図 2-7 年間収入の累積相対度数折れ線

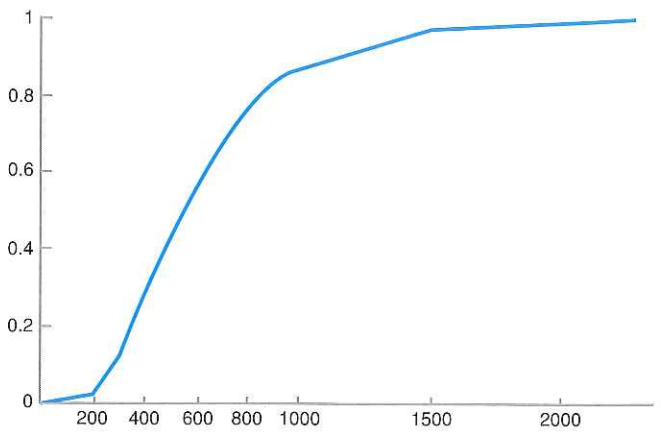
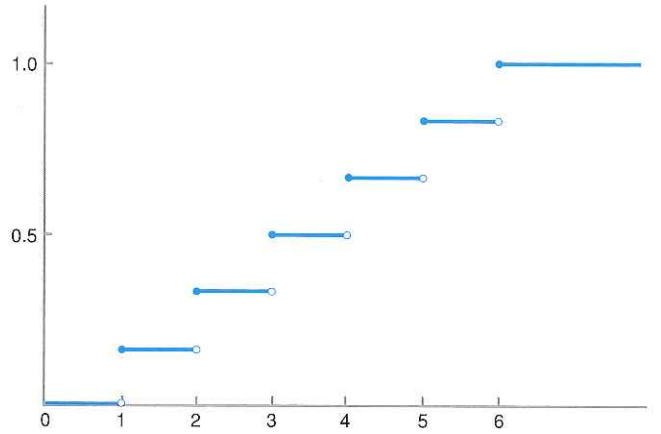


図 2-8 不連続な累積相対度数折れ線



目	1	2	3	4	5	6	合 計
度 数	15	16	18	18	16	17	100

このときの累積相対度数折れ線は図 2-8 となる。階段関数の場合には，この例のように，ジャンプする点ではジャンプ後の値をとるようにする。

□ 2-4 ローレンツ曲線とジニ係数 □

度数分布は分布の集中度あるいは分散度の観点から見る場合が多い。例えば，本章の最初に取り上げたＡ大学の新入生の出身地域別分布（表 2-1）も，特定の地域に集中しているのか，あるいは全国的に散らばっているのか，という点に興味がある。また，表 2-5 や表 2-6 に示した年間収入の分布については，所得分配の散らばり，あるいは集中の程度に興味がある。

ローレンツ曲線は，このように，集中，散らばり，不平等などの度合を観察するために度数分布表から作られる曲線である。例えば，表 2-5 の年間収入についていえば，横軸に低収入の世帯からの世帯数の累積相対度数，縦軸に収入の累積相対度数を目盛り，対応する点を順次結んだものである。したがって，ローレンツ曲線は累積相対度数折れ線と同様に，0 から 1 まで単調に増加する曲線となる。

例題 2.4 表 2-5 の年間収入の度数分布表からローレンツ曲線を描け。

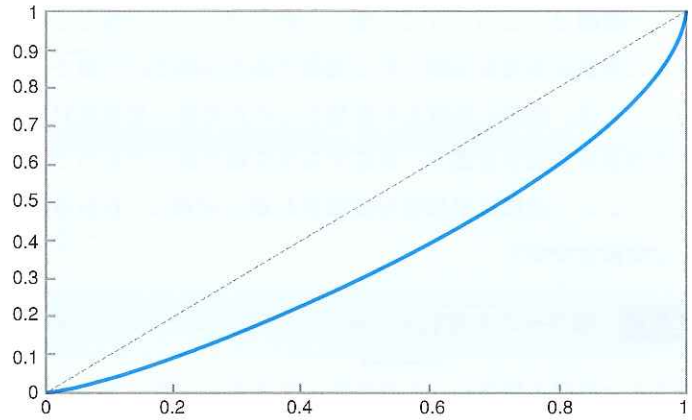
（解） 表 2-5 に低収入階級からの世帯数の累積相対度数があるので，あとは階級ごとの収入の全収入に対する割合（階級収入比）を計算し，その累積比（階級収入累積比）を求めればよい。表 2-7 に，階級収入比，階級収入累積比の値を，表 2-5 の相対度数，累積相対度数とともに示した。ローレンツ曲線は世帯数の累積相対度数，階級収入累積比のペアをプロットして順次結んだ折れ線であり，図 2-9 のようになる。

この場合，ローレンツ曲線は 45 度線よりも常に下側にある。それは，表 2-7 において階級収入累積比が世帯数の累積相対度数より常に小さいことと同等である。このことは，低所得階層においては世帯数の割合の方が収入の割合よりも大きく，高所得階層では逆になることを意味する。

▶表 2-7　世帯数の相対度数，累積相対度数および階級収入比（%）

階　級	1	2	3	4	5	6	7	8	9	10	11
相対度数	2.4	10.4	17.3	15.5	12.9	10.2	7.9	6.1	4.6	9.7	3.0
累積相対度数	2.4	12.9	30.1	45.6	58.5	68.7	76.6	82.7	87.3	97.0	100.0
階級収入比	0.6	4.3	9.7	11.1	11.2	10.6	9.5	8.3	6.9	18.4	9.5
階級収入累積比	0.6	4.9	14.6	25.7	36.9	47.5	57.0	65.3	72.2	90.5	100.0

図 2-9　年間収入のローレンツ曲線



もし，すべての世帯が同一の収入であればローレンツ曲線は 45 度線と一致する。この 45 度線のことを完全平等線という。逆に，完全に不平等な状況は，所得が単一の世帯に独占されている場合であり，このときローレンツ曲線は正方形の左下から右上までの辺上を通る直角の折れ線となる。この折れ線は完全不平等線と呼ばれる。

集中度の観点を強調する場合には，ローレンツ曲線は 45 度線の上側にあるように描くこともできる。上の年間収入の例でいえば，高収入階級から累積することになる。別の例，例えば，同一産業内の企業規模（＝従業員数）の分布を調べる場合には，規模の大きい方の企業から累積することになる。

この場合の横軸は，総企業数に対して，各企業を 1 つずつ累積したときの相対度数であり等間隔で増えていく。表 2-1 あるいは表 2-2 からローレンツ曲線を描くときの横軸も同様である（例題 2.4 を参照）。

ローレンツ曲線が完全平等線から離れるに従って，集中，独占，不平等などの度合が大きくなる。このことに注目して，その乖離の程度を測る特性値として考案されたのがジニ係数である。それは，

ジニ係数＝2×（完全平等線とローレンツ曲線で囲まれた面積）

により定義され，0（＝完全平等線に一致する場合）から 1（＝完全不平等線に一致する場合）までの値をとりうる無名数である。

ジニ係数を計算する簡単な方法を説明しよう。世帯収入の場合を考えることにして，次の変数を定義する。

p_i：第 i 階級に属する世帯の相対度数

q_i：第 i 階級までの世帯の累積相対度数

r_i：第 i 階級の世帯の収入比

これらの値は，百分比ではなく小数で表すことにする。以下，説明の便宜上，階級数を 3 として，ローレンツ曲線が 45 度線のどちら側にあるかで場合分けして考えよう。

［1］ローレンツ曲線が 45 度線の下側にある場合

図 2-10 で示したローレンツ曲線の場合を考えよう。まず，次の 2 つの面積を求める。

A＝色付けされた長方形の面積の総和

$$=p_1r_1+(p_1+p_2)r_2+(p_1+p_2+p_3)r_3=q_1r_1+q_2r_2+q_3r_3$$

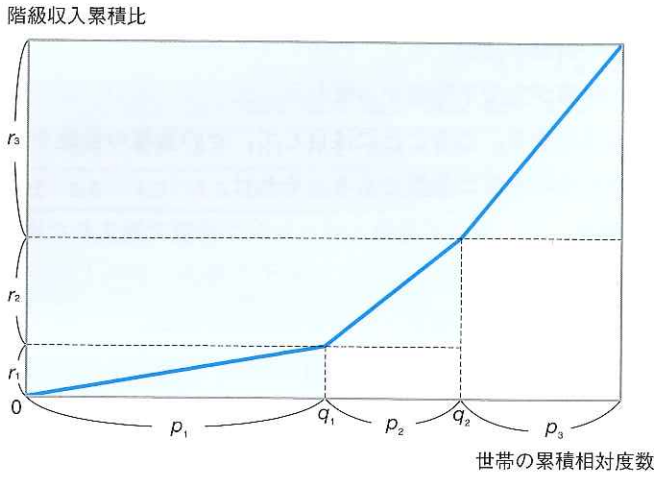
B＝A の中でローレンツ曲線の右側にある直角三角形の面積の総和

$$=\frac{1}{2}(p_1r_1+p_2r_2+p_3r_3)$$

したがって，この場合のジニ係数は，面積 A と B を使って，

$$\text{ジニ係数}=2\times\left(A-B-\frac{1}{2}\right)=2\sum_{i=1}^{3}q_ir_i-\sum_{i=1}^{3}p_ir_i-1$$

図 2–10 ジニ係数の計算方法（その 1）



で与えられる。

［2］ローレンツ曲線が 45 度線の上側にある場合

次に，図 2–11 のように，ローレンツ曲線が 45 度線より上にある場合を考えよう。このとき，

C＝色付けされた長方形の面積の総和

D＝斜線で示された直角三角形の面積の総和

とすれば，ジニ係数は，

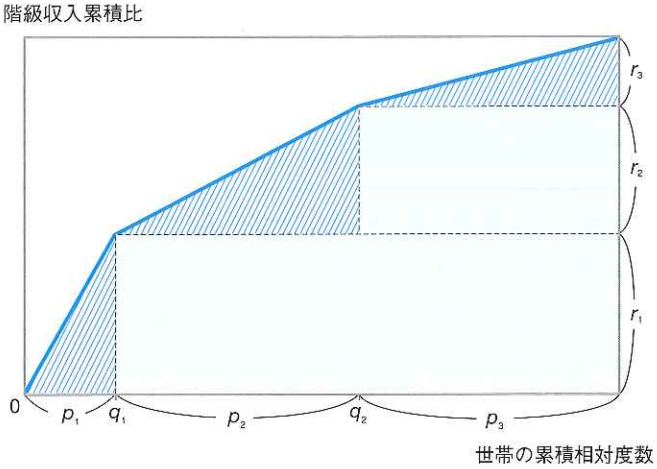
$$2\times\left(C+D-\frac{1}{2}\right)=2\times\left(\sum_{i=1}^{3}(1-q_i)r_i+\frac{1}{2}\sum_{i=1}^{3}p_ir_i-\frac{1}{2}\right)$$

と表すことができる。そして，これを整理すると，

$$\text{ジニ係数}=2\times\sum_{i=1}^{3}r_i-2\sum_{i=1}^{3}q_ir_i+\sum_{i=1}^{3}p_ir_i-1=-2\sum_{i=1}^{3}q_ir_i+\sum_{i=1}^{3}p_ir_i+1$$

を得る。この表現は，ローレンツ曲線が 45 度線よりも下にある場合の表現の符号を変えたものになることがわかる。

図 2–11 ジニ係数の計算方法（その 2）



以上が，度数分布表にまとめられた場合のジニ係数の計算方法である。これらの結果を階級数が m 個の場合に一般化すると，次のようにまとめることができる。なお，カッコ内の計算式は個票の場合（表 2–1 のような質的データの場合を含む）であり，この結果は，$p_i=1/m$，$q_i=i/m$ とすれば得られることに注意されたい。

●ジニ係数の計算式

以下，階級数を m，第 i 階級に属する世帯の相対度数を p_i，累積相対度数を q_i，階級収入比を r_i とする。

［1］ローレンツ曲線が 45 度線の下側にある場合

$$\text{ジニ係数}=2\sum_{i=1}^{m}q_ir_i-\sum_{i=1}^{m}p_ir_i-1$$

$$\left(=\frac{2}{m}\sum_{i=1}^{m}i\times r_i-\frac{m+1}{m}\right)$$

［2］ローレンツ曲線が 45 度線の上側にある場合

$$\text{ジニ係数} = \left| 2\sum_{i=1}^{m} q_i r_i - \sum_{i=1}^{m} p_i r_i - 1 \right|$$

$$\left(= \left| \frac{2}{m}\sum_{i=1}^{m} i \times r_i - \frac{m+1}{m} \right| \right)$$

一般に，単独のジニ係数の値から不平等度や集中度の程度をあれこれ議論することは困難である。比較可能な他のデータから得られるジニ係数と比べることにより，はじめて意味をもつということができる。例えば，異なる年度あるいは国々の年間収入のジニ係数を比較したり，同じ年度の収入と貯蓄のデータから得られるそれぞれのジニ係数を比較することなどが考えられる。

例題 2.5 表 2-1 および表 2-2 の新入生の出身地域別度数分布からジニ係数を求め，2 つの大学を比較せよ。

(解) まず，ローレンツ曲線は図 2-12 のようになる。この図から，A 大学のジニ係数の方が小さいことは明らかである。ジニ係数を計算するためには，

図 2-12 A 大学と B 大学の新入生の出身地に関するローレンツ曲線

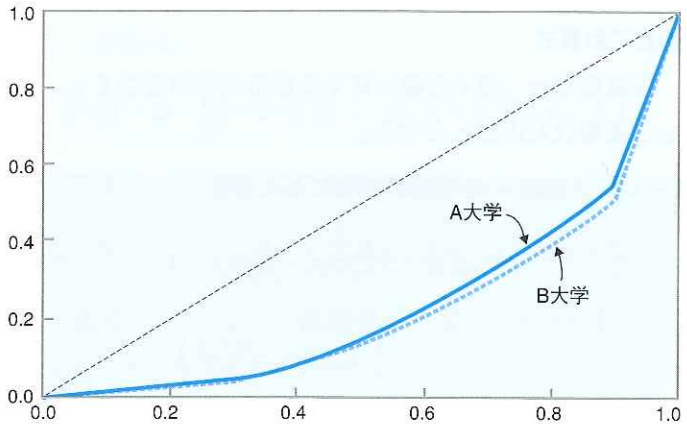


表 2-1 と表 2-2 を相対度数の小さい方から並び替えて，次のような度数分布表を作る。

階　級	1	2	3	4	5	6	7	8	9	10
A大学	0.01	0.0125	0.0125	0.045	0.0625	0.085	0.095	0.1075	0.12	0.45
B大学	0.01	0.012	0.0125	0.045	0.045	0.068	0.0875	0.105	0.135	0.48

このとき，ジニ係数の計算式［1］で個票の場合を使って，A 大学のジニ係数は，

$$\frac{2}{10}(1\times 0.01 + 2\times 0.0125 + \cdots + 10\times 0.45) - \frac{11}{10} = 0.536$$

となる。同様にして，B 大学のジニ係数は 0.570 となる。したがって，統計的には，A 大学の方が規模は小さいがより全国型であるといえる。

●練習問題

1．次のデータは 50 人の試験の点数である。度数分布表，ヒストグラム，累積相対度数折れ線を作成せよ。また，度数分布表から平均とメディアンを求めよ。

51	60	52	55	53	55	56	50	24	53	60	47	47	58	41	53	43
43	46	42	54	72	40	47	33	65	53	76	48	41	49	61	56	42
50	44	24	52	55	45	48	46	53	67	43	56	60	48	47	37	

2．次の 50 個のデータは，ある都市における 50 日間にわたる毎日の事故件数である。このデータに対して，問 1 と同様の問題に答えよ。また，問 1 のデータとの違いについて述べよ。

8	7	6	9	3	7	6	3	8	7	7	2	7	2	4	6	2
5	5	3	6	1	2	6	5	4	2	3	2	6	9	10	4	6
7	5	8	5	6	4	7	8	3	9	6	5	7	4	4	5	

3．次のデータは，ある会社の電話の通話時間（単位：分）に関する 50 個のデータである。このデータに対して，問 1 と同様の問題に答えよ。また，問 1，問 2 のデータとの違いについて述べよ。

0.7 1.2 2.0 0.2 9.4 1.3 1.6 9.4 0.4 0.7 1.2 10.4 1.3 10.8 5.3 1.5 14.2
3.8 2.9 7.9 1.9 22.0 10.8 1.6 3.7 6.4 14.5 8.0 10.9 1.9 0.2 0.1 5.2 1.9
1.0 3.9 0.4 3.0 1.8 4.3 0.7 0.7 10.4 0.2 2.4 2.8 1.2 5.7 5.6 3.6

4．n 人の所得が $x_1, \cdots, x_n$ であるとき，

$$\text{ジニ係数}=\frac{1}{2n^2\bar{x}}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i-x_j|$$

と表されることを示せ。

5．5 人の所持金が次の場合にジニ係数を求めよ。

2,000 3,000 6,000 9,000 20,000

6．表 2-5 および表 2-6 の世帯の年間収入のジニ係数を求めよ。

7．次の表は，例題 2.2 で扱った「全国消費実態調査」における男女別単身世帯の年間収入を合算した度数分布表（階級と階級値の単位は万円）である。この結果から，ジニ係数および相対的貧困率を求めよ。

階　級	～100	～150	～200	～250	～300	～350	～400	～500	～600	600～
階級値	65.8	127.4	173.2	223.3	272.2	319.9	370.7	440.6	542.0	789.6
度　数	397	556	720	677	548	482	373	527	271	390

8．次の表は 2004 年と 2009 年の調査による世帯の年間収入の五分位階級別度数分布表である。階級値（単位：万円）には平均が使われている。それぞれの年の累積相対度数折れ線とローレンツ曲線を描き，ジニ係数を求めよ。

階　級	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	合 計
階級値（2004 年）	279	422	566	760	1,210	3,237
階級値（2009 年）	270	402	537	725	1,211	3,145

（資料）「家計調査年報」。

9．次のデータは，都道府県別の人口密度（人/km²）のデータである。度数分布表とヒストグラムを作成せよ。また，人口密度の高い方から集計したローレンツ曲線を描き，ジニ係数を求めよ。

72 93 106 116 118 135 151 153 154 159 176 191 195
195 196 197 197 201 228 249 257 264 267 271 278 302
304 309 309 310 336 360 373 382 467 472 540 546 564
645 969 1,078 1,300 1,687 3,310 4,640 5,430

第 3 章

データの特性値

度数分布表やヒストグラムはデータ全体の分布を見やすく整理したものであるが，本章ではデータをさらに縮約して，分布の様子を単一の数値（特性値）で表すことを考える。特性値には，分布の中心を表すものをはじめとして，ばらつき，ゆがみ，とがりなどを測る代表値があるが，以下では中心とばらつきを測る特性値に焦点を当てる。