応用心理学I

データ収集後探索的解析（テキストマイニング）　その1.5

人文社会系学生のための R ブートキャンプ 事前学習のキャッチアップ + R実習

2022.06.25

佐藤浩輔

位置づけ

- 6/25まで: 事前準備＋事前学習
- 6/25午後 **事前学習のキャッチアップ＋ R 実習 ←この講義**
- 6/26午前 質的データの量的コーディング
  - 中分遥先生（高知工科大学）担当
- 6/26午後 テキストマイニング講義
  - （前半）テキストマイニング理論編
  - （後半）テキストマイニング実践編

講義にあたって

- 自己紹介
- 講義について
  - 概要
  - 扱う内容
- 講義の進め方
  - 講義の形式について
  - Zoomの使い方
  - 講義時間
- 課題について
- サポートサイトについて

講義について：概要

- ①事前学習: Rブートキャンプ
  - コンピュータを用いた基本的なファイル操作・R を用いたデータ処理の基礎を学ぶ
- ②事前学習のキャッチアップ＋ R 実習 (この講義)
  - 事前学習を踏まえたうえで簡単なR実習
- ③テキストマイニング 入門：講義編
  - テキストマイニング（計量テキスト分析）と自然言語処理の概要を学び、何ができるかを知る
  - テキストを用いた研究のデザイン、研究計画の立て方を学ぶ
- ④テキストマイニング 入門：実習編
  - 実際にデータを扱いながら学ぶ

講義について：扱う内容

目的: 事前学習で学んだことの復習と実習

- 事前学習のふりかえり
  - 基本的には事前学習のサポート時間
- 実際のデータを使って分析してみる

講義の進め方①

- この講義ではRブートキャンプに引き続き、Rを使います
- 実習部分を試すには R がインストールされている環境が必要です
  - Windows環境を想定しています
- オンライン講義なので講義動画を見返す前提 の構成です
  - ×全部記憶する
  - ○必要な時に検索・参照する
- スライド・R コードは講義終了後公開します
- ハンズオン形式にはしません
  - i.e., 皆さんが操作するのを待つ時間はとりません
  - あとで復習しながら試してください

講義の進め方②

Zoomの使い方(再掲)

- マイク/ビデオについて
- 喋っていないときはマイクをミュート
    - ビデオ表示（カメラ）はオフ
    - 帯域節約のため基本的にオフで
- 講義中にコメント/質問等あるときは
    - ①Zoomのチャット機能を使う
    - ②直接発言する …どちらもOK

講義の進め方③

- 講義時間：13:00~16:35（予定）
- ところどころ休憩を入れます
- 無理のないように受講してください
  - 飲み物を補給する（熱中症予防）
  - 画面を凝視しすぎない
  - 休憩中に体を動かす

課題について(再掲)

課題があります

- 6/25-6/26の講義で学んだことを使ってテキストを分析する内容です
- 課題はR以外の言語を使ってもOKです
    - ただし分析の過程がわかるものに限る

※Rの知識を直接問うような課題ではありません

詳細は詳細は6/26の講義の終了時にお伝えします

サポートサイトについて(再掲)

- GitHub上の講義ページ https://github.com/satocos135/lecture2022
  - このページに講義資料をアップロードします
    - スライド
    - 分析用データ
    - などなど
- Discourse
  - 掲示板形式のWebアプリケーション
  - 受講生・聴講生限定で閲覧および書き込みができます
    - 事前に送った招待リンクからアクセスするとアカウントを作ってログインできます
  - 技術的サポートは基本こちらで

事前学習のふりかえり

実習① 質的変数の解析: Titanic dataset

タイタニック号沈没事故

- 1912年4月15日、北大西洋で起きた海難事故
  - 当時世界最大の客船であったタイタニック号が氷山に衝突し、沈没
  - 乗船していた2,224人のうち1,513人が亡くなる
  - 文化的影響も大きく、事故以降小説や映画など様々な作品が作られる

データセット

Titanic Dataset - https://hbiostat.org/data/repo/titanic.html

- タイタニック号の乗客の生存状態に関する公開データセット
  - タイタニック号の乗客1309名に関するデータ
    - 800名あまりいた船員のデータは含まれない
  - データ出典: Encyclopedia Titanica https://www.encyclopedia-titanica.org/
    - タイタニック号に関する総合的な参考資料集
    - 研究者が分析用のデータセットとして整備
  - Kaggleの入門コンペとしても有名
- 今回の実習ではこのデータを使って「探索的に」データ分析をする

問: タイタニック号において何が乗客の生存につながったのか？

14の変数:

`pclass` / `survived` / `name` / `sex` / `age` / `sibsp` / `parch` / `ticket` / `fare` / `cabin` / `embarked` / `boat` / `body` / `home.dest`

これらの変数のうち、何が生存につながったのか？

分布の確認

- 変数の値の分布を確認していく
  - 質的(カテゴリカル)変数には `table()` 関数を使うとわかりよい
    - `useNA` オプションを指定するとNAの数も表示してくれる
      - `useNA='always'` ...NAの頻度を表示
      - `useNA='ifany'` ...NAの頻度を表示(NAがない場合は表示しない)
  - 量的変数（特に値が連続のとき)にはヒストグラム(`hist()` 関数)を使って図示するとよい

２変数の記述統計

- 「何が生存につながったのか」を検討するために各変数と生存との関係性をみていく

質的変数×質的変数

- 性別(2値)×生存(2値)のような質的変数同士の関連を見て行く
  - クロス表/分割表
    - χ^2 検定: 独立かどうか検定できる
    - 連関の指標
  - モザイクプロット
    - クロス表を積み上げグラフとして図示
    - 面積が各セルに比例

分割表

- 変数の組み合わせについて頻度を集計したもの
- 連関: 変数間の相互関係
  - 「変数Aの値によって変数Bの値が変わるか」
  - 相互関係がない=独立である

	X	Y
A	10	10
B	50	50

連関がないケース

	X	Y
A	50	10
B	10	50

連関があるケース

連関

- χ^2 検定
  - 独立性の検定: 帰無仮説「行と列は独立」
- 連関の指標
  - ϕ係数: 2×2のクロス表のときのみ使える
    - 2値変数間の相関係数に等しい
  - ユールのQ
    - 関連が強いほど絶対値が1に近づく
  - クラメールのV
    - 1に近いほど関連が強い

水準の統合

- χ^2 検定の適用基準: 期待度数が5未満のセルが全体の20%を超えてはいけない (Cochran, 1954; Cochran's rule)
    - 期待度数は周辺度数から計算する
- 水準を統合するかフィッシャーの正確確率検定を使う

量的変数×質的変数の図示

- 箱ひげ図 `boxplot()`
  - 質的変数の違いによって量的変数の値の分布がどう異なるか
- ヒストグラムの比較
  - 2水準くらいまでなら直接比較できる

量的変数と量的変数の図示

- 散布図 など

★三変数以上の解析

- 層別解析
  - 値や階級別に分析する(e.g., 年齢階級別)
- 回帰モデル
  - 複数の変数を同時に検討
  - 重回帰分析
  - ロジスティック回帰(従属変数がカテゴリカル変数の場合) など

Titanic datasetのまとめ

- データからわかること
  - 救命ボートに乗れたかどうかが運命の分かれ道
    - 女性や子供は優先的に救命ボートに乗れた
    - 旅客等級が高い方が救命ボートに乗れた
  - それ以外の要因
    - 性差: 男性の方が死亡率が高い
      - 身体的特徴(e.g., 体脂肪)の違い?
- 分析上の注意
  - 分布や変数の特徴をつかむことが重要
  - データ内の情報だけではわからないことがある→外の世界の情報も使う

実習② 集計のレッスン: 新型コロナ関連データ

データセット

- 新規陽性者数の推移 https://www.mhlw.go.jp/stf/covid-19/open-data.html
  - 厚生労働省の公開するオープンデータ
  - 日別の都道府県別新規陽性者のデータ(報告日ベース)
- 新型コロナワクチン接種状況 https://info.vrs.digital.go.jp/dashboard
  - デジタル庁の公開するオープンデータ
  - 日別の都道府県別接種状況のデータ(集計日ベース)
  - 医療従事者接種数は含まれていない(06/25現在)
- これらのデータを使いのデータセットの結合・集計の練習を行う

※ともにデータ形式を改変したものを分析に使用

filter, select, group_by, join(merge)

- `select(列名)`
  - 抽出する列を選択
- `filter(条件1, 条件2, ...)`
  - 条件によってデータを抽出
- `group_by(変数)`
  - 変数の値によってグルーピング
  - `summarise(変数1 = 集約関数, 変数2 = ...)` で集計
- `join(データセット, by=キーになる列名)`
  - データセットをマージ
  - `inner_join()`, `left_join()`, `full_join()` などがある

集計のレッスンのまとめ

- 頻度データ: 様々な観点から分析できる
  - 集約 / 結合
  - 比較

→ 自在にデータを加工できるようなスキルセットが重要