現代日本学演習 II

統計分析の基礎

田中重人 (東北大学文学部准教授)

3 年生対象：2022 年度 前期＜金 4 ＞ Google Classroom クラスコード **sbm375y**

1 『講義概要』記載内容 ＋α

◆ 講義題目：統計分析の基礎

◆ 授業の目的と概要：意識調査・テスト・実験などのデータはどのように分析すればいいでしょうか。この授業では、小規模の標本調査を念頭において、統計分析の基礎的な手法を学びます。これまで統計的な分析をおこなったことのない人を対象に、初歩から講義します。同時に、コンピュータを実際に使って、データ分析の実習をおこないます。

◆ 到達目標: (1) 統計分析の基礎を理解する; (2) 実際にデータ分析ができるようになる

◇ テキスト：吉田寿夫、1998『本当にわかりやすいすごく大切なことが書いてあるごく初歩の統計の本』北大路書房。

◇ 成績評価の方法：授業中の課題と宿題 (70%) と期末レポート (30%) を合計して評価する。

※ 卒業論文等で質問紙調査を予定している者は、現代日本学演習 I 「質問紙調査の基礎」(前期 金 5) および現代日本学演習 V 「実践的統計分析法」(後期 金 4) も受講することがのぞましい。

2 授業予定 (Google Meet 会議による)

(1) イントロダクション [4/15]
(2) 統計分析の基礎 [4/22, 5/6]
(3) 度数分布表とグラフの利用 [5/13–27]
(4) 復習と進度確認 [6/3]
(5) 平均値の比較 [6/10 – 7/1]
(6) 推測統計 [7/8–29]
(7) 期末レポート提出期限 [8/12]

[ ] 内の日付は、学期前のおおよその計画をあらわしているが、 実際の授業の進行状況によって前後にずれることがある。

3 課題とフィードバック

この授業では、ほぼ毎回、課題を出します。提出期日は、毎週**木曜日の正午** です。

- Google Classroom 「現代日本学演習 II」に登録しておくこと。
- 課題に関する質問は随時受け付ける。Google Classroom での質問を推奨するが、電子メール (下記参照) その他の手段で質問を出してもよい。
- 提出された課題の内容によっては、再提出を指示することがある。また、特に指示がない場合も、書き直したものを再提出してよい。
- 課題は 1 回につき 6 点。最初に提出された内容でいったん点数をつけるが、再提出された場合には加点することがありうる。
- 何を調べてもよいし、誰と相談してもよいが、それらの情報源について解答の中で説明すること。

これとは別に期末レポート (8/12 締切) があります。課題はつぎのとおり。

> クロス表と平均値の比較の両方について適当な分析をして結果を解釈する。それぞれ推測統計（区間推定または統計的検定）の結果もつけること。図・表は読みやすく整形し、論文としての体裁を整えること。授業で配布した以外のデータを使ってもよいが、その場合はデータについての解説をレポート中にふくめること。

4 受講環境について

4.1 PC 等の準備状況

この授業では、つぎの 2 種類のソフトウエアが必要です。

- 統計分析
- グラフ作成

前者については PSPP、後者については Google スプレッドシートを使うことを考えています。ただし、初回授業時におこなう受講者各自の利用環境調査の結果、これらのソフトウエアを利用できない受講者がいる場合は、変更することがありえます。

4.2 PSPP について

PSPP は、SPSS (という広く普及している有料の統計分析ソフト) に似せて作られた無料のソフトウエアです。試しにインストールしてみて、うまく動くかどうか確認して、結果をお知らせください。

- Windows の場合、https://sourceforge.net/projects/pspp4windows/files/ から「Download Latest Version」をクリックしてください。
- 他の OS については、https://www.gnu.org/software/pspp/get.html を見てください

5 数学的知識の調査

Google Classroom に課題を出しているので、答えておいてください (木曜正午まで)。これは受講者の予備知識とレディネスについて知るためのもので、採点対象外です。わからない場合は「わからない」と書いておいてください。

現代日本学演習 II「統計分析の基礎」

第 2 講　PSPP 入門

田中重人 (東北大学文学部教授)

[テーマ]　データ配布 ; PSPP の基礎知識

1　前回課題について

予備知識の調査について解説

2　模擬データ入力実習

配布した架空の回答票 (別紙) をもとに、データを入力してみよう。

まず PSPP を起動する。通常、「データエディタ」ウインドウの「変数ビュー」タブが表示された状態になるはず。

ここで、まず「変数」を定義する。

- 変数名を必要なだけつくる。 今回は a, b, ..., e とでもしておこう。 変数名は自分がわかればどんなものでもよい。 日本語も使える。 なお、変数名以外のフィールドは入力しなくてよい
- 書き終わったら「データビュー」タブに切り替えて、いちばん上の行に変数名がならんでいることを確認する。

つづいてデータを入力していく。 今回は 3 人分のデータを用意してあって、変数は 5 個なので、 3×5 の行列型のデータができるはずである。

適当な名前で保存してみる。

- PSPP データファイル (なんとか.sav) ができていることをたしかめる。
- このデータファイルは授業終了時に削除すること。 (次回以降の授業ではつかわないので、おいておく必要はない。)

この方式は PSPP でデータを入力するときのいちばん簡便な方法であるが、 大きなデータはあつかいにくい。実際の調査データの入力では、Excel ファイルやテキストファイルでデータを用意しておいて、PSPP に読み込むのがふつうである。

3　データ配布

この授業で使用するのは、1995 年 SSM 調査 B 票の一部。調査については、配布資料のほか、『日本の階層システム』(2000 年、全 6 巻、東京大学出版会) を参照。

- 全国から 70 歳以下の有権者を層化 2 段無作為抽出
- 訪問面接法

ただし、配布したのはこの調査データの一部に限定したものである。

- 意識項目と基本的属性に限定 (調査票の×印はデータセットにない項目)
- 250 ケースをランダムに抽出
- 菅野剛さん (日本大学) による変数ラベルが入っている

毎回の授業で使うので、忘れないこと (調査票も)。

このデータは、この授業でのみ使用を許可されているものである。データが流出しないように注意すること。また、期末レポート提出時に、データを削除すること。

なお、自分の研究用のデータがある人は、課題などではそれを使ってもよい。ただし事前に相談すること。

4 PSPP の基礎知識

4.1 データ・セット

PSPP のデータ (「データエディタ」ウインドウで見られる) は、ケース × 変数の行列型になっている。

- 「ケース」は、個々の調査回答者にあたる
- 変数には「変数名」がついている (歴史的事情により、英数字 8 文字以内)。これだけだとわかりにくいので、変数名以外に「ラベル」をつけるのがふつう
- 無回答などの欠損値はどうなっているか?

4.2 ウインドウ構成

- データ・エディタ (上記)
- 出力ビューア (→ 分析結果やエラーメッセージなど)
- シンタックス・エディタ (プログラムを直接編集するときに使う)

4.3 分析の一般的な手続き

「データエディタ」のメニューの使いかた

(1) 分析手法をえらぶ
(2) 変数を指定
(3) 必要なオプションを指定
(4) 「OK」をクリック

結果は別ウインドウ (出力ビューア) に表示される

- 左側に目次、右側に出力内容
- エラー表示もここに出る
- PSPP のプログラム (シンタックス) も表示される

4.4 度数分布表を出してみる

- データエディタのメニュー →「分析」→「記述統計量」→「度数分布表」
- 左側の変数リストから、分析対象とするものを選択して、右側のパレットに移動させる
- 下側の「統計」のチェックをぜんぶ外す
- 「OK」

4.5 他のアプリケーションとの連携

PSPP の出力はあまりきれいでないので、レポートを作成するときなどは、Excel や Word に表を貼り付けて整形することになる。が、出力ビューアの**表をそのままコピーすることができない。**

そこで、出力結果をいったん HTML で保存するなどして変換する。

- 出力ビューアのメニューから「ファイル」→「書き出し」を選択
- ファイル名の最後を 「.html」にして保存
- 該当ファイルをブラウザで開く
- 該当部分をコピーして、他のアプリケーションに貼り付ける

アプリケーションの種類によっては、「オープンドキュメント」「CSV」等での保存も使える可能性がある。

川口秀樹 (2016)「[PSPP]インポート、エキスポート、マージ」<https://note.com/xinzuzhai/n/n63b900f0bb86> などを参照。

5 変数値の再割り当て

ウインドウ上部のメニューバーから操作する

- 「変換」→「他の変数への値の再割り当て」
- 変換先変数の名前をつけ、「変更」を押す。名前は英数字だけにしておくのが無難 (記号や日本語を使うと、問題がおきることがある)
- 「今までの値と新しい値」の組を順次指定する。「今までの値」は範囲で指定することも、単一の値を指定することもできる
- 値の組を指定したら「続行」を押す (元の画面に戻る)
- 「OK」ボタンを押して実行する
- 出力ビューアを右端までスクロールして、新変数ができていることを確認
- 度数分布を確認
- 問題がなければ、名前をつけてデータセットを保存 (どこに保存されるかを確認しておくこと)
- 再割り当ての手順を示したシンタックスが出力ビューアに出るので、それも保存しておくこと

6 宿題

配布したデータを使い、年齢についての度数分布表を出力する。ただし、適当な年齢幅に区切ること。結果の表を、年齢幅の設定などがわかるよう整形して、どの年齢層が多いかなどのコメントをつけて提出。また、課題の途中でどこでつまずいたかなどの経過について書いてもよい。木曜日正午までに Google Classroom に提出。

ほかの人と自由に相談してよい。

教科書のほか、つぎの資料を参考にしてよい。

- 小木曽道夫「SPSS の使い方」<http://www2.kokugakuin.ac.jp/~ogiso/spss/>
- 浦上昌則「SPSS おたすけマニュアル」<http://www.ic.nanzan-u.ac.jp/~urakami/u-spss/SPSS_f.html>
- 保田時男「SPSS 操作メモ 岩井・保田（2007）準拠版」 <http://www2.itc.kansai-u.ac.jp/~tyasuda/files/2013/methoda/spss_memo_2.pdf>

ただ、これらは SPSS についての説明であるため、PSPP 操作とは一部くいちがいがある。
これら以外の資料を使ったときは、課題中に書いておくこと。

現代日本学演習 II「統計分析の基礎」

第 3 講　統計分析の基礎

田中重人 (東北大学文学部教授)

[テーマ]　度数の利用と統計分析の基礎

1　前回課題について

操作については 前回資料 後半部分を参照

- 用紙上部に番号・名前を記載
- 70 代はなぜ少ないのか
- 年齢層別の人口の変動 → 1995 年の人口ピラミッド <http://www.ipss.go.jp/site-ad/TopPageData/1995.png> と出生力の変動 <http://www.ipss.go.jp/syoushika/tohkei/Popular/P_Detail2020.asp?fname=G04-01.gif>
- 変数ラベルの利用：値の再割り当ての変数を指定する際に、変数名とは別に「ラベル」をつけることができる。あとから「データビュー」の「変数ビュー」でも
- 値ラベルの利用：「データビュー」の「変数ビュー」タブで、変数の値に「ラベル」をつける →分析結果出力に表示される
- シンタックス (syntax) の利用

2　度数分布表の読みかた

- 度数
- 相対度数 （%）
- 累積度数・累積相対度数
- 欠損値のあつかい

(教科書 p. 27–31)

3　データセットを分割する方法

PSPP には、特定の変数の値によってデータセットを分割するコマンドがある

- メニューから「データ」→「ファイルの分割」を選ぶ
- 適当な変数を選び、「グループの比較」を選び、OK

いったんこの操作をすると、それ以降は、すべての分析が、その変数の値ごとに別々におこなわれる。
元に戻すときは、「データ」→「ファイルの分割」→「すべてのケースを分析」

4　データ収集から分析まで

(1) データの収集 (実験／観察)
(2) 分析可能な形に加工
(3) データ・セット作成
(4) クリーニング
(5) データの特徴を少数の数値に要約 ＝ 記述統計
(6) 誤差の評価 (この手続きの一部が推測統計)

(教科書 p. 1–6)

5　標本抽出

標本抽出の 4 段階モデル

- 理論母集団 = 興味の対象となる人や事物の全体
- 調査母集団 = 調査の対象とする具体的な範囲
- 計画標本 = 母集団から抽出した対象者のこと
- 有効標本 = 調査の結果あつまった有効なデータ

「無作為抽出」(random sampling) とは：

- 母集団から計画標本を選ぶ際に、母集団にふくまれるすべての個体の抽出確率が等しくなるように抽出する
- この結果として、**「確率標本」** (probability sample) がえられる

統計的な推測のための理屈は、確率標本を前提として組み立てられている。母集団の人口がわかっていて、全個体を網羅した台帳がないと、無作為抽出はできない。実際にはそういうことはないので、いろいろ工夫して無作為抽出に近づける。

「層化 2 段無作為抽出」はその方法のひとつ：

- まず「地点」を抽出 (第 1 次抽出)
- その際、地域・都市規模等で地点抽出数を割り当てておく (層化)
- その地点の台帳から個人を抽出 (第 2 次抽出)

6　宿題

(1) 教科書 pp. 7–16 を元に、「データの種類」の分類 (名義尺度、順序尺度、間隔尺度、比率尺度) についてまとめよ
(2) SSM 調査の質問項目のうち、比率尺度に当たるものはどれか
(3) 累積パーセントに意味があるのはどの種類のデータか
(4) 「収入」や「学歴」を比率尺度として分析するにはどのようにすればよいか

木曜正午までに Google Classroom に提出。