交流協会　　2009/03/07

統計の基礎とグラフ作成

立教大学　淡江大学　　村瀬　洋一

1　エクセルの基礎的操作
2　統計関数入門　－社会調査データを用いた分析
3　グラフ作成
4　統計分析の基礎　－分散とは何か
5　統計的検定と分散分析

1.1.分析の目的とは

■ 1)データの要約
- 平均、分散（ばらつき具合、標準偏差）

まずは、標準偏差を自分で計算できるように！

■ 2)関連の解明（と将来予測）

例１　少子化の原因とは

分析の結果、結婚年齢が高い県ほど子供が少ない

→　将来どの県で、少子化が進か予測可能

例２　キリスト教の国は産業化が進む

国を豊かにするためには、教会を造ればよい？

日本では？

1.2.分析の手順

■ 調査目的を設定

■ 仮説を作る

原因と結果を明確に

仮説とは因果関係を含む文である

調査したが分析せず、とならないよう

■ 分析の基本

- 木を見て森を見ず、に注意
- 予断と偏見　分析前の思いこみが反映しないか注意

■ 最近、日本語に興味を持つ人は多い

結果だけで原因がない

■ 若い人ほど日本語に興味を持つ

年齢と興味という2つの要素　－　よい仮説

■ 仮説ははずれてもよい

■ はずれてから考えることこそが、重要！

－　当たり前の結果より意外な結果

1.3.データ行列とは何か

- 行が個人、列（カラム、けた）が変数となる数字の行列

- 具体例　　3人分のデータ行列の例
  - 5カラム目までがサンプル番号
  - 00101番の人が、問1で2、問2で1、問3で4と答えている。

00101 21412508 2421111111
00102 21611402 1221213132
00103 12714806 1222212121

1.4.疑似相関とは何か

- 表面的な関連（架空例）
  低学歴　→　　性役割を肯定
  ところが、学歴が低い人は高齢
- 真の関連
  高年齢　→　　性役割を肯定

- 生態学的相関（マクロレベルでの疑似相関）
  - 表面的な関連
    米の生産の大きい県　　→　　自殺が多い
  - 真の関連
    一人暮らしの老人が多い県　　→　　自殺が多い

2エクセルの統計関数

■ セル　K2～K201　についての計算

任意のセルの中に、以下のように書く

- 合計　　=SUM(K2:K201)
- 平均　　=AVERAGE(K2:K201)
- 標準偏差　=STDEV(K2:K201)

半角英数字で入力！

- 挿入をクリックし、関数→すべて表示、で選択してもできますが、最初はまず、自分の手で入力してみてください。

■ データの場所　http://shakaichousa.net/　→SPSSなどPDF資料

変数の種類

■ 離散変数（名義尺度）

例えば　仙台調査問１９

■ 連続変数（量としてとらえるもの）

順位、間隔、比率尺度

3. よいグラフとは

- グラフのみを見て、第 3者が理解できること
  - 適切なタイトル
  - 数字の意味を軸で説明　－　%か、人数かなど
  - グラフラベル
    グラフ全体をクリック　→　グラフのオプション

- 見ても分からないグラフ
  - 質問内容を書いていない
  - 数字が大きいほど賛成か反対か不明
  - かっこいいが見にくい3次元グラフ

4.分散とは何か

- 散らばり具合

A組　60　80　62　78
B組　68　72　67　73

どちらも平均は 70 ($\bar{X} = 70$)

- 平均からの距離　… d　($X - \bar{X}$)

A組　-10　10　-8　8
B組　-2　2　-3　3

分散の定義式

$$S^2 = \frac{d^2 \text{の合計}}{\text{人数N}} \qquad \frac{100+100+64+64}{}$$

S が標準偏差（分散の平方根）

なおデータ分析では、Nでなく(N-1)で割る（不偏分散）
標本の散らばり具合は、母集団より小さくなるので。
実際には、Nが大きい場合、両者はほぼ同じ値になる。

5.検定の考え方

- 2つの平均値に差があった場合、
  本当に差があるのか、誤差なのか
- 学生全体に調査はできない。
  データは一部（標本）
- 標本から母集団を推測する（統計的推定）

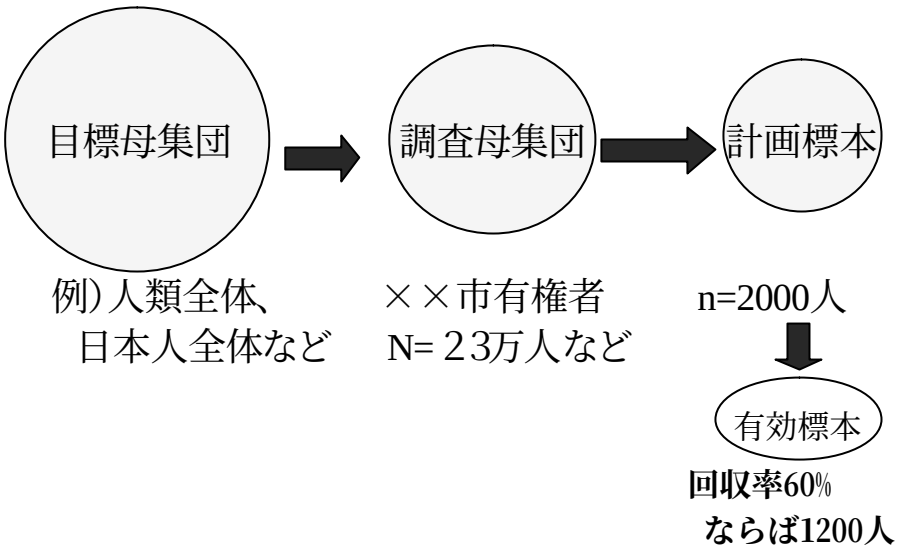


図　母集団と標本の関係（『社会調査演習』より）

5.1.平均値の差の検定　分散分析

- 具体例　あるクラスの点数平均値
  ー　男が68点、女が70点。
- 意味のある差か、誤差か。

- 結果：　点数
- 原因:　性別
  この例だと、
  説明変数は性別という1つだけ（一元配置）

- 目的　ー　被説明変数（従属変数）Yと関連する
  説明変数Xは何かを解明。

■ Yは連続変数（量）、Xは離散変数（カテゴリー）。
　重回帰分析　XもYもすべて量的変数

■ 分析法の考え方
　－ 他のXの影響を統計的に取り除いても、
　　Yに対するXの効果があるか解明

■ 具体的には2つ以上の平均値の差を検定する
・帰無仮説　各グループ（級、組、群、カテゴリー）の平均値は全て等しい。
・対立仮説　各平均値の間に差がある。

■ 統計的検定の帰無仮説
　原因と結果が「無関連」という意味の仮説

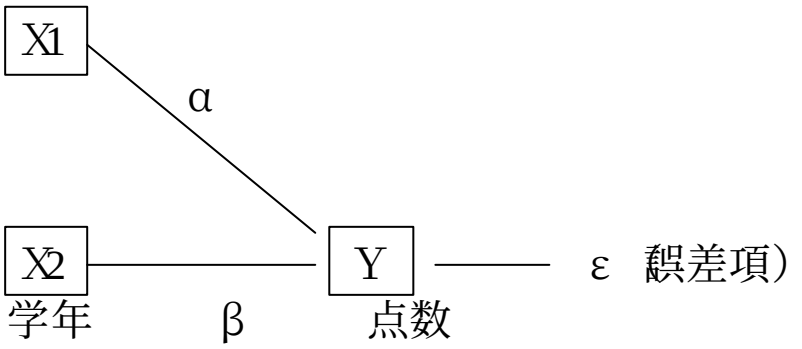
■ 統計的検定は、分析の初歩
　関連の有無を見るだけ　関連の大きさは言及しない
　有無について1，0で考えている
　関連の大きさについては、検定以外の方法で分析

■ 分散分析
　少人数についても可能
　ただし検定のみ　関連の大きさは考えない
　Xの数は3つ程度

F値（F比）の考え方

$$F = \frac{MS_{BETWEEN}\text{(級間平均平方)}}{MS_{WITHIN}\text{(級内平均平方)}} = \frac{\text{モデルによって説明できる分散}}{\text{モデルによって説明できない分散}}$$

分散分析の基本モデル（グラフによる表現）



分散分析の基本モデル

$y = \mu + \alpha + \varepsilon$

y：各個人の測定値
μ：全体平均
α：グループに属する効果
ε：誤差

5.2.分析結果の見方

■ 資料参照　分散分析表