

URL: <http://www.nik.sal.tohoku.ac.jp/~tsigeto/statg/>
作成: 田中重人 (講師) <tsigeto@nik.sal.tohoku.ac.jp>

比較現代日本論研究演習 I

大学院生対象: 2003 年度前期
<木2>コンピュータ実習室 (文学部本館 7F 711-2) 授業コード=LM14203

『講義概要』 p. 398 記載内容

- ◆講義題目: 統計分析入門
- ◆授業内容: 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、データの特徴を要約する記述統計の手法を中心に、統計分析の基礎を学びます。統計解析パッケージ SPSS を使ってデータ分析の実習を毎回おこないます。
◇実習室で使用できるコンピュータ台数が限られているため、受講人数の制限をおこなうことがある。
◇テキスト: 吉田寿夫、1998『本当にわかりやすいすぐ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: 各回の授業中の課題 (50%)、中間試験 (20%)、期末レポート (30%) を合計して評価する。

授業の概要 (予定)

目次

1. イントロダクション (4/10)
2. SPSS 入門 (4/17)
3. 統計分析の基礎 (4/24, 5/8)
4. 度数分布とクロス表 (5/15~6/5)
5. 中間試験 (6/12)
6. 平均値の比較 (6/19~7/17)

※ () 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。

1. イントロダクション

- この授業の概要・スケジュール・評価方法
- 部屋とコンピュータの使いかた
- SPSS の起動
- 他のソフトウェアについて (電卓, Excel, Word)
- 印刷

2. SPSS 入門・データ配布

- SPSS の概要
- データ行列 (データセット)
- 模擬データ入力実習
- データの配布
- SPSS コマンド・シンタックス
- メニューによるシンタックス作成
- 変数値の再割り当て

3. 統計分析の基礎

- 実験と観察
- データの記述
- データの種類
 - 名義・順序・間隔・比例
 - 順序尺度と間隔尺度の変換
 - 正規分布とは
- 標本抽出の 4 段階モデル
- サンプルングの概念と手順

- 新聞・雑誌・論文などにみられる調査の母集団・標本などについて各自報告 (5/8)

4. 度数分布とクロス表

4.1. 度数分布表

- frequencies コマンド
- 相対度数 (パーセンテージ)
- 棒グラフ
- ヒストグラム・度数ポリゴン
- Excel で整形, グラフ作成

4.2. クロス表

- 度数分布表のグループ化
- クロス表表記
- 行と列の%
- 周辺度数 (marginal distribution)
- crosstabs コマンドとそのオプション

4.3. 無関連状態と期待度数

- Φ 係数
- 期待度数・残差・連関係数
- クロス表の書きかた

5. 中間試験

6. 平均値の比較

6.1. 平均と分散

- データの種類: 復習
- 平均値
- 分散と標準偏差
- 分布と外れ値
- ノンパラメトリックな代表値 (中央値と四分位偏差)

6.2. 平均値の層別比較

- 平均の差と差の平均
- エフェクト・サイズ
- 相関比から分散分析へ
- 公表に際してなにを書くべきか

2003.4.10

比較現代日本論研究演習 I (田中重人) 受講登録フォーム

氏名：

学年：

学籍番号：

所属（文学部日本語教育以外の場合）：

興味のあること（非学術的な話題も可）：

・自宅でパソコンを使えますか？ **ある / ない**

・SPSS 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（ ）

・確率・統計または類似の授業を受けたことがありますか？

ある / ない

カードをとって
適当なところに着席

電源はまだ入れない

0

比較現代日本論研究演習 I

統計分析入門

東北大学大学院文学研究科 2003 年度
田中 重人 (講師)

1

【目的】

統計分析の基礎的な手法の習得

- SPSS の操作 (4 月)
- クロス表分析 (5-6 月)
- 平均値の比較 (6-7 月)

2

★ 推測統計は後期の
「研究演習 II」であつかう。

→ 修士論文等で統計的手法を使う予定の場合
は連続履修することがのぞましい

3

【教科書】

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。

※生協文系書籍部に入荷済み

4

受講登録フォーム記入

【コンピュータ実習室について】

- ★ 入室に**学生証**が必要
(ない場合は一時利用カードを教務掛で借りる)
- ★ 土足・飲食・喫煙 **厳禁**
- ★ 退出時は必要事項を紙に書く
(書けるところを書いてみよう)
- ★ ドアが開かなくなったときは電話で連絡

6

【コンピュータの起動と終了】

- ・ 本体とディスプレイの電源を ON
- ・ 表示されるお知らせの内容をよく読む
- ・ シャットダウンしたら、
ディスプレイの電源を切る

7

【ファイルの保存場所】

授業でつかうファイルは、
授業開始時に My Document
フォルダにコピーして使う。
授業終了時に削除してかえること。

★ 内蔵 Disk にデータは置けない

8

必要なデータは各自でフロッピー
にコピーして持ち帰る

→ フロッピーディスクを
各自で購入しておくこと。

9

【SPSS】

データ解析用ソフトウェア

- ★ Windows での開発に
特に力を入れている
- ★ 購入しやすい

10

【この授業で使用するデータ】
1995 年 SSM 調査 B 票の一部

cf. 『日本の階層システム』(全 6 巻)
東京大学出版会、2000 年。

11

【その他のソフトウェア】

- 表計算 (Excel)
- 電卓 (「アクセサリ」のなか)
- ワードプロ (Word)

12

【印刷】

モノクロプリンタ 2 台

- ★ 電源の入れかた
- ★ 出力先の切り替えかた
- ★ ジョブの確認・取り消し
- ★ 印刷前にプレビューで確認
- ★ タイル印刷 (2 面, 4 面, ...) の方法

13

第2回「SPSS入門」目次

1. 模擬データ入力実習
2. データの配布
3. SPSSのウィンドウ構成
4. メニューとシンタック
5. 変数値の再割り当て
6. 出力の読みかた・印刷

1

【模擬データ入力実習】

- まず変数を定義
 - ・ 「データエディタ」ウィンドウの下の「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
 - ・ 「データビュー」タブに切り替え、確認

2

- データを入力
- My Document 内に保存
- 「エクスプローラ」で確認

※ このファイルは授業終了時に削除 (フロッピーにコピーする必要はない)

3

【データの配布】

1995年SSM調査B票の一部
 ★ 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

cf. 『日本の階層システム』(全6巻)
 東京大学出版会、2000年。

4

- ★ 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
- ★ 未公開のデータなので流出しないように
- ★ 変数ラベルは菅野剛 (日本大学) 氏による

5

- ★ 毎回の授業で使うので、忘れないこと
- ★ レポート提出時に返却

6

【データ・セット】

- ★ ケース × 変数
- ★ 変数は変数名で管理
- ★ 変数名以外に「ラベル」
- ★ 無回答などの欠損値 (.)

7

【SPSSのウィンドウ構成】

- データ・エディタ
- シンタックス・エディタ
- 出力ビューア

8

【メニューとシンタックス】

- ★ 分析手法をえらぶ
- ★ 必要なオプションを指定
- ★ 「貼り付け」をクリック
- ★ シンタックスの必要部分を選択して実行 (▶)

9

【変数値の再割り当て】

- データエディタのメニューバーで
- 「変換」→「値の再割り当て」→「他の変数へ」
 - 変換先変数の名前をつける

10

- 「今までの値と新しい値」
- 値の組を指定したら「続行」
- シンタックスを貼付けて実行
- 新変数の度数分布を確認
- 問題がなければデータセットを保存する

11

【出力ビューア】

- ★ 左側に目次、右側に出力内容
- ★ エラー表示もここに出る

【印刷】

- ★ 左側の目次で選択
- ★ 印刷前にプレビュー
- ★ タイル印刷 (2面, 4面, ...)

12

【データの保存】

- シンタックス・エディタ
 - 出力ビューア
 - データ・エディタ
- (上書きしないこと: 変更がなければ保存不要)

13

シンタックスはテキストファイル
 → メモ帳などのエディタで読める

他のファイルは、SPSS でのみ読み書き可能

14

【実習】

本人年収 (Q44_1)を5~7程度の
 適当な間隔に区切って度数分布表
 を出力し、印刷して提出

15

1. データ収集から分析まで
2. 標本抽出

1

【データ収集から分析まで】

- データの収集 (実験／観察)
- データの特徴を少数の数値に要約して記述 = **記述統計**
- 誤差の評価
(この手続きの一部が**推測統計**)

(教科書 p. 1-6)

2

【標本抽出の4段階モデル】

- ユニバース (universe)
- 母集団 (population)
- 計画標本 (designed sample)
- 有効標本 (valid sample / case)

3

★ 伝統的な統計学では4段階にわけずに、2段階で考えるのがふつう：

母集団 = Universe + population

標本 = (designed/valid) sample

4

【無作為抽出】

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 (random sampling)

➡ 「**等確率標本**」

5

つぎの条件が必要：

- ★ 母集団の人口が既知
- ★ 個体を網羅した「台帳」

※ 個体によって抽出確率が違う場合も、事後的に調整して等確率標本と同様の統計処理をおこなうことは可能

※ 「台帳」が完備してない状況でも、工夫次第で無作為抽出に近づけることができる

6

統計的な推測は、**等確率標本を前提とする**

実際の調査で理想的な標本抽出ができることはまずない。

また計画標本のなかから無効回答があるので、

無作為ではない誤差がかならず発生する。

この誤差は**統計的には処理できない**ので、個別に推測する

- ・ どの層を過剰に代表しているかを把握する
- ・ おなじ母集団を対象にした調査と比較する

7

【宿題】

論文や新聞・雑誌記事で使われている調査データについて、標本抽出の4段階にそって紹介する。

人数分コピーを用意してきて、次回授業時に報告。

8

1. 変数の種類
2. 尺度の変換
3. 度数分布表
4. 棒グラフとヒストグラム

1

【変数の種類】

- 比率尺度 (ratio scale)
- 間隔尺度 (interval —)
- 順序尺度 (ordinal —)
- 名義尺度 (nominal —)
(質的変数とも)

(教科書 p. 8)

2

【尺度の変換】

- ★ 上位の尺度のほうがあつかえる演算が豊富
- ★ 上位の尺度は下位の尺度の特徴を兼ね備えている

→分析手法の選択幅がひろい

3

私たちが測定するものはたいてい
順序尺度以下である

- ★ 上位の尺度への変換には
一定の理論的根拠が必要

4

【度数分布表】

Frequencies コマンドを使う

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

(教科書 p. 27-31)

5

【累積%とパーセンタイル】

- 順序尺度以上の場合のみ意味を持つ
- percentile (= %点)
- 中央値 (median) = 50%点
- 「割り切れてしまう」場合は中点をとる
(教科書 p. 43)
- 同じ値が並ぶ場合は多少の操作が必要
(森敏昭・吉田寿夫(編)(1990)『心理学のための
データ解析テクニカルブック』北大路書房. p. 15)

6

【棒グラフとヒストグラム】

- 棒グラフ……棒同士の間を空白をあける。高さ(長さ)をよむ。
- histogram (柱グラフ)……柱の間隔をあけない。面積をよむ。

※縦軸は度数または%

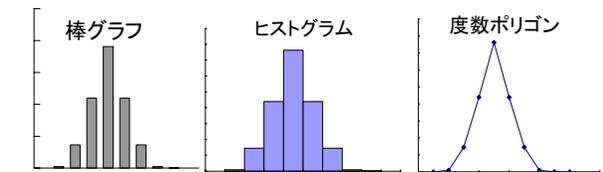
7

- ★ 連続量を階級分けした場合
→ ヒストグラム
- ★ それ以外の場合 (離散量/
名義尺度) → 棒グラフ

※度数多角形 (polygon) は複数の変数の分布を比較するとき便利。

(教科書 p. 32-36)

8



SPSS では histogram が書きにくい。
★ recode で整形した上で度数分布表のメニューで「図表…」指定。棒グラフを書く
★ グラフ→インタラクティブ→ヒストグラム
では等間隔の区間に分割してくれる

9

【実習】

- (1) 本人年齢の度数分布表を出力し、中央値と上側 20% 点に印をつけよ
- (2) 適当な変数について棒グラフまたはヒストグラムを作成

【キーワード】

行 (row) 列 (column) セル (cell)

周辺度数 (marginal frequency)

行% (row percent) 列% (column percent)

1

【度数分布表の比較】

- データエディタのメニューで
「データ」→「ファイルの分割」
→「グループの比較」

- 度数分布表を出力

2

- 「データ」→「ファイルの分割」
→「すべてのケースを分析」
でもとにもどしておく

3

【クロス表の基本型】

質的変数 (名義尺度) 同士の関連
についての基本的な分析法

		β			
α		1	2	3	合計
行	1	a	b	c	a+b+c
	2	d	e	f	d+e+f
	3	g	h	i	g+h+i
合計		a+d+g	b+e+h	c+f+i	N
		列			周辺度数

4

5

【Crosstabs コマンド】

性別 × 「性別による不公平」
のクロス表を書いてみよう

「分析」 → 「記述統計」 → 「クロス集計表」

6

【行%と列%】

「クロス集計表」メニューで「セル」にパー
センテージ (行・列) を追加

- ★ 行%, 列%のつかいわけは
説明→被説明の関係に対応
行→列の説明をすることが多い
- ★ 周辺度数の%とも比較する

7

【グラフを書いてみる】

- ★ クロス表は積み上げ棒グラフ
で表現することが多い
SPSS ではうまくかけない。コピーして
Excel に貼付けてグラフを書くのがよい
- ★ 度数にも注意

8

【課題】

性別 × 適当な変数でクロス表作成、
グラフも書いて印刷して提出

9

第 8 回 「φ 係数」

1. 自由度 (degree of freedom)
2. クロス表分析のふたつの系列
3. 2×2 クロス表の性質
4. φ 係数 (phi coefficient)

1

【自由度】

2×2 クロス表では、周辺度数が所与なら、1つのセル度数が決まればほかも決まる

α	β		合計
	1	2	
1	a	g-a	g
2	i-a	h-i+a	h
合計	i	j	N

2

3×3 クロス表：セル度数が 4 つ決まれば…

α	β			合計
	1	2	3	
1				f
2				g
3				h
合計	i	j	m	N

k×l クロス表の自由度 (degree of freedom)

$$d.f. = (k-1)(l-1)$$

3

【クロス表分析の 2 つの系列】

- 「%の差」系 (期待度数との差) = 連関係数
- オッズ比系 (乗法モデル) = 対数線形分析、ロジット分析

この授業で取り上げるのは前者だけ

4

【2×2 クロス表の性質】

以下、つぎの記号法を使う

α	β		合計
	1	2	
1	a	c	g
2	b	d	h
合計	i	j	N

5

(1) 行%は 1 列について比較すればよい：

$$\frac{a}{g} - \frac{b}{h} = \frac{d}{h} - \frac{c}{g}$$

(2) 行%の差がゼロなら列%の差もゼロ

(3) g=i なら行%の差と列%の差は同じ：

$$\frac{a}{g} - \frac{b}{h} = \frac{a}{i} - \frac{c}{j}$$

6

(例 1) 行%の差 = 8%

60%	40%	100%
52%	48%	100%

(例 2) 行・列とも%に差なし

52	48	100
52.0%	48.0%	100.0%
66.7%	66.7%	
26	24	50
52.0%	48.0%	100.0%
33.3%	33.3%	
78	72	150
52.0%	48.0%	100.0%

(例 3) 行・列とも 10%の差

70	30	100
70.0%	30.0%	100.0%
70.0%	60.0%	
30	20	50
60.0%	40.0%	100.0%
30.0%	40.0%	
100	50	150
52.0%	48.0%	100.0%

7

【φ 係数】

2×2 クロス表の「連関」の尺度

$$\phi = \frac{ad - bc}{\sqrt{ghij}}$$

この係数の意味は？

(分子だけ取り出して考えてみよう)

8

【SPSS での φ 係数の計算】

「クロス集計表」の

「統計」で

「ファイとクラマーの V」をチェック

9

【キーワード】

連関 (association), 独立 (independence),
期待度数 (expected frequency),
クラメールの連関係数 (Cramer's V)

1

【 ϕ 係数の性質】

1. $\phi = \text{交差積の差} / \sqrt{\text{周辺度数の積}}$
2. $\phi = \text{相関係数の特殊ケース}$
3. $|\phi| = \text{行\%差と列\%差の中間の値}$
4. $\phi^2 = \text{標準残差の総計} / N$
($\rightarrow 2 \times 2$ 以上のクロス表に拡張できる)

2

【期待度数と ϕ 係数】

※記号法は前回と同じ

独立 (無関連) : $a/b = c/d$

期待度数 (expected frequency)

周辺度数を固定しておいて独立なクロス表
を作ったとき、各セルに入る度数 :

$$\frac{gi/N}{hi/N} \quad \frac{gj/N}{hj/N}$$

3

独立なクロス表の例

52	48	100
52.0%	48.0%	100.0%
66.7%	66.7%	
26	24	50
52.0%	48.0%	100.0%
33.3%	33.3%	
78	72	150
52.0%	48.0%	100.0%

4

- ★ 期待度数はたいてい小数になる
- ★ 期待度数について行%と列%を計算すると、周辺度数の%とおなじになる

観測度数 各セルに入る実際の度数
残差 (residual) 観測度数と期待度数の差
標準残差 (standardized ---) 残差/ $\sqrt{\text{期待度数}}$

$$\text{ex. } A = \frac{a - gi/N}{\sqrt{gi/N}}$$

5

χ^2 (chi-square) 標準残差の平方和
各セルに入る標準残差を A, B, C, D とする

$$\chi^2 = A^2 + B^2 + C^2 + D^2 = N \left(\frac{a^2}{gi} + \frac{b^2}{hi} + \frac{c^2}{gj} + \frac{d^2}{hj} - 1 \right)$$

χ^2 を人数で割った値が **ϕ の2乗** に等しい

$$\phi^2 = \frac{\chi^2}{N} \quad \text{すなわち} \quad |\phi| = \sqrt{\frac{\chi^2}{N}}$$

6

【クラメールの連関係数 V 】 $k \times l$ 表への ϕ 係数の拡張 (教科書 p. 114-117)

- ★ k と l のうち小さいほうを m とする
- ★ 2×2 表と同様に期待度数・残差を求める
- ★ χ^2 を求める
- ★ χ^2 を N と $(m-1)$ で割って平方根をとる

$$V = \sqrt{\frac{\chi^2}{N(m-1)}}$$

7

【 V の性質】

- ★ 行・列変数が独立のとき $V = 0$
- ★ 関連が強くなると大きくなる
- ★ 最大値は 1

8

【SPSS で実習】

クロス表のオプションを指定 :

- 「セル」… 度数(観測/期待)
残差(標準化なし/標準化)
- 「統計」… カイ 2 乗
ファイとクラメールの V

9

中間試験

2003.6.19

【回答上の注意】

- ① 他の回答者の画面が見えないよう、互いに離れて座ること
- ② コンピュータで回答を書き、印刷して提出
- ③ Word が使えることを確認
- ④ 小数の回答については、小数第 1 位まで書くこと
- ⑤ 何を持ち込んで参照してもよいが、人に相談してはならない

問 1 年齢が 50 歳以下のグループと 51 歳以上の 2 グループにデータセットを分割して分析したい。SPSS でこの操作をするときに必要なシンタックスを書け。ただし年齢の変数名は q1_2a である。

問 2 次の対概念について、それぞれどういう違いがあるかを簡単に説明せよ。

- (1) 「実験」と「観察」
- (2) 「記述統計」と「推測統計」
- (3) 「順序尺度」と「間隔尺度」
- (4) 「母集団」と「計画標本」

問 3 男性 250 人、女性 300 人を対象にしたある調査結果によると、よくお酒を呑む者の率は男性では 76.0%、女性では 65.7%であった (欠損値はないものとする)。この結果に基づいて、次のようなクロス表を作成せよ (ただし%のところには行%を書くこと)。

	よく呑む	呑まない	合計
男性	人数	人数	人数
	(%)	(%)	(%)
	期待値	期待値	
	残差	残差	
女性	人数	人数	人数
	(%)	(%)	(%)
	期待値	期待値	
	残差	残差	
合計	人数	人数	人数
	(%)	(%)	(%)

中間試験 解答例

2003.6.19

問 1

```
RECODE
  q1_2a
  (Lowest thru 50=1) (51 thru Highest=2) INTO age2 . ← 新変数名はなんでもよい
EXECUTE .
```

```
SORT CASES BY age2 .
SPLIT FILE
  LAYERED BY age2 .
```

問 2 次の対概念について、それぞれどういう違いがあるかを簡単に説明せよ。

- (1) 条件を人工的に統制するのが「実験」、しないのが「観察」
- (2) データの特性を要約して示すのが「記述統計」。
データに含まれる誤差を推測するのが「推測統計」。
- (3) 「順序尺度」の値の配列には一定の順序があるが、値の和や差に意味はない。
「間隔尺度」の場合、値の和や差をとることに意味がある。
- (4) 母集団：検討しようとしている対象の集団全体で、その範囲が確定しているもの
計画標本：母集団から何らかの方法で抜き出した実際の調査対象

問 3

	よく呑む	呑まない	合計
男性	190	60	250
	(76.0)	(24.0)	(100.0)
	175.9	74.1	
	14.1	-14.1	
女性	197	103	300
	(65.7)	(34.3)	(100.0)
	211.1	88.9	
	-14.1	14.1	
合計	387	163	550
	(70.4)	(29.6)	(100.0)

1. 他人に見せる表
2. 表と図のあつかい
3. 表の書きかた

1

【他人に見せる表】

- 資料としての表…データを詳細に再現したものがよい
- プレゼンテーション用の表…わかりやすく情報を圧縮する
→どう圧縮するかがセンスの見せどころ

2

【他人に見せられない表】

- ★ セル数が多すぎて周辺度数が偏っているもの
期待度数が5未満のセルがあると、
V係数は無意味
→適切なカテゴリー統合を行う必要

※資料としての意味はまた別である

3

- ★ カテゴリーの並べ順や行列のくみあわせをわかりやすく
- ★ 変数とカテゴリーの命名
- ★ 表のタイトル

4

【表と図】

表 (table) …活字と罫線で行列型に組む。

図 (figure) …活字・罫線以外の要素を含む。グラフのほか、概念図や写真を使うことも

5

【表と図の約束ごと】

- ★ 「表1」「図1」のようにそれぞれ通し番号をつけて参照
- ★ 表のタイトルは上、図のタイトルは下
- ★ 「それだけでわかる」ように

6

【表に書くべき要素】

- 各セルの行(列)%
- 行(列)合計の度数と「100.0%」
- 列(行)合計の%
- 全体の度数
- Cramer の V (または ϕ)
- 欠損数とその原因

7

- ★ 行→列の因果を想定するのがふつうだが、列→行でもよい。(％の「100.0」で区別)
- ★ 全度数が1000人以下であれば、％は小数第1位まで
- ★ V や ϕ などの係数は小数第3位まで
- ★ 2列表の場合は1列の％だけ示してもよい
- ★ 統計的検定をした場合は、その結果も

8

- ★ 縦罫線はなるべく引かない
- ★ 文字列は左揃え、数字は小数点揃えが基本
- ★ タイトル、表本体、注釈を読めばそれだけでわかるように書く
→タイトルと行・列頭の見出し (heading) を工夫する

9

授業資料

表1 性別と性別による不公平感との関連

性別	性別による不公平			合計	(人)
	「大いにある」	「少しはある」	「ない」		
男性	36.0	50.5	13.5	100.0	(111)
女性	27.3	56.8	15.9	100.0	(132)
合計	31.3	53.9	14.8	100.0	(243)

Cramer's $V=0.094$ 。無回答=7。

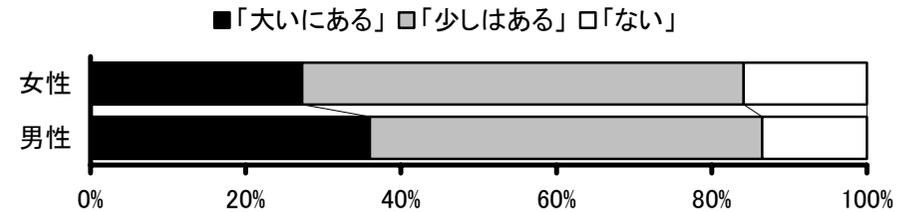


図1 性別と性別による不公平感との関連

表2 県や市町村の部課長以上の役人に知り合いがいる比率の男女差

性別	%	(人)
男性	46.0	(113)
女性	27.6	(134)
合計	36.0	(247)

$\phi=0.191$ 。無回答=3。

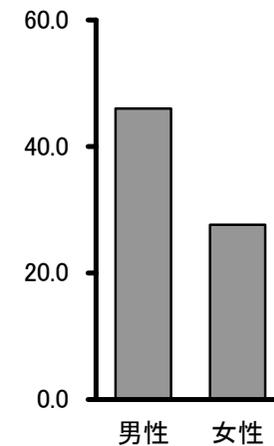


図2 県や市町村の部課長以上の役人に知り合いがいる%の男女差

第10回「平均値と標準偏差」

1. 尺度水準と分析法
2. 代表値と散布度
3. 平均値と標準偏差
4. SPSS のコマンド
5. 平均値を使うときの注意事項

1

【尺度水準と分析法】

名義×名義 → クロス表

名義×間隔 → 平均値の比較

2

【代表値と散布度】

★ 平均値 (mean) — 標準偏差 (SD)
(間隔尺度以上)

★ 中央値 (median) — 四分位偏差 (Q)
(順序尺度以上)

(教科書 p. 42-51)

3

【平均値】

総和をデータ数で割ったもの

【標準偏差】

平均値からの偏差の2乗値の平均が「分散」
分散の平方根が「標準偏差」

★ 平均値と標準偏差はセットで使う

4

★ 次のデータの平均と SD は?

{0, 1, 4, 5, 7}

5

【SPSS のコマンド】

「記述統計」 → 「度数分布表」

→ 「統計」 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」でもよい

6

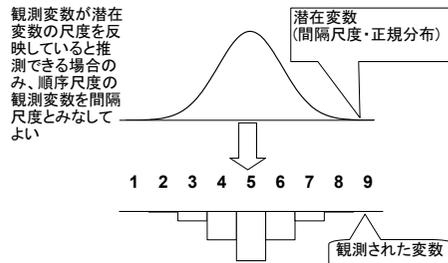
【平均値を使うときの注意事項】

- ★ 平均値ははずれ値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%を取りのぞいたデータセットで
計算する (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

7

- ★ 平均値・標準偏差は間隔尺度以上のデータ
に対してしか意味をもたない。
順序尺度の平均値をとっていいのは
 - ・ 潜在的には間隔尺度のはず
 - ・ 測定のポイントが一定間隔
- という2条件をともに満たす場合
※ 2値の変数は間隔尺度とみなせるが、若干の注意が必要。

8



9

- 具体的には
- 4点以上の尺度
 - 正規分布に近似 (教科書 p. 53-59):
 - ・ 単峰性
 - ・ 左右対称性 (歪度)
 - ・ 中央への集中度 (尖度)
- ヒストグラムを描いて検討するとよい。
- 正規分布との乖離度を統計的に検討する手法もある

10

歪度・尖度は「度数分布表」の「統計」オプションで指定できる

正規分布のとき0、
絶対値が大きくなるほど、正規分布から外れる

11

- 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)
 - 順位に変換したり中央値を使って分析

12

※ 間隔尺度のデータでも、
左右対称でないものについては
平均値よりも中央値のほうが
適当であることが多い

典型例: 収入・人口など

13

【課題】

適当な変数について、度数分布表を出力し、
そこに平均と標準偏差を書き入れて提出

14

【期末レポート】

- 期限: 8/5 (火) 17:00
- 提出先: 田中研究室 (文法合同棟 2F)。
田中が不在のときは 205 室のレターケースへ
- 内容: クロス表・平均値の比較の両方を使い、適当な分析をして結果を解釈する。
- 備考: 後期の授業「比較現代日本論研究演習 II」を受講しない者は、SSM データのディスクをレポートと一緒に提出。
データのコピーをすべて消去すること。

15

1. 平均値の層別比較
2. SPSS のコマンド
3. エフェクト・サイズ
4. 分散分析と相関比

1

【平均値の層別比較】

ふたつの層の間の平均値の比較

- ★ 平均値の差をもとめる
(層別平均)
- ★ 標準偏差を基準にして差を評価
(effect size; 相関比)

2

【SPSS のコマンド】

「平均の比較」→「グループの平均」

- 従属変数=平均値を求める変数
(間隔尺度)
- 独立変数=層を指定する変数
(名義尺度)

3

【エフェクト・サイズ】

$$ES = \text{平均値の差} / \text{標準偏差}$$

- ★ 正式には層別 SD の重みつき平均のような数値 (併合 SD) をつかう (教科書 p. 137)

4

【例】

性別による生活全般満足度の違い

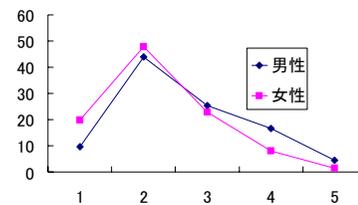
	平均	SD	(人数)
男性	2.62	1.02	(114)
女性	2.24	0.91	(136)
合計	2.41	0.98	(250)

平均の差=0.39 併合 SD=0.97
ES=0.401

※ ES は SPSS では計算してくれない

5

性別による生活満足度の違い



6

【ES の特徴と問題点】

- ★ 各層の人数を考慮せず平均値だけ比較
→ 大きさがちがう場合は？
- ★ 2 層間の比較だけ
→ 3 つ以上の層を比較したい場合は？

7

【相関比】

- ★ 各層の個体が全員その層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」。η (イータ) であらわす
- ★ 相関比の 2 乗 η² を「決定係数」「分散説明率」などという
※ η² を「相関比」ということもある

8

- ★ SPSS では「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
- ★ η 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 は最小値 0、最大値 ∞

9

- ★ 3 層以上で平均値を比べる場合にも相関比が使える。
- ★ このように、層別平均値をあてはめて仮想分散を求める分析法を「分散分析」(ANOVA: ANalysis Of VAriance) という。

10

【注意事項】

層別の平均値を分析する場合、各層の人数は一定以上必要 (最低 20 人?)

→ カテゴリ統合が必要になることがある

11

【ES と η の関係】

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

特に、2 層の大きさが同じ (n₁ = n₂) なら、

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

層の大きさがちがえば、ES はこれより大きくなる

12

※ このように ES と η は互いに変換できる。

→ 両方示すのは冗長

13

【ダミー変数】

2 値の変数に (0, 1) の値を割り当ててつかう場合、「ダミー変数」(dummy variable) という。

- ★ ダミー変数の平均値は「値が 1 をとる人の比率」をあらわす
- ★ ダミー変数についての相関比 η はクラメル の 連 関 係 数 V に 等 しい

14

【課題】

適当な変数の平均値について、男女別の平均値の差と ES, η を求める。
表に ES と η を書き込んで提出。

15