現代日本学演習 V

実践的統計分析

田中重人 (東北大学文学部准教授)

3 年生対象：2020 年度 2 学期 (6 セメスタ) ＜金 4 ＞ Google Classroom クラスコード **csk6ctx**

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

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: 5 セメスタ開講の 現代日本学演習 II「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 (10/2〜10/16)
(2) 相関係数 (10/23〜11/13)
(3) 対応のあるデータの分析 (11/20〜12/4)
(4) 多変量解析 (12/11〜1/22) 〔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〕
(5) 期末レポート (2/5 提出期限)

3 復習事項

3.1 PSPP 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
- 中央値と四分位の求め方

3.2　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

3.3　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

3.4　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、帯グラフ、誤差範囲つき折れ線グラフの書きかた

4　連絡先

田中重人 (東北大学文学部現代日本学研究室)

〒: 980-8576 仙台市青葉区川内 27-1 文学部棟 6F
Homepage: http://tsigeto.info/officej.html

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、Google Classroom または電子メールによる。

現代日本学演習 V「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。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:　解答を木曜正午までに Google Classroom 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＝○, 裏＝▲ であらわすと

▲ ▲ ▲ ▲ (x=0)
▲ ▲ ▲ ○ (x=1)
▲ ▲ ○ ▲ (x=1)
▲ ▲ ○ ○ (x=2)
. . . . . .
○ ○ ○ ○ (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: 解答を木曜正午までに Google Classroom 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

現代日本学演習 V「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)。

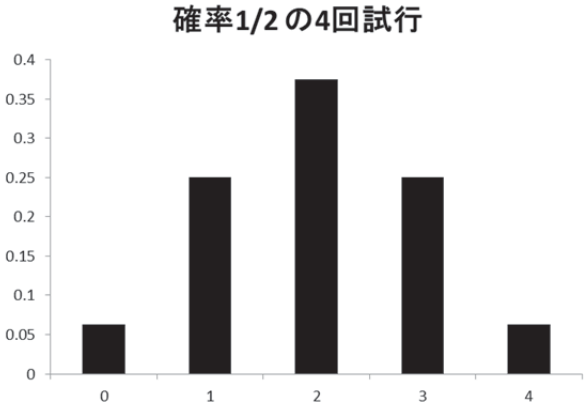
母平均の 95%信頼区間：

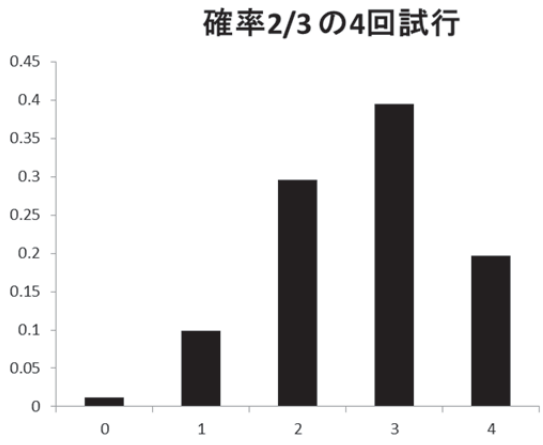
$$m \pm \text{臨界値}\frac{\text{SD}}{\sqrt{n}} \tag{2}$$

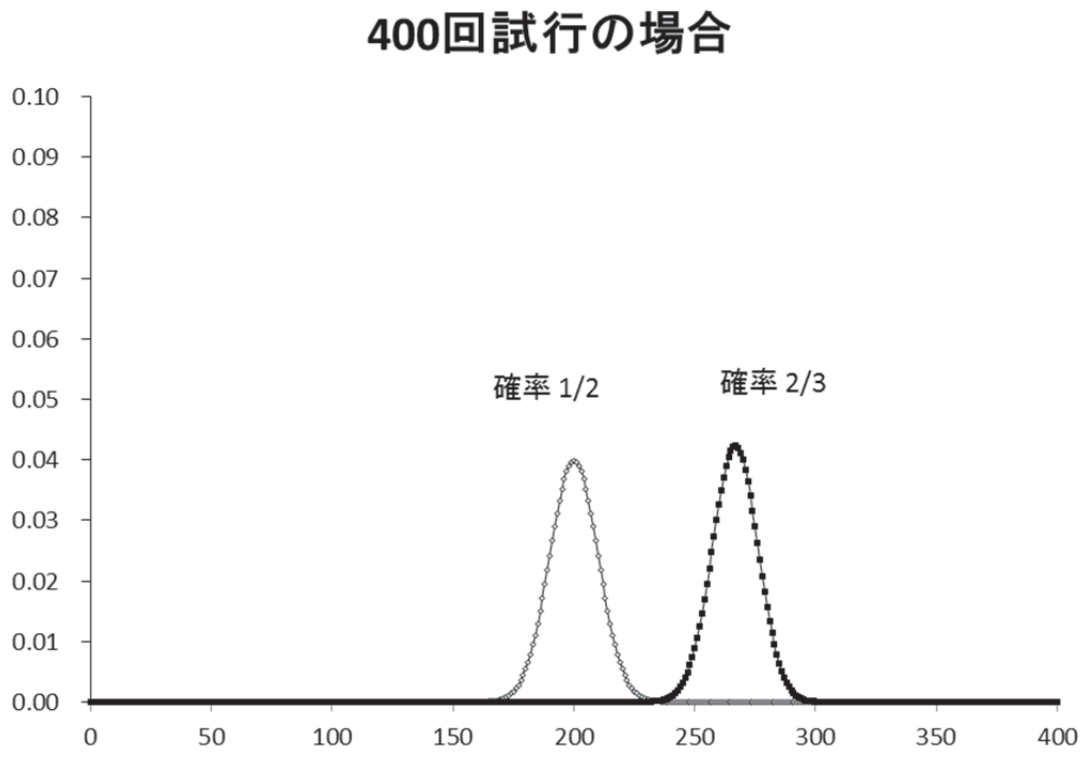
臨界値は自由度と危険率によって変化する (数表で調べる)。標本規模 200 以上で信頼率 95%なら、1.96 と考えてよい。

5　課題

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。


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

現代日本学演習 V「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 11 講

2　統計的検定

復習：

- 平均値の差の検定の方法 (数式と PSPP 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 をひとつ持つ (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 をふたつ持つ (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.

現代日本学演習 V「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

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 PSPP コマンド

- 「分析」→「記述統計量」→「クロス集計表」
- 変数を指定
- 「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.

現代日本学演習 V「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

PSPP 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 を参照)

例題:　前回宿題 (Kendall の順位相関係数) のデータを標準化してみよう。

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　PSPP 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。

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.

現代日本学演習 V「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　PSPP コマンド

- メニューの「分析」→「2 変量相関」を選択
- 変数を指定する

1.2　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.

現代日本学演習V (田中重人)

2020.11.20授業資料 (2020.11.27改訂)

問27 相関係数行列

	a	b	c	d	e	f	g	h
a. 評価高い職業	1.00*	0.41*	0.50*	0.19*	0.20*	0.19*	0.35*	0.56*
	(241)	(239)	(239)	(238)	(233)	(237)	(241)	(237)
b. 高い収入	0.41*	1.00*	0.42*	0.17*	0.16*	0.09	0.47*	0.34*
	(239)	(245)	(242)	(242)	(237)	(239)	(245)	(240)
c. 高い学歴	0.50*	0.42*	1.00*	0.13*	0.22*	0.20*	0.37*	0.51*
	(239)	(242)	(245)	(243)	(235)	(239)	(245)	(239)
d. 家族の信頼尊敬	0.19*	0.17	0.13	1.00*	0.43*	0.16*	0.07	0.07
	(238)	(242)	(243)	(245)	(235)	(239)	(245)	(239)
e. ボランティアや町内会	0.20*	0.16*	0.22*	0.43*	1.00*	0.42*	0.17*	0.13
	(233)	(237)	(235)	(235)	(238)	(236)	(238)	(234)
f. 趣味サークル	0.19*	0.09	0.20*	0.16*	0.42*	1.00*	0.31*	0.33*
	(237)	(239)	(239)	(239)	(236)	(242)	(242)	(238)
g. 多くの財産	0.35*	0.47*	0.37*	0.07	0.17*	0.31*	1.00*	0.50*
	(241)	(245)	(245)	(245)	(238)	(242)	(248)	(242)
h. 高い地位	0.56*	0.34*	0.51*	0.07	0.13	0.33*	0.50*	1.00*
	(237)	(240)	(239)	(239)	(234)	(238)	(242)	(242)

表1　相関係数行列

	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)

Pearson の積率相関係数．*: p <0.05.

表2 相関係数行列 (ケース数が全部同じの場合)

	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

Pearson の積率相関係数．*: p <0.05．N=105．

小数点をそろえるのが大変。
スペースで微調整する。

現代日本学演習Ⅴ「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　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：

PSPP では「分析」→「記述統計量」→「記述統計量」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

問 27 について、「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?

クロス表を出力して数えてみよう (セルに「合計」のパーセントを入れるとよい)。

→　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 ± 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

3.3　PSPP コマンド

- 「ノンパラメトリック検定」→「2 related samples」
- 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。

現代日本学演習 V「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「合計」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。

PSPP では、あらかじめ 2 つの変数の差を求め、「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」→「2 項」を使う。

- 「変換」→「計算」
- 「目標変数」に適当な名前 (たとえば DIFF とする)
- 数式を作成 (DIFF = 変数 x − 変数 y) して「実行」
- 「変換」→「他の変数への値の再割り当て」
- 変換元の変数 (DIFF) と変換先の変数名 (たとえば SIGN) とを指定
- 「今までの値と新しい値」の組み合わせを指定して 2 値変数に変換する (ゼロをどちらに入れるか注意)
- 実行

シンタックスを直接書く場合は、たとえばつぎのようになる。

```
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)
- 実行
- 度数分布 (「統計量」オプションで平均、分散、SD、平均の標準誤差を出力)

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　PSPP 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、PSPP で次のふたつの分析をおこない、結果が同じになることを確かめる

- 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。)

現代日本学演習V　（田中重人）
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				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	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=0.07 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)。

現代日本学演習 V「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、重回帰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。具体例は太郎丸 (2000) 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)。

- 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
- 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

この授業では前者をあつかう。

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　回帰分析

PSPP で、「分析」→「回帰」→「線形」を選択。

変数は次のように指定する

- 従属変数 = Q39g
- 独立変数 = Q1_2a (満年齢)

「統計」オプションで「係数」「信頼区間」「R」「分散分析」を指定

(学歴は変数変換が必要なので、次回以降)

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

$$\text{Q39g} = \text{切片} + \text{B} \times \text{Q1\_2a}$$

文献

太郎丸博 (2000)「階層性の神話」高坂健次 (編)『日本の階層システム 6: 階層社会から新しい市民社会へ』東京大学出版会, 161–180.

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.

現代日本学演習 V「実践的統計分析」

第 10 講　回帰分析

田中重人 (東北大学文学部准教授)

[テーマ]　回帰分析の基礎

1　変数間の関連を確認するための加工

年齢や学歴のような変数は、こまかくわかれていたり、順序に統一性がなかったりするので、そのままでは使いにくい。

適当なカテゴリーに統合して、クロス表を見る：

- 年齢 →10 歳刻みなど
- 学歴 →初等・中等・高等の 3 区分 (前回資料 参照)

間隔尺度あるいは順序尺度としてあつかい、相関係数を見る：

- 年齢はそのまま比率尺度として使える
- 学歴は、「教育年数」(その学歴取得に必要な標準的年限) に変換することが多い (前期第 5 講資料 参照)

PSPP のシンタックスはつぎのようになる：

```
recode q1_2a
 ( 20 thru 29 = 20 )
 ( 30 thru 39 = 30 )
 ( 40 thru 49 = 40 )
 ( 50 thru 59 = 50 )
 ( 60 thru 70 = 60 )
 into age10.

recode q6_1
 (1 thru 2 = 1) (3 thru 5 = 2) (6 thru 7 = 3)
 (12 = 1) (13 = 2) (14 thru 17 = 3)
 into edu3.

recode q6_1
 (1 = 6 ) ( 2 = 8 ) ( 3 thru 5 = 11 ) ( 6 = 14 ) ( 7 = 17 )
 ( 12 = 9 ) ( 13 = 12 ) ( 14 = 14 ) ( 15 = 16 ) ( 16 = 18 )
 into eduyear.
```

2　回帰分析のモデルとパラメータ

2.1　独立変数がひとつだけのモデル (単回帰分析)

従属変数として Q39g を、独立変数として Q1_2a を投入して回帰分析を実行 (前回資料参照)。

- 切片 (A) と回帰係数 (B)
- 標準化回帰係数 (β) と Pearson の積率相関係数 (r)

2.2 最小2乗法

回帰分析では、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つ増えたとき Y がどれだけ増えるか

教科書 78–81 頁参照

2.3 独立変数が複数の場合 (重回帰分析)

学歴を「教育年数」(上記参照) に変換したものを、独立変数に追加。この場合、回帰係数 (B) が独立変数の数だけあることになる。

$$\text{Q39g} = \text{切片} + B_1X_1 + B_2X_2 \tag{1}$$

やはり最小2乗法で係数を求めるので、$\sum(Y-A-B_1X_1-B_2X_2)^2$ が最小になるように A と B_1 と B_2 の組み合わせを求める。

- 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
  - 擬似相関 (教科書 pp. 86–91)
- 分散分析表から独立変数の影響力の大きさを読む

3 宿題

Q39g を従属変数とした回帰分析を3種類おこなって、結果がどうちがうかを説明する

- 年齢だけを独立変数とする
- 教育年数だけを独立変数とする
- 年齢と教育年数の両方を独立変数とする

4 期末レポート

期限: 2/5 (金)

提出先: Google Classroom

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。

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.

現代日本学演習 V「実践的統計分析」

第 11 講　疑似相関と媒介効果

田中重人 (東北大学文学部准教授)

[**テーマ**]　回帰分析の結果を解釈する一般的な手順

1　独立変数の「効果」をどう読むか

1.1　疑似相関

X1 が X2 と Y の両方に影響をもたらしているが、X2 は Y に対しては影響力を持たない場合を考える

```
X1 → X2
X1 → Y
X2    Y
```

このとき、X1 を無視して X2 と Y だけの関連を分析すると、X2 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

X1 によって X2 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。

X1 → X2 → Y

このとき、X1 は Y に**直接** 影響を与えているのではなく、X2 が「媒介効果」をもたらしていることになる。

1.3　直接効果

以上のようなことを考えながら、独立変数間の関連と、従属変数に対する直接的な効果を同時に把握する。

例題:　前回宿題の、3 つの回帰分析の結果から、3 変数間のどこに「直接効果」「媒介効果」「疑似相関」がみられるかを考える。

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。決定係数は、分散分析でいう「相関比」η の 2 乗に相当するので、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。これは、回帰分析では「重相関係数」と呼ばれ、R であらわす。

PSPP が出力する「分散分析」表では、「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が表示される (「有意水準」の欄)。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　係数の推定値

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、「係数」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 (「有意水準」の列をみて判断してもよい)。

3　欠損値処理とケース数

回帰分析をふくめ、多変量解析では、投入したすべての変数についてひとつも欠損値のないケースだけを分析に使う。このような欠損値処理方法を、「表単位」(listwise) の欠損値除去という。

この処理の結果として、多くの変数を投入すると、それだけケース数が小さくなるので注意。

4　課題

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。

現代日本学演習 V「実践的統計分析」

第12講　ダミー変数

田中重人 (東北大学文学部准教授)

[**テーマ**]　回帰分析での名義尺度変数の利用

1　2値変数の使いかた

たとえば、性別 (男性 =1, 女性 =2 の 2 値変数) を独立変数とする場合、そのまま使ってもよい。この場合には、「性別」の回帰係数は、値がひとつ増えることの効果をあらわすので、そのまま男性と女性の差を表すことになる。特に、独立変数がひとつだけのばあいには、男女の平均値の差が回帰係数と等しくなる。

ただし、2 値変数を独立変数とするときには、一方を 0、他方を 1 にしたほうが結果の解釈が簡単になるので、そういう変数をつくって使うことが多い。

```
recode Q1_1 ( 2=0 ) ( else=copy ) into MALE.
```

このような、特定の条件を満たす場合に 1、それ以外の場合にゼロをとるような変数のことを「ダミー変数」(dummy variable) という。

[**例題 1**]　性別による平均値の差の分析と、性別のダミー変数を投入した回帰分析をおこない、結果の対応関係を考える。特に、「定数」の値が何を意味するか考えること。

2　3つ以上の値のある変数をダミー変数に変換する

ある変数が k 個の値を持つとする。この変数を回帰分析で使いたい場合は、$k-1$ 個のダミー変数を作成する。

たとえば、学歴を 3 区分した変数 EDU がある場合、つぎのようにして、ふたつのダミー変数を作る。

```
recode EDU ( missing=sysmis ) ( 1 = 1 ) ( else = 0 ) into EDU_1.
recode EDU ( missing=sysmis ) ( 3 = 1 ) ( else = 0 ) into EDU_3.
```

この例では、EDU=2 の場合にはダミー変数をつくっていないことに注意。

[**例題 2**]　これらふたつのダミー変数を投入した回帰分析の結果と、学歴による平均値の差の分析を照合して、どの数値がどれに対応しているかを考える。

現代日本学演習 V「実践的統計分析」

第13講　講義まとめ

田中重人 (東北大学文学部准教授)

[テーマ]　回帰分析および講義の全体について復習とまとめ

1　前回課題について

前回の課題の回帰分析の推定結果では、Q39g の値が次の式で近似されていることになる：

$$\text{Q39g} = \text{定数} + B_1X_1 + B_2X_2 \tag{1}$$

ただし、

- X1 は初等教育のものについて 1、それ以外は 0 とする
- X2 は高等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 (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
定数と係数は、カテゴリ別の平均とつぎのような関係にある：

初等:　3.60 − 0.71 = 2.89
中等:　3.60 (= 定数)
高等:　3.60 − 0.01 = 3.59

2　ダミー変数の使いかた：補足

このように、k 個の値を持つ変数を回帰分析に投入するときは、$k-1$ 個のダミー変数に変換して使う。このとき、すべてのダミー変数がゼロになるカテゴリーがひとつ出てくることになるが、このカテゴリーを「基準」と呼ぶことがある。「基準」のカテゴリーの平均値が回帰分析結果の「定数」となり、各係数 (B) は、基準カテゴリーとの平均値の差をあらわす。

どのカテゴリーを基準にしてもよいのだが、通常は、つぎのどちらかにすることが多い：

- いちばんケース数の多いカテゴリー
- 平均値が最大 (または最小) のカテゴリー

このようなダミー変数は、本来はひとつの変数だったものなので、まとめてどの程度の影響をあたえているかを判断したいことがある。その場合、ダミー変数を投入した結果としなかった結果との間で決定係数 (R2 乗) を比較して、どれくらい増えたかを見ることがある。

3　ダミー変数をふくむ重回帰分析

ダミー変数も、普通の変数と同様に使ってよいので、複数の独立変数を投入して回帰分析をすることができる。

4　標準化係数 (ベータ)

独立変数・従属変数の両方を標準化した場合の係数を「標準化係数」(standardized coefficient) と呼び、β で表す (PSPP では「基準化係数」と書いている)。独立変数の効果の相対的な大きさをみたい場合に使える (積率相関係数 r とおなじ感覚で評価できる)。

5　回帰分析結果の書きかた

係数の推定値の表だけを書く

- 推定値と標準誤差を書くのがふつう
- ダミー変数の場合、ひとまとまりであることと基準のカテゴリーがわかるように工夫する
- 検定結果をアスタリスク (*) で示す
- 表の下に決定係数 R^2 とその検定結果、人数を示す (人数は、「分散分析」表の「合計」の自由度に 1 を足すとわかる)

6　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
- ダミー変数

7　発展のための新しいトピック

- 被験者内効果を含むモデル (変量効果 → random effect)
- マルチレベルデータや時系列データ
- 因果関係の分析をめぐる近年の展開 (厳密な無作為化実験への接近)
- 複雑なモデルを推定する方法 (ベイズ統計学とシミュレーション)

表 1　政治効力感の一般線型モデル

独立変数	係数	標準誤差
定数	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

$R^2 = 0.225$*.　N=239.　* : p<0.05