ローレンツ曲線とジニ係数に関する覚え書き

A Note On the Lorenz Curve and the Gini Coefficient

井　上　勝　雄

Since a relatively early stage in the history of statistical economic analysis, the Lorenz curve has been used as a visual representation of income inequality, and the Gini coefficient has been derived from it as a quantitative indicator. Because of this background, the Lorenz curve tends to be considered as a method developed in order to display income inequality. However, this is not an accurate interpretation. In general, the Lorenz curve or the Gini coefficient is one of the quantitative characteristics of the data, which indicates the degree of dispersion, as well as the variance, standard deviation and coefficient of variation. The purpose of this paper is to describe some of propositions about the Lorenz curve and the Gini coefficient.

Katsuo Inoue

JEL：C00, C10

キーワード：ローレンツ曲線、ジニ係数、格差指標

Key words：the Lorenz curve, the Gini coefficient, degree of dispersion

はじめに

所得の格差を表すとき、数量経済分析の比較的早い時期から、ローレンツ曲線を描き、数量的な指標としてジニ係数を算出することがなされてきた。このような事情からか、ローレンツ曲線を描くのは所得格差を表示するものと想われがちである。しかし、ローレンツ曲線とそれに関わって計算されるジニ係数は、一般的に、データの持つ数量的内容に関して個体間の格差や散らばり程度あるいは、ばらつきの程度を表すものであって、その限りでは、その意味する

ところは分散や標準偏差あるいは変化係数と同様なのである。本稿において、ローレンツ曲線とそれから導かれるジニ係数とに関わって、いくつかの命題を述べておきたい。

個別データによる分析　〜その枠組み

いま、分析のためのデータが n 個あるとしよう。もちろん、それらのデータは所得に関するデータであっても、全く所得とは関係のないデータであってもよい。ともかく分析対象になっている変量の観測値が n 個あるとする。

分析対象である変量の n 個のデータを、$\{x_1,\ x_2,\ x_3,\ \cdots,\ x_n\}$ で表し、一般性を失うことなく、

$$x_1 \leqq x_2 \leqq x_3 \leqq \cdots \leqq x_n$$

であるとしてよい。

ローレンツ曲線と完全平等曲線

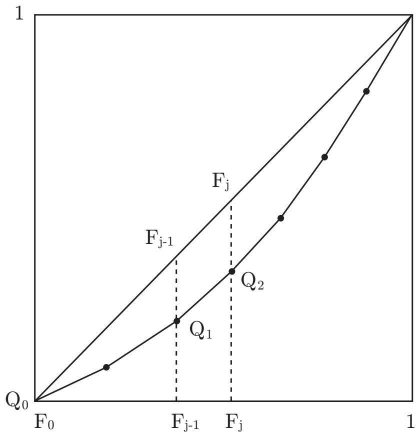
第 j 番目の変量データは x_j であるが、このデータまでの累積相対度数を F_j で示し、変量に関する累積値を Q_j で表す。そうすると、

$$F_j = \frac{j}{n} \qquad Q_j = \frac{\displaystyle\sum_{k=1}^{j} x_k}{\displaystyle\sum_{i=1}^{n} x_i} \qquad (j = 1, 2, \cdots, n)$$

である。ここで、$F_0 = 0 \quad Q_0 = 0$ と定義しておこう。そして、横軸に累積相対度数 F_j をとり、縦軸に変量に関する累積値 Q_j をとった平面において、$(n+1)$ 個の点 $(F_0,\ Q_0) = (0, 0)$, $(F_1,\ Q_1)$, $(F_2,\ Q_2)$, $\cdots$ $(F_n,\ Q_n) = (1, 1)$ を結んでできる図形がローレンツ曲線である。

そして、$(0, 0)$ と $(1, 1)$ とを結んでできる直線は完全平等曲線と呼ばれる。n 個の変量データすべてが等しい値のとき、この一組のデータからローレンツ曲線を描こうとすると、結局、完全平等曲線を描くことになる。

図 1：ローレンツ曲線と完全平等曲線



ローレンツ曲線と変量データの格差

先に定義したように、手元にある n 個の変量データから描いたローレンツ曲線と完全平等曲線とを比較して、ローレンツ曲線が完全平等曲線を離れて下に描かれるほど、変量データは格差が大きいと言える。ローレンツ曲線が横軸に近いほど、分析対象である変量データは格差あるいはバラツキ程度は大きい。

そこで、ローレンツ曲線と完全平等曲線とで囲まれる図形の大きさを、格差あるいは不平等度の指標にする。実際、格差あるいは不平等の程度を表す指標は、この図形の面積の 2 倍とし、これをジニ係数と呼ぶ。ローレンツ曲線と完全平等曲線とで囲まれる図形の最小値はゼロであり、最大値は $\frac{1}{2}$ である。不平等の程度を表す指標の最小値をゼロ、最大値を 1 とするため、不平等の指標は、図形面積の 2 倍と定義しているのである。

ジニ係数

さて、ローレンツ曲線と完全平等曲線とで囲まれる図形は、最も左と最も右に位置するのは三角形であるが、これを台形の特殊ケースと見なせば、n 個の台形に分かれる。したがって、図形の面積は、

$$\sum_{j=1}^{n} \frac{1}{2}\left[(F_{j-1}-Q_{j-1})+(F_j-Q_j)\right](F_j-F_{j-1})$$
$$=\frac{1}{2}\sum_{j=1}^{n}\left[(F_{j-1}-Q_{j-1})+(F_j-Q_j)\right]\frac{1}{n}$$

である。したがって、いま、ジニ係数の値を S で表すと、

$$\begin{aligned}
S &= \sum_{j=1}^{n}\left[(F_{j-1}-Q_{j-1})+(F_j-Q_j)\right]\frac{1}{n}\\
&= \sum_{j=1}^{n}\left\{\left(\frac{j-1}{n}-\frac{\sum_{k=1}^{j-1}x_k}{\sum_{i=1}^{n}x_i}\right)+\left(\frac{j}{n}-\frac{\sum_{k=1}^{j}x_k}{\sum_{i=1}^{n}x_i}\right)\right\}\frac{1}{n}\\
&= \sum_{j=1}^{n}\left\{\frac{2j-1}{n}-\frac{\sum_{k=1}^{j-1}x_k+\sum_{k=1}^{j}x_k}{n\bar{x}}\right\}\frac{1}{n} \qquad \left(\because \quad \sum_{j=1}^{n}x_i=n\bar{x}\right)\\
&= \frac{1}{n^2\bar{x}}\sum_{j=1}^{n}\left[(2j-1)\bar{x}-\sum_{k=1}^{j-1}x_k-\sum_{k=1}^{j}x_k\right]\\
&= \frac{1}{n^2\bar{x}}\sum_{j=1}^{n}\left[2\left(j\bar{x}-\sum_{k=1}^{j}x_k\right)+x_j-\bar{x}\right]\\
&= \frac{2}{n^2\bar{x}}\sum_{j=1}^{n}\left(j\bar{x}-\sum_{k=1}^{j}x_k\right)+\frac{1}{n^2\bar{x}}\sum_{j=1}^{n}(x_j-\bar{x})\\
&= \frac{2}{n^2\bar{x}}\sum_{j=1}^{n}\left(j\bar{x}-\sum_{k=1}^{j}x_k\right) \qquad \left(\because \quad \sum_{j=1}^{n}(x_j-\bar{x})=0\right)\\
&= \frac{2}{n^2\bar{x}}\left[\sum_{j=1}^{n}j\bar{x}-\sum_{j=1}^{n}\sum_{k=1}^{j}x_k\right]
\end{aligned}$$

ところで[1)]、

1) 以下の展開で、一見、行列の演算に見えるが、実は、そうではない。通常の演算式であるが、演算のルールが見通せるよう、行列の書式を維持している。

$$\sum_{j=1}^{n}\sum_{k=1}^{j}x_k=\begin{pmatrix}+x_1 & & & & \\ +x_1 & +x_2 & & & \\ +x_1 & +x_2 & +x_3 & & \\ \vdots & \vdots & \vdots & \cdots & \\ +x_1 & +x_2 & +x_3 & \cdots & +x_n\end{pmatrix}=\begin{pmatrix}+x_1 & +x_1 & +x_1 & \cdots & +x_1 & +x_1 \\ & +x_2 & +x_2 & \cdots & +x_2 & +x_2 \\ & & +x_3 & \cdots & +x_3 & +x_3 \\ & & & \cdots & \vdots & \vdots \\ & & & & +x_{n-1} & +x_{n-1} \\ & & & & & +x_n\end{pmatrix}$$

$$=\sum_{j=1}^{n}(n+1-j)x_j$$

$$=(n+1)\sum_{j=1}^{n}x_j-\sum_{j=1}^{n}jx_j=(n+1)n\bar{x}-\sum_{j=1}^{n}jx_j \quad \cdots\cdots\cdots \quad (1)$$

であることより、

$$\sum_{j=1}^{n}j\bar{x}-\sum_{j=1}^{n}\sum_{k=1}^{j}x_k=\bar{x}\sum_{j=1}^{n}j-(n+1)n\bar{x}+\sum_{j=1}^{n}jx_j$$

$$=\frac{n(n+1)}{2}\bar{x}-n(n+1)\bar{x}+\sum_{j=1}^{n}jx_j$$

$$=\sum_{j=1}^{n}jx_j-\frac{n(n+1)}{2}\bar{x} \quad \cdots\cdots\cdots\cdots\cdots \quad (2)$$

である。ゆえに、$S=\dfrac{2}{n^2\bar{x}}\left[\sum_{j=1}^{n}jx_j-\dfrac{n(n+1)}{2}\bar{x}\right]$

データの差の平均

変量データの格差あるいは散らばり程度の指標として、一般的には、分散、標準偏差、変化係数などがある。さらに、上述のローレンツ曲線を描き、それから導くジニ係数も、変量データの格差や散らばり程度の指標なのである。そうして、今ひとつ、格差や散らばり、バラツキあるいは不平等度の指標として、各データの差の平均がある。

2 つのデータの差は $|x_j - x_k|$ と表せる。ここで、$j = 1, 2, \cdots, n$ および $k=1, 2, \cdots, n$ である。したがって、それ自身も含めると、変量データの差は、総計で n^2 個できるので、変量データの差の平均を $D=\dfrac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}|x_j-x_k|$ と定義する。

この値 D が標本平均の 2 倍に対する比を、分析対象である変量データの格差や散らばり、バラツキ程度の指標 G にする。そして、結論を先取りして言えば、実は、この指標 G は先述のジニ係数の値に等しくなるのである。

さて、上述のように指標 G を定義すれば、

$$G=\frac{1}{2\bar{x}}D=\frac{1}{2n^2\bar{x}}\sum_{j=1}^{n}\sum_{k=1}^{n}|x_j-x_k|$$

である。ここで、

$$\sum_{j=1}^{n}\sum_{k=1}^{n}|x_j-x_k|=\begin{pmatrix} +|x_1-x_1| & +|x_1-x_2| & +|x_1-x_3| & \cdots & +|x_1-x_{n-1}| & +|x_1-x_n| \\ +|x_2-x_1| & +|x_2-x_2| & +|x_2-x_3| & \cdots & +|x_2-x_{n-1}| & +|x_2-x_n| \\ +|x_3-x_1| & +|x_3-x_2| & +|x_3-x_3| & \cdots & +|x_3-x_{n-1}| & +|x_3-x_n| \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ +|x_n-x_1| & +|x_n-x_2| & +|x_n-x_3| & \cdots & +|x_n-x_{n-1}| & +|x_n-x_n| \end{pmatrix}$$

$$=2\times\begin{pmatrix} +(x_1-x_1) & & & & & \\ +(x_2-x_1) & +(x_2-x_2) & & & & \\ +(x_3-x_1) & +(x_3-x_2) & +(x_3-x_3) & & & \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ +(x_n-x_1) & +(x_n-x_2) & +(x_n-x_3) & \cdots & +(x_n-x_{n-1}) & +(x_n-x_n) \end{pmatrix}$$

$$=2\sum_{j=1}^{n}\sum_{k=1}^{j}(x_j-x_k)=2\sum_{j=1}^{n}\{(x_j-x_1)+(x_j-x_2)+\cdots+(x_j-x_j)\}$$

$$=2\sum_{j=1}^{n}\{jx_j-\sum_{k=1}^{j}x_k\}=2\left[\sum_{j=1}^{n}jx_j-\sum_{j=1}^{n}\sum_{k=1}^{j}x_k\right]$$

$$=2\left[\sum_{j=1}^{n}jx_j-(n+1)n\bar{x}+\sum_{j=1}^{n}jx_j\right]$$

(先述の (1) より、$\sum_{j=1}^{n}\sum_{k=1}^{j}x_k=(n+1)n\bar{x}-\sum_{j=1}^{n}jx_j$ である)

$$=2\left[2\sum_{j=1}^{n}jx_j-2\frac{n(n+1)}{2}\bar{x}\right]=4\left[\sum_{j=1}^{n}jx_j-\frac{n(n+1)}{2}\bar{x}\right]$$

ゆえに、$G=\dfrac{2}{n^2\bar{x}}\left[\sum_{j=1}^{n}jx_j-\dfrac{n(n+1)}{2}\bar{x}\right]$ ………………………… (3)

以上に示したことは、ローレンツ曲線と完全平等曲線とで囲まれる図形の面積の 2 倍 S と各データの差の平均が標本平均の 2 倍に対する比 G とは等し

く、それをジニ係数と定義できることが証明された。

ジニ係数の別表現

さて、データの順序数列 $(1, 2, 3, \cdots, n)$ の平均および分散、標準偏差を、それぞれ、$\bar{J}$および s_J^2、s_J と表すと、

$$\bar{J} = \frac{1}{n}\sum_{j=1}^{n} j = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$

$$s_J^2 = \frac{1}{n}\left[\sum_{j=1}^{n} j^2 - n\bar{J}^2\right] = \frac{1}{n}\left[\frac{1}{6}n(n+1)(2n+1) - n(\frac{n+1}{2})^2\right] = \frac{n^2-1}{12}$$

である。また、分析対象である変量データ $\{x_1, x_2, x_3, \cdots, x_n\}$ の平均および分散、標準偏差は、$\bar{x}$、s_x^2、s_x と表す。

さらに、順序数列 J と変量データ x との共分散 $\mathrm{Cov}(J, x)$ は、

$$\mathrm{Cov}(J, x) = \frac{1}{n}\sum_{j=1}^{n}(j - \bar{J})(x_j - \bar{x}) = \frac{1}{n}\left[\sum_{j=1}^{n} jx_j - n\bar{J}\bar{x}\right]$$

$$= \frac{1}{n}\left[\sum_{j=1}^{n} jx_j - \frac{n(n+1)}{2}\bar{x}\right]$$

である。

これらを利用すると、ジニ係数は、次のように展開できる。

$$G = S = \frac{2}{n^2\bar{x}}\left[\sum_{j=1}^{n} jx_j - \frac{n(n+1)}{2}\bar{x}\right]$$

$$= \frac{2s_J}{n} \cdot \frac{s_x}{\bar{x}} \cdot \frac{1}{s_J s_x}\frac{1}{n}\left[\sum_{j=1}^{n} jx_j - \frac{n(n+1)}{2}\bar{x}\right]$$

$$= \sqrt{\frac{n^2-1}{3n^2}}CV_x \cdot R_{Jx} \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (4)$$

ここで、

$$CV_x = \frac{s_x}{\bar{x}} \qquad R_{Jx} = \frac{1}{s_J s_x}\mathrm{Cov}(J, x)$$

であり、それぞれ変量データ x の変化係数と、順序数列 J と変量データ x の

相関係数[2)]である。

上の式 (4) から分かることは、ジニ係数は、変量データの変化係数と順序数列と変量データの相関係数との積に比例するということである。

度数分布表による分析　〜その枠組み

大量のデータを扱うとき、データ全体の有り様を見やすくするため、しばしば度数分布表にまとめる。度数分布表によって、(1) データ全体の情報を階級値と階級度数に集約する。様々な値を持つ膨大な数のデータを、各階級度数の個数だけ当該階級値のデータが存在すると縮約する。そうして、(2) 分析対象である変量の数量的指標を昇順に並べて、全体の分布状況を見易すくする。

一般的に、全部で n 個あるデータを m 個の階級に分類した度数分布表は、次の表 1 の形式をとる。

表 1: 度数分布表の一般型

	階級			階級値	度数	相対度数	累積相対度数				
				Y	n	f	F	Yf	Y^2f	q	Q
1)	Y_1^L	$\sim$	Y_1^H	Y_1	n_1	f_1	F_1	Y_1f_1	$Y_1^2f_1$	q_1	Q_1
2)	Y_2^L	$\sim$	Y_2^H	Y_2	n_2	f_2	F_2	Y_2f_2	$Y_2^2f_2$	q_2	Q_2
3)	Y_3^L	$\sim$	Y_3^H	Y_3	n_3	f_3	F_3	Y_3f_3	$Y_3^2f_3$	q_3	Q_3
		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k)	Y_k^L	$\sim$	Y_k^H	Y_k	n_k	f_k	F_k	Y_kf_k	$Y_k^2f_k$	q_k	Q_k
		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m)	Y_m^L	$\sim$	Y_m^H	Y_m	n_m	f_m	F_m	Y_mf_m	$Y_m^2f_m$	q_m	Q_m
$\sum$					n	1.0000		$\bar{Y}$	$\sum Y^2f$	1.0000	

2) 順序数列 J と変量データ x の相関係数は、変量データの等間隔性の程度である。例えば、分析対象の変量データが等差数列的な値であるとき、$R_{Jx}=1$ になる。

度数分布表において、相対度数 f および累積相対度数 F は、次の通りである。

$$f_k = \frac{n_k}{n}, \qquad F_k = F_{k-1} + f_k \qquad k = 1,\ 2,\ \cdots,\ m$$

ただし、$f_0 = F_0 = 0$ と定義する。

また、一般に、分析対象である変量データ全体の平均値 $\bar{Y}$ や分散 s_Y^2、標準偏差を計算するため、Yf の欄、$Y^2 f$ の欄を追加する。周知の通り、$\bar{Y} = \sum Y_k f_k$ であり、$s_Y^2 = \sum Y_k^2 f_k - \bar{Y}^2$ である。

さらに、度数分布表からローレンツ曲線を描くには、上記のように、q 欄および Q 欄を加えるのが適切である。q 欄の各値は、各階級の変量データの合計が全データの総和に占める割合が記される。そうして、Q 欄には、その累積値が記載される。

したがって、q 欄および Q 欄の q_k、Q_k は次の通りである。

$$q_k = \frac{Y_k f_k}{\bar{Y}} = \frac{Y_k f_k \times\ n}{\bar{Y} \times\ n} = \frac{Y_k n_k}{\sum Y_t n_t} \qquad k = 1,\ 2,\ \cdots,\ m$$

$$Q_k = Q_{k-1} + q_k = \sum_{j=1}^{k} q_j = \frac{\displaystyle\sum_{i=1}^{k} Y_i f_i}{\bar{Y}} \qquad k = 1,\ 2,\ \cdots,\ m$$

ただし、$q_0 = Q_0 = 0$ と定義する。

上述の枠組みのもとで、ローレンツ曲線は、個別データによる分析のケースと同様に、$(F_0, Q_0) = (0,\ 0)$, (F_1, Q_1), (F_2, Q_2), (F_3, Q_3), $\cdots$, $(F_m, Q_m) = (1.0,\ 1.0)$ を結んでできる図形である。

ローレンツ曲線から導くジニ係数

度数分布表からローレンツ曲線を描く手順を上述した。次に、このローレンツ曲線と完全平等曲線とで囲まれる図形の面積を利用して、変量データの格差指標であるジニ係数を導出する。

個別データによるケースでは、ローレンツ曲線と完全平等曲線とで囲まれる図形は、いくつかの台形に分かれることを利用した。ここでは、後述内容のため、次のように考える。

ローレンツ曲線と完全平等曲線とで囲まれる図形の面積は、完全平等曲線と横軸および直線 $F=1$ とででき、面積が $\dfrac{1}{2}$ の直角 2 等辺三角形から、ローレンツ曲線と横軸および直線 $F=1$ とでできる図形の面積を差し引いて求める。

ところで、ローレンツ曲線と横軸および直線 $F=1$ とで囲まれてできる図形は、(F_1, Q_1), (F_2, Q_2), $(F_3, Q_3)\cdots(F_{m-1}, Q_{m-1})$ の各点から横軸へ垂線をおろせば、最も左側の図形は三角形であるが、m 個の台形に分かれる。したがって、ローレンツ曲線と完全平等曲線とで囲まれる図形の面積は、

$$\frac{1}{2}-\sum_{j=1}^{m}\frac{1}{2}(Q_{j-1}+Q_j)(F_j-F_{j-1})=\frac{1}{2}-\frac{1}{2}\sum_{j=1}^{m}(Q_{j-1}+Q_j)f_j$$

である。したがって、変量の格差指標であるジニ係数 S は、上に求めた図形面積の 2 倍であるから、

$$S=1-\sum_{j=1}^{m}(Q_{j-1}+Q_j)f_j \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (5)$$

である。

度数分布表から導くデータの差の平均

度数分布表による分析では、手元にあるデータは、値が Y_j であるデータが n_j 個あり、値が Y_k であるデータが n_k 個あると、データ全体を縮約している。したがって、データの差が $|Y_j-Y_k|$ となるのは、$n_j\times n_k$ だけの度数があると見なさなければならない。このことに注意して、個別データによる分析のケースと同様に定義する。つまり、ジニ係数は、データの差の平均が標本平均 $\bar{Y}$ の 2 倍に対する比率である。そうすると、ジニ係数 G は次の通りである。

$$G=\frac{1}{2\bar{Y}}\cdot\frac{1}{n^2}\sum_{j=1}^{m}\sum_{k=1}^{m}|Y_j-Y_k|n_jn_k=\frac{1}{2\bar{Y}}\sum_{j=1}^{m}\sum_{k=1}^{m}|Y_j-Y_k|f_jf_k$$

ところで、$\displaystyle\sum_{j=1}^{m}\sum_{k=1}^{m}|Y_j-Y_k|f_jf_k=D$ とすると、

$$D=\begin{pmatrix} +f_1|Y_1-Y_1|f_1 & +f_1|Y_1-Y_2|f_2 & +f_1|Y_1-Y_3|f_3 & \cdots & +f_1|Y_1-Y_m|f_m \\ +f_2|Y_2-Y_1|f_1 & +f_2|Y_2-Y_2|f_2 & +f_2|Y_2-Y_3|f_3 & \cdots & +f_2|Y_2-Y_m|f_m \\ +f_3|Y_3-Y_1|f_1 & +f_3|Y_3-Y_2|f_2 & +f_3|Y_3-Y_3|f_3 & \cdots & +f_3|Y_3-Y_m|f_m \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ +f_m|Y_m-Y_1|f_1 & +f_m|Y_m-Y_2|f_2 & +f_m|Y_m-Y_3|f_3 & \cdots & +f_m|Y_m-Y_m|f_m \end{pmatrix}$$

$$=2\begin{pmatrix} +f_1(Y_1-Y_1)f_1 & & & & \\ +f_2(Y_2-Y_1)f_1 & +f_2(Y_2-Y_2)f_2 & & & \\ +f_3(Y_3-Y_1)f_1 & +f_3(Y_3-Y_2)f_2 & +f_3(Y_3-Y_3)f_3 & & \\ \cdots & \cdots & \cdots & \cdots & \\ +f_m(Y_m-Y_1)f_1 & +f_m(Y_m-Y_2)f_2 & +f_m(Y_m-Y_3)f_3 & \cdots & +f_m(Y_m-Y_m)f_m \end{pmatrix}$$

$$=2\begin{pmatrix} +f_1Y_1f_1 & & & & \\ +f_2Y_2f_1 & +f_2Y_2f_2 & & & \\ +f_3Y_3f_1 & +f_3Y_3f_2 & +f_3Y_3f_3 & & \\ \cdots & \cdots & \cdots & \cdots & \\ +f_mY_mf_1 & +f_mY_mf_2 & +f_mY_mf_3 & \cdots & +f_mY_mf_m \end{pmatrix}$$

$$-2\begin{pmatrix} +f_1Y_1f_1 & & & & \\ +f_2Y_1f_1 & +f_2Y_2f_2 & & & \\ +f_3Y_1f_1 & +f_3Y_2f_2 & +f_3Y_3f_3 & & \\ \cdots & \cdots & \cdots & \cdots & \\ +f_mY_1f_1 & +f_mY_2f_2 & +f_mY_3f_3 & \cdots & +f_mY_mf_m \end{pmatrix}$$

$$=2D_1-2D_2$$

さて、上述の D_1 において、第 j 列の列和は、

$$\begin{aligned} & f_jY_jf_j+f_{j+1}Y_{j+1}f_j+f_{j+2}Y_{j+2}f_j+\cdots+f_mY_mf_j \\ &= (f_jY_j+f_{j+1}Y_{j+1}+f_{j+2}Y_{j+2}+\cdots+f_mY_m)f_j \\ &= \left(\sum_{k=1}^{m} f_kY_k-\sum_{k=1}^{j-1} f_kY_k\right)f_j=\left(\bar{Y}-\sum_{k=1}^{j-1} f_kY_k\right)f_j \\ &= \bar{Y}(1-Q_{j-1})f_j \end{aligned}$$

だから、D_1 における要素の総和は、$\displaystyle\sum_{j=1}^{m}\bar{Y}(1-Q_{j-1})f_j=\bar{Y}\left(1-\sum_{j=1}^{m}Q_{j-1}f_j\right)$ である。

他方、D_2 において第 j 行における行和は、

$$f_jY_1f_1+f_jY_2f_2+f_jY_3f_3+\cdots+f_jY_jf_j$$
$$=f_j\left(Y_1f_1+Y_2f_2+Y_3f_3+\cdots+Y_jf_j\right)=f_j\sum_{k=1}^{j}Y_kf_k=\bar{Y}f_jQ_j$$

であるから、D_2 における要素の総和は、$\bar{Y}\sum_{j=1}^{m}Q_jf_j$ である。

したがって、

$$D=2\bar{Y}\left(1-\sum_{j=1}^{m}Q_{j-1}f_j\right)-2\bar{Y}\sum_{j=1}^{m}Q_jf_j=2\bar{Y}\left[1-\sum_{j=1}^{m}(Q_{j-1}+Q_j)f_j\right]$$

になる。

ゆえに、$G=\dfrac{1}{2\bar{Y}}\sum_{j=1}^{m}\sum_{k=1}^{m}|Y_j-Y_k|f_jf_k=1-\sum_{j=1}^{m}(Q_{j-1}+Q_j)f_j$ になるのである。

以上に示したことより、ローレンツ曲線から導くジニ係数 S とデータの差の平均から導く格差指標 G は等しい値になる事が証明された。