URL: http://www.sal.tohoku.ac.jp/~tsigeto/statg/2.html
作成：田中重人（講師）

比較現代日本論研究演習 II

大学院生対象：2008 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

授業の概要

（『講義概要』p. 402 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

1 学期開講の 比較現代日本論研究演習 I「統計分析入門」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

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

※（）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 χ2 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

未定（受講者の興味と必要性によります）

6. 期末レポート

比較現代日本論研究演習 II (田中重人)

第 1 回「推測統計の基礎」(2008.10.2)

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. **比率の区間推定**

【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)

【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)

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……?

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

→ 2 項分布 (次回)

【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)

【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

全世界のうどん好き人口の比率 (95% 信頼区間) ＝

$$0.6 \pm 1.96 \times \sqrt{(0.6 \times 0.4 / 400)}$$

答: 55.2 ～ 64.8 %

【次の場合は?】

うどん が好き: 30 人
そば が好き: 20 人

→ 人数が重要

【統計的検定】

全世界の人口のなかで、
うどん好きとそば好きはどちらが多いか?

2008.10.2

比較現代日本論研究演習 II 課題

氏名：

学籍番号：

次の 2 つの場合について、硬貨を実際に投げて、表が出る回数を数える。

・ 4 枚×4 回

・ 20 枚×40 回

(1) 下の表を完成させ、平均値を求める。

(2) 「表の枚数」を横軸、「回数」を縦軸にとって、棒グラフを描く

4 枚×4 回の場合

回数　　表の枚数

1 回目

2 回目

3 回目

4 回目

【平均値】

20 枚×40 回の場合

回数　　表の枚数

1 回目

2 回目

3 回目

4 回目

5 回目

6 回目

7 回目

8 回目

9 回目

10 回目

11 回目

12 回目

13 回目

14 回目

15 回目

16 回目

17 回目

18 回目

19 回目

20 回目

21 回目

22 回目

23 回目

24 回目

25 回目

26 回目

27 回目

28 回目

29 回目

30 回目

31 回目

32 回目

33 回目

34 回目

35 回目

36 回目

37 回目

38 回目

39 回目

40 回目

【平均値】

比較現代日本論研究演習 II (田中重人)

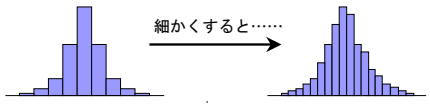
第 2 回「確率密度と理論分布」(2008.10.9)

1. ヒストグラムと確率密度
2. 確率の理論分布
3. 2 項分布
4. 正規分布

【ヒストグラム】

Histogram

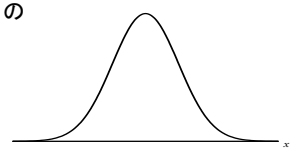
連続量を階級分けして度数分布を示したもの



【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



【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 →

【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?

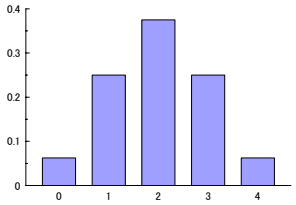
【計算方法】

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。

$n = 4$, 確率= 0.5 の 2 項分布



【期待値】

Expected value

値 (x) に確率 (p) を掛けたものの総和：

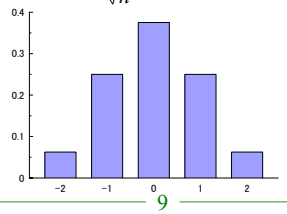
$$E = \Sigma(x \times p)$$

※「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?

【標準化】

$Z = (x - E)\frac{2}{\sqrt{n}}$ **に変換すると**



【標準正規分布】

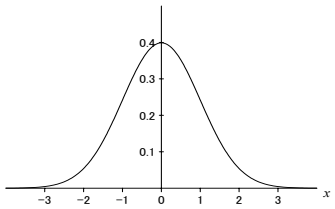
Standard normal distribution

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

標準正規分布の確率密度のグラフ：



※ 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

※ 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 を増加させると正規分布に近づく

【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

【文献】

宮川公男（1999）『基本統計学［第 3 版］』有斐閣。

比較現代日本論研究演習 II (田中重人)

第 3 回「比率と平均の区間推定」(2008.10.16)

1. 母比率と標本比率
2. 区間推定の原理
3. 比率の区間推定
4. t 分布
5. 平均値の区間推定

【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 $N(\quad , \quad)$ である

【母比率と標本比率】

母集団における比率を M とする。
そこから n 人の標本を抽出する。
標本における比率 m は?

M=0.5 で n=400 とすると……

$Z \quad = \quad (x - 200) / \sqrt{(400 \times 0.5 \times 0.5)}$

が標準正規分布 $N(0, 1)$ で近似できる。このとき

標本比率 $m \quad = \quad x/n \quad =$

は正規分布 $N(\quad , \quad)$ で近似できる。

したがって、95%の確率で、標本比率は

の範囲にあるといえる

一般に、標本比率は

正規分布 $N\left(M, \sqrt{\dfrac{M(1-M)}{n}}\right)$

にしたがう

【母比率が不明のとき】

標本比率 m はわかっているが母比率 M が不明

たとえば m=0.6 のとき M は? (n =400 とする)

もし M＝0.5 なら……

もし M＝0.55 なら……

もし M＝ なら……

95%以上の確率で m=0.6 になりうる M の範囲は?

【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。

(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

【比率の区間推定】

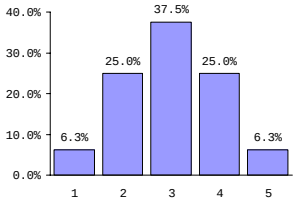
標本の規模がじゅうぶん大きく ($n > 30$)、
比率があまり偏っていない ($0.1 < m < 0.9$) **とき、**

95%信頼区間は

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

標準誤差
(standard error)

【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**

【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？

【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...d)を数える

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。

【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。

平均値の 95%信頼区間のおおよその値：

$$\underbrace{m}_{\text{標本平均}} \pm \underbrace{1.96}_{t\text{ 臨界値}} \times \overbrace{\frac{SD}{\sqrt{n}}}^{\text{標準誤差}}$$

※ 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)。

【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

比較現代日本論研究演習 II　(田中重人)

第 4 回「統計的検定」(2008.10.23)

1. 平均値の差の推定
2. 区間推定と統計的検定

【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

$$\underset{\text{平均値の差}}{d} \pm t\text{臨界値} \times \text{併合SD} \times \underset{\text{標準誤差}}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

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

【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

【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、という仮設

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

危険率 (critical level): 1－信頼率

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「p 値」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）

比較現代日本論研究演習 II （田中重人）

第 5 回「統計的検定のいろいろ」(2007.10.23)

1. 区間推定と統計的検定
2. 分散分析と F 検定
3. クロス表の独立性の検定

【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

【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

【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

【クロス表の独立性の検定】

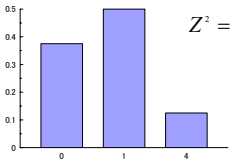
「クロス集計表」の「統計」で
「カイ 2 乗」を指定。
出力の「Pearson」の列の右端が有意確率

※ V（または $|\phi|$）の信頼区間で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

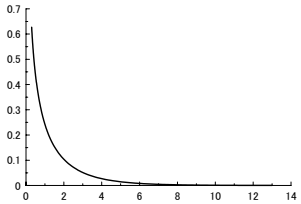
【自由度 1 の χ^2 分布】

2 項分布にしたがう変数の 2 乗を考える：



$$Z^2 = \left((x-E)\frac{2}{\sqrt{n}}\right)^2$$

n が増加すると、Z^2 の確率分布は
自由度 1 の χ^2 分布に近づく



【χ^2 分布の一般形】

硬貨を n 回投げる作業を c 回繰り返す。

それぞれについて表が出る回数 x_i を数え、それを標準化して 2 乗して総和を求める：

$$\chi_c^2 = \sum_{i=1}^{c}\left((x_i-E)\frac{2}{\sqrt{n}}\right)^2$$

n が大きければ、自由度 c の χ^2 分布に近似

【χ^2 分布の応用上の意義】

期待値や平均値からのずれを予測するときに使う

【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

【F 分布】

(1) 硬貨を n 回投げる作業を c 回おこない、それぞれについて表が出る回数 x_i (i=1 ... c)を数える

(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 ... d)を数える

このとき

$$F_{(c,d)} = \frac{\sum_{i=1}^{c}(x_i-E)^2}{\sum_{j=1}^{d}(y_j-E)^2}\times\frac{d}{c}$$

の確率分布は、n が大きければ、
自由度(c, d) の F 分布で近似できる。

※ $\sqrt{F_{(1,d)}} = t_d$ である。

※また、d が大きければ、$F_{(c,d)}$ は χ^2_c に近似する。

【F 分布の応用上の意義】

平均値からのずれの大きさを
比較するときに使う

【相互関係】

正規分布 ― (2 乗) → χ^2 分布

↑ (自由度∞) ｜ ↑ (自由度∞)

t 分布 ― (2 乗) → F 分布

【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)

【課題】

・平均値の差の t 検定
・クロス表の独立性の検定
・分散分析の F 検定
をそれぞれ適当な変数についておこなう

【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。

2008.10.30 比較現代日本論研究演習 II (田中重人)

授業資料

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				
	「大いにある」	「少しはある」	「ない」	合計	（人）
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)

ϕ=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)

η =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)

η =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)

η =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)

η =0.086^{ns}.　ns: 5%水準で非有意.

無回答＝７.

比較現代日本論研究演習 II (田中重人)

第 7 回「サンプルサイズの決定」(2008.11.13)

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. 「真の値」と誤差
6. 非標本誤差

【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

【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係数とおなじ

順位相関係数類 → 後日

【「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」(sampling error) がその典型

【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

★ 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.

2008.11.20 比較現代日本論研究演習 II

第 8 回「相関係数」(田中重人)

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

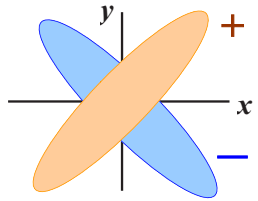
Spearman の r_s または ρ

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



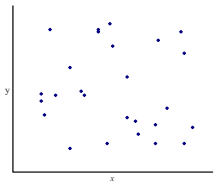
－1～＋1 の範囲の値をとる：

- 無関連のときゼロ
- 完全な関連のとき±1

※ 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)

【相関図】

または「散布図」(scattergram)



【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 または ρ 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。

2008.11.27 比較現代日本論演習

第 9 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 の信頼区間は ϕ と同じ方法で求められる

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=±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

【中間試験】

・再来週 (12/4) 授業観察室

・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

・前半が試験、後半はふつうの授業

比較現代日本論研究演習II (田中重人)
2008.11.27 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。
スペースで微調整する。

2008.12.4 比較現代日本論研究演習 II

第 10 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

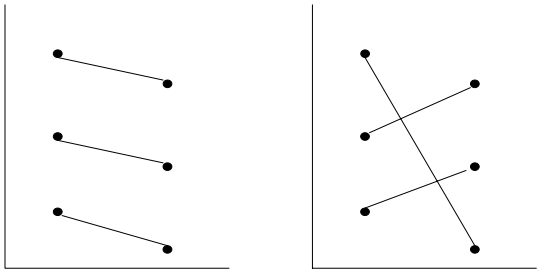
【対応のあるケース】

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

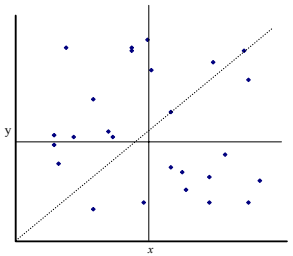
→変数をキーとした分析

（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】



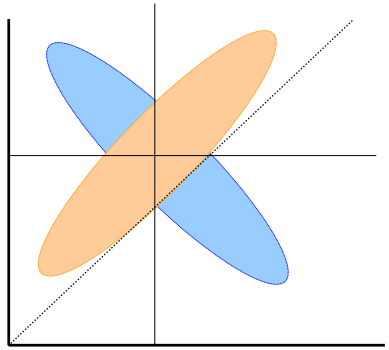
★ 平均値の差はどう表現されるか

平均値の差 ＝ 差の平均

★ 相関係数との関連

相関係数が大きいほど SD が小さい

差の $SD^2 = SD_1^2 + SD_1^2 - 2\,r\,SD_1\,SD_2$



【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

● 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

2008.12.11 比較現代日本論研究演習 II

第 11 回「対応のある *t* 検定」(田中重人)

1. 平均値の差の統計的推測
2. 対応のある *t* 検定

【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」に適当な名前を
- 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)

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

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

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 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信頼区間：

平均値の差 ± *t* 臨界値 × 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

【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

2008.12.18 比較現代日本論研究演習 II

第 12 回「方向性の一致度」(田中重人)

1. 方向性の一致度
2. 符号検定
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?

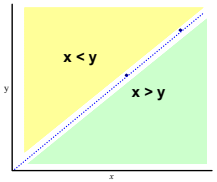
★ 2 項目間の一定の順序付け
(好き嫌い・適切さなど) が
どの程度共有されているかを
問題にしたい場合

【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】



【「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?

【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

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)

【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

【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)

【0.5 以外を基準値とする場合】

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」
で出る表とおなじ

統計的検定:
「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 (上記の基準値=a にあたる)

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:

【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。

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)

【期末レポート】

期限: 2/5 (木) 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。

比較現代日本論研究演習II （田中重人）
2008.12.18 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 (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)。

2009.1.8 比較現代日本論研究演習 II

第 13 回「多変量解析入門」(田中重人)

1. 多変量解析のツボ
2. 因果関係の設定
3. 回帰分析

【多変量解析のツボ】

★ **目的** (類似関係か因果関係か?)

★ モデル構造 ($y=b_1x_1+b_2x_2+\cdots+b_nx_n+a$)

★ 係数のポリシー
(モデルが説明する分散を最大化)

★ 係数の算出 (最小 2 乗法)

★ **結果の検討** (Fit and meaning)

(大野, 1998, pp. 56–61)

【ふたつの目的】

● **類似関係型**

因子分析, クラスター分析……

● **因果関係型**

回帰分析, 判別分析……

(大野, 1998, p.48–56)

【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

- 主成分法(principal component)
  ……特別にあつかって「主成分分析」とも
- 主因子法(principal factor)
- 最尤法 (maximum likelihood)
- その他

クラスター分析 (cluster analysis)

……さまざまな尺度の(非)類似行列を使う

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…

【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

【因果関係の設定のルール】

★ 時間的な順序関係

★ (実験の場合) 操作の順序

★ 先行研究でのあつかい

★ 一般的常識

つまるところ、絶対的なルールはない
→分析者が恣意的に決めるもの

【偏差】

適当な代表値 L について

$= \Sigma(\text{個々のケースの値} - L)^2$

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

【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?

【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 $\Sigma(Y-A-BX)^2$ が最小になるように

A と B の組み合わせを求める。

※ X と Y を入れ替えると結果が変わることに注意

【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)

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

【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

【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 であらわす。
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

【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。

$$\Sigma(Y- A - B_1X_1 - B_2X_2 - B_3X_3 -\ldots)^2$$

のような式を最小化するような係数 A, B_1, B_2, B_3 …
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 … を
「偏回帰係数」ということがある

【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
比較するときに使う

【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.

2009.1.15 比較現代日本論研究演習 II

第 14 回「回帰分析」(田中重人)

1. 単回帰分析と重回帰分析
2. 多重共線性
3. 統計的推測

【単回帰分析】

独立変数がひとつだけの回帰分析

$$r = R = \beta = B\, s_x / x_y$$

【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 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

【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

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

【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

【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

【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 ± t 臨界値 × 標準誤差

有意確率も表示される

【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

【宿題】

擬似相関、媒介効果の具体的な例を考えてくること

2009.1.22 比較現代日本論研究演習 II

第 15 回「回帰分析の結果の表示」(田中重人)

1. 重回帰分析におけるケース数
2. ダミー変数の利用
3. 表の作成

【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
→ 「分散分析」表の全体の自由度 + 1

リストワイズの相関行列を出してケース数を
確認するとよい

【ダミー変数】

dummy variable

0 と 1 でふたつのカテゴリーを表す変数。
a 個で a+1 カテゴリーを表すことができる

	中学	高校	短大	大学	大学院
X_1	0	1	0	0	0
X_2	0	0	1	0	0
X_3	0	0	0	1	0
X_4	0	0	0	0	1

全ダミー変数が 0 のカテゴリーを
「基準カテゴリー」(reference) という

ダミー変数を使った回帰分析
は分散分析とおなじ結果になる。

**ただし、各係数は、基準カテゴリーと
そのカテゴリーとの平均値の差になる。
(β に意味はない)**

recode でダミー変数をつくる:

```
recode q1_2a (40 thru 55 =1 ) (missing=sysmis)
    (else=0) into age40.
recode q1_2a (56 thru 70 =1 ) (missing=sysmis)
    (else=0) into age56.
```

このダミー変数を回帰分析に投入できる。
他の変数と同時に投入してよい。

「一般線形モデル」→「1 変量」で
「固定因子」にカテゴリー変数を指定
→ 自動的にダミー変数がつくられる

係数の推定結果は
「オプション」→「パラメータ推定値」

【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

表 1 階層帰属意識の重回帰分析

説明変数	係数	(標準誤差)	β
(定数)	6.787**	(0.570)	
性別 (基準：女性)	0.032	(0.203)	0.120
年齢 (基準：−39)			
40−55	0.002	(0.007)	0.000
55−70	0.025*	(0.009)	0.131
家族収入	−0.117**	(0.029)	−0.270
R^2=0.09**	N=238		

目的変数：階層帰属の主観的評価 (10 段階，逆転)
**：1%水準で有意 *: 5%水準で有意
無印：5%水準で非有意

★ 小数第 3 位まで
★ 小数点をそろえる

**基礎的な統計量として、各変数の
平均値、SD、相関行列を示すとよい。**

**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。**