URL: http://www.nik.sal.tohoku.ac.jp/~tsigeto/stata/2001/s010411.html
作成：田中重人（講師）<tsigeto@nik.sal.tohoku.ac.jp>

[現代日本論演習 I 統計分析入門]

第 1 回（2001-04-11）

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。 「DOS 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、ISBN 4-563-00888-5。

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 Q35A, Q35B, ... Q35F 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 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 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。

TANAKA Sigeto (tsigeto@nik.sal.tohoku.ac.jp)
Created at 2001-04-09. Last updated at 2001-04-09. Sorry to be Japanese only (encoded in accordance with MS-Kanji: "Shift JIS").

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

2001.4.11

現代日本論演習 I

統計分析の基礎

東北大学文学部 2001 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　(4 月)
- クロス表分析　(5 月)
- 平均値の比較　(6 月)
- 標本誤差の推定(7 月)

【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。

※ 増刷中。生協には 4/20 ごろ入荷予定

【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）

【コンピュータの起動と終了】

【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。

受講フォームを配布

【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

2001.4.11

現代日本論演習 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) 「関数」とは何か。簡単に説明せよ。

(2) 1 次方程式 $y = 0.5x + 1.2$ をグラフに書いたとき、傾きと切片はそれぞれいくつか。

傾き ＝＿＿＿＿＿＿＿ ； 切片 ＝＿＿＿＿＿＿

(3) 「必要十分条件」とは何か。簡単に説明せよ。

(4) 「平均」とは何か。簡単に説明せよ

(5) つぎの数式の値を求めよ：

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

数学的予備知識の調査：解答のポイント

(1)「関数」とは

X がきまればそれに対応して Y が**一意に定まる**

(2) 1 次方程式 $y = 0.5x + 1.2$ をグラフに書いたとき…

↓ 傾き　↓ 切片

(3) 「必要十分条件」とは

X という条件が**あるときはかならず**、そして**その時にかぎって** Y である…

(4) 「平均」とは

・全員分を足して個体数で割ったもの

・ひとりあたり～

(5) つぎの数式の値：

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

2001.4.18 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

★ 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 (大阪大学) 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

★ 印刷前にプレビュで確認

2001.4.25 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)

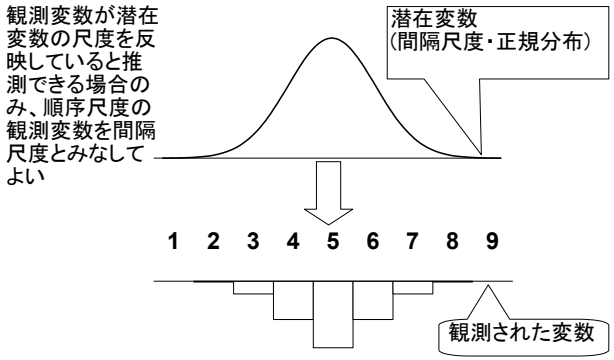
【変数の種類】

- ●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】

- ★ 上位の尺度は下位の尺度の特徴を兼ね備えている
- ★ 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】

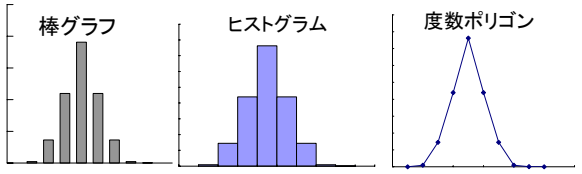
- ●棒グラフ……棒同士の間に空白をあける。高さ (長さ) をよむ。
- ●histogram (柱グラフ)……柱の間隔をあけない。面積をよむ。

※縦軸は度数または% (状況次第)

- ★ 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 (質的変数／離散量) → 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 histgram は書きにくい。

★ 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

2001.5.2 現代日本論演習 I (田中重人)
第 4 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)

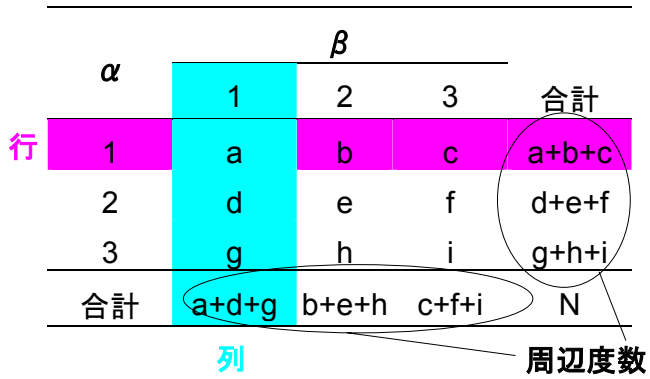
【度数分布表の比較】

- 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」
- そのあとで度数分布表を書いてみる

- 「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

2001.5.9 現代日本論演習 I (田中重人)
第 5 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	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	2	3	合計
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)**

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

【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	2	合計
1	a	c	g
2	b	d	h
合計	i	j	N

(1) 行%は 1 列についてだけ比較すればよい:

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

(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

【φ係数の別定義】

Pearson 積率相関係数 (教科書 72 頁) の特殊ケース
(四分点相関係数 4-fold point correlation)

φの絶対値 $|\phi|$ を使うこともある (教科書 110 頁)

【%の差とφ係数】

φ の絶対値＝行・列の比率差の幾何平均：

$$|\phi| = \sqrt{\left|\frac{a}{g}-\frac{b}{h}\right|\left|\frac{a}{i}-\frac{c}{j}\right|} = \sqrt{\left|\frac{c}{g}-\frac{d}{h}\right|\left|\frac{b}{i}-\frac{d}{j}\right|}$$

行 % の差　　列 % の差

周辺度数がつりあっていれば ($g=i$)、行%の差と列%の差は等しい。

現代日本論演習 I （田中 重人）

2001.5.9 課題

氏名：
学年：
所属：
学生番号：

I. 2×2 クロス表の性質に関して、つぎの事項を証明せよ。記号法は別紙参照。

(1) 行%は 1 列についてだけ比較すればよい： $\frac{a}{g}-\frac{b}{h}=\frac{d}{h}-\frac{c}{g}$

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

II. 周辺度数、%、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）

2001.5.9 課題 解答例

I. 2×2 クロス表の性質に関して、つぎの事項を証明せよ。記号法は別紙参照。

(1) 行%は 1 列についてだけ比較すればよい： $\frac{a}{g}-\frac{b}{h}=\frac{d}{h}-\frac{c}{g}$

$hg = h(a+c) = (b+d)g$

$(a+c)/g = (b+d)/h$

$a/g + c/g = b/h + d/h$

$a/g - b/h = d/h - c/g$

(2) 行%の差がゼロなら列%の差もゼロ

$a/g = b/h$

$a/(a+c) = b/(b+d)$

$a(b+d) = b(a+c)$

$ab + ad = ab + bc$

$ad = bc$

$ac + ad = ac + bc$

$a(c+d) = c(a+b)$

$aj = ci$

$a/i = c/j$

(3) $g=i$ **なら行%の差と列%の差は同じ：** $\frac{a}{g}-\frac{b}{h}=\frac{a}{i}-\frac{c}{j}$

$g=i$ **なら** $h = N - g = N - i = j$ **かつ** $c = g - a = i - a = b$

したがって $a/g - b/h = a/i - c/j$

II. 周辺度数、%、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

2001.5.16 現代日本論演習 I (田中重人)
第 6 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. $|\phi|$ ＝ 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

★ 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 (standardised ---) 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】

$k \times l$ 表への ϕ 係数の拡張

★ $k \times l$ 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

現代日本論演習 I （田中 重人）

2001.5.16 課題

氏名：
学年：
所属：
学生番号：

期待度数、残差、標準残差、標準残差の 2 乗、χ^2、ϕ を計算して下の表に書き入れよ。

α	β		
	1	2	合計
1	52	61	113
期待度数			
残差			
標準残差			
標準残差 ^{2}			
2	37	97	134
期待度数			
残差			
標準残差			
標準残差 ^{2}			
合計	89	158	247
(%)	(36.0)	(64.0)	(100.0)

$\chi^2=$

$\phi=$

現代日本論演習 I （田中 重人）

2001.5.16 課題 解答

期待度数、残差、標準残差、標準残差の 2 乗、χ^2、ϕ を計算して下の表に書き入れよ。

α	β		
	1	2	合計
1	52	61	113
期待度数	40.717	72.283	
残差	11.283	−11.283	
標準残差	1.768	−1.327	
標準残差 ^{2}	3.127	1.761	
2	37	97	134
期待度数	48.283	85.717	
残差	−11.283	11.283	
標準残差	−1.624	1.219	
標準残差 ^{2}	2.637	1.485	
合計	89	158	247
(%)	(36.0)	(64.0)	(100.0)

$\chi^2=9.010$

$\phi=0.191$

2001.5.23 現代日本論演習 I (田中重人)
第 7 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
4. グラフの書きかた
5. クロス表を説明する文章

【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 係数には意味がなくなる、とされている
→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

【Excel による作表】

★SPSS の出力をコピーして必要なところを残す

★「書式」→「セル」メニューの「表示形式」「配置」「罫線」で整形

【グラフの書きかた】

クロス表は積み上げ棒グラフで表現するのがふつう

(2 列表の場合は棒グラフでも)

(度数多角形を重ねる場合もある)

★グラフは細かい数字がのせにくいし、紙幅を食う。基本的には表を使用して、特に視覚的インパクトを狙う場合に限ってグラフを使うのがよい

【Excel によるグラフ作成】

★注意点：

- データ系列の順序
- 凡例の表示
- 区分線の表示
- 棒同士の間の距離

【クロス表を説明する文章】

★まず周辺度数分布を説明

★ 線形の関連か？

大小のある変数で、行によって大きいほうまたは小さいほうに偏っている関連

★どのセルに注目するか？

特徴的なセルをみつけて、それをほかの適当なセル（周辺度数）と比較する。カテゴリー統合のセンスがものをいう

★ %とポイント差：

「Y は X より 10%大きい」とは

Y=1.1X　それとも　Y=0.1+X？

後者の場合は「10 ポイント大きい」と書く

★表に示される事実と、自分が加える解釈との峻別

【解釈のツボ】

★実感に照らして納得できるか

★測定・分析のミスではないか

★標本の偏りで説明できないか

(→統計的検定)

【先行研究と照合】

★命題を立てる

★既存のデータと整合的か

★ (後知恵ではないかたちで)

理論的に説明できるか

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。

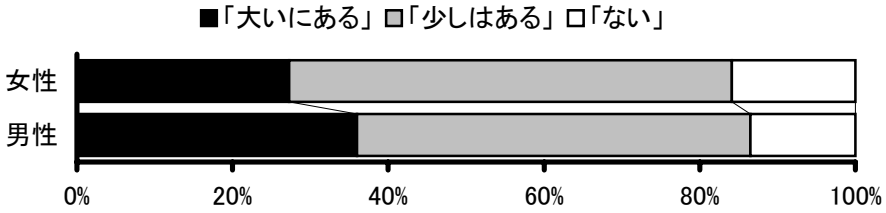


図1　性別と性別による不公平感との関連

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.

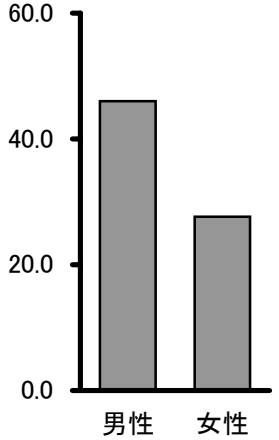


図2　県や市町村の部課長以上の役人に知り合いがいる%の男女差

2001.6.6 現代日本論演習 I (田中重人)
第 9 回「平均値と標準偏差」

1. 代表値と散布度
2. 平均値
3. 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

【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

【SPSS のコマンド】

「記述統計」→「記述統計」

→ 変数とオプションを指定

【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

➡ 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)

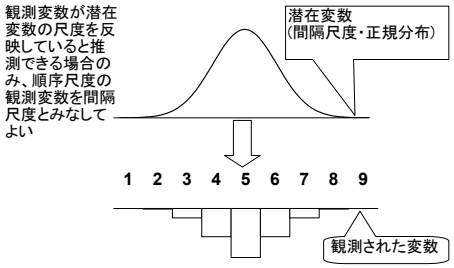
● 順位に変換したり中央値を使って分析

※ 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

【課題】

適当な変数について、ヒストグラムの軸上に平均値と標準偏差を（手書きで）書き入れたものを作って提出



2001.6.20 現代日本論演習 I (田中重人)
第 10 回「平均値の比較」

1. 平均値の比較
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)

【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】

性別による不公平

	平均	SD	(人数)
男性	1.77	0.67	(111)
女性	1.89	0.65	(132)
合計	1.84	0.66	(243)

平均の差＝0.11　併合 SD＝0.66
ES=0.17　η ＝0.08　η^2=0.01

※ 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ような状況を仮定して 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、独立変数の影響力が大きいほど大きくなる

★ 同じ大きさの 2 層で平均値を比べる場合、

$$ES^2 = \frac{4\eta^2}{1-\eta^2}$$ という関係がある。

層の大きさがちがえば、ES はこの式よりも大きくなる

※ 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。

2001.6.27 現代日本論演習 I (田中重人)

第 11 回「平均値と標準偏差の性質」

1. 全体と層別の平均値・標準偏差
2. ダミー変数の平均値
3. 表の書きかた

【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 $U = V - P$

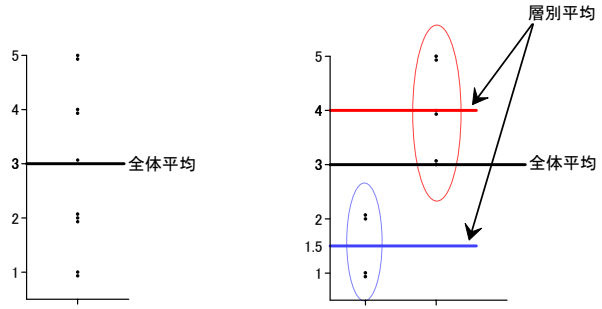
【相関比の意味】

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



平均値の差と層間分散との関係:

$$(m_1 - m_2)^2 = UN^2 / n_1 n_2$$

エフェクト・サイズの定義 (前回資料) から

$$ES^2 = \frac{(m_1 - m_2)^2}{P} = \frac{UN^2}{(V-U)n_1 n_2} = \frac{\eta^2}{1-\eta^2} \times \frac{N^2}{n_1 n_2}$$

※ $n_1 = n_2 = N/2$ のときは $ES^2 = 4U / P$ が成り立つ

【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」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

【表に書くべき要素】

- 各層と全体の平均値と標準偏差
  (素データの測定水準の 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

【例】

表 1 「性別による不公平感」の性別によるちがい

	平均	標準偏差	(人)
男性	1.77	0.67	(111)
女性	1.89	0.65	(132)
合計	1.84	0.66	(243)

η = 0.08。無回答 = 7。

現代日本論演習I (田中 重人)

2001.6.27 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	−2	4
	1	3	−2	4
	2	3	−1	1
	2	3	−1	1
	3	3	0	0
	5	3	2	4
	4	3	1	1
	5	3	2	4
	4	3	1	1
	3	3	0	0
合計	30	30		
平均	3	3		
			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
1.5	3		
1.5	3		
1.5	3		
1.5	3		
4	3		
4	3		
4	3		
4	3		
4	3		
4	3		
30	30		
3	3		
		SD=	

層別の平均と標準偏差

		層別平均	偏差	偏差^2
	1	1.5		
	1	1.5		
	2	1.5		
	2	1.5		
合計	6	6		
平均	1.5	1.5		
			SD=	
	3	4		
	5	4		
	4	4		
	5	4		
	4	4		
	3	4		
合計	24	24		
平均	4	4		
			SD=	

現代日本論演習I (田中 重人)

2001.6.27 課題 解答

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

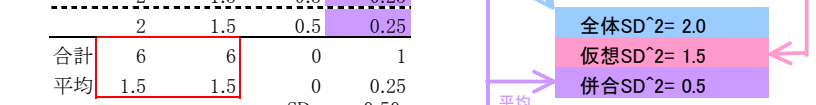
		平均	偏差	偏差^2
	1	3	−2	4
	1	3	−2	4
	2	3	−1	1
	2	3	−1	1
	3	3	0	0
	5	3	2	4
	4	3	1	1
	5	3	2	4
	4	3	1	1
	3	3	0	0
合計	30	30	0	20
平均	3	3	0	2
			SD=	1.41

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
1.5	3	−1.5	2.25
1.5	3	−1.5	2.25
1.5	3	−1.5	2.25
1.5	3	−1.5	2.25
4	3	1	1
4	3	1	1
4	3	1	1
4	3	1	1
4	3	1	1
4	3	1	1
30	30	0	15
3	3	0	1.5
		SD=	1.22

層別の平均と標準偏差

		層別平均	偏差	偏差^2
	1	1.5	−0.5	0.25
	1	1.5	−0.5	0.25
	2	1.5	0.5	0.25
	2	1.5	0.5	0.25
合計	6	6	0	1
平均	1.5	1.5	0	0.25
			SD=	0.50
	3	4	−1	1
	5	4	1	1
	4	4	0	0
	5	4	1	1
	4	4	0	0
	3	4	−1	1
合計	24	24	0	4
平均	4	4	0	0.66667
			SD=	0.82
			併合SD=	0.70711



2001.7.4 現代日本論演習 I (田中重人)
第 12 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

【記述統計と推測統計】

記述統計＝データ (**標本**) の特徴を数値や図表でまとめる (教科書 p. 5)

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 p.148)

【「真の値」と測定値】

測定値 ＝ 真の値 ＋ 誤差

記述　　推測

【誤差 (error) の種類】

- 測定上の誤差
  - 計器の故障・測定精度の問題 (教科書 p. 18)
  - 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

【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、かたよりの方向性を想定しておく

★ 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)

SSM 調査の場合について、それぞれの段階をあてはめてみよう

★ **伝統的な推測統計学では 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

2001.7.11 現代日本論演習 I (田中重人)

第 13 回「無作為抽出と区間推定」

1. 無作為抽出の理論と実際

2. 平均値の推定

【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。

★ **層別抽出法**＝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

※「層化抽出法」「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

【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。

➡ 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。

★ 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 **(点推定)**

【平均値の信頼区間】

※「母集団では正規分布」の仮定が必要

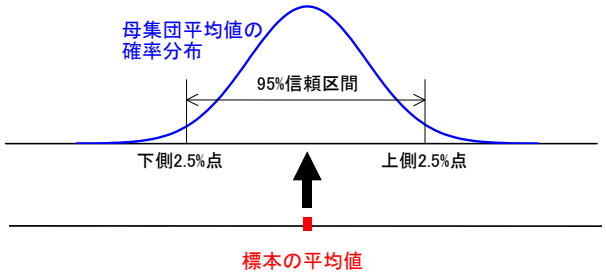
★ 標本の平均値が母集団平均値からはずれる確率は正規分布にしたがう

➡ 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　　$\frac{SD}{\sqrt{n}}$：標準誤差　　1.96：t 臨界値

(教科書 p. 281)

【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

【課題】

適当な変数について

- **全標本**
- **男女別**

の平均値と信頼区間をもとめ、グラフを描く

【期末レポートについて】

期限: 8/1 (水)　　**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: SSM データのディスクをレポートと一緒に提出。データのコピーはすべて消去すること

【参考文献】

谷岡 一郎 (2000)『「社会調査」のウソ』文藝春秋。

森岡 清志 (1998)『ガイドブック社会調査』日本評論社。

盛山 和夫 + 近藤 博之 + 岩永 雅也 (1992)『社会調査法』放送大学教育振興会。

鈴木 達三 + 高橋 宏一 (1998)『標本調査法』朝倉書店。

2001.7.18 現代日本論演習 I (田中重人)

第 14 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. 検定結果の表示

【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

➡標本を均等に分けたほうが 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】

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 を含む⇔「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。教科書 6 章を参照

【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を前提と

★ 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】

例 1

	平均	標準偏差	(人)
男性	1.77	0.67	(111)
女性	1.89	0.65	(132)
合計	1.84	0.66	(243)

η =0.086。p > 0.05。無回答＝ 7 。

例 2

	平均	標準偏差	(人)
男性	2.62	1.02	(114)
女性	2.24	0.91	(136)
合計	2.41	0.98	(250)

η =0.198*。*: 5%水準で有意

URL: http://www.nik.sal.tohoku.ac.jp/~tsigeto/stata/repfeed.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」

2001.8.29

レポート返却のお知らせ

期末レポートの採点結果を返却します。日本語教育学の学生については、研究室のメールボックスに入れておきます。その他の受講者は、田中研究室（文学部・法学部合同研究棟2F）までとりにきてください。

講評

レポートは 60 点満点で、

- 記述統計・推測統計の両方が使いこなせているか。
- 図表がわかりやすく書けているか

を中心に採点しています。

つぎの 2 点は理解のあやうい人がいたので、念のために追記しておきます。

1. 統計的検定の結果が「有意でない」ということは、分析結果の数値（たとえばクロス表の行%の差）が誤差の範囲内であるために、結論をくだせない、ということです。したがって、有意ではない分析結果については、解釈は本来できません。
2. ファイ係数（Φ）は 2×2 のクロス表にだけ使います。それ以外の場合はクラメールの連関係数 V を使ってください。

このレポートの点数に毎回の課題（40 点）の評価を加えて、成績評定をおこないました。

TANAKA Sigeto (tsigeto@nik.sal.tohoku.ac.jp)
Created: 2001-08-29. Updated: 2001-09-20. Sorry to be Japanese only (encoded in accordance with MS-Kanji: "Shift JIS").