SAHUARUS
Revista Electrónica de Matemáticas,
Departamento de Matemáticas, Universidad de Sonora,
Vol. 4, No. 1. Septiembre, 2019, pp. 19–33.

Estimación del Coeficiente de Gini utilizando Distribuciones Tipo Fase

Luz Judith R. Esparza

Cátedra Conacyt- Universidad Autónoma Chapingo
e-mail: judithr19@gmail.com

Resumen

En este trabajo, se considera una nueva metodología para estimar el coeficiente de Gini utilizando las distribuciones tipo fase, en particular, utilizando las distribuciones de momento tipo fase. Para estimar dicho coeficiente, primero se obtienen los estimadores de máxima verosimilitud de las distribuciones tipo fase utilizando el algoritmo EM, para así después utilizar las distribuciones de momentos y obtener el coeficiente de Gini. Ilustramos la eficiencia del método propuesto calculando el coeficiente de Gini de México considerando tres años: 1995, 2000 y 2005.

1 Introducción

Es bien sabido que los ingresos son uno de los principales factores para tener mejores oportunidades en la vida. Si un país tiene una distribución inequitativa de sus ingresos, esto genera sociedades diferenciadas que afectan su desarrollo.

En economía, un tema importante es cómo medir el progreso humano a través de indicadores, como el Producto Interno Bruto (PIB), que evalúa la afluencia económica en términos de ingresos. Existen otros índices que de hecho usan el PIB, como el índice de bienestar (ver [12, 13]). El índice de Desarrollo Humano (IDH) ha sido uno de los más populares en este campo. Además, el PIB per cápita se usa comúnmente como indicador, ya que sirve para medir el progreso de un país en función de los ingresos. También permite una comparación y una evaluación del desempeño de las autoridades y los agentes económicos.

El ingreso de un país representa la llamada relación de intercambio, y también representa una unidad con el fin de adquirir todo tipo de bienes y servicios que contribuyan al bienestar de las personas. De ahí la importancia de estudiar su distribución.

Una de las consecuencias de una distribución desigual en un país es la falta de recursos para continuar el proceso económico, por lo tanto, una retribución inequitativa limita los recursos para las personas, empobreciéndolas en diversos niveles, como alimentos, sus capacidades y pobreza patrimonial. No obstante, un porcentaje importante de personas tiene necesidades que no solo exhiben el carácter inhumano de la visión económica, sino que también muestran la ineficiencia al llevarse a millones de consumidores potenciales.

En las últimas décadas, ha sido interesante no solo conocer los ingresos, sino también la forma de su distribución. La brecha entre un país rico y un país pobre se encuentra en su nivel más alto en la mayoría de los países de la OCDE (Organización para la Cooperación

y el Desarrollo Económico) en 30 años. En estos días, el 10% de la población más rica en el área de la OCDE gana 9.5 veces más que el 10% de los más pobres.

Un índice de desigualdad es una medida que resume la forma en que una variable se distribuye entre entidades o individuos. En el caso particular de la desigualdad económica, la medición está asociada a los ingresos. Un indicador muy utilizado para estudiar la desigualdad es el índice de Gini ([9]). Este índice es una medida que representa cómo se distribuye el ingreso en una población, y toma valores de 0 a 1. Un valor aproximado a 1 indica que hay una mala distribución del ingreso, mientras que un valor cercano a 0 indica que hay una buena distribución del ingreso.

Hasta ahora, el índice de Gini no se ha estimado utilizando ninguna clase de distribución estadística. Un método muy utilizado es usar deciles. En este artículo, proponemos una metodología muy útil y eficiente para estimar este índice, utilizando las distribuciones tipo fase ([10, 11]), en particular las distribuciones de momento tipo fase ([4]).

Este artículo está organizado de la siguiente forma. En la sección 2 se presenta un breve resumen de la curva de Lorenz y el índice de Gini. Dado que la estimación de este índice será a través de distribuciones tipo fase, la teoría detrás de esta clase de distribuciones se presenta en la sección 3, incluyendo la teoría de la estimación. Una aplicación utilizando datos reales se presenta en la sección 4. Finalmente, algunas observaciones se muestran en la sección 5.

2 Antecedentes

En esta sección discutiremos sobre la distribución del ingreso, la curva de Lorenz y el índice de Gini, sus ventajas y desventajas. Para tener más detalles sobre estos temas, recomendamos al lector revise [7, 8].

2.1 Distribución del ingreso

La distribución del ingreso se refiere a la forma en que la riqueza generada en una región o un país se distribuye entre diferentes segmentos de la población, en un período determinado. Si bien la pobreza se mide en términos absolutos (cuantificación), la distribución del ingreso sí lo hace en términos relativos. Por lo tanto, la distribución del ingreso nos permite ubicar las condiciones de desigualdad y los niveles de concentración que presenta una sociedad.

La distribución del ingreso refleja cómo es un segmento de las familias (clasificado generalmente por deciles), con respecto a otro segmento de la población, dependiendo de sus niveles de ingresos; es decir, su participación en el ingreso nacional generado.

El análisis de la distribución del ingreso se realiza mediante la aplicación de algunas herramientas estadísticas que nos permiten diferenciar los ingresos de la población en función de una clasificación en orden ascendente de los ingresos obtenidos por las familias.

2.2 Curva de Lorenz y coeficiente de Gini

La curva de Lorenz es la representación gráfica de la distribución de la riqueza desarrollada por Max Lorenz en 1905 ([8]). En la Figura 1, la recta a 45 grados, representa la igualdad perfecta de la distribución de la riqueza; la curva de Lorenz se encuentra debajo de ella, mostrando la realidad de la distribución de la riqueza. La diferencia entre la línea recta y la línea curva es la cantidad de desigualdad de la distribución de la riqueza, una cifra descrita por el coeficiente de Gini. Matemáticamente, el coeficiente de Gini (G) se obtiene de la curva de Lorenz $L(\cdot)$ como sigue

$$G = 1 - 2\int_0^1 L(u)du.$$

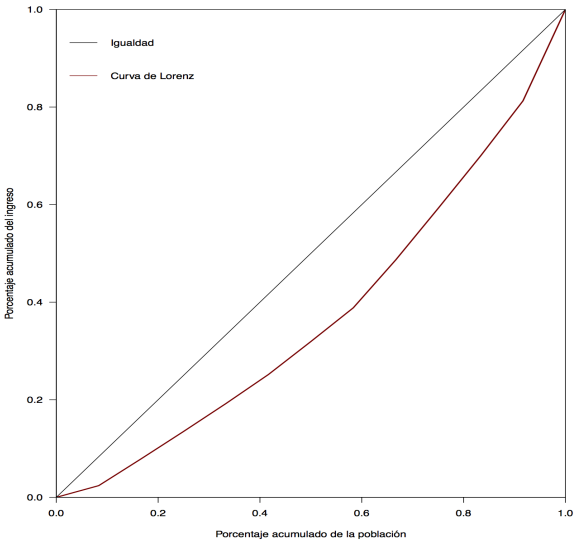


Figura 1: Ejemplo de la curva de Lorenz

La curva de Lorenz se puede usar para mostrar qué porcentaje de los residentes de una nación posee qué porcentaje de la riqueza de esa nación. El coeficiente de Gini se calcula en función de la discrepancia entre la línea diagonal y la curva de Lorenz, dividiendo esa cifra entre el total de la riqueza que se mantiene dentro de un país en particular. Esto permite comparar varias economías cuando se examina la distribución de la riqueza entre las naciones individuales. El coeficiente de Gini puede estar en cualquier lugar desde 0, representando igualdad completa (es decir, todos tienen el mismo ingreso), hasta 1, representando la desigualdad completa (es decir, una persona tiene todos los ingresos, mientras que todos los demás tienen ingresos 0).

La curva de Lorenz y el coeficiente de Gini son populares en economía porque permiten una riqueza negativa. Si cierta porción de la población tiene riqueza negativa, la curva de Lorenz puede moverse por debajo del eje x.

La principal ventaja del índice de Gini es que es una medida de la desigualdad que se obtiene a través de un análisis de razones. Se puede usar para comparar distribuciones de ingresos entre diferentes sectores de la población y países. De hecho, el coeficiente de Gini puede utilizarse para indicar cómo ha cambiado la distribución del ingreso dentro de un país durante un período de tiempo, por lo que es posible ver si la desigualdad está aumentando o disminuyendo. Este índice no considera el tamaño de la economía, la forma en que se mide, o si es un país rico o pobre en promedio.

Por otro lado, el coeficiente de Gini de un país grande y económicamente diverso, generalmente resulta en un coeficiente mucho más alto que el que cada una de sus regiones tiene individualmente. También se debe tomar en cuenta que la comparación de la distribución del ingreso entre los países puede ser difícil porque los sistemas de beneficios pueden diferir. De hecho, la medida dará resultados diferentes cuando se aplica a individuos en lugar de hogares.

En cuanto a todas las estadísticas, habrá errores sistemáticos y aleatorios en los datos. El significado del coeficiente de Gini disminuye a medida que los datos se vuelven menos precisos. Además, los países pueden recopilar datos de manera diferente, lo que dificulta la comparación estadística entre países. Las economías con ingresos similares y coeficientes de Gini pueden tener muy diferentes distribuciones de ingresos, esto se debe a que las curvas de Lorenz pueden tener diferentes formas y aún así producir el mismo coeficiente de Gini.

A pesar de todas estas desventajas como medida de la desigualdad, el coeficiente de Gini es, por mucho, uno de los índices más importantes para medir el ingreso en todos los países. Por esta razón, decidimos estudiar este importante concepto utilizando una clase relativamente nueva de distribuciones para estimar el índice de Gini: las distribuciones tipo fase ([11]). En la siguiente sección presentamos estas distribuciones en su forma matemática, también se incluye una sección de la estimación de los parámetros de estas distribuciones utilizando el algoritmo Esperanza-Maximización (EM) ([1]).

3 Distribuciones tipo fase

Sea $\{X(t)\}_{t\geq 0}$ un proceso de saltos de Markov (MJP, por sus siglas en inglés) con espacio de estados finito $E = \{1, 2, \ldots, m, m+1\}$ donde $\{1, 2, \ldots, m\}$ son estados transitorios y $\{m+1\}$ es un estado absorbente. Entonces, la matriz de intensidad tiene la forma

$$\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{T} & \boldsymbol{t} \\ \mathbf{0} & 0 \end{pmatrix}$$

donde $\boldsymbol{T}$ es una matriz de dimensión $m \times m$ tal que $t_{ii} < 0$ y $t_{ij} \geq 0$ para $i \neq j$; $\boldsymbol{t}$ es un vector columna de dimensión $m \times 1$. Ya que los renglones deben sumar 0, entonces $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{e}$, donde $\boldsymbol{e}$ es un vector de 1's.

Sea $\alpha_i = \mathbb{P}(X(0) = i)$ y $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ denota las probabilidades iniciales del MJP. Así, tenemos la siguiente definición (ver [10, 11]).

Definición 3.1 *El tiempo de absorción de la cadena de Markov*

$$\tau = \inf\{t \geq 0 : X(t) = m+1\}$$

tiene una distribución tipo fase (PH, por sus siglas en inglés), y escribimos $\tau \sim PH(\boldsymbol{\alpha}, \boldsymbol{T})$.

Características:

- La densidad de τ está dada por: $f(s) = \boldsymbol{\alpha} \exp(\boldsymbol{T}s)\boldsymbol{t}$.
- La distribución acumulada de τ es: $F(s) = 1 - \boldsymbol{\alpha} \exp(\boldsymbol{T}s)\boldsymbol{e}$.
- El n-ésimo momento de τ está dado por: $\mathbb{E}(\tau^n) = (-1)^n n! \boldsymbol{\alpha} \boldsymbol{T}^{-n} \boldsymbol{e}$.
- La función generadora de momentos de τ está dada por: $\mathbb{E}(e^{s\tau}) = \boldsymbol{\alpha}(-s\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{t}$.

Ver [2] para la demostración de estas propiedades.

Algunos ejemplos de las distribuciones PH son los siguientes.

1. **Distribución Exponencial:** Sea $X \sim \exp(\lambda)$ con $\lambda > 0$, entonces, la densidad de X es $f(x) = \lambda e^{-\lambda x}$, y una representación PH está dada por:

$$\boldsymbol{\alpha} = [1]; \quad \boldsymbol{T} = [-\lambda]; \quad \boldsymbol{t} = [\lambda].$$

2. **Distribución Erlang:** Supongamos $Z = \sum_{i=1}^{m} X_i$ con $X_i \sim \exp(\lambda_i)$, una representación PH está dada por:

$$\boldsymbol{\alpha} = (1, 0, \ldots, 0), \quad \boldsymbol{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \ldots & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -\lambda_{m-1} & \lambda_{m-1} \\ 0 & 0 & 0 & \ldots & 0 & -\lambda_m \end{pmatrix}.$$

3. **Distribución Hiper-exponencial:** (Mezcla de exponenciales).

   Sean m variables aleatorias $X_i \sim \exp(\lambda_i)$ y supongamos que Z tiene el valor de X_i con probabilidad p_i. La distribución de Z, se llama hiper-exponencial, y una representación PH está dada por:

$$\boldsymbol{\alpha} = (p_1, \ldots, p_m), \quad \boldsymbol{T} = \begin{pmatrix} -\lambda_1 & 0 & \ldots & 0 \\ 0 & -\lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & -\lambda_m \end{pmatrix}.$$

   Sin pérdida de generalidad se puede considerar que $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_m$.

4. **Distribución Coxian:** Esta distribución tiene la siguiente representación PH:

$$\boldsymbol{\alpha} = (1, 0 \ldots, 0), \quad \mathbf{T} = \begin{pmatrix} -\lambda_1 & t_{12} & 0 & \ldots & 0 \\ 0 & -\lambda_2 & t_{23} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -\lambda_m \end{pmatrix}, \quad \boldsymbol{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{pmatrix}$$

donde $\lambda_i = t_i + t_{i,i+1}$ para $i = 1, \ldots, m-1$, y $\lambda_m = t_m$.

Si la distribución inicial es $\boldsymbol{\alpha} = (p_1, \ldots, p_m)$, entonces tenemos la distribución General Coxian.

A continuación, presentamos una clase de distribuciones PH, llamadas distribuciones de momentos tipo fase, a partir de las cuales se podrá obtener la curva de Lorenz y el índice de Gini.

3.1 Distribuciones de momentos tipo fase

Las distribuciones de momentos PH fueron consideradas en un inicio por [4], y se han extendido al caso discreto en [6].

Consideremos la función de densidad f de una variable aleatoria (v.a) no negativa X, entonces

$$f_i(x) = \frac{x^i f(x)}{\mu_i}, \quad \text{where} \quad \mu_i = \int_0^\infty x^i f(x) dx$$

son densidades de v.a's no negativas $X^{(i)}$ siempre y cuando μ_i existe. Decimos que f_i es la densidad de la distribución del i-ésimo momento de f.

En [4] encontramos el siguiente resultado que establece que si f es PH, entonces la distribución del primer momento f_1 también es PH.

Teorema 3.1 *Considere una distribución PH con representación $(\boldsymbol{\alpha}, \boldsymbol{T})$. Entonces, la distribución del primer momento es tipo fase. Una representación es $PH(\boldsymbol{\alpha}_1, \boldsymbol{T}_1)$, donde*

$$\boldsymbol{\alpha}_1 = (\boldsymbol{t}' \Delta(\boldsymbol{m}_2), \mathbf{0})$$

y

$$\boldsymbol{T}_1 = \begin{pmatrix} \Delta^{-1}(\boldsymbol{m}_2)\boldsymbol{T}'\Delta(\boldsymbol{m}_2) & \rho_1^{-1}\Delta^{-1}(\boldsymbol{m}_2)\Delta(\boldsymbol{m}_1) \\ \mathbf{0} & \Delta^{-1}(\boldsymbol{m}_1)\boldsymbol{T}'\Delta(\boldsymbol{m}_1) \end{pmatrix}$$

con $\rho_i = \boldsymbol{\alpha}(-\boldsymbol{T})^{-i}\boldsymbol{e}$ y $\boldsymbol{m}_i = \rho_{i-1}^{-1}\boldsymbol{\alpha}(-\boldsymbol{T})^{-i}$.

Aquí $\Delta(\cdot)$ denota la matriz diagonal.

Además, si F es una función de distribución y F_1 la función de distribución de la correspondiente distribución de momento de primer orden, luego la curva paramétrica $\gamma : t \to (F(t), F_1(t))$ para $t \in [0, \infty)$ se llama curva de Lorenz. Como se mencionó en las secciones

anteriores, teniendo esta curva, podemos obtener el índice de Gini. Cuanto mayor es el índice de Gini, mayor es la desigualdad de ingresos. Si γ es la línea recta $y = x$, entonces habría igualdad completa con el índice de Gini siendo 0.

Si F es tipo fase, [4] proporciona fórmulas explícitas para obtener tanto para la curva de Lorenz como para el índice de Gini.

Teorema 3.2 *Sea F la función de distribución de una variable aleatoria distribuida PH con la representación $(\boldsymbol{\alpha}, \boldsymbol{T})$. Entonces la curva de Lorenz está dada por la fórmula:*

$$\gamma : t \to \left(1 - \boldsymbol{\alpha} e^{\boldsymbol{T}t}\boldsymbol{e}, 1 - \frac{\boldsymbol{\alpha}\boldsymbol{T}^{-1}}{\boldsymbol{\alpha}\boldsymbol{T}^{-1}\boldsymbol{e}}(e^{\boldsymbol{T}t}\boldsymbol{e} + te^{\boldsymbol{T}t}\boldsymbol{t})\right) \tag{1}$$

y el índice de Gini G está dado por:

$$G = 2(\boldsymbol{\alpha} \otimes \boldsymbol{\alpha}_1)(-(\boldsymbol{T} \oplus \boldsymbol{T}_1))^{-1}(\boldsymbol{t} \otimes \boldsymbol{e}) - 1. \tag{2}$$

Aquí $\otimes$ y $\oplus$ denotan el producto y la suma de Kronequer, respectivamente.

Después de definir las distribuciones PH, y especialmente las distribuciones de momentos PH, que son piezas clave para obtener la curva de Lorenz y el índice de Gini, estamos listos para presentar otra parte muy importante: la estimación.

Teniendo como dato el PIB per cápita y utilizando la teoría de estimación que a continuación se presentará, se podrá estimar el coeficiente de Gini utilizando la ecuación (2).

3.2 Estimación de las distribuciones tipo fase

En esta sección consideraremos el algoritmo de Esperanza-maximización (EM) para estimar las distribuciones PH.

Un problema muy importante en estadística es la maximización, donde encontramos las estimaciones de máxima verosimilitud. El algoritmo EM fue introducido por Dempster en 1977 ([5]) para resolver un problema de maximización que tenía la función de verosimilitud. Una de las principales ventajas de este algoritmo es que se trata de un algoritmo iterativo para resolver problemas de máxima verosimilitud cuando no se observan "algunas" variables aleatorias (información incompleta). La idea general de este algoritmo es la siguiente:

1. Sustituir los datos faltantes por valores estimados.
2. Estimar los parámetros considerando la información completa con los valores estimados.
3. Repetir los pasos anteriores hasta la convergencia.

Supongamos que tenemos el vector aleatorio Y con densidad $f(Y;\theta)$ donde $\theta \in \Theta \subset \mathbb{R}^m$. El algoritmo EM propone una solución para maximizar $\ell(\theta; Y)$ (log-verosimilitud correspondiente a los datos completos) de forma iterativa reemplazando los datos faltantes con el

valor esperado de los datos observados (Y_{obs}). Esta esperanza es calculada con respecto a la distribución de los datos completos evaluados en el valor real del parámetro θ, es decir, si $\theta^{(0)}$ es el valor inicial de θ, luego la primera iteración debe calcularse de la siguiente manera:

$$Q(\theta, \theta^{(0)}) = \mathbb{E}_{\theta^{(0)}}[\ell(\theta; Y)|Y_{obs}]$$

entonces $Q(\theta, \theta^{(0)})$ se maximiza como una función de θ, i.e., encontramos $\theta^{(1)}$ tal que

$$Q(\theta, \theta^{(0)}) \leq Q(\theta, \theta^{(1)})$$

para todo $\theta \in \Theta$.

Ahora, estamos listos para presentar el algoritmo formalmente.

Algoritmo 1 Algoritmo EM

Este algoritmo consiste básicamente en dos pasos, una vez iniciado el parámetro $\theta^{(0)}$ y haciendo $n = 1$, entonces:

1: Paso E: Calcular

$$Q(\theta, \theta^{(n)}) = \mathbb{E}_{\theta^{(n)}}[\ell(\theta; Y)|Y_{obs}]$$

2: Paso M: Encontrar $Q(\theta, \theta^{(n+1)})$, tal que

$$Q(\theta, \theta^{(n)}) \leq Q(\theta, \theta^{(n+1)})$$

para todo $\theta \in \Theta$.

Para la estimación de las distribuciones PH que utilizan el algoritmo EM, recomendamos al lector revise [1, 3]. A continuación, presentamos un breve resumen.

3.2.1 Algoritmo EM para las distribuciones PH

Sea $y_1, \ldots, y_M$ una realización de M variables aleatorias *iid* de $PH(\boldsymbol{\alpha}, \boldsymbol{T})$. Estamos en el caso de datos faltantes, ya que solo tenemos los tiempos de absorción, no la trayectoria completa. Sea $\mathbf{y} = (y_1, \ldots, y_M)$ y $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{T}, \mathbf{t})$, donde $\mathbf{t} = -\boldsymbol{T}\mathbf{e}$. La función de verosimilitud de los datos incompletos está dada por:

$$L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{k=1}^{M} \boldsymbol{\alpha} e^{\boldsymbol{T} y_k} \mathbf{t}, \tag{3}$$

entonces, la log-verosimilitud es $\ell(\boldsymbol{\theta}; \mathbf{y}) = \sum_{k=1}^{M} \log f(y_k)$, donde $f(y_k) = \boldsymbol{\alpha} e^{\boldsymbol{T} y_k} \mathbf{t}$.

Para encontrar los estimadores de máxima verosimilitud de $\boldsymbol{\theta}$, con base en los datos observados, consideramos $\mathbf{x} = \{\mathbf{x}_i\}_{i=1,\ldots,M}$, que denotan los datos completos de los M tiempos de absorción, por lo tanto, las $\mathbf{x}_i$ son las trayectorias de los MJP's.

La función de verosimilitud de los datos completos está dada por:

$$L_f(\boldsymbol{\theta};\mathbf{x}) = \prod_{i=1}^{m} \alpha_i^{B_i} \prod_{i=1}^{m}\prod_{j\neq i}^{m} t_{ij}^{N_{ij}} e^{-t_{ij}Z_i} \prod_{i=1}^{m} t_i^{N_i} e^{-t_i Z_i}, \tag{4}$$

donde B_i es el número de procesos que empiezan en el estado i, N_i es el número de procesos que saltan al estado absorbente a partir del estado i, N_{ij} es el número de saltos del estado i al estado j dentro de todos los procesos, y Z_i es el tiempo total que los procesos estuvieron en el estado i antes de la absorción.

Ya que los datos $\mathbf{y} = (y_1, \ldots, y_M)$ son incompletos, utilizaremos el algoritmo EM (ver [1]). La función de log-verosimilitud de los datos completos viene dada por:

$$\begin{aligned} \ell_f(\boldsymbol{\theta};\mathbf{x}) &= \sum_{i=1}^{m} B_i \log(\alpha_i) + \sum_{i=1}^{m}\sum_{j\neq i}^{m} N_{ij}\log(t_{ij}) \\ &\quad - \sum_{i=1}^{m}\sum_{j\neq i}^{m} t_{ij}Z_i + \sum_{i=1}^{m} N_i \log(t_i) - \sum_{i=1}^{m} t_i Z_i. \end{aligned} \tag{5}$$

Usando los multiplicadores de Lagrange, encontramos que los estimadores de máxima verosimilitud de los parámetros están dados por:

$$\hat{\alpha}_i = \frac{B_i}{M}; \quad \hat{t}_{ij} = \frac{N_{ij}}{Z_i}, \quad \hat{t}_i = \frac{N_i}{Z_i}.$$

Sean $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0, \boldsymbol{T}_0, \mathbf{t}_0)$ los parámetros iniciales, por lo tanto, el algoritmo EM sería el siguiente:

1. Paso E: Calcular la función
$$h : \boldsymbol{\theta} \to \mathbb{E}_{\boldsymbol{\theta}_0}(\ell_f(\boldsymbol{\theta};\mathbf{x})|\mathbf{Y} = \mathbf{y}).$$
2. Paso M:
$$\boldsymbol{\theta}_0 = \text{argmax}_{\boldsymbol{\theta}} h(\boldsymbol{\theta}).$$
3. Ir a (1).

Como (5) es una función lineal con respecto a las estadísticas suficientes B_i, Z_i, N_i y N_{ij}, entonces solo tenemos que calcular las esperanzas condicionales de estas estadísticas. Sean B_i^k, Z_i^k, N_i^k, y N_{ij}^k las correspondientes estadísticas de la k-ésima observación, entonces

$$B_i = \sum_{k=1}^{M} B_i^k, \quad Z_i = \sum_{k=1}^{M} Z_i^k, \quad N_i = \sum_{k=1}^{M} N_i^k, \quad N_{ij} = \sum_{k=1}^{M} N_{ij}^k,$$

para $i, j = 1, \ldots, m$, $i \neq j$, entonces $\mathbb{E}_{\boldsymbol{\theta}}(S|\mathbf{Y} = \mathbf{y}) = \sum_{k=1}^{M} \mathbb{E}_{\boldsymbol{\theta}}(S^k|Y_k = y_k)$, donde $S \in \{B_i, Z_i, N_i, N_{ij}\}$. Así, la principal tarea es calcular $\mathbb{E}_{\boldsymbol{\theta}}(S^k|Y_k = y_k)$.

En [1] tenemos el siguiente resultado.

Teorema 3.3 *Para $i, j = 1, \ldots, m$, $i \neq j$, entonces*

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}(B_i^k|Y_k = y_k) &= \frac{\alpha_i \boldsymbol{e}_i' \exp(\boldsymbol{T}y_k)\boldsymbol{t}}{\boldsymbol{\pi}\exp(\boldsymbol{T}y_k)\boldsymbol{t}} \\
\mathbb{E}_{\boldsymbol{\theta}}(Z_i^k|Y_k = y_k) &= \frac{\int_0^{y_k} \boldsymbol{\alpha}\exp(\boldsymbol{T}u)\boldsymbol{e}_i\boldsymbol{e}_i'\exp(\boldsymbol{T}(y_k-u))\boldsymbol{t}du}{\boldsymbol{\alpha}\exp(\boldsymbol{T}y_k)\boldsymbol{t}} \\
\mathbb{E}_{\boldsymbol{\theta}}(N_i^k|Y_k = y_k) &= \frac{t_i\boldsymbol{\alpha}\exp(\boldsymbol{T}y_k)\boldsymbol{e}_i}{\boldsymbol{\alpha}\exp(\boldsymbol{T}y_k)\boldsymbol{t}} \\
\mathbb{E}_{\boldsymbol{\theta}}(N_{ij}^k|Y_k = y_k) &= \frac{t_{ij}\int_0^{y_k} \boldsymbol{\alpha}\exp(\boldsymbol{T}u)\boldsymbol{e}_i\boldsymbol{e}_j'\exp(\boldsymbol{T}(y_k-u))\boldsymbol{t}du}{\boldsymbol{\alpha}\exp(\boldsymbol{T}y_k)\boldsymbol{t}}.
\end{aligned}$$

Teniendo este teorema, estamos listos para utilizar el algoritmo EM para estimar los parámetros de la distribución PH.

Finalmente, una estimación de la curva de Lorenz y el índice de Gini se obtiene a través del siguiente algoritmo.

Algoritmo 2 Metodología para estimar el índice de Gini y curva de Lorenz.

1: **Recolección de los datos:** PIB per cápita.
2: **Estimación:** Considerando los datos, encontrar los estimadores de máxima verosimilitud $(\boldsymbol{\alpha}, \boldsymbol{T})$ de la distribución PH que mejor ajuste a los datos utilizando el algoritmo EM (ver [1]) .
3: **Cálculo de la curva de Lorenz y el índice de Gini.** Utilizando los estimadores de máxima verosimilitud, calcular las ecuaciones presentadas en el Teorema 3.1 para obtener los parámetros del primer momento. Esos serán usados en el Teorema 3.2 para estimar la curva de Lorenz y el índice de Gini (ver ecuaciones (1) y (2)).

4 Aplicación

En esta sección, como una aplicación de la metodología presentada, consideraremos los datos reales del PIB per cápita de México (basado en 1993) y la población de este país para los años: 1995, 2000 y 2005 (página de inicio: http: //www.chapingo.mx/dicifo/demyc/idh/new/). Esta información se obtuvo a nivel estatal, por lo que estos datos se consideran como tiempos de absorción para la parte de estimación.

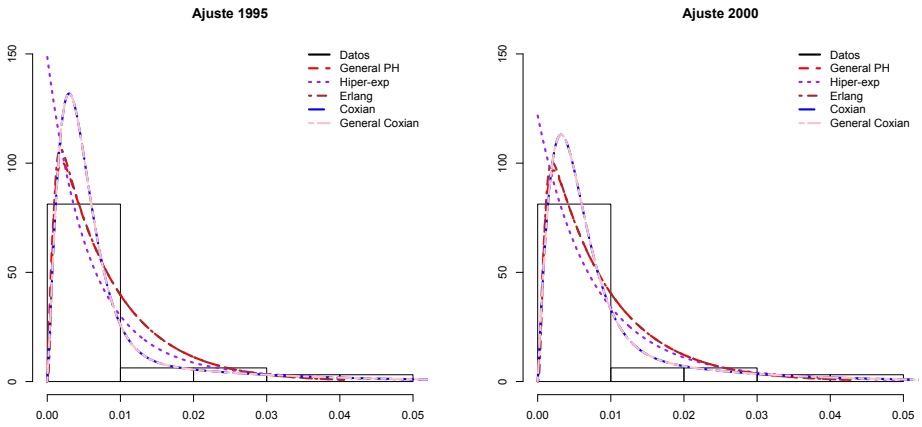
De acuerdo a [1], las distribuciones tipo fase propuestas para el ajuste de los datos son: General PH, Hiper-exponencial, Erlang, Coxian y General Coxian. Así pues, veremos cuál distribución y qué dimensión ajustan mejor a los datos.

En primer lugar, se consideran las dimensiones 4 y 5 para obtener la log-verosimilitud (LL). Ver Tabla 1.

Tabla 1:
Log-verosimilitud (LL) considerando las distribuciones General PH, Hiper-exp., Erlang, Coxian y General Coxian.

Año	***Distribución***	***LL: dim=4***	***LL: dim=5***
1995	General PH	121.224584	121.524502
	Hiper-exp.	120.452397	120.452397
	Erlang	121.803215	121.916592
	Coxian	124.591064	124.688606
	General Coxian	124.591011	125.099822
2000	General PH	119.441245	119.742096
	Hiper-exp.	117.978967	117.978967
	Erlang	120.011986	120.134267
	Coxian	121.207841	121.318630
	General Coxian	121.207827	121.623843
2005	General PH	122.667385	122.790271
	Hiper-exp.	120.684049	120.684024
	Erlang	122.998055	123.088883
	Coxian	123.694720	123.778368
	General Coxian	123.694684	124.018809

Respecto a la log-verosimilitud, la distribución hiper-exponencial tuvo los peores resultados, las otras distribuciones tuvieron resultados similares entre ellas, de hecho, las diferencias en sus verosimilitudes son pequeñas. Corroboramos qué distribución ajustó mejor los datos al presentar los gráficos (ver Figuras 2 y 3) para los años 1995, 2000 y 2005, considerando todas las distribuciones y las dimensiones 4 y 5.



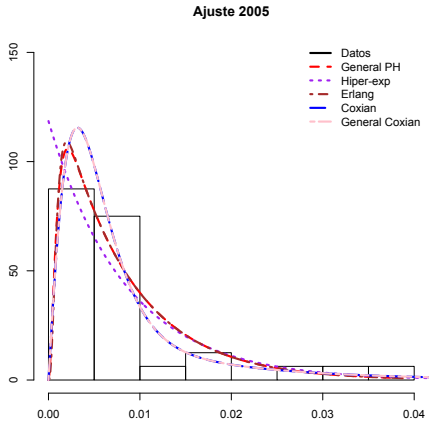


Figura 2: Ajuste considerando dimensión 4

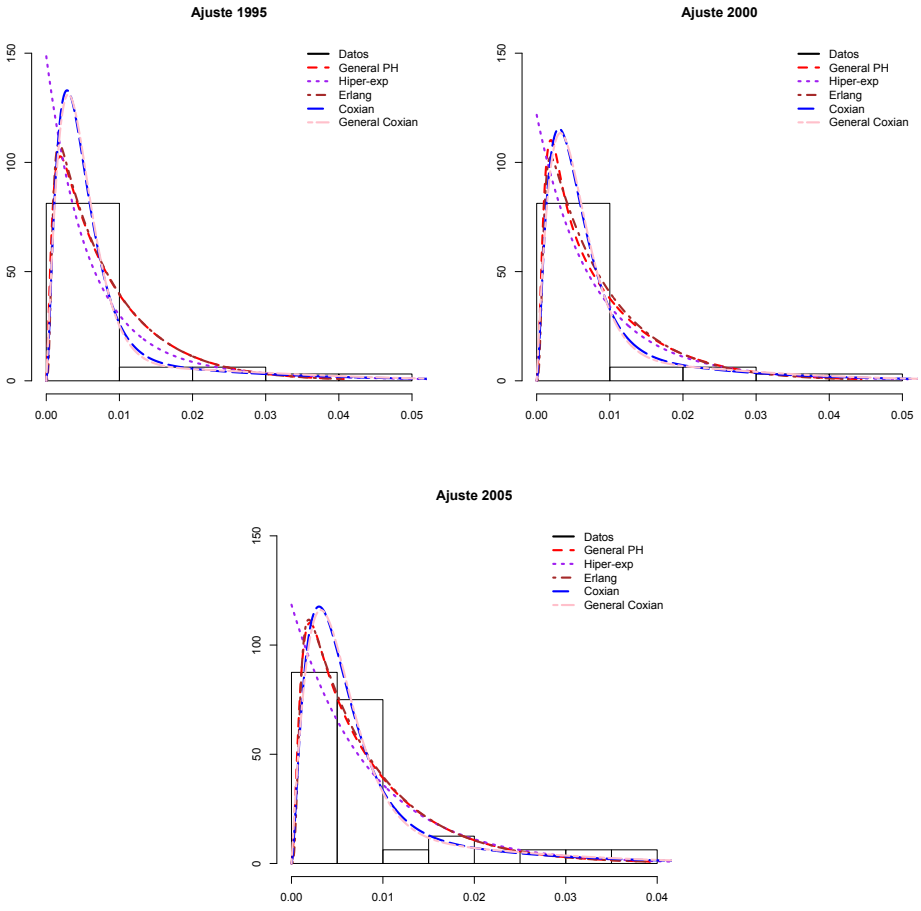


Figura 3: Ajuste considerando dimensión 5

Como podemos ver en las Figuras 2 y 3, el General PH ajustó mejor los datos para todos los años, por lo que consideraremos esta distribución y la dimensión 4 para obtener la curva de Lorenz y el índice de Gini.

Una vez que encontramos los estimadores de máxima verosimilitud PH para cada año, encontramos los parámetros de primer momento (ver Teorema 3.1) y los sustituimos por ecuaciones (1) y (2). Los resultados se muestran en las Figuras 4, 5 y 6.

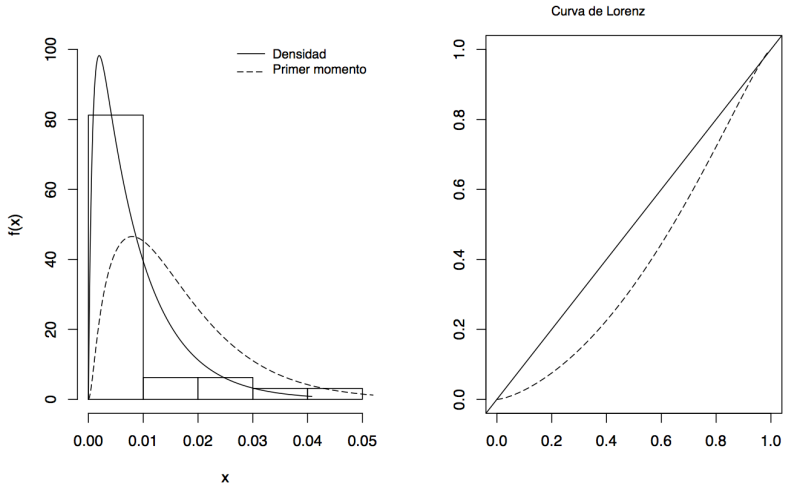


Figura 4: Año 1995, Gini=0.4605576

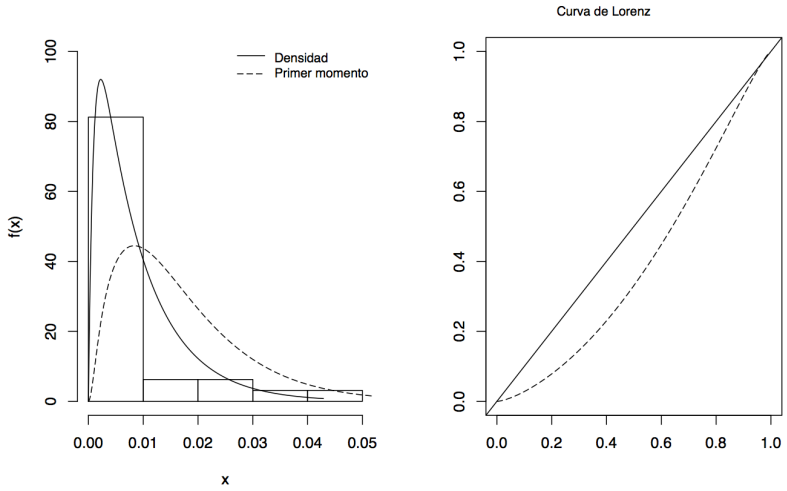


Figura 5: Año 2000, Gini=0.456076

4.1 **Comparación**

Ahora comparamos nuestros resultados con algunos que se brindan en la web. En primer lugar, comparamos con los resultados que la Universidad Autónoma Chapingo tiene en su

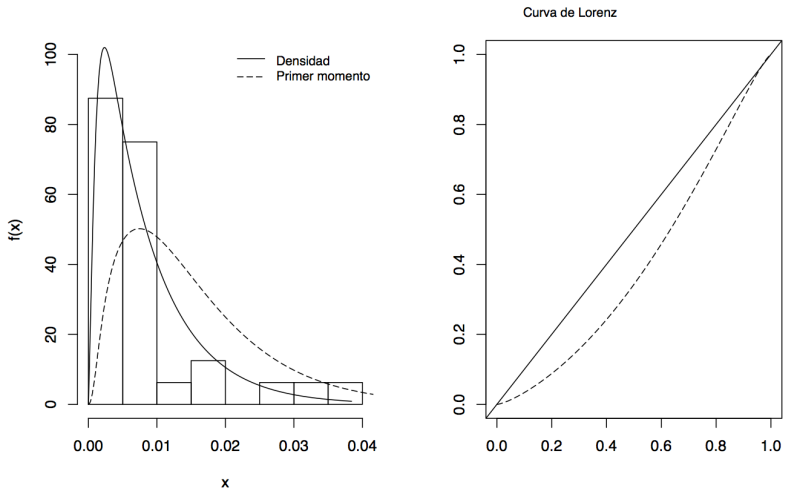


Figura 6: Año 2005, Gini=0.4488076

página de inicio (http: //www.chapingo.mx/dicifo /demyc/idh/new/). En segundo lugar, con el índice de Gini presentado por el Banco Mundial (http: //data.worldbank.org/indicator /SI.POV.GINI?locations=MX), desafortunadamente no hay información del índice de Gini para el año de 1995 y 2005, en su lugar, presentamos para el 1996 y el 2004. Ver Tabla 2.

Tabla 2:
Comparación del índice de Gini

Año	***Propuesto***	***Chapingo***	***The World Bank***
1995	0.4605	No disponible	0.4847 (año 1996)
2000	0.4560	0.44	0.5167 (año 2000)
2005	0.4488	No disponible	0.4603 (año 2004)

5 Conclusiones

En este artículo se ha propuesto una metodología nueva y eficiente para estimar el índice de Gini de cualquier país (lugar) utilizando el Producto Interno Bruto per cápita. Es bien sabido que las distribuciones de tipo fase tienen una amplia aplicación en el área de finanzas. Aquí, utilizando una subclase de las distribuciones tipo fase, llamada distribución de momentos tipo fase, pudimos obtener una estimación del índice de Gini en México.

Referencias

[1] S. Asmussen, O. Nerman, and M. Olsson. Fitting phase-type distributions via the EM algorihtm. *Scandinavian Journal of Statistics*, 23:419–441, 1996.

[2] M. Bladt. A Review on Phase–Type Distributions and Their Use in Risk Theory. *Astin Bulletin*, 35(1):145–161, 2005.

[3] M. Bladt, L. J. R. Esparza, and B. F. Friis. Fisher Information and Statistical Inference for Phase-type distributions. *J. Appl. Prob.*, 48A:277–293, 2011.

[4] M. Bladt and B. Friis. Moment Distributions of Phase Type. *Stochastic Models*, 27:651–663, 2011.

[5] A. P. Dempster, D. B. Rubin, and N. M. Laird. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. B*, (39):1–38, 1977.

[6] L. J. R. Esparza. On size-biased matrix-geometric distributions. *Performance Evaluation (PEVA)*, 70:639–645, 2013.

[7] Joseph L. Gastwirth. The estimation of the Lorenz Curve and Gini index. *The Review of Economics and Statistics*, 54(3):306–316, 1972.

[8] M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905.

[9] F. Medina. Consideraciones sobre el índice de Gini para medir la concentración del ingreso. *Comisión Económica para América Latina*, 2001.

[10] M. F. Neuts. Probability distributions of phase-type. *In Liber Amicorum Prof. Emeritus H. Florin*, pages 173–206, 1975.

[11] M. F. Neuts. *Matrix Geometric solutions in stochastic models*, volume 2. Johns Hopkins University Press, Baltimore, Md., 1981.

[12] S. Seth. A class of distribution and association sensitive multidimensional welfare indices. *The Journal of Economic Inequality*, 2(11):113–162, 2013.

[13] J. Van den Bergh and M. Antal. Evaluating alternatives to GDP as measure of social welfare/progress. page 56, 2014.