研究ノート

調査資料とローレンツ面積

― アンケート・データを検定するために ―

四　方　健　雄
餅　田　敬　司
平　井　孝　治

キーワード：ジニ係数　ローレンツ曲線　アンケート　離散量データ　連続量データ　判定値　最大斜線

目　　次

は　じ　め　に

数量的なアンケート・データの歪みを扱うために，筆者平井等は先に「判定値 T」なる概念を提起してきた（『立命館経営学』（第 47 巻第 6 号））。しかしこの判定値は，当初の目的からして離散量データは扱えるが，連続量データには適用できない。そこで所得格差を計量するかの有名な「ジニ係数 G」の考え方を援用する「ローレンツ面積 L」なるものを考案するに至った。

この小論で導入・定義する L 値は，いままで計算してきた判定値 T ときわめて高い正の相関を示し，その相関係数は実に $r = 0.92$ を超える値となった。

ジニ係数 G と違って，判定値 T やローレンツ面積 L は分散も定義できるが，この事実は，異なる二つの母集団から標本を採取した時に，たとえほとんど母平均が変わらない場合でも，集団間にゆがみやひずみに有異な差があるか否かを検定することを可能とする。

このような利点を有するローレンツ面積を導入するために，まずはジニ係数の再定義から論を起こすことにする。

第 1 章　ジニ係数 G とローレンツ面積 L の考え方

第 1 節　ジニ係数の再定義（Redefine）

統計学の基礎として，平均値や中央値は中学の数学でも習う程度に広く知られている。そして，さまざまなデータの特性が平均値や中央値，標準偏差として数値化され示されている。これらの数値の意味するところは，分布の中心からどの程度のバラつきを持っているのかを知るのには重要であるが, どれだけの歪を持って分布しているのかを知るのには不十分である。よって，基本的な統計量である平均値や標準偏差だけでは，分析の対象となったデータ全体のバラつきを，分布という側面から評価することが至難である。すなわち，A という集団と B という集団の平均値や標準偏差が同じであっても，それぞれの集団のデータのバラつき具合が違うことは往々にして見受けられる。

このような場合には，別の方法として，変動係数（CV，Coefficient of Variation）や四分位分散係数などを利用することが知られている。しかし，大学で専門的に統計学を学んだ者であれば，データのひずみやゆがみ，はずれ値などがデータに影響すると判断し，それなりの補正を考慮して知見を出すことになる。変動係数については，詳しくは成書に譲るとして，平均値や中央値，標準偏差からだけでは，分布の歪みやデータのバラつき具合は見えにくい。そのために技術的に四分位分散などの統計計算を行うよりも，データの分布を図で示す方法が普通に利用されている。

医学的な保健衛生動向の分野でデータを考察する場合，例えば，同じ平均寿命である A 県と B 県であっても，死亡年齢の分布まで同じであるとは言い難い。男女比や年齢比を比較することで，その地域固有の疾患や病状が見える事も大いに有りうる。データを分析した値から知見を出す際には，データ全体を俯瞰する意味で，分布を知ることはとても重要である。

そこで, 我々はジニ係数 G（Gini Coefficient）とローレンツ曲線 Lorenz Curve（以下, LC と略す）の考え方を応用して，分布の偏りや散らばりを簡潔に比較するための数量的手法を考案するに至った。そこで，格差評価する上でよく利用されている曲線 LC について簡単に説明する。格差という視点からデータのばらつき具合を計量するための LC は，まず，データを小さいものから大きい順に並べ，並べられた度数の構成要素を累積したデータを横軸に示し，分配されたデータの数量の累積比率を縦軸に表した下に凸なカーブで表された曲線である。

もし，分配される全ての人に対して均等に分配された場合は，以下のグラフ（図 A）の直線，すなわち最大斜線で示されることになる。一方で，格差が最大となる場合とは，他の誰にも分配されずに最後のたった一人に 100% 分配された場合で，底辺の線とグラフの右側の縦の直線で示した太線が格差最大の分配線になる。一般に LC は，最大斜線と格差最大の分配線の間に下に凸なカーブする曲線で示すことができる。すなわち，格差が小さい場合は，最大斜線に近

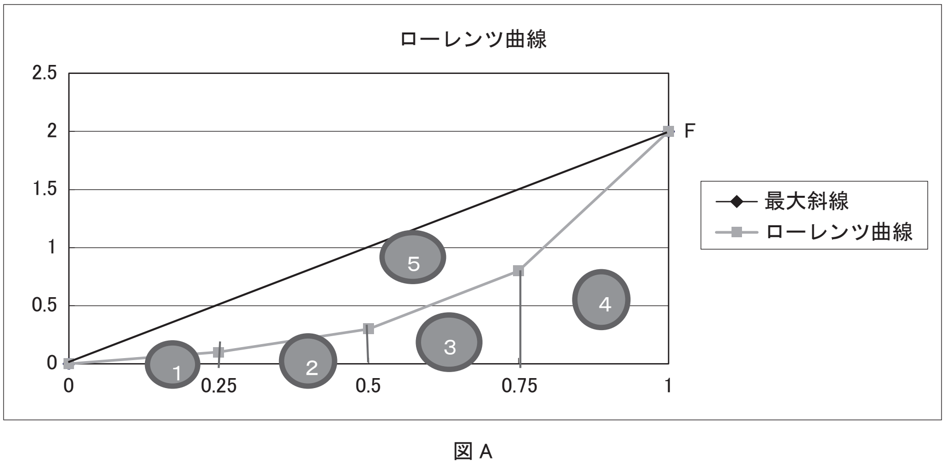
づき，格差が大きい場合は，最大斜線から遠ざかるカーブになる。

ローレンツ曲線LCは視覚的なグラフによって格差をみる事ができるのを特徴としているが，次に示すジニ係数Gは，計量的に格差の程度を表した数量的な指標と言える。もともと，単位がないためにあらゆるものと比較でき，国際間やグループ間の所得格差を示すのに用いられている。格差が均等に分配された時には,「0（ゼロ）」となり,完全な不均衡分配の時には「1」となる。すなわち，ジニ係数Gが示す範囲は，$0 \leqq G \leqq 1$の間の値となる。しかし，一般的に不均等である時には，より小さな（ネガティブな）値で示すほうが自然である。

「1」よりも「0」の方が不均等であると表示するほうが受け入れやすく，「1」に近い方が肯定的な（より平等な）状態とするほうが，数値と形容詞が一致して理解しやすい。ジニ係数のように，数値とそれを示す形容詞が反転している例は他にもある。例えば，為替レートもそうである。円高だと報道される際は，為替レートは小さくなっている。円の価値があがったのであれば，千円が何ドルかのように，数値も大きくなる方が理解しやすい。

そこで,当該指標を求める際に,当該数値が「1」に近い数値のほうがより均等を示すように,ジニ係数Gに換えて,ローレンツ面積Lを以下のように定義する（定義式(＊)）。この式の特徴は,縦軸を初めから横軸の値の2倍に取っておくことで，三角形の面積（最大斜線下の面積）を1にした点である。後の便宜のためこの論文では，従来「均等分配線」とか「均等曲線」と称していたものを，『原点Oから出て終点F（1，2）にいたる線分』で定義し，『最大斜線』と称することにする。

一般的に一目で判断できるようにするには，頭の中であれこれ変換する煩わしさを省いた方がいい。そこで，定義式（＊）のように，最大斜線と格差最大の分配線で囲まれた面積をはなから「1」としておき，そこからローレンツ面積Lを引いた値（数式＊＊）をジニ係数Gとした方が解かりやすい。ただし，このように再定義しても，ジニ係数のローレンツ曲線は原点を



図A

出発して終点は F（1, 2）になるが，後述するアンケート・データなど調査データによってはローレンツ曲線 LC の終点は F（1，2）とは限らないことに留意を要する。

定義式，ローレンツ面積　L ＝①＋②＋③＋④　　………（＊）

再定義，ジニ係数　G ＝ 1 － L　……………………………（＊＊）

このように数値 L とそれに対応する形容詞をそろえておくことは，データ分析した後の知見を導き出す際にも混乱が生じない。その利点と言える例を第 2 節で説明することにして，ローレンツ面積 L を（＊）のように定義しておく。すなわち，L が 0 に近いほど不均等であり，1 に近いほど均等である。

第 2 節　離散量データだけでなく連続量データへの応用

前節で定義したローレンツ面積 L が連続量データでも利用できることを以下で説明する。（詳細は，後の第 3 章で述べる事にするが，L は筆者平井などが以前に開発した判定値 T との正の相関が極めて高い値となった。）何故ローレンツ面積 L を導入するに至ったかについては，アンケートや成績得点などの離散量データはもちろんのこと，医療現場や自然観測などで多数利用されている連続量データにおいても利用可能にしたかったからである。（ちなみに，判定値 T は連続量データでは使えない。）その実例をいくつか紹介するので，参考にしてもらいたい。

まず，一般的なアンケート調査で，ある 5 択の質問に対して，10 人の回答者の全てが中立の「3」を選択した場合を想定する。この場合，サンプル数は n ＝ 10 で，ワースト（1）とベスト（5）の「レンジ r」が「4」なので，「LC 乗数」は「k ＝ 2 / (nr) ＝ 0.05」となる。するとローレンツ曲線 LC は図 B のようになり，LC 下の面積は L ＝ 0.5 となる。

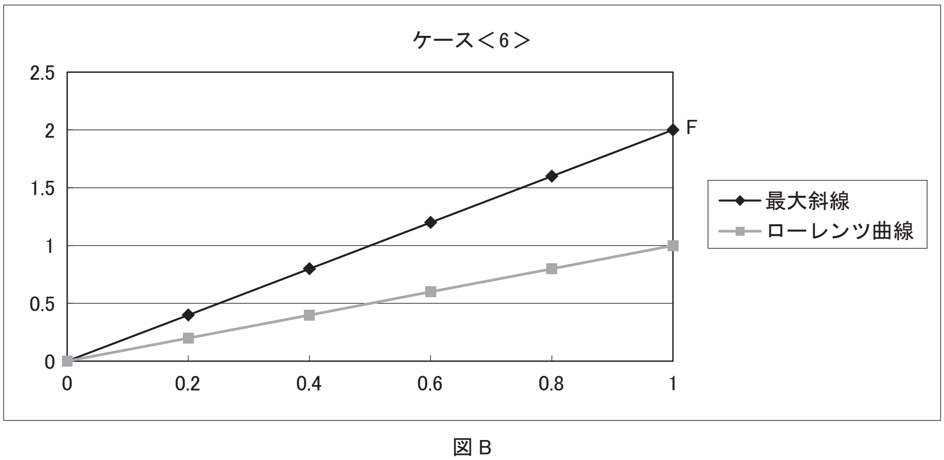
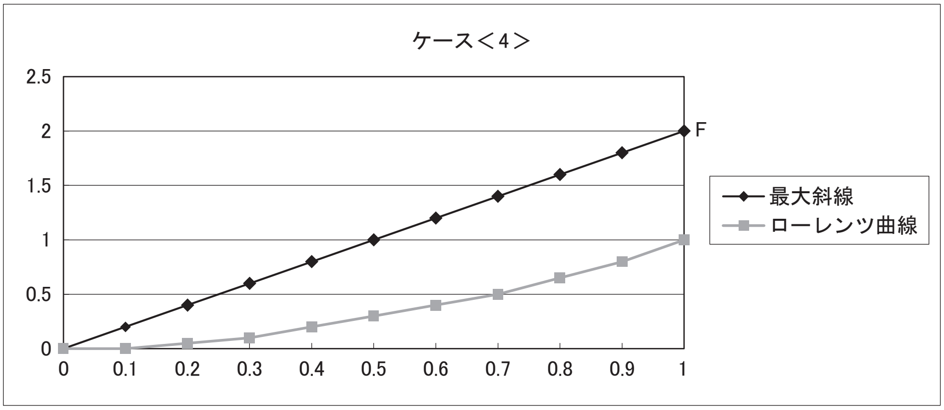


図 B

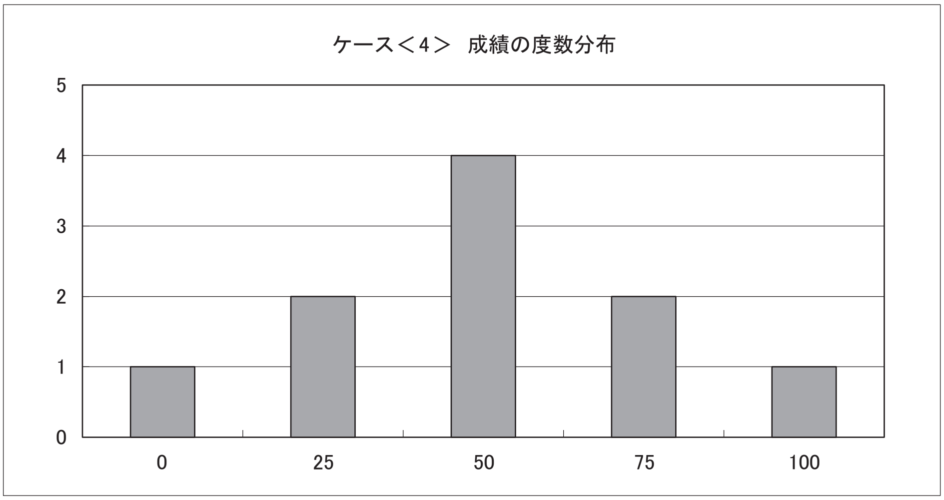
2つ目の例は，5択の質問に10人が回答した分布が，「1」を選択した者は1名，「2」を選択した者は2名，「3」を選択したものは4名，「4」を選択した者は2名で，「5」を選択した者は1名であった場合である。この場合のL曲線は次の図Cのようになる。



図C

この場合のローレンツ面積はL＝0.35となる。

次に，ある集団（10名）に100点満点の試験を課したところ，0点が1名，25点が2名，50点が4名，75点が2名，100点が1名の得点分布になったとしよう。この時のヒストグラムは図Dのようになる。



図D

このケースもn＝10で，レンジはr＝理論上のMax－理論上のMin＝100なので，LC乗数はk＝2/(nr)＝0.002となる。するとL曲線は図Cと同じグラフとなり，曲線下の台

形たちの面積は L ＝ 0.35 と計算される。

この例はアンケートのような離散量データだけではなく，身長のような連続量データでも，「低〜高」へソートしさえすれば，ローレンツ面積 L が求められることを含意している。しかしながら，身長のような連続量データの場合は，試験の成績のように理論的な最大値や最小値が無い場合も考えられる。そのような時でも，求められたデータ群の中での最大値と最小値で代理すれば事足りる。

第 2 章　積上げ方式によるローレンツ面積 L と ρ の定義（Algorithm）

第 1 節　積上げ方式によるローレンツ面積の算定

ローレンツ面積 L についての考え方を前章で導入したので，この章では積上げ方式によるローレンツ面積 L とその分散に対応する ρ^2 の算法（アルゴリズム）について考察する。

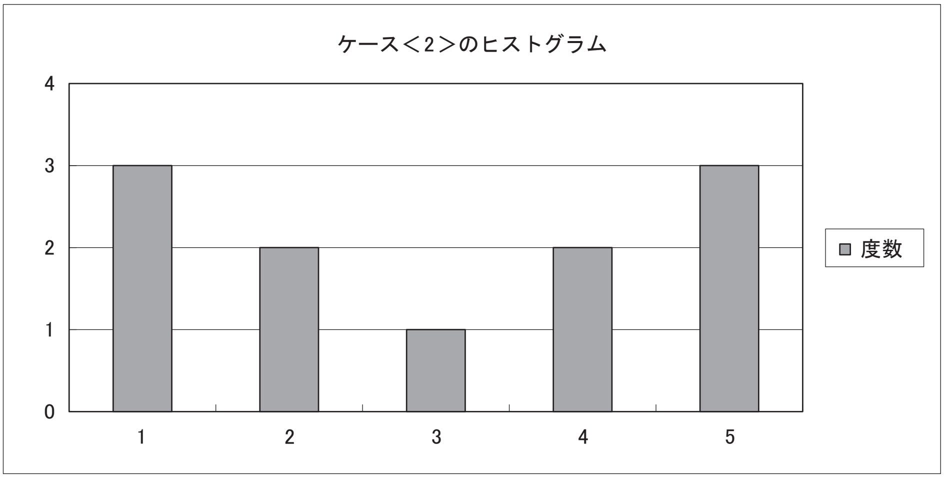


図 2a

図 2a は 5 択のアンケートで，サンプル数が n ＝ 11 で，図のような分布を示している。このときのローレンツ面積 L は，表 0 を用いて計算することが出来る。

選択肢	1	2	3	4	5
得点 y	0.0	0.5	1.0	1.5	2.0
度数	3	2	1	2	3
割合 p	0.273	0.182	0.091	0.182	0.273
Σ割合	0.273	0.455	0.545	0.727	1.000
yp	0.000	0.091	0.091	0.273	0.545
Σyp	0.000	0.091	0.182	0.455	1.000

表 0

表 0 で，各選択肢に対する得点は，原点 O から出発して終点 E（1，2）で終わる最大斜線に整合させるよう，ワーストの選択肢（1）には「0」を，ベストの（5）には「2」を割り当て，他の選択肢には均等になるよう得点を案分している。また，「Σ割合 p」は横座標，**「Σyp」は縦座標を表している。**この表を用いてローレンツ曲線を描くと，次の図 2b のようになる。

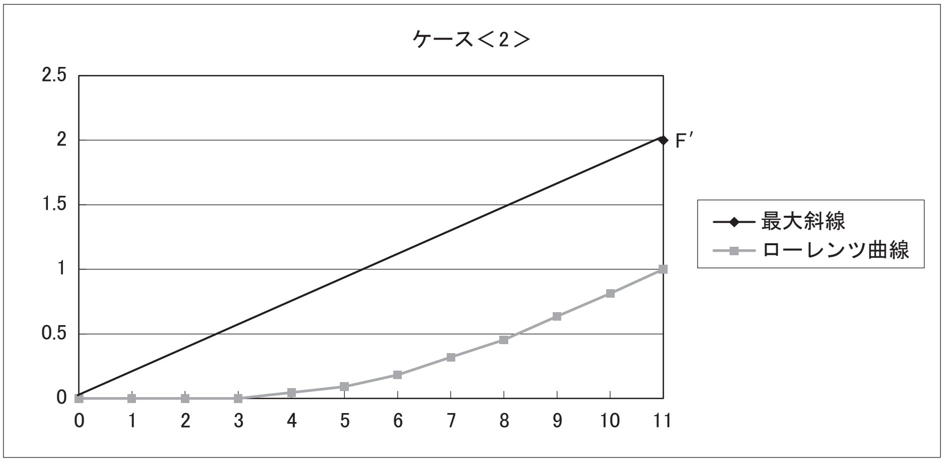


図 2b

この時，ローレンツ曲線下の面積は 1 つの三角形と 7 つの台形の和で，ローレンツ面積は，L ＝ 0.25 である。これを「積上げ方式」で計算すると次のようになる。次の表 2 は，データベースからローレンツ面積 L を計算するためにしつらえたテンプレートである。これは L を求めるアルゴリズムでもあるので，ローレンツ面積の定義と解しても一向にさしつかいない。その際，ついでに判定値 T も同時に求めることにする。

LC 乗数	
n ＝	11
r ＝ M − m	4
k ＝ 2 / (nr)	0.04545

選択肢	判定値
1	−1.000
2	−0.425
3	0.100
4	0.575
5	1.000

			辺長 Y	寄与値 y			個別 判定値
No.	選択肢	−min	0	k ＊ Y			
1	1	0	0	0.0000			−1.0000
2	1	0	0	0.0000			−1.0000
3	1	0	0	0.0000			−1.0000
4	2	1	1	0.0455			−0.4250
5	2	1	2	0.0909			−0.4250
6	3	2	4	0.1818			0.1000
7	4	3	7	0.3182			0.5750
8	4	3	10	0.4545			0.5750
9	5	4	14	0.6364			1.0000
10	5	4	18	0.8182			1.0000
11	5	4	11	0.5000			1.0000
	ローレンツ面積 L ＝平均			0.2769		判定値	0.0364
			Var (L)	0.0831	標本分散	Var (T)	0.6485
			ρ	0.2735	母標準偏差	σ	0.7640

表 2

積上げ方式でローレンツ面積 L を求めるには，まずサンプル・データをワーストからベストにソートしておく。次に n はサンプル数で，レンジは r ＝ Max － min で，LC 乗数を k ＝ 2 / (nr) として求めておく。－ min の列は，最大斜線を原点から出発するための変換である。辺長 Y 列のセル値 Yi は直上と直左を加算し累加して行くのであるが，最後のサンプルについてのみ，その値を 1 / 2 しておく。

というのは，「ローレンツ曲線の原型の面積 L_0」は，台形面積の和より

$$L_0 = \{(0+Y_1)+(Y_1+Y_2)+(Y_2+Y_3)+\cdots\cdots\cdots+(Y_{n-2}+Y_{n-1})+(Y_{n-1}+Y_n)\}/2$$
$$= Y_1+Y_2+\cdots\cdots\cdots+Y_{n-1}+Y_n/2$$

だからである

もしすべてのサンプルが Max なら Y_i ＝ ir となり，y_n ＝ k × Yn ＝ 2 となり，最大斜線の終点の座標はこのままでは（n，2）であることになる。

そこで，更に　$L_1 = y_1+y_2+y_3+\cdots\cdots\cdots+y_{n-2}+y_{n-1}+(y_n/2)$ をサンプル数 n で割ると，即ち平均を取ると，最大斜線の終点は無事 F（1，2）となる。以上の仕掛けをテンプレートにしたのが表 2 である。

このように，ローレンツ面積 L は寄与値 y_i の平均値として求まるので，これには標本分散 Var（L）や母標準偏差 ρ の存在が付随することになる。このケース＜2＞の場合はそれぞれ，L ＝ 0.2769，Var ＝ 0.0831，ρ ＝ 0.2735 と求められる。

なお，筆者平井等は既に『立命館経営学』（第 47 巻第 6 号）で「判定値 T」なるものを導入したが，テンプレートの右側はそれをデータベースから求める Algorithm を示したものである。

第 2 節　その他のケースと判定値との相関係数

その他のケースとして，同じアンケートで，次のような分布（図 5a）の場合のテンプレート（表 5）と，ローレンツ曲線を示しておこう。

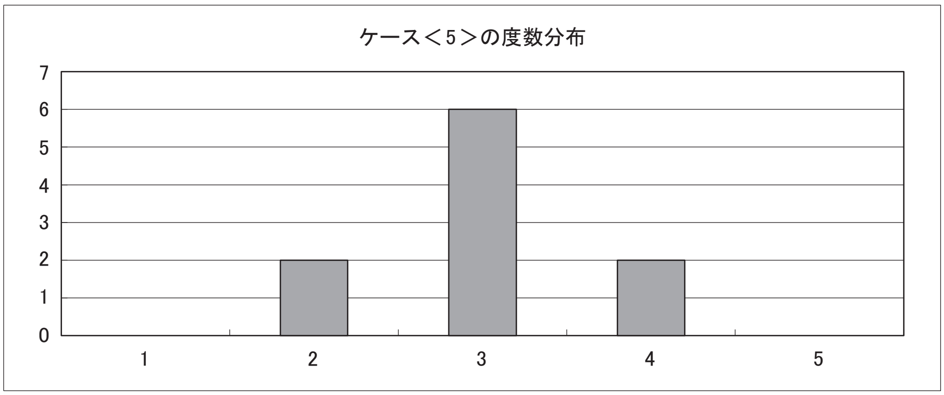


図 5a

LC 乗数	
n ＝	10
r ＝ M － m	4
k ＝ 2 / (nr)	0.0500

選択肢	判定値
1	−1.000
2	−0.425
3	0.100
4	0.575
5	1.000

No.	選択肢	−min	辺長 Y 0	寄与値 y k ＊ Y			個別 判定値
1	2	1	1	0.0500			−0.4250
2	2	1	2	0.1000			−0.4250
3	3	2	4	0.2000			0.1000
4	3	2	6	0.3000			0.1000
5	3	2	8	0.4000			0.1000
6	3	2	10	0.5000			0.1000
7	3	2	12	0.6000			0.1000
8	3	2	14	0.7000			0.1000
9	4	3	17	0.8500			0.5750
10	4	3	10	0.5000			0.5750
	ローレンツ面積 L ＝平均			0.4200		判定値	0.0900
			Var (L)	0.0679	標本分散	Var (T)	0.1113
			ρ	0.2472	母標準偏差	σ	0.3165

表 5

この表からケース＜5＞の場合，L ＝ 0.4200, Var ＝ 0.0679, ρ ＝ 0.2472 となることが判る。なお，判定値 T については，Excel の Vlookup 関数を用いて個別判定値を求め，その平均を取ると自動的に求まる。

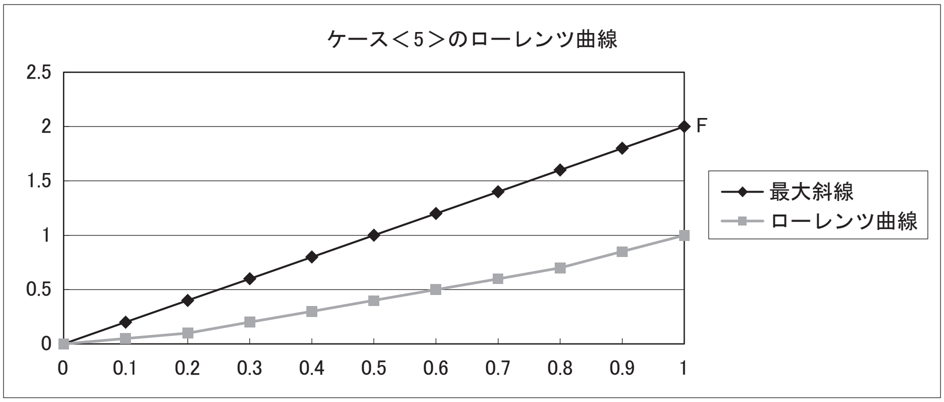


図 5b

今までの紹介から漏れていたケース＜1＞とケース＜3＞の度数分布については，ヒストグラムだけを，図 1a，図 3a として次に示しておく。

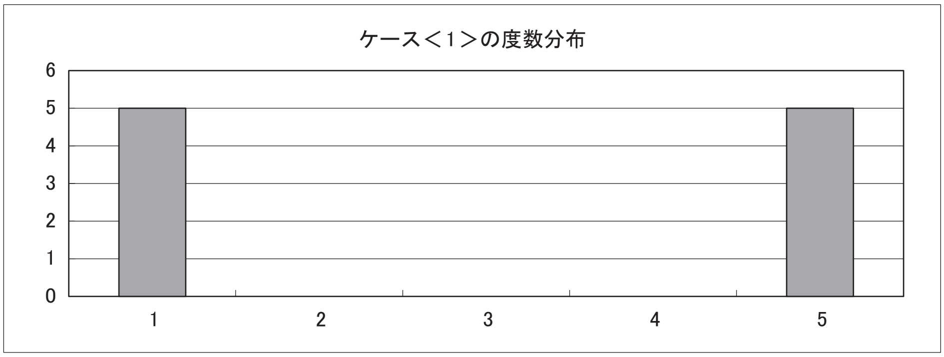


図 1a

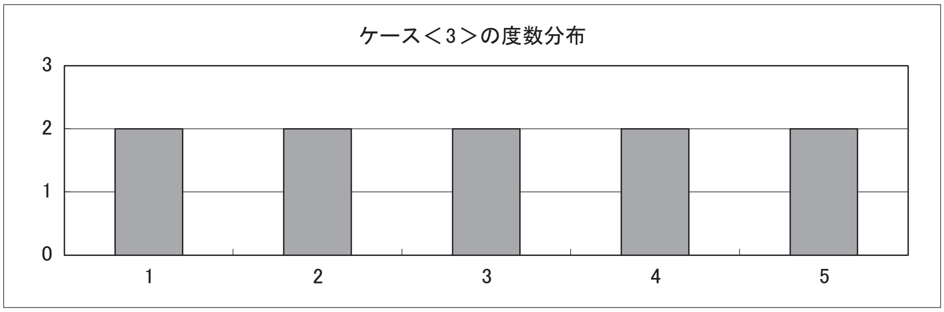


図 3a

以上，六つのケースのローレンツ面積をまとめると，次の表 7 のようになる。

ケース	＜1＞	＜2＞	＜3＞	＜4＞	＜5＞	＜6＞	相関係数
ローレンツ面積 L	0.2500	0.2769	0.3000	0.3500	0.4200	0.5000	0.9286
判定値 T	0.0000	0.0364	0.0500	0.0700	0.0900	0.1000	
L の標本分散	0.0917	0.0832	0.0772	0.0711	0.0679	0.0667	
ρ	0.2872	0.2751	0.2636	0.2530	0.2472	0.2449	0.9216
σ	1.0000	0.7987	0.7083	0.5490	0.3165	0.0000	

表 7

ここで観察されるように，ローレンツ面積 L と判定値 T の間にはきわめて強い正の相関がある。その上さらに，二つの母標準偏差 ρ と σ の間にも同様の事情が現象していることが見て取れる。両者とも元々はアンケート・データの歪みを解析するために開発したものであるから，どちらを使って解析してもいいようなものである。しかしローレンツ面積 L の方は連続量データにも使える点で優れている。これが判定値 T だけではなく，ローレンツ面積を開発した所以で，その例を次章で示す。

第3章　連続量データと検定への応用

第1節　等分散を仮定した離散量データの検定

アンケート・データのような離散量データを検定にかけることもしばしばある。前節までに例示したケース＜1＞とケース＜6＞は，平均値は同じ「3」でも分布は極端に異なっている。このような場合には,「平均値の差の検定」をしても無意味であることはいうまでもない。また，ケース＜6＞の標本分散は明らかにゼロなので，この時の統計検定量Fは無限大となる。

二つのケースのローレンツ面積Lにつき, $L_1 = 0.25$ と $L_6 = 0.50$ の「差の検定」を, 以下, 有意水準12.8%のもとで実施してみると，次のようになる。

	＜1＞	＜6＞
ローレンツ面積L	0.2500	0.5000
自由度	9	9
Lの標本分散	0.0917	0.0667

両側F検定，ケース＜1＞と＜6＞

仮説	母分散に差がない	
有意水準	α	0.1280
確率	α/2	0.0640
閾値	F INV.RT	2.9046
	＜1＞	＜6＞
偏差変動	0.8250	0.6000
自由度	9	9
標本分散	0.0917	0.0667
検定量	分散の比	1.3750

棄却できない

即ち『母分散に特段の差はない』

等分散を仮定した差のt検定

仮説	ローレンツ面積に有意差がない			
	解説	計算		
有意水準	α	0.1280		
確率	α/2	0.0640		
閾値	T.INV.2T	1.9734		
	＜1＞	＜6＞	計算	
ローレンツ面積	0.2500	0.5000	0.2500	←差
偏差変動	0.8250	0.6000	1.4250	←和
加重平均	変動和÷自由度の和		0.0792	←S
サンプル数	10	10		
1/n	0.1000	0.1000	0.2000	←T
S＊T			0.0158	
平方根ε			0.1258	
検定量	差/ε		1.9868	

棄却する

『ローレンツ面積に有意な差がある』

表8

この表の左側では母分散につきF検定を，右側では「L値の差のt検定」の計算過程を示している。その結果,「平均値では差のない」データ群の間でも，ローレンツ面積Lならば有意な差があるとみなされることになる。即ち，歪に違いがある。ちなみに，今日ではExcel 2010のような優れた表計算ソフトがある時代には，自由度が360度を超えるまではt検定を実施すべきであろう。

第2節　非等分散を仮定した連続量データの検定

ローレンツ面積は身長のような連続量でもそれが計れることに利点がある。ここでは身長に関し,「A群からは40名」の,「B群からは25名」の標本を採取し，次のテンプレートにあるような表9のデータを得たとする。

LC 乗数		
サンプル数 n	a	40
	b	25
最大 M		198.0
最小 m		148.0
r ＝ M－m		50.0
k ＝ 2 / (nr)	A	0.00100
	B	0.00160

A			辺長 Y	寄与値 y
No.	身長	－min	0	k ＊ Y
1	148.0	0.0	0.0	0.0000
2	152.0	4.0	4.0	0.0040
3	152.0	4.0	8.0	0.0080
⋮	⋮	⋮	⋮	⋮
38	186.0	38.0	1400.0	1.4000
39	189.0	41.0	1441.0	1.4410
40	190.0	42.0	741.5	0.7415
		ローレンツ面積 L ＝平均		0.7200
		標本分散	Var (L)	0.1168

B			辺長 Y	寄与値 y
No.	身長	－min	0	k ＊ Y
1	150.0	2.0	2.0	0.0032
2	153.0	5.0	7.0	0.0112
3	155.0	7.0	14.0	0.0224
⋮	⋮	⋮	⋮	⋮
23	190.0	42.0	1000.0	1.6000
24	195.0	47.0	1047.0	1.6752
25	198.0	50.0	548.5	0.8776
		ローレンツ面積 L ＝平均		0.8350
		標本分散	Var (L)	0.0640

表 9

このような場合は理論的な最大値 M や最小値 m が無いので，A，B の両群に渡る最大値と最小値でもってその代理と（Surrogate）する。この代理値でレンジ r を量り，テンプレートを用いてローレンツ面積を計算する。このようにして A 群の L 値は La ＝ 0.72 でその標本分散は Va＝0.1168 となったとする。また B 群の L 値と標本分散はそれぞれ L_b＝0.835 と V_b＝0.64 となったとの想定のもとで，先ほどと同じ有意水準 12.8% の F 検定や t 検定をした計算過程が，次の表 10 である。

この表によれば，F 検定の結果，A 群寄与値の母分散と B 群のそれとは有意な差が認められ，非等分散を仮定した t 検定に先立ち，Welch の考案した自由度 f を求めている。これがなんと四捨五入して 63 度となり，A 群と B 群の自由度の和と一致することとなった。

即ちこの場合，（有意水準が何であれ）等非等に限らず，同じ自由度でもって t 検定に臨むことになった珍しい事例である。

それはともあれ，連続量データである両群の L 値の差＝ 0.835 － 0.72 ＝ 0.115 には危険率 12.8% のもとで有意な差が認められたということである。このように，連続量の場合でも，ローレンツ面積には検定ができるという利点がある。

なお，平均概念と独立なジニ係数は「検定」という概念になじまないことを指摘しておく。

	A	B
ローレンツ面積	0.7200	0.8350
自由度	39	24
標本分散	0.1168	0.0640

Welch の自由度

	A	B	計
標本分散	0.1168	0.0640	0.1808
加重係数	0.6460	0.3540	1.0000
加係の平方 S	0.4173	0.1253	
自由度	39	24	↓和
S / 自由度	0.0107	0.0052	0.0159
和の逆数			62.806

よって自由度を 63 度とみなす

両側 F 検定，ケース＜A＞と＜B＞

仮説	母分散に有意差がない	
有意水準	α	0.1280
確率	$\alpha / 2$	0.0640
閾値	F.INV.RT	1.8057

	A	B
偏差変動	4.5552	1.5360
自由度	39	24
標本分散	0.1168	0.0640
検定量	分散の比	1.8250

棄却する

即ち『母分散に有意な差がある』

非等分散を仮定した差の t 検定

仮説	ローレンツ面積に有意差がない		
	解説	計算	
確率	$\alpha / 2$	0.0640	
自由度	Welch の f	63	
閾値	T.INV.2T	1.5423	
	A	B	計算
ローレンツ面積	0.7200	0.8350	0.1150
標本分散 V	0.1168	0.0640	↑差
サンプル数	40	25	↓和
V / n	0.0029	0.00256	0.00548
平方根 ε			0.0740
検定量	差 / ε		1.5535

棄却する

『ローレンツ面積に有意な差がある』

表 10

お　わ　り　に

この小論の主題であるローレンツ面積を論ずるに当たり，この論文ではアンケートの結果が中立の選択肢（3）を軸として，左右対称となっている六つのケース扱ってきた。即ち，平均値が同じでも分布が異なる場合，それをどのように計量し比較・検討したらいいのかを，議論するための素材として，この六つのケースを取り上げた。

ジニ係数のみならずローレンツ面積（や判定値）を計算をするのは，それなりに厄介なものではある。この小論では Excel 2010 など表計算ソフトのワークシート上に然るべき「テンプレート」を設定することで，その煩わしさを回避している。このテンプレートは，ローレンツ面積 L を算定するアルゴリズムを示すのみならず，その分散 Var（L）や標準偏差 ρ もまた同時に計算できる。

この分散が然るべく定義できなければ，検定することは到底かなわなかったのであるが，この「ローレンツ面積 L の分散」を定義しえたのは，筆者四方の貢献である。

記述的な統計を学んだことのある人は誰でもジニ係数を理解しているものと思われるが，遺憾ながらこれは「検定」にはなじまない。このジニ係数を検定に利用できないかと問題を提起したのは筆者餅田であった。この提起を受け，三人の筆者で議論を重ね，たどり着いたのが「ローレンツ面積L」である。

なお，再定義したジニ係数Gのローレンツ曲線の終点は必ずE（1，2）であるが，我々の考案したローレンツ面積Lの終点F（1，y）のyは$0 \leqq y \leqq 2$である点に留意してもらいたい。