統計学第1講

明治大学情報ｺﾐｭﾆｹｰｼｮﾝ学部
後藤　晶
akiragoto@meiji.ac.jp

今日のお話

第1講：イントロダクション

教員の自己紹介

講義の概要

授業で使うアプリケーションの紹介

統計学やデータサイエンスを学ぶ意義を考える

基本的な操作方法

Rを使って計算してみる

第1講：イントロダクション

教員の自己紹介

軽く自己紹介を．．．

- 名前：後藤 晶（ごとう あきら）
- 1984年7月13日生まれ
  - 37歳，独身＞＜；
- 出身地：横浜
  - いわゆる「ハマっ子」です．
  - 高校は横浜駅から歩いて通ってました．

軽く自己紹介を．．．

- 趣味：読書，音楽聞いたり．一人旅
  - イロイロ読みます．
  - 割と激し目の音楽が好き
- 専門分野：行動経済学・社会情報学・実験/計算社会科学
  - 自発的貢献行動（協力行動）の促進／抑制要因：ゲーム理論に関する実験を中心として
  - 行動経済学の観点からの「政策」評価：行動経済学を政策に活かせないか？
  - クラウドソーシングを用いた経済ゲーム実験環境の構築

講義の概要

授業概要

本講義では仮説をデータに基づいて統計的に検証したり、データ解析の結果から、新たな事実を発見したりするときに役立つ統計的手法を身に付ける。

講義はデータの整理の仕方、平均、分散等を求める記述統計学より始め、確率、母集団、標本抽出、確率分布を学び、最終的に推定、検定といった推測統計学を解説、演習する。本講義は、理学療法学研究法、作業療法研究法、卒業研究の基礎となる科目である。

ざっくりいうと

- データとの付き合い方を学ぶために，Rを使って分析手法を学ぶ.
  - Rを使いやすくするためにRStudioを使用する.
- 基本的なデータの扱い方を整理して，グラフの可視化などを学ぶ.
- 分析手法として，**回帰分析**，**t検定**，**分散分析**，**重回帰分析**を学ぶ.
- 最後はグループプレゼンテーションを評価する.
  - もしくはレポートなど，個人での演習となる場合もある.
  - 様子を見ながら，プレゼンテーション資料をもとにしたレポート資料についても評価する.

到達目標

1. 各統計手法について，その目的と意義を説明することができる．
2. 各統計手法について，各自で分析を実行できる．
3. 分析結果について，適切に他者に説明できる．

授業内容

- 【第1講】イントロダクション
- 【第2講】基本的な操作法と記述統計量の算出①
- 【第3講】基本的な操作法と記述統計量の算出②
- 【第4講】基本的な操作法と記述統計量の算出③
- 【第5講】データの可視化・実証分析の手続き
- 【第6講】確率分布とシミュレーション
- 【第7講】カテゴリデータの分析①
- 【第8講】カテゴリデータの分析②

授業内容

- 【第9講】回帰分析・差の検定①
- 【第10講】回帰分析・差の検定②
- 【第11講】１要因分散分析・２要因分散分析①
- 【第12講】１要因分散分析・２要因分散分析②
- 【第13講】２要因分散分析・モデル選択①
- 【第14講】２要因分散分析・モデル選択②
- 【第15講】問題演習（テスト代わり）

※ただし，履修者の状況により内容を一部変更することがある.

評価

- 平常点：30%
  - 毎回の授業での取組状況を評価します.
- 講義内の課題：50%
  - 毎回の講義の中で演習問題を出します．こちらの課題を提出してください.
- 演習問題：20%
  - 最後に演習問題を課します.

課題に対するフィードバックの方法

- フィードバックとして，課題に対して全体的なコメントを返します．
  - 具体的には，前の講義での演習問題について，次回の講義でコメントをします．

教科書

- 必要な資料をオンラインで配布します.
  - 配布URLは以下のとおりです.
  - https://akrgt.github.io/2021statistics/

参考図書

- Navarro, D.J, & D.R.Foxcroft（著）芝田征司（訳）,
  『jamoviで学ぶ心理統計』

統計について

- 小杉考司, 2019, 『言葉と数式で理解する多変量解析入門』, 北大路出版
  - 非常に平易な言葉で多変量解析について説明がなされている．ただし，一歩戻って基本的RStudioの使い方と基本的な統計(t検定や回帰分析，カイ2乗検定など)を学ぶのであれば，小杉考司, 2019, 『Rでらくらく心理統計 RStudio徹底活用』などもよいでしょう．
- 星野匡郎, 田中久稔, 2018, 『Rによる実証分析 ―回帰分析から因果分析へ―』, 裳華房
  - Rを用いた様々な分析手法について記載されている．春学期の内容も含まれるが，主に秋学期の内容をカバーしている．

- 川端一光, 岩間徳兼, 鈴木雅之, 2018, 『Rによる多変量解析入門 - データ解析の実践と理論』, オーム社
  - 基本的な技術を学ぶことができる．かなり良書．
- 森田果, 2014, 『実証分析入門 - データから「因果関係」を読み解く作法』, 日本評論社
  - 法律系の先生が書かれた本だが，なかなかわかりやすく面白い．どちらかというと秋学期の内容に関わる．

- 地道正行, 2018,『データサイエンスの基礎 Rによる統計学独習』, 裳華房
  - 細かく書かれているが，数式も多い．講義内で数学的説明の時間を十分に確保できないため，興味のある方はこちらへ．

再現性の議論について

- 高橋康介, 2018, 『再現可能性のすゝめ (Wonderful R 3)』, 共立出版
- 江口哲史(編), 2018『自然科学研究のためのR入門―再現可能なレポート執筆実践― (Wonderful R 4)』, 共立出版
  - いずれの本も再現可能性のためにRStudioとRMarkdownの使い方について紹介した本.

分析の一連の流れについて

- 松村優哉, 湯谷啓明, 紀ノ定保礼, 前田和寛, 2018, 『RユーザのためのRStudio[実践]入門–tidyverseによるモダンな分析フローの世界』, 技術評論社
  - RStudioを用いたデータ収集，データ整形，可視化，レポーティングといった一連の分析の流れに関する本.

参考web資料

- からだにいいもの
  - Rに関する様々な情報が掲載されている．多少応用的なトピックが多い．
- marketechlabo
  - ちょっと新しいパッケージ等が紹介されていて興味深い

アクティブ・ラーニング

- グループワークおよびプレゼンテーションを行う.

留意事項

- R，RStudioという統計ソフトを使って，実践的に分析を進めながら統計学を学んでいきます．
- 各自が参考図書等に目を通しながら，積極的に学ばれることを期待します．可能であれば，自宅のPCにもインストールをしてください．
- インストール方法は，簡単に動画でまとめてあるので自宅に帰ってからご確認ください．

授業で使うアプリケーションの紹介

授業で使うアプリケーション

- R＆RStudio：基本的な分析手法から高度な分析手法まで，必要な分析は何でもできる．
  - しかもフリーと来たもんだ！

この授業でフリーのアプリケーションを使う理由：

- 様々な有料のアプリケーションは存在する.
  - SPSS，Stata，SAS，Eviews，などなど．．．
  - 企業に就職してもこれらのアプリケーションが使えるとは限らない.
  - 家で勉強しようにも有料で困る＞＜；
- フリーソフトであれば，いつでもどこでも必要に応じて自分で分析ができる.
  - 卒論に限らず，就職してからも使えるスキルとして身につけられる.
  - そういった観点から，この授業ではRを中心としたフリーソフトを使います.

統計学やデータサイエンスを学ぶ意義を考える

統計学とは？

- データサイエンス：データから有用な情報・知識を引き出したり，新たな価値あるデータを創造するための基本な考え方
  - 平均・分散を算出したり，全体的な傾向を把握するためにヒストグラムを作るのは有用な情報・知識を引き出すため.
  - プログラムを組んで，新たに価値あるものを作る.
- 統計学を学ぶと何がどうなる？
  - 「データ」に基づいた思考方法が身につく
  - 「見せかけの類似性」に騙されなくなる.
  - 日常生活・ビジネスへの応用可能性が広がる

「血液型占い」：血液型によって性格が異なる.

- 「データ」に基づくと，血液型によって性格が異なるとはいえない（参考）.

縄田 健悟 (2014).
血液型と性格の無関連性――日本と米国の大規模社会調査を用いた実証的論拠――
心理学研究, 85, 148-156. [**PDF**]

【タイトル】

血液型と性格の無関連性――日本と米国の大規模社会調査を用いた実証的論拠――

【要約】

日本の社会では，ABO式血液型と性格に関連があるという俗説が広く信じられている一方で，心理学の実証研究ではその関連性は認められていない。本研究は，血液型と性格が無関連であるより積極的な実証的根拠を提示することを目的としている。そのために，大規模調査の二次分析を行った。大規模な無作為標本抽出で日本とアメリカから合計10000以上 の標本を分析した。日本のデータセットは2004年 (N = 2878–2938)と2005年 (N = 3618–3692) であり，アメリカのデータセットは2004年 (N = 3037–3092)である。分析の結果，3つのデータセット全ての68項目中65項目で血液型間に有意な違いは確認されなかった。また，効果量η^2は.003以下であり，血液型の効果は全分散の0.3%以下しか説明しなかった。以上の結果は，血液型と性格の無関連であることを示している。

【もっと簡単な説明】

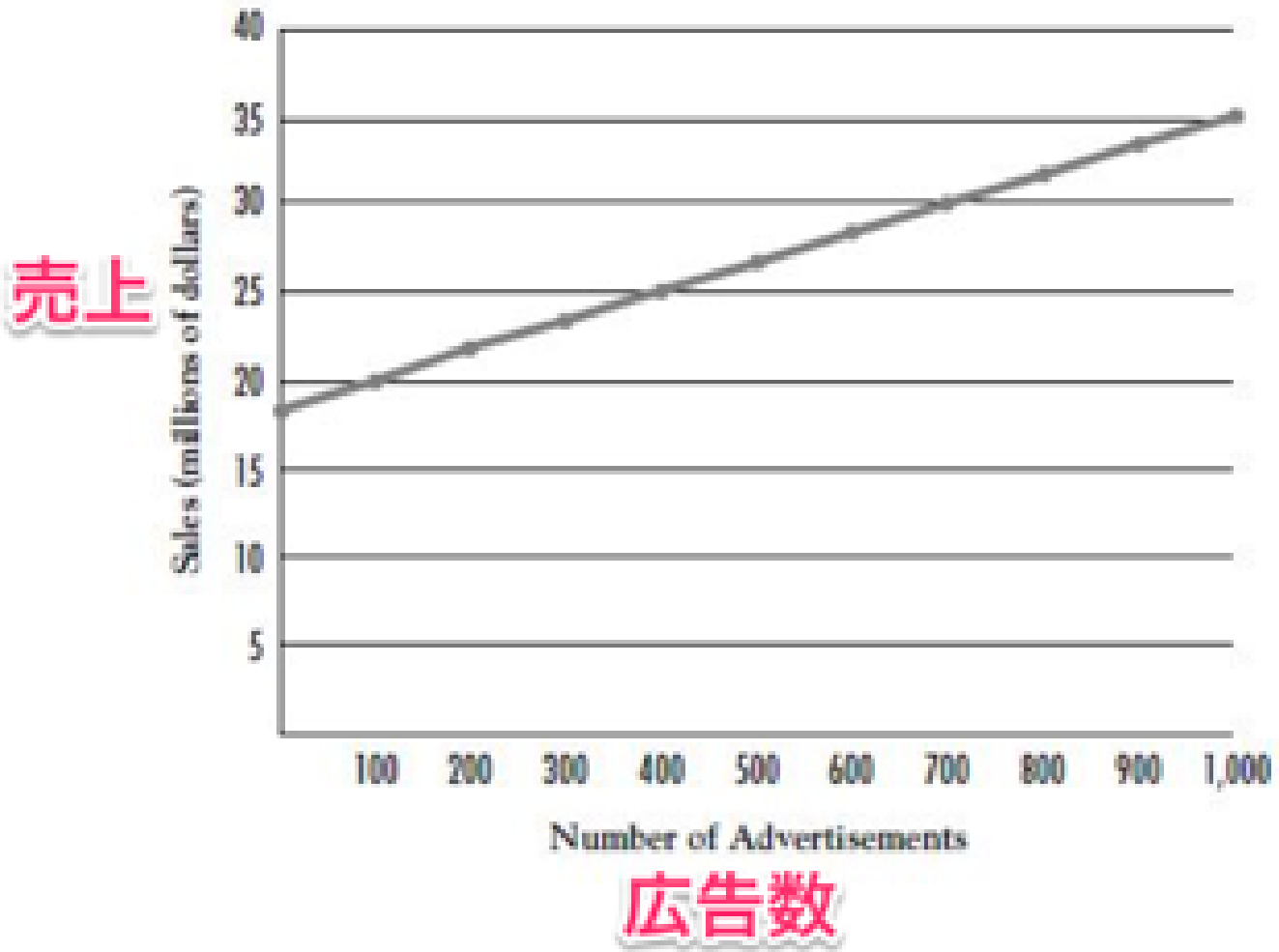
・これまでの心理学研究でも血液型と性格に関連性が見られないことは研究されてきた。
・本研究は血液型と性格の無関連性をより積極的に示すため，
　日本とアメリカの合計１００００人以上の規模のアンケートデータを元に，検討した。
・わずかな差でも検出できる大規模データにもかかわらず，一般的な生活に対する態度に関する**合計６８項目中６５項目**で血液型間に統計的に意味のある差が見られなかった。残り３項目も偶然の範囲。
・質問項目の個人差のうち血液型が説明した割合は，一番大きなもので**たった0.3%**。ほぼゼロだと見なせる。

見せかけの類似性：相関関係と因果関係.

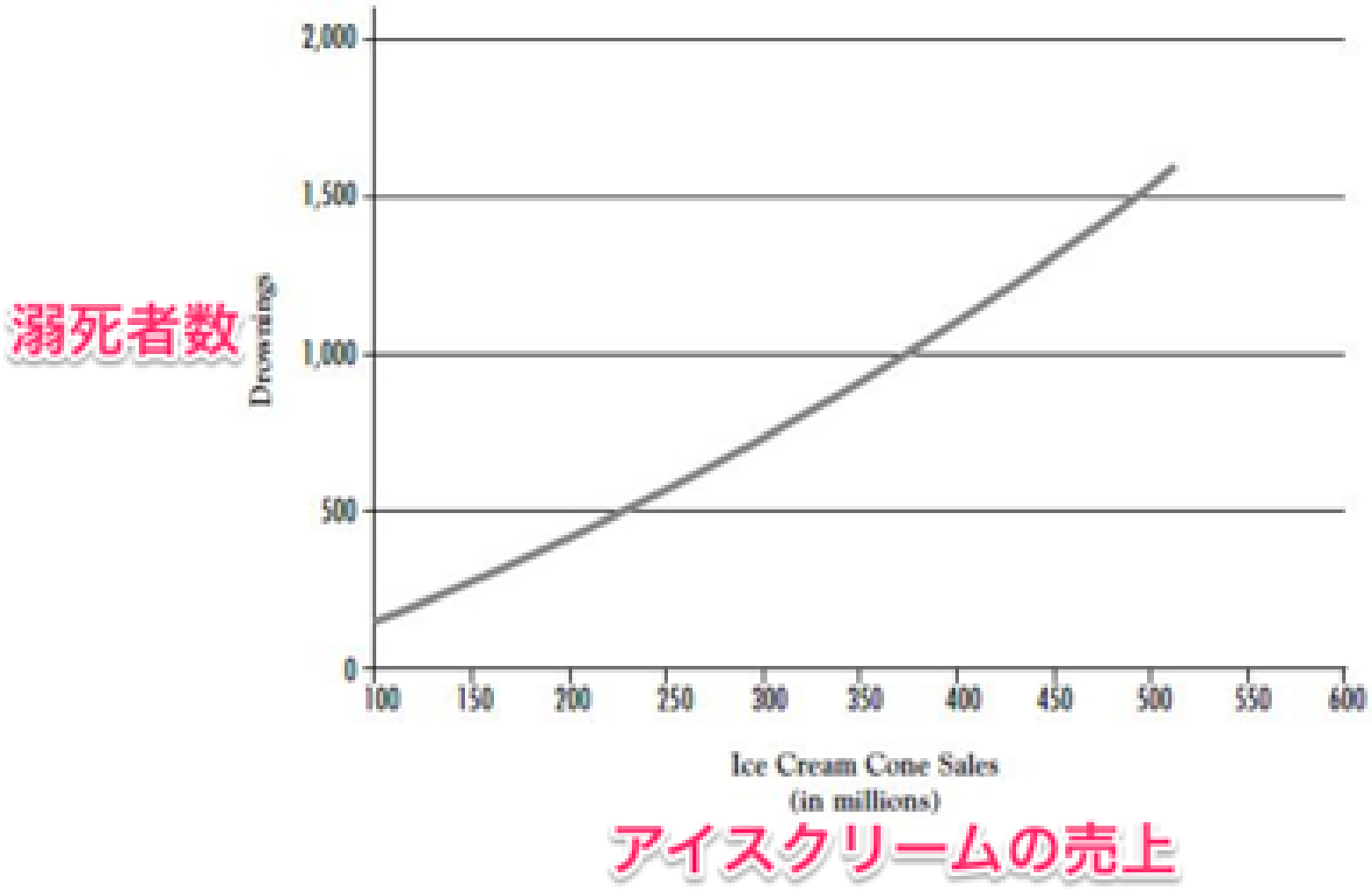
- 「データ」はモノを考えるのに大事なことだが，「数字」や「見た目」に騙されてはいけない.
- その他の要因が影響している可能性がある（参考）.

広告数と売上の関係

Relationship Between Advertisements and Sales



アイスクリームの売上と溺死者の関係



日常生活・ビジネスへの応用可能性が広がる

- ビジネスにおいて，データを利用し，企業活動を改善・開拓するのに必要な3要素
- データを効率的に収集・処理する
  - 企業活動におけるデータは莫大であり，このようなデータを処理するためには情報技術の利用が必要
- データを適切に取り扱い，妥当かつ汎用的な成果を残す
  - データ全体から，その傾向の妥当性を検討する．
- データをビジネスの枠組みの中にうまく組み込む
  - 効率的に処理をし，有用な結論を導き出してビジネスへの応用をしていく．

基本的な操作方法

Rを使って計算しよう

今回の目標：

- RStudioの準備をしよう.
- Rを使った基本的な計算方法に慣れよう.
- 分散・標準偏差を手計算で算出してみよう.

RStudioを使うための準備

- RStudioで使うディレクトリを決める
  - 初回のみ準備が必要
- ".R"ファイルを作成する
  - 毎回準備が必要

RStudioで使うディレクトリを決める

- 最初にRStudioで使うディレクトリ（フォルダ）を決めます．この時，【フォルダまでの間に2バイト文字（日本語）やスペースが入らないように気をつけて下さい】．
  - 使えない文字：あいうえおかきくけこ「」【】今日の天気は晴れ！１２３４５６７８９０
  - 使える文字：abcdefghijkelmnopqrstuvwxyz ABCDEFGHIJKLMNOPQRSTUVWXYZ1234567890
- また，windowsでもmacでもこの作業は変わりません．

“.Rproj”ファイルを作成する

- 毎回，授業の際にRで作業をする際には，指定した作業用フォルダにある“.Rproj”ファイルから開きます．
- 作業フォルダを決める（必要に応じてフォルダを作成する）
- RStudioを起動する．
- 右上の現在いるフォルダを示している文字をクリックする．
  - 場合によっては“(None)”と表示されているかもしれません．
- “New Project…”をクリックする．
- “Existing Directory”をクリックする．

- “Browse…”をクリックして，先程決めた作業フォルダまでたどり着く
- “Open”をクリックする.
- “Create Project”をクリックする.
- 完了

“.R”ファイルを作成する

- 授業中に作成したファイルを保存するために，“.R”ファイルを作成する．
  - これは毎回行います．
- 左上にある白い四角の左上に緑のプラスが書いてあるヤツをクリックする．
- “R Script”をクリックする．
- “Untitled 1”と書かれたファイルができる．
- “ctrl + s”を押すと保存ができる．
  - Macで作業している場合は“cmd + s”

- ファイル名として“今日の日付_学籍番号”を設定する.
  - ex.“20200929_学籍番号.R”とする
  - 学籍番号を入れてもらうのは，後ほど提出してもらうため.
- “Save”をクリックする.
- また，このファイルに書いた数式は“ctrl + Enter”（Macの場合は“cmd + Enter”）でその行の計算をRに読み込ませることができます.

実際に計算をしてみよう

加減乗除

```
123 + 123   #
```

```
## [1] 246
```

```
123 - 123    #
```

```
## [1] 0
```

```
23 * 123     #
```

```
## [1] 2829
```

```
123 / 123    #
```

```
## [1] 1
```

```
123 ^ 2   #
```

```
## [1] 15129
```

加減乗除の基本はこのような形です．少し演習問題を解いてみましょう．

注意です.

- 「#」から始まるとコメントアウトができます.
- 要は，「計算式」ではなく，「ただの文字」として認識されます.

演習問題1

- 113 + 987
- 2135 + 231
- 9832 - 3422
- 12348 - 8976
- 17 * 16
- 3298 * 5
- 285195 / 5
- 12387 * 33
- 324 ^ 2
- 89 ^ 4

その他の計算

```
sqrt(144)#
```

```
## [1] 12
```

```
1234 %/% 123  #
```

```
## [1] 10
```

```
1234 %% 123  #
```

```
## [1] 4
```

演習問題2

- (34×2)+(43-12)
- 23×(92-9)
- (53+23)の5乗
- (334-56)÷90
- (34×2)+(43-12)
- (3221+239)÷(87+27)の整数商
- (751×90)÷(5412/32)の剰余

オブジェクト指向

- 「オブジェクト」：データやモデル式などを入れる「何でも箱」
  - Rではモデル式，データなどをオブジェクトに入れて考える
  - 数式やデータをいちいち書くのは大変．．．
  - オブジェクトに入れることを「代入する」と言う

オブジェクトに入れて計算する：

```
x <- 5 #x      5
x #x
```

```
## [1] 5
```

```
(y <- 3) #(  )
```

```
## [1] 3
```

```
x + y
```

```
## [1] 8
```

```
x * y
```

```
## [1] 15
```

```
x - y
```

```
## [1] 2
```

よくないオブジェクト：

```
x <- 5
x <- 16
x
```

```
## [1] 16
```

- これで出力すると，xに16が代入されてしまっている．
- 基本的には違う語を使うようにしたい．

```
# NA <- 3
```

「NA <- 3 でエラー: 代入の左辺が不正 (do_set) です」と怒られる．これは「NA」がデータがないことを示す記号として指定されているため．　他にも指定されている語がいくつかあるが，怒られたら違う文字を割り振れば良い．

オブジェクトには文字を入れることが可能

```
A <- "Pen"
B <- "Pineapple"
C <- "Apple"
paste(A, B, C, A)
```

```
## [1] "Pen Pineapple Apple Pen"
```

- paste：複数の文字列を結合して，一つの文字列にする関数

今日の課題

今日の課題

- 各ページにある演習問題を行う.
- 今日作業した.Rファイルを提出してください.
- 提出方法は案内する.

おうちでやっておこう

以下の動画を見て作業をしてください.

- RおよびRStudioのインストールをしておきましょう（いずれも明治大学データ解析論Iの講義資料より）.
  - Rのインストール：https://www.youtube.com/watch?v=tJQmdNXkeso
  - RStudioのインストール：https://www.youtube.com/watch?v=9AjdZaYOQp4

次回の案内：

次回の案内：

- 記述統計量の算出方法を学びます.

Rでデータを扱う時に注意すべきこと

Rでデータを扱う時に注意すべきこと

- 必ず数字／文字は半角で入力する.
- 日本語は使わずにローマ字を使用する.
- コメントアウト（コードではなく，関係ないメモを入れること）をするときは半角の「#」から始める.
  - メモする内容は全角でもよい.
- ファイル名およびパスには決して全角の文字（ひらがな，カタカナ，漢字，全角スペースなど）を入れてはいけない.
  - 半角英数字だけにする.
- 慌てずに落ち着いて操作すれば，決して難しくない.
  - 1つずつ落ち着いて作業することを心がける.
- 「わからない」ことを恐れない
  - 周りの友人に聞いたり，教員に確認したりしよう.