経済統計概論　2

データの整理 (1)

コラム　1

- シュワちゃん知事になる！
  - 投票は8日正午頃終わった．そして，それから程なく，FOXなど4大ネットワークは，「リコール成立，シュワ知事誕生」と報じた．
  - なぜ開票が始まるか，始まらないうちに，マスコミは当選や当選確実を報じることができるか？
  - 出口調査が鍵である．
    - 出口調査は，投票所の出口で全有権者のうちの少数 (おそらく1500人程度) に誰に投票したか聞く．
    - その結果を統計学を使って解析．
    - 誤差を考慮しても，リコール成立，かつ，シュワの得票率が最大となる確率が非常に高いとなれば，その時点で当選と報じる．

コラム　2

- 今年のノーベル記念経済学賞
  - ロバート･エングル
  - クライブ･グランジャー
- 時系列分析
  - 株価，GDPなどのデータの解析法
  - エングル
    - ARCH (アーチ) Model　　分散の変動をモデル化
  - グランジャー
    - グランジャー因果性　　データ間の先行遅延関係を検出
  - 共同
    - 共和分関係の検出法
      複数のデータ間の線形関係 (Y=aX+b)を検出

0．今日の目標

- 様々な代表値を理解する
- データの分布，散らばりを把握するためのデータ整理法を理解する
  - 度数分布，ヒストグラム

1．　データの中心の代表値
1．1　平均

- 平均の例
  - 四月の統一地方選で当選した府議 (百十二人) と大阪市議 (八十九人) の資産が七日、公開された。一九九六年施行の資産公開条例に基づき、土地･建物の面積や課税標準額、預貯金などを報告した。有価証券を除いた土地･建物の課税標準額と預貯金、金銭信託を合計した**平均**額は府議が二千八百五十二万円、市議が四千百九十万円。一億円を超えた府議は九人、市議は八人いた。 (2003年10月8日 読売大阪版)

1．2　中央値

- 中央値の例
  - [東京，2003年9月2日] **主要エコノミスト、欧州中銀の年内再利下げを予想～ロイター調査、次回会合では様え置き観測が大半～**
    - ロイターが主要エコノミスト62人に実施した調査によると、欧州中央銀行 (ECB: European Central Bank) が今年後半に再度の金利引き下げを実施するとの予想が半数以上に達しました。しかし、経済の先行きに楽観的な見方が出ているため、大幅な引き下げの可能性 は縮小しています。．．．一方，調査に答えたエコノミストのうち25人は、世界の株価が上昇し、米国経済の回復を示すデータが増えていることを踏まえ、金利がすでに2%で底を打ったと予測しています．2003年末時点の金利水準については1．0%から2．0%まで予測の幅があり、**中央値**は1．75%でした．2004年末に関しては0．75%から3．25%に分布しており、**中央値**は2．25%となっています。
- 中央値
  - データを大きさ順に並べて真ん中をとる．しかし，データ数が偶数の場合は，真ん中のデータ2つの平均．
  - 中位数，メディアンとも呼ばれる．
- なぜ例では平均ではなく中央値をとったのか？
  - 平均の場合，極端な予測のエコノミストの影響を受けてしまう．しかし，中央値の場合，極端な予測には影響を受けない．

1．3．最頻値

- 最頻値の例
  - PC Watch読者環境調査2003年7月
    - 今回の調査では、CPUクロックの上昇がめざましく、初めて「～2.4GHz」が最頻値となった。また、HDD、メモリの大容量化傾向も続いている。周辺機器では、マルチフォーマット対応の書き込み/書き換えDVDドライブと、17インチLCDディスプレイが大幅に伸びている。接続環境はADSL/CATVの普及が一服したが、FTTHは引き続きポイントを伸ばしている。
    - **[今回の最頻値]**
      - 所有台数 2台、自作のデスクトップPC/AT互換機
      - OS：Windows XP Professional
      - CPU：Pentium 4
      - クロック：～2.4GHz
      - メモリ：～512MB
      - HDD：～200GB
      - ディスプレイ：17インチCRT、解像度～1,280×1,024ドット
      - インターネット接続：xDSL(フレッツADSL除く)
      - 家庭内LAN：構築している

    (http://pc.watch.impress.co.jp/docs/2003/0806/eng23.htm)
- 最頻値
  - データのなかでもっともよく現れた値．
  - 極端な値の影響を受けづらい
  - 数値化されていないデータでも得ることができる．

1．4　平均，中央値，最頻値

- 平均
  - 極端な値（はずれ値，異常値）の影響大
  - 計算が容易
  - 分散との関わり
- 中央値
  - はずれ値（異常値）の影響を受けない
  - 一度順序で並べなければいけない．多数のデータの場合大変
  - 平均偏差との関わり
- 最頻値
  - はずれ値（異常値）の影響を受けない
  - 最頻データ以外を完全に無視している
- 3つの関係
  - ヒストグラムの山が1つのとき，平均－最頻値　3×（平均－中央値）＜ピアソンの式＞がほとんどの場合成り立つ．
  - 中央値≒(2平均+最頻値)/3，つまり，中央値は平均と最頻値の間を1：2に内分する．

2．散らばりの代表値
2．1　分散，標準偏差，範囲

- 標本分散（標本が調査対象の一部の場合）

$$\frac{(\text{データ}-\text{平均})^2\text{の合計}}{\text{標本数}-1}$$

- 分散（標本が調査対象と一致する場合）

$$\frac{(\text{データ}-\text{平均})^2\text{の合計}}{\text{標本数}}$$

- （標本）標準偏差　$\sqrt{(\text{標本})\text{分散}}$
- 分散，標準偏差になると経済ニュースにさえも出てこない．これでいいんだろうか？
- 範囲＝最大値ー最小値

2．2．チェビシェフの不等式

- 標準偏差からデータの分布がある程度分かる
- チェビシェフの不等式（どんなデータにも当てはまる）
  - $|\text{データ値}-\text{標本平均}| \geq k \times \text{標本標準偏差}$ を満たすデータの割合は全標本数の $1/k^2$ 以下である．
  - 例えば標本平均から標準偏差の2倍以上離れているデータは全体の1/4以下である．
    - 偏差値70以上は絶対1/4以下しかいない

2．3．その他散らばりの代表値

- 四分位点（しぶんいてん）
  - 第1四分位点
    - 順序どおりに並べ替え，データを4分割する．その4分割点のうち最も小さい値．もし，その分割点にデータがなければ，補間近似する．具体的には$(n+1)/4$ の整数部番目のデータとその次のデータの半分，または，$(n+1)/4$ の小数部を重みにした加重平均
  - 第2四分位点　中央値と同じ
  - 第3四分位点
    - 順序どおりに並べ替え，データを4分割する．その4分割点のうち最も大きい値．もし，その分割点にデータがなければ，補間近似
  - 四分位範囲　第3四分位点－第1四分位点
- 10分位点　データを10分割
- パーセント点（百分位点）　データを百分割

2．4　例の解説

- 例1．1
  - 教科書参照
- 例1．2
  - 中間階層はどこの地域でも似たり寄ったりの収入である．
- 例1．3
  - ピアソンの式から各地域の貯蓄額の最頻値の近似値を求めてみよう．
  - 例　北海道　最頻値≒平均－3(平均－中央値)

$$=1007-3(1007-698)=80\text{万円}$$

3．度数分布表とヒストグラム
3．1　作成法

- 作成法は教科書参照
  - ヒストグラムの面積は度数を表す．
- 階級数の決め方
  - スタージェスの公式　　$1+3.3\log_{10} n$
    nは標本数（データ数）
- 図1．7の解説
  - 平均－最頻値　3×（平均－中央値）＜ピアソンの式＞が成立していない．このことは，貯蓄の分布のヒストグラムの山が1つではないなど，かなり変形された分布であることを示している．

3.2 度数分布表による代表値の計算

- 平均

$$\frac{(\text{階級値}\times\text{度数})\text{の合計}}{\text{標本数}}=(\text{階級値}\times\text{相対度数})\text{の合計}$$

- 標本分散

$$\frac{\{(\text{階級値}-\text{平均})^2\times\text{度数}\}\text{の合計}}{\text{標本数}-1}$$

3.3 度数分布表とパーセント点算出
3.3.1　算出法

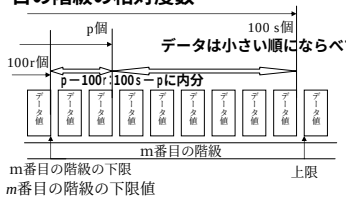
- pパーセント点
  - m番目の階級で初めて累積相対度数がp/100を超えるとする．

$$m\text{番目の階級の下限値}+\left\{\frac{\text{標本数}\times p}{100}-(m-1\text{番目の階級の累積度数})\right\}\times\frac{\text{階級幅}}{m\text{番目の階級の度数}}$$

- 中央値は50%点，第1四分位点は25%点

3．3．2　説明

- ある階級に属すデータが階級内に均等分布
  - m-1番目の累積相対度数はrとm番目はs，s-rはm番目の階級の相対度数



m番目の階級の下限値

$$+\frac{p/100-(m-1\text{番目の階級の累積相対度数})}{m\text{番目の階級の相対度数}}\times\text{階級幅}$$

3．4　度数分布表による最頻値算出

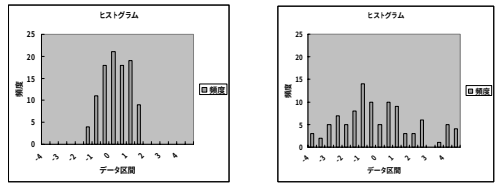
- 二つの方法

$$\text{最大度数の階級の下限値}+\frac{\text{その後の階級の度数}}{\text{前後の階級の度数の合計}}\times\text{階級幅}$$

$$\text{最大度数の階級の下限値}+\frac{\text{最大度数の階級とその直前の階級の度数の差}}{2\times m\text{番目の階級の度数}-\text{前後の階級の度数の合計}}\times\text{階級幅}$$

3．5　データの散らばりとヒストグラム

- 散らばりが小さいデータのヒストグラムと大きいデータのヒストグラム



- 今後はデータの散らばりをヒストグラムの形状で表すことがある

4． 発展したデータの代表値 (1)

- 刈り込み平均
  - 最大値，最小値，およびその周辺の値を平均値の算出からはずす
- 加重平均

$$\sum_{i=1}^{n} w_i x_i \qquad \sum_{i=1}^{n} w_i = 1, \quad w_i \geq 0$$

  - 平均は重み $w_i = 1/n$ となっている加重平均
  - 刈り込み平均は最大値，最小値，およびその周辺の値に関する重みを0にした加重平均

4．発展したデータの代表値 (2)

- 幾何平均
  - 成長率などその効果が積によって累積していく指標の平均に利用される
  - $\sqrt[n]{\text{データ全部の積}}$
  - 容易に計算する方法

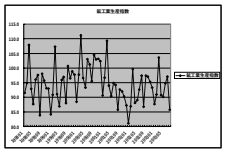
$$\log \text{幾何平均} = \log \sqrt[n]{\text{データ全部の積}}$$

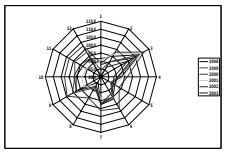
$$= \frac{1}{n}\log(\text{データ全部の積}) = \frac{1}{n}\{\log(\text{データ})\text{の合計}\}$$

    - $10^{(\log_{10}\ \text{データ})\text{の平均}}$
    - 度数分布表を使用する場合は，log値の平均を出す

5．季節調整
5．1 季節変動と季節調整

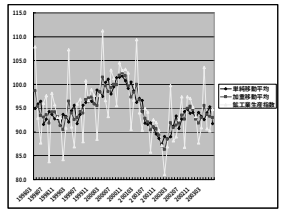
- 季節変動
  - 鉱工業生産の場合1月8月，5月に落ち込む
    - 休みなどのため
  - これでは，景気の指標として鉱工業生産指数を利用することはできない．
- 季節調整
  - 季節変動を取り除く





5．2 季節調整と移動平均

- 季節調整
  - 単純移動平均
    - あるデータと前後同一数のデータの平均
    - 例えば前後2期のデータも考えて平均をとる
  - 加重移動平均
    - 単純平均だと前後のデータも同一の比重で考慮
    - 現在のデータに比重をおく
    - 単純平均ではなく加重平均をとる
    - 加重の仕方には色々ある．（教科書は一例）
  - これらは単純な方法で初歩的分析に適する



6．変動係数

- 散らばりを比較する
  - 平均が同じくらい
    - 分散や範囲，四分位範囲を比較すればよい
  - 平均が異なるデータの比較
    - 平均で標準化した標準偏差＝変動係数 $\dfrac{\text{標本標準偏差}}{\text{標本平均}}$
    - 四分位分散係数
      - 四分位範囲の半分を中央値で割ったもの
        - » 四分位の半分は片側の変動幅を示す
    - 十分位分散係数

$$\frac{\text{第９十分位数} - \text{第１十分位数}}{2 \times \text{中央値}}$$