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：2019 年度 前期＜火 3 ＞コンピュータ実習室 (文学部棟 7F)

## 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%、うち 20%は 5 月 28 日予定の進度確認課題) と期末レポート (30%) を合計して評価する。

◇ その他：実習室のコンピュータ台数が限られているため、受講人数を制限することがある。

※ 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 4) および現代日本学演習 V 「実践的統計分析法」(後期 火 3) も受講することがのぞましい。

## 2 授業予定

(1) イントロダクション (4/9)
(2) SPSS 入門 (4/16)
(3) 統計分析の基礎 (4/23) [序章]
(4) 度数分布表とクロス表 (5/7–21) [1 章, 4 章]
(5) 復習と進度確認 (5/28)
(6) 平均値の比較 (6/4–18) [2 章, 5 章]
(7) 推測統計 (6/25 – 7/23) [6 章, 8 章]
(8) 期末レポート (8/14 提出) → (9/5 以降に返却)

[ ] 内は、教科書の参照箇所。( ) 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 受講者との連絡とフィードバック

- 毎回の課題・宿題は、コメントをつけて返却します (再提出を求めることもあります)。
- 期末レポートは、採点後に返却します。
- 課題・宿題は、特に指示のあるものをのぞき、ISTU (東北大学インターネットスクール: https://istu3g.dc.tohoku.ac.jp) のレポート機能による提出とします。 提出期限は、原則として 授業前日 (月曜) 正午 (12:00) です。
- ISTU の「現代日本学演習 II」に「受講申請」をしておいてください （受講者情報の自動的登録は、履修登録完了以降)

現代日本学演習 II「統計分析の基礎」

# 第 1 講　イントロダクション

田中重人 (東北大学文学部准教授)

## 1　受講者の興味と数学的知識の調査

→　別紙

## 2　コンピュータ実習室について

- 入室・退室に学生証が必要 (正規の学生以外は、登録申し込みが必要。ない人は、教務係で臨時カードを借りること)。 文学部正規学生以外 (研究生や他学部の学生など) は登録しておくこと。
- 土足・飲食・喫煙厳禁。
- 退出時には必要事項を紙に記入。

## 3　コンピュータの起動と終了

使いはじめるときは……

- コンピュータ本体の電源を入れる
- 表示されるお知らせをひととおりよむこと

使い終えるときは……

- 「マイドキュメント」などに保存してある自分のファイルを削除
- 画面左下の「スタートメニュー」から「シャットダウン」を選択
- コンピュータ本体の電源が切れたことを確認
- USB スティック・メモリなどをわすれないこと

ファイルの保存場所について

- 教室のコンピュータの内蔵ディスクに、個人のファイルを置いておくことはてできない。
- 授業中に必要なファイルは「マイドキュメント」フォルダに一時的に保存してよいが、 授業が終わったら自分のスティック・メモリ等にコピーして、 内蔵ディスクのほうのファイルは削除すること。

# 4　ISTU への登録

https://istu3g.dc.tohoku.ac.jp にログイン (東北大 ID が必要)

- 現代日本学演習 II (田中重人) を探す
- 受講申請する

この授業では毎回の課題や中間試験、期末レポートをISTUを通して提出するので、使いかたを覚えておくこと。

なお、ISTUに受講申請するのは、この授業の資料にアクセスしたり課題を提出したりするためであり、正規の履修登録とは関係ないので注意。**履修登録は、各自「学務教育システム」で別途おこなう** こと。また、正規に履修しない受講者も、ISTUには登録しておくこと。

# 5　模擬データ入力実習

## 5.1　SPSS の起動

- スタートメニューから「プログラム」→「IBM SPSS Statistics」→「IBM SPSS Statistics 24」で起動する。(※ここで何かエラーメッセージが出るかもしれないが、気にせず「続行」または「OK」する。)
- 何をするか選ぶ画面が出る場合は、「新規データに入力」をクリック

## 5.2　データ入力

配布した架空の回答票をもとに、データを入力してみよう。
まず「変数」を定義

- 「データエディタ」ウインドウのいちばん下の「変数ビュー」タブに切り替える
- 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、 $3 \times 5$ の行列型のデータができるはずである。

適当な名前で「マイドキュメント」内に保存してみる。

- 「マイドキュメント」を開いて、SPSS 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、コピーしておく必要はない。)

この方式は SPSS 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、SPSS に読み込むのがふつうである。

**現代日本学演習 II　(田中重人)**

## 受講登録フォーム

氏名 (よみがな) :

学年 :

学籍番号 :

所属 (現代日本学研究室以外の場合) :

研究内容 :

・自宅でパソコンを使えますか?　**ある / ない**

・ISTU を使った経験がありますか?　**ある / ない**

・SPSS を使った経験がありますか?　**ある / ない**

・その他の統計ソフトを使った経験がありますか?　**ある / ない**

・コンピュータ・プログラムを作成したり、プログラミングの授業を受けたりしたことがありますか?　**ある / ない**

**ある場合 → 言語名 (　　)**

以下は採点用

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 宿題 | | | | | | | | | | | | | |
| 課題 | | | | | | | | | | | | | |
| 参加 | | | | | | | | | | | | | |

| | | | |
|---|---|---|---|
| 進度 | | | |
| 期末 | | | |

現代日本学演習 II「統計分析の基礎」

# 第2講　SPSS入門

田中重人 (東北大学文学部准教授)

[テーマ]　SPSS の基本的な操作

## 1　今回の課題

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。結果を Word に貼り付け、年齢幅の設定などがわかるように整形して、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。ISTU で月曜日正午まで。

周囲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/>
- 森際孝司「SPSS の基本操作 2」<http://www.koka.ac.jp/morigiwa/sjs/les10201.htm>
- 森際孝司「データ変容」<http://www.koka.ac.jp/morigiwa/sjs/les10401.htm>
- 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>

SPSS バージョンの違いなどにより、実習室 PC の操作と上記資料の説明に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。

## 2　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 (次回説明)
- 訪問面接法

調査票は <http://srdq.hus.osaka-u.ac.jp/PDF/SSM95BJ.pdf> にもある。

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。

# 3 SPSS の基礎知識

## 3.1 データ・セット

SPSSのデータ (「データビューア」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?

## 3.2 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)

## 3.3 分析の一般的な手続き

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
- Ver. 19 以降では SPSS のプログラム (シンタックス) も表示される

印刷

- 左側の目次で、印刷したいものを選択
- 印刷前にプレビュすること
- 実習室のプリンタについて、電源の入れかた、ジョブの確認・取り消し、タイル印刷 (2 面, 4 面, ...) の方法を習得しておくこと
- 実習室ではプリンタ用紙を供給していないので、紙は自分で調達する。また、印刷枚数に制限があるので注意すること。

### 3.4 その他のアプリケーション

実習室の PC では、Microsoft Office (Word や Excel など) が使える。
SPSS の出力ビューアから表を Excel や Word に貼り付ける方法を覚えておくこと。

## 4 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

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 (東北大学文学部准教授)

[テーマ]　累積度数の利用と統計分析の基礎

## 1　前回課題について

- 用紙上部に番号・名前を記載
- 40 代はなぜ多いのか →1995 年の人口ピラミッド <http://www.ipss.go.jp/site-ad/TopPageData/1995.png>
- 70 代はなぜ少ないのか
- 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

「再」マークがついている人は再提出 (来週月曜正午まで)

## 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　今回の課題

年齢（カテゴリ統合していない元の変数）の度数分布から、中央値と四分位を求めよ (提出は不要)。
参考資料：

- 教科書 p. 43
- 総務省統計局「なるほど統計学園高等部：データの特性を見よう」<http://www.stat.go.jp/koukou/howto/process/proc4_3_1.htm>
- 船津好明「統計計算の方法」<http://www.wwq.jp/stacal.htm>

また、任意のパーセンタイル (percentile) を求める方法を考えること。

## 4　発展問題 (余裕のある人のみ)

次の情報を参考にして、カテゴリ統合した後の年齢の度数分布表から中央値を求める方法を考える

- Yahoo! 知恵袋 <http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q1214126522> の回答
- 青木繁伸「中央値 (M_e)」<http://aoki2.si.gunma-u.ac.jp/lecture/Univariate/median.html>「One more step!」以降
- 森・吉田 (1990, p.15)

## 5　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)

## 6　標本抽出

標本抽出の 4 段階モデル

- 理論母集団 (universe) = 興味の対象となる人や事物の全体
- 調査母集団 (population) = 調査の対象とする具体的な範囲
- 計画標本 (designed sample) = 母集団から抽出した対象者のこと
- 有効標本 (valid sample / case) = 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)

## 7　宿題

(1) 教科書 pp. 7–16 を元に、「データの種類」の分類についてまとめよ
(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) 「中央値」「四分位」など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

ISTU で 5/6 (月) 正午までに提出。

現代日本学演習 II「統計分析の基礎」

# 第4講　度数分布表とグラフの利用

田中重人 (東北大学文学部准教授)

[**テーマ**]　グラフの種類と、それらの適切な使い分け

## 1　前回課題について

(1) 「データの種類」の分類について

- 尺度水準によって使える計算方法が違う (= 使える分析法が違う) ことを理解しておくこと
- 測定している対象そのものの性質ではなく、データにおいてどのような数値が割り当てられているかが問題である
- 上位の尺度は下位の尺度の性質を兼ね備える (たとえば間隔尺度のデータは順序尺度としても分析できる)

(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 → 人数、年数など

(3) 「中央値」「四分位」などに意味があるのはどの尺度水準か → 順序尺度以上

(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 → 「変数値の再割り当て」で適当な値に変換：

- 「収入」については、各カテゴリを適当な金額に変換すればよい (たとえば各階級の真ん中の値をとる)
- 「学歴」については、その学歴を取得するのに必要な標準的年限で置き換えることが行なわれている（単に「教育年数」とよぶことが多い)。

```
 1 → 6
 2 → 8
 3, 4, 5 → 11
 6 → 14
 7 → 17
12 → 9
13 → 12
14 → 14
15 → 16
16 → 18
```

「再」マークがついている人は再提出 (来週月曜正午まで)。どこをどう修正したかがわかるようにすること。

## 2　データセットを分割する方法

SPSS 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」

## 3　今回の課題

つぎの 3 種類の度数分布について、適切なグラフを描け。SPSS または Excel を利用すること。白黒で印刷することを念頭に置いて作成する。Word などに貼り付け、コメントをつけて提出 (ISTU に月曜 12:00 まで)

(1) 男女比
(2) 本人年収の分布
(3) 本人年収の分布の男女比較

教科書 32–37 ページを参照。

## 4　グラフの利用

分析結果は、通常、表またはグラフで示す。

**表 (table):**　正確な数値がわかるが、全体の傾向を読み取るには熟練が必要
**グラフ (graph/chart):**　全体の傾向が簡単に読み取れるが、正確さは犠牲になる

初心のうちは、表とグラフの両方を作成して読んでいくのがよい

## 5　度数分布 (の比較) をあらわすグラフの種類

- 円グラフ (半数を超えているかの判別に便利)
- 棒グラフ (離散量のそのままの分布を示す)
- ヒストグラム (連続量を階級に区切って示す)
- 度数ポリゴン (度数多角形とも。複数の分布の比較に便利。教科書 p. 34)
- 帯グラフ (積み上げ棒グラフとも。教科書 p. 106)

現代日本学演習 II「統計分析の基礎」

# 第 5 講　クロス表分析の基礎

田中重人 (東北大学文学部准教授)

[**テーマ**]　クロス表の書きかたと読みかた

## 1　前回課題について

グラフは大きく 2 種類にわかれる：

(1) 一定の面積を分割して割合を示す: 円グラフ、帯グラフ、ヒストグラム、度数ポリゴンなど
(2) 位置または長さで量を示す: 棒グラフ、折れ線グラフ、散布図など

構成比 (全部足すと 100%) を示すには (1) のグラフを使うのが原則……だが、実際にはそうでないことも多い。

- 円グラフは「半数」を基準としてみるときに使えるが、それ以外の目的には不適当
- 連続量の度数分布は適当な階級幅に分けてヒストグラムを書くのが本来であるが、Excel などでは描きにくい (棒グラフの距離をゼロにして見た目をヒストグラム風にする)
- 複数の分布を比較するには度数ポリゴンがよい (実際には折れ線グラフとして書く)。二つの分布の比較では棒グラフを並べてもよいが、棒の色をはっきり違えないと、識別しにくい
- 人数がゼロのところがあるので注意
- 棒グラフ・折れ線グラフでは、縦軸の数値、目盛り、範囲に注意すること
- 3 次元 (3D) グラフは正確な数値がつかみにくい
- Excel では、凡例や軸数値やタイトルのほか、各カテゴリの人数なども表示できる
- カラーで作成すると、白黒印刷では読みにくくなることが多い
- 欠損値のあつかい

## 2　今回の課題

「性別」と「性別による不公平」について、次の手順で「クロス表」(cross table) を作成する：

(1) メニューから「分析」→「記述統計」→「クロス集計表」
(2) 変数を「行」「列」にひとつずつ指定
(3) 「セル」にパーセンテージ (行・列の両方) を追加

出力を元に、次のことを考える (参照：教科書第 4 章)

- この表から何がわかるか
- 「行」の%と「列」の%は何を表しているか。またこのクロス表を解釈するときはどちらを見るのが適切か
- このクロス表をわかりやすく表示するにはどのようなグラフが適当か考え、実際に作成してみる (Excel を使用)

提出は、ISTU で月曜日正午まで。

現代日本学演習 II「統計分析の基礎」

# 第6講　連関係数とクロス表の解釈

田中重人 (東北大学文学部准教授)

[**テーマ**]　連関係数と%の関係を理解する

## 1　前回課題について

- 「行」と「列」の区別
- 行% と列% の使い分け: 原因→結果に対応
- SPSS では「○○ の %」と表示される (○○ は変数ラベル)
- 論文等に表を載せる場合は、行%か列%どちらか一方、適切なほうだけを書く
- グラフにする場合は、帯グラフ (積み上げ棒グラフ) で合計 100%になるようにするのが標準 (折れ線グラフまたは度数ポリゴンでもよい)
- Excel の「積み上げ棒グラフ」ではカテゴリー順序が逆転するので注意 (もとどおりにしたいときは、シート上の順序をいれかえる)
- 列%によるグラフになってしまう場合は、右クリック→「データの選択」で行／列を入れ替える
- 「レイアウト」→「線」で「区分線」を指定するとよい。
- 「全体」のグラフは不要

## 2　今回の課題

「性別」と「性別による不公平」のクロス表を作成する。ただし、「セル」「統計量」オプションで「観測度数」「期待度数」「残差」「標準残差」「カイ 2 乗」「Phi」「Cramer V」の数値を指定すること。

出力と教科書 (pp. 108, 116–117) をもとに、つぎのことを考える：

- 連関係数「Cramer の V」と「Pearson のカイ 2 乗」の間の数学的な関係 [式 4-19]
- 式 [4-17] のなかに、「Pearson のカイ 2 乗」「観測度数」「期待度数」「残差」「標準残差」はどのように表れているか
- 連関係数 $V$ の最小値・最大値はそれぞれいくつか。またどのような場合に最小値・最大値をとるか。

提出は、ISTU で月曜日正午まで。

なお、余力があれば、次のことも考えてみる：

- $2 \times 2$ クロス表におけるファイ係数 ($\phi$：教科書 p.110 [式 4-10]) は Cramer の $V$ とどのような関係にあるか

## 3　キーワード

**独立 (無関連 = independent):**　すべての列について行%が等しい (またはすべての列について行%が等しい) 状態

**周辺度数 (marginal frequency):**　クロス表の右端・下端に書く「合計」の度数

**期待度数 (期待値 = expected frequency):**　周辺度数を固定しておいて、独立な (架空の) クロス表をつくった場合、各セルに入る (と期待される) 度数

**観測度数 (frequency):**　各セルに入っている実際の度数

**残差 (residual):**　観測度数 − 期待度数

**標準残差 (standard residual):**　残差を期待度数の平方根で割ったもの

**$\chi$ 2 (chi-square):**　標準残差の平方和

**クラメールの連関係数 $V$:**　$\chi$ 2 を全度数で割り、セル数を調整したものの平方根

行・列の数が多いクロス表では、各セルの%を比較するのが大変である。また、%の差が大きいように見えても、度数が少ない場合には、実質的には大差ないと考えるべきであるが、そのようなことを判断するのもむずかしい。そこで、まずクロス表全体について「連関係数」を見ることで、行変数と列変数の「連関の強さ」を判断し、そのうえで細かく%を比較していくのが定石になっている。

## 4　今後の予定

5/28 進度確認。出題範囲は、今週の授業内容まで。持ち込み可 (ただし通信・相談禁止)。コンピュータで解答を作成して、ISTU で提出。

試験後は、通常通り授業。

現代日本学演習 II「統計分析の基礎」

# 第7講　平均と分散

田中重人 (東北大学文学部准教授)

[**テーマ**]　平均値と標準偏差の定義と計算

## 1　進度確認課題返却

問 1=3 点、問 2=4 点、問 3=4 点、問 4=3 点、問 5=6 点 (合計 20 点)

## 2　復習事項

### 2.1　SPSS の操作

- データエディタにおける「変数ビュー」の使いかた
- 「欠損値」(missing value) とは何か
- シンタックス (syntax) とは何か
- 変数値の再割り当ての方法
- グループに分割する方法
- 度数分布表における「パーセント」と「有効パーセント」のちがい
- 度数分布表における「累積パーセント」の利用法
- 中央値、四分位、パーセンタイルの求め方

### 2.2　統計分析の基礎など

- 尺度水準とは何か。それはなぜ重要か。
- 「母集団」(population) と「標本」(sample)
- Excel による棒グラフ、帯グラフ、折れ線グラフの書きかた

### 2.3　クロス表

- 「行」「列」「セル」「周辺度数」
- 「行%」と「列%」の使い分け
- クロス表をグラフにするときは、どのような種類のグラフが適切か

## 3　代表値と散布度

教科書 pp. 42–52 を読んで、「中央値」「四分位偏差」「平均」「標準偏差」の計算方法を理解する。特に、表 2-1 (p. 48) で何が計算されているかを考えること。

## 4　平均値と標準偏差

**平均 (mean):**　総和をデータ数で割ったもの

**分散 (variance):**　平均値からの偏差の 2 乗値の平均

**標準偏差 (standard deviation):**　分散の平方根 (SD と書くことが多い)

教科書の表 2-1 (p. 48) で何が計算されているかを理解する

- 平均と標準偏差はセットで使う
- 尺度水準による制限

## 5　宿題

教科書 p. 52 の練習問題 2-3 について、平均値と標準偏差を計算せよ。計算の途中経過がわかるように解答すること。ISTU で来週月曜正午まで。

現代日本学演習 II「統計分析の基礎」

# 第 8 講　平均値の比較

田中重人 (東北大学文学部准教授)

[テーマ]　ふたつのグループ間での平均値の比較

## 1　度数分布表のオプション

度数分布表の「統計量」オプションで「平均値」と「標準偏差」をチェック。

- 「記述統計」→「記述統計」でも出力できる。
- SPSS などの統計ソフトは、すこしちがう計算式で「標準偏差」を計算している (教科書 p. 48 注 6)。データが大きくなれば (およそ 200 以上なら) このことによるちがいはほとんどなくなるが、小さいデータ (たとえば 10 人程度) では大きなちがいになるので注意。

練習問題：前回宿題について、SPSS にデータを打ち込み、平均値と標準偏差を出力してみよう。

## 2　順序尺度の変数の「平均値」

平均値は、本来は、間隔尺度以上の水準の変数にしか使えない。しかし、実際には、一定条件を満たせば、順序尺度についても平均値をとっていいとする基準が使われている。

- 潜在的には間隔尺度のはず
- 測定のポイントが一定間隔

具体的には、4 点以上の尺度であって、正規分布に近似している場合 (教科書 p. 53–59)。これは、「偶然の積み重ねで形成されるものは正規分布にしたがう」という仮定による。

「正規分布に近似」しているかどうかは、通常、つぎの 3 点で判断する。

- 単峰性
- 左右対称性 (歪度)
- 中央への集中度 (尖度)

SPSS でヒストグラムを描いて検討するとよい。

「度数分布表」の「統計量」オプションで「歪度」「尖度」を指定すると、正規分布との乖離度を統計的に検討できる。これらの値は、正規分布のとき 0 をとり、絶対値が大きくなるほど、正規分布から外れる。およそ $\pm 2$ の範囲を超えていれば、正規分布からのずれが無視できない。

これらの条件を満たさない場合は非線形変換 (教科書 p.142–144) をおこなったり、順位に変換したりすることがある。あるいは、平均値を使わずに中央値を使って分析することもある。

なお、2 値の変数は、この条件にかかわらず間隔尺度とみなしてよいが、一定以上のデータ数があり、あまり偏っていないことが必要。

## 3　平均値の欠点

平均値は「はずれ値」(outlier) の影響を受けやすい。あまりにかけはなれたケースがあるときは

- 上下数%を取りのぞく (調整平均：教科書 p. 46)
- 順位に変換したり中央値を使って分析

などの方法を使うことがある。

また、極端なはずれ値がなくとも、左右非対称の分布の変数 (所得、人口、めったに起こらない現象の経験回数など) では、平均値より中央値の方が適切な代表値であることが多い。

## 4　ふたつのグループ間での平均値の比較

データをグループに分けて、それぞれ平均値 (＝層別平均) を求め、それらの間の差をもとめる。この差の大きさを、標準偏差を基準にして評価する。具体的には、effect size (ES) または 相関比 ( $\eta$ ：イータ) という統計量を使う。

### 4.1　エフェクト・サイズ

Effect size (ES)：一般には「Cohen の $d$」と呼ばれる。

$$\mathrm{ES} = \frac{\text{グループ別平均の差}}{\text{併合 SD}} \tag{1}$$

「併合 SD」の計算については教科書 p. 137 を参照。大雑把には、グループ別の SD の中間の値と考えてよい。

ES は、計算が簡単であり、直感的に把握しやすい。しかし、各グループの人数を考慮せず平均値だけ比較するため、グループの人数が大きくちがう場合でも、同じ人数に 2 等分されている場合でも、その間のちがいは ES の値に反映しない。また、2 グループ間の比較だけを行うものであるため、3 つ以上のグループを比較するのにはつかえない。

### 4.2　相関比 (correlation ratio)

- 各グループの個体が全員そのグループの平均値を持つ状況を仮定して SD を求める。
- この仮想 SD を実際の SD で割った数値が「相関比」である。数式では $\eta$ (eta) であらわす

### 4.3　SPSS コマンド

メニューの「分析」から「平均の比較」→「グループの平均」を開く。

- 「従属変数」に平均値を計算する変数を指定
- 「独立変数」にグループの変数を指定
- 「オプション」の「第 1 層の統計」で「分散分析表とイータ」をチェックする。

イータ ( $\eta$ ) は 0〜1 の範囲の値をとり、独立変数の影響力をあらわす

ES は SPSS では計算できない。

## 5　課題

(1) 適当な変数について、度数分布表・平均・標準偏差を出力（全体と男女別）
(2) (1) の変数について、性別による平均値の比較をおこなう。イータも出力すること。
(3) ES を (手計算で) 求める。
(4) 性別でわけて度数分布をグラフに表す (度数ポリゴンまたは折れ線グラフ)
(5) これらの分析結果から何が言えるか、解釈を書く。

現代日本学演習 II「統計分析の基礎」

# 第 9 講　分散分析

田中重人 (東北大学文学部准教授)

[**テーマ**]　分散分析 (ANOVA) の考えかたと計算方法を理解する

## 1　前回課題について

- 選択肢が 4 つ以上の項目を選ぶこと
- 正規分布に近似しているか確認すること (単峰性、偏り、集中度)
- グラフの縦軸の単位
- (ES の計算で) 割り算を間違えない

## 2　分散分析の考えかた

グループ別の平均値を当てはめて仮想の分散を求める分析法を「分散分析」(ANOVA：ANalysis Of VAriance) という。

- 従属変数 (dependent variable) と独立変数 (independent variable)

相関比 (イータ) の性質：

- 最小値:
- 最大値:

大きさの評価基準は、Cramer の連関係数 V と同様。
なぜ相関比を求めると、平均値を比較していることになるのか?

## 3　課題

次のデータ (10 人) について、分散分析を行なう

**男性:**　1, 2, 3, 3, 4
**女性:**　2, 3, 4, 4, 5

まず手計算 (または Excel) で考えてみて、そのあと、SPSS にデータを入力して検算する。

(1) 全体の平均値と SD を求める
(2) 男女別の平均値を求める
(3) 男性の平均値×5 人と女性の平均値×5 人からなる仮想データを考えて SD を求める
(4) (3) の SD を (1) の SD で割ったものが相関比 $\eta$

この相関比がなぜ「平均値の比較」の指標になるかを考えること。

## 4　相関比とエフェクトサイズの関係

相関比 $\eta$ とエフェクトサイズ ES の間にはつぎの関係がある ($n_1$, $n_2$ は各グループの度数、$N = n_1 + n_2$ は全体の度数)。

$$\mathrm{ES}^2 \quad = \quad \frac{\eta^2}{1-\eta^2} \times \frac{N^2}{n_1 n_2} \tag{1}$$

特に、2 グループの度数が等しい ($n_1 = n_2$) なら、この式は次のようになる。

$$\mathrm{ES}^2 \quad = \quad \frac{4\eta^2}{1-\eta^2} \tag{2}$$

(グループの度数が違えば、ES はこれより大きくなる)

さらに、$\eta$ があまり大きくない ( $\eta$ <0.4 程度) 場合であれば、次のような単純な式で近似できる：

$$\mathrm{ES} = 2\ \eta$$

## 5　モデルとデータの乖離

相関比 $\eta$ は、モデルとデータの乖離を表した値と解釈できる

- 「モデル」は何か？
- データとの乖離はどうやって計算しているか?
- 係数の取りうる値の範囲は?

## 6　表の書きかた

- 各層と全体の平均値と標準偏差 (測定水準の 2 桁下まで)
- 各層と全体の人数
- 相関比またはエフェクトサイズ (小数第 3 位まで)
- 欠損数とその原因

## 7　グラフの書きかた

平均値をプロットし、上下に SD を表示する。誤差範囲 (error bar; 別名「ヒゲ」) には SD 以外を書く場合もあるので、必ず「±標準偏差」であることを明記する。

Excel では

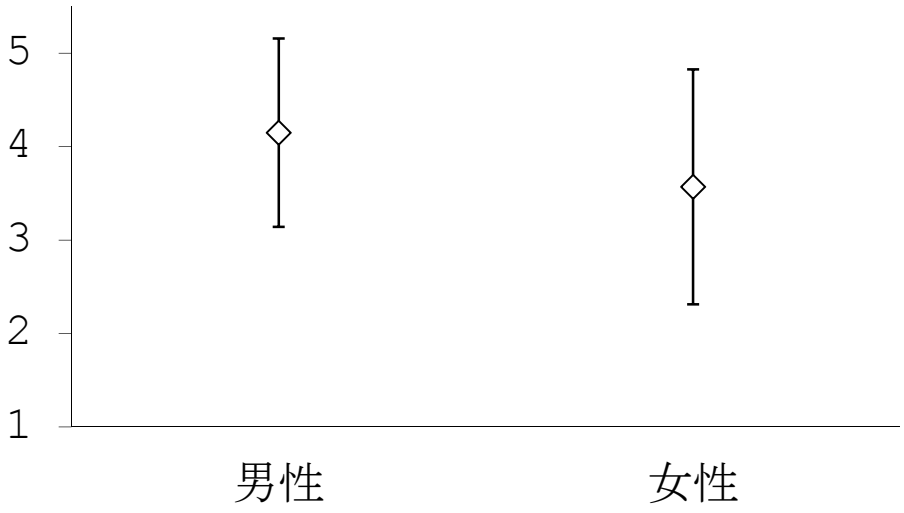
- SPSS 出力をシートにはりつける
- 折れ線グラフを描く
- メニューの「レイアウト」から「誤差範囲」→「その他の誤差範囲オプション」をえらぶ
- 「ユーザ設定」→「値の指定」
- 「正の誤差の値」「負の誤差の値」に SD が入っているセル範囲を指定 (おなじものでよい)

より詳細に分布の違いを検討したいときは、グループ別に度数ポリゴン (または折れ線グラフ) を描いてもよい。

**表 1 保守的意識の男女差**

| | 平均 | 標準偏差 | (人) |
|---|---|---|---|
| 男性 | 4.15 | 1.01 | (109) |
| 女性 | 3.57 | 1.26 | (130) |
| 合計 | 3.83 | 1.18 | (239) |

「以前からなされていたやり方を守ることが、最上の結果を生む」に対する回答:「1. そう思う」～「5. そう思わない」
相関比 η＝0.244. 無回答＝11.




「以前からなされていたやり方を守ることが、最上の結果を生む」
に対する回答:「1. そう思う」～「5. そう思わない」
相関比 η＝0.244. N=239. 無回答＝11.

**図 1 保守的意識の男女差 (平均±標準偏差)**

現代日本学演習 II「統計分析の基礎」

# 第 10 講　推測統計の基礎と区間推定

田中重人 (東北大学文学部准教授)

[テーマ]　推測統計の基礎

## 1　前回課題について

**男性:**　1, 2, 3, 3, 4 → 平均 2.6
**女性:**　2, 3, 4, 4, 5 → 平均 3.6
**全体の平均 (SD):**　3.1 (1.14)

グループ別平均値を当てはめた「仮想」データの平方和は、つぎのようになる。下線部に注意。

$$\text{グループ間平方和} \quad = \quad 5(\underline{2.6-3.1})^2+5(\underline{3.6-3.1})^2=2.5 \tag{1}$$

これを $N$ (=10) で割って平方根をとると標準偏差が得られる。

$$\text{仮想 SD} \quad = \quad \sqrt{\frac{2.5}{10}}=0.5 \tag{2}$$

$$\eta \quad = \quad \frac{\text{仮想 SD}}{\text{実際の SD}}=\frac{0.5}{1.14}=0.44 \tag{3}$$

ただし、SPSS では平方和を $N-1$ (=9) で割って「標準偏差」を求めているので、注意。度数がある程度大きくなれば (およそ $N>200$ の場合)、このことによる違いは気にしなくてよい。

分散分析の実際の計算では、平方和どうしで割り算して $\eta$ を求める ($N$ で割らずに済み、平方根を求めるのも一度で済むため)。SPSS 出力の「分散分析表」参照。

## 2　基礎知識

- 記述統計と推測統計 (教科書 pp. 3–5)
- 母集団と標本 (第 3 講資料)
- 無作為抽出と確率標本

## 3　統計的推測のふたつの方法

- 袋のなかに色つきの玉がたくさん入っている。ここから 8 個取り出したところ、すべて赤であった。
- 全世界から 8 人を無作為抽出して麺類の好みをきいたところ、全員が「うどんが好き」と答えた。

このような情報 (= 標本統計量) から、母集団における統計量 (= 母比率) を推測する

**区間推定:** ある統計量の母集団における値について、確率的な推測によって範囲を求める →母比率はたぶん ○○ から ×× の範囲にある

**統計的検定:** ある統計量の母集団における値について何らかの「帰無仮説」(null hypothesis) を設け、それが棄却できるかを判定する →母比率が 0.5 だと考えてよいか?

統計的検定のほうが計算が簡単であるため、よくつかわれている。区間推定を論文等で目にする機会はあまりないが、きちんと理解するにはまず区間推定の考え方をおさえるのがよい。

## 4 母比率の区間推定

### 4.1 区間推定の原理

(1) 「信頼率」を決めておく (たとえば 95%)
(2) (1 −信頼率) の確率を両極の事象に設定する (高いほう、低いほうからそれぞれ 2.5%ずつを除く)
(3) 母集団における値がいくつであれば、この両極端を除いた区間に測定値が入るかを計算する。統計量の性質に応じて、これを計算するための式と数表があるので、それを利用する。

このようにして求めた、母集団においてありうる値の集合が「信頼区間」である。通常、最初に決めた「信頼率」を明示して、「95%信頼区間」などのようにいう。

### 4.2 母比率の区間推定

標本の規模 $n$ がじゅうぶん大きく ($n$>30)、比率 $m$ があまり偏っていない (0.1<$m$<0.9) とき、母比率の 95%信頼区間は次の式で求められる：

$$m \pm 1.96\sqrt{\frac{m(1-m)}{n}} \tag{4}$$

### 4.3 課題

全世界から 400 人を無作為抽出してある意見に対する賛否を聞いたところ、「賛成」と答えた人が 240 人であった (欠損値はないものとする)。このとき、母集団 (全世界の人々) における賛成の比率の 95%信頼区間を求めよ。

## 5 宿題

教科書 pp. 156–162 を読み、統計的検定の手続きをまとめよ

## 6 期末レポート

**期限:** 8/14 (水) 17:00

**提出先:** ISTU 「期末レポート」にファイルを提出

**内容:** 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。

**備考:** レポート提出後に、データのコピーをすべて消去すること。

現代日本学演習 II「統計分析の基礎」

# 第 11 講　統計的検定

田中重人 (東北大学文学部准教授)

[テーマ]　平均値の区間推定と統計的検定の方法

## 1　前回課題について

### 1.1　課題 1: 母比率の区間推定

- 母比率の区間推定においては、95%信頼区間は、$n$=100 で± 10%、$n$=400 で± 5%程度
- 母集団の規模は関係ない (無限母集団の仮定)

### 1.2　課題 2: 統計的検定の手続き

- 背理法的思考 (「帰無仮説」とは)
- 「臨界値」はどうやって計算するか (→教科書末尾の数表)
- 「有意でない」ことの意味
- 区間推定との関係 (「有意水準」と「信頼率」「危険率」)

## 2　母平均の区間推定

間隔尺度以上の変数の場合には、「母集団においては正規分布している」という仮定を置けば、平均値の区間推定が可能。標本における平均 $m$ と標準偏差 SD から、母集団における平均 $M$ を推測する。

95%信頼区間は次のようになる：

$$m \pm \text{臨界値} \frac{\text{SD}}{\sqrt{n}} \tag{1}$$

臨界値は、$t$ 分布を使って求める (数表で調べる)。「自由度」($\text{df} = n - 1$) と危険率 (= 1 −信頼率) によって変化する。標本規模 200 以上で信頼率 95%なら、臨界値は 1.96 と考えてよい。

## 3　平均値の差の区間推定

ふたつのグループの間の平均値を比較するときは、平均値のグループ間の差についての信頼区間を直接求める方法をとる。標本における 2 グループ間の平均値の差を $d$ とすると、95 %信頼区間は

$$d \pm \text{臨界値} \times \text{併合 SD} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{2}$$

ただし $n_1, n_2$ はそれぞれのグループの人数。「臨界値」は自由度 $(n_1 + n_2 - 2)$ の $t$ 分布にしたがって求める。

## 4　SPSS コマンド

### 4.1　母平均の区間推定

「分析」→「記述統計」→「探索的」

- 「従属変数」を指定
- パネル左下の「統計」だけをチェック

信頼率を変更するには「統計」オプション。「因子」を指定すると、グループ別に分析できる。

### 4.2 平均値の差の区間推定

「平均の比較」→「独立したサンプルの $t$ 検定」

- 「グループ化変数」は、数値を指定しないといけない
- 連続量を一定の値で切ることもできる
- 出力は「独立サンプルの検定」の 1 行目「等分散を仮定する」を見る (この場合、「母集団で正規分布」「2 層間で SD が等しい」ということが前提になる)

## 5 統計的検定 (statistical test)

特定の値 $x$ (0 にすることが多い) を設定して、その値が信頼区間に含まれているかどうかを判定する。

### 5.1 統計的検定用語 (教科書 pp. 156–158, 165–166)

**帰無仮説 (null hypothesis):** 母集団における統計量が「特定の値」に等しい、という仮説
**有意 (significant):** 「特定の値」が信頼区間に**入っていない** ことをあらわす

### 5.2 平均値の差の検定の場合

「5%水準で有意」とは……

- 95%信頼区間が $x$ をふくまない
- すくなくとも 95 %の確率で、母集団において平均値の差があるといえる

「5%水準で非有意」とは……

- 95%信頼区間が $x$ をふくむ
- 母集団においては平均値の差はないかもしれない

### 5.3 有意確率とは

信頼区間の幅は、危険率 ($= 1 -$ 信頼率) を下げると広くなる。危険率を下げて信頼区間をひろげていくと、どこかで $x$ をふくむようになる。このときの危険率のことを「有意確率」または「p 値」という。
分析の際は、前もって危険率を設定しておき (通常は 5 %)、有意確率がその値を**下回っているかどうか** 判別する。

- 有意確率が 0.007 → 5%水準で有意
- 有意確率が 0.023 → 5%水準で有意
- 有意確率が 0.088 → 5%水準で非有意

## 6 区間推定と統計的検定

区間推定と統計的検定の間に本質的なちがいはない。ただし、区間推定は、統計量によっては、すごくむずかしい場合がある。統計的検定のほうが計算が簡単なので、統計的検定を使うことが多い (分野によってちがう)。

## 7 課題

適当な変数の平均の男女間の差について統計的検定を行い、結果にコメントをつけて提出

現代日本学演習 II「統計分析の基礎」

# 第12講　さまざまな検定手法

田中重人 (東北大学文学部准教授)

[テーマ]　相関比と連関係数の検定 (F 検定、カイ 2 乗検定)

## 1　前回宿題について

- 平均値を求めてよい変数かどうか、尺度水準について吟味すること
- 「等分散を仮定する」とは何か
- 有意でない場合の解釈

## 2　信頼区間と有意確率について補足

SPSS「独立したサンプルの $t$ 検定」では、「オプション」で信頼率を変更できる (「信頼区間のパーセント」)。適当な値に変更してみて、「有意確率 (両側)」との対応を確認してみよう。

「差の標準誤差」を 1.96 倍すると、95%信頼区間の幅の半分になる (ケース数が 200 以下の場合や、95%以外の信頼率の場合は、$t$ 分布表から求めた臨界値を使う)。

## 3　分散分析と F 検定

**帰無仮説:**　母集団においては $\eta = 0$

SPSS では「平均値の比較」→「グループの平均」を選択。オプション「分散分析表とイータ」を指定出力「分散分析表」の右端「有意確率」を見る。

2 グループの比較なら、平均値の差の $t$ 検定と同じ結果。

必要とする前提も $t$ 検定と同様 (母集団では正規分布しており、SD が全グループで等しい)。

## 4　クロス表の「独立性の検定」

**帰無仮説:**　母集団においては $V$=0

SPSS では、「クロス集計表」の「統計」で「カイ 2 乗」を指定。出力の「Pearson」の列の右端が有意確率 (各セルの期待度数が 5 以上であることを前提とする。この前提が満たされない場合は警告が出る)

2 × 2 クロス表では、$\chi$ 2 の値が大きめに出る (= 有意になりやすい) ため、種々の調整を要求されることがある。

## 5　課題

クロス表の「独立性の検定」と分散分析を、それぞれ適当な変数について行い、有意確率が 0.05 未満になるものを探す。その時の連関係数 $V$ と相関比 $\eta$ の値を確認すること。