Statistical Consulting Report

Statistical Consulting Service
Department of Biostatistics and Statistics
University of Washington

To: Linn Gould
Department: Health Services
From: Nathaniel Derby
Additional Consultants who contributed to this report: Jackie Benedetti, Paul Scheet
Report Date: May 17, 2003
Consulting Dates: April 24 & May 8, 2003
Re: Mathematical Definition of the Gini Index

This report explains the mathematical definition of the Gini as given in Brown (1994).

Derivation of the Gini Index

Suppose we have a sample 5-point data set $\{X_0, X_1, X_2, X_3, X_4\}$ and $\{Y_0, Y_1, Y_2, Y_3, Y_4\}$, as displayed in Figure 1, which is similar to Figure 1 of Brown (1994). Equation (3) of this reference defines the Gini index G as

$$G = 1 - \sum_{i=0}^{k-1} (Y_{i+1} + Y_i)(X_{i+1} - X_i) \tag{1}$$

In our case, $k = 5$. Where does this definition come from?

On page 1247 of Brown (1994), Brown states "Defined graphically, the Gini coefficient formally is measured as the area between the equality curve and the Lorenz curve, divided by the area under the equality curve." Referring to Figure 1, the equality curve is the top curve, and the Lorenz curve is the bottom curve. We can calculate the Gini index using this geometric definition for our sample data set and see if it matches equation (1) above.

Looking at the bottom graph of Figure 1, we have that G is equal to the area A between the curves divided by the area underneath the top line (i.e., the equality curve), which is equal to $\frac{1}{2}$*:

$$G = \frac{A}{1/2} = 2A \quad \Leftrightarrow \quad A = \frac{G}{2} \tag{2}$$

We can compute A by dividing the graph region outside of the in-between curve region into rectangles and triangles, as shown in the bottom graph of Figure 1. The area of the entire graph region is $1 \times 1 = 1$, since X and Y both have a range between 0 and 1. Thus, the area A between the curves is equal to the total graph area (1) minus the areas of the rectangles and triangles shown on the bottom of Figure 1:

*This area is equal to the area of the triangle underneath the upper line in Figure 1, which is $\frac{1}{2}(1 \times 1) = \frac{1}{2}$.

$$A = 1 - \text{Areas of rectangles and triangles from Figure 1}$$

Now, the upper triangle of Figure 1 has an area of $\frac{1}{2}$, so we have this, using the labels of the lower rectangles and triangles from the bottom of Figure 1:

$$A = 1 - \frac{1}{2} - \sum_{i=1}^{4} \text{Area}(T_i) - \sum_{i=1}^{3} \text{Area}(R_i)$$

Now, if we number the triangles from left to right as T_1 through T_4, we get these:

$$\begin{aligned}
\text{Area}(T_1) &= \tfrac{1}{2}(X_1 - X_0)(Y_1 - Y_0) \\
\text{Area}(T_2) &= \tfrac{1}{2}(X_2 - X_1)(Y_2 - Y_1) \\
\text{Area}(T_3) &= \tfrac{1}{2}(X_3 - X_2)(Y_3 - Y_2) \\
\text{Area}(T_4) &= \tfrac{1}{2}(X_4 - X_3)(Y_4 - Y_3)
\end{aligned}$$

Thus, $\displaystyle\sum_{i=1}^{4} \text{Area}(T_i) = \frac{1}{2} \sum_{i=0}^{3} (X_{i+1} - X_i)(Y_{i+1} - Y_i)$. For the rectangles, we have

$$\begin{aligned}
\text{Area}(R_1) &= Y_1(X_2 - X_1) \\
\text{Area}(R_2) &= Y_2(X_3 - X_2) \\
\text{Area}(R_3) &= Y_3(X_4 - X_3)
\end{aligned}$$

So that $\displaystyle\sum_{i=1}^{3} \text{Area}(R_i) = \sum_{i=1}^{3} Y_i(X_{i+1} - X_i) = \sum_{i=0}^{3} Y_i(X_{i+1} - X_i)$ [Since $Y_0 = 0$, we can add a term for $i = 0$ and not change the total sum].

Putting the above all together, we have

$$\begin{aligned}
A &= 1 - \tfrac{1}{2} - \textstyle\sum_{i=1}^{4} \text{Area}(T_i) - \sum_{i=1}^{3} \text{Area}(R_i) \\
&= \tfrac{1}{2} - \tfrac{1}{2} \displaystyle\sum_{i=0}^{3} (X_{i+1} - X_i)(Y_{i+1} - Y_i) - \sum_{i=1}^{3} Y_i(X_{i+1} - X_i) \\
&= \tfrac{1}{2} - \displaystyle\sum_{i=0}^{3} (X_{i+1} - X_i) \left(Y_i + \frac{Y_{i+1} - Y_i}{2} \right) \\
&= \tfrac{1}{2} - \tfrac{1}{2} \displaystyle\sum_{i=0}^{3} (X_{i+1} - X_i)(Y_i + Y_{i+1}) \\
&= \tfrac{1}{2} \left(1 - \displaystyle\sum_{i=0}^{k-1} (Y_{i+1} + Y_i)(X_{i+1} - X_i) \right)
\end{aligned}$$

Since we showed in equation (2) that $A = \frac{G}{2}$, we must have that $G = 1 - \displaystyle\sum_{i=0}^{k-1} (Y_{i+1} + Y_i)(X_{i+1} - X_i)$, and equation (1) is proven.

A Couple Caveats

In the derivation of the Gini index, we assumed that the Lorenz curve was below the equality line (i.e., the line $Y_i = X_i$). That is, we assumed that for each point (X_i, Y_i), that $Y_i \leq X_i$.

If this is not the case, the Gini index will no longer be the area between the two curves divided by $\frac{1}{2}$. Consider the two cases shown in Figure 2. On the top, we have that for each point, $Y_i \geq X_i$. On the bottom, we have that $Y_i \leq X_i$ for some points, and $Y_i \geq X_i$ for others. In the first case (i.e., $Y_i \geq X_i$ for each point), the Gini index is equal to the negative of the area between the two curves. In the second case, the Gini index is equal to the area below the equality curve minus the area above the equality curve.

Thus, when computing the Gini index, in order to make sure you really are computing the area between the two curves, divided by the area underneath the equality curve, it's a good idea to ascertain that $Y_i \leq X_i$ for each point. Checking into this for the document `SJC gini decibles.xls`, you are computing the Gini coefficient with $Y =$ `Cum % Inc` (column K) and $X =$ `Deciles` (column L). A glance at the data verifies that, indeed, $Y_i \leq X_i$ for each point.

One more point: The formula for the Gini index is not correct if either X or Y has a range other than between 0 and 1. That is, beware if you ever have a data set where it is NOT the case where $0 \leq X_i \leq 1, 0 \leq Y_i \leq 1$ for each point. This is because when we divide by the area underneath the equality curve, we would be dividing by a number other than $\frac{1}{2}$. However, this is not the case with your data, Linn.

References

Brown, Malcolm C. (1994), Using Gini-style indices to evaluate the spatial patterns of health practitioners: Theoretical considerations and an application based on Alberta data, *Social Science Medicine* 38:9, 1243-1256.

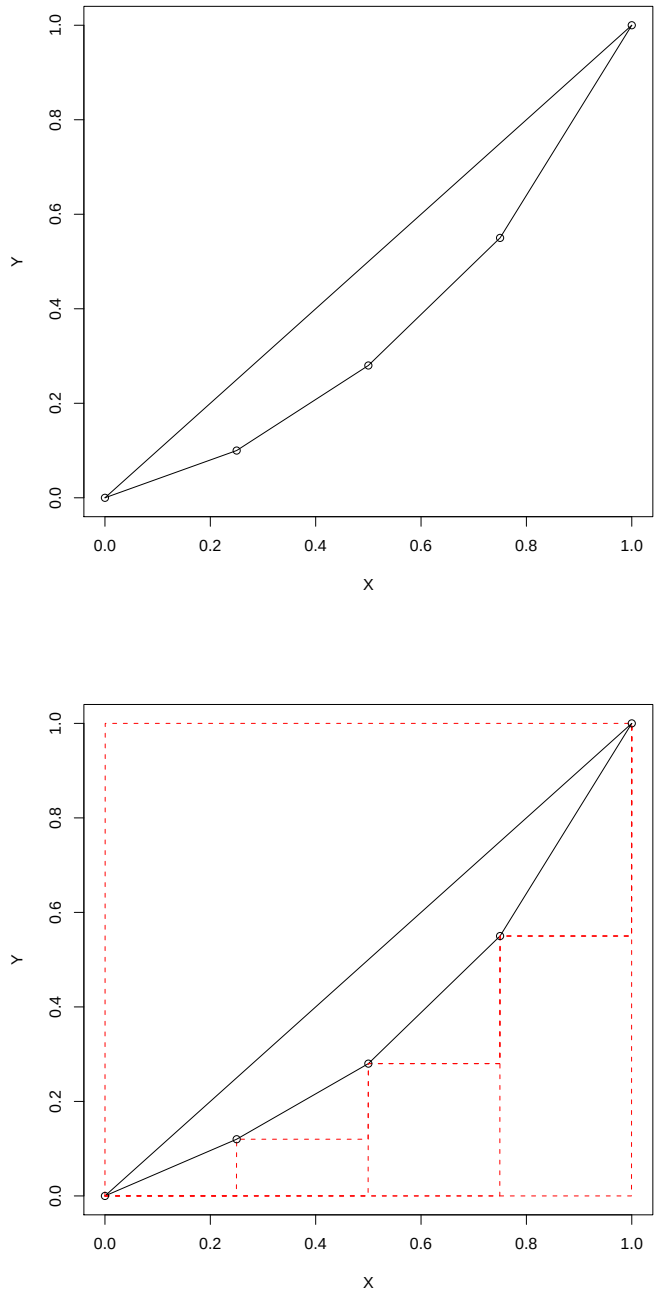

Figure 1: For sample 5-point data $\{X_i\}$ and $\{Y_i\}$, the equality curve (top curve) and the Lorenz curve (bottom curve), similar to Figure 1 in Brown (1994). The bottom plot divides the area outside of the between-curve region into triangles and rectangles.

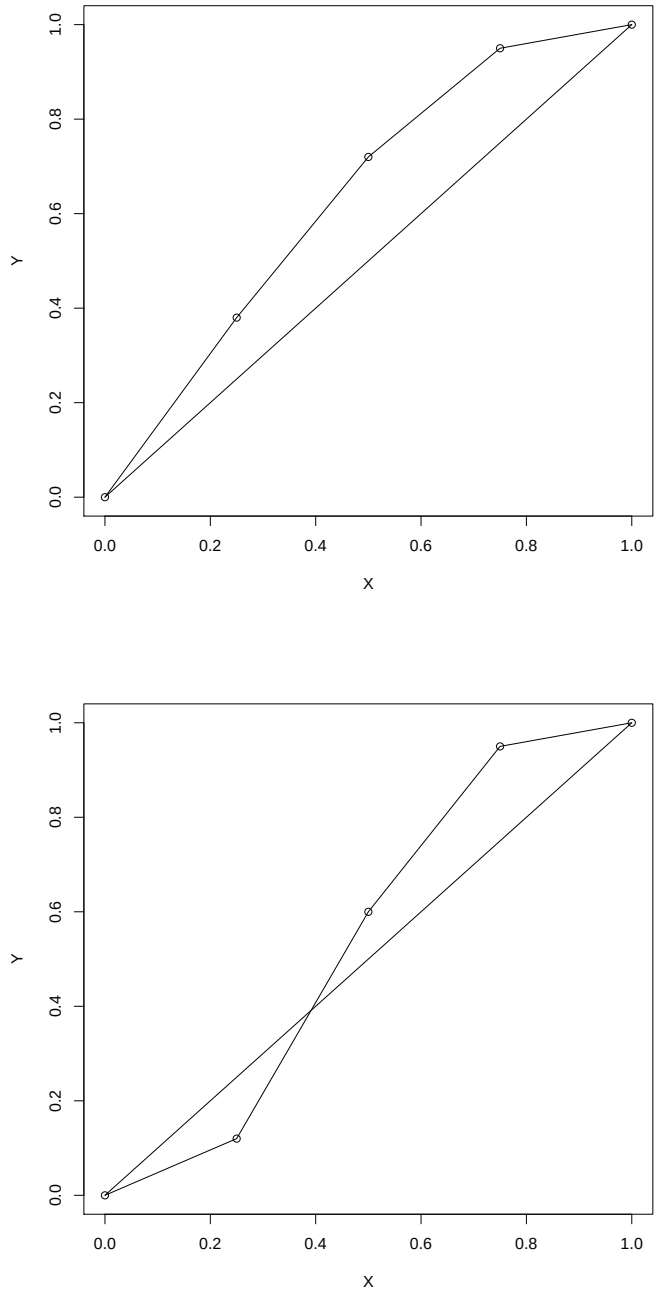

Figure 2: Curves similar to those shown in Figure 1, except that $Y_i \leq X_i$ for each point (X_i, Y_i) (above), or $Y_i \leq X_i$ for some points and $Y_i \geq X_i$ for others (below).