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

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：2011 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,LB64206

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

1. 推測統計（10/6〜10/27）
2. 相関係数（11/10〜11/17）
3. 中間試験（11/24）
4. 変数をキーにした分析（12/1〜12/8）
5. 多変量解析（12/15〜1/19）
6. 期末レポート（2/2 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 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**
   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**

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
E-mail: tanakas2009@sal. tohoku. ac. jp
WWW: http://tsigeto.blog.fc2.com/blog-category-3.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、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://www.sal.tohoku.ac.jp/~tsigeto/newsj.html を参照。

比較現代日本論研究演習 III／現代日本論演習 (田中重人)

第 1 講「推測統計の基礎」(2011.11.7)

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**

【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 100 個入っている:

赤色: 60 個
青色: 20 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○○は何%ふくまれるか?

→ 2 項分布

【母集団特性の推定】

全世界から 8 人を無作為抽出:

うどん が好き: 8 人
そば が好き: 0 人

うどんが好きな人の比率は?

【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

【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

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=1 のとき M は? (n =8 とする)

多めに出てしまっている可能性:

M を適当に仮定して

8 回「うどん」が出る確率 を計算する

もし M=0.9 なら……

もし M=0.8 なら……

もし M= なら……

m = 1 になる確率が 2.5%以上である M の範囲は?

すくなめに出ている可能性も同様に:

m = 1 になる確率が 2.5%以上になる M の範囲は?

【もっと複雑な例】

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?

【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?

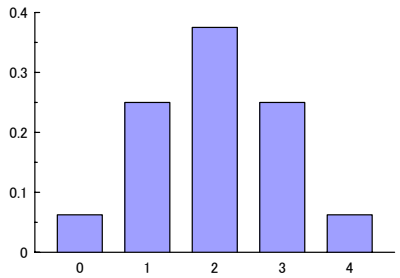
【計算方法】

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



【宿題】

タコの Paul 氏の 2010 年 FIFA ワールドカップでの活躍について、推測統計の観点から論じよ。

比較現代日本論研究演習 III／現代日本論演習（田中重人）

第 2 講「確率密度と理論分布」(2011.10.13)

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**

【宿題のポイント】

・結論は、**信頼率しだい**である
　→ 適当に決める
・95%信頼区間は 0.631 ～ 1
・確率 1/2 の繰り返しではなさそう
・およそ 3 回に 2 回は「あたり」になる程度の偏りがあったと考えてよい
・偏りの原因は不明

【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 8 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。

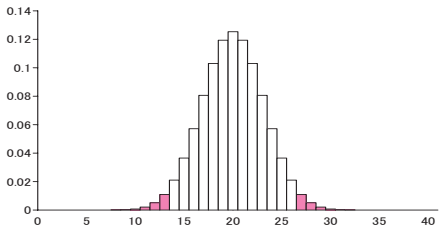
(1) 「危険率」(α) を決める

(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) 棄却域と採択域の境界を「臨界値」という

α は 0.05 にすることが多い。

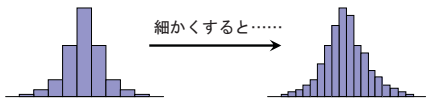
確率=0.5, n=40 の 2 項分布の場合



【ヒストグラム】

Histogram

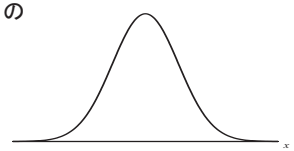
連続量を階級分けして度数分布を示したもの



【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



【期待値】

Expected value

値 (x) に確率 (p) を掛けたものの総和：

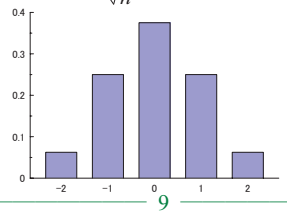
$$E = \Sigma (x \times p)$$

※「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】

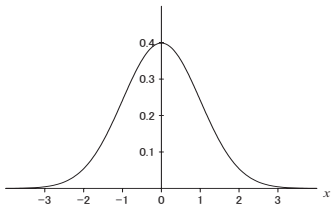
Standard normal distribution

***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：



※ 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

【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

【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

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。

α=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。

【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)

比較現代日本論研究演習 III / 現代日本論演習 (田中重人)

第 3 回「平均値の区間推定」(2011.10.20)

1. 平均値の区間推定
2. *t* 分布

【前回の課題】

m=0.6, n=400 のとき

正規分布で近似して 95%信頼区間を求めると

→ 標準誤差 (SE) = $\sqrt{(0.6 \times 0.4 / 400)} = 0.0245$

→ 95%信頼区間 = 0.6 ± **1.96**×0.0245 = 0.6 ± 0.0480

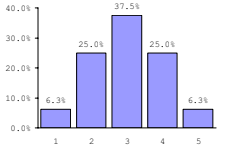
答：55.2% — 64.8%

【信頼区間の解釈】

「400 の標本を無作為抽出したときに
95%の信頼率に対応する採択域に
標本比率 0.6 が入るような
母比率の値の集合」

「母比率は 95%の確率で 55.2~64.8%」**ではない** ので注意

【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：

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

m：標本平均　1.96：*t* 臨界値　$\frac{SD}{\sqrt{n}}$：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

【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

他の人の意見をもらうこと（その人の名前を書くこと）

比較現代日本論研究演習 III / 現代日本論演習 (田中重人)

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

・ 【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

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

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

【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

前提：「母集団で正規分布」「2 層間で SD が等しい」

【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)

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
＝母集団においては平均値の差はない
かもしれない

【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。

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

【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

【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**

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。

【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

【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】

(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

【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。

2011.10.27 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				
	「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.

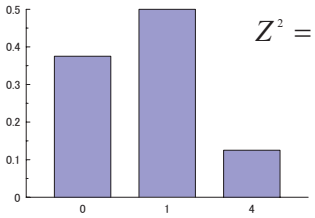
比較現代日本論研究演習 III / 現代日本論演習 (田中重人)

第 5 回「推測統計の応用」(2011.11.10)

1. χ^2 分布
2. F 分布
3. 検定力
4. サンプル・サイズの決定
5. 標本誤差と非標本誤差

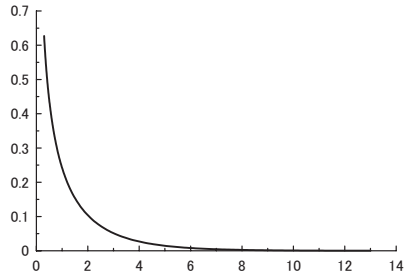
【自由度 1 の χ^2 分布】

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

→ 教科書巻末の数表の読みかた

【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

→ 教科書巻末の数表の読みかた

【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

【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

【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.

【今後の予定】

12/1　中間試験

2011.11.17 比較現代日本論研究演習 III / 現代日本論演習

第 6 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. Kendall の τ_b
4. SPSS コマンド

【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

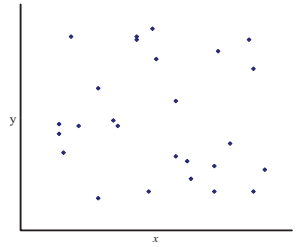
間隔×間隔 → 積率相関係数

(product-moment correlation coefficient)

Pearson の r

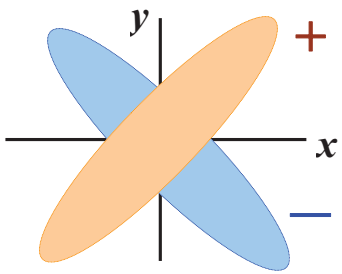
【相関図】

または「散布図」(scattergram)



【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

(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)

【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

【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

【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

2011.11.24 比較現代日本論研究演習 III / 現代日本論演習

第 7 回「さまざまな相関係数」(田中重人)

1. Kendall の τ_b
2. Pearson の積率相関係数 r
3. Spearman の r_s
4. 相関係数の使い分け

【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力し、クロス表の%を見て解釈する

【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)

【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】

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。

【SPSS コマンド】

クロス表の「統計」オプション

→「相関係数」を選択

【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

【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

【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.

2011.12.8 比較現代日本論研究演習 III / 現代日本論演習

第 8 回「相関係数行列」(田中重人)

1. 相関係数の推定と検定

2. 相関係数行列

【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)

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】

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

【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)

【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

【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

比較現代日本論研究演習III / 現代日本論演習 (田中重人)
2011.12.15 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。

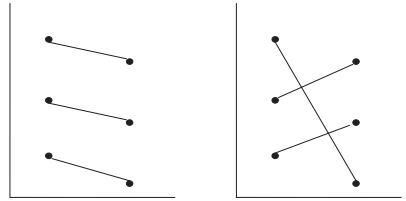
1. 対応のあるケース
2. 相関図とクロス表
3. 方向性の一致度
4. 2 項検定

【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

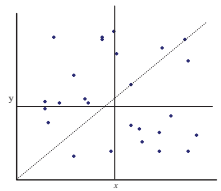
（実験の場合）被験者内要因か
被験者間要因か

- **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**

【相関図による表現】



- **相関図の書きかた：**

「グラフ」→「図表ビルダー」

「単純散布図」を指定
「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

- **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**」、

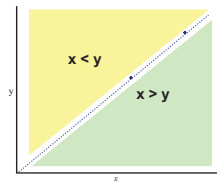
「統計」で「相関係数」をチェック

【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

【相関図で考えると】



【「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】

標本における値が
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）

【期末レポート】

期限：2/2 (月) 17:00
提出先：PDF ファイルをメールで tanakas2009@sal....あて
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。

PDF ファイルの作成には、次のような方法がある

- Adobe Acrobat を購入するか、使える場所を探す
- Microsoft Word 2010 では、PDF 形式で保存することができる
- 無料で PDF ファイルを作成できるソフトが各種存在
- オンラインで PDF ファイルを作成するサービスもある（Google Document など）

いずれの場合も、作成の形式として
「PDF/A」が指定できれば、それがのぞましい。

2011.12.22 比較現代日本論研究演習 III / 現代日本論演習

第 10 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
3. 表の書きかた

【符号検定の結果】

対角セルより右上／左下のセル度数の意味

度数分布表（周辺度数）との比較

Z = (比率の差 – 0.5) / 標準誤差

この Z が標準正規分布にしたがう
という前提で検定をおこなう

【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

・対角セルへの集中度は相関係数によってかわる
→ クロス表で Pearson の相関係数

【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

【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)

【課題】

(1) 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
を計算して出力し、解釈をつける
(2) 対応のある t 検定では、
なぜ相関係数を使うのか。説明せよ

比較現代日本論研究演習II （田中重人）
2011.12.22 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)。

2012.1.12 比較現代日本論研究演習 III / 現代日本論演習

第 11 回「多変量解析入門」(田中重人)

1. 多変量解析とは
2. 因果関係の設定
3. 第 3 変数の統制
4. 一般線型モデル

【多変量解析とは】

3 つ以上の変数をつかう分析法

- **類似関係型**
  因子分析，クラスター分析など
  似た変数同士をまとめる(潜在因子の抽出)
- **因果関係型**
  回帰分析，分散分析，一般線型モデル……

(大野, 1998, p.48–56)

【因果関係型の分析】

従属変数 (dependent variable)
　結果になる変数 (ひとつ): 目的変数とも

独立変数 (independent variables)
　原因になる変数 (複数可): 説明変数とも

従属変数と独立変数は、しばしば Y と X であらわされる

【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

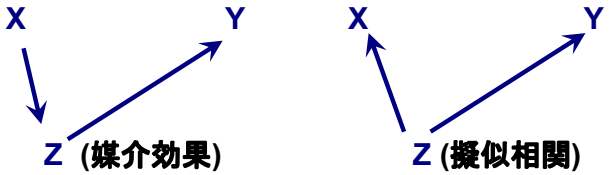
→ どのようなことがいえそうか?

※学歴は 3 分割 :
初等 (1,2,12); 中等 (3-5,13); 高等 (その他)

２変数での分析

独立変数 **X** ⟶ **Y** 従属変数

背後の因果構造



(教科書　p. 87)

【第 3 変数の統制】

Control
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

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

【一般線型モデル】

「分析」→「一般線型モデル」→「1 変量」

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

【パラメータ推定値を読む】

Q39g の値は、つぎの式で近似できる :

初等教育 :
Q39g = 切片 + B_1X_1 + B_2X_2

中等教育 :
Q39g = 切片 + B_1X_1 + B_3X_3

高等教育 :
Q39g = 切片 + B_1X_1

【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。

結果を印刷し、パラメータ推定値を比較して、なにがわかるかを考察。

【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.

2012.1.19 比較現代日本論研究演習 III / 現代日本論演習

第 12 回「一般線形モデル」(田中重人)

1. 固定因子と共変量
2. 独立変数を増やすと何が変わるか
3. 分散分析表と決定係数
4. 表の書きかた

【固定因子と共変量】

- **固定因子 ＝名義尺度の変数**
  - 自動的にカテゴリーに分割され、
  - 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

【固定因子ひとつのモデル】

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 ← 基準

【分散分析表】

edu3 + 誤差 ＝ 修正総和

決定係数 R^2 = edu3 / 修正総和

$\sqrt{R^2}$ = 相関比 η

【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

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。

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

【独立変数が複数の場合】

・「コントロール」することの意味

- 独立変数がひとつの場合と何が変わるか?
- 分散分析表から独立変数の影響力の大きさを読む

【統計的推測】

推定された係数それぞれについて、区間推定と統計的検定がおこなわれる

【注意事項】

固定因子が複数あると、それらを組み合わせたカテゴリーごとに係数が推定される

→ 「モデル」オプションで「交互作用」をはずす

相関の極端に高い共変量を投入してはならない
(およそ $r > 0.7$)