URL: http://www.sal.tohoku.ac.jp/~tsigeto/statu/
作成：田中重人（准教授）

現代日本論演習「統計分析の基礎」

3 年生対象：2010 年度前期（5 セメスター：授業コード=L52405）
＜火４＞コンピュータ実習室（文学部本館 7F 711-2）

『講義概要』p. 168 記載内容

◆講義題目：統計分析の基礎
◆到達目標：(1) 統計分析の基礎的な手法を理解する；(2) 実際に統計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、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、毎回データ分析の実習をおこないます。
◇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。
◇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。

卒業論文等で質問紙調査を予定している者は、 5 セメスタ開講の現代日本論演習「質問紙法の基礎」（火 5）および 5 セメスタ開講の現代日本論演習「応用統計分析」（木 2：大学院と合同）も受講することがのぞましい。

授業の概要

1. イントロダクション（4/13）
2. SPSS 入門（4/20～4/27）
3. 統計分析の基礎（5/11）
4. 記述統計(1)：クロス表の分析（5/19～5/25）
5. 中間試験（6/1）
6. 記述統計(2)：平均値の比較（6/8～6/29）
7. 推測統計（7/6～7/20）
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。

講師連絡先

田中重人（東北大学文学部日本語教育学研究室）
〒980-8576 仙台市青葉区川内 27-1 文学部・法学部合同研究棟 2F

オフィス・アワーは定めていません。適当な時間に予約をとってください。

現代日本論演習
「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

2010.4.13

現代日本論演習

統計分析の基礎

東北大学文学部 2010 年度
田中 重人（准教授）

【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。

【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】

5 セメスタ

・現代日本論演習「質問紙法の基礎」（火 5）

6 セメスタ

・現代日本論演習「実践的統計分析法」
（木 2）…大学院と合同

質的研究法の授業？

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
→研究生などは、オンラインで登録
(他学部の場合は文書で申請)
★ 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
- 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/21coe/ssm/ 参照

2010.4.20 現代日本論演習 (田中重人)
第 2 講「SPSS 入門・データ配布」目次

1. データの配布
2. 標本抽出
3. SPSS のウインドウ構成
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. (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

★ 毎回の授業で使うので、忘れないこと (調査票も)

★ 期末レポート提出時に返却

【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)

➔「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、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

【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)

【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 (▶)

【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, ...)

【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

【実習】

満年齢 (Q1_2a)を 10 才刻みに区切って

度数分布表を出力し、印刷して提出

【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)

SPSS の出力ビューアから表を Excel や Word に貼り付けられる

2010.4.27 現代日本論演習 (田中重人)
第 3 講「度数分布表」

1. 度数分布表
2. 累積%とパーセンタイル
3. グラフの利用

【度数分布表】

Frequencies コマンド

「分析」
→「記述統計」
→「度数分布表」

出力：

★ 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】

○ 順序に意味がある場合のみ有効（→次回）
○ 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)

【実習】

世帯収入 (q44_3) について、度数分布表を出力し、中央値、25%点、75%点を求めよ

【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】

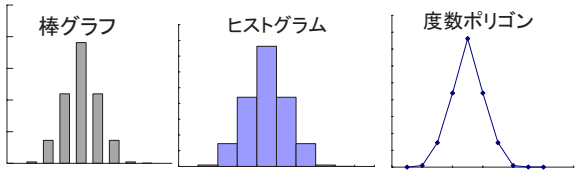
●棒グラフ……棒同士の間に空白をあける。高さ (長さ) をよむ。
●histogram (柱グラフ)……柱の間隔をあけない。面積をよむ。

※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
★ 必要なら変数値のラベルをつける
(横軸に表示される)
★ ヒストグラムや度数多角形のときは
両端に度数 0 の行をつくる
★ グラフを作成

棒グラフをヒストグラム風にするには

★ グラフの棒の上で右クリック
→「データ系列の書式設定」
→「オプション」
→「棒の間隔」を 0 にする

※ 見た目がそれらしくなるだけなので、横軸のラベルや階級幅の調整はむずかしい。本当のヒストグラムを書くには、グラフ専用のソフトウエアを使う。

【実習】

年齢について 5 歳刻みの
ヒストグラムを作成する:

(21–25, 26–30, … のようにラベルをつける)

【次回】

授業観察室 (文・法合同棟 2F)
でおこないます

URL: http://www.sal.tohoku.ac.jp/~tsigeto/family/

現代日本論演習「統計分析の基礎」 2010 年度

第 4 講　統計分析の基礎 (5/11)

田中重人 (東北大学文学部准教授)

[**今回のテーマ**]　データの性質に関する基本的な事項を理解する

1　データ収集から分析まで

1.1　データの収集

実験 (experiment) とは:
観察 (observation) とは:

1.2　分析可能な形への加工

- 分析の単位
- 変数の同定
- 変数値の付与 (coding)

1.3　データ・セット作成

- データ入力
- クリーニング

2　記述と推測

「統計をとる」ことの第 2、第 3 段階 (教科書 p. 1–6)

- データの特徴を少数の数値に要約 = 記述統計 (descriptive statistics)
- 誤差の評価 (この手続きの一部が推測統計 inferential statistics)

科学的な研究においては、分析結果の「正しさ」についての最終的な決着は、追試の繰り返しによる再現性のチェックによって行われるはずである。しかし、実際には……

- 費用や人的資源の不足などから、頻繁に追試がおこなわれない分野のほうが多い
- 厳密な追試が原理的に不可能であることも多い (歴史的な研究など)

このため、分析結果を公表する際には、誤差に関する情報をできるかぎり公表することが仁義となっている。推測統計は、この目的のために使われる標準的な手法のひとつ。

3　尺度水準

教科書 p. 8

- 比率尺度 (ratio scale)
- 間隔尺度 (interval scale)
- 順序尺度 (ordinal scale)
- 名義尺度 (nominal scale) → 「質的変数」と呼ばれることもある

上位の尺度のほうがあつかえる演算が豊富であり、また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 (SSM 調査の調査票参照)。

→　上位の尺度への変換には一定の理論的根拠が必要

実際には、本来は順序尺度のはずの変数について平均値を求めて分析する、といった類のことが広くおこなわれている。

2010.5.18 現代日本論演習 (田中重人)

第 5 講「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)

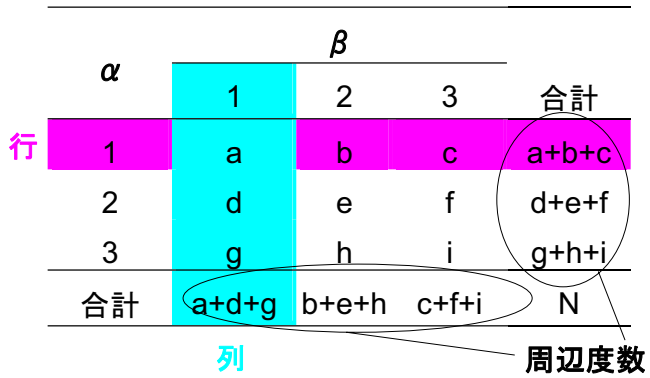
【度数分布表の比較】

- 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、%からわかることをコメントする。グラフも書いて印刷して提出

2010.5.25 現代日本論演習 (田中重人)

第 6 講「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

$k \times l$ 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	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) 行%の差が 100 なら列%の差も 100

(4) $g=i$ or $g=j$ なら行%の差と列%の差は同じ：

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

(5) 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)

2010.6.1 現代日本論演習 (田中重人)

第 7 講「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
   (→ VI セメスタ授業)
3. |φ| ＝ 行%差と列%差の中間の値
   (教科書 p. 103 表 4-1 について計算してみよう)

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

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

2010.6.8 現代日本論演習　(田中重人)

第 8 講「連関係数とクロス表の解釈」

1. 連関係数の性質
2. SPSS で実習
3. 尺度水準 (復習)
4. 代表値と散布度

【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：

「統計」で

「カイ 2 乗」「ファイと Cramer の V」

※「セル」で「度数」(観測／期待) と

「残差」(標準化なし／標準化)を指定することもできる

つぎの変数についてクロス表を解釈：

・性別 (q1_1) × 性別役割意識 (q35a)

・年齢 10 歳階級 × 性別役割意識 (q35a)

・生活水準の変化 (q36) × 満足度 (q37)

V がどれくらいか

→ どこに%の差があるか?

【注意事項】

期待度数の小さいセルがある場合、
連関係数は適切な指標にならない

→ 期待度数 < 5 のセルがないか、
カイ 2 乗値の表の下の警告で確認

【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】

★ 中央値 (median) ― 四分位偏差 (Q)

(順序尺度以上)

★ 平均値 (mean) ― 標準偏差 (SD)

(間隔尺度以上)

(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

【予告】

来週は中間試験

・出題範囲は、「クロス表」まで

・コンピュータで答案を作成、印刷して提出

・何でも持ち込み可
ただし通信と相談は禁止

・試験後は通常の授業

2010.6.29 現代日本論演習 (田中重人)

第 9 講「平均値の層別比較」

1. SPSS での平均値と標準偏差の計算
2. 層別 (group 別) 比較
3. 平均値を使うときの注意事項
4. Effect Size
5. 相関比

【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)

【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

【平均値を使うときの注意事項】

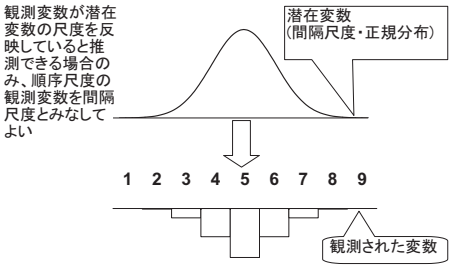
★順序尺度の平均値をとっていいのは

・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**

★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**

★左右対称でないデータ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

【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 コマンド】

「平均の比較」→「グループの平均」

「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。

【期末レポート】

期限 : 8/10 (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 : クロス表と平均値の比較について適当な分析をして結果を解釈する。統計的推測の結果をふくめること。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。

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。

2010.7.6 現代日本論演習（田中重人）

第 10 講「分散分析」

1. 相関比の意味
2. エフェクト・サイズと相関比
3. 推測統計の基礎
4. 区間推定

【相関比の意味】

ある個体の値を x，全体平均を M，層別平均を m とすると、全体平均との差（偏差）は

$$x - M = (x - m) + (m - M)$$

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}

★【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

【ES と η の関係】

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

特に、2 層の大きさが同じ（$n_1 = n_2$）なら、

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

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**

【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

【記述統計と推測統計】

記述統計（descriptive statistics）
＝データ（**ケース**）の特徴を
数値や図表にまとめる

推測統計（inferential statistics）
＝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)

【無作為抽出】

random sampling
母集団から計画標本を選ぶ際に、
すべての個体の抽出確率が等しくなる
ように抽出する

➔「**等確率標本**」(probability sample)

袋のなかに色つきの玉が 60 個入っている:

赤色: 30 個
青色: 30 個

玉を n 個取り出したとき、その色は……?

→ 全世界から n 人を無作為抽出したとき、そのなかに ○○ の人は何%ふくまれるか?

【区間推定】

interval estimation

「答えは **たぶん** この範囲内にある」
↓
信頼率（confidence level）を適当に設定して
信頼区間（confidence interval）を求める

全世界から 400 人を無作為抽出:

うどん が好き: 240 人
そば　が好き: 160 人

うどんが好きな人の比率は?

$$0.6 \pm 1.96 \times \sqrt{(0.6 \times 0.4 / 400)}$$

答:　　〜　　%（95%信頼区間）

【比率の区間推定】

標本の規模がじゅうぶん大きく（$n > 30$）、
比率があまり偏っていない（$0.1 < m < 0.9$）とき、

95%信頼区間は

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

標準誤差
(standard error)

【平均値の区間推定】

母集団における平均値についても、同様の計算ができる。
ただし、正規分布を仮定：

$$\underbrace{m}_{\text{標本平均}} \pm \underbrace{1.96}_{t\text{ 臨界値}} \times \underbrace{\frac{\mathrm{SD}}{\sqrt{n}}}_{\text{標準誤差}}$$

※「t 臨界値」は n によって変化するが、n>200 で 1.96 に収束（教科書 p. 281）。

現代日本論演習　(田中重人)

第 11 講「推測統計の基礎」(2010.7.13)

1. 推定と検定
2. 平均値の差の区間推定
3. 平均値の差の検定
4. 有意確率
5. その他の検定

【推定と検定】

限られたデータに基づいた
合理的意思決定のための統計的基準

- ある統計量の母集団における値について確率的な推測を行なうのが「推定」
- 母集団における統計量についてなんらかの「帰無仮説」を設定して、それを棄却できるか判断するのが「検定」

(教科書 p. 151)

1/2 の確率で当たるくじを８回ひいたとき、
すべて当たる確率は?
→ 統計的検定

確率不明のくじを８回ひいたところ、すべて当たりであった。このとき、当たりくじの確率はどれくらいだとかんがえるのが合理的か?
→ 区間推定

★ **「区間推定」と「統計的検定」の方法の間に本質的なちがいはない**

★ **区間推定のほうが直感的に理解しやすい**

★ **実際の計算は、区間推定のほうがむずかしいことが多い**

★ **慣習的に統計的検定を使うことが多い (分野によってちがうが)**

標本について計算できる統計量については、すべて統計的推測が可能である

(ただし、計算方法はさまざま)

【平均値の差の区間推定】

「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

【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

【統計的検定用語】

帰無仮説 (null hypothesis):
母集団における統計量が
この「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

危険率 (critical level): 1－信頼率

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 0 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 0 をふくむ
＝母集団においては平均値の差がない
という可能性を無視できない
→ 平均値の差があるとはいえない

【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「p 値」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 →
有意確率が 0.023 →
有意確率が 0.088 →

【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

【クロス表の独立性の検定】

V または $|\phi|$ の信頼区間にゼロ
(=独立の状態) がふくまれるかを判別する。

「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**

【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

【表の書きかた】

★ 検定の結果は表の下端の注釈に書く
★ 検定の対象になる統計量を必ず書く
★ $p < 0.05$ のように書くか、
統計量右肩にアステリスク (*) をつける
★ 有意でなければ $p > 0.05$ のように書くか、
統計量右肩に ^{ns} と書く (= not significant)

2010.7.20 現代日本論演習 (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.

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.

表 3　生活全般満足度の男女差 (1)

性別	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.　p < 0.05.

表 4　生活全般満足度の男女差 (2)

性別	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.

表 5　性別役割意識の男女差 (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.

表 6　性別役割意識の男女差 (2)

	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^{ns}.　ns: 5%水準で非有意.

無回答＝７.