

# 筑波大学における教材の内容

今倉暁(筑波大学)

# もくじ

- データサイエンス授業概要
  - 講義全体の概要
  - 学習の範囲
  
- 教材内容について
  - 講義
  - 演習
  - ビデオ講義

# データサイエンス授業概要

# 全体の概要

- 目的
  - データサイエンスの基礎を、座学およびエクセルを用いた演習を通して、しっかり学ぶこと
- そのため、
  - 実社会への応用については深くは取り扱わない
    - 各分野に対応した応用については、ビデオ講義を用意
  - 演習であつかうデータは、エクセルで扱える範囲に限る
    - データには人工データ、オープンデータの他、自分たちのデータも利用する
    - プログラミングに興味がある人は、各学類の講義を(開講されていれば)受講を勧める

# 全体の概要

- 授業形態(150分/週 x 10週)
  - 講義、演習、ビデオ講義
  - 小テスト、レポート
- 全学必修であるための工夫
  - 講義資料および演習内容については、下記の3種類を用意
    - 簡易版: 数理に馴染みのない文系クラスを想定
    - 標準版: 数理に馴染みのある文系クラスを想定
    - 発展版: 理系クラスを想定
  - 多様な学生の興味と動機の向上を目的とした各分野に対応したビデオ講義を用意(詳細は後述)
  - 自分たちのデータを自分たちで解析する実践的な演習課題(詳細は後述)



# 教材内容について

# 講義内容・キーワード

週	講義内容	キーワード
1	データサイエンス概論	データサイエンスの概要
		データサイエンス導入ビデオ講義
2	データの種類と収集1	データの種類、研究倫理、データの収集、エクセル基本操作
3	データの種類と収集2	データサイエンスの基本プロセス、データの前処理
		データの再利用性、エクセル応用操作(統計)
4	データの種類と収集3	データ管理の意義と目的、データ収集項目の設定
		情報構造と表現の分離、エクセル応用操作(データ操作)
5	高度なデータ管理 達成度評価(中間)	ビッグデータビデオ講義
6	データの可視化と質的データの解析	可視化の意義と目的、視覚的表現の選び方、データ可視化演習
		1つの質的データ(度数分布表)、2つの質的データ(クロス集計表)
7	1つの量的データの解析	1つの量的データの理解、ヒストグラム、記述統計、平均値の違い
		統計検定
8	2つの量的データや時系列データの解析	2つの量的データの理解, 相関と相関係数, 時系列データ, (回帰直線)
9	実際のデータの分析	様々なデータの分析、ネットワークデータ、相関と因果
10	高度なデータ分析 達成度評価(期末)	人工知能ビデオ講義

# データの種類と収集

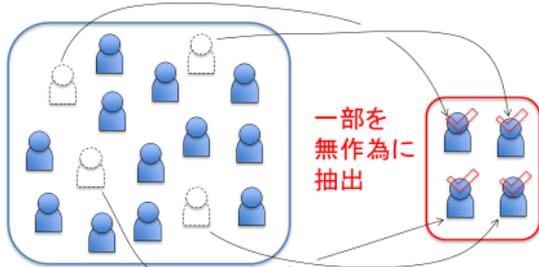
## 内容

- データの種類と尺度水準、収集方法と調査方法、研究倫理、エクセルの基本操作
- データサイエンスの基本プロセス、特にデータの収集法と前処理法

適切な前処理は？

### データの収集：調査

- 標本調査  
母集団に含まれる一部の標本を対象とした調査



母集団  
(例) 全ての筑波大生

標本  
(例) 学籍番号で  
無作為抽出した  
学生を調査 ✓

データ収集・調査方法

### データ欠損・重複・不整合

- データ欠損
  - あるべきデータがないこと
  - 例: センサーの故障による値の欠損、記入忘れ
- データ重複
  - 同一対象のデータが重複すること
  - 例: 同一人物のデータが重複して登録されている
- データ不整合
  - 型の不整合
    - 例: 電話番号を表す文字列データと数字データが混在
  - 表記ゆれ
    - 例: 年が和暦と西暦混ぜこぜで表現されている
    - 例: 全角・半角が混ざっている
  - データ依存性違反
    - 例: 郵便番号と都道府県といった依存関係のあるデータに一貫性がない
      - 住所は茨城県つくば市なのに郵便番号が004-0001 (北海道札幌市)と明らかに異なる

年月日	最高気温
2020/8/1	32
2020/8/2	32
2020/8/3	30
2020/8/3	30
令和2年8月4日	34

欠損  
重複  
不整合(表記ゆれ)

26

データ前処理

# データの種類と収集

- 内容

- データの収集と管理

- 文書構造(データ構造)と視覚表現(ビュー)を分離することの意義
    - 質問票の設計

データ処理しやすく  
整理されたデータとは

見やすいがデータ処理にくい

データ処理しやすい

セルが結合されている

学群	学類	入学定員
人文・文化学群	人文学類	120
	比較文化学類	80
	日本語・日本文化学類	40
	計	240
社会・国際学群	社会学類	80
	国際総合学類	80
	計	160
人間学群	教育学類	35

学類の列に小計が並んでいる

16

## データの管理

質問票設計：調査の概要

調査を計画するために調査の概要を作成する

- 目的：なぜ調査を行うのか
- 対象：誰を対象とするのか(母集団の設定)
- 時期：調査の実施期間
- 方法：調査対象の選出の仕方(全数調査 or 標本調査)や、調査票の配布から回収までの流れなど

配布・回収方法

1. 調査員調査(面接調査・留置(とめおき)調査)
2. 郵送調査
3. 電話調査
4. インターネット調査

✓全数調査か、標本調査か、標本調査の場合は、母集団は何か、無作為抽出になっているか、抽出により偏りが生じていないか等を検討。

30

## 質問表の設計

# 講義内容・キーワード

週	講義内容	キーワード
1	データサイエンス概論	データサイエンスの概要 データサイエンス導入ビデオ講義
2	データの種類と収集1	データの種類、研究倫理、データの収集、エクセル基本操作
3	データの種類と収集2	データサイエンスの基本プロセス、データの前処理 データの再利用性、エクセル応用操作(統計)
4	データの種類と収集3	データ管理の意義と目的、データ収集項目の設定 情報構造と表現の分離、エクセル応用操作(データ操作)
5	高度なデータ管理 達成度評価(中間)	ビッグデータビデオ講義
6	データの可視化と質的データの解析	可視化の意義と目的、視覚的表現の選び方、データ可視化演習 1つの質的データ(度数分布表)、2つの質的データ(クロス集計表)
7	1つの量的データの解析	1つの量的データの理解、ヒストグラム、記述統計、平均値の違い 統計検定
8	2つの量的データや時系列データの解析	2つの量的データの理解, 相関と相関係数, 時系列データ, (回帰直線)
9	実際のデータの分析	様々なデータの分析、ネットワークデータ、相関と因果
10	高度なデータ分析 達成度評価(期末)	人工知能ビデオ講義

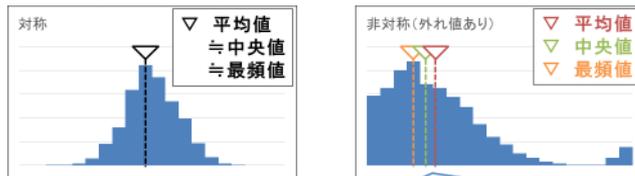
# データの分析

## 内容

- 1つの量的データの代表値やヒストグラムの読み方、異なる母集団の平均値の違い (Z検定)
- 2つの量的データの可視化や関係性の解析方法、時系列データの可視化や解析方法

### データの分布と記述統計量

- データの分布と記述統計量の大小関係
  - 分布が左右対称
    - 「平均値 ≡ 中央値 ≡ 最頻値」となる
  - 分布が非対称、外れ値あり
    - 「平均値 > 中央値 > 最頻値」(もしくは「平均値 < 中央値 < 最頻値」)となることが多い

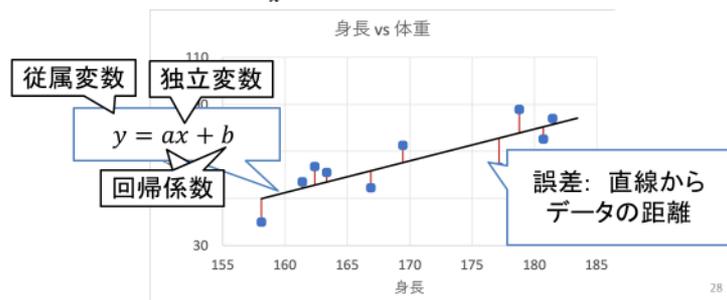


記述統計量から分布を予想できる場合もあるが、実際には複数のピークがある場合など、記述統計量だけでは分布を十分に説明できないデータもあるので、ヒストグラムを描いて確認することが重要です。

データの分布

### 回帰直線

- データとの誤差が一番小さい直線  
傾き  $a = \sigma_{xy} / \sigma_x^2$ , 切片:  $b = \bar{y} - a\bar{x}$   
 $\sigma_x^2$  は  $x$  の分散



2つのデータの関係性

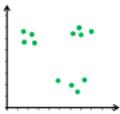
# データの分析

- 内容

- 可視化: データに適したグラフ(チャート)の選び方や描き方、不適切なグラフとその理由
- 実際のデータ解析の流れの学習
- Excel では扱えないような複雑なデータ解析の事例についても紹介

### 視覚属性

値を表現できる視覚的な属性

位置 

色相 

模様(きめ) 

大きさ(長さ、太さ) 

方向(向き) 

明度 

形 

17

可視化で使用する視覚属性



e-Stat 統計で見る日本  
政府統計の総合窓口

377年のデータ

データセット: アイスクリーム

統計分類 (大分類)	提供統計名・提供分類	調査年月	公開 (更新) 日	表示・ダウンロード
前編で絞り込み	牛乳乳製品統計調査	2016年	2020-03-05	DB API
政府統計名で絞り込み	牛乳乳製品統計調査 (全国・北海道・都府県) (月別)			
生産・出荷・中間調査	牛乳乳製品統計調査 / 産報 / 平成27年中	2015年	2020-03-05	DB API
家計調査	牛乳乳製品統計調査 (全国・北海道・都府県) (月別)			
小売物価統計調査	牛乳乳製品統計調査			

13

e-Statからのデータ収集

# 演習

## 概要

- エクセルを使用して、各講義内容の演習に取り組む
  - エクセルの基本操作、データの前処理
  - 可視化、データの入手、データの分析など

### E03-01: 総合演習A

- シート「E03-01」は各都道府県の人口および面積のデータである

	A	B	C	D
1	都道府県	人口 (単位: 人)	面積 (単位: km <sup>2</sup> )	人口密度 (単位: A/km <sup>2</sup> )
2	東京	13,843,403,000	2133.96	
3	神奈川	9,179,835	2416.16	
4	大阪	8,824,566	1905.29	
5	愛知	7,539,185	5172.96	
6	埼玉		3797.75	
7	青森	1,308,265	9645.64	
8	神奈川	9,179,835	2416.16	
9	広島	2,219,962	8479.61	
0	茨城	2,182,943	6097.23	
1	千葉	6,268,585	5157.61	
2	京都	2,591,779	4612.20	
3	熊本	1,756,442	7,409.50km <sup>2</sup>	

- 各都道府県の人口密度を集計したい
- 次の手順でデータを処理しなさい  
欠損値や重複を含む行を削除しなさい
  1. D列で都道府県ごとの「人口密度」(=「人口」/「面積」)を計算しなさい
  2. 2.の結果、正しく計算できない都道府県が存在する、正しく計算できるように表記揺れを修正しなさい
  3. 日本の全人口は約1.2億人ということをふまえると、記入ミスと思われる大きく外れた値が一つある、該当する値が含まれる行を削除しなさい
  4. D列の書式を変更し小数点以下の桁数表示を2桁に揃えなさい

29

### 演習課題 E08-03-01

- 課題1
  - 茨城県のデータについて2001年の人口を基準とした各年の人口の伸び率(指数)を求めよ
- 課題2
  - 求めた伸び率を折れ線グラフにプロットせよ
- 課題3
  - 描画した移動平均の折れ線グラフについて考察せよ(穴埋め問題)

64

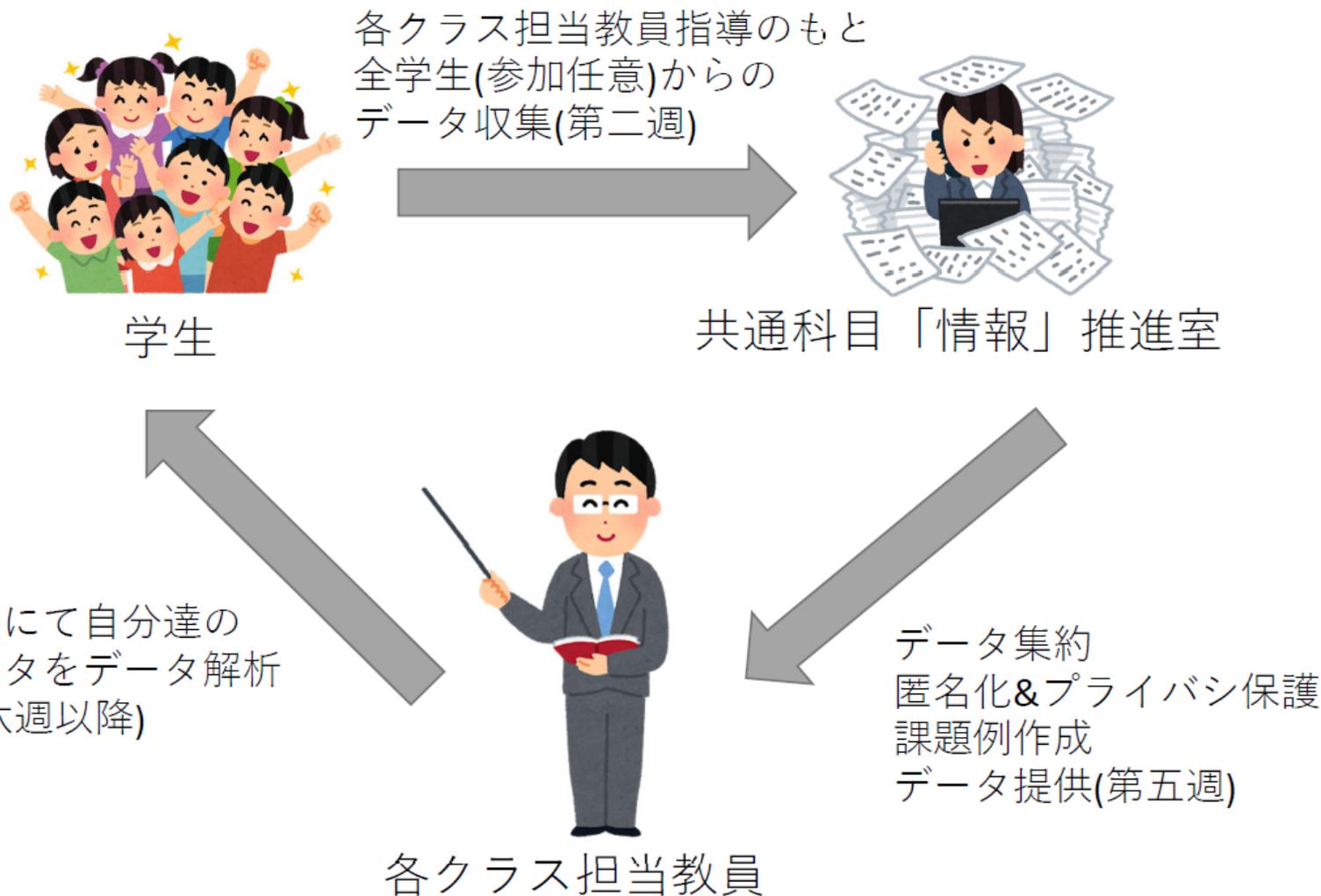
データ前処理

データの分析

# 演習

- 使用するデータ
  - 人工データ
  - オープンデータ
    - タイタニック、カリフォルニアハウジング
    - 気象庁、e-Stat
  - 自分たちで集めたデータ
    - アンケートデータ

# 自分たちのデータを自分で解析



# 質問項目例

2. 2015年に国際連合にて193ヶ国の首脳により署名された持続可能な開発目標(SDG)、またはグローバル目標はご存知ですか。↓  
(ア)はい↓  
(イ)いいえ ↓
3. SDGの17個のグローバル目標のうち、どのグローバル目標が、あなたとあなたの家族にとって目前に迫っている課題ですか？(6つ選択してください。) ↓



- (ア) 貧困をなくそう↓  
・あらゆる場所で、あらゆる形態の貧困に終止符を打つ ↓



- (イ) 飢餓をゼロに↓  
・飢餓に終止符を打ち、食料の安定確保と栄養状態の改善を達成するとともに、持続可能な農業を推進する ↓



- (ウ) すべての人に健康と福祉を↓  
・あらゆる年齢のすべての人の健康的な生活を確保し、福祉を推進する ↓



- (エ) 質の高い教育をみんなに↓  
・すべての人に包摂的かつ公平で質の高い教育を提供し、生涯学習の機会を促進する ↓



- (オ) ジェンダー平等を実現しよう↓  
・ジェンダーの平等を達成し、すべての女性と女児のエンパワーメントを図る ↓

(国際連合広報センター MY WORLD2030)

<将来の仕事に関する質問> ↓

## 7. 職業選択の重視点 ↓

仕事を選ぶ際に、どのようなことを重視しますか。この中からいくつかでも選んでください。 ↓

- (ア) 収入 ↓
- (イ) 労働時間 ↓
- (ウ) 通勤の便 ↓
- (エ) 仕事内容 ↓
- (オ) 職場の雰囲気 ↓
- (カ) 仕事の社会的意義 ↓
- (キ) 事業や雇用の安定性 ↓
- (ク) 将来性 ↓
- (ケ) 専門的な知識や技能が活かせること ↓
- (コ) 能力を高める機会があること ↓
- (サ) 自分を生かすこと ↓
- (シ) 自分の好きなことや趣味を生かせること ↓
- (ス) その他 ↓
- (セ) わからない ↓

(内閣府 我が国と諸外国の若者の意識に関する調査(平成30年度)から) ↓

問題意識(SGD)、キャリア意識、  
およびITリテラシーに関する質問項目から構成

# 講義内容・キーワード

週	講義内容	キーワード
1	データサイエンス概論	データサイエンスの概要 データサイエンス導入ビデオ講義
2	データの種類と収集1	データの種類、研究倫理、データの収集、エクセル基本操作
3	データの種類と収集2	データサイエンスの基本プロセス、データの前処理 データの再利用性、エクセル応用操作(統計)
4	データの種類と収集3	データ管理の意義と目的、データ収集項目の設定 情報構造と表現の分離、エクセル応用操作(データ操作)
5	高度なデータ管理 達成度評価(中間)	ビッグデータビデオ講義
6	データの可視化と質的データの解析	可視化の意義と目的、視覚的表現の選び方、データ可視化演習 1つの質的データ(度数分布表)、2つの質的データ(クロス集計表)
7	1つの量的データの解析	1つの量的データの理解、ヒストグラム、記述統計、平均値の違い 統計検定
8	2つの量的データや時系列データの解析	2つの量的データの理解, 相関と相関係数, 時系列データ, (回帰直線)
9	実際のデータの分析	様々なデータの分析、ネットワークデータ、相関と因果
10	高度なデータ分析 達成度評価(期末)	人工知能ビデオ講義

# ビデオ講義



筑波大学オープンコースウェア  
UNIVERSITY OF TSUKUBA OPENCOURSEWARE

- 授業内容の客観評価を高めるには学生の興味と動機向上が重要
- 様々なバックグラウンドを持つ学生の特性を考慮した23本のビデオ講義
- オープンコースウェアとして誰でも視聴可能(<https://ocw.tsukuba.ac.jp/>)

## 筑波大学 10学群(学生の背景)

- 人文・文化学群
- 社会・国際学群
- 人間学群
- 生命環境学群
- 理工学群
- 情報学群
- 医学群
- 体育専門学群
- 芸術専門学群
- 総合学域群

## ビデオ講義

各分野におけるデータの収集、  
管理および活用について学ぶために使用

### データサイエンスの導入

- ヒューマンインタラクション
- 臨床医学・社会医学とデータサイエンス
- 現代サッカーボールの空力特性
- 生命科学とデータサイエンス
- デジタル・ヒューマニティーズ
- 人工知能における倫理的・法的・社会的問題
- データ駆動型社会における津波高即時予想
- サッカーの上達にデータを生かす
- データサイエンスと社会科学
- センサム137全球データベースおよび環境放射能データの検索と公開サイト

### 高度なデータの管理と活用

- ビックデータとIoT/GPS
- 人工知能と機械学習

### データサイエンスの発展

- 仮説検定入門

全学群の  
学生が視聴

計測センサーと大規模シミュレーションを  
組み合わせた、津波予測

地震計、水圧計センサー、GPSセンサー

スマートフォン等で位置

予測表示システム

https://toshitaka-baba.wixsite.com/index/tsunami-prediction-on-hpc

メリット：正確に予測することができる

デメリット：膨大な計算コスト / 予測までにある程度の時間が必要

11/26

ビデオ講義一例(データ駆動型社会における  
津波高即時予測より)

ビデオ教材で**専門に近い動画**が用意されている  
学類の学生は、ビデオ講義で**学習動機が向上**  
(授業アンケート結果より)

# まとめ

- データサイエンス講義の目的
  - データサイエンスの基礎を、**座学**および**エクセル**を用いた演習を通して、しっかり学ぶこと
- 教材内容
  - データサイエンスの基礎
    - **データ収集、前処理、可視化、分析**
  - 学習意欲の向上への取り組み
    - **各種分野でのデータ解析の応用に関するビデオ講義**
    - **自分たちのデータを演習で利用**