

比較現代日本論研究演習 III / 現代日本論演習

大学院生対象: 2010 年度後期
<木2> コンピュータ実習室 (文学部本館 7F 711-2) 授業コード=LM24206, L64206

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 「統計分析入門」 / 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もをを対象とする。

教科書

吉田寿夫 (1998) 『本当にわかりやすいすぐ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

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。

授業の予定

目次

1. 推測統計 (10/7~10/28)
2. 相関係数 (11/4~11/18)
3. 中間試験 (11/25)
4. 変数をキーにした分析 (12/2~12/9)
5. 多変量解析 (12/16~1/20)
6. 期末レポート (2/7 提出期限)

※ () 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

- 推測統計の基礎
- 確率密度と理論分布
- 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 χ^2 検定
- サンプル・サイズと検定力
- 誤差の対策

2. 相関係数

- 尺度水準について復習
- 相関図
- 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
- 2 項検定
- ハッセ図の利用

5. 多変量解析

重回帰分析? (受講者の興味と必要性によります)

6. 期末レポート

第 1 講「推測統計の基礎」(2010.10.7)

1. 記述統計と推測統計
2. 無作為抽出
3. 点推定と区間推定
4. 2 項分布

_____ 1 _____

【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)

_____ 2 _____

【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)

_____ 3 _____

袋のなかに色つきの玉が 60 個入っている:

赤色: 10 個

青色: 8 個

黄色: 4 個

.....

この袋から玉をひとつ取り出したとき、その色は……?

_____ 4 _____

全世界から n 人を無作為抽出したとき、
そのなかに〇〇は何%ふくまれるか?

→ 2 項分布

_____ 5 _____

【母集団特性の推定】

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?

→ 点推定 (point estimation)

_____ 6 _____

【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

_____ 7 _____

【確率の理論分布】

特定の仮定から

理論的に導出された確率の分布

例: 硬貨を投げるとき

表が出る →

裏が出る →

_____ 8 _____

【2 項分布】

Binomial distribution

硬貨を n 回投げる。

表が出る回数を x とする。

$n = 4$ のとき、 x はどのような値を
どのような確率でとるか?

_____ 9 _____

【計算方法】

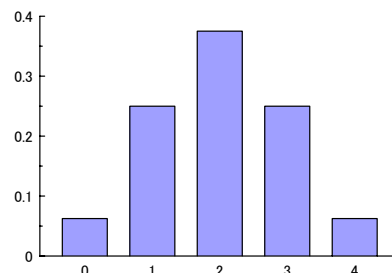
表=1, 裏=0 であらわすと

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)

の16通り。それぞれ等しい確率 (1/16) で起こると考える。

10

$n = 4$, 確率=0.5 の2項分布



11

【練習問題】

全世界から無作為に4人を抽出したところ、2人が「うどん好き」。

→ 全世界における「うどん好き」の比率は?

12

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

$m=0.5$ のとき M は? ($n=4$ とする)

13

多めに出てしまっている可能性:

M を適当に仮定して

$m \geq 0.5$ になる確率 を計算する

もし $M=0.3$ なら……

もし $M=0.2$ なら……

もし $M=$ なら……

$m \geq 0.5$ になる確率が2.5%以上になる M の範囲は?

14

すくなめに出ている可能性も同様に:

M を適当に仮定して

$m \leq 0.5$ になる確率 を計算する

$m \leq 0.5$ になる確率が2.5%以上になる M の範囲は?

15

【区間推定の原理】

- (1) 「信頼率」を決めておく (たとえば95%)
- (2) データから統計量を計算する
- (3) 母集団分布についていろいろなケースを想定する。その想定のもとでの標本統計量の確率分布を計算し、95%の確率で出現する範囲を確定する。

16

(4) この範囲のなかに、データから求めた統計量の値がふくまれるかを調べる

(5) (4) の条件を満たす想定ケースのすべてについて統計量を求める

(6) (5) で求めた値の集合が「95%信頼区間」である

17

【宿題】

全世界から無作為に4人を抽出したところ、全員が「うどん好き」。

→ 全世界における「うどん好き」の比率は?

※途中経過がわかるように書くこと

18

第2講「確率密度と理論分布」(2010.10.14)

- 1. 棄却域と採択域
- 2. 確率密度
- 3. 正規分布
- 4. 比率の区間推定

1

【棄却域と採択域】

2 項分布では、極端なケースほど起こる確率が低い。

硬貨を 40 枚投げて全て裏が出る確率＝

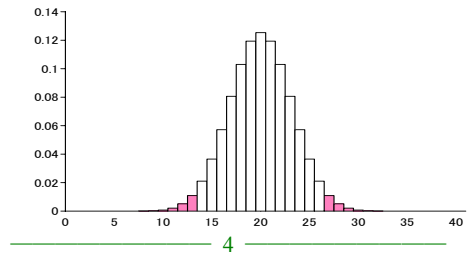
非常に確率が低いはずの極端な事象を観測したときは、理論分布の仮定を疑う。

2

- (1) 「危険率」(α) を決める
 - (2) ヒストグラムの上下の端から、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) 棄却域と採択域の境界を「臨界値」という
- α は 0.05 にすることが多い。

3

確率=0.5, n=40 の 2 項分布の場合

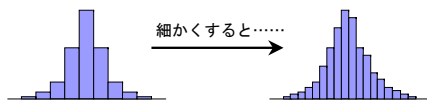


4

【ヒストグラム】

Histogram

連続量を階級分けして度数分布を示したもの



5

【確率密度のグラフ】

Probability density

連続量に対応して、連続的に変化する確率を表したものの



6

【期待値】

Expected value

値 (x) に確率 (p) を掛けたものの総和:

$$E = \sum(x \times p)$$

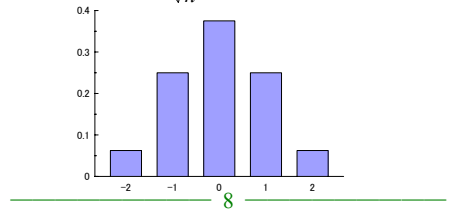
※「平均値」と呼ばれることもある

n = 4 の 2 項分布の期待値は?

7

【標準化】

$Z = \frac{(x - E)}{\sqrt{n}}$ に変換すると



8

【標準正規分布】

Standard normal distribution

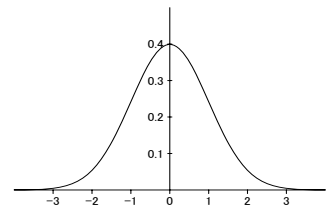
n が大きければ、Z は標準正規分布の確率密度関数

$$\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]$$

で近似できる

9

標準正規分布の確率密度のグラフ:



10

※ 0.5 以外の確率による 2 項分布でも、適当な標準化を行って n を増加させると標準正規分布に近づく

※ 標準正規分布に定数による加減乗除を加えたものを総称して「正規分布」(normal distribution) という

11

【正規分布の表記法】

標準正規分布の横の縮尺を s 倍に拡大して右に u だけずらしたものを

$$N(u, s)$$

と表記する。

標準正規分布は N(,) である

12

【正規分布の応用上の意義】

偶然による現象の生起確率や、その組み合わせで決まる物事は、正規分布 (またはそのファミリー) で近似できることが多い

※ 無作為抽出 = 等確率の繰り返し

13

標準正規分布の棄却域と採択域

数表が用意されている。教科書巻末参照。

$\alpha = 0.05$ のときの臨界値は?

n が大きければ 2 項分布は正規分布に近似 → 数表から臨界値を求めることができる。

14

【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)

15

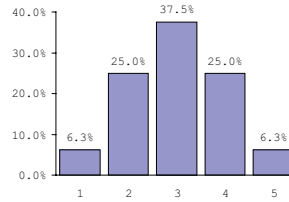
第3回「比率と平均の区間推定」(2010.10.21)

1. 平均値の区間推定

2. *t* 分布

1

【平均値の場合】



この母集団から 400 人の標本を抽出したとき
標本平均 m は、正規分布 $N(3, 0.05)$ にしたがう

2

【母平均が不明の場合】

母集団における分布についてあらゆるケースを想定して
計算するのは不可能



正規分布にしたがうことを仮定

$N(u, s)$ ただし u と s は不明

このとき u はいくらか?

3

【*t* 分布】

Student's *t* distribution

平均とばらつきの両方を予測するとき使う

(1) 硬貨を n 回投げる作業を 1 回おこない、表が出た回数を x とする

(2) 硬貨を n 回投げる作業を d 回繰り返し、それぞれについて表が出る回数 y_j ($j=1\dots d$) を数える

4

このとき

$$t_d = (x - E) \sqrt{\frac{d}{\sum_{j=1}^d (y_j - E)^2}}$$

の確率分布は、 n が大きければ、
自由度 d の *t* 分布で近似できる。

d が大きければ、標準正規分布で近似できる。

5

【平均値の区間推定】

母集団における正規分布という仮定の下では、
母集団平均値は、

t 分布を

- ・横方向に SD / \sqrt{n} 倍して
- ・右に m 移動させたもの

で推測できる。

6

平均値の 95%信頼区間のおおよその値：

$$m \pm 1.96 \times \frac{SD}{\sqrt{n}}$$

標準平均 (under m), 標準誤差 (under SD), *t* 臨界値 (under 1.96)

※ *t* 臨界値は自由度 ($n-1$) によって変化するが、
 $n > 200$ で 1.96 に収束する (教科書 p. 281)。

7

【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

8

【実習】

適当な変数について平均値の区間推定をおこなう。

「因子」を指定して層別の分析をおこない、
結果についてコメントをつけて提出

9

第 4 回 「統計的検定」 (2010.10.28)

1. 平均値の差の推定
2. 区間推定と統計的検定
3. さまざまな統計量の検定
4. 表の書きかた

1

【前回課題について】

- ・ 平均値をとってよい変数か → 尺度水準
順序尺度を間隔尺度とみなす際の条件
- ・ 層別の区間推定の結果からわかるのは、
「差があるか」であって「差が大きいか」
ではない

2

【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

$$\bar{d} \pm t_{\text{臨界値}} \times \text{併合SD} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

標準誤差

ただし n_1, n_2 はそれぞれの層の人数
 t 臨界値は自由度 (n_1+n_2-2) の t 分布にしたがって求める

3

【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イケない。

連続量を一定の値で切することもできる

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

前提：「母集団で正規分布」「2 層間で SD が等しい」

4

【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

5

【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

(教科書 pp. 156-158)

6

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

7

「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はない
かもしれない

8

【有意確率とは】

危険率を下げて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。

9

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 → 5%水準でも 1%水準でも有意
有意確率が 0.023 → 5%水準では有意だが 1%水準では非有意
有意確率が 0.088 → 5%水準でも 1%水準でも非有意

10

【区間推定と統計的検定】

- ★ 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
- ★ 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)
- ★ 統計量によっては、区間推定はすごくむずかしい場合がある

11

【むずかしい区間推定】

ϕ 係数 → 「Fisher の z' 変換」をおこない標準正規分布を利用 (相関係数と同じ) → 森・吉田 (1990, p. 225)

連関係数 V → 非心 χ^2 分布を利用

相関比 η → 非心 F 分布を利用

12

【クロス表の独立性の検定】

帰無仮説: 母集団においては $V=0$

「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

13

2×2 クロス表では $V=|\phi|$ なので、「母集団においては $\phi=0$ 」という帰無仮説を、上記の方法で検定できる。

ただし、独立性の検定で使う χ^2 の値が大きめに出るため、種々の調整を要求されることがある。

14

【分散分析と F 検定】

帰無仮説: 母集団においては $\eta=0$

「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」

※ 2 層の場合、 t 検定と同じ結果
※ 必要とする前提も t 検定と同様

15

【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)

16

【課題】

(1) 次の表から、平均値の差、エフェクト・サイズ、標準誤差、95%信頼区間を求めよ

	平均	SD	人数
男性	3.1	1.0	50
女性	2.9	1.0	50

(2) 人数が 4 倍 (それぞれ 200 人) の場合について同様に

17

【文献】

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。

18

2010.11.2 授業資料

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 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			合計 (人)
	「大いにある」	「少しはある」	「ない」	
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)

$\phi=0.191^*$. 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)

$\eta = 0.198$. $p < 0.05$.

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)

$\eta = 0.198^*$. *: 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)

$\eta = 0.086$. $p > 0.05$. 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)

$\eta = 0.086^{ns}$. ns: 5%水準で非有意.

無回答 = 7.

第 6 回「サンプルサイズの決定」(2010.11.11)

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
5. 非標本誤差

1

【検定力】

power (of a statistical 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

2

【 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

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

3

【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 → $3.84 / 0.2^2 = 96$
- 16% 差 →
- 14% 差 →
- 12% 差 →
- 10% 差 →
- 5% 差 →
- 1% 差 →

4

【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

5

【その他の係数の場合】

Pearson の相関係数 → ϕ 係数とおなじ

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

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

6

相関比 η → 次の式を使う (k はカテゴリ数):

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

- ※ $k \times 2$ 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 係数とおなじ

順位相関係数類 → 後日

7

【非標本誤差】

つぎのような誤差は、統計的に推測できない

- ・ 測定上のさまざまなエラー
- ・ 無作為でない標本抽出

- 測定の段階でできるだけ排除
- 分析・解釈の段階で配慮

8

★ 無作為でない標本についても、
統計的推測は必ずおこなうこと

→ 人数が少なすぎないか

【文献】

永田靖 (2003) 『サンプルサイズの決め方』 朝倉書店.

9

0. 尺度水準：復習
1. 尺度水準と分析法
2. 相関係数とは
3. Kendall の τ_b
4. SPSS コマンド
5. Pearson の r
6. Spearman の r_s

1

【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

2

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 または ρ

間隔×間隔 → 積率相関係数

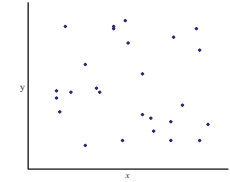
(product-moment correlation coefficient)

Pearson の r

3

【相関図】

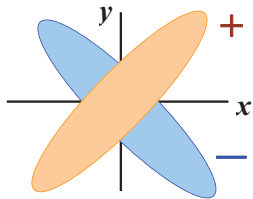
または「散布図」(scattergram)



4

【相関係数とは】

正(+)^の関係か、負(-)^の関係か



5

-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

(教科書 p. 75)

※ ϕ 係数は「4 分点相関係数」と呼ばれることがある
(Pearson の積率相関係数とおなじ方法で計算できる)

6

【Pair】

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

7

【Kendall の順位相関係数】

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

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

8

【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

(教科書 pp. 129, 130)

9

【SPSS コマンド】

クロス表の「統計」オプション

→ 「Kendall のタウ b」を選択

10

【実習】

Kendall の τ_b がプラスになる表とマイナスになる表を出し、クロス表の%を見て解釈する

11

【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

12

【Spearman の順位相関係数】

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。

r_s または ρ であらわす。

13

【SPSS コマンド】

クロス表の「統計」オプション

→ 「相関係数」を選択

14

【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

15

【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。

16

1. 各種相関係数の性質
2. 相関係数の推定と検定
3. 相関係数行列

1

【各種相関係数の性質】

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 τ_b
- ・ Pearson の r
- ・ Spearman の r_s

2

【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 の信頼区間は ϕ と同じ方法で求められる

(森・吉田 1990)

3

この信頼区間に $r=0$ が含まれるかを検定すればよい

信頼区間を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 $r=\pm 0.2$

N=400 で $r=\pm 0.1$

4

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

5

【相関係数行列】

correlation matrix

総当たりで相関係数を並べたもの

【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

6

【欠損値の処理】

- 対単位 (pairwise) の除去
個々の組み合わせごとに欠損ケースを除く

- 表単位 (listwise) 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)

7

【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

8

【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

【文献】

Bohrstedt, 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。

9

【中間試験】

- ・ 来週 (12/2) コンピュータ実習室
- ・ 試験範囲は、後期の授業開始から「相関係数」まで
- ・ 概念の説明／計算
- ・ 何でも持ち込み可

10

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

1. 対応のあるケース
2. 相関図とクロス表
3. 方向性の一致度
4. 2項検定

1

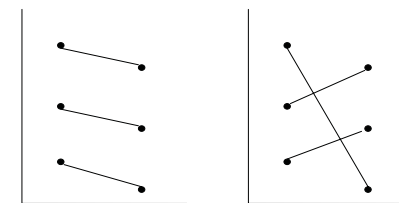
【対応のあるケース】

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

2

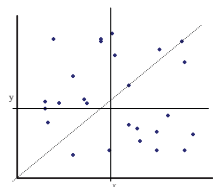
- ・類似の内容をちがう側面から測定
- ・おなじ項目を時間を置いて反復して測定



対応を考慮しないのはもったいない

4

【相関図による表現】



5

● 相関図の書きかた：
 「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

6

● クロス表の書きかた：
 「分析」→「記述統計」→「クロス集計表」
 「セル」で「パーセンテージ：全体」、
 「統計」で「相関係数」をチェック

7

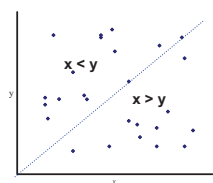
【方向性の一致度】

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

8

【相関図で考えると】



9

【「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 50%) を超えているか？

10

【統計的推測】

基準値 (たとえば 50%) 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

11

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)

12

【2項検定】

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

基準値 = 0.5 のときは特に「符号検定」という

13

【SPSS のコマンド】

「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)

14

【期末レポート】

期限：2/7 (月) 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：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：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。

15

1. 平均値の差の統計的推測
2. 対応のあるt検定
3. 表の書きかた

1

【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
- ・ 相関係数の意味

2

【平均値の差の統計的推測】

- 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
 - ・ 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)

3

信頼区間:

平均 ± t 臨界値 × 標準誤差

この区間に0が含まれているか? → t検定

4

対応関係にかかわらず、平均の値そのものはおなじ

ただし、相関係数によってSDが変わる
→ 標準誤差が変わる

5

標準誤差は次の式で求められる:

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

(ただし SD₁, SD₂ 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信頼区間:

平均値の差 ± t 臨界値 × 標準誤差

6

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」

7

●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

8

【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:

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 (基準比率を明記)。
x=y ケースの処理について明記。

9

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)

符号検定 / 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)

【課題】

適当な変数の組について、

- ・クロス表・相関係数
- ・差の変数の度数分布・平均・SD
- ・対応のある t 検定

を計算して出力

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 (5.4)	1 (0.4)	0 (0.0)	1 (0.4)	15 (6.3)
2. どちらかといえばそう思う	35 (14.6)	12 (5.0)	2 (0.8)	0 (0.0)	49 (20.5)
3. どちらかといえばそう思わない	79 (33.1)	37 (15.5)	9 (3.8)	0 (0.0)	125 (52.3)
4. そう思わない	32 (13.4)	15 (6.3)	3 (1.3)	0 (0.0)	50 (20.9)
合計	159 (66.5)	65 (27.2)	14 (5.9)	1 (0.4)	239 (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 (符号検定)。

符号検定の場合

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ケースを含む)。

2項検定 (x=yケース含む) 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

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)。

1. 多変量解析とは
2. 因果関係の設定
3. 回帰分析

1

【多変量解析のツボ】

3 つ以上の変数をつかう分析法

- 類似関係型
因子分析, クラスタ分析……
- 因果関係型
回帰分析, 判別分析……

(大野, 1998, p.48-56)

2

【類似関係型の分析】

因子分析 (factor analysis)

……間隔尺度の相関行列を使う

クラスタ分析 (cluster analysis)

……さまざまな尺度の(非)類似行列を使う

3

類似関係型の分析でできること :

- ★ 似た変数同士をまとめる
→ cluster
- ★ 潜在的要因を抽出
→ factor
- ★ 少数の変数に縮約
→ component, axis, vector, score…

4

【因果関係型の分析】

目的変数 (dependent variable)

結果になる変数 (ひとつ): 従属変数とも

説明変数 (independent variable)

原因になる変数 (複数可): 独立変数とも

目的変数と説明変数はしばしば Y と X であらわされる

5

【偏差】

適当な代表値 L について

$$= \sum(\text{個々のケースの値} - L)^2$$

4 人分のデータ {0, 5, 3, 8} について、
適当な値を L に代入して、
上の式を求めてみよう

6

【平均値と最小 2 乗解】

偏差の 2 乗和が最小になる値

= 最小 2 乗解

ふたつのグループ A, B について、
目的変数 Y の平均が 1.4 と 2.2

A グループは X=1 とし、
B グループは X=2 とすると、
次の式で近似できる :

$$Y = 0.6 + 0.8X$$

→ X が連続的な場合は?

8

【最小 2 乗法】

ordinal least square method

適当な直線 $A + BX$ によって Y の値を近似する方法。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と。

※ X と Y を入れ替えると結果が変わることに注意

9

【回帰分析とは】

Regression analysis

X の値によって Y が決まる、と考えて説明する

- Y をうまく説明できるような「回帰直線」を引く
(最小 2 乗法)
- 直線のパラメータ (とくに傾き) を評価する
(回帰係数)
- 回帰直線からのずれを評価する
(決定係数 = R^2)

10

A を「定数」または「切片」(intercept),
B を「回帰係数」(regression coefficient) という。

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

11

【SPSS コマンド】

「分析」→「回帰」→「線型」
従属変数と独立変数を指定する

「係数」の「B」の列に定数と回帰係数が出る

12

【決定係数】

回帰分析によって予測される Y と

実際の Y との相関係数を「重相関係数」と呼び、R とあらわす。

 R^2 を「決定係数」という。

分析結果のデータへのあてはまりのよさを表す

13

【重回帰分析】

Multiple regression analysis

ふたつ以上の変数を独立変数とする回帰分析。

$$\sum(Y - A - B_1X_1 - B_2X_2 - B_3X_3 - \dots)^2$$

のような式を最小化するような係数 $A, B_1, B_2, B_3 \dots$

を最小 2 乗法で推定する。

重回帰分析で推定された $B_1, B_2, B_3 \dots$ を

「偏回帰係数」ということがある

14

【標準化回帰係数】

変数を事前に標準化してから回帰分析を行った場合の
回帰係数を「標準化回帰係数」と呼び、 β であらわす

独立変数が従属変数に与える「効果」の大きさを

比較するときを使う

15

【宿題】

「生活全般満足度」を従属変数とし、「性別」「階層帰属意識(10)」をそれぞれ独立変数として、それぞれ回帰分析をおこない、重回帰分析の結果と比較する。
また、これら 3 つの変数間の相関係数行列を出力し、相関係数と回帰係数との関連について考える

【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。

16

1. 単回帰分析と重回帰分析
2. 多重共線性
3. 統計的推測

1

【今後の予定】

- 2/3 補講
- 2/9 17:00 レポート提出期限 (変更)
提出方法は 12/9 資料とおなじ
- 2/22 13:00-17:00 レポート返却
(田中研究室)

2

【単回帰分析】

独立変数がひとつだけの回帰分析

$$r = \beta = B \times SD_x / SD_y$$

$$|r| = R$$

3

【分散分析表】

- ★ すべての個体が回帰直線上にある状況を仮定して「平方和」を求める
- ★ この仮想平方和を実データの平方和で割った数値が R^2 に等しい
- ★ 独立変数が2値変数なら、層別の平均値による相関比 η と R は等しい

4

【重回帰分析】

ふたつの独立変数を投入した場合

もし独立変数同士が無相関なら、 β_1, β_2 はそれぞれ単回帰分析したときの β に等しい

$$\beta_1^2 + \beta_2^2 = R^2$$

5

【多重共線性】

multi-colinearity

説明変数間に相関があると、それぞれの効果を完全には分離できない。このため

$$\beta_1^2 + \beta_2^2 > R^2$$

となる

6

【直接効果】

重回帰分析では特定の独立変数の β のことを「直接効果の大きさ」ということがある。

そのほかの独立変数は「統制変数」(control variables) という。

例：擬似相関, 媒介効果

7

【統計的推測】

- 決定係数の検定
「分散分析表」に F 検定の結果が出る。
- 回帰係数の区間推定
 $B \pm t$ 臨界値 \times 標準誤差
有意確率も表示される

8

【多重共線性 (2)】

独立変数間に相関があると、回帰係数の標準誤差が大きくなる

→ 独立変数の効果を確定しにくくなる

およそ $r > 0.6$ のときは気をつけること

9

1. SPSS 19 について
2. 媒介効果と擬似相関
3. 表の書きかた

1

【SPSS 19】

使い方は基本的に以前と同じ

分析オプションやシンタックスエディタの見た目がかなりかわった

難しい分析法が使えるようになった

旧版の出力 (.spo ファイル) は開けないかも

2

【媒介効果と擬似相関】

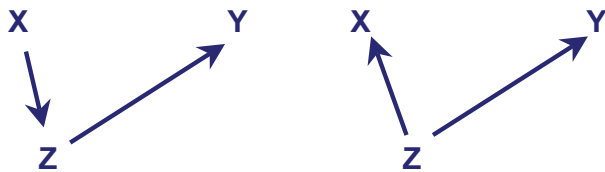
- A 階層帰属意識(10) → 生活全般満足度の単回帰分析
- B これにつぎの独立変数を追加
 全体的な不公平(Q29_1)
 生活水準の変化(Q36)
 自分には良い点があると思う(Q28E)

3

単回帰分析

独立変数 X → Y 従属変数

背後の因果構造



4

因果の方向は、分析結果からはわからない

理論的に決定する

- ・ 時間的な順序関係
- ・ 先行研究でのあつかい

5

【ケース数】

重回帰分析では、欠損値をふくむケースはリストワイズで除かれる

SPSS の回帰分析では、ケース数が出ない
 → 「統計」オプションで「記述統計量」
 または「分散分析」表の全体の自由度 + 1

6

【表に書くべき事項】

パラメタの推定値 (回帰係数および定数)

※標準誤差

※標準化回帰係数 β

●各係数の t 検定の結果

(B の推定値に * などをつける)

● R^2 と F 検定の結果 (* など)

●ケース数

7

表 1 生活満足度の重回帰分析

説明変数 (定数)	係数 (標準誤差)	β
	-0.037 (0.317)	
階層帰属意識	0.103* (0.041)	0.158
全般的な不公平	0.207* (0.077)	0.166
生活水準の変化	0.316* (0.063)	0.311
自分の良い点	0.157* (0.065)	0.144

$R^2=0.253^*$ $N=223$ *: 5%水準で有意

- ★ 小数第 3 位まで
- ★ 小数点をそろえる

8

基礎的な統計量として、各変数の
 平均値、SD、相関行列も示すとよい。

いずれも、欠損値をふくむケースを
 リストワイズでのぞく。

回帰分析の「統計」で「記述統計量」
 「部分/偏相関」を指定するのが簡便。

9