometric mean)

24

- すべてのデータの積をとり，そのN 乗根を求める．
- データの増加率を議論するときに有用である．
- 例：
  - ある年のデータは前年に比べて α 倍になる．
  - ５年間の増加を組み合わせると

$$\alpha = \alpha_1 \alpha_2 \alpha_3 \alpha_4 \alpha_5$$

平均増加率を求めるには，これの5乗根を使うことになる．

分布のばらつきを表す値

25

- **二つのグループでの成績の分布**

  group A: 1, 5, 5, 5, 5, 5, 5, 5, 5, 9

  group B: 1, 1, 2, 4, 5, 6, 7, 7, 8, 9

  どちらも平均点は5**点．**
  しかし，A**グループでの**9**点と**B**グループでの**9**点では，**
  「成績の顕著さ」が異なる．

  この違いを定量的に評価するにはどうすればよいか？

  →　分布のばらつきを尺度にする必要がある．

分散(variance)

26

- **平均値からのずれを評価する量**
  1. **平均値との差**
  2. **平均値との差の絶対値**
  3. **平均値との差の**2**乗**

  数学的取扱いの楽な 3. **の考え方がよく使われる．**

- **分散の定義** (variance)**：ずれの大きさの平均値**

$$V = S_X^2 = \frac{1}{N} \Sigma_i (X_i - \bar{X})^2$$

$$= \frac{1}{N} \Sigma_i f_i (X_i - \bar{X})^2$$ **：度数分布で与えた場合**

（X_i **は階級値，**f_i**は度数）**

標準偏差 (standard deviation)

27

- 分散はデータの2乗の次元をもつ．
  - データと同じ次元にするには，平方根をとる必要がある．
- 標準偏差 (standard deviation)

$$S_X = \sqrt{S_X^2}$$

- 例題：

  最初の例題で，二つのグループの分散と標準偏差を求めてみる．

分散と標準偏差がもつ性質

28

- すべてのデータにAだけ足しても，

  分散，標準偏差ともに値は変わらない．
- すべてのデータがA倍になると，

  分散は A^2倍になり，標準偏差は A倍になる．
- 計算上の便法

  （分散）＝（データの2乗の平均値）－（平均値の2乗）

$$= \frac{1}{N} \Sigma_i X_i^2 - (\frac{1}{N} \Sigma_i X_i)^2$$

z-score

29

- 標準偏差で正規化したデータの表し方
  - 標準偏差はデータのばらつき具合いを表している．
  - データと平均の差を標準偏差で割ると，そのデータが平均からどの程度離れているかを評価できる．

$$z = (X_i - \bar{X}) / S_X$$

- z-scoreの平均と分散はそれぞれ0，1になる．
  平均と標準偏差のもつ性質からすぐに導ける．
- 例題：二つのグループの9点の学生のz-scoreを求める．

その他の標準値

30

- 偏差値　　平均 50, 標準偏差 10
- 米国留学試験の点数　平均 500, 標準偏差 100
- 知能指数(IQ)　　平均 100, 標準偏差 15

z-scoreとの関係

平均点をa，標準偏差をbとすると

$$p = a + bz$$

z-scoreについて注意すべきこと

31

- 分布形が異なる集団間でz-scoreを比べてはいけない
- 例：

group A: 2, 3, 3, 3, 4, 4 ,4 ,8, 9, 10
group B: 0, 1, 2, 6, 6, 6, 7, 7, 7, 8

- どちらも平均は5, 標準偏差は2.72.
  → 点数8の人のz-scoreはいずれでも1.10
- percentile rank は，A群では75, B群では95.
  → B群に属する人の方が成績がよいと判断すべき.