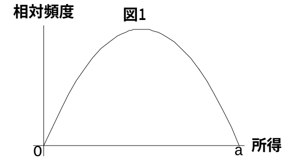
経済統計概論　3

データの整理 (2)

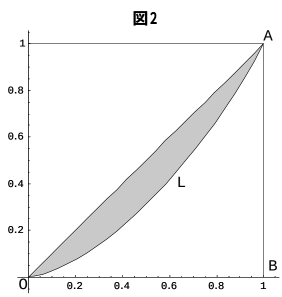
2001年国家1種経済職問題より

- 所得分布のグラフは，図1のように与えられ，横軸に所得を，縦軸にその所得を得ている人数の全人数に対する割合 (相対頻度，相対度数) を取り，グラフに表したものである．ただし，図中のaは最高所得額を示す．



つづき

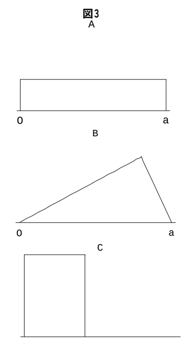
- 所得分布の小さい方から考えて，横軸に累積人数の全人数に対する割合を，縦軸に累積所得の額の全所得額に対する割合を取って描いたグラフをローレンツ曲線という．図2の曲線Lはローレンツ曲線の一例である．
- さらに，図2において斜線部の面積のΔOABに対する割合をジニ係数と定義する．このとき，ジニ係数は所得に関する不平等度の一つの尺度となる．



つづき

- 図3に与えられる所得分布A，B，Cに対し，それらのジニ係数を G_A, G_B, G_C と表すとすると，このときのジニ係数の大小関係として正しい物はどれか．

$1.\ G_A > G_C > G_B,\ 2.\ G_A = G_C > G_B,$
$3.\ G_B > G_A > G_C,\ 4.\ G_B > G_A = G_C,$
$5.\ G_B > G_C > G_A$



1．ローレンツ曲線 (1)

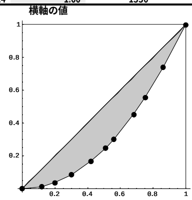
- (正の) 経済変量の集中度を表すのに使用される
- データ値による描き方
  - データ値を小さい順に並べる．
  - 各データについて，(順位/標本数) と (それ以前のデータに関する経済変量の合計/全データに関する合計) の組を求めておく．
  - この組み合わせを (相対順位，相対変量) と呼ぶ
  - その組み合わせに対応した点をプロットし，折れ線グラフを書く．

1．ローレンツ曲線 (2)

- 度数分布表による描き方
  - 各階級について，(累積相対度数) と (その階級までの<階級値×度数>の合計/全データに関する<階級値×度数>の合計) の組を求めておく．
  - その組み合わせに対応した点をプロットし，折れ線グラフを書く．

ローレンツ曲線の実例

階級	階級値	度数	相対度数	累積相対度数	変量	累積変量	相対変量	累積相対変量
			度数/標本数	その階級までの相対度数の合計	階級値×度数	その階級までの変量の合計	変量/総変量	累積変量／変量合計
0～10	5	12	0.12	0.12	60	60	0.012	0.012
10～20	15	8	0.08	0.20	120	180	0.024	0.035
20～30	25	10	0.10	0.30	250	430	0.049	0.084
30～40	35	12	0.12	0.42	420	850	0.082	0.167
40～50	45	9	0.09	0.51	405	1255	0.079	0.246
50～60	55	5	0.05	0.56	275	1530	0.054	0.300
60～70	65	12	0.12	0.68	780	2310	0.153	0.453
70～80	75	7	0.07	0.75	525	2835	0.103	0.556
80～90	85	11	0.11	0.86	935	3770	0.183	0.739
90～100	95	14	0.14	1.00	1330	5100	0.261	1.000
				横軸の値				縦軸の値



2. ジニ係数（1）

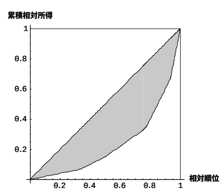
- 相対順位
  - あるデータ値の順位（小さい順）のデータ数に対する割合
  - 度数分布表の場合ある階級の累積相対度数
- 相対変量
  - それ以下のデータ値の合計を全てのデータの合計で割った値
  - 度数分布表の場合ある階級の（階級値×相対度数）/全データの合計
  - 全データ値の合計は（階級値×相対度数）の合計

ジニ係数（2）

- 定義

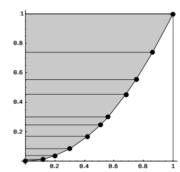
$$G=\{(\text{各階級とその一つ前の相対順位の和}\times\text{相対変量})\ \text{の全階級にわたる合計}\}-1$$

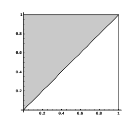
- 図形的意味
  - 右図のように各階級毎にその相対順位と相対累積所得に組み合わせに点を打っていき，それつなぎ合わせるとローレンツ曲線ができる．
  - ジニ計数値は45度線とローレンツ曲線との間の面積の2倍



2. ジニ係数（3）

- 第1の計算式の根拠
  - 青の部分の台形の集まりの面積から，上三角形の面積を引いて，2倍する．
  - 上図の台形の高さは各階級の相対変量である．
  - 計算式（Kは階級数）





$$\begin{aligned}&(0+1\text{番目の累積相対度数})\times 1\text{番目の相対変量}+\\&(1\text{番目の累積相対度数}+2\text{番目の累積相対度数})\times 2\text{番目の相対変量}+\\&\cdots+(K-2\text{番目の累積相対度数}+K-1\text{番目の累積相対度数})\times K-1\text{番目の相対変量}+\\&\quad(K-1\text{番目の累積相対度数}+1)\times K\text{番目の相対変量}-1\end{aligned}$$

2. ジニ係数（4）

- 教科書の式の導出
  - もし，度数がすべて等しければ．．．．
    - そうなるのは，例えばデータそのものを元にした場合や，4分位点，5分位点，10分位点などの等分位点を元に計算する場合

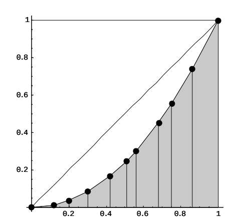
$$i-1\text{番目の累積相対度数}=i\text{番目の累積相対度数}-1/K$$

–

$$\begin{aligned}&(2\times 1\text{番目の累積相対度数}-1/K)\times 1\text{番目の相対変量}+\\&(2\times 2\text{番目の累積相対度数}-1/K)\times 2\text{番目の相対変量}+\\&\cdots+(2\times(K-1\text{番目の累積相対度数})-1/K)\times K-1\text{番目の相対変量}+\\&\quad(2\times 1-1/K)\times K\text{番目の相対変量}-1\\&\quad=2(\text{累積相対度数}\times\text{相対変量の合計})-(\text{相対変量の合計})/K-1\\&\quad=2(\text{累積相対度数}\times\text{相対変量の合計})-1/K-1\end{aligned}$$

2. ジニ係数（5）

- 縦に面積を計算すると．．．
  - 累積相対度数＝相対順位の差は相対度数であることに注意
  - 三角形から青色の部分の面積を引いて2倍

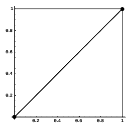


$$1-\left(\begin{matrix}\text{ある階級とその一つ前の累積相対変量の和}\\\times\text{相対度数}\end{matrix}\right)\text{の合計}$$

2．ジニ係数(6)

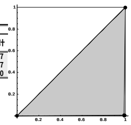
- 極端なローレンツ曲線とそのときのジニ係数
  - すべてが同一値，または，同一階級の場合(平等ケース)

階級	階級値	度数	相対度数	累積相対度数	変量	累積変量	相対変量	累積相対変量
			度数/標本数	その階級までの相対度数の合計	階級値×度数	その階級までの変量の合計	変量/総変量	累積変量／変量合計
0～10	5	0	0.00	0.00	0	0	0.000	0.000
10～20	15	0	0.00	0.00	0	0	0.000	0.000
20～30	25	0	0.00	0.00	0	0	0.000	0.000
30～40	35	0	0.00	0.00	0	0	0.000	0.000
40～50	45	0	0.00	0.00	0	0	0.000	0.000
50～60	55	0	0.00	0.00	0	0	0.000	0.000
60～70	65	100	1.00	1.00	6500	6500	1.000	1.000
70～80	75	0	0.00	1.00	0	6500	0.000	1.000
80～90	85	0	0.00	1.00	0	6500	0.000	1.000
90～100	95	0	0.00	1.00	0	6500	0.000	1.000
				横軸の値				縦軸の値



    - ジニ係数は0
  - 1データだけが極端に大きい場合

階級	階級値	度数	相対度数	累積相対度数	変量	累積変量	相対変量	累積相対変量
			度数/標本数	その階級までの相対度数の合計	階級値×度数	その階級までの変量の合計	変量/総変量	累積変量／変量合計
0～10	5	99	0.99	0.99	495	495	0.047	0.047
10～20	15	0	0.00	0.99	0	495	0.000	0.047
10000～	10000	1	0.01	1.00	10000	10495	0.953	1.000
				横軸の値				縦軸の値



    - ジニ係数は1

2．ジニ係数(7)

- ジニ係数の特徴
  - 端のデータが真ん中に移ると，ジニ係数は小さくなる(だから不平等の指標)
  - データの単位を変えてもジニ係数は変わらない
  - 均等な度数分布の場合　　最小階級の下限が両方0なら

階級	階級値	度数	相対度数	累積相対度数	変量	累積変量	相対変量	累積相対変量
			度数/標本数	その階級までの相対度数の合計	階級値×度数	その階級までの変量の合計	変量/総変量	累積変量／変量合計
0～10	5	10	0.10	0.10	50	50	0.010	0.010
10～20	15	10	0.10	0.20	150	200	0.030	0.040
20～30	25	10	0.10	0.30	250	450	0.050	0.090
30～40	35	10	0.10	0.40	350	800	0.070	0.160
40～50	45	10	0.10	0.50	450	1250	0.090	0.250
50～60	55	10	0.10	0.60	550	1800	0.110	0.360
60～70	65	10	0.10	0.70	650	2450	0.130	0.490
70～80	75	10	0.10	0.80	750	3200	0.150	0.640
80～90	85	10	0.10	0.90	850	4050	0.170	0.810
90～100	95	10	0.10	1.00	950	5000	0.190	1.000
				横軸の値				縦軸の値

階級	階級値	度数	相対度数	累積相対度数	変量	累積変量	相対変量	累積相対変量
			度数/標本数	その階級までの相対度数の合計	階級値×度数	その階級までの変量の合計	変量/総変量	累積変量／変量合計
0～5	2.5	10	0.10	0.10	25	25	0.010	0.010
5～10	7.5	10	0.10	0.20	75	100	0.030	0.040
10～15	12.5	10	0.10	0.30	125	225	0.050	0.090
15～20	17.5	10	0.10	0.40	175	400	0.070	0.160
20～25	22.5	10	0.10	0.50	225	625	0.090	0.250
25～30	27.5	10	0.10	0.60	275	900	0.110	0.360
30～35	32.5	10	0.10	0.70	325	1225	0.130	0.490
35～40	37.5	10	0.10	0.80	375	1600	0.150	0.640
40～45	42.5	10	0.10	0.90	425	2025	0.170	0.810
45～50	47.5	10	0.10	1.00	475	2500	0.190	1.000
				横軸の値				縦軸の値

最大階級がどこにあるかに関わらずジニ係数は同じ

min:最小の階級値

max:最大の階級値

ジニ係数＝

$$\frac{(1+k)(\max-\min)}{3k(\max+\min)}$$

問題の答え

- AとCの比較

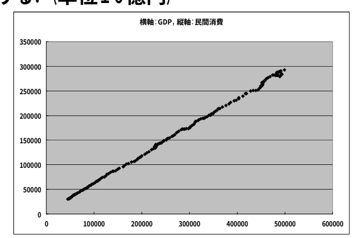
  Aはmin=0,max=a

  Bはmin=0,max=a c　(c＞0)

  kは十分大きいとするとどちらも1／3にほぼ等しい
- AとBの比較
  - BはAの端のデータを真ん中に寄せたと見なすことができる．
  - $G_A > G_B$

3．2変数データの整理
3．1　散布図

- 2変数の関係を視覚的に捉える＝散布図
  - データ値の組み合わせを座標と考えて，データ毎に点をプロットする．(単位10億円)



3．2　標本相関係数(1)

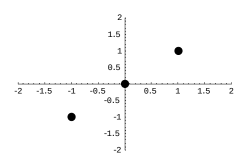
- 2変数の関係の深さを数値化する．
- 共分散

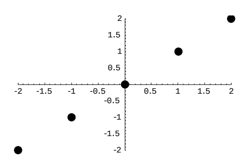
$$標本共分散=\frac{\left\{\begin{array}{l}(x変数の値-xの標本平均)\times(y変数の値-yの標本平均)\\ のすべての標本に関する合計\end{array}\right\}}{n(またはn-1)}$$

  - 分散を計算するのにnで割ったかn-1で割ったかによってきめる
  - 共分散の性質
    - xの値がその平均を上回るときにはyもその平均を上回る様な場合，共分散の式の分子は正になる(右上がりの傾向)
    - xの値がその平均を上回るときにはyはその平均を下回る様な場合，共分散の式の分子は負になる(右下がりの傾向)
    - xの値とyの値に関係がない場合は，0になる

3．2　標本相関係数(2)

- 共分散の困った点
  - 右の上下の図はともにxとyに線形の関係がある場合の散布図
  - しかし，上の場合の共分散は2／3で，下は2
  - つまり，共分散の絶対値が相関の強さを表していない．
  - xやyの散らばり方が大きい方が共分散は大きくなってしまう．





3．2　標本相関係数（3）

- 共分散をxとyの標準偏差で標準化
  - なぜ分散で標準化しないのか
    - 単位をそろえる．分子は，2乗の単位
    - 分母も2乗の単位でないと困る．すると，x，yともに1乗の単位にならなければならない．
    - 散らばりを表す1乗の単位の代表例が標準偏差
- 標本相関係数

$$r_{xy} = \frac{\text{標本共分散}}{\text{x の標本標準偏差} \times \text{y の標本標準偏差}}$$

3．2　標本相関係数（4）

- 標本相関係数の性質
  - 符号は相関の方向を表す
    - xが平均より大きかったらyも平均より大きい傾向
      - 符号は正，正の相関を持つと呼ぶ
    - xが平均より大きかったらyは平均より小さい傾向
      - 符号は負，正の相関を持つと呼ぶ
  - 絶対値は相関の強さ，関係の深さを表す
    - 大きいほど関係は深い
    - 0～1の間
  - 標本相関係数はー1以上＋1以下

3．2　標本相関係数（5）

- ＧＤＰと消費の例では標本相関係数は0.999
- 相関係数の限界
  - 相関係数は線形関係が完全に成立している場合を基準にしている＜相関係数を計算する前に散布図を書こう＞
    - 非線形関係，例えば，2次式の関係などは表せない．
    - 教科書の例では，相関係数が0なのに非線形の関係がある
  - 見せかけの相関
    - 二つの変数とも，第3の変数と相関を持っていて，その結果，両変数に関係があるように見える場合がある．
    - 例えば，時系列データの場合に，二つの変数が時間とともに増加，または，減少している場合

3．3　標本偏相関係数

- 第3の変数と両方とも相関を持っていて，そのために2つの変数の相関が強いように見える場合の対策
  - 第3の変数（zとする）の影響を取り除いたうえで，二つの変数の相関を調べる．
  - 標本偏相関係数

$$r_{xy|z} = \frac{r_{xy} - r_{xz} r_{zy}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{zy}^2\right)}}$$

  - 「子のつく名前は頭がいい」
    ←女子高校の子の付く名前の生徒の比率とその高校の偏差値に相関
    しかし，第3の変数と関係があるだけではないか？
      第3の変数候補，その家庭の所得，親の資産，皇室への関心度
      このような変数との相関が容易には調べられないのをいいことに，さも直接的な関係があるように見せているのでは？
- 分割表は教科書参照のこと