□ 2-1 度数分布表 □

データは，そのままでは数字や文字の羅列であり，それらの並びから即座に意味のある情報を引き出すことは困難である。データを前にして最初に考えるべきことは，データをいかに縮約するか，ということである。本節で扱う度数分布表は，そのための一つの方法である。以下では，度数分布表の作り方を，質的データと量的データのそれぞれの場合について説明する。

■質的データの場合

例として，ある年のA大学全体の新入生の出身地域別の度数分布表を作ることを考えよう。対象とするデータは，名義尺度の質的データである。最初の問題は，出身地域をどのレベルで考えるかということである。市町村レベルで考えるとカテゴリーが大きくなりすぎるであろう。そこで，外国人留学生（帰国子女を含む）の出身地域は一括して「外国」とし，日本人については全国を後述の9つの地域に分け，卒業した高校（大検の者は中学校）の所在地をこれらの地域とする。したがって，カテゴリーは多くとも10である。

この場合の度数分布表の作成は，きわめて単純である。各データを該当する地域に割り振って，各地域に属する度数を数え上げればよい。その結果として表2-1が得られたものとしよう。

この度数分布表の第1列は階級であり，地域を表す10個のカテゴリーからなっている。名義尺度のデータではカテゴリーの順番は特に決まっていないが，ここでは北の地域から並べ，最後に外国とした。第2列の度数は各階級に属する人数であり，第3列の相対度数は総度数に対する各度数の割合を百分率表示したものである。

表2-1から，関東出身者が最も多く，45%と全体の半分近くを占めてい

▶表2-1 A大学新入生の出身地域の度数分布表

階 級	度 数	相対度数(%)
北海道	50	1.25
東北	40	1.00
関東	1,800	45.00
北陸	250	6.25
中部	480	12.00
近畿	380	9.50
中国	340	8.50
四国	180	4.50
九州・沖縄	430	10.75
外国	50	1.25
計	4,000	100.00

▶表2-2 B大学新入生の出身地域の度数分布表

階 級	度 数	相対度数（%）
北海道	100	1.00
東北	120	1.20
関東	1,350	13.50
北陸	680	6.80
中部	1,050	10.50
近畿	4,800	48.00
中国	450	4.50
四国	450	4.50
九州・沖縄	875	8.75
外国	125	1.25
計	10,000	100.00

るものの，出身地域は全国にまたがっており，A大学は全国型の大学といえそうである。ここで，最大度数を与えるカテゴリーはモード（最頻値）と呼ばれる。「関東」はA大学新入生の出身地域のモードである。

たリーマン・ショック（米国の名門投資銀行の破綻によりもたらされた世界的な金融危機）によるものである。階級数については，データ数が84であることと範囲が約30であることを勘案して，ここでは8つの階級を考え，階級の幅を5とする。そして，階級の限界点を−25，−20，−15，−10，−5，0，5，10，15とする。このようにして作られた度数分布表が表2-4である。この表において，第2列の階級値は階級の中点である。また，最後の列の累積相対度数は，相対度数を上から順々に加えたものである。

この度数分布表からは次の事実を読みとることができる。極端に小さな負の値をとるデータがあるものの，全体としては0を中心として分布している。ただし，正の値をとるデータの方が多く，その割合は55%程度である。次に，モードは，階級の幅が等間隔ならば，最大度数を与える階級の階級値（あるいは階級の中点）で定義されるので，この場合のモードは第6階級の階級値の2.5となる。ただし，等間隔でない場合には，度数を基準幅当たりの度数，あるいは相対度数に換算した上で考えなければならない（例題2.1を参照）ので，注意が必要である。

量的データの度数分布表においては，モード以外にもさまざまな特性値を計算することができる。累積相対度数が50%となる値をメディアン（中央

▶表2-4　TOPIXの月次収益率の度数分布表

階　級	階級値	度　数	相対度数(%)	累積相対度数(%)
−25以上 — −20未満	−22.5	1	1.2	1.2
−20 — −15	−17.5	0	0.0	1.2
−15 — −10	−12.5	1	1.2	2.4
−10 — −5	−7.5	11	13.1	15.5
−5 — 0	−2.5	25	29.8	45.2
0 — 5	2.5	31	36.9	82.2
5 — 10	7.5	14	16.7	98.8
10 — 15	12.5	1	1.2	100.0
合　計		84	100.0	

値）という。すなわち，データを小さい方から並べたとき，下に半分，上に半分となる真ん中の値である。表2-4から，メディアンは，モードと同様に第6階級に属すことがわかる。

メディアンを求めるには，次のように考える。メディアンが第i階級に属すものとし，階級の上限をa_iとする。また，第i階級までの累積相対度数（%）をq_iとする。このとき，図2-2にあるように，この階級内ではデータが均等に分布しているものと想定し，比例配分の考え方を使って，

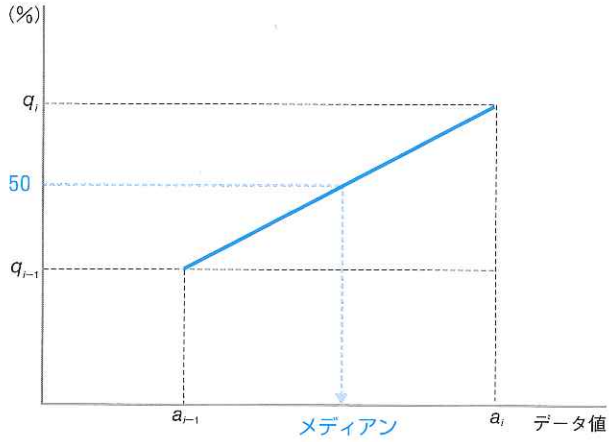
$$\text{メディアン} = a_{i-1} + (a_i - a_{i-1}) \times \frac{50 - q_{i-1}}{q_i - q_{i-1}} \qquad (2)$$

$$= 0 + (5-0) \times \frac{50 - 45.2}{82.2 - 45.3} = 0.65$$

となる。

量的データの度数分布表からは，平均も計算することができる。ただ，階級にまとめられているので，同じ階級に属すデータはすべて階級値をとるものと考える。今，全部でm個の階級があり，第i階級の階級値をx_i，度数をf_i，総度数をnとすれば，

図2-2　度数分布表からメディアンを求める方法



$$\text{平均} = \bar{x} = \frac{1}{n}\sum_{i=1}^{m} f_i x_i \qquad (3)$$

で計算される。ここで，$\bar{x}$ はエックスバーと読む。表 2-4 の度数分布表の場合は，

$$\bar{x} = \frac{1}{84} \times (-22.5 \times 1 + (-17.5) \times 0 + \cdots + 12.5 \times 1) = 0.18$$

となる。

モード，メディアン，および平均は，いずれも度数分布の中心を表す特性値である。今の例の場合には，それぞれ，2.5，0.65，0.18 となった。これらの値は，分布の形状により特定の大小関係をもつことがある。この点については次節で説明する。また，中心の特性値以外にも，分布のばらつきなどを表す特性値を考えることができるが，これらについては次章のトピックとしたい。

ところで，官庁から公表されているデータは官庁統計と総称されるが，プライバシーの保護や標本サイズが大きいなどの理由から，個々のデータ（個票）ではなく，それらが集計されて度数分布表として提供されているのが普通である。実際の例が表 2-5 に示されている。これは，2009 年の「家計調査年報」（総務省統計局作成）から抜粋した 2 人以上からなる一般世帯の年間収入の度数分布表である。母集団は全国約 3,000 万の一般世帯からなり，ここで選ばれた 7,831 の調査世帯は層別 3 段抽出法により抽出された。第 1 次抽出単位は市町村，第 2 次抽出単位は国勢調査のために設定された調査単位区，第 3 次抽出単位は世帯である。

この度数分布表は 11 の階級からなっているが，最後の階級は上限が明記されないオープン・エンドの階級である。それ以外の階級では，幅の長さが，第 1 階級が 200 万円，第 2 階級から第 9 階級までは 100 万円，第 10 階級は 500 万円となっており，等間隔ではない。極端に大きい，あるいは小さい値（異常値）が混在しているようなデータでは，オープン・エンドの階級を設けたり，階級の幅を不等間隔にせざるをえないことが多い。

▶表 2-5 一般世帯の年間収入（2009 年）

階級（万円）	階級値（万円）	度 数	相対度数（%）	累積相対度数（%）
0 — 200	157	190	2.4	2.4
200 — 300	255	817	10.4	12.9
300 — 400	349	1,352	17.3	30.1
400 — 500	447	1,215	15.5	45.6
500 — 600	545	1,004	12.9	58.5
600 — 700	644	801	10.2	68.7
700 — 800	745	619	7.9	76.6
800 — 900	844	479	6.1	82.7
900 — 1,000	944	356	4.6	87.3
1,000 — 1,500	1,173	763	9.7	97.0
1,500 —	1,968	235	3.0	100.0
合 計		7,831	100.0	

（資料）「家計調査年報」。

例題 2.1 表 2-5 の度数分布表から，年間収入のモード，メディアン，平均を求めよ。

（解） 階級の幅が等間隔でない場合には，モードを求める際に注意が必要である。単に，最大度数を与えている階級の階級値をモードとしてはならない。モードを求めるためには，基準幅当たりの相対度数に換算して考えねばならない。ここでは，基準幅を 100 万円として考えると，結局，モードは第 3 階級にあることがわかり，この階級の階級値 349 万円がモードとなる。また，メディアンは第 5 階級にあり，（2）の計算式を使って，

$$\text{メディアン} = 500 + (600 - 500) \times \frac{50 - 45.6}{58.5 - 45.6} = 534$$

を得る。最後に，平均は（3）を使って，

$$\bar{x}=\frac{1}{7831}(190\times157+\cdots+235\times1968)=623$$

となる。

この例題のように，階級値に階級平均が使われている場合には，度数分布表から計算される平均は個票の場合の平均と同一である。なお，モード＜メディアン＜平均の大小関係が成立していることに注意されたい。この理由については次節で説明する。

次の例題も，度数分布表から意味のある情報を取り出すための応用例である。例題 2.1 では，メディアンの求め方が問題となっていた。それは，一般的には，累積相対度数を与えて，それに対応する収入を求める問題である。次の例題では，逆に，与えられた収入以下，あるいは以上となる割合を求めることが問題となる。その応用例として，個人の収入分布において，メディアン収入の半分を下回る収入の人の割合により定義される相対的貧困率，および，ここで定義する相対的富裕率（メディアン収入の 1.5 倍を上回る収入の人の割合）を求めることを考える。

例題 2.2 表 2-6 は，2004 年の「全国消費実態調査」（総務省統計局作成）から抜粋した男女別単身世帯の年間収入の度数分布表である。階級値の単位は万円，p は相対度数（%），q は累積相対度数（%）である。この分布表から，男女別の相対的貧困率と相対的富裕率を求めよ。

（解） まず，メディアン収入は，次のようになる。

$$\text{男性}：350+50\times2.7/10.6=362.7$$

$$\text{女性}：200+50\times5.4/16.6=216.3$$

したがって，相対的貧困率は，男性では 181.35 万円以下，女性では 108.15 万円以下の収入の人の割合となり，表 2-6 から，次の結果が得られる。

$$\text{男性}：\ 8.5+6.9\times31.35/50=12.8$$

$$\text{女性}：10.9+14.7\times8.15/50=13.3$$

▶表 2-6 男女別単身世帯の年間収入（2004 年）

	男性				女性			
階級（万円）	階級値	度数	p	q	階級値	度数	p	q
0 — 100	66.1	58	3.2	3.2	65.7	339	10.9	10.9
100 — 150	127.7	96	5.3	8.5	127.2	460	14.7	25.6
150 — 200	171.0	126	6.9	15.4	174.0	594	19.0	44.6
200 — 250	221.9	161	8.8	24.2	223.9	516	16.5	61.2
250 — 300	270.3	199	10.9	35.1	273.7	349	11.2	72.4
300 — 350	319.3	222	12.2	47.3	320.5	260	8.3	80.7
350 — 400	372.2	193	10.6	57.9	368.7	180	5.8	86.5
400 — 500	438.5	334	18.3	76.3	445.1	193	6.2	92.7
500 — 600	543.1	176	9.7	85.9	539.1	95	3.0	95.7
600 —	793.5	256	14.1	100.0	778.3	134	4.3	100.0
合　計		1,821	100.0			3,120	100.0	

（資料）「全国消費実態調査」。

他方，相対的富裕率は，男性では 544.05 万円以上，女性では 324.45 万円以上の収入の人の割合となり，次のようになる。

$$\text{男性}：100-(76.3+9.6\times44.05/100)=19.5$$

$$\text{女性}：100-(72.4+8.3\times24.45/50)=23.5$$

女性は，男性に比べて，相対的貧困率も富裕率も高い。特に，相対的富裕率が高くなっている。その理由は，女性の収入分布は，メディアン収入が低いにもかかわらず，右スソを長く引いており，ばらつきの大きな分布になっていることによる。

度数分布表は，記述的な統計分析を行う場合の出発点となる非常に重要なものである。本章の残りの節では，度数分布表に基づくさまざまな分析方法を説明したい。度数分布表は，実際に作成するとなると，結構むずかしいものである。その理由は，階級数や階級の幅，階級の上限，下限などを決めるための統一的な公式がないからである。度数分布表作成の一般的な手順をま

とめると，次のようになるであろう。

●度数分布表の作り方（量的データの場合）

範囲（レンジ）を求める。

⬇

階級数と階級の幅，階級値，下限，上限を定める。階級数は5から15程度とする。階級値は階級の中点とするのが普通であり，区切りのよい数値となるように下限，上限を決める。ただし，オープン・エンドの階級に対してはデータの平均を階級値とする。

⬇

各階級に属するデータの度数を数え上げる。

⬇

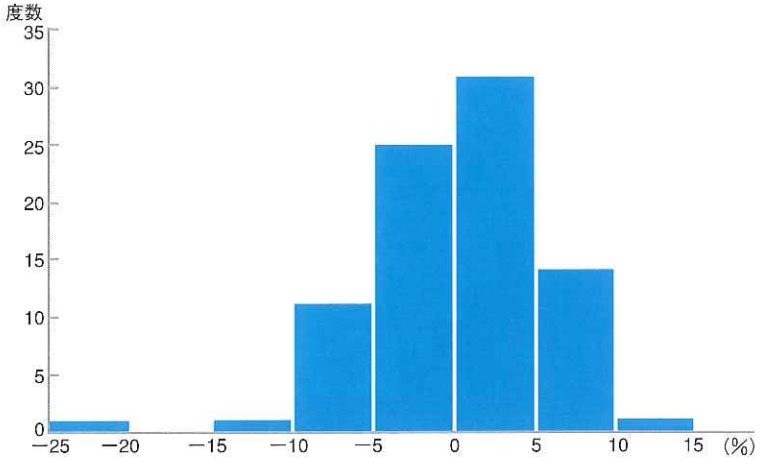
相対度数，累積相対度数などを求める。

□ 2-2 ヒストグラム □

ヒストグラム（あるいは柱状グラフ）は，各階級の相対度数を長方形を使ってグラフ表示したものであり，度数分布表から簡単に作ることができる。質的データでは，各カテゴリーごとに，同一の長さを底辺として，各カテゴリーの度数あるいは相対度数を高さとする長方形を順次作ればよい。また，量的データでも，度数分布表が同じ幅の階級からなっている場合には同様である。実際，表2-4のTOPIXの月次収益率の度数分布表からは，図2-3のようなヒストグラムを作ることができる。ここで，高さには度数そのものを使っている。

しかし，階級の幅が等間隔でない場合には事情が異なる。長方形の面積は，度数そのものではなく，相対度数を表す。このことから，区間が等間隔でな

図2-3 TOPIXの月次収益率のヒストグラム

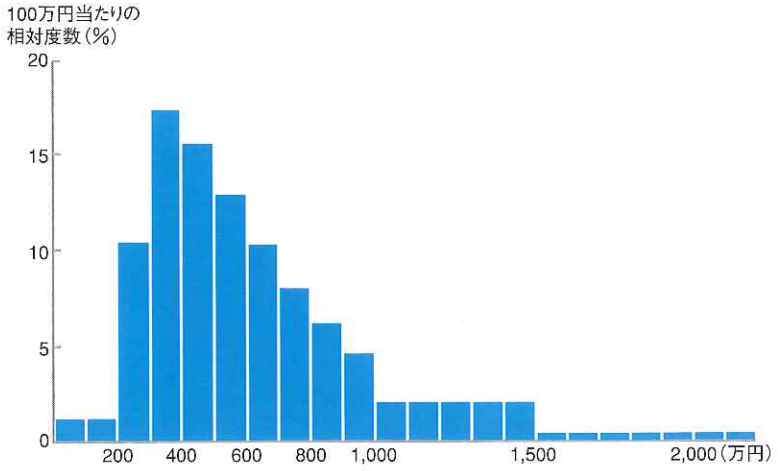


いヒストグラムを作成する場合には，高さを基準幅当たりの相対度数に換算する必要がある。

例題2.3 表2-5の年間収入の度数分布表からヒストグラムを作成せよ。

(解) 度数分布表にはオープン・エンド以外の階級は全部で10個あり，幅が100のものが8個，200と500のものがそれぞれ1個ある。基準幅を100とすれば，幅が500の第10階級の相対度数は基準幅当たり $9.7/5 = 1.94\%$ となるので，この値を長方形の高さとする。高さとして，9.7そのものを使ってはならない。なぜなら，底辺は500であるから，幅が100の階級の5倍であり，面積も5倍になってしまうからである。幅が200の第1階級も同様に考えて，高さを $2.4/2 = 1.2$ とすることになる。オープン・エンドの階級は幅の概念をもたないが，しいて階級平均を階級の中点と考えれば，その幅は936となる。したがって，基準幅当たりの相対度数は $3.0/9.36 = 0.32$ となる。以上の手続きに従って作られた年間収入のデータのヒストグラムが図

図 2-4 年間収入のヒストグラム



2-4 に示されている。縦軸は相対度数であるが，その目盛りは基準幅 100 万円当たりの相対度数（%）であることに注意されたい。

上で説明したヒストグラムの作り方をまとめると次のようになる。

●度数分布表からヒストグラムを作る方法（量的データの場合）

[1] 階級幅が同一で，オープン・エンドの階級がない場合

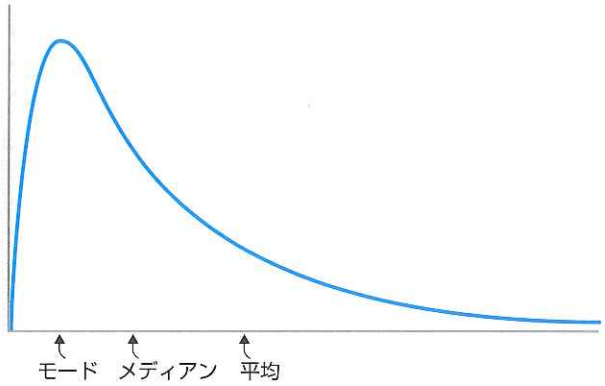
底辺は同一の長さとし，高さは各階級の度数あるいは相対度数とする長方形を各階級ごとに作る。

[2] 階級幅が異なっていたり，オープン・エンドの階級がある場合

まず，基準幅を設定し，相対度数を基準幅当たりの相対度数に換算する。オープン・エンドの階級は，階級平均を階級の中点とするような幅を求め，他の階級と同様に基準幅当たりの相対度数を計算する。そして，各階級の幅を底辺とし，基準幅当たりの相対度数を高さとする長方形を作る。

ヒストグラムは相対度数をグラフ表示したものであり，視覚に訴える点で

図 2-5 右にゆがんだ分布と特性値との関係



は度数分布表よりも優れている。前節で説明したモードなどの中心の特性値も，ヒストグラムと関連付けて考えれば，それらの性質がより明らかになる。例えば，ヒストグラム（を連続的に近似したもの）が図 2-5 のような形であるとしよう。この分布は，右にスソを引いており，右にゆがんだ分布と呼ばれる。右にゆがんだ分布は，データに下限があり，上限が非常に大きいような場合，例えば，所得や貯蓄などの経済データによく見られる。

ヒストグラムの形状が図 2-5 のような場合，モードは峰に対応する点，メディアンは全体を半々に分ける点となることは前に説明した通りである。右にゆがんだ分布では，メディアンはモードよりも大きくなる。他方，平均は，このヒストグラムを下から支えたときにバランスする点，すなわち重心である。平均の計算式（3）は重心の計算式にほかならない。重心はシーソーの支点のようなものであるから，バランスするためには，平均は右側遠方にある異常値に引っ張られて，メディアンよりもさらに大きくなるであろう。図 2-4 に示した年間収入のヒストグラムも右にゆがんだ分布である。したがって，例題 2.1 で計算した 3 つの特性値の大小関係は当然の帰結である。

もし，分布が左にゆがんでいれば，これら 3 つの特性値の大小関係が逆転することが了解できよう。すなわち，分布の形状と 3 つの特性値の間には，

次の関係が成立する。

●分布のゆがみと中心の特性値との関係

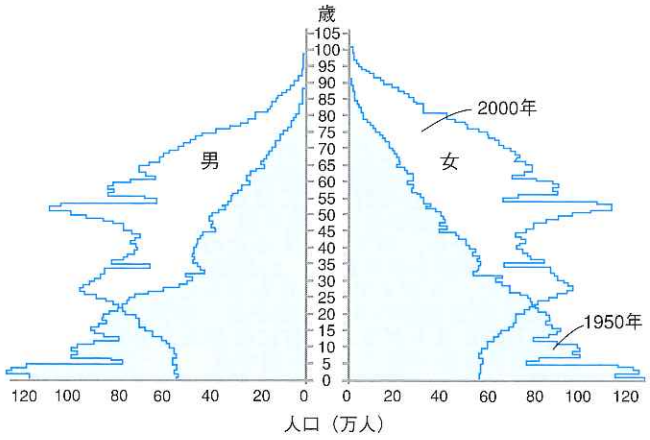
分布が右にゆがんでいる ➡ 平均 > メディアン > モード

分布が左にゆがんでいる ➡ 平均 < メディアン < モード

ここで注意すべきことは，分布の右あるいは左方向へのゆがみは特性値の大小関係を示唆するが，逆は必ずしもいえない，ということである。分布の形状は，対称，右方向へのゆがみ，左方向へのゆがみ，の3つに限定されるわけではないからである。中心の特性値については，次章でもデータそのものの観点から再度取り上げることにする。

度数分布表をグラフ表示するには，ヒストグラム以外にも帯グラフ，円グラフ，人口ピラミッドなどがある。帯グラフと円グラフは質的データの相対度数分布をグラフ表示する場合に使われることが多い。例えば，表2-1や表2-2の度数分布表からは，ヒストグラムよりも円グラフを作る方が普通である。人口ピラミッドは年齢階級ごとの人口を男女別に示したヒストグラムで

図2-6 人口ピラミッド



（出典） 国立社会保障・人口問題研究所（http://www.ipss.go.jp/site-ad/Top Page Data/pyra.html）。

あるが，通常のヒストグラムが垂直型グラフであるのに対して，図2-6のように，水平型グラフで表示するものである。

□ 2-3 累積相対度数折れ線 □

ヒストグラムが相対度数をグラフ表示するものであるのに対して，累積相対度数折れ線は，累積相対度数をグラフ表示したものであり，ヒストグラムと異なり，量的データに対してのみ有効である。折れ線は0から1に増加するように描かれるが，階級の作り方に応じて若干の違いがあり，具体的には次のように作られる。

●累積相対度数折れ線の作り方

［1］階級が区間で与えられている場合

（階級の限界値，累積相対度数）の点をプロットして直線で結ぶ。

［2］各階級が1点からなる場合

累積相対度数を高さとする水平線を当該階級の値から次の階級の値の所まで順次引く。

累積相対度数折れ線は［1］の場合には連続となる。図2-7には表2-5の年間収入の累積相対度数折れ線が示されている。厳密にいえば収入は離散的であるが，とりうる値が多いので連続的とみなしてよい。右にゆがんだ分布の累積相対度数折れ線は，この例のように，最初に大幅な増加，あとはゆっくりと増加する形状となる。

［2］の場合の累積相対度数折れ線は，階級点でジャンプする階段関数となり，とりうる値が離散的で，しかもあまり多くない場合に使われる。例えば，サイコロを100回振ったときに出た目が次のようであったとしよう。

図 2-7 年間収入の累積相対度数折れ線

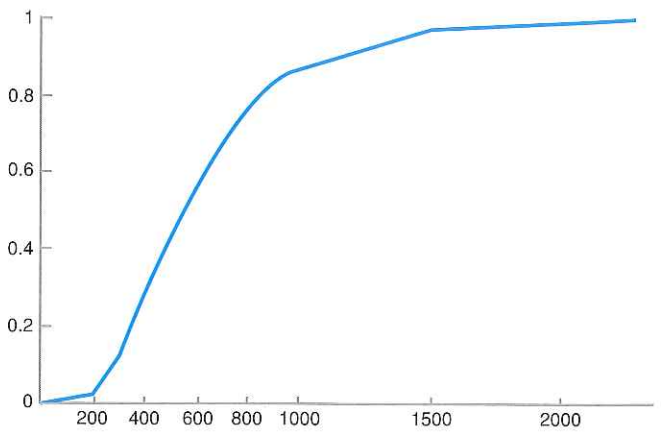
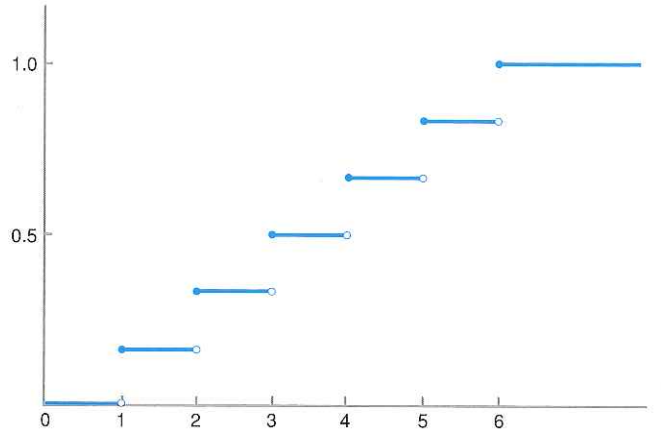


図 2-8 不連続な累積相対度数折れ線



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。

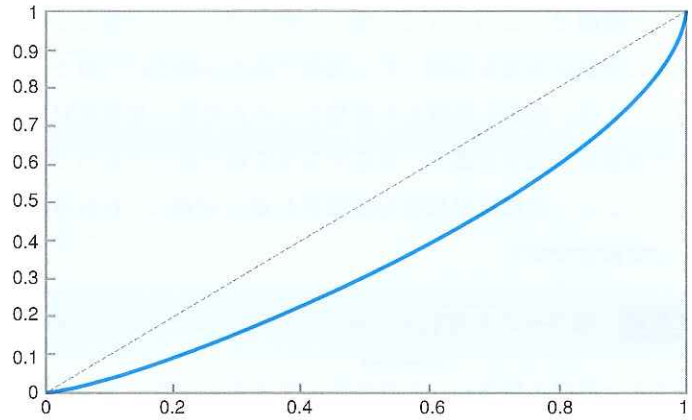
（解） 表 2-5 に低収入階級からの世帯数の累積相対度数があるので，あとは階級ごとの収入の全収入に対する割合（階級収入比）を計算し，その累積比（階級収入累積比）を求めればよい。表 2-7 に，階級収入比，階級収入累積比の値を，表 2-5 の相対度数，累積相対度数とともに示した。ローレンツ曲線は世帯数の累積相対度数，階級収入累積比のペアをプロットして順次結んだ折れ線であり，図 2-9 のようになる。

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

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

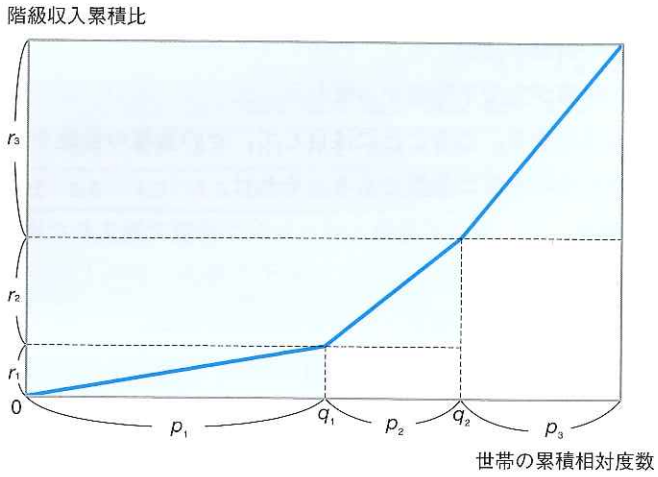
B＝A の中でローレンツ曲線の右側にある直角三角形の面積の総和

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

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，

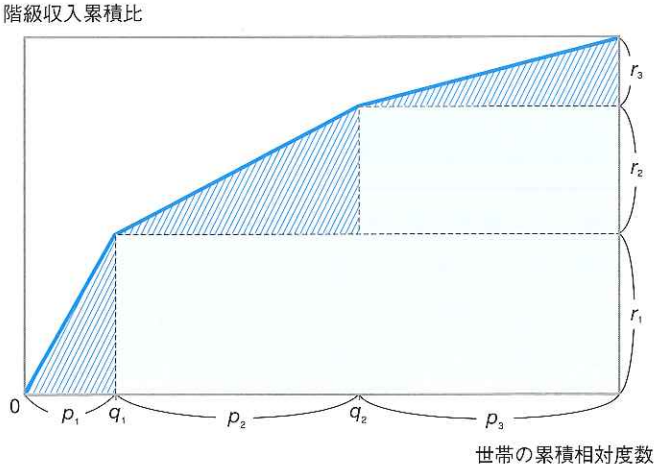
$$2\times\left(C+D-\frac{1}{2}\right)=2\times\left(\sum_{i=1}^{3}(1-q_i)r_i+\frac{1}{2}\sum_{i=1}^{3}p_ir_i-\frac{1}{2}\right)$$

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

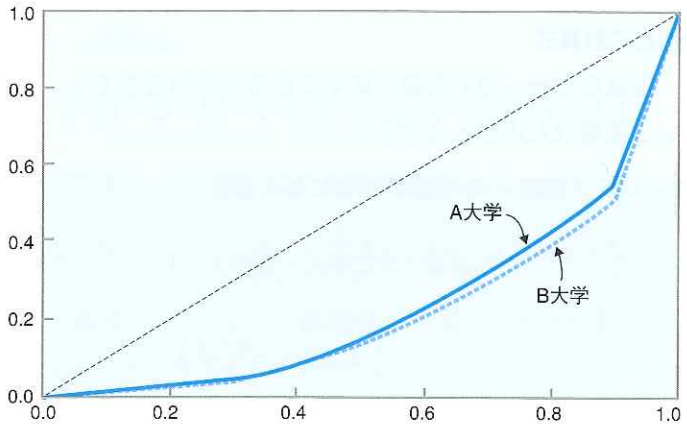


表 2-1 と表 2-2 を相対度数の小さい方から並び替えて，次のような度数分布表を作る。

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。