

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：2016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 (10/6～10/20)
- (2) 相関係数 (10/27～11/17)
- (3) 復習と進度確認 (11/24)
- (4) 対応のあるデータの分析 (12/1～12/8)
- (5) 多変量解析 (12/15～1/26) ※重回帰分析を予定しているが、受講者の希望を受け付ける
- (6) 期末レポート (2/6 提出期限) → 返却

3 ISTU への登録

ISTU <http://www.istu.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
- ϕ , V , χ^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://www.sal.thoku.ac.jp/~tsigeto/welcomej.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、基本的に、授業においてまたは文学部 2F 教務係前の掲示板においておこなう。ただし、休講などで緊急を要する連絡は、田中の個人ブログ (School カテゴリの記事) に掲載することがある。 <http://tsigeto.info/newsj.html>

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

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

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%以下である → 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：

- 表裏の組み合わせは 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 $\alpha = 0.05$ 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:

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

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:

$$m \pm \text{臨界値} \frac{SD}{\sqrt{n}} \quad (2)$$

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<<http://ja.wikipedia.org/wiki/正規分布>>

7 課題 2

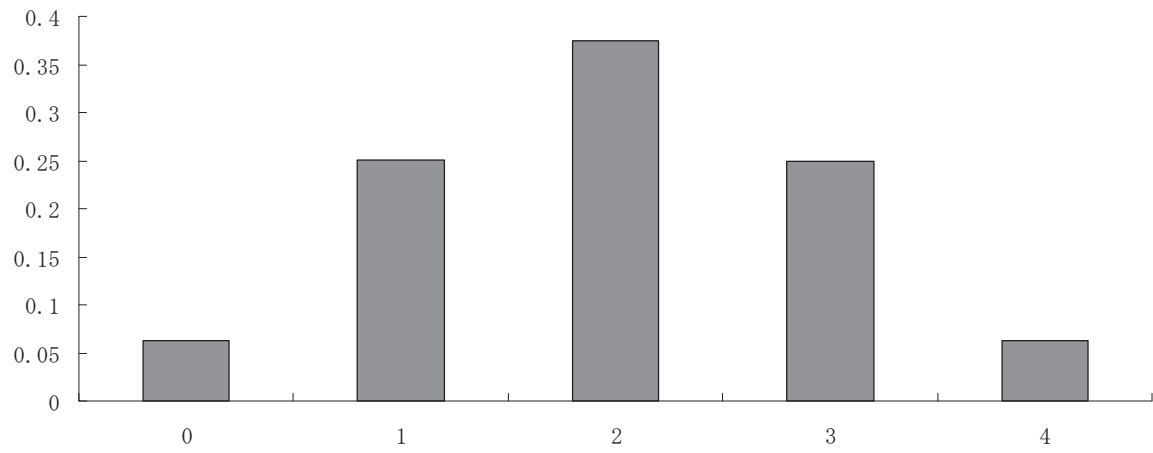
SPSS で、つぎのふたつの分析をおこなう

- (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

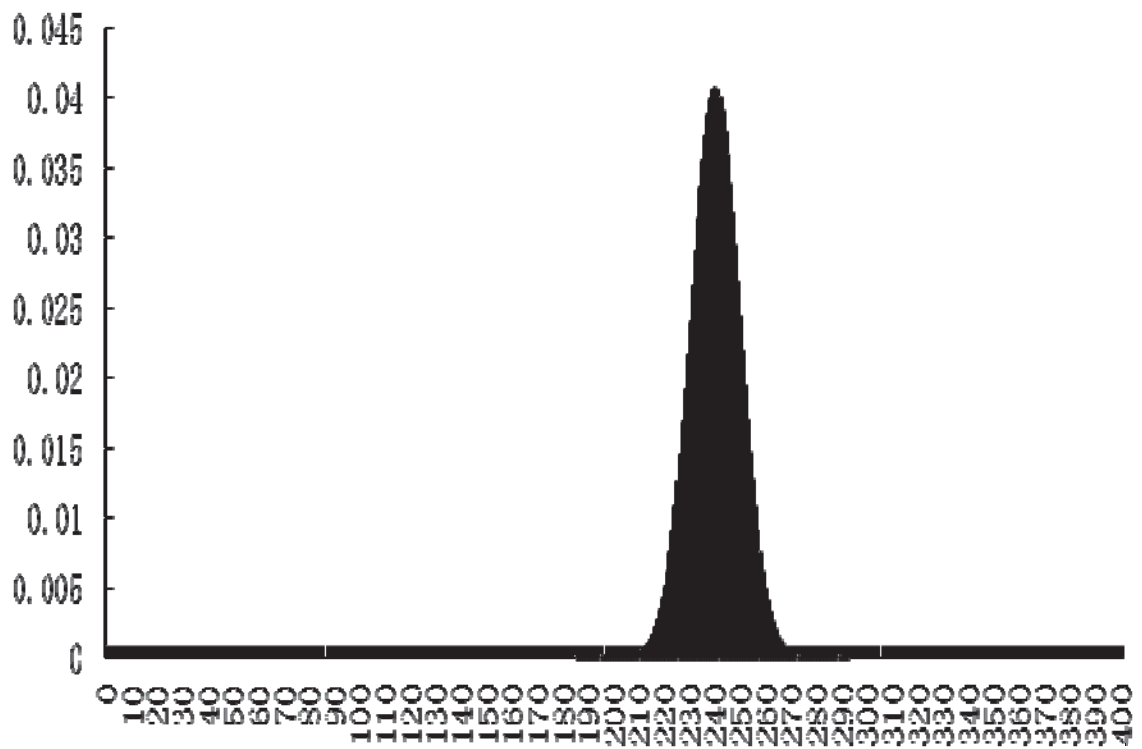
これらの結果についてコメントをつけて提出 (ISTU で水曜正午まで)。他の人の意見をもらうこと (その人の名前を書く)

なお、前回課題未提出または内容が間違っていた人は、それもやり直して提出すること。ファイルは今回の課題とおなじでよい。

二項分布 ($p=0.5, n=4$)



二項分布 ($p=0.6, n=400$)



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 ($DF = \text{ケース数} - 1$)。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 ($\eta = 0$ 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 $\pm 10\%$ の精度で推定するにはどれくらいのケース数が必要か。 $\pm 5\%$ 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なようになるか。

6 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.

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

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。

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

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

平均値の差については、前期第12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

$$x > 2 \times 1.96 \times \text{SD} \times \sqrt{\frac{1}{n/2} + \frac{1}{n/2}} \quad (3)$$

$$x > 4 \times 1.96 \times \frac{\text{SD}}{\sqrt{n}} \quad (4)$$

$$n > 16 \times 3.84 \times \frac{\text{SD}^2}{x^2} \quad (5)$$

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

$$d > 2 \times 1.96 \times \frac{\text{SD}}{\sqrt{n}} \quad (6)$$

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

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：

- 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、 $-1 \sim +1$ 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

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

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

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

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.

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:

$$r = \frac{\sum XY}{N} \quad (1)$$

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。

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: 個々の組み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ふとしたら、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。

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 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。

2 「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。

3 ふたつの変数での大小の比較

3.1 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。

3.2 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 \pm 臨界値 \times 標準誤差 で信頼区間を求めていると考えてもよい

3.3 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

4 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

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こと
- 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

- $x = y$ 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」に適切な名前を
- 数式を作成 (compute 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。)

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、重回帰分析の基本的な考えかた

1 前回課題について

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

この授業では後者だけをあつかう。

4 用語

従属変数 (dependent variable): 結果になる変数のこと。通常、ひとつの分析についてひとつだけ。「目的変数」「被説明変数」ということもある。

独立変数 (independent variables): 原因になる変数のこと。ひとつの分析に複数あってよい。「説明変数」ということもある。

従属変数と独立変数は、しばしば Y と X であらわされる

5 課題 1

Q39g (……指導者や専門家……) と Q1.1a (満年齢), Q6.1(学歴) の関連を確認し、これらの間の因果関係についてどのようなことがいえそうかを考える。

ただし、学歴の変数は次のように 3 分割すること：初等 (1,2,12); 中等 (3-5,13); 高等 (その他)

6 第 3 変数の統制 (control)

複数の要因が影響を与えていると想像される場合、因果関係を確定するには、ある変数の効果を「一定に保った」状態をつくったうえで、別の変数の効果を推定する必要がある。

実験の場合: 被験者の割り当ての時点で統制する (無作為割り当てなど)

観察の場合: 分析の段階で、多変量解析をおこなう

たとえば、データセットを学歴で 3 分割して、年齢と Q39g との相関分析を行ってみるだけでも、かなりのことがわかる。このような発想を洗練させたものが多変量解析である。

7 一般線型モデル

SPSS で、「分析」→「一般線型モデル」→「1 変量」を選択。

変数は次のように指定する

- 従属変数 = Q39g
- 固定因子 = 学歴 (3 分割)
- 共変量 = 満年齢

「オプション」で「記述統計」「パラメータ推定値」を指定

8 結果の読みかた

Q39g の値は、つぎの式で近似できることになる

初等教育: $Q39g = \text{切片} + B1X1 + B2$

中等教育: $Q39g = \text{切片} + B1X1 + B3$

高等教育: $Q39g = \text{切片} + B1X1$

9 課題 2

上記の一般線型モデルを変形して、年齢だけ、学歴だけを独立変数とする分析をそれぞれおこなう。結果出力のパラメータ推定値を比較して、なにがわかるかを考察。水曜正午までに ISTU に提出。

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。

第10講 一般線型モデル

田中重人 (東北大学文学部准教授)

[テーマ] 固定因子と共変量

1 前回課題について

年齢と Q39G は間隔尺度と考えてよいが、学歴（3段階）は順序尺度。

- 年齢 × Q39G → 相関係数 (Pearson / Spearman / Kendall)
- 学歴 × Q39G → クロス表、平均値の比較、または順位相関係数 (Pearson の積率相関係数は使えない)
- 学歴 × 年齢 → クロス表 (年齢を適当にカテゴリにわけ)、平均値の比較、または順位相関係数 (Pearson の積率相関係数は使えない)

クロス表には適切な%を表示すること。

2 モデルとパラメータ

前回の一般線形モデルの推定結果では、Q39g の値が次の式で近似されていることになる：

$$Q39g = \text{切片} + B_1X_1 + B_2X_2 + B_3X_3 \quad (1)$$

ただし、

- X_1 は年齢
- X_2 は初等教育のものについて1、それ以外は0とする
- X_3 は中教育のものについて1、それ以外は0とする

推定された係数 (切片とB) それぞれについて、区間推定と統計的検定がおこなわれる

3 固定因子と共変量

固定因子: 名義尺度の変数。自動的にカテゴリに分割され、そのうちひとつが「基準」になる。推定される係数は、カテゴリ数-1。

共変量: 間隔尺度の変数。そのままの値が投入される。推定される係数はひとつだけ。

3.1 固定因子ひとつだけのモデル

カテゴリ別平均から係数が計算される

初等: $3.591 - 0.700 = 2.891$

中等: $3.591 + 0.011 = 3.602$

高等: $3.591 + 0.000 = 3.591 \leftarrow$ 基準

「分散分析表」に表示されるものは、平均値の比較の際に使われるものと同様であるが、用語が少し違う：

- 決定係数 $R^2 =$ 相関比 η の 2 乗 $=$ edu3 / 修正総和
- edu3 + 誤差 $=$ 修正総和

おなじ変数について平均値の比較をおこない、結果を照らし合わせてみよう。

3.2 共変量ひとつのモデル

最小2乗法 (least square method) で係数を求める。これは、適当な直線 $A + BX$ によって Y の値を近似する方法であり、 Y と $A+BX$ とのずれの大きさを評価するために差の2乗和をとる。この2乗和 $\sum(Y - A - BX)^2$ が最小になるように A と B の組み合わせを求める。

回帰係数 B の意味: X が 1 単位増えたとき Y がどれだけ増えるか

3.3 独立変数が複数の場合

- 独立変数がひとつの場合と何が変わるか?
- 「コントロール」することの意味
 - 媒介効果
 - 疑似相関
- 分散分析表から独立変数の影響力の大きさを読む

4 期末レポート

期限: 2/6 (月)

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。すべての分析について、推定または検定結果をつける。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。

文献

吉川徹・轟亮 (1996) 「学校教育と戦後日本の社会意識の民主化」『教育社会学研究』58:87-101.

第11講 モデルとパラメータ

田中重人 (東北大学文学部准教授)

[テーマ] 多変量解析を利用して、変数間の因果関係について考察する

1 前回課題について

独立変数ひとつだけのモデルとふたつ以上のモデルでは何が変わるか

- パラメータ推定値
- 平方和 (「被験者間効果の検定」の表)
- R2乗 (「被験者間効果の検定」の表の下)

共変量のみモデルについて

- 最小二乗法 (ordinary least square)
- パラメータ推定値に、独立変数 SD と従属変数 SD との比をかけると、Pearson の積率相関係数 r になる
- R2乗と積率相関係数の関係: $R=r$

2 「分散分析」(ANOVA) と「一般線型モデル」(GLM)

おなじ変数について分散分析と一般線型モデル (固定因子ひとつのみ) の両方の分析をおこなう。

- 分散分析のカテゴリ別平均値が、一般線型モデルのパラメータ推定値 B から計算できることを確認
- 「分散分析表」と「被験者間効果の検定」の表の「平方和」の欄を比較:

- (a) 「グループ間」 = 「修正モデル」
- (b) 「グループ内」 = 「誤差」
- (c) 「合計」 = 「修正総和」 = 分散 \times (人数 - 1)
- (d) R2乗 = イータの2乗 = 「決定係数」
- (e) $R = \eta = a/c$

複数の独立変数を投入した一般線型モデルでは、平方和の値を比較することで、変数の影響力の大きさを比較できる。

3 宿題

自分の興味のある分野で、「媒介効果」と「疑似相関」の例をひとつずつ考えよ。教科書 168-169 ページも参照。ISTU で来週水曜正午まで。

第12講 一般線型モデルの解釈

田中重人 (東北大学文学部准教授)

[テーマ] 一般線型モデルの出力の読みかた

1 前回宿題について

1.1 疑似相関

X_1 が X_2 と Y の両方に影響をもたらしているが、 X_2 は Y に対しては影響力を持たない場合を考える

$$\begin{array}{l} X_1 \rightarrow X_2 \\ X_1 \rightarrow Y \\ X_2 \quad Y \end{array}$$

このとき、 X_1 を無視して X_2 と Y だけの関連を分析すると、 X_2 が Y に対して影響を与えているように見えるので、因果関係を見誤ることがある。このような現象を「疑似相関」という。

1.2 媒介効果

X_1 によって X_2 が決まり、それによって Y が影響を受けるという因果関係を考える。

$$X_1 \rightarrow X_2 \rightarrow Y$$

このとき、 X_1 は Y に**直接** 影響を与えているのではなく、 X_2 が「媒介効果」をもたらしていることになる。

2 モデルの評価と係数の検定

2.1 モデル全体の評価

一般線形モデルのデータに対するあてはまりのよさをあらわすのが決定係数 R^2 である。これは相関比 η の2乗に等しいので、平方根をとれば、 η と同様の感覚で、「そのモデルによって、従属変数がどの程度説明できているか」を評価できる。

SPSS の出力では、「母集団においては決定係数がゼロである」(= どの独立変数も、従属変数に対して効果を持たない) という帰無仮説について検定を行った結果 (有意確率) が表示される。この結果が有意でなければ、モデル全体について説明力があるとはいえないことになる。

2.2 パラメータ推定値

決定係数が有意であれば、モデル内の各独立変数の効果について解釈していく。

各変数にかかる係数については、「パラメータ推定値」の表に、95%信頼区間が表示される。このなかにゼロが含まれているかどうかで、5%水準で有意な効果があるかどうかを判断できる（有意確率の列をみて判断してもよい）。

ただし、SPSSは、固定因子については、その変数の中のいちばん大きい値のカテゴリーを「基準」としてあつかい、その基準と他のカテゴリーとの効果の差を係数として表示している。これらの効果全部（たとえば5カテゴリーの変数であれば、4つの効果があることになる）でどの程度従属変数に影響力を与えているかについては、「被験者間効果」として表示される分散分析表を見たほうがよい。

2.3 平方和の分解

各独立変数の効果の大きさは、「被験者間効果の検定」の表の「平方和」の列に表示される。「修正総和」の行の数値が、従属変数の平方和（分散×（ケース数−1））である。各変数の行の平方和をこれで割ると、従属変数に対する相対的な影響力の大きさがわかる。

→ 分散分析表のよみかたについて復習

3 推定周辺平均

「オプション」の「推定周辺平均」「平均値の表示」に、カテゴリー別平均値を表示したい独立変数をえらぶ。

→ 他の固定因子については基準カテゴリー、共変量には平均値を代入して計算した値が表示される

「主効果の比較」をチェックする（「Bonferroni」で信頼区間を調整）と、どのカテゴリー間に有意な違いがあるかを比較できる。（教科書 p. 205–207 「多重比較」の項参照）

4 欠損値処理とケース数

一般線形モデルでは、欠損値は listwise で処理される。このため、多くの変数を投入すると、ケース数が小さくなるので注意。

5 課題

固定因子1つ、共変量1つ以上の一般線型モデル分析をおこない、その結果について、上記の「モデル全体の評価」から「欠損値処理とケース数」までの各項目について解釈する。

第13講 主効果と交互作用効果

田中重人 (東北大学文学部准教授)

[テーマ] 固定因子2つ以上をふくむ一般線型モデル

1 前回課題について

推定周辺平均の考え方：

- 一般線型モデルの「パラメータ推定値」を使って方程式をつくる
- これに各変数の値を適当に代入すると、その値の組み合わせを持つ人についての「予測値」がえられる
- これを使って、固定因子の各カテゴリに該当する人の予測値を求めたのが「推定周辺平均」である
- 共変量についてはその変数の平均値が代入される
- (注目しているもの以外の) 固定因子については、基準カテゴリが代入される

ペアごとの比較 (多重比較) は、カテゴリごとの人数の影響を大きく受ける (人数の多いカテゴリ間での比較は有意になりやすい)。

2 主効果と交互作用効果

「固定因子」に複数の変数を指定した場合、それらのカテゴリのすべての組み合わせごとに効果が計算される。この効果を「相互作用効果」または「交互作用効果」(interaction effect) という。交互作用効果は、カテゴリを表す変数同士の積であらわされる。

→ パラメータ推定値と推定周辺平均を見て、どのように計算されているかを理解してみよう

カテゴリ組合せを考慮しない、カテゴリによる従属変数への影響を「主効果」(main effect) という。通常の一般線形モデルでは、まず主効果を計算し、それで説明できない「残差」を交互作用効果として取り出す。

「モデル」で、各独立変数とそれらの交互作用効果のどれを使用するかを選ぶことができる。「主効果」だけのモデルや、逆に交互作用効果だけのモデルなども指定することができる。

※ 交互作用効果をふくむモデルで、主効果も交互作用効果も有意にならない場合は、交互作用効果をふくまないモデルも計算してみることを。

3 課題

教科書 pp. 213-222 を読み、次の設問に答えよ：

- (1) 「主効果」「交互作用効果」とはそれぞれどのようなものか。簡単にまとめよ。
- (2) 例題 (p. 217) について、SPSS データセットをつくって一般線型モデルで分析し、表 8-12 や図 8-2 とおなじ結果になることを確かめよ

記述統計量

従属変数: 指導者や専門家に頼れ

性別	edu3	平均値	標準偏差	N
1	1.00	3.25	1.410	20
	2.00	3.60	1.262	50
	3.00	3.68	1.035	41
	総和	3.57	1.211	111
2	1.00	2.69	1.278	35
	2.00	3.60	1.236	68
	3.00	3.44	1.261	25
	総和	3.32	1.304	128
総和	1.00	2.89	1.343	55
	2.00	3.60	1.241	118
	3.00	3.59	1.123	66
	総和	3.44	1.265	239

被験者間効果の検定

従属変数: 指導者や専門家に頼れ

ソース	タイプ III 平方和	自由度	平均平方	F 値	有意確率
修正モデル	30.130 ^a	6	5.022	3.323	.004
切片	224.690	1	224.690	148.676	.000
q1_2a	3.995	1	3.995	2.644	.105
q1_1	4.801	1	4.801	3.177	.076
edu3	7.297	2	3.649	2.414	.092
q1_1 * edu3	3.220	2	1.610	1.065	.346
誤差	350.615	232	1.511		
総和	3201.000	239			
修正総和	380.745	238			

a. R2 乗 = .079 (調整済み R2 乗 = .055)

パラメータ推定値

従属変数: 指導者や専門家に頼れ

パラメータ	B	標準誤差	t 値	有意確率	95% 信頼区間	
					下限	上限
切片	3.805	.333	11.426	.000	3.149	4.461
q1_2a	-.011	.007	-1.626	.105	-.024	.002
[q1_1=1]	.333	.317	1.050	.295	-.291	.957
[q1_1=2]	0 ^a
[edu3=1.00]	-.522	.352	-1.484	.139	-1.216	.171
[edu3=2.00]	.262	.294	.890	.374	-.317	.841
[edu3=3.00]	0 ^a
[q1_1=1] * [edu3=1.00]	.259	.466	.556	.579	-.660	1.178
[q1_1=1] * [edu3=2.00]	-.323	.390	-.829	.408	-1.091	.445
[q1_1=1] * [edu3=3.00]	0 ^a
[q1_1=2] * [edu3=1.00]	0 ^a
[q1_1=2] * [edu3=2.00]	0 ^a
[q1_1=2] * [edu3=3.00]	0 ^a

a. このパラメータは冗長であるためゼロに設定されます。

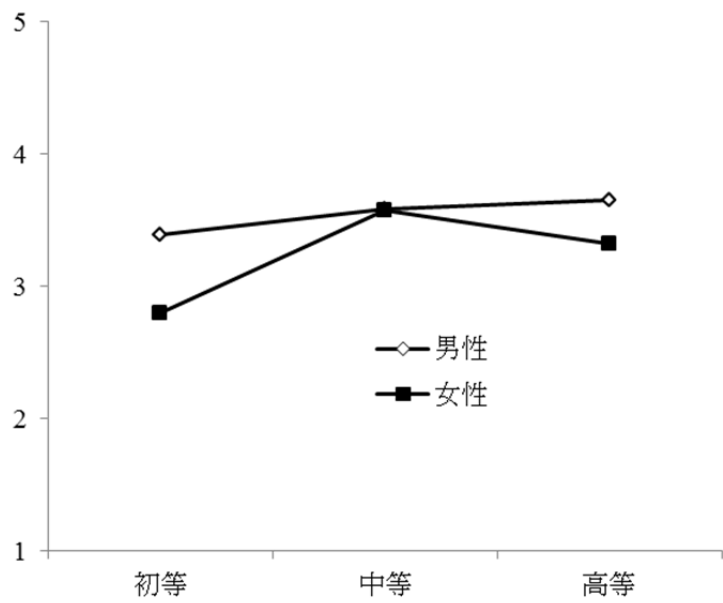
推定周辺平均

3. 性別 * edu3

従属変数: 指導者や専門家に頼れ

性別	edu3	平均値	標準誤差	95% 信頼区間	
				下限	上限
1	1.00	3.387 ^a	.288	2.821	3.954
	2.00	3.589 ^a	.174	3.246	3.932
	3.00	3.650 ^a	.193	3.270	4.031
2	1.00	2.795 ^a	.218	2.365	3.226
	2.00	3.579 ^a	.150	3.284	3.874
	3.00	3.318 ^a	.257	2.811	3.824

a. このモデルにある共変量は、満年齢 = 45.50 の値を基に評価されます。



第14講 一般線型モデルのまとめ

田中重人 (東北大学文学部准教授)

1 一般線型モデルの結果の書きかた

パラメータ推定値の表だけを書く (全部の変数の効果を正確に示したい場合)

- 推定値と標準誤差を書くのがふつう
- 固定因子のパラメータの示しかた
- 検定結果をアスタリスク (*) で示す
- 表の下に R² とその検定結果、人数を示す
- 基本的に小数第3位まで。小数点をそろえる

分散分析表と推定周辺平均 (特定の固定因子に注目する場合)

- 教科書 pp. 202–207, 213–222 も参照
- グラフを描く場合は、前回資料参照

多重比較結果の示しかたには、決まった様式がない。表ではなく、論文中に文章で説明したり、グラフに書き込んだりすることも多い。有意水準の調整方法を示すこと。

2 一般線型モデルのまとめ

- 因果関係とは
- 疑似相関と媒介効果
- 剰余変数の制御 (control)
- 最小二乗法
- 固定因子のあつかい
- 交互作用効果
- 他の多変量解析への応用

表 1 政治効力感の一般線型モデル

独立変数	推定値	標準誤差
切片	2.508*	0.336
年齢	0.017*	0.007
男性	0.609	0.322
学歴 (基準：高等教育)		
初等教育	-1.728*	0.351
中等教育	-1.497	0.297
男性×初等教育	0.853	0.461
男性×中等教育	-0.182	0.395

$R^2 = 0.225$. $N=239$. * : $p < 0.05$

表 2 政治効力感の分散分析表

変動因	TYPE III 平方和	自由度	平均平方	F
モデル全体	109.870	6	18.312	11.707*
年齢	10.215	1	10.215	6.530*
性別	36.084	1	36.084	23.070*
学歴	38.618	2	19.309	12.345*
性別×学歴	10.552	2	5.276	3.373*
誤差	378.524	242	1.564	

$R^2 = 0.225$. $N=239$. * : $p < 0.05$

表 3 政治効力感に対する学歴と性別の交互作用効果

学歴	男性	女性
初等教育	3.008	1.546
中等教育	3.204	2.778
高等教育	3.883	3.275

年齢 = 45.68 の場合の推定周辺平均.