経済統計分析 3
よく使う記述統計量

事務連絡

- Webclassを使ってみようと思います.
  - 登録できる人はしておいてください.
  - 宿題をwebclass経由で回収・返却する予定です.
  - じつはすでにデータをアップロードしています.

- MS-Word, Excelが使えますか?
  - VBAとかできなくてもいいです.
  - 宿題をこれらで出していただけると, 採点しやすいです.
  - 互換機能(校閲機能含む)があればいいです.

今日のおはなし.

- 記述統計, ただし1変数, ちょっと2変数
  - データの状況をおおまかに表す／伝える
  - たくさんあるデータをいくつかの数値で代表して表現する
- 「ふつー」ってなんだ?
  - いくつかの「平均」
- 指数, ふたたび
- 散らばりと分位点, 不平等度尺度

- 今日のタネ
  - 中村隆英ほか. 1984.『統計入門』東大出版会, 第3章
  - 飯田泰之. 2007. 考える技術としての統計学. NHKブックス1101.

見ただけで分かるか.

- あるひとつの事柄についてのデータの状況を伝えたい
  - ある1変数の分布を伝えたい
  - ヒストグラムは視覚に訴える
  - 正確さを求めるなら, 度数分布表を用いる
- 度数分布表やヒストグラムでは?
  - 度数分布表はまだデータ量が多い
  - ヒストグラムは違いを表すにはよいが, 類似は示しにくい
- 記述統計
  - データの分布の状態をいくつかの数値で表現すること
  - それらの指標をまとめて「特性値」と呼ぶ
  - 「ふつう」と「ちらばり」をあらわす特性値が基本中の基本

「ふつう」もいろいろ.

- データの状況を数値1つで代表させるには?
  - 例:日本人の所得ってどれくらい?
  - 「ふつう」な値を1つ使う
  - それだけ「情報を捨てている」
- 「ふつう」をあらわすいくつかの指標
  - 平均値
    - 算術平均, 幾何平均, 調和平均
    - 加重平均
    - 切り落とし平均
  - 中位値／中央値
  - 最頻値

算術平均 average, mean

- 定義
  - 値の総和を観測値数（データのサイズ）で割ったもの

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - いわゆる「平均値」といえば，算術平均を指すことが多い
- 特徴
  - 個々の観測値の値が分からなくても，サイズと総和から計算可能
  - 例：1人当たりGDP = GDP / 人口
  - 逆に，平均とサイズから総和を計算できる
  - 「平均値」をもつ観測値は存在しない（ことが多い）
  - 例：試験の平均点が59.7点であっても，各点数は整数値
  - 「率」は質的変数の平均値と解釈できる

算術平均の性質

- 偏差の和がゼロ.
  - 偏差 = 各観測値と平均値との差

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

- 平均値の一次変換は, 一次変換の平均値に等しい

$$a\bar{x} + b = \overline{ax + b}$$

  - 平均値の計算の簡単化（暗算）によく用いられる
  - 例：点数の平均値を求めるとき
- 主体が同じであれば, 平均の和は和の平均に等しい

$$a\bar{x} + b\bar{y} = \overline{ax + by}$$

  - 例：平均収入額と平均支出額の差 = 平均黒字額

加重平均

- 「重みweight」をつけた和（加重和）
  - 重みの和が1になるようにしておく
    - 単純平均は，すべての重みが $1/n$ であるような加重平均
  - 例：2グループのそれぞれの単純平均がわかっているとき

$$\text{全体の平均 } \bar{x} = \frac{n_1}{n_1 + n_2}\bar{x}_1 + \frac{n_2}{n_1 + n_2}\bar{x}_2$$

- 度数分布からの平均値の計算
  - 階級内の平均値の，相対頻度をウェイトとする加重平均
  - 階級内平均値が分からないときには，階級値で代理

$$\text{全体の平均 } \bar{x} = \sum_{j=1}^{k}\left(\frac{f_j}{n}\right)\bar{x}_i$$

伸び率の平均は単純平均でいい?:幾何平均

原数値	伸び率	原数値	伸び率
100.00		100.00	
130.00	30.00	101.00	1.00
91.00	-30.00	99.99	-1.00
118.30	30.00	100.99	1.00
82.81	-30.00	99.98	-1.00
107.65	30.00	100.98	1.00
75.36	-30.00	99.97	-1.00

- 近似的に「伸び率の単純平均」が用いられることも多い.
- 「複利計算」の恐ろしさ

- 左の例では
  - 伸び率の単純平均:0%
  - 最後 / 最初÷6 = - 4.11%
  - 最後 / 最初の6乗根 = -4.61%
- 幾何平均
  - 積の n 乗根をとったもの
  - 一般に幾何平均のほうが小さい
  - 伸び率の平均値によく用いる
  - 対数変換値の算術平均に等しい

時速の平均のばあい?:調和平均

- 例:片道10kmの道のりを,行きは平均時速10kmで,帰りは平均時速5kmで往復したときの平均時速は?
  - 往復20kmに合計3時間かかっているから,6.7km
  - 算術平均(7.5km)より小さい
  - 一般に調和平均は幾何平均より小さい
- 定義

$$\text{幾何平均} = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$$

$$\text{調和平均} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

例：金融資産保有額

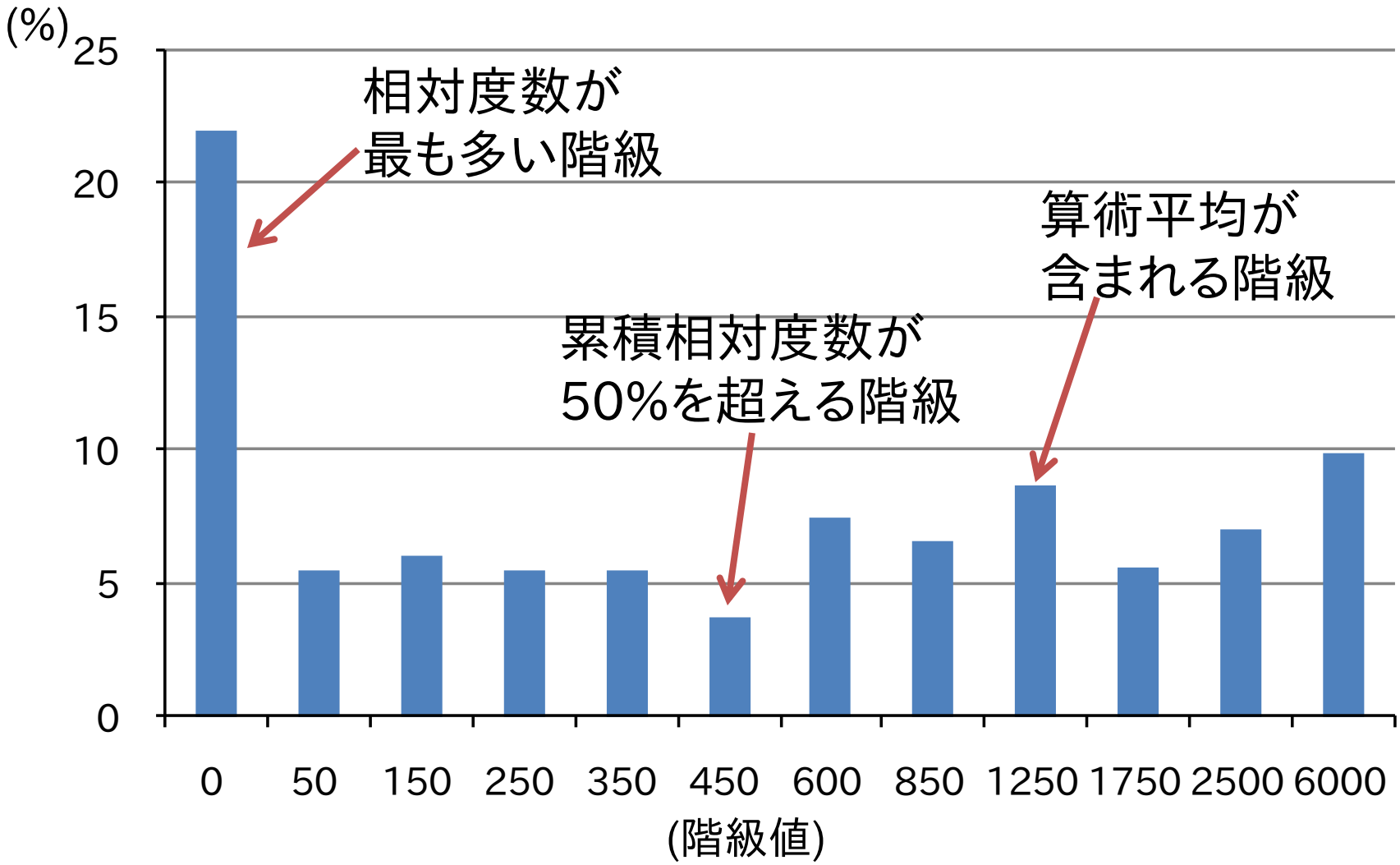
（日本銀行金融広報中央委員会，家計の金融行動に関する世論調査
［二人以上世帯調査］平成20年）

	頻度	相対頻度	階級値
0	858	22.08	0
0 - 100	213	5.48	50
100 - 200	237	6.10	150
200 - 300	212	5.46	250
300 - 400	215	5.53	350
400 - 500	145	3.73	450
500 - 700	291	7.49	600
700 - 1000	255	6.56	850
1000 - 1500	336	8.65	1250
1500 - 2000	220	5.66	1750
2000 - 3000	272	7.00	2500
> 3000	386	9.93	6000
N.A.	246	6.33	
合計	3886	100.00	1111.55

平均の計算では無回答（N.A.）は除去している．

- 公表されている平均値は1,152万円
- しかしそれは少し多いのではない？

例：金融資産保有額（続き）


(%)
25
20
15
10
5
0
相対度数が
最も多い階級
累積相対度数が
50%を超える階級
算術平均が
含まれる階級
0
50
150
250
350
450
600
850
1250
1750
2500
6000
(階級値)

「ふつう」を表す他の特性値

- 中位値, 中央値, median
  - データを大きさ順に並べたときの真ん中の値
  - 累積相対度数が50%になる観測値の値
  - 中位値からの偏差の絶対値を最小化する
- 最頻値, mode
  - 相対度数が最も大きくなる階級の階級値
- 平均値・中位値・最頻値の関係
  - ヒストグラムが左右対称ならすべて等しい
  - 右に歪んだ分布：最頻値 ＜ 中位値 ＜ 平均値
    - 所得・消費・資産など, 右に歪んだ分布は多い
    - 金融資産保有額の中位値は430万円

中位値によく似た他の特性値

- 中位値の別名:50%分位点
  - 「下」から数えて50%のところにあるから.
- q%分位点 percentile
  - 累積相対度数がq%になる観測値
  - 例:1%分位点より小さな値を取る観測値は全体の1%
- 四分位点quartile
  - 25%分位点が第1四分位点, 75%分位点が第3四分位点
- 十分位decile
  - 10%, 20%, …, 90%分位点のこと.
  - 公表統計では階級が十分位に分けられていることもある

外れ値 outlier

- 算術平均は極端な値の影響を受けやすい
  - 中位値は「外れた」値の影響が小さい
  - しかし, 算術平均でも「外れた」値を外せば使えるのでは?
  - 注意! 「異常値」ではない
  - 例: 日本の都道府県データでの北海道や東京都

- 切り落とし平均 trimmed mean
  - たとえば, 両側1%（1%分位点より小さいデータと99%分位点より大きいデータ）を除去した残りについての算術平均
  - 3点平均 trimean: （第1四分位 + 中位値の2倍 + 第3四分位）を4で割った値

指数:「ふつう」がどう変化しているか

- 全体的な状況の変化を大雑把に知りたい
  - 各時点における「ふつう」がどう変化しているか
  - 指数:「平均値」が時間によってどう動いているか
  - 例:物価指数は各時点の平均的な物価を示す
  - 例:株価指数は各時点の平均的な株価を示す
- 「各時点のふつう」をどう定義するか?
  - 物価指数は, 単に値段の算術平均でよいのか?
  - あまり買わないものの値段が変化しても「実感に合わない」
  - 各時点で, なんらかの加重平均を使おう　→購入量で
  - 値段が変わらなくて購入量が変化したら指数も変化
    →重みは変化させない
  - どの時点での重みを使うの?　→ラスパイレス, パーシェ, . . .

「散らばり」の大きさ

- 使われる機会は比較的少ないものの，簡単なもの
  - 計算がめんどう，数学的な扱いがめんどう
- 平均偏差
  - 偏差（平均との差）の絶対値の算術平均
- レンジ range（範囲）
  - 最大値と最小値の差
  - 外れ値の影響を受けやすい
- 四分位範囲
  - 第3四分位と第1四分位の差
  - 外れ値の影響が小さい
  - 範囲内の散らばり方についてはなにも言えない

よく使う「散らばり」の指標：分散 variance

- 「散らばっている」とは?
  - 「平均値」の周りに集まっているかどうか
  - 偏差の平均値を取ればよい?　←偏差の合計は常にゼロ
- 分散
  - 偏差を2乗して正の値に直してその平均をとったもの

$$\text{分散}\quad s^2 = \frac{(x_1-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$$

  - 観測値がすべて同じ値を取ればゼロ
  - 分散の公式の分子の部分を「変動」とも呼ぶ
  - 「単位」はもとのデータの単位の2乗
  - 絶対値が出てこないので数学的にも扱いやすい

標準偏差 standard deviation

- 定義：分散の2乗根

$$\text{標準偏差 } s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}$$

- 性質
  - 標準偏差は「単位」がもとのデータと同じ
  - 1次変換（$ax + b$）したデータの標準偏差はそのまま
  - 1次変換（$ax + b$）したデータの標準偏差は2乗される

$$s_{ax+b}^2 = a^2 s^2, s_{ax+b} = |a|\, s$$

  - いずれも，定数 b に依存しない
  - 平均から標準偏差 k 個分の範囲内に入らないデータの相対度数は $(1/k^2)$ より小さい：チェブシェフの不等式

変動係数

- 標準偏差は「単位」を持つ
  - 平均を中心に, $\pm 3s$ の外にある観測値の相対度数は 1/9 以下
  - とはいえ, 他のデータとの比較は難しい
  - 例：日本は他の国と比べて所得や資産の散らばりが大きいのか

- 変動係数：標準偏差を平均で割った値
  - 単位を持たない（無名数）
  - データの単位が異なっても比較できる
  - 例：日本は他の国と比べて所得の分散が大きいのか
  - 例：日本の所得分布は広がってきたのか：インフレの影響を除去

例:金融資産保有額

階級値	相対度数	平均	分散
0	23.57	0.00	291236
50	5.85	2.93	65942
150	6.51	9.77	60199
250	5.82	14.56	43231
350	5.91	20.67	34256
450	3.98	17.93	17434
600	7.99	47.97	20920
850	7.01	59.55	4792
1250	9.23	115.38	1769
1750	6.04	105.77	24636
2500	7.47	186.81	144054
5000	10.60	530.22	1603387
		1111.55	2311859

- 「平均」
  - 階級値と相対度数/100の積
  - すべて足すと算術平均
- 「分散」
  - 階級値と平均の差の2乗に, 相対度数/100 をかけたもの
  - すべて足すと分散
  - 分散の2乗根が標準偏差
  - 標準偏差 = 1520.48
  - 変動係数 = 1.37

データの標準化

- ここでは，それぞれのデータに注目．
- 標準偏差を使うと，「平均からどれくらい離れているか」をそれぞれのデータについて計算できる
  - 各観測値から平均を引いて，標準偏差で割る

$$\text{標準化されたデータ} = \frac{\text{もとの値} - \text{平均}}{\text{標準偏差}} = \frac{x_i - \bar{x}}{s}$$

  - 標準化されたデータの平均はゼロ，標準偏差は1
  - 異なるデータの「位置」を比較できる

  - 偏差値：平均50，標準偏差10に標準化した値

$$\text{偏差値} = 50 + 10 \times \frac{x_i - \bar{x}}{s}$$

「不平等」指標

- ローレンツ曲線（Lorenz curve）
  - 所得や資産の小さい順に観測値を並べ替え，下から x %の人たちが全体の y %を保有している，という関係を $(x\text{-}y)$ 平面にプロットしたもの
  - 累積相対度数と，累積保有比率のプロット
  - (0, 0)と(1, 1)を通るが，すべてが同じ量だけ保有しているとき，(0, 0)と(1, 1)を結ぶ45度線になる（完全平等線）
  - 一般に，45度線の右下にふくらんだ線となり，右下にふくらむほど不平等とされる
  - 単位に依存しないので，異なる集合の比較が可能．ただし，曲線が交差するときは順位をつけられない

「不平等」指標

- ジニ係数（Gini coefficient）
  - 定義はややこしいので省略.
  - ローレンツ曲線と完全平等線（45線）で囲まれた弓形の面積の2倍に等しい
  - ローレンツ曲線が交差するケースでも順位付けが可能
- ハーフィンダール指数（Herfindahl Index）
  - 集中度の尺度として知られる
  - 企業の市場占有率の2乗の和
  - 例：複占で，シェアがともに50%のとき，$0.5^2 + 0.5^2 = 0.5$
- その他「不平等」議論で使われる指標
  - タイル尺度
  - 貧困率

例:金融資産保有額

階級値	累積度数	累積資産
0	23.57	0.00
50	29.42	0.26
150	35.93	1.14
250	41.76	2.45
350	47.66	4.31
450	51.65	5.92
600	59.64	10.24
850	66.65	15.60
1250	75.88	25.98
1750	81.92	35.49
2500	89.40	52.30
5000	100.00	100.00