URL: http://www.nik.sal.tohoku.ac.jp/~tsigeto/statu/
作成：田中重人（講師）〈tsigeto@nik.sal.tohoku.ac.jp〉

現代日本論演習 I

3 年生対象：2003 年度前期 (5 セメスター：授業コード=L52504)
＜火５＞コンピュータ実習室 (文学部本館 7F 711-2)

『講義概要』p. 164 記載内容

◆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、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。それと同時に、コンピュータを実際に使って、毎回データ分析の実習をおこないます。
◇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。

授業の概要 (予定) 4/8 現在

目次

※（）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
- データ行列 (データセット)
- 模擬データ入力実習

2. データ配布・SPSS 入門

- データの配布
- SPSS の概要
- 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
- 他のソフトウエアについて (電卓, Excel, Word?)
- 印刷

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

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
- 順序尺度と間隔尺度の変換
- 平均値
- 分散と標準偏差
- 分布と外れ値

6.2. 平均値の層別比較

- 層別平均
- エフェクト・サイズ
- 相関比から分散分析へ
- 表とグラフの書きかた

7. 推測統計

7.1. 誤差の評価

- データの記述と誤差の評価：復習
- Case, Sample, Population, Universe
- 無作為抽出
- 非標本誤差
- 標本誤差の統計的推測

7.2. 平均値の推定

- 平均値の点推定
- 区間推定と t 分布
- 平均値の差の区間推定
- エフェクトサイズ・相関比と区間推定

7.3. 統計的検定

- 区間推定の簡易表記としての有意水準
- 平均値の差の t 検定
- 連関係数の $\chi 2$ 検定
- 分散分析と F 検定
- 検定結果の表記

8. 期末レポート

URL: http://www.nik.sal.tohoku.ac.jp/~tsigeto/statu/u030408.html
作成：田中重人（講師）〈tsigeto@nik.sal.tohoku.ac.jp〉

現代日本論演習 I 「統計分析の基礎」

第 1 回（2003-04-08）

この授業の概要・スケジュール・評価方法

コンピュータ実習室について

入室・退室

カードが必要。

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

ディスプレイの電源を落とすのを忘れないこと。

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 My Document 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ディスクにコピーして、 内蔵ディスクのほうのファイルは削除すること。

フロッピー（3.5 インチ）は各自購入しておくこと。 「Windows フォーマット」のものが便利である。

受講者の興味と数学的知識の調査

→別紙

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 10.0J」→「SPSS for Windows 10.0J」で起動する。

「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 q7a, q7b, ..., q7e 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いじらなくてよい
- 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 My Document 内に保存してみる。

「エクスプローラ」で My Document 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。（フロッピーにコピーする必要はない。）

※ 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。

2003.4.08

現代日本論演習 II (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（　　）

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) 「対偶」とは何か。簡単に説明せよ。

(3) 「平均」とは何か。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

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

数学的予備知識の調査：解答のポイント

(1)「乱数」とは

すべての数字が、**おなじ確率**で偶然に出てくる。

事前の規則性を持たない

(2) 命題 「A ならば B」に対して、

「B でなければ A でない」を「対偶」(contrapositive)

という。

もとの命題と対偶命題は論理的に同値である

(3) 「平均」とは

・全員分を足して個体数で割った値

・ひとりあたり～

(4) つぎの数式の値：

$$\sum_{k=1}^{10} k = 1+2+3+4+5+6+7+8+9+10 =$$

現代日本論演習 I
「統計分析の基礎」

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

2003.4.8

現代日本論演習 I

統計分析の基礎

東北大学文学部 2003 年度
田中 亘人（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
- 推測統計の手法

【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。

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
★ 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

【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

【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

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。

【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。

模擬データ入力実習

2003.4.15 現代日本論演習 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. メニューとシンタックス
4. 変数値の再割り当て
5. 出力の読みかた・印刷

【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。

★ 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

【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, ...)

2003.4.22 現代日本論演習 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. 変数の分類
3. 度数分布表とヒストグラム

【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)

【変数の種類】

●名義尺度 (nominal scale)

(質的変数とも)

●順序尺度 (ordinal —)

●間隔尺度 (interval —)

●比率尺度 (ratio —)

(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

【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）

【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。

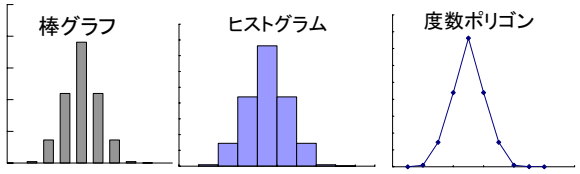
※縦軸は度数または%

★ 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

【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

2003.5.13 現代日本論演習 I (田中重人)

第 5 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)

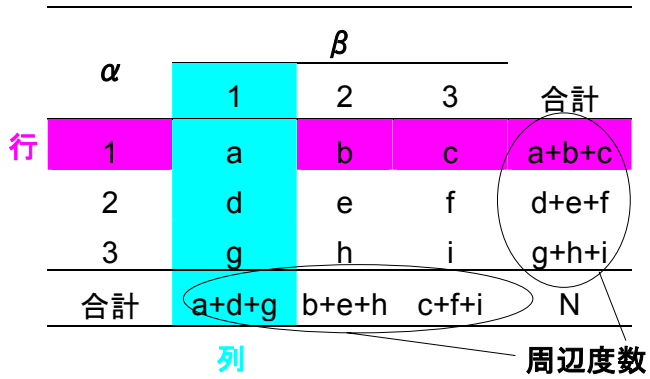
【度数分布表の比較】

- 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

2003.5.20 現代日本論演習 I (田中重人)

第 6 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：

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

(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
「ファイとクラマーの V」をチェック

現代日本論演習 I （田中 重人）

2003.5.20 課題

氏名：
学年：
所属：
学生番号：

周辺度数、%、%の差、ϕ を計算して下の表に書き入れよ。

α	β 1	β 2	合計
1	52	61	
2	37	97	
合計			

行%の差＝

列%の差＝

ϕ ＝

現代日本論演習 I （田中 重人）

2003.5.20 課題 解答例

周辺度数、%、%の差、ϕ を計算して下の表に書き入れよ。

α	β 1	β 2	合計
1	52	61	113
	46.0	54.0	100.0
	58.4	38.6	
2	37	97	134
	27.6	72.4	100.0
	41.6	61.4	
合計	89	158	247
	36.0	64.0	100.0

行%の差＝18.4

列%の差＝19.8

ϕ ＝0.191

2003.5.27 現代日本論演習 I (田中重人)

第 7 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

独立なクロス表の例

52 52.0% 66.7%	48 48.0% 66.7%	100 100.0%
26 52.0% 33.3%	24 48.0% 33.3%	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. $A = \dfrac{a - gi/N}{\sqrt{gi/N}}$

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
　　ファイとクラマー の Ｖ

中間試験

2003.6.3

【回答上の注意】

① コンピュータで解答を書き、印刷して提出
② 小数の解答については、小数第１位まで書くこと
③ 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 44 歳以下のグループと 45 歳以上の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。

【ヒント】 ２値の変数をつくってから分割処理をする

問 2　「記述統計」とはなにか。簡単に説明せよ。

問 3　４つの尺度水準について、それぞれの性質を簡単に説明せよ。

問 4　男性 246 人、女性 326 人を対象にしたある調査結果によると、クラシックコンサートによく行く者の率は男性では 28.9%、女性では 50.0%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)。

	よく行く	行かない	合計
男性	人数 (%) 期待値 残差	人数 (%) 期待値 残差	人数 (%)
女性	人数 (%) 期待値 残差	人数 (%) 期待値 残差	人数 (%)
合計	人数 (%)	人数 (%)	人数 (%)

中間試験　解答例

2003.6.3

問 1　年齢が 44 歳以下のグループと 45 歳以上の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。

```
RECODE
  q1_2a
  (Lowest thru 44=1)  (45 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　「記述統計」とはなにか。簡単に説明せよ。

データの特性を数値に要約して示すこと

問 3　４つの尺度水準について、それぞれの性質を簡単に説明せよ。

名義尺度：値が区別できるだけで、順序に意味がない
順序尺度：順序は一意に並べられるが、和や差をとることができない
間隔尺度：値の差に一定の意味があるため、和と差をとることができるが、ゼロ点に意味がないため、積や商はとれない
比率尺度:ゼロ点に意味があり、値の差と比に一定の意味がある。通常の演算がすべておこなえる。

問 4　男性 246 人、女性 326 人を対象にしたある調査結果によると、クラシックコンサートによく行く者の率は男性では 28.9%、女性では 50.0%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）。

	よく行く	行かない	合計
男性	71	175	246
	28.9	71.1	100.0
	100.6	145.4	
	-29.6	29.6	
女性	163	163	326
	50.0	50.0	100.0
	133.4	192.6	
	29.6	-29.6	
合計	234	338	572
	40.9	59.1	100.0

2003.6.3 現代日本論演習 I (田中重人)

第 8 回「クロス表の書きかた」

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

★ 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

【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

★ 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

★ 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

2003.6.3 現代日本論演習 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。無回答=7。

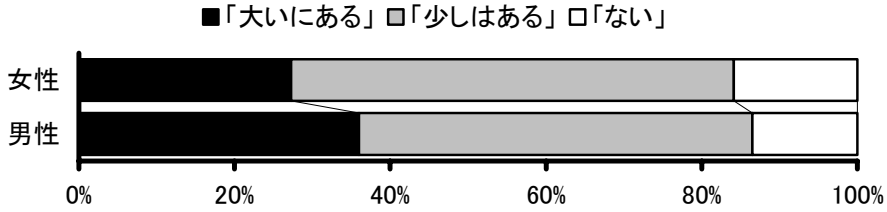


図1　性別と性別による不公平感との関連

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.

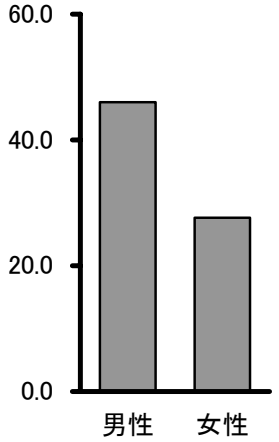


図2　県や市町村の部課長以上の役人に知り合いがいる%の男女差

2003.6.10 現代日本論演習 I (田中重人)

第 9 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?

$\{0, 1, 4, 5, 7\}$

【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

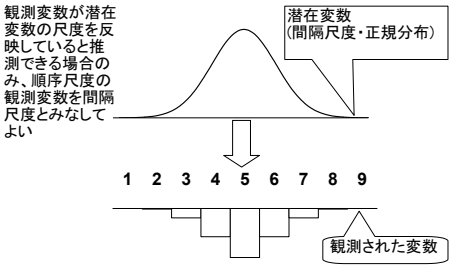
- **上下数%を取りのぞいたデータセットで計算する (調整平均：教科書 p. 46)**
- **順位に変換したり中央値を使って分析**

★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

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

2003.06.17 現代日本論演習 I (田中重人)

第 10 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)

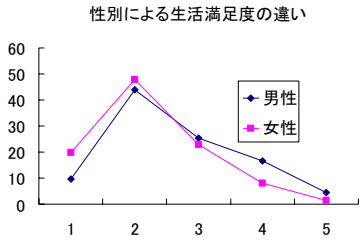
【例】

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＝0.39　併合 SD＝0.97
ES=0.401

※ ES は SPSS では計算してくれない



【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

【課題】

適当な変数の男女別平均値について平均値の差と ES を求める。表に ES を書き込んで提出。

2003.6.24 現代日本論演習 I (田中重人)

第 11 回「測定値と誤差」

1. 記述統計と推測統計
2. 「真の値」と測定値
3. 誤差の種類と対策
4. 標本抽出のプロセス
5. 期末レポートについて

【記述統計と推測統計】

記述統計＝データ（ケース）の特徴を数値や図表にまとめる

推測統計＝確率的な誤差を考慮して、母集団の特徴を推測する

(教科書 pp. 3–5)

【「真の値」と測定値】

測定値 ＝ 真の値 ＋ 誤差

記述　　推測

(教科書 pp. 17–20)

【誤差 (error) の種類】

- 測定上の誤差 (妥当性の問題)
  計器の故障・測定精度の問題
  回答者の間違い・虚偽の回答
  調査員の間違い・不正
  調査票の不備・入力ミス
- 対象者の選択に起因する誤差
  (サンプリングの問題)

【誤差への対策】

誤差の発生メカニズムを想定して対処する

★ 特定の方向へのかたより (bias)
→ できるだけ起こらないようにするか、かたよりの方向を想定して補正

★ 方向性を持たない (狭義の error)
→ できるだけ小さくする。
誤差の範囲を考慮してデータ解釈

【統計学があつかえる誤差】

- **発生メカニズムが既知**
- **誤差の範囲が確率的に決まる**

無作為標本抽出にともなう「標本誤差」がその典型である

【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)

*：一般的な用語ではないので注意

★ 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

【無作為抽出の実際】

★ **2 段抽出** ＝ 2 段階の抽出単位を設定

例：市町村→住民、学校→生徒

- 確率比例抽出法：その抽出単位が含む個体数に抽出確率を比例させる。
- 等確率抽出法：上位抽出単位の抽出確率は一定にしておき、個体の抽出数のほうを調整。

★ **系統抽出** ＝「台帳」から等間隔に抽出。

- スタート番号は乱数で決める
- 抽出間隔は次のことを考えてきめる
  (1) 台帳のもつ周期性と同調しない
  (2) 台帳全体をカバーできる
  具体的には 台帳人数／計画標本数に近い素数をえらぶのがよい。

★ **層化抽出法**＝母集団を層別にわけ、各層の人数に比例して標本数を割り当てる

- 結果に影響を与えそうな重要な属性についておこなう：性別・年齢・地域など
- 抽出単位や個体がどの層に属しているかを台帳から判断できないと使えない

※「層別抽出法」「比例割当抽出法」ともいう

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】

(1) 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、
- その記事等の標本抽出がわかる部分のコピー
- その記事等の出典がわかる文献情報を書いたもの
- 標本抽出の 4 段階にそった解説

を提出 (次回授業時)

(2) 次回までに、教科書の pp. 56–59, 147–152, 259–260 を読んでおいてください

【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。記述統計的な分析と推測統計的な分析の両方をふくんでいなければならない。

備考：後期の授業「現代日本論演習 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。

2003.7.1 現代日本論演習 I (田中重人)

第 12 回「区間推定」

1. 標本誤差の推定
2. 平均値の推定

【標本誤差の推定】

「標本誤差」(sampling error)
＝無作為抽出による誤差

★ 方向性をもたない
★ 確率的に決まる
★ 標本数が大きいほど誤差の範囲が小さい

➨「統計的推測」によって範囲を推定できる

【無限母集団の仮定】

母集団がある程度大きければ、統計的推測のうえでは、母集団は無限大とみなしてよい。

厳密にいうと、$\frac{N-n}{(N-1)n} \approx \frac{1}{n}$ の場合

➨ 無限大の母集団から n 個の標本を無作為に選んだ場合について考える

【母集団平均値の推定】

★ 等確率標本の平均値は、母集団の平均値より高くなったり低くなったりする。
★ そのばらつきは、母集団におけるばらつきが小さく、標本数が大きいほど小さくなる
★ しかし**平均的にみれば**母集団の平均値に一致すると期待できる

【平均値の信頼区間】

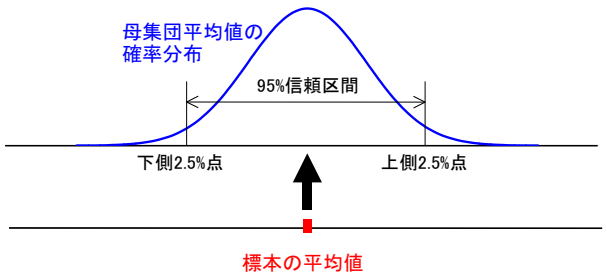
※「母集団では正規分布」の仮定が必要

★ 標本の平均値が母集団平均値からはずれる確率は正規分布にしたがう
➨ 標本平均値から逆算すれば、母集団の平均値の確率分布 (t 分布) がわかる

★ 母集団の平均値の確率分布から両端を α%分だけ切り落としてえられる区間を $(100-\alpha)$%の「**信頼区間**」という。

α を「**危険率**」、$(100-\alpha)$ を「**信頼率**」という。
この値は自由に決めていいのだが、通常は
α＝5% として、95%信頼区間を求める。

信頼区間のもとめかた



★ 平均値の信頼区間のおおよその値：

$$m \pm 1.96 \times \frac{SD}{\sqrt{n}}$$

m：標本平均
1.96：t 臨界値
$\frac{SD}{\sqrt{n}}$：標準誤差

【SPSS コマンド】

「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」を選択
※ 「因子」を指定すると層別に分析できる

2003.7.15 現代日本論演習 II (田中重人)

第 13 回「統計的検定」

1. 平均値の差の推定
2. 区間推定と統計的検定
3. 分散分析と F 検定
4. クロス表の独立性の検定
5. 検定結果の書きかた

【平均値の差の推定】

2 層間の 平均値の差 についても
平均値そのものと同様の区間推定ができる：
このとき 95%信頼区間はおよそ

$$\underbrace{d}_{\text{平均値の差}} \pm 1.96 \times \text{併合SD} \times \underbrace{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}_{\text{標準誤差}}$$

ただし n_1, n_2 はそれぞれの層の人数

各層の人数が多いほど
平均値の差の信頼区間が狭くなる

➡標本を均等にわけるほうが信頼性が高い

【SPSS のコマンド】

「平均値の比較」→「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

【区間推定と統計的検定】

Statistical test
統計的検定＝特定の値を設定して、その値が信頼区間に含まれているかどうかを判定する

0 に設定するのがふつう

95%信頼区間が 0 をふくまない
⇔「5%水準で有意」

※ 統計的検定の論理は本当はもっと複雑である。

【有意確率とは】

**信頼区間をひろげていくと、
どこかでゼロ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「有意水準」(level of significance)という。

分析の際は、
・前もって危険率を設定しておく
(通常は 5%または 1%)
・有意確率がその値を
下回っているかどうか判別する

例:
有意確率が 0.007 → 1%水準で有意（5%水準でも有意）
有意確率が 0.023 → 1%水準で非有意（5%水準では有意）
有意確率が 0.088 → 1%水準で非有意（5%水準でも非有意）

【統計的検定のいろいろ】

★ 平均値の差の t 検定
コマンドの指定は区間推定とおなじ。
出力の「有意確率 （両側)」を見る

※ 2 層の間の差の検定にしか使えない
※ 「母集団では正規分布」を前提とする
※ 2 層の間で分散が等しいことが前提

★ 分散分析と F 検定
「平均値の比較」→「グループの平均」
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」

※ 3 層以上の場合に使う。
η の信頼区間を使って判断するのと同じである。
※ 2 層の場合にも使えるが、t 検定と同じ結果になる
※ 必要とする前提も t 検定と同様

★ クロス表の独立性の検定
「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

※ **Vの信頼区間を使って判断するのとおなじ**
※ **各セルの期待度数が 5 以上であることを前提とする**

【検定結果の表示】

表の下に書く

(1) $p < 0.xx$ 形式：
平均値の差＝0.04 ($p > 0.05$)
相関比 η ＝0.198 ($p < 0.05$)
Cramer's V＝0.258 ($p < 0.01$)

(2) 「xx%水準で有意」 形式：
平均値の差＝0.04 (5%水準で非有意)
相関比 η ＝0.198 (5%水準で有意)
Cramer's V＝0.258 (1%水準で有意)

(3) 星印 (*) 形式：
平均値の差＝0.04^{ns}
相関比 η ＝0.198*
Cramer's V＝0.258**

→ 注釈をどこかに書いておく
(ns: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$)