URL: http://www.nik.sal.tohoku.ac.jp/~tsigeto/statg/2.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/現代日本論演習 II

学部 3 年生以上・大学院生対象：2003 年度後期
＜木２＞コンピュータ実習室 (文学部本館 7F 711-2)

授業の概要

授業内容

研究の現場で必要となる統計分析手法は、分析の目的とデータの特徴によってさまざまです。この授業の前半では、推測統計学の基本的な概念について解説し、統計的推定および検定の方法について学びます。後半では、さまざまな分析手法をとりあげて、それらの特徴と使い方を習得していきます。どのような分析手法をとりあげるかについては、受講者の関心と必要性を考慮します。統計解析パッケージ SPSS を使ってデータ分析の実習をおこないます。

履修要件

前期開講の 現代日本論演習 I 「統計分析の基礎」 / 比較現代日本論研究演習 I「統計分析入門」を履修済みであるか、それと同等の知識を習得済みであること。

テキスト

なし

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。

備考

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 (卒業論文／修士論文等で統計分析をおこなう予定の者を優先する)。

授業の予定

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ますが、 実際の授業の進行状況によって前後にずれることがあります。

---

1. 推測統計

- 誤差の対策 [10/2 提示資料 (PDF 形式) 130KB]
- 標本誤差の推定
- 平均値の点推定・区間推定
- 平均値の差の区間推定と t 検定
- 連関係数の区間推定と $\chi 2$ 検定
- サンプルサイズと検定力

2. 相関係数

- 尺度水準についての復習
- 散布図
- Kendall の順位相関係数
- Spearman の順位相関係数
- Pearson の積率相関係数
- 相関係数行列
- 欠損値の処理 (pairwise/listwise)

3. 中間試験

4. 変数をキーにした分析

- 個体間変動と変数間変動
- 対応のある分析
- 一致率の計算

5. 多変量解析

未定 (因子分析またはクラスター分析?)

6. 期末レポート

比較現代日本論研究演習/現代日本論演習 II (田中重人)

第 1 回「測定値と誤差」(2003.10.2)

1. 「真の値」と測定値
2. 誤差の種類と対策
3. 標本抽出のプロセス

【「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
  - 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：科学的原則論】

誤差はゼロにはならない。

→追試を通じた再現性のチェック

しかし実際には追試はめったに行われない
- 研究資源の問題
- 時間の問題

【現実的な対策】

誤差の発生原因と
その大きさについて推定・公表

→ 追試をおこなう人の助けになる

→ 追試がなくても誤差について見当がつく

【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)

★ 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

【無作為抽出について】

系統抽出、多段抽出、層化抽出...

比較現代日本論研究演習/現代日本論演習 II (田中重人)

第 2 回「無作為抽出と区間推定」(2003.10.09)

1. 中心極限定理
2. 平均値の区間推定

【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

【中心極限定理】

central limit theorem

★ 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)**

★ 標本サイズが大きいほど、母集団の平均とのずれが小さくなる

別紙の乱数表から、1 桁の数字を
10 個と 20 個抜き出して、
それぞれ平均値を求めてみよう

【平均値の信頼区間】

※「母集団では正規分布」の仮定が必要

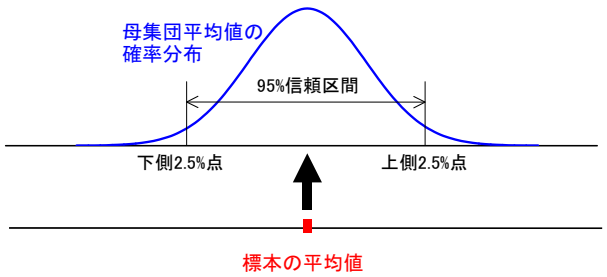
★ 標本の平均値が母集団平均値からはずれる確率は正規分布にしたがう

➨ 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



【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

★ 無限母集団からの標本の場合の
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　　t 臨界値　　$\frac{SD}{\sqrt{n}}$：標準誤差

比較現代日本論研究演習/現代日本論演習 II (田中重人)

第 3 回「統計的検定」(2003.10.23)

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**

【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

【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

※ 統計的検定の論理は本当はもっと複雑である。

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
＝母集団において平均値の差がない
という確率が 5%以上ある

【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】

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)

$\eta = 0.086$。$p > 0.05$。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)

$\eta = 0.198$*。

*: 5%水準で有意。

比較現代日本論研究演習/現代日本論演習 II (田中重人)

第 3 回「サンプルサイズの決定」(2003.10.30)

1. 検定力
2. ϕ 係数と%の差
3. ϕ 係数と χ^2 臨界値
4. サンプルサイズと検定力

【検定力】

power of test

母集団における一定の大きさの関連を
どれくらいの危険率で検出できるか

→ サンプル・サイズに依存

【ϕ 係数と%の差】

2×2 クロス表の%の差

＝周辺度数がバランスしていれば、
ϕ 係数に等しい

【ϕ 係数と χ^2 臨界値】

2×2 クロス表で独立性の検定が５%有意

$$\chi^2 = N\phi^2 > 3.84$$

【サンプルサイズと検定力】

ある%差を 5%水準で検出するのに
必要なサンプルサイズ：$N > 3.84/\phi^2$

20%差 → $3.84 / 0.2^2 \fallingdotseq 96$
16%差 →
14%差 →
12%差 →
10%差 →
5%差 →
1%差 →

【サンプルサイズの決定】

- 変数の測定法・分析法をきめる
- どの程度の強さの関連を検出できれば
  よいかを決める
- 必要なサンプルサイズを決める
- 分析のキーとなるカテゴリに均等分配
  した場合を最低限度とする

※不均等な配分を前提として厳密に求めることも可能

【その他の係数の場合】

Pearson の相関係数 → ϕ 係数とほぼおなじ

連関係数 V →χ^2 臨界値が自由度で変わる。
またカテゴリ数(少ない方)を考慮する。
一般に $N > \chi^2$ 臨界値 $/ (m - 1)V^2$

たとえば 3×3 クロス表なら

$$N > 9.49 / 2V^2$$

相関比 η → 次の式を使う (k はカテゴリ数)：

$$\frac{\eta^2}{1-\eta^2} \times \frac{N-k}{k} > F\text{臨界値}$$

※ k×2 クロス表の V 係数とほぼおなじ

※ 2 グループ間の平均比較ならΦ係数とほぼ同じ

順位相関係数類 → 後日

0. 尺度水準：復習
1. 尺度水準と分析法
2. 相関係数とは
3. 散布図
4. Goodman-Kruskal の γ と Kendall の τ_b
5. Pearson の r
6. Spearman の r_s

【尺度水準と分析法】

名義×名義 → クロス表

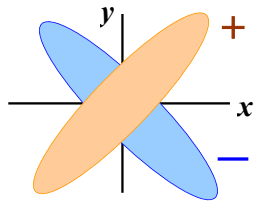
名義×間隔 → 分散分析・平均値の比較

順序×順序 → 順位相関係数
(rank correlation coefficient)
Goodman-Kruskal の γ
Kendall の τ_b
Spearman の r_s

間隔×間隔 → 積率相関係数
(product-moment correlation coefficient)
Pearson の r

【相関係数とは】

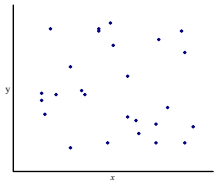
正(＋)の関係か、負(－)の関係か



－1～＋1 の範囲の値をとる：

- 無関連のときゼロ
- 完全な関連のとき±1

【散布図】



【ペア】

散布図上の任意の 2 点を直線で結んだとき

- 右上がり → Concordant
- 左上がり → Discordant

それぞれのペアの個数を C, D とする。

Goodman-Kruskal の $\gamma = \frac{C-D}{C+D}$

同順位ペアをうまく扱えないので、あまり使われない

【Kendall の順位相関係数】

Kendall の順位相関係数 $\tau_b = \frac{C-D}{\sqrt{KL}}$

K : x について同順位でないペア数
L : y について同順位でないペア数

同順位ペアがなければ、Goodman-Kruskal の γ と同じ

【変数の標準化】

(間隔尺度の場合)
平均＝0，標準偏差＝1 になるよう変換する。

$$X = \frac{x - \text{平均}}{\text{SD}}$$

これで単位を気にせずに比較できるようになる

【相関係数】

Pearson の積率相関係数

標準化済みの変数 X, Y について

$$r = \frac{XY\text{の総和}}{N}$$

単に「相関係数」といえばこの r をさす

欠点：はずれ値や歪みに弱い

【Spearman の順位相関係数】

r_s であらわす。

各変数を順位に変換した上で、
Pearson の積率相関係数を求める。

【相関係数類の使いわけ】

順序尺度の場合 → Kendall の τ_b
　　または Spearman の r_s

間隔尺度の場合
　正規分布なら → Pearson の r
　歪みや外れ値 → Spearman の r_s

【SPSS コマンド】

「相関」→「2 変量」
変数を指定する
相関係数の種類をチェック

Goodman-Kruskal の γ は出ない
（クロス表のオプションで出せる）

【相関係数行列】

3 変数以上について総当たりで出すこともできる(correlation matrix)

【欠損値の処理】

- 対単位 (pairwise) の除去
  個々の組合わせごとに欠損ケースを除く
- 表単位 (listwise) の除去
  分析に使う変数に**ひとつでも**欠損のあるケースを除く
  （**「オプション」で「リストごとに除去」をえらぶ**）

【文献】

池田 央 (編) (1989)『統計ガイドブック』新曜社

森敏明・吉田寿夫 (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。

2003.11.20 比較現代日本論研究演習／現代日本論演習 II

第 7 回「相関係数 (2)」(田中重人)

1. 相関係数の推定と検定
2. 相関係数行列の書きかた

【相関係数の推定と検定】

母集団において 2 変量正規分布のとき
r の 95%信頼区間は次式に近似：

$Z = \ln\frac{1+r}{1-r}$, $c = 2\times 1.96\sqrt{\frac{1}{N-3}}$ とおいて

上限：$\frac{\exp[Z+c]-1}{\exp[Z+c]+1}$

下限：$\frac{\exp[Z-c]-1}{\exp[Z-c]+1}$

この信頼区間に r=0 が含まれるかを
検定すればよい

信頼区間の求めるのが面倒なので、
通常は t 分布を利用した検定をおこなう (数表参照)。

相関係数の検定力 (5%水準)：

N=100 で r=±0.2

N=400 で r=±0.1

Spearman の順位相関係数 r_s も、
r と同じ方法で推定・検定できる。

Kendall の順位相関係数 τ_b の推定・検定は
別の方法を用いる (省略)。
r よりも検定力が低い

【相関係数行列の書きかた】

★ 線対称なので、
右上／左下の三角部分だけを書けばよい。

★ 小数第 3 位までが原則

★ 小数点の前につくゼロは省略してもよい

★ 検定の結果にしたがって * をつける

★ 小数点をそろえること

【文献】

Bohrnstedt, G. W. and Knoke, D. (1992)『社会統計学』(海野道郎・中村隆監訳、学生版) ハーベスト社。

森敏明・吉田寿夫 (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。

【課題】

５つ以上の変数を使って pairwise, listwise の
相関係数行列をそれぞれ出力し、整形して
印刷して提出

【来週の試験】

・試験範囲は、後期の授業開始から
今日までに習ったことすべて

・概念の説明／計算 (PC を使ってよい)

・何でも持ち込み可

比較現代日本論研究演習／ 現代日本論演習II (田中重人)
2003.11.20 資料

表1 順位相関係数行列 (listwise)

	変数名１	変数名２	変数名３	変数名４	変数名５	変数名６	変数名７
変数名２	.133						
変数名３	.203*	.200*					
変数名４	.054	.102	.076				
変数名５	.134	.186	.015	. 032			
変数名６	.110	.261*	－.002	.099	.319*		
変数名７	.195*	.132	－.124	.016	.185	－.165	
変数名８	.132	.205*	－.012	－.233*	－.022	.057	.084

Spearman の順位相関係数．*: p<0.05．N=105．

表2 順位相関係数行列 (pairwise)

	変数名１	変数名２	変数名３	変数名４	変数名５	変数名６	変数名７
変数名２	.133						
	(110)						
変数名３	.203*	.200*					
	(119)	(111)					
変数名４	.054	.102	.076				
	(120)	(110)	(116)				
変数名５	.134	.186	.015	. 032			
	(110)	(112)	(113)	(112)			
変数名６	.110	.261*	－.002	.099	.319*		
	(112)	(118)	(118)	(111)	(115)		
変数名７	.195*	.132	－.124	.016	.185	－.165	
	(110)	(118)	(118)	(116)	(110)	(115)	
変数名８	.132	.205*	－.012	－.233*	－.022	.057	.084
	(110)	(114)	(118)	(110)	(112)	(113)	(115)

Spearman の順位相関係数．*: p<0.05．()内は人数

小数点をそろえるのが大変。
「書式」→「セル」で表示形式を「文字列」にしておいて、「配置」とスペースで微調整する。

学籍番号：　　氏名：

比較現代日本論研究演習／現代日本論演習 II　(田中重人)

中間試験 (2003.11.27)

【回答上の注意】

① 小数の解答については、小数第 2 位まで書くこと
② 計算の問題の解答は、計算のプロセスがわかるように書くこと。
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

1. 「等確率標本」とはなにか。簡単に説明せよ。

2. 「計画標本」と「有効標本」のちがいについて簡単に説明せよ。

3. クロス表の独立性の検定をおこなう際に必要な前提を 2 つあげよ。

4. 平均値＝3.33, 標準偏差＝1.53, 標本数＝400 のとき、平均値の 95%信頼区間を求めよ。

5. x, y の値がつぎの組み合わせであるような 5 人の標本があるとする：
(1, 1) (2, 4) (3, 2) (4, 5) (5, 3)
(1) x, y それぞれの平均と SD を求めよ。

(2) Pearson の積率相関係数を求めよ。

学籍番号：　　氏名：

比較現代日本論研究演習／現代日本論演習 II　(田中重人)

中間試験 解答例 (2003.11.27)

1. 「等確率標本」とはなにか。簡単に説明せよ。

すべての個体がおなじ確率で選ばれるようにする抽出法（無作為抽出）で選ばれた標本

2. 「計画標本」と「有効標本」のちがいについて簡単に説明せよ。

調査対象として選ばれたすべての個体が「計画標本」。
そこから、実際の調査の過程で無効な標本となったものを除いたのが「有効標本」。

3. クロス表の独立性の検定をおこなう際に必要な前提を 2 つあげよ。

① 無作為抽出（等確率標本），
② すべてのセルの期待度数がじゅうぶん大きいこと（通常 5 以上）

4. 平均値＝3.33, 標準偏差＝1.53, 標本数＝400 のとき、平均値の 95%信頼区間を求めよ。

$3.33 \pm 1.96 \times 1.53/\sqrt{400} \fallingdotseq 3.33 \pm 0.15$

95%信頼区間： $3.18 \sim 3.48$

5. x, y の値がつぎの組み合わせであるような 5 人の標本があるとする：
(1, 1) (2, 4) (3, 2) (4, 5) (5, 3)
(1) x, y それぞれの平均と SD を求めよ。

x: 平均 ＝ (1+2+3+4+5)/5 = 3.00 ；
$5 \times SD^2 = (1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 = 10$
したがって $SD = \sqrt{(10/5)} = \sqrt{2} \fallingdotseq 1.41$
y: x とおなじ（平均=3.00; SD=1.41）　　**【不偏分散を使った別解もある】**

(2) Pearson の積率相関係数を求めよ。

各自の値から平均を引き、SD で割って標準化する:
(−1.41, −1.41) (−0.71, 0.71) (0, 0.71) (0.71, 1.41) (1.41, 0)
$5 \times r = 1.41 \times 1.41 - 0.71 \times 0.71 + 0 + 0.71 \times 1.41 + 0 = 2 - 0.5 + 1 = 2.5$
$r = 2.5/5 = 0.5$

2003.12.4 比較現代日本論研究演習／現代日本論演習 II

第 9 回「変数をキーにした分析」(田中重人)

1. 対応のあるケース
2. 散布図による表現
3. 対応のある平均値の差の推測

【対応のあるケース】

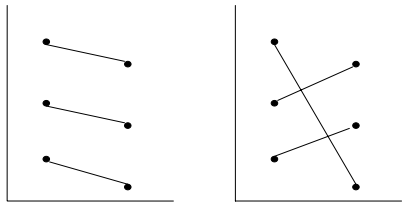
ふたつの変数のうち、どちらのほうが高いか

＝対応のあるケース

→変数をキーとした分析

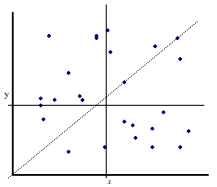
（実験の場合）被験者内要因か

被験者間要因か



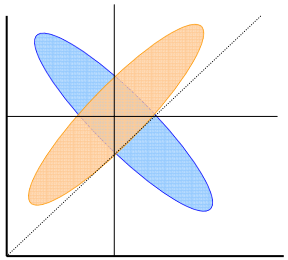
対応を考慮しないのはもったいない

【散布図による表現】



★ 平均値の差はどう表現されるか

★ 相関係数との関連



【記述の方法】

★ 平均・標準偏差だけでなく
相関係数も示す

★ できれば、散布図またはクロス表を示す

● **クロス表の書きかた：**

「分析」→「記述統計」→「クロス統計表」

「セル」で「パーセンテージ：全体」、

「統計」で「相関係数」をチェック

● **散布図の書きかた：**

「グラフ」→「散布図」→「単純」

「Y 軸」と「X 軸」の変数を指定

※ データエディタから必要な列を

Excel にコピーしてグラフを書く手もある

【平均値の差の統計的推測】

※ 平均値の差 ＝ 差の平均

つぎの式で標準誤差を求める：

$$\text{標準誤差} = \sqrt{\frac{SD_1^2 + SD_2^2 - 2rSD_1SD_2}{N-1}}$$

（ただし SD_1, SD_2 は各変数の標準偏差、r は相関係数）

対応のある平均値の差の 95%信頼区間：

$$d \pm 1.96 \times \text{標準誤差}$$

信頼区間の幅は、

- 人数が多いほど
- 標準偏差が小さいほど
- 相関係数が大きいほど

狭くなる。

この区間に 0 が含まれているか？

→対応のある t 検定

●**対応のある t 検定：**

「平均値の比較」→「対応のあるサンプルの T 検定」

２変数を選択してからでないとパレットに入れられない

【注意点】

対応のある分析は、**同一の尺度** で測られた変数同士でないと意味がない

2003.12.10 比較現代日本論研究演習／現代日本論研究演習 II

第 10 回「方向性の一致度」(田中重人)

1. 方向性の一致度
2. データの変容
3. 2 項検定

【平均値の比較の問題点】

★ 順序尺度の変数の比較は?

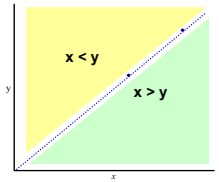
★ 2 項目間の一定の順序付け
(好き嫌い・適切さなど) が
どの程度共有されているかを
問題にしたい場合

【方向性の一致度】

2 変数 x, y の差の方向性は、
ケース中の何%で一致しているか

x > y
x = y
x < y

【散布図で考えると】



【差のとりかた】

新しい変数をつくる:
データエディタの「変換」→「計算」で

新変数名 ＝ 変数 x － 変数 y

「変換」→「値の再割り当て」→「同一の変数」で

負の値 ＝ －1
正の値 ＝ ＋1

この新変数について度数分布表を出せばよい

※ クロス表を見て、ケース数を確認すること

※「分析」→「ノンパラメトリック検定」
→「2 個の対応サンプルの検定」→「符号検定」
でも同様の表が出せる

【「一致度」の計算】

x>y のケース (または x<y のケース) の比率

★ 全ケース中の比率

★ x=y のケースを除いて、
差が出ているケースの中での比率

適当な基準 (例えば 80%) を超えているか？

$x=y$ **ケースを分析からのぞくには、**

「値の再割り当て」の際に

recode 変数名 (lowest thru -1=-1)
(0 = sysmis) (1 thru highest=1).

とする。

【統計的推測】

基準値（たとえば 80%）を上回っていても、
それが母集団に当てはまるかどうかは別問題

比率の標準誤差は、母集団での比率 a と
ケース数 N できまる：

$$\text{標準誤差} = \sqrt{\frac{a(1-a)}{N}}$$

($0.05<a<0.95$ かつ $N>30$ の場合の近似式)

a ± 1.96×標準誤差 ＝ 測定値
となる a を探せば、95%信頼区間が定まる。

→ a を適当な基準値 (たとえば 0.8) に設定
して、測定値が a＋1.96×標準誤差
をうわまわっているかを検定する

＝2 項検定

【SPSS のコマンド】

「ノンパラメトリック検定」→「2 項」

★ 再割り当てした新変数を指定
★ 「分割点」を指定 (ゼロ未満とゼロ以上に
分割したいなら、-0.1 などと指定)
★ 「検定比率」を指定 (上記の a)

2 項検定では、「分割点」以下の値を持つ
ケースの比率と「検定比率」とが比較される

→ **「分割点」以上の比率を検定するには、**
(1－基準値) を検定比率にする

検定比率＝0.5 のときは「符号検定」と同じ：

「分析」→「ノンパラメトリック検定」
→「2 個の対応サンプルの検定」

★ 元の変数の対を指定

★「符号検定」をチェック

(ただし x=y ケースがのぞかれる)

天井効果・床効果に注意
測定上の上限／下限に偏っている場合

2003.12.18 比較現代日本論研究演習／現代日本論演習 II

第 11 回「変数をキーにした分析：結果の表記」(田中重人)

【結果の書きかた】

クロス表 (または散布図) が基本:
各セルには度数と 全体での% を書く。

統計量などは表の下に:

対応のある t 検定 → 相関係数、平均値の差、
有意水準 （対応のある検定であることを明記）

2 項検定 → $x>y$ ケースと $x<y$ ケースの比率、
有意水準 (基準比率と検定法を明記)。
$x=y$ ケースを除いた場合はその旨明記。
50%基準なら単に「符号検定」と書いてもよい。

圧縮した書きかた:

対応のある t 検定 → 各変数の平均・SD の表
表の下に、人数、相関係数、平均値の差、
有意水準（対応のある検定であることを明記）

2 項検定→ $x>y$, $x=y$, $x<y$ 各ケースの比率の表
表の下に、有意水準 (基準比率と検定法を明記)

多数の変数を総当りで比較する場合：

Hasse diagram (ハッセ図)

★ $x>y$ か $x<y$ か「どちらともいえない」か

★ 上下関係に従って並べ、順位付け可能な組を線で結ぶ

★ 上から下に向かって線をたどれれば、
2 変数間に順序付け可能である

何を基準としたのかを明記すること

比較現代日本論研究演習II／現代日本論演習II （田中重人）
2003.12.18 参考資料

表1　自分にとって大切なこと

高い地位を得ること(x)	家族の信頼・尊敬を得ること (y) 1	2	3	4	合計
1. そう思う	13	1	0	1	15
	(5.4)	(0.4)	(0.0)	(0.4)	(6.3)
2. どちらかといえばそう思う	35	12	2	0	49
	(14.6)	(5.0)	(0.8)	(0.0)	(20.5)
3. どちらかといえばそう思わない	79	37	9	0	125
	(33.1)	(15.5)	(3.8)	(0.0)	(52.3)
4. そう思わない	32	15	3	0	50
	(13.4)	(6.3)	(1.3)	(0.0)	(20.9)
合計	159	65	14	1	239
	(66.5)	(27.2)	(5.9)	(0.4)	(100.0)

度数　（全体%） を示す。

平均値の差=1.48 (x=2.88, y=1.40), p<0.01 (対応のある t 検定による)。r=0.073。　対応のあるt検定の場合

x>yケース84.1%, x<yケース1.7%, p<0.01 (80%を基準とする2項検定)。　2項検定の場合

x>yケース84.1%, x<yケース1.7%, p<0.01 (80%を基準とする2項検定、x=yケースを除く)。　2項検定　(x=yケース除く) の場合

x>yケース84.1%, x<yケース1.7%, p<0.01 (符号検定)。　2項検定　(x=yケース除く、基準=50%) の場合

表2　自分にとって大切なこと

	平均	SD
高い地位を得ること	2.88	0.81
家族の信頼・尊敬を得ること	1.40	0.62

平均値の差=1.48, p<0.01 (対応のあるｔ検定による)。r=0.073。N=239。

表3　自分にとって大切なこと

	N	(%)
x>y	201	(84.1)
x=y	34	(13.6)
x<y	4	(1.7)
合計	239	(100.0)

x:　高い地位を得ること, y: 家族の信頼・尊敬を得ること。

p>0.05 (x>yケース80％を基準とする2項検定)。

2003.1.10 比較現代日本論研究演習／現代日本論演習 II

第 12 回「クラスター分析(1)」(田中量人)

1. 多変量解析
2. 類似度行列の並べ替え
3. クラスター分析の一般的手続き
4. グループ間平均連結法
5. デンドログラム

【多変量解析】

Multivariate analysis

3 つ以上の変数を同時にあつかう分析

- **因果分析型** (回帰分析／分散分析)

  因果関係を設定する…独立変数と従属変数

  (グループ別分析を洗練させたもの)

  →事前に統制できない変数の影響を事後的に排除

  →交互作用効果

- **類似関係型**

  「似ている」変数を見つける (全変数が同レベル)

  ・因子分析 (EFA/CFA)
  ・多次元尺度構成法 (MDS)
  ・林の数量化 (I 類～IV 類)
  ・クラスター分析

【類似度行列と距離行列】

相関係数＝変数間の類似度を表す

・その他、いろいろな類似度の係数がある

距離 ＝ 変数間の非類似度

【行列の並べ替え】

似ている変数を見つけるための簡便な手法

類似した変数が隣り合わせになるように
行と列を並べ替える (別紙参照)

【クラスター分析の手続き】

・データの準備 ・標準化

- **類似度 (距離) 行列を作成**
- **類似した変数同士を順次クラスター化する**
- **樹状図 (デンドログラム) を描く**

・ クラスター化による「ゆがみ」の評価

【グループ間平均連結法】

UPGMA

・いちばん「近い」変数同士を連結する

・連結してできた「クラスター」について
２変数の平均を代入して類似度を再計算

このステップを繰り返していって、最終的に
全変数が１クラスターになるまでつづける

【デンドログラム】

クラスター化の各ステップで、
どれだけの類似度のものを連結したか

→ 変数を適当に並べ替えて
「デンドログラム」を書く

【SPSS コマンド】

「分類」→「階層クラスタ」

- 「クラスタ対象」を「変数」に
- 「統計」オプションで
  「クラスタ凝集経過工程」「距離行列」をチェック
- 「作図」オプションで「デンドログラム」チェック、
  「つららプロット」を「なし」に
- 「方法」オプションで「測定方法」を
  「間隔」の「Pearson の相関」に

複数のコマンドが出力されるので注意

（類似度行列作成／クラスター分析／作業ファイル削除）

【今日の課題】

データファイル中の
ひとまとまりの変数群 (問 27 以外) について、
クラスター分析をおこなう

【参考文献】

古谷野 亘 (1988)『数学の苦手な人のための多変量解析ガイド』川島書店。

大野 高裕 (1998)『多変量解析入門』同友館。

三土 修平 (2001)『数学の要らない因子分析入門』日本評論社。

Romesburg, H. C. (1992)『実例クラスター分析』内田老鶴圃。

【期末レポート】

期限：2/4 (水) 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：相関係数、変数をキーにした分析、クラスター分析を使い、適当な分析をして結果を解釈する。いずれかの分析で、統計的推測をおこなうこと。

備考： SSM データのディスクをレポートと一緒に提出。
データのコピーをすべて消去すること。

2004.1.15　　比較現代日本論研究演習／現代日本論研究演習II　資料

相関係数行列

	評価高い職業	高い収入	高い学歴	家族の信頼尊敬	volunteer町内会	趣味サークル	多くの財産
評価高い職業							
高い収入	0.429						
高い学歴	0.524	0.426					
家族の信頼尊敬	0.209	0.174	0.144				
volunteer、町内会活動	0.202	0.172	0.238	0.424			
趣味サークル	0.190	0.130	0.213	0.146	0.413		
多くの財産	0.367	0.470	0.394	0.089	0.183	0.357	
高い地位	0.554	0.383	0.518	0.074	0.124	0.311	0.525

	評価高い職業	高い地位	高い学歴	高い収入	多くの財産	家族の信頼尊敬	volunteer町内会
評価高い職業							
高い地位	0.554						
高い学歴	0.524	0.518					
高い収入	0.429	0.383	0.426				
多くの財産	0.367	0.525	0.394	0.470			
家族の信頼尊敬	0.209	0.074	0.144	0.174	0.089		
volunteer、町内会活動	0.202	0.124	0.238	0.172	0.183	0.424	
趣味サークル	0.190	0.311	0.213	0.130	0.357	0.146	0.413

2004.1.22 比較現代日本論研究演習 II

第 13 回「クラスター分析(2)」(田中重人)

1. クラスター分析の選択肢
2. クラスター数の決定
3. 合成尺度と信頼性
4. R分析と Q 分析
5. 結果の書きかた

【クラスター分析の選択肢】

- **類似度 (距離) の選択**
- **標準化・係数の変換**
- **クラスター化の方法**

【SPSS コマンド】

「方法」オプション

- **クラスタ化の方法**
- **測定方法 ……変数の種類と類似度 (距離)**
- **値の変換**
- **測定方法の変換**

【類似度 (距離) のいろいろ】

間隔尺度の場合

- Pearson の相関 (説明省略)
  - ・もともと標準化された係数 (単位に無関係)
  - ・絶対値変換をするか？
- ユークリッド距離 ……多次元空間上の「距離」
  - ・変数の単位に依存する
  - ・あらかじめ標準化しておくか?
  - ・「係数の絶対値をとる」ことはできない

２値変数の場合

- **φ係数**
- **単純マッチング**
- **Jaccard の類似性係数**

順序尺度の場合

- **順位相関係数を使う （SPSS 対象外）**

【クラスター化の方法】

- グループ間平均連結法
- 最近隣法
- 最遠隣法
- Ward 法

【選択肢の選択基準】

- **理論的根拠**
- **先行研究との比較**
- **複数の方法を試してみること (結果の頑健性)**

【クラスター数の決定】

デンドログラムをどこで「切る」か

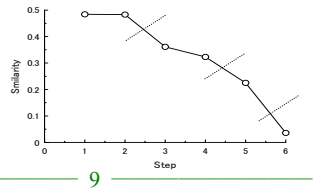
- **類似度の低いものを連結しないほうがよい**
- **クラスター数が少ないほどよい**

→クラスタをひとつ増やしたときの類似度の下がり具合を考える

【スクリープロット】

Scree plot

類似度（距離）の折れ線グラフ



解はひとつとは限らない。

→合成変数を作ってさらに分析を続ける場合は、信頼性を検討する

【信頼性】

類似度の低い変数同士を合成するのはまずい
→信頼性の問題

合成尺度の SD を利用して確認する
→ Cronbach の信頼性係数 α

「尺度」→「信頼性分析」で変数を指定。
「統計」オプションで「記述統計」の 3 項目をチェック

アルファ係数は大きいほど信頼性が高い。
大まかな目安：

- 0.8 以上 …… OK
- 0.6〜0.8 …… 要検討
- 0.6 以下 …… だめ

【合成変数】

データエディタの「変換」→「計算」で合成変数をつくる

※ 度数分布を確認すること

【R 分析とＱ分析】

- R分析 …… 変数を対象とする
- Q分析 …… ケースを対象とする

クラスター分析ではどちらも可能

SPSS では「クラスタ対象」で指定する

【表の書きかた】

- 類似度 (距離) 行列
  係数の種類、サンプル数を明記すること
- クラスタ凝集過程またはデンドログラム
  クラスタ化の方法も書くこと
  デンドログラムのスケールに注意

※ 紙幅が限られているときは、適宜省略してよい
※ SPSS のバージョンを記載しておくとよい