現代日本論演習／比較現代日本論研究演習 III

実践的統計分析

田中重人 (東北大学文学部准教授)

3 年生／大学院生対象：2018 年度後期 ＜木 2 ＞コンピュータ実習室 (文学部本館 7F 711-2)

1 授業の概要

(『講義概要』記載内容)

授業題目: 実践的統計分析法／応用統計分析

学習目標: さまざまな統計分析手法を理解し、使いこなせるようになる

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を使ってデータ分析の実習をおこないます。

履修要件: 1 学期／5 セメスタ開講の 比較現代日本論研究演習 I ／現代日本論演習「統計分析の基礎」を履修済みか、それと同等の知識を習得済みの者を対象とする。

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.

成績評価の方法: 授業中の課題と宿題 (70 %)、期末レポート (30%) を合計して評価する。

2 授業の予定

(1) 推測統計 (10/4〜10/18)
(2) 相関係数 (10/25〜11/15)
(3) 復習と進度確認 (11/22)
(4) 対応のあるデータの分析 (11/29〜12/6)
(5) 多変量解析 (12/13〜1/24) ※一般線型モデルを予定しているが、受講者の希望を受け付ける
(6) 期末レポート (2/5 提出期限) → 返却 (2/14 の予定)

3 ISTU への登録

ISTU https://istu3g.dc.tohoku.ac.jp の「比較現代日本論研究演習 III」に受講申請しておくこと。

4　復習事項

4.1　SPSSの操作

- データエディタにおける「変数ビュー」の使いかた
- 「欠損値」とは何か
- シンタックスとは何か
- 度数分布における「パーセント」と「有効パーセント」のちがい
- 変数値の再割り当ての方法
- グループに分割して集計する方法
- 中央値とパーセンタイルの求め方

4.2　クロス表

- 「行%」と「列%」の使い分け
- 「独立」とはどういう意味か
- 期待度数と残差の計算方法
- V と $\chi 2$ の計算方法
- クロス表をグラフにするときは、どのような種類のグラフが適切か

4.3　平均値

- 平均値を計算してよいのはどのような場合か
- 標準偏差の計算方法
- エフェクト・サイズと相関比の計算方法

4.4　その他

- 尺度水準とは何か。それはなぜ重要か
- 分析結果を表にするときの一般的な書式
- Excel による棒グラフ、度数ポリゴン、帯グラフ、誤差範囲つき折れ線グラフの書きかた
- ISTU を利用したレポート提出

5　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

Homepage:　http://tsigeto.info/officej.html

Blog:　http://b.tsigeto.info/school

オフィス・アワーは定めていない。質問等がある場合は、あらかじめ適当な時間に予約をとること。受講者への連絡は、文学部 2F 教務係前の掲示板または学務情報システムによる。

現代日本論演習／比較現代日本論研究演習 III「実践的統計分析」

第1講　推測統計の基礎

田中重人 (東北大学文学部准教授)

[テーマ]　推測統計の基礎; 母比率の区間推定

1　復習

- 記述統計と推測統計 (教科書 pp. 3–5)
- 母集団と標本
- 無作為抽出
- 区間推定と統計的検定の考えかた

2　標本比率 m はわかっているが母比率 M が不明の場合の区間推定

つぎのような情報 (= 標本統計量) から、母集団における統計量 (= 母比率) を推測する → 母比率はたぶん ○○ から ×× の範囲にある (区間推定)

> 袋のなかに色つきの玉がたくさん入っている。ここから8個取り出したところ、すべて赤であった。→袋のなかの玉のうち、赤玉の占める比率はどれくらいか?

この例題では、 m=1 であることがわかっているが、 M が不明である (n=8)。このとき、95%信頼区間を求めるには、M を適当に仮定し、その仮定の下で m=1 になる確率を計算することを繰り返す：

- もし $M = 0.9$ なら……
- もし $M = 0.8$ なら……
- もし $M =$ 　 なら……

このようにして、m=1 になる確率が **2.5%以上** である M の範囲を求める。(母集団は無限大の規模であると考えてよい。)

課題1:　解答を水曜正午までに ISTU に提出。プロセスがわかるように書くこと。

累乗 (0.9 の8乗など) を求めることが必要になる。Windowsの「電卓」ではメニューから［表示］→［関数電卓］に切り替えるとよい。Excelでは ^ という演算子が使える (掛け算を8回繰り返してもよい)。

3　もっと複雑な例

全世界から 400 人を無作為抽出してある意見を訊いたところ、「賛成」と答えた人が 240 人であった。このとき、母集団 (全世界の人々) における賛成の比率の 95%信頼区間を求めよ (欠損値はないものとする)。

原理的には上記とおなじやりかたで計算できるが、計算量が膨大になるので実際的でない。このような問いに答えるためには、「二項分布」(binomial distribution) の知識を利用する。

4　二項分布の簡単な例題

硬貨を 4 回投げて、そのうち表が出る回数 x を数える。

```
表=1, 裏=0 であらわすと
0 0 0 0 (x=0)
0 0 0 1 (x=1)
0 0 1 0 (x=1)
0 0 1 1 (x=2)
................
1 1 1 1 (x=4)
```

どれも等しい確率 (1/16) で起こるとすると、つぎのそれぞれの場合の確率が求められる：

表が 1 回も出ない（x=0）確率:
表が 1 回出る（x=1）確率:
表が 2 回出る（x=2）確率:
表が 3 回出る（x=3）確率:
表が 4 回出る（x=4）確率:

課題 2:　解答を水曜正午までに ISTU に提出。プロセスがわかるように書くこと。

参考資料

- Wikipedia の「二項分布」の項 <http://ja.wikipedia.org/wiki/二項分布 >
- 高校までの数学の教科書で、順列・組合せと確率・統計をあつかった部分

現代日本論演習／比較現代日本論研究演習 III「実践的統計分析」

第2講　正規分布の利用

田中重人 (東北大学文学部准教授)

[テーマ]　二項分布から正規分布へ、正規分布の性質、数表の利用、比率と平均値の区間推定

1　復習と宿題のポイント

- 無作為抽出とは / 区間推定の考えかた / 二項分布とは
- なぜ信頼率 0.95 に対して確率 0.025 が基準になるか

宿題 1 について：

- 95%信頼区間は 0.631 〜 1
- ということは、袋 (= 母集団) のうち 2/3 程度かそれ以上は赤玉だと考えてよい。(この推測はまちがいかもしれないが、その可能性は 5%未満である → 危険率)
- 信頼率は適当に決めている。信頼率 0.95 ＝ 危険率 0.05 にするのが通例だが、根拠は特にない。

宿題 2 について：

- 表裏の組合わせは 16。全部書いて考えてもよいし、「組合せ」(combination) 公式を利用してもよい。
- グラフをえがいてみると?

このような、一定の確率 (課題 2 の場合は確率 0.5) で偶然起こる出来事を n 回繰り返したとき、その出来事が起こる回数を理論的に予測した理論分布が「二項分布」(binomial distribution) である。

→　確率が 0.5 でない場合はどうなるか?

2　棄却域と採択域

理論分布:　一定の仮定の下での確率の分布を理論的に計算したもの

二項分布では、極端なケース (硬貨を 8 回投げて 6 回以上表、など) は起こる確率が低い。非常に確率が低いはずの極端な事象を観測したときは理論分布の仮定を疑う、というのが統計的推測の基本 (教科書 160 頁)。

(1)　「危険率」(α) を決める ($\alpha = 0.05$ にすることが多い →信頼率 0.95 に対応)

(2) 理論分布の上下の端から、確率が $\alpha/2$ を下回る領域を「棄却域」、それ以外の領域を「採択域」とする
(3) 棄却域と採択域との境界を「臨界値」という

区間推定の場合、「一定の仮定」を変化させながら、そのつど臨界値を計算し、実際の観測値と比較することになる。

3　正規分布

二項分布は、試行回数を増やすと、一定の形状に近づいていく (グラフを描くと、左右対称で真ん中にピークを持つなだらかな曲線になる)。試行回数が無限大 (∞) のときの二項分布のことを「正規分布」(normal distribution) という。

真ん中 (= 平均値) が 0 で標準偏差 (SD) が 1 になるように単位を調整して正規分布を描いたものを「標準正規分布」といい、$N(0,1)$ のようにあらわす。これを s 倍して m を足したものもやはり正規分布であり、$N(m,s)$ であらわす。

標準正規分布については、臨界値の表が用意されている (教科書巻末)。

例題:　標準正規分布の α =0.05 に対応する棄却域と採択域を教科書の数表から求めよ。

母比率の推測の場合、それほど比率が偏っていなくて ($0.1 < M < 0.9$)、サンプルサイズが大きければ ($n > 30$)、正規分布で近似できるものと考えて代用することが多い。通常、「比率の区間推定」といえば、この方法を指す。(実際には、平均値の区間推定 (後述) の方法で代用することが多い。)

母集団から無作為に n 人を抽出したところ、標本比率が m であった場合、母比率 M の 95%信頼区間はつぎの式で求められる：

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

この式の $\sqrt{\frac{m(1-m)}{n}}$ の部分を「標準誤差」(standard error) という。

臨界値 1.96 は危険率 0.05 に対するものである。この値は、危険率によって変わる (数表で調べる)。

例題:　標本規模 n=400 で標本比率 m=0.6 の場合、母比率 M の 95%信頼区間は?。

4　平均値の区間推定

値がいくつもある (たとえば 1–5) 変数の場合は?

→　すべての組合せについて理論分布を求めることは、事実上不可能

間隔尺度以上の変数の場合には、「母集団においては正規分布している」という仮定を置けば、平均値の区間推定が可能。つまり、標本における平均 m と標準偏差 s から、母集団における平均 M を推測する。この推測プロセスでは、母集団における平均と標準偏差の 2 つを推測しなければならないため、正規分布ではなく、t 分布 (Student's t distribution) を使う。

t 分布の性質 (教科書巻末参照)：

- かたちは標準正規分布に似ているが、正規分布より幅が広い
- 「自由度」(degree of freedom: DF) を持つ。これは標本規模によってきまる (df=$n-1$)。
- 自由度が大きくなると、標準正規分布に近づく (df>200 なら標準正規分布と同じと考えてよい)。

母平均の 95%信頼区間：

$$m \pm 臨界値 \frac{\mathrm{SD}}{\sqrt{n}} \tag{2}$$

臨界値は自由度と危険率によって変化する (数表で調べる)。標本規模200以上で信頼率95%なら、1.96と考えてよい。

5 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」オプション。「因子」を指定すると、グループ別に分析できる

6 課題1

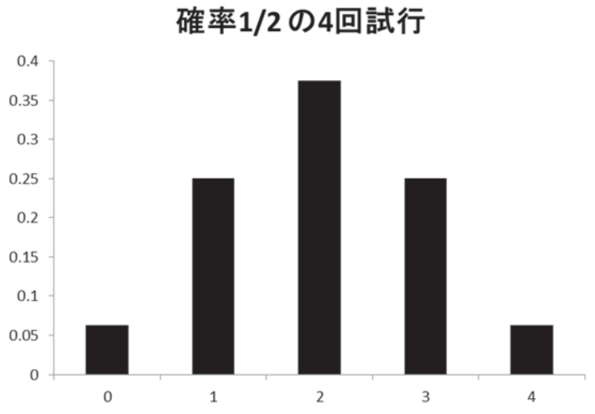
Wikipedia の「二項分布」の項 <http://ja.wikipedia.org/wiki/二項分布> と「正規分布」の項を読んで、これらの関係を理解する。

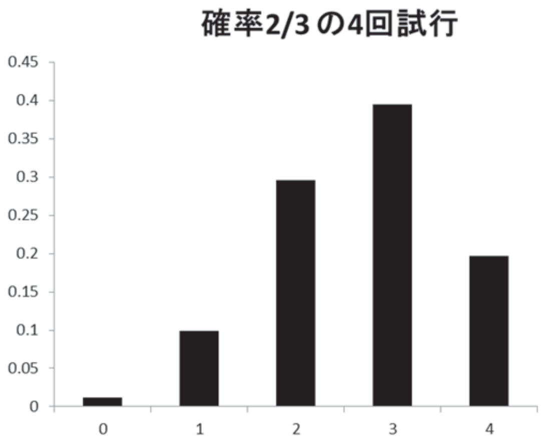
7 課題2

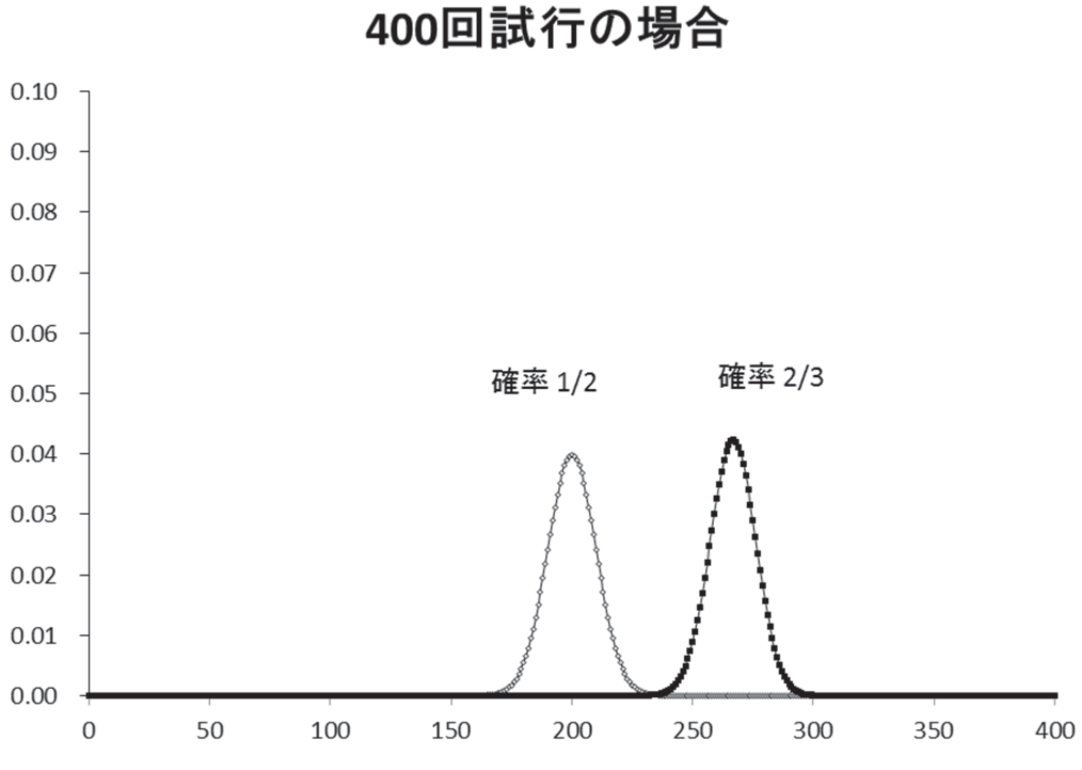
SPSS で、つぎのふたつの分析をおこなう

(1) 適当な変数について平均値の区間推定
(2) 同じ変数について、「因子」を指定して男女別の分析

これらの結果についてコメントをつけて提出 (ISTU で水曜正午まで)。他の人の意見をもらうこと（その人の名前を書く）


確率1/2 の4回試行
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1
2
3
4


確率2/3 の4回試行
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1
2
3
4


400回試行の場合
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
確率 1/2
確率 2/3
0
50
100
150
200
250
300
350
400

現代日本論演習／比較現代日本論研究演習 III「実践的統計分析」

第 3 講　統計的検定と検定力

田中重人 (東北大学文学部准教授)

[テーマ]　区間推定、統計的検定と検定力、サンプルサイズ

1　平均値の差の推定

ふたつのグループで別々に信頼区間を求めた場合：

- 信頼区間が重なっていなければ、差があると結論できる
- 信頼区間が重なっていれば、差があるかの判断は困難 (「同時分布」を考慮しなければならない)

通常は、「グループ間の平均値の差」について、母集団における値の信頼区間を求める方法をとる。→ 前期第 12 講

2　統計的検定

復習：

- 平均値の差の検定の方法 (数式と SPSS コマンド)
- 「臨界値」はどうやって計算するか
- 「有意確率」の解釈
- 「有意な差がある」「有意な差がない」ことの意味

3　カイ 2 乗分布と F 分布

推測統計手法で正規分布を使った推定・検定をおこなうことはあまり多くない。よく使うのは、正規分布を変形した t 分布、χ 2 分布、F 分布である。いずれも「自由度」(degree of freedom: DF) と呼ばれるパラメータを持ち、それによって形が変わる。

t **分布:**　DF をひとつもつ (DF ＝ ケース数 － 1)。正規分布に似た形をしているが、ちょっと幅が広い。自由度が増えると正規分布に接近していき、およそ DF>200 で標準正規分布とほぼ同じものになる。平均と分散の両方を推定・検定する場合に使う。

χ **2 分布:**　クロス表の独立性の検定で使う。DF によって形が変わる (DF は行・列のカテゴリ数からそれぞれ 1 を引いて求める)

F 分布: 分散分析 (η =0 を帰無仮説とした検定) で使う。DF をふたつもつ (カテゴリ数－1 と ケース数－1)

標準正規分布に従う変数の 2 乗は、DF = 1 の χ 2 分布に従う。
t 分布に従う変数の 2 乗は、第 1DF = 1 の F 分布に従う。

4 検定力

「検定力」(power of a statistical test) とは…… 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

→ 標本の規模 (＝ケース数) できまる
→ ○○ の差を危険率 xx% で検出するには、どれくらいのケース数が必要か?

信頼区間の幅がどれくらいになるかを、標本の規模を変化させて計算してみるとよい。

5 課題

検定力について、つぎの計算をせよ

- 母比率を ± 10% の精度で推定するにはどれくらいのケース数が必要か。± 5% なら?
- SD=1 である変数について、人数の等しいふたつのグループ間で平均値の差の区間推定をおこなう場合、信頼区間の幅を x 以下にするには、どれくらいのケース数が必要か。x の値を適当に設定して計算せよ。また、SD=0.5 の場合、SD=2 の場合はそれぞれどのようになるか。

文献

森敏昭・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.
永田靖 (2003)『サンプルサイズの決め方』(統計ライブラリー) 朝倉書店.

現代日本論演習／比較現代日本論研究演習 III「実践的統計分析」

第 4 講　順位相関係数

田中重人 (東北大学文学部准教授)

[テーマ]　順序尺度の相関を測る方法

1　前回課題について

※　「再」マークがついている人は再提出

比率の差については、第 2 講資料 の正規分布を利用した信頼区間の幅が x より狭くなる条件を求めればよい。

$$x > 2 \times 1.96\sqrt{\frac{m(1-m)}{n}} \tag{1}$$

$$n > 4 \times 3.84 \times \frac{m(1-m)}{x^2} \tag{2}$$

平均値の差については、前期第 11 講 の平均値の差の信頼区間の公式をあてはめればよい。n 人を半分ずつの人数 (=$n/2$ 人ずつ) にわけるとすると、95 %信頼区間の幅は

$$x > 2 \times 1.96 \times \mathrm{SD} \times \sqrt{\frac{1}{n/2} + \frac{1}{n/2}} \tag{3}$$

$$x > 4 \times 1.96 \times \frac{\mathrm{SD}}{\sqrt{n}} \tag{4}$$

$$n > 16 \times 3.84 \times \frac{\mathrm{SD}^2}{x^2} \tag{5}$$

なお、標本における平均値の差 d がこの信頼区間の幅の半分より大きいと、検定結果は 5%水準で有意になる。このための条件は、式 (4) より

$$d > 2 \times 1.96 \times \frac{\mathrm{SD}}{\sqrt{n}} \tag{6}$$

$$n > \left(2 \times 1.96 \times \frac{\mathrm{SD}}{d}\right)^2 = 15.37\frac{1}{\mathrm{ES}^2} \tag{7}$$

第 2 講資料 の「平均値の信頼区間」の幅が x より小さい、と考えて求めても同じ値になる。

※　これは簡略な求め方で、実際には、人数が均等に分かれていなかったり、自由度や併合 SD などの問題がある。さらに、本来は、**母集団における** 一定以上の大きさの平均値の差を検出するための条件を求めなければならないので、正確に計算するのは相当面倒である。永田 (2003) を参照。

調査規模の目安として、つぎのことをおぼえておくとよい (5%水準の場合)：

- 100 ケースで有意差がでるのは: 比率の 20%以上の差、エフェクト・サイズが 0.4 以上になるような平均値の差
- 400 ケースで有意差がでるのは: 比率の 10%以上の差、エフェクト・サイズが 0.2 以上になるような平均値の差

2　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 (ρ と書くこともある)
- 間隔×間隔 → 積率相関係数 (product-moment correlation coefficient)
    - Pearson の r

3　相関係数とは

ふたつの変数どうしが正 (＋) の関係にあるか、負 (−) の関係にあるかを、−1 〜 ＋1 の範囲の値であらわす。

- 無関連のときゼロ
- 完全な関連のとき±1

「相関図」(または「散布図」(scattergram) ともいう) を描いて考えるとよい (教科書 p. 75)。

4　順位相関係数

4.1　Pair

相関図上の任意の 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。

4.2　グッドマンとクラスカルの「ガンマ」係数

$$\text{Goodman-Kruskal's } \gamma = \frac{C-D}{C+D} \tag{8}$$

同順位ペアをうまく扱えないので、あまり使われない

4.3　ケンドールの順位相関係数 (タウ b)

- 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

$$\text{Kendall's } \tau_b = \frac{C-D}{\sqrt{KL}} \tag{9}$$

同順位ペアがなければ、Goodman-Kruskal の γ と同じ値になる。

4.4　SPSS コマンド

クロス表の「統計量」オプション →「Kendall のタウ b」を選択

5　課題

(x, y) の値がつぎの組み合わせであるような 6 人の標本があるとする：

(1, 2) (2, 4) (2, 4) (4, 3) (4, 5) (5, 5)

この標本について、Kendall の順位相関係数タウ b を求めよ。

6　次回予習

教科書の第 3 章、第 8 章 7 節を読んでおくこと。

文献

永田 靖 (2003)『サンプルサイズの決め方』朝倉書店.

現代日本論演習／比較現代日本論研究演習 III「実践的統計分析」

第5講　積率相関係数

田中重人 (東北大学文学部准教授)

[テーマ]　ピアソンの積率相関係数と相関係数の統計的検定

1　課題

SPSS の「クロス集計表」で、Kendall のタウ b がプラスになる表とマイナスになる表を出力し、クロス表の％を見て解釈する

2　積率相関係数類

2.1　変数の標準化

平均＝ 0, 標準偏差＝ 1 になるよう変換する。これで単位を気にせずに、変数同士の値を比較できるようになる

具体的には:　(その個体の値 − 平均値) / SD (→ 教科書 pp. 129, 130)

2.2　Pearson の積率相関係数

標準化済みの変数 X, Y について、それらの積の平均をとったもの：

$$r = \frac{\sum XY}{N} \tag{1}$$

通常、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

2.3　Spearman の順位相関係数

先に各変数を順位に変換しておく。あとの計算は、Pearson の積率相関係数とおなじ。
r_s または ρ (rho: ロー) であらわす。

2.4　SPSS コマンド

クロス表の「統計量」オプションで「相関係数」を選択。

3　相関係数類の使いわけ

- 順序尺度の場合: Kendall のタウ b または Spearman の ρ
- 間隔尺度の場合
  - 正規分布なら → Pearson の r
  - 歪みや外れ値 → Spearman の ρ

相関係数が 0 または ±1 になるのはどのような場合か?

- Goodman-Kruskal の γ :
- Kendall のタウ b:
- Pearson の r:
- Spearman の ρ :

4　相関係数の検定

Pearson の r の信頼区間は、「Fisher の z 変換」と呼ばれる方法で求められる (森・吉田 1990)。この信頼区間に r=0 が含まれるかを判断すれば、統計的検定がおこなえる。

ただし、この方法で正確に信頼区間を求めるのは面倒なので、通常は t 分布を利用した検定だけをおこなう (教科書巻末の数表参照)。Spearman の順位相関係数 ρ についても、おなじ方法が使える。

Kendall の順位相関係数タウ b についての推定・検定は別の方法を使う (Bohrnstedt and Knoke, 1992) が、省略。r に関する t 検定より検定力が低いことに注意。

5　予告

次々回の授業中に進度確認の課題をおこないます。範囲は、次回の授業内容まで。なんでも持込可 (ただしオンラインで何かを調べるのは禁止)。授業で使っているデータと調査票を持ってくること。

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。

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.

現代日本論演習／比較現代日本論研究演習 III「実践的統計分析」

第 6 講　相関係数行列の利用

田中重人 (東北大学文学部准教授)

[テーマ]　相関係数行列

1　相関係数行列

3 つ以上の変数について、総当たりで相関係数を並べた表を「相関係数行列」 (correlation matrix) という。

1.1　SPSS コマンド

- メニューの「分析」→「相関」→「2 変量」を選択
- 変数を指定する ／ 相関係数の種類をチェック

1.2　欠損値の処理

- 対単位 (pairwise) 除去: 個々の組合わせごとに欠損ケースを除く
- 表単位 (listwise) 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)

多変量解析の前段階の分析として相関係数行列を使うときは、必ず listwise 除去をおこなうこと。そうでない場合でも、係数どうしを比較するときには、listwise で欠損値をふくむケースを除去する (すべての係数について使っているケースを統一する) のが普通である。ただし、多くの変数を使った分析で listwise 除去をおこなうと、ケース数がかなり少なくなることがあるので注意。

この方法のどちらを取るかで結果が大きく違うとしたら、部分的に欠損値を持っているケースの挙動が特殊であることを意味する。その場合には、特定のケースで妙な回答パターンになっていないか、チェックすること。

1.3　相関係数行列の整形

- 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

2　課題

5 つ以上の変数をつかって相関係数行列を出力

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。

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.

比較現代日本論研究演習III / 現代日本論演習 (田中重人)

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。
スペースで微調整する。

現代日本論演習／比較現代日本論研究演習 III「実践的統計分析」

第7講　符号検定

田中重人 (東北大学文学部准教授)

[テーマ]　対応のある項目間の比較

1　進度確認課題結果

20 点満点 (4 + 6 + 4 + 6)

2　変数間の関連と比較

- 類似性・因果関係 → 相関係数行列
- どの変数がより高い／低いか? → 変数間の比較 (対応のある分析、被験者内要因)
  - 例：10 歳から 20 歳の間の身長の変化
  - 例：授業の前後での知識の変化
  - 例：「見れる」と「起きれる」ではどちらのほうが受容度が高いか?
  - 例：問 27 の 8 項目のうち、最も「重要」と評価されているもの／されていないものはどれか?

とりあえず、平均値を並べて比較するには：

SPSS では「分析」→「記述統計」→「記述統計」で多くの変数の平均値 (と標準偏差) を並べて出力できる。

3　「対応」とは

同一ケースが複数の項目に回答している場合、**項目間に「対応」がある** という。

このような項目の比較には、対応を考慮した分析法を使う必要がある。

例題: つぎのような 3ケース×2変数 のデータについて、どちらの変数が大きくなる傾向にあるかを考えてみよう：

	変数 A		変数 B		差
ケース X	4	-	5	=	
ケース Y	3	-	4	=	
ケース Z	5	-	1	=	

平均値

多数決をとると?

このように、ケース間の**異質性が大きい** 場合は、対応を考慮して分析しないと、データの特徴をつかみそこねる可能性がある。

4　ふたつの変数での大小の比較

4.1　例題

「B. 高い収入」 vs. 「D. 家族からの信頼・尊敬」 …… どちらが大切?

→　B>D の人と B<D の人のどちらが多いか?

これら 2 つ以外に B=D の人がいるから、これをどうあつかうかが問題になる。

4.2　符号検定

「符号検定」(sign test) とは…

1. A = B のケースを除外
2. 帰無仮説:「母集団では同数」(=50%)
3. 正規分布を利用して検定をおこなう

※　この考えかたは、比率の区間推定とおなじものなので、比率 ± 臨界値×標準誤差 で信頼区間を求めていると考えてもよい

4.3　SPSS コマンド

- 「ノンパラメトリック検定」→「過去のダイアログ」→「2 個の対応サンプルの検定」
- 比較したい変数をペアで指定する
- 「符号検定」をチェック

5　課題

適当な 2 変数について、符号検定を行う。クロス表 (または相関図) も出力して、結果を解釈すること。ISTU で来週水曜正午まで。

参考 URL

http://aoki2.si.gunma-u.ac.jp/lecture/Average/sign-test.html 青木繁伸：符号検定の簡潔な説明

http://kusuri-jouhou.com/statistics/fugou.html 役に立つ薬の情報〜専門薬学: 架空データによる符号検定の説明：小標本／大標本

現代日本論演習／比較現代日本論研究演習 III「実践的統計分析」

第8講　対応のある平均値の差の検定

田中重人 (東北大学文学部准教授)

[**テーマ**]　二項検定; 対応のあるサンプルの平均値の差の分析

1　前回課題について

- 順序尺度以上の変数であって、**同一の測定方式** の変数の組み合わせを選ぶ
- 符号検定は個々の対応ケースにおいて値の比較をおこなうので、そのことがわかる文章表現にする
- 回答選択肢の値の順序によっては、解釈を勘違いしやすいので、クロス表を見てよく考えること
- 検定統計量 (Z) は、正規分布を用いた比率の検定
- この種の分析の場合は、クロス集計表のセルに「全体」のパーセントを入れるとよい

2　二項検定

「符号検定」は、比率を使った検定の一種である

- $\mathrm{x} = \mathrm{y}$ のケースを除外しないで比率を求めることもできる
- 帰無仮説は、比率 =0.5 でなくても、任意の比率を設定できる

このように、さまざまな比率の計算方法と帰無仮説の設定で統計的検定をおこなう一群の方法を「二項検定」と呼ぶ。

SPSS では、あらかじめ 2 つの変数の差を求め (下記参照)、「変数値の再割り当て」で 2 値変数を作って「分析」→「ノンパラメトリック検定」→「過去のダイアログ」→「2 項」を使う。

```
compute DIFF = X - Y.
recode DIFF (lowest thru 0 = -1) (1 thru highest = 1) (missing=sysmis) into SIGN.
```

3　差の平均値の統計的推測

二つの変数の差について新たな変数を作ってみる：

- 「変換」→「計算」
- 「目標変数」に適当な名前を
- 数式を作成 (新変数 ＝ 変数 x − 変数 y)
- シンタックス貼付、実行
- 度数分布 (「統計」オプションで平均、分散、SD、標準誤差を出力)

上で求めた「ふたつの変数間の差」の平均値について、信頼区間を求めるにはどうすればよいか?

4　平均値の差の統計的推測

二つの変数 x と y の平均値は母集団においてはどちらのほうが高いか?

→　差の変数 $x - y$ を作って、その平均値について区間推定するのと同じことになる

実際には、x と y の平均値、標準偏差とそれらの間の Pearson の積率相関係数 r を使って計算できるので、そのやりかたがふつうつかわれる。

[**課題**]　教科書 p. 192–197 の説明を読み、「対応」の有無によって計算方法がどのように変わるかを考える

5　SPSS コマンド

- 「平均値の比較」→「対応のあるサンプルの t 検定」
- 2 変数の組を選択してパレットに入れる

6　宿題

適当な変数について、SPSS で次の二つの分析を行い、結果が同じになることを確かめる (ISTU で水曜正午まで)

- 差の変数を作って、その平均値の区間推定をおこなう
- 「対応のある」t 検定をおこなう

7　今後の予定

次回以降は「多変量解析」に入ります。とりあげてほしい多変量解析手法がある場合は、田中まで。(希望がなければ、重回帰分析をとりあげます。)

現代日本論演習／比較現代日本論研究演習 III「実践的統計分析」

第 9 講　多変量解析入門

田中重人 (東北大学文学部准教授)

[テーマ]　多変量解析の種類と、因子分析の基本的な考えかた

1　前回課題について

「差」の変数をつくったときの「平均」「標準誤差」の意味。

対応のあるデータの場合、平均値の差の信頼区間を求める際の数式の標準誤差 (standard error) を、相関係数を用いて調整する。この点が、通常の (対応のない) 平均値の場合と異なる。

→　相関図を描いて考えてみよう

2　対応のある分析について: 結果の書きかた

2.1　個々の結果を表示する十分なスペースがある場合

クロス表 (または相関図) をいちいち示すのが基本 (別紙参照)。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書く。必要な統計量は分析法によって違うので注意。

- 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）
- 符号検定 → x>y ケースと x<y ケースの比率、有意水準

2.2　スペースがあまりない場合

対応のある t 検定であれば、各変数の平均と SD の表をのせる。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）を書く。

符号検定であれば、x>y, x=y, x<y 各ケースの比率の表をのせる。表の下に、有意水準 (符号検定であることを明記) を書く。

2.3　多数の変数間の関連を示す場合

ハッセ図 (Hasse diagram) が使える。平均値などの高い順に変数を並べ、有意な差がある変数どうしを線でむすぶ。具体例は別紙参照。

3　多変量解析とは

3つ以上の変数をつかう分析法を「多変量解析」(multivariate analysis) という。次の2種類に分けられる (大野, 1998, p.48-56)。

- 類似関係型: 似た変数同士をまとめたり、潜在因子を取り出したりするもの。因子分析, クラスター分析など
- 因果関係型: 原因と結果の関係を追究するもの。回帰分析, 分散分析, 一般線型モデルなど

この授業では前者をあつかう。

4　課題

相関係数行列を出力して、「似ている」変数のグループを探す。

文献

大野高裕 (1998)『多変量解析入門』同友館.

三土修平 (1997)『初歩からの多変量統計』日本評論社.

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)				合計
	1	2	3	4	
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<0.05 (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<0.05 (符号検定)。

符号検定の場合

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<0.05 (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 (符号検定)。

現代日本論演習／比較現代日本論研究演習 III「実践的統計分析」

第10講　因子分析の基礎

田中重人 (東北大学文学部准教授)

[テーマ]　「主成分」の抽出法

1　前回課題について

問 27 の相関係数行列から、8 つの変数が 2 つのグループにわけられることを確認

2　主成分 (principal component)

2 変数の分布を「うまく説明する」直線を定める方法

(1) 変数を標準化しておく
(2) 散布図 (相関図) を描く
(3) 各点からの距離の 2 乗の合計がいちばん小さくなるような直線を引く (＝その直線上に投影した値の SD が最大になる)

この直線を「第 1 主成分」という。第 1 主成分に直交する直線を「第 2 主成分」という。

※　これは結局、元の座標軸を回転させるのとおなじ
→　3 変数以上の場合も、同様にして、変数の個数と同じだけの主成分を抽出できる。

3　固有値

主成分の分散 (SD の 2 乗) のことを固有値 (eigenvalue) という

- 固有値の最大値は変数の個数
- 固有値を変数の個数で割った値を「寄与率」という

通常、固有値が 1 未満の主成分は無視して、1 以上の主成分だけで解釈を考える。

4　SPSS コマンド

「次元分解」→「因子分析」で変数を指定したうえで、つぎのオプションを設定する。

記述統計:　1 変量の記述統計量、初期の解、「相関行列」の「係数」にチェック
回転:　「バリマックス」を選択
オプション:　「サイズによる並び替え」にチェック

出力から読みとる情報

- 人数
- 「共通性」(因子抽出後) ＝ 主成分と元の変数の相関の 2 乗和
- 「初期の固有値」の値と累積寄与率 (%)
- 回転後の成分行列 (負荷量)

文献

大野 高裕 (1998)『多変量解析入門』同友館.
高橋 信 (2006)『マンガでわかる統計［因子分析編］』オーム社.

現代日本論演習／比較現代日本論研究演習 III「実践的統計分析」

第11講　因子分析のいろいろ

田中重人 (東北大学文学部准教授)

[テーマ] 因子分析におけるさまざまな方法とその選択

1　因子分析の考えかた

因子分析においては、観測された変数群の背後には、直接観測できない潜在的な「因子」(factor) があると仮定する。変数間の相関関係からその背後の潜在的因子を推測するのが、因子分析である。

心理学でよく使われる。具体的には、能力や性格の研究など。

例: 子供の「学力」(= 因子) を、複数の試験の点数 (観測変数) から推測する

因子分析にはさまざまな方法が混在していて、どれを選択すればいいかの一律の基準は存在しない。一般に、「色々やってみていちばんよさそうなものを選ぶ」というやりかたがとられている。また、統計的検定などをふつうは行わないという特徴がある。

2　因子構造の決定

2.1　因子抽出法

主成分法 (principal component)、主因子法 (principal factor) などがよく使われる。

主成分法については前回資料参照。

主因子法は、主成分法と同様のことを最初におこなったあと、繰り返し計算で、一定の基準を満たすところまで解を変形する方法である。計算の細かいところで、いろんなパラメータを指定できる。主成分法との結果の違いは、**2 つ以上の変数に大きな負荷量をもつ因子だけが最終的に残る** というところ。

2.2　回転法

- 直交回転 (Varimax など): 因子同士の相関がゼロ
- 斜交回転 (Promax など): 因子同士の相関あり

社交回転では、因子負荷量の解釈は「パターン行列」をみておこなう。また、因子間相関の表も出てくるので、それも確認すること。

2.3　因子数の決定

因子の数をいくつに設定するかは、分析結果におおきな影響をあたえる。通常は、**固有値が 1 以上の主成分の数** 以下にすることを大前提として、その範囲内で、いろいろ試してみて、「良い」ものを選ぶ。

- 累積寄与率の高さ
- 回転後の因子の解釈 (因子負荷量をみる)

累積寄与率の伸びかたをみて、伸びが鈍るところ (よく「肘」と呼ばれる) までを採用する、という基準が補助的に用いられることがある (→ scree plot)。

2.4　変数の取捨選択

どの因子にも負荷量の大きくない変数は、結果に影響をほとんどあたえていないので、分析から削除することが多い。抽出した因子によって各変数がどれくらい説明されているかは「共通性」(communality) でわかるので、この値が低い (たとえば 0.3 未満) の変数は削除するなど。

3　SPSS での因子分析

メニューの「分析」から「次元分解」→「因子分析」をえらぶ。変数を指定したうえで、つぎのオプションを設定する。

記述統計:　1 変量の記述統計量、初期の解、「相関行列」の「係数」にチェック

因子抽出:　「方法」を指定。「抽出の基準」で因子数を決める (デフォルトでは「固有値が 1 以上」になっている)

回転:　「バリマックス」「プロマックス」などを選択

オプション:　「サイズによる並び替え」にチェック

4　課題

5 項目以上の意識セット変数を使って、主成分法と主因子法による因子分析を、バリマックス回転とプロマックス回転の両方でおこない、結果がどのように変わるかを考察する。

5　期末レポート

期限:　2/5 (火)

提出先:　ISTU

内容:　相関係数、対応のある分析、多変量解析について、それぞれ適当な分析をして結果を解釈する。相関係数と対応のある分析については、推定または検定の結果をつける。データは何を使ってもよいが、SSM データ以外のものを使うときはデータについての説明をつけること。

備考:　レポート提出後に、SSM データのコピーをすべて消去すること。レポートは、 採点後に返却する 。

現代日本論演習／比較現代日本論研究演習 III「実践的統計分析」

第12講　因子得点

田中重人 (東北大学文学部准教授)

[テーマ]　因子得点を別の分析に使う

1　前回課題について

- 結果がよくわからないときは、相関係数行列を見て考えるとよい
- はっきりと相関係数の高い変数の「かたまり」があれば、因子分析の手法による差はあまり出てこない →心理尺度の構成
- プロマックス回転など斜交解を求めるときは、因子間相関に注意
- 共通性の低い変数は分析から除くことを考える
- 主因子法のほうが主成分法より「厳しい」
- 累積寄与率がある程度 (50%以上?) 高い場合は、因子数を減らしてやってみる

2　因子得点

因子分析の結果をもとに、個々のケースについて、因子の値を計算したものを「因子得点」(factor score) という。この値を変数として保存して、ほかの分析に使うことができる。

SPSS では、「得点」オプションで「変数として保存」をえらぶ。方法は「回帰法」(デフォルト) のままでよい。変数名は指定できない (SPSS が自動的につける) ので、必要ならデータエディタを開いて名前を変更する。

因子得点は、

(1) 標準化した変数値に
(2) 負荷量をもとに算出した係数をかけて
(3) それらを足し合わせる

という手続きで求めている。SPSS では、「得点」オプションで「因子得点係数行列を表示」をえらぶと、(2) の係数がわかる

因子得点の平均は 0 であり、各変数との相関が因子負荷量と一致する。

因子得点同士の相関係数は、直交回転 (バリマックスなど) の場合はゼロであるが、斜交回転 (プロマックスなど) の場合は因子間相関に等しい。

3　課題

適当な変数群について因子得点を算出して、それと性別・年齢との関係を分析する。分析結果の出力と、その解釈を書いて提出。