LibreOffice Calcで統計解析をやってみよう

安部武志
<tabe@fixedpoint.jp>
2017-06-17

統計解析の2つの側面

1. 記述統計(descriptive statistics)
   データそのものの性質について知りたい
2. 推測統計(inferential statistics)
   データの背後にある母集団について知りたい

データそのものの性質を把握するための記述統計

- より基本的
- データについての情報をコンパクトに表す尺度(統計量)を見る
- 例: 国勢調査 ← 全数調査

記述統計の具体例

- 総務省統計局が公開しているデータ

  http://www.stat.go.jp/data/
- XLSやCSVのファイル
- 「日本の統計」>「第1章　国土・気象」>「1- 6　都道府県別行政区画と面積（エクセル：35KB）」

  http://www.stat.go.jp/data/nihon/zuhyou/n170100600.xls

  … 記述統計の対象

Calcステータスバーにあるサマリーが便利

- LibreOffice 5.2からの新機能: 複数の関数を同時に選べる

平均

- =AVERAGE(X1, X2, ..., X30)
- 算術平均を計算する
- 「真ん中がどのあたりか」を示す
- 外れ値に弱いことがある
  ← 面積でいうと北海道が平均を引き上げる
- だめな例: 平均年収?

Count or CountA?

- 数値とテキストの扱いの違いがある
  - "COUNT(Value1; Value2; ...; Value30)"は引数のリストにいくつ数値があるかを返す

    ... テキストは無視
  - "COUNTA(Value1; Value2; ...; Value30)"は引数のリストにいくつ値があるかを返す。

    ... テキストも含めて数える。空文字も1つに数える。参照の形の引数において空のセルは数えない。

中央値

- =MEDIAN(X1, X2, ..., X30)
- 「真ん中がどのあたりか」を示す
- 外れ値に強いと言われる

  ←北海道を含めても平均のように大きく影響されない

最頻値

- "=MODE(X1, X2, ..., X30)"
- 「一番よく出てきた値」を示す
- 暗にサンプルデータとして扱っている
  ... 母集団の確率分布を想定している

最頻値(続)

- =MODE.SNGL()

  "=MODE()"と同じで、とりあえず1つ返してくれる。
- =MODE.MULT()

  複数個該当する値があったら全部返してくれる。

分散

- 「個々のデータ間のばらつきの大きさ」
- 実は2つの異なる分散がある(標本分散 vs 不偏分散)ので、どちらかの話をしているか注意が必要。
- "=VAR()"

  N-1で割る。"=VARA()"はテキストを無視してくれるので便利。
- "=VAR.P()"

  Nで割る。母集団が与えられているとしているから。
- "=VAR.S()"

  N-1で割る。サンプルデータが与えられているとしているから。

標準偏差

- 分散の平方根
  平方根を取るので、単位がデータの軸のものと揃う
- 分散と同じようにこれにも2種類(4関数)

  "=STDEV()"

  "=STDEVA()"

  "=STDEV.P()"

  "=STDEV.S()"

結局、記述統計はデータの要約

- 度数分布(ヒストグラム)の性質を要約して伝えることができる。

  例: 平均と中央値と最頻値が一緒ならどんな度数分布?

推測統計: サンプルデータから何か分かるの?

- どうなったら母集団の性質が分かったといえるか
  1.パラメトリックな分布を前提にする場合
    パラメーターの値(もしくは推定値)を知る
  2.統計的仮説が成り立つかどうかを調べる

1.推定(estimation)

2.検定(test)

応用例: 選挙速報の「当選確実」

Calcで推測統計解析するときの便利な機能

- Statistics Wizard (Excelの"Analysis ToolPak"相当)
  - LibreOffice 4.2から
  - メニューの「データ(D)」>「統計(H)」
  - 分散分析(ANOVA)、相関分析や各種検定のための入力がウィザード形式でできる

分散分析(ANOVA)の例

例: 学校での3教科の試験の点数

- 出題者としては3教科とも平均点が同じになるか知りたい

- 入力 alpha: 有意水準
- F検定
  - 「帰無仮説が正しければ統計量はF分布に従う」
  - 目的: 2つ以上のサンプルで平均を比較する
  - 帰無仮説: 2つ以上のグループのサンプルが同じ平均値を持つ母集団から取られた
  - 分散比が高ければ帰無仮説棄却

    ←F境界値より大きかったらその有意水準では棄却
  - P値を見てもいい

一元配置分散分析(single-factor ANOVA or one-way ANOVA)の仮定

以下が全て満たされているという仮定:

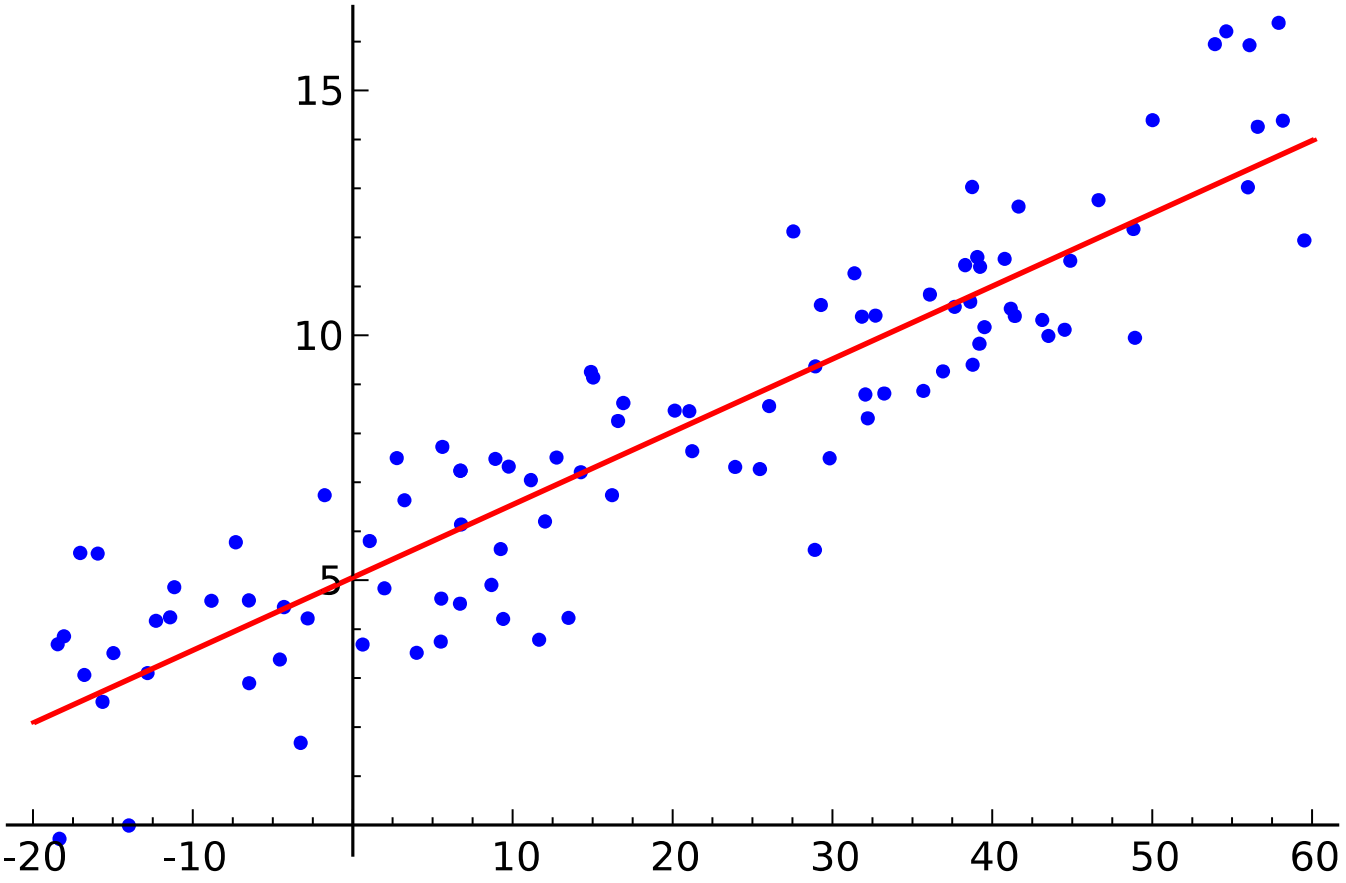
- (数値データ)
- 応答関数残差は正規分布する
- 母集団の分散は等しい
- 各グループに対する応答はi.i.d.

回帰分析

回帰はregressionのことだが、バグとしてのregressionとは別の意味合い

- 回帰は条件つき平均である
- 回帰分析は分散を分解すること

線形回帰



線形回帰の例

例: 時刻と計測値の関係(デモ)

(説明変数が1つの場合の)線形回帰の仮定

以下が全て満たされているという仮定:

- 説明変数は誤差なし
- 応答変数の平均は回帰係数と説明変数の和になっている

「本物の統計家はスプレッドシートを使わない」

Rを使う。

- アメリカ統計学会(ASA)の学術雑誌の1つに"Journal of Statistical Software"があるが、掲載されている大部分ははRのパッケージについての論文。
- MATLAB、Python、S-PLUS、SAS、SPSSもぼちぼち

統計解析の3つの落とし穴

深刻な順に:

1.仮定や確率モデルの間違い
- 無理のある仮定
- 前提に合っていないモデル、など

2.精度の悪い近似

3.計測誤差や浮動小数点数計算の誤差

Abraham Wald

- ハンガリー出身の数理統計学者。
  - ナチス政権下からアメリカに亡命して第二次世界大戦中に統計による軍事研究を行う機関に所属
- 軍からの調査依頼

  「戦闘機のどこに装甲をつけたら効果的か」

与えられたデータ

実際に戦闘に出撃した戦闘機のうち、帰還した戦闘機から弾痕のデータを取ったもの。
どの部分に装甲をつけたら敵からの銃撃から戦闘機をより効果的に守れるか?

機体の部分	平方フィートあたりの弾痕数
エンジン	1.11
胴体	1.73
燃料系統	1.55
その他	1.8

見えない弾痕

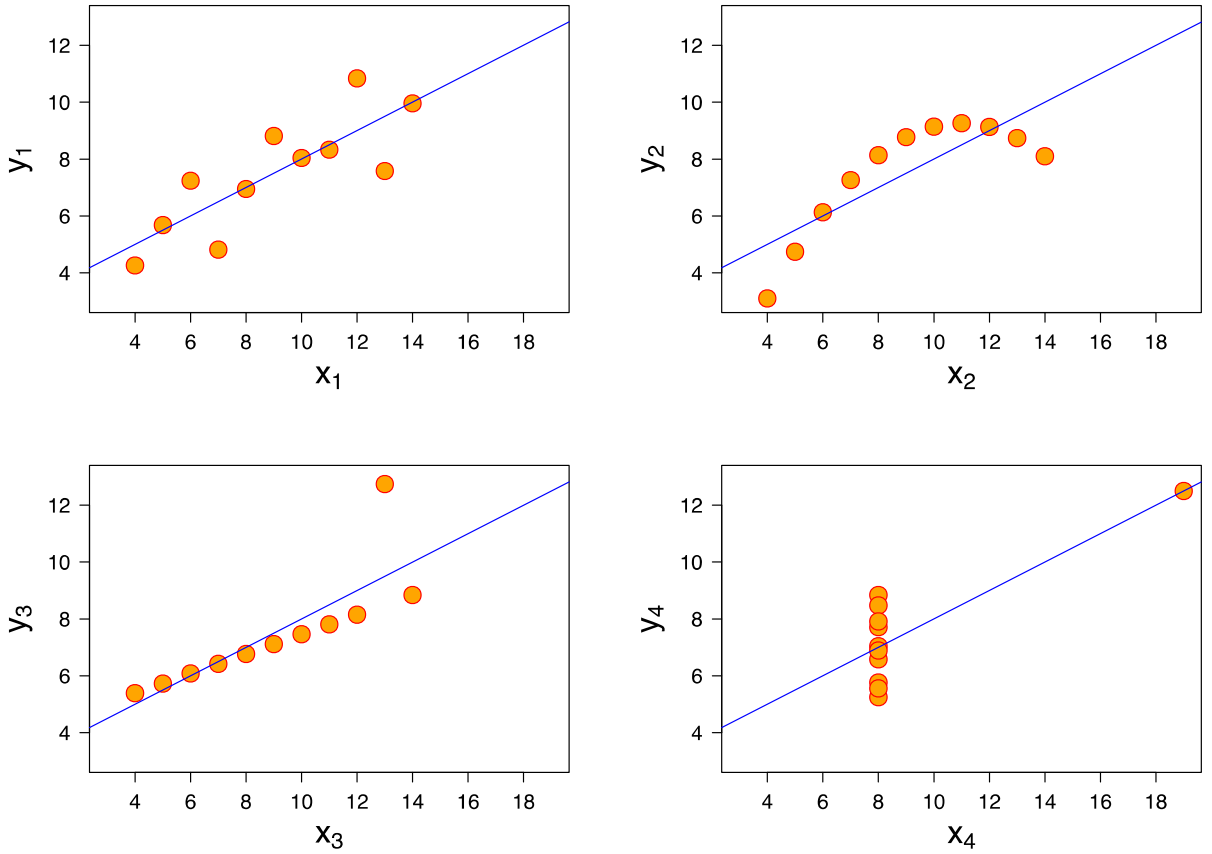
- Waldの勧告:

  「エンジンに装甲をつけろ」

  ...一見すると被弾が他の部分より少ないように見えるが、これはエンジンに被弾した場合他の部分を被弾した場合より多く撃墜され帰還できなかったということ。
- データのサンプリングが母集団からの無作為抽出でなく、何らかのバイアスがある場合は注意

Anscombe's quartet

傾向の違うデータに線形回帰を当てはめても、違いが出ないケース。



スプレッドシートにおける浮動小数点数計算の誤差についての研究

- 200X年代には、OOo Calcの方がExcelよりも統計関数の精度が良かったという研究:

  http://www.tandfonline.com/doi/abs/10.1198/tas.2011.09076
- 現在はExcelの実装も改善してきている:

  http://homepages.ulb.ac.be/~gmelard/rech/compsstat37.pdf
- 今後の調査にも期待

関連: 表示される数値の精度

LibreOffice 5.4の新機能

https://wiki.documentfoundation.org/ReleaseNotes/5.4#Calc

Calculate with "Precision as shown" option (Laurent Balland-Poirier) now works also with

- fraction format tdf#105657
- several subformats tdf#106052
- engineering notation tdf#106252
- thousands divisors tdf#106253

Calcで統計解析をする場合の注意点

- 統計関数はExcel互換の仕様になっていることが多い
  - データ中のテキストの扱い、空文字やゼロの扱いがややこしい
- 進んだ解析は今後の課題
- 浮動小数点数計算の精度はそれなり

お勧めの統計解析ステップ

1.データをCalcで予備解析

　手軽さを活かす。

2.興味深い統計量が得られたらR等で本番解析

- 予備解析で見た数字を再現できるかどうかを確認。
- 確認できたら先に進む。

3.予備解析で見た数字と齟齬があったら、理由を検証

- もしCalcの癖が原因なら、記録しておくと後々に役立つ。
- もしCalcのバグのせいなら、bugzillaに報告する。

4.1に戻る

まとめ

- Calcは手軽に統計解析するのに便利。
- 母集団からのサンプルしたデータで推測する場合は、どのようにして得られたデータかという前提に基づいて仮定を定めるべき。
- Calcの統計関数がよく分からない振舞いをするときは、仕様を確認するとき。
- LibreOfficeのバージョンが上がるにつれて、より便利になってきている。

参考資料

- 統計の初歩についてのお薦めのテキスト
  - 宮川公男「基本統計学[第4版]」(有斐閣)
  - 涌井良幸、涌井貞美「統計解析がわかる」(技術評論社)
- お薦めの読み物
  - Jordan Ellenberg (2014) "How Not to Be Wrong: The Power of Mathematical Thinking," THE PENGUIN PRESS

    邦訳:「データを正しく見るための数学的思考」(日経BP社)
  - Michael Blastland and Andrew Dilnot (2007) "The tiger that isn't: seeing through a world of numbers," Profile Books

    邦訳:「統計数字にだまされるな―いまを生き抜くための数学」(化学同人)
- 統計解析の検証に使える統計データ
  - Statistical Reference Datasets (StRD)

    http://www.itl.nist.gov/div898/strd/