第2回 記述統計

2007年9月18日

データの平均・期待値

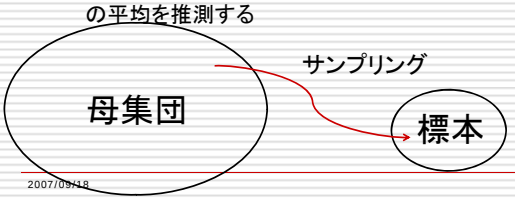
- このクラスでテストをした時の平均
  - 全員の点数を足して，人数で割る
- サイコロを振った時に出る目の平均（期待値）
  - 500回くらいサイコロを振って出た目を平均する
  - それぞれの目は1/6の確率で出るので計算で求める
- 20歳男性の身長の平均

母集団と標本

- 20歳男性の身長の平均
  - 20歳男性を全員集めてきて身長を測定する！
    - 不可能とは言わないが現実的ではない
  - 通常は標本抽出（サンプリング）を行って標本調査を行う
    - 真の平均とはずれている可能性はあるが，だいたいの値はわかる
      - 推定・検定の話につながる

サンプリング

- つまり
  - 「20歳男性」という母集団があり，
  - その中から一部を抽出して標本とし，その標本の「身長」を測定
  - 標本の平均（標本平均）を計算することで母集団の平均を推測する



基本的な統計量（1）

- 中心位置を推測するための統計量
  - 平均
  - メディアン（中央値）
    - データを小さい順に並べたときに中央にくる値
  - 四分位（第1四分位，第3四分位）
    - データを小さい順に並べたときに，25%，75%にくる値
  - モード（最頻値）
    - データで最も現れる回数の多い値

基本的な統計量（2）

- 広がり具合を推測するための統計量
  - 分散
    - 平均からのずれ
  - 標準偏差
  - 範囲（レンジ）
    - 最大値－最小値

平均・分散

- データを x_i
  - 平均 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - 分散 $V = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - 標準偏差 $s = \sqrt{V}$
    - 皆さん大好きな(？)シグマの計算がここで出てくる

難しい話は置いといて

- 実際は・・・
- 母集団や標本を意識することより，測定によって与えられたデータを目の当たりにすることの方が多い(気がする)
  - そのデータで様々な統計量を求めることが多い
- 実際のデータでいろいろ計算をしてみる
  - http://www.ae.keio.ac.jp/~satoru_y/shouei/02.xls
  - 教科書17ページの問題1.1の1のデータ

データを見てまず最初に

- データの様子を知るためにまず度数分布表・ヒストグラムを作成する
  - 度数分布表・ヒストグラムでデータの形状を見る
    - 何か見えてくるものがある・・・かも
- 平均・分散などの統計量を計算する
  - 度数分布表・ヒストグラムではわからない傾向を知る

度数分布表

- 度数分布表の作り方
  - データをいくつかの区分に分割する
    - 区分の数はいくつかの基準があるが，おおむね$\sqrt{n}$が一つの目安(であることが多いようである)
      - テキストにはスタージェスの公式が記載されている
  - その区分に入るデータの数を数える
    - 「度数」と呼ぶ
  - ちなみに，区分のことを「階級」，区分の幅のことを「階級の幅」，区分の平均値を「階級値」と呼ぶ
  - データの個数のことは「標本の大きさ」「サンプルサイズ」などと呼ぶ

度数分布表・ヒストグラム

- 度数分布表では，各階級の度数のほか，
  - 相対度数(度数を標本の大きさで割ったもの)
  - 累積度数(そこまでの階級に現れた度数の和)
  - 累積相対度数(累積度数を標本の大きさで割ったもの)

  を計算することが多い
- 度数分布表を棒グラフにしたものがヒストグラム

度数分布表・ヒストグラムの際に使うExcelの関数

- 頻度を計算するFREQUENCYや階級の数を数えるために使っているCOUNTIFなどの関数がある
- 実際は分析ツールのヒストグラムを使ったほうが楽！
  - 度数分布表・ヒストグラムを描くときには事前に階級の値を用意しておく必要がある
    - 90，100，110と用意した場合には，
      90以下の度数，90より大100以下の度数，100より大110以下の度数，110より大の度数
      が求まる
      - 通常は90以上100未満が普通の気もするが・・・

度数分布・ヒストグラムを眺めて・・・

- どこの階級の度数が一番多いか
- どこら辺が真ん中か
- ヒストグラムの形はどうか
- 最大・最小はどこらへんか

などいろいろわかる

- 細かい値を知りたい場合は，平均，分散などの統計量を求める

統計量の計算に用いるExcelの関数

- 平均 AVERAGE
- 分散 VAR
- 標準偏差 STDEV
- 最大値・最小値 MAX, MIN
- 中央値 MEDIAN
- 四分位 QUARTILE（最大値・最小値・中央値としても使うことができる）
- 最頻値 MODE

分析ツールの基本統計量を使ってもよい

統計量と度数分布・ヒストグラムの比較

- 実際に求めた各統計量を基に，度数分布表やヒストグラムを見直してみる
  - 実際に求めた平均や中央値とヒストグラムで予測した値は同じか違うか
  - ヒストグラムで度数が一番大きかった階級と実際に求めた最頻値は同じか違うか
  - ヒストグラムの形状と分散（標準偏差）との関係はどうか
  - etc.

統計でよく使う記号について（補足）

- データ x_i
- 母集団の平均（母平均） μ
- 母集団の分散（母分散） σ^2
- 母集団の標準偏差（母標準偏差） σ
- 標本平均 $\bar{x}$
- 標本分散 V
- 標準偏差 s
- 標本の大きさ n

分散について（補足）

- 分散の式はテキストによっては

$$V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$ となっている

  - これに対応するExcelの関数はVARP（標準偏差はSTDEVP）
  - 記述統計ではよく見られる
    - 教科書の1章ではこの式を，6章ではn-1で割った式を用いている
  - なぜ2種類あるのかはいろいろあるが，今後，母集団の推定・検定を行う上では，n-1で割った分散（不偏分散）を使うことが多い

参考文献

- 勝野・井川，Excelによるメディカル／コ・メディカル 統計入門，共立出版（教科書）
- Excelを使った統計の本
  - 酒折・大石・竹内，Excel徹底活用 推測統計入門，秀和システム
- 統計学の入門書
  - 永田靖，入門 統計解析法，日科技連