推定と検定

(復習)記述統計と推測統計

- 統計解析は大きく２つに分けられる
  - 記述統計
  - 推測統計
- 記述統計
  - 観察集団の特性を示すもの
  - 代表値（平均値や中央値）や、ばらつきの指標（標準偏差など）
  - 図表を効果的に使う
- 推測統計
  - 観察集団のデータから母集団の特性を「推定」する
    - 平均／分散／係数値などの推定(点推定)
    - 点推定値のばらつきを調べる(区間推定)
  - 検定統計量を用いた検定

全数調査と標本調査

- 全数調査(国勢調査，人口動態統計など)
  - 母集団全員に調査をしてデータを得る
  - コストがかかる
    - 平成12年度の国勢調査→約690億円!
  - 集計結果に標本誤差は含まれない(精度が高い)

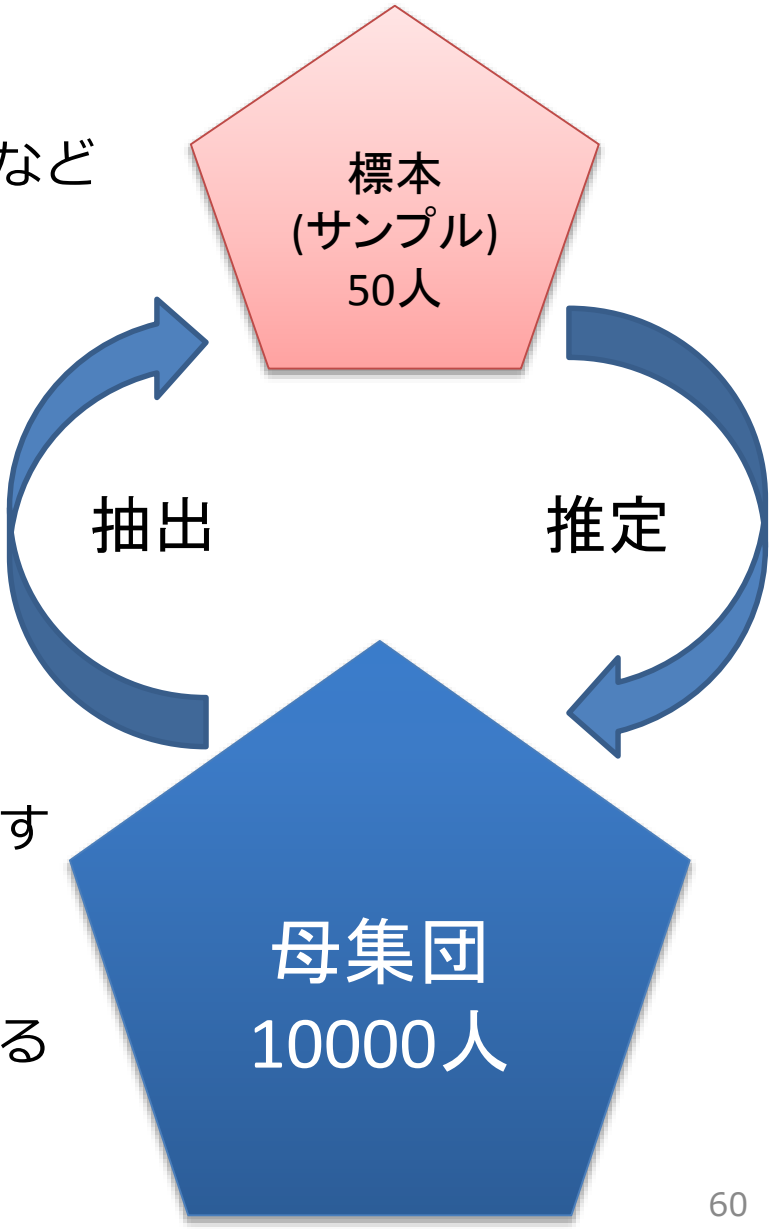
  精度は高いがコストが膨大

  →精度はそこまで高くなくていいからコストを抑えたい→標本調査

- 標本調査(サンプリング調査)
  - 母集団の一部(標本)に調査をしてデータを得る
  - 全数調査と比較するとコストが低い
  - 集計結果に標本誤差が含まれる

標本調査から母集団の特性を推定する

- 母数（パラメータ）
  - 母集団の特性値(平均，分散，相関係数など)
- 推定
  - 標本のデータから母数(パラメータ)を推し量ること
- 推定には大きく分けて2種類
  - 点推定
    - 母集団の特性値に最も近い値を推定する
  - 区間推定
    - 点推定値の誤差やばらつきを推定する



いろいろな点推定値

- 母平均(母集団での平均)の点推定値
  - 標本平均
    - 標本調査のデータから計算できる平均
- 母比率(母集団での比率)の点推定値
  - 標本比率を使う
    - 標本調査のデータから計算できる比率
- 母分散(母集団での分散)の点推定値
  - 不偏分散
    - 分散を計算する時の分母にN(標本数)-1を使ったもの
- 母標準偏差(母集団での標準偏差)の点推定値
  - 不偏標準偏差
    - 不偏分散の平方根をとったもの

点推定の例

学生	得点 (x)	偏差 (x−m)	偏差^2 $(x-m)^2$
A	61	−9	81
B	74	4	16
C	55	−15	225
D	85	15	225
E	68	−2	4
F	72	2	4
G	64	−6	36
H	80	10	100
I	82	12	144
J	59	−11	121
平均(m)	70	不偏分散(s^2)	106.2
		不偏標準偏差(s)	10.3

- 左の１０人のサンプルの例・・・高校生全国共通試験を受けた人のうち，１０人分のデータ

- 標本平均＝70.0
- 不偏分散＝106.2
- 不偏標準偏差＝10.3

推定には誤差がつきもの

- 誤差を定量化したい－標準誤差(Standard Error; SE)
  - 点推定値の分布のばらつき
  - 何度も母集団からサンプリングした時の、点推定値の標準偏差

母標準偏差がわかっている場合

$$\text{標準誤差} = \frac{\text{母標準偏差}}{\sqrt{\text{標本数}}}$$

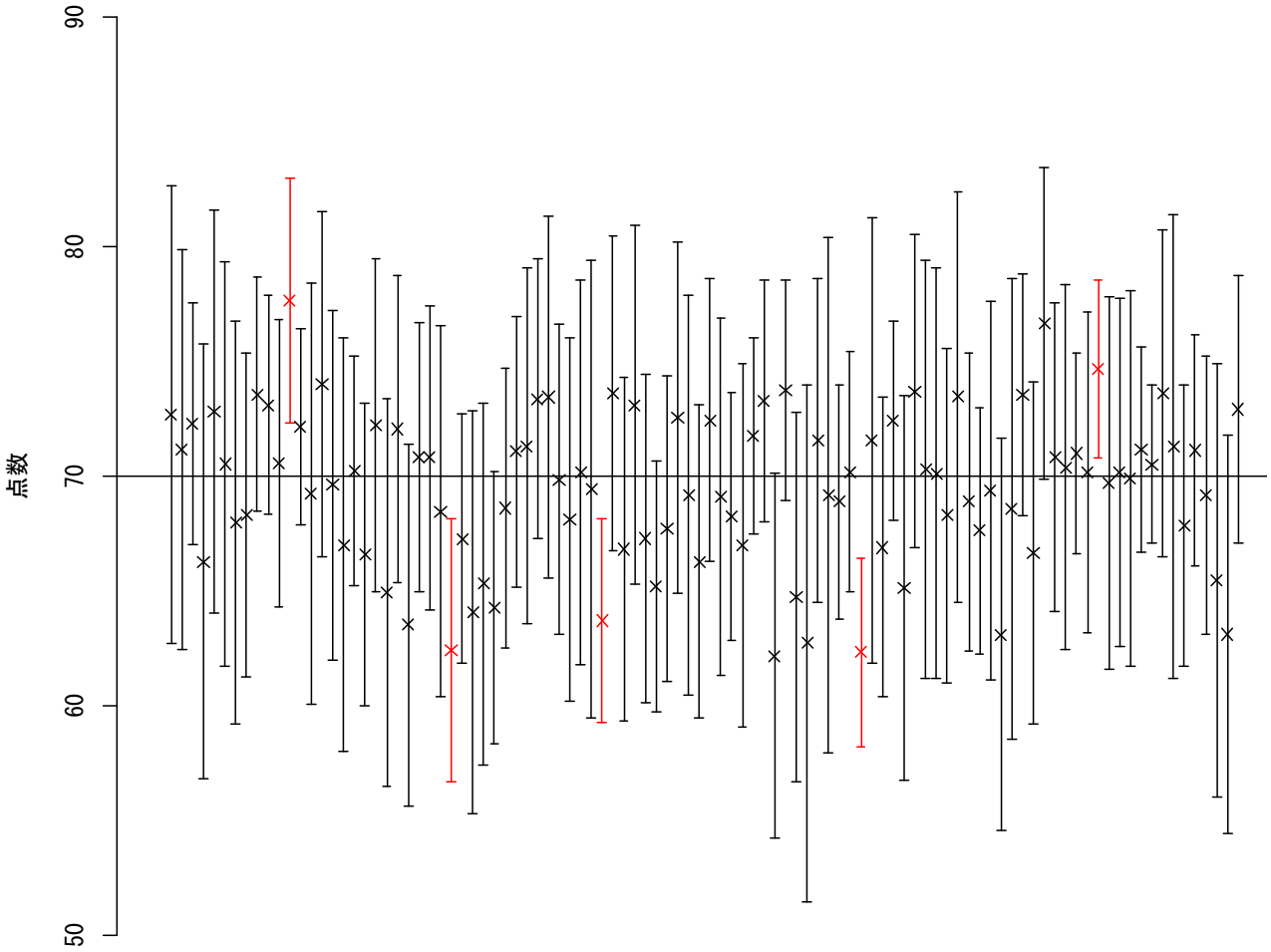
母標準偏差がわからない場合
→母標準偏差の代わりに母標準偏差の推定値の不偏標準偏差を使う

$$\text{標準誤差} = \frac{\text{不偏標準偏差}}{\sqrt{\text{標本数}}}$$

標本数(サンプル数)が多くなるほど標準誤差は小さくなる

区間推定・信頼区間

- 区間推定
  - 母数が入る区間を推定
- 信頼度
  - 区間推定が的中する確率
  - 区間推定をする際に自分で決める
  - 90%、95%、99%が使われることが多い
- 信頼区間
  - 区間推定で求められる区間
  - 信頼度と合わせて〇〇%信頼区間という使い方をする
  - 信頼度が95%の信頼区間なら95%信頼区間
  - 同じ母集団から同じ数の標本を抽出して区間推定することを繰り返した時に、信頼度の確率で母数が含まれる区間
    - 先ほどの共通試験の例でいうと、全受験者から10人分のデータを抽出してきて95%信頼区間を出すのを100回繰り返すと5回は95%信頼区間に全受験者の平均点が含まれない
  - 信頼区間が狭いほど、推定の精度が高い



10000人の母集団から
10人の標本を抽出して
95%信頼区間を出すことを
100回繰り返した時の例

母集団での平均は70

×印は標本平均
上下の線は信頼区間

赤色のものは信頼区間が母集団の平均値を含まないもの

100回中5回、95%信頼区間に母平均の70が含まれていない

平均値の区間推定－母分散がわかっている場合

- 平均の点推定値Mの95%信頼区間は

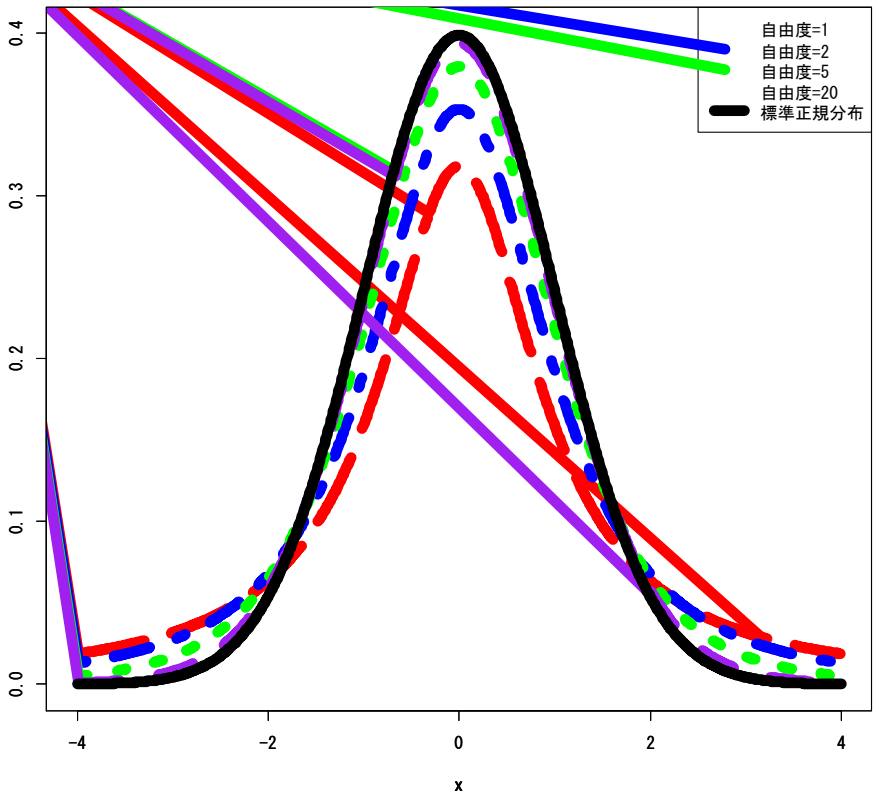
  （M-1.96×標準誤差～M+1.96×標準誤差）

  - この信頼区間は母集団での母分散を用いて標準誤差を出し、区間推定をしている

- ふつうは母集団の分散はわからないことがほとんど
- そのとき、母集団の分散の推定値として分散を計算するときに、偏差の2乗の和をn-1で割って計算すると母集団の分散の推定値になると言われている・・・「不偏分散」という

平均値の区間推定-母分散がわからない場合

- 不偏分散を使って平均が0、分散が1になるように標準化すると標準正規分布でなくt分布と呼ばれる正規分布に似た分布に従う
- T分布は標本数nから1を引いた値（自由度）によって形が異なる
  - 自由度が無限大になると標準正規分布になる
  - 自由度30くらいではほぼ正規分布と考えて良い



95%信頼区間は

$m-t_{n-1}(2.5\%)\times SE \sim m+t_{n-1}(2.5\%)\times SE$

平均値の区間推定の例

学生	得点 (x)	偏差 (x−m)	偏差^2 (x−m)^2
A	61	−9	81
B	74	4	16
C	55	−15	225
D	85	15	225
E	68	−2	4
F	72	2	4
G	64	−6	36
H	80	10	100
I	82	12	144
J	59	−11	121
平均(m)	70	不偏分散(s^2)	106.2
		不偏標準偏差(s)	10.3

- 不偏分散は106.2
- 点推定値は70
- 標準誤差＝ $10.3 \div \sqrt{10} \fallingdotseq 3.3$
- $t_9(2.5\%)=2.26$ (t分布表より)
- 95%信頼区間は (62.5～77.5)

※tの値は古くはt分布表を参照したが、近年はコンピュータが計算してくれる

信頼区間の幅

- 信頼区間は基本的には
  - 点推定値±1.96×標準誤差で囲まれた区間
- 標準誤差が小さくなると区間は狭くなる
- 点推定値が10、不偏分散が8
  - n=4　　(2.16～17.84)
  - n=9　　(4.77～15.23)
  - n=25　(6.86～13.14)
  - n=400 (9.22～10.78)
- 信頼区間の幅が狭くなるほど正確な推定になる

→標本数(サンプルサイズ)が大きくなると正確に推定できやすい

仮説(統計的)検定

- 母集団の特性についての予想(仮説)が正しいか間違っているかを標本調査のデータから判断する方法
  - 母集団全体のデータが取れる全数調査では検定は必要ない

- 仮説が正しいかどうかをどのように判断する？
  - 日本に住んでいる人は男性と女性どちらが多いか？
    - 住んでいる人全員の性別を調べる(→国勢調査 690億円)
  - もう少しコストを抑えて判断したい
    - 標本調査のデータから判断する(仮説検定)
    - 十分な数の標本を母集団から無作為に抽出すれば一定の精度で可能
    - 判断が間違っていることもある
      →αエラーとβエラー(後述)

帰無仮説と対立仮説

- 帰無仮説(H_0)
  - 母数の値を明確に指定する仮説
  - 帰無仮説の例
    - 新しく開発した血圧を下げる薬の効果は従来の薬の効果と同じ
    - 日本に住んでいる男性の比率と女性の比率は同じ(男女比は1：1)
- 対立仮説(H_1)
  - 帰無仮説の正反対の内容の仮説
  - 対立仮説の例
    - 新しく開発した血圧を下げる薬の効果は従来の薬の効果と同じではない
    - 日本に住んでいる男性の比率と女性の比率は同じではない

仮説検定の考え方

- 対立仮説が正しいことを直接示すのは難しい
  - 母数がひとつに定まっていない
    - 新しく開発した血圧を下げる薬の効果は従来の薬の効果と同じではない
      - 新しい薬のほうが10mmHg血圧が下がる、新しい薬のほうが20mmHg血圧が下がる、、、
    - 日本に住んでいる男性の比率と女性の比率は同じではない
      - 男性40%女性60%、男性70%女性60%、、、

→母数が定まっている帰無仮説が正しいかどうか検討して、正しくなければ対立仮説が正しいということにしよう

- どうやって帰無仮説が正しいかを検討する?
  - 帰無仮説が正しいと仮定して、標本のデータが偶然得られる確率(=有意確率)を計算
    - 確率分布がわかっている指標(=検定統計量)を使って有意確率を計算
  - 有意確率が一定水準(=有意水準)を下回ったら帰無仮説が間違っていると判断する
    - 有意水準は5%がよく使われる

仮説検定の手順

1. 帰無仮説を設定する
   - 帰無仮説を設定すれば対立仮説も決まる
2. 有意水準を決める
3. 検定統計量を計算する
   - 検討する仮説によって統計量は変わる
     - 平均の差を検定する場合は t 値、クロス表の検定ならカイ二乗値など
4. 検定統計量から有意確率を求める
5. 帰無仮説を棄却するか判断する
   - 有意確率が有意水準を下回れば帰無仮説を棄却(＝帰無仮説が間違っていると考える)
   - 有意確率が有意水準以上であれば帰無仮説を採択(＝帰無仮説が間違っているとはいえない)

検定で生じる2つの誤りの確率

- 第1種の過誤(αエラー)
  - 帰無仮説が本当は正しかったが棄却してしまったこと
  →母集団では「差がない」のに「差があった」としてしまった
  - 第1種の過誤が生じる確率をα（アルファ）という
  - アルファは有意水準と同じ

- 第2種の過誤(βエラー)
  - 対立仮説が正しかったが帰無仮説を棄却できなかったこと
  →母集団では「差がある」のに「差がない」としてしまった
  - 第2種の過誤が生じる確率をβ（ベータ）という
  - 通常の検定ではベータはあまり相手にされない

両側検定と片側検定

- 「差がない」という帰無仮説の対立仮説には2通りが考えられる
  - 日本に住んでいる男性の比率と女性の比率は同じではない
    - 男性比率＜女性比率
    - 男性比率＞女性比率

- 両方共ありうると考えて検定をするのが両側検定
  - 一般には両側検定を行う

- どちらか一方しかありえないと考えて検定をするのが片側検定
  - 両側検定より帰無仮説が棄却されやすい