URL: http://www.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：2008 年度前期
＜木２＞コンピュータ実習室（文学部本館 7F 711-2）

『講義概要』p. 402 記載内容

◆ 講義題目：統計分析入門
◆ 到達目標：(1) 統計分析の基礎を理解する；(2) 実際にデータ分析をできるようになる
◆ 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を使ってデータ分析の実習を毎回おこないます。
◇ 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：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。
◇ その他：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、受講人数を制限することがある。

授業の概要

目次

1. イントロダクション（4/10）
2. 統計分析の基礎（4/17）
3. SPSS 入門・データ配布（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）
7. 予備日（7/24）
8. 期末レポート（8 月中旬提出）

※ () 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。

修士論文等で質問紙調査を予定している者は、

- 1 学期開講の比較現代日本論研究演習 I「質問紙法調査の理論と実践」（火 5：鈴木教授）
- 2 学期開講の比較現代日本論研究演習 II「実践的統計分析法」（火 4：田中講師）

も受講することがのぞましい。

1. イントロダクション

- 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

2. 統計分析の基礎

- 記述統計と推測統計
- SPSS の起動
- データ行列（データセット）とは
- 模擬データ入力実習
- その他のソフトウエアについて

3. SPSS 入門・データ配布

- データの配布と説明
- データの種類
- 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
- 標本調査とは

4. 記述統計 (1)：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
- クロス表とグラフの書きかた

5. 中間試験

6. 記述統計 (2)：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
- 層別平均
- 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

比較現代日本論研究演習 I
「統計分析入門」

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

2008.4.10

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 2008 年度
田中 重人 (講師)

【目的】

統計分析の基礎的な手法の習得

- SPSS の操作
- クロス表分析
- 平均値の比較

【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。

【成績評価】

- 授業中の課題 (50%)
- 中間試験 (20%)
- 期末レポート (30%)

【関連する授業】

1 学期
・比較現代日本論研究演習 I
「質問紙調査の理論と実践」(火 5)
2 学期
・比較現代日本論研究演習 II
「実践的統計分析法」(火 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 (ない人は教務掛で)
★ 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)
★ ドアの開けかた

【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
- 「NumLock」ランプ点灯を確認
- 終了するときは、ディスプレイの電源を切ることをわすれないように

【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 マイドキュメント
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

必要なデータは各自で
フロッピーかスティックメモリ
にコピーして持ち帰る

→ 各自で購入しておくこと。

【SPSS】

データ解析用ソフトウエア
★ Windows での開発に
特に力を入れている
★ 購入しやすい

【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。

SSM 調査については http://www.sal.tohoku.ac.jp/coe/ssm/ 参照

URL: http://www.nik.sal.tohoku.ac.jp/~tsigeto/statg/g080410.html
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 統計分析入門］

第1回（2008-04-10）

受講者の興味と数学的知識の調査

→別紙

コンピュータ実習室について

入室・退室

学生証が必要（ない人は、教務係で臨時カードを借りること）。

土足・飲食・喫煙厳禁。

退出時には必要事項を紙に記入。

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
- ディスプレイの電源を入れる（2-3秒押しつづけないと入らないので注意）
- 表示されるお知らせをひととおりよむこと
- キーボード右上の「NumLock」ランプがついているか確認

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「終了オプション」→「電源を切る」を選択
- コンピュータ本体の電源が切れたことを確認
- ディスプレイの電源を切る
- フロッピーディスク、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は、個人のファイルを置いてはならない。 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フロッピーかスティック・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。

コンピュータ実習室で使えるリムーバブルメディアはつぎのふたつ。 各自どちらかを購入しておくこと。

- フロッピーディスク（3.5インチ）……「Windows フォーマット」のものが便利。安いがよく故障する。容量が小さい。
- フラッシュメモリ ……「USB2.0対応」のもの。値段は高いが容量が大きい。とりはずすときは画面右下の「ハードウエアの安全な取り外し」アイコンをクリックして、「USB大容量記憶装置」を停止させてから、メモリ本体を引き抜く。

模擬データ入力実習

SPSS について

参考書：宮脇典彦・和田悟・阪井和男（2000）『SPSS によるデータ解析の基礎』培風館。

SPSS の起動

スタートメニューから「プログラム」→「SPSS for Windows」→「SPSS for Windows 12.0J」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。）

「どのような作業を行いますか?」ときかれたら「データに入力」をチェックして「OK」。

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6個なので、 3×6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。（ほかのフォルダには保存できません。）

「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ので、テキストファイルでデータを用意しておくのがふつうである。

2008.4.10

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：

学生番号：

所属（文学研究科日本語教育学専攻以外の場合）：

研究内容：

・自宅でパソコンを使えますか? **ある / ない**

・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（ ）

数学的予備知識の調査 （成績評価には関係ありません）

(1) 「乱数」とは何か。簡単に説明せよ。

(2) 「必要十分条件」とは何か。簡単に説明せよ。

(3) 「偏差値」はどういう目的のために使われるか。またどうやって求めるか。簡単に説明せよ

(4) つぎの数式の値を求めよ。計算のプロセスがわかるように解答すること

$$\sum_{k=1}^{10} k =$$

2008.4.17 比較現代日本論研究演習 I (田中重人)

第 2 回「SPSS 入門」目次

1. データの配布
2. SPSS のウインドウ構成
3. 変数値の再割り当て
4. 出力の読みかた・印刷
5. その他のプログラム

【データ回収から入力まで】

★ ID 付与とエディティング
★ コード表とコーディング
★ ファイル作成
★ クリーニング

【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. (2000)『日本の階層システム』(全 6 巻) 東京大学出版会。

★ 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
★ 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

【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 (.)

【SPSS のウインドウ構成】

● データ・エディタ
● シンタックス・エディタ
● 出力ビューア

【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 (▶)

【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

● 「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

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

【その他のアプリケーション】

● 文書作成 (Word)
● 表計算 (Excel)
● 電卓 (アクセサリ)

【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

教科書の序章・第 1 章を読んでおくこと

2008.4.24 比較現代日本論研究演習 I (田中重人)

第 3 回「統計分析の基礎」目次

1. データ収集から分析まで
2. 標本抽出
3. 尺度水準
4. 度数分布表

【データ収集から分析まで】

- データの収集 (実験／観察)
- 分析可能な形に加工
  - ・分析の単位
  - ・変数の同定
  - ・尺度水準
- データ・セット作成

- データの特徴を少数の数値に要約
  ＝ 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)

【標本抽出の 4 段階モデル】

目標母集団 (universe)

調査母集団 (population)

計画標本　(designed sample)

有効標本　(valid sample / case)

★ 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

【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**」

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

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でるので、
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

【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。

(1) その論文等の書誌情報と、目標母集団・調査母集団・
　計画標本・有効標本について簡単にまとめたもの

(2) その論文等のなかで、データについての説明の部分

人数分コピーを用意してきて、次回授業時に報告。

【層化 2 段無作為抽出】

・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)

【尺度水準】

●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)

【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

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

【実習】

SSM 調査の調査票中で、比率尺度とみなせるものはどれか

【度数分布表】

Frequencies コマンドを使う

★ 度数

★ 相対度数 （%）

★ 累積度数・累積相対度数

★ 欠損値のあつかい

(教科書 p. 27–31)

【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)

【変数値の再割り当て】

前回資料を参照

2008.5.15 比較現代日本論研究演習 I (田中重人)

第 5 回「グラフの利用」目次

1. グラフの利用
2. 棒グラフとヒストグラム

【グラフの利用】

- 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

【棒グラフとヒストグラム】

- 棒グラフ……棒同士の間に空白をあける。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。

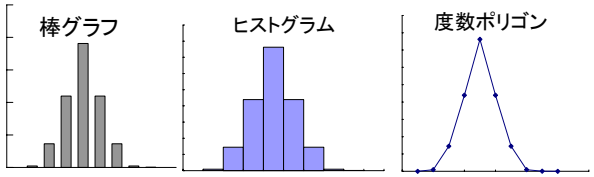
※縦軸は度数または%

★ 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に便利。

(教科書 p. 32–36)



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

Excel を使う場合：

★ recode で整形した上で度数分布表を出力
★ 表を Excel にコピーする
★ 棒グラフを作成
★ グラフの棒の上で左クリック
→「データ系列の書式設定」
→「オプション」
→「棒の間隔」を 0 にする

【実習】

(1) 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

(2) 年齢の度数分布表を出力し、中央値と 80%点に印をつけよ

(3) 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

2008.5.22 比較現代日本論研究演習 I (田中重人)

第 6 回「クロス表分析の基礎」

【キーワード】

行 (row)　列 (column)　セル (cell)

周辺度数 (marginal frequency)

行% (row percent)　列% (column percent)

【度数分布表の比較】

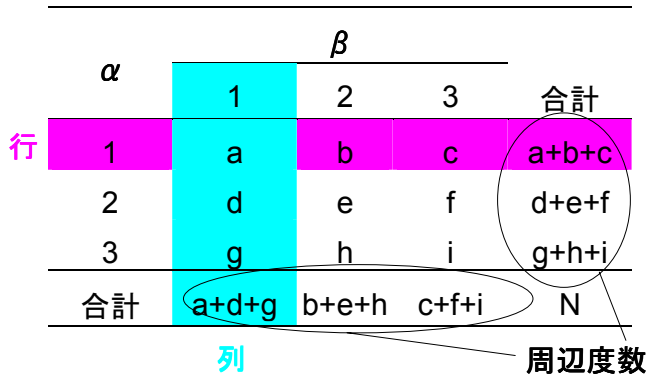
● データエディタのメニューで
「データ」→「ファイルの分割」
→「グループの比較」

● 度数分布表を出力

● 「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

【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



【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」

【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**

【グラフを書いてみる】

★ **クロス表は帯 (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**

【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

2008.6.5 比較現代日本論研究演習 I （田中重人）

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)

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

3×3 クロス表：セル度数が 4 つ決まれば…

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

$k \times l$ クロス表の自由度 (degree of freedom)

$$\text{d.f.} = (k-1)(l-1)$$

【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

【2×2 クロス表の性質】

以下、つぎの記号法を使う

α	β 1	β 2	合計
1	a	c	g
2	b	d	h
合計	i	j	N

(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：

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

(4) これら以外の場合、行%の差と列%の差はちがう値になる

(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%

【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)

【SPSS でのφ係数の計算】

「クロス集計表」の「統計」で
「ファイと Cramer の V」をチェック

【宿題】

適当な変数の組み合わせでクロス表（度数、行%、列%）とφ係数を出力。表からわかることを書く。

%とφとの関連を考えておくこと。

2008.6.12 比較現代日本論研究演習 I (田中重人)

第 8 回「連関係数」

【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*)

【φ係数の性質】

1. φ＝交差積の差 / √(周辺度数の積)
2. φ＝ 相関係数の特殊ケース
3. |φ| ＝ 行%差と列%差の中間の値
4. **ϕ^2 ＝ 標準残差の 2 乗の総計 / *N***
   (→ 2×2 以上のクロス表に拡張できる)

【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 :

gi/N	gj/N
hi/N	hj/N

各セルの期待度数は?

		100 100.0%
		50 100.0%
78 52.0%	72 48.0%	150 100.0%

★ 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/√期待度数

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

観測度数が下記の場合、

各セルの残差と標準残差は?

40	60	100 100.0%
38	12	50 100.0%
78 52.0%	72 48.0%	150 100.0%

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

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

【クラメールの連関係数 *V*】

***k*×*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

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

【*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**

【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 Cramer の V

【宿題】

別紙

【予告】

再来週 (6/26) は中間試験

- **何でも持ち込み可**
- **出題範囲は、6/19 授業まで**

練習問題

男性 220 人、女性 200 人を対象にしたある調査結果によると、お酒を呑む者の率は男性では 60.0%、女性では 45.5%であった (欠損値はないものとする)。

(1) 次のようなクロス表を作成せよ。小数の値は、小数第 1 位まで書くこと。

	呑む	呑まない	合計
男性	人数 (%) 期待度数 残差 標準残差	人数 (%) 期待度数 残差 標準残差	人数 (%)
女性	人数 (%) 期待度数 残差 標準残差	人数 (%) 期待度数 残差 標準残差	人数 (%)
合計	人数 (%)	人数 (%)	人数 (%)

(2) χ^2 と ϕ 係数を求めよ。計算のプロセスがわかるように書くこと

2008.6.19 比較現代日本論研究演習 I (田中重人)

第 9 回「表・図の書きかた」

1. 他人に見せる表
2. 表と図のあつかい
3. 表の書きかた

【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

【他人に見せられない表】

★セル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

★グループの人数が少なすぎるもの
→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

★ 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

【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

【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

【クロス表】

- 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

★ 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

★ 統計的検定の結果
(「P<0.05」とか「5%水準で有意」と書く)

★ 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

2008.6.19 比較現代日本論研究演習 I (田中重人)

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				
	「大いにある」	「少しはある」	「ない」	合計	（人）
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　　$p < 0.05$　　無回答=7

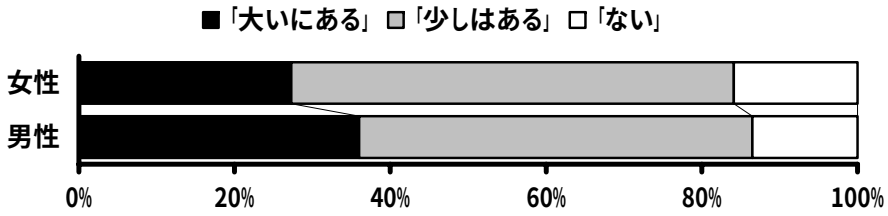


図1　性別と性別による不公平感との関連

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)

φ=0.191*　無回答=3

*: 5%水準で有意

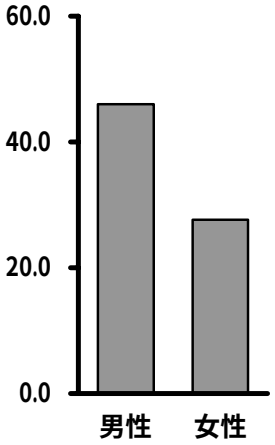


図2　県や市町村の部課長以上の役人に知り合いがいる%の男女差

2008.7.3 比較現代日本論研究演習 I (田中重人)

第 11 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

【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

【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)

【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

★次のデータの平均と SD は?

{0, 1, 4, 5, 7}

値	偏差	偏差^2
0		
1		
4		
5		
7		

平方和＝

分散＝

SD ＝

【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

【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

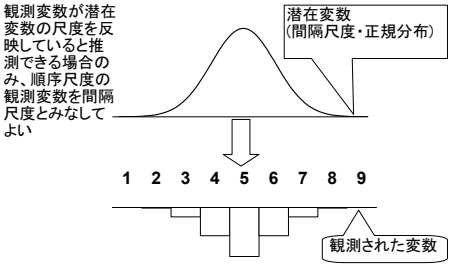
★平均値・標準偏差は間隔尺度以上のデータに対してしか意味をもたない。
順序尺度の平均値をとっていいのは

・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。



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

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

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**

※ 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

【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

【期末レポート】

期限：8/12 (火) 17:00

提出先：日本語教育学研究室 (文法合同棟 2F)
205 室の田中のレターケース

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。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。

備考： 後期の授業を受講しない者は、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。

2008.7.10 比較現代日本論研究演習 I (田中重人)

第 12 回「平均値の層別比較」

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

【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)

【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)

【エフェクト・サイズ】

ES ＝ 平均値の差／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)

【例】

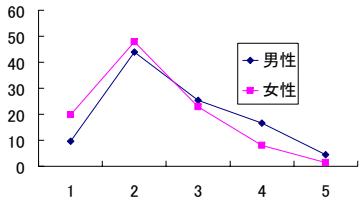
性別による生活全般満足度の違い

	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＝
併合 SD ≒
ES ＝

※ ES は SPSS では計算してくれない

性別による生活満足度の違い



【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？

【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 η^2 を「決定係数」「分散説明率」などという

※ η^2 を「相関比」ということもある

★ 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∞

★ 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。

【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

【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 ($n_1 = n_2$) 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

※ 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

【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 はクラメールの連関係数 V に等しい

男性＝１，女性＝２

のような変数の平均値の意味は?

【課題】

(1) 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
(併合 SD のかわりに層別 SD の単純平均を使ってよい)。
表に ES と η を書き込んで提出。

(2) 平均と SD の表から η を求める方法を考える

2008.7.17 比較現代日本論研究演習 I (田中重人)

第 13 回「平均値の比較結果」

1. 全体と層別の平均値・標準偏差
2. モデルとデータの乖離
3. 表とグラフの書きかた

【層別の平均値】

次のデータの平均値と SD は?

{1, 1, 2, 2, 3, 5, 4, 5, 4, 3}

これをふたつの層に分割すると：

{1, 1, 2, 2}　　{3, 5, 4, 5, 4, 3}

全体の平均と分散： M, V

層別の平均と分散： m_1, m_2, v_1, v_2

各層の人数： n_1, n_2　　全人数：$N = n_1 + n_2$

★ $M = (n_1 m_1 + n_2 m_2) / N$

★ 併合分散 $P = (n_1 v_1 + n_2 v_2) / N$

★ 層別平均値による仮想分散 $U = V - P$

【相関比の意味】

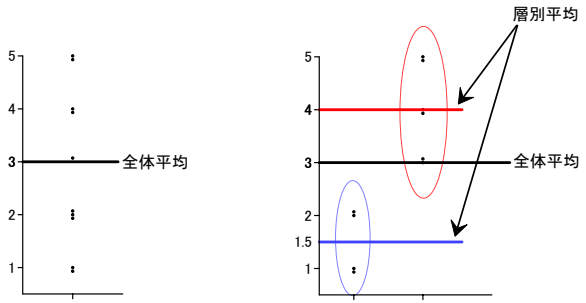
分散の分解：$V = U + P$

（ 全分散 ＝ 層間分散 ＋ 層内分散 ）

層の違いで説明できる　　できない

層間分散と全分散の比が相関比の 2 乗：

$$\eta^2 = \frac{U}{V}$$



【モデルとデータの乖離】

連関係数も相関比も、モデルとデータの乖離を表した値と解釈できる

- 特定の仮定 (モデル) の下で予測される値 (期待度数・仮想 SD) を求める
- 実際のデータの値と比較する
- 0～１の範囲の係数になるように調整する

多くの統計手法がこのタイプに属する

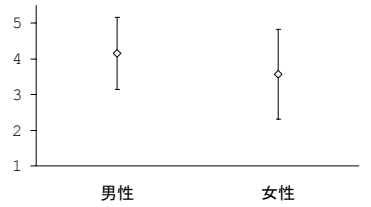
【表に書くべき要素】

- 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

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)

比較現代日本論研究演習I (田中 重人)

2008.7.17 課題

氏名:
学年:
所属:
学生番号:

次の3つの表の網掛け部分を埋めよ

全体についての平均と標準偏差

		平均	偏差	偏差^2
	1	3.00	−2.00	4.00
	1	3.00	−2.00	4.00
	2	3.00	−1.00	1.00
	2	3.00	−1.00	1.00
	3	3.00	0.00	0.00
	5	3.00	2.00	4.00
	4	3.00	1.00	1.00
	5	3.00	2.00	4.00
	4	3.00	1.00	1.00
	3	3.00	0.00	0.00
合計	30	30.00		
平均	3.00	3.00		
			SD=	

層別平均を当てはめた仮想データセットの平均と標準偏差

層別平均	全体平均	偏差	偏差^2
1.50	3.00		
1.50	3.00		
1.50	3.00		
1.50	3.00		
4.00	3.00		
4.00	3.00		
4.00	3.00		
4.00	3.00		
4.00	3.00		
4.00	3.00		
4.00	3.00		
30.00	30.00		
3.00	3.00		
		SD=	

層別の平均と標準偏差

		層別平均	偏差	偏差^2
	1	1.50		
	1	1.50		
	2	1.50		
	2	1.50		
合計	6	6.00		
平均	1.50	1.50		
			SD=	
	3	4.00		
	5	4.00		
	4	4.00		
	5	4.00		
	4	4.00		
	3	4.00		
合計	24	24.00		
平均	4.00	4.00		
			SD=	

η =
η^2=

比較現代日本論研究演習I (田中 重人)

2008.7.17 課題

氏名:
学年:
所属:
学生番号:

次の3つの表の網掛け部分を埋めよ

全体についての平均と標準偏差

		平均	偏差	偏差^2
	1	3.00	−2.00	4.00
	1	3.00	−2.00	4.00
	2	3.00	−1.00	1.00
	2	3.00	−1.00	1.00
	3	3.00	0.00	0.00
	5	3.00	2.00	4.00
	4	3.00	1.00	1.00
	5	3.00	2.00	4.00
	4	3.00	1.00	1.00
	3	3.00	0.00	0.00
合計	30	30.00	0.00	20.00
平均	3.00	3.00	0.00	2.00
			SD=	1.41

層別平均を当てはめた仮想データセットの平均と標準偏差

層別平均	全体平均	偏差	偏差^2
1.50	3.00	−1.50	2.25
1.50	3.00	−1.50	2.25
1.50	3.00	−1.50	2.25
1.50	3.00	−1.50	2.25
4.00	3.00	1.00	1.00
4.00	3.00	1.00	1.00
4.00	3.00	1.00	1.00
4.00	3.00	1.00	1.00
4.00	3.00	1.00	1.00
4.00	3.00	1.00	1.00
30.00	30.00	0.00	15.00
3.00	3.00	0.00	1.50
		SD=	1.22

層別の平均と標準偏差

		層別平均	偏差	偏差^2
	1	1.50	−0.50	0.25
	1	1.50	−0.50	0.25
	2	1.50	0.50	0.25
	2	1.50	0.50	0.25
合計	6	6.00	0.00	1.00
平均	1.50	1.50	0.00	0.25
			SD=	0.50
	3	4.00	−1.00	1.00
	5	4.00	1.00	1.00
	4	4.00	0.00	0.00
	5	4.00	1.00	1.00
	4	4.00	0.00	0.00
	3	4.00	−1.00	1.00
合計	24	24.00	0.00	4.00
平均	4.00	4.00	0.00	0.67
			SD=	0.82
			併合SD=	0.71

平均

η = 0.866
η^2= 0.750

全体SD^2= 2.00
仮想SD^2= 1.50
併合SD^2= 0.50