UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y
DE LA ESPECIALIDAD EN ESTADÍSTICA APLICADA

INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS

Estimación de la Distribución del Ingreso Corriente de los Hogares de la ENIGH 2016 Ajustado de Acuerdo al Ingreso de Cuentas Nacionales

TESINA

QUE PARA OPTAR POR EL GRADO DE:
ESPECIALISTA EN ESTADÍSTICA APLICADA

PRESENTA:
ROSA TOMASA SALINAS DÍAZ

Directora de Tesina
M. en E. Leticia Eugenia Gracia Medrano Valdelamar
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

Ciudad Universitaria. Cd. De México, 30 de agosto de 2019

AGRADECIMIENTOS

Agradezco a mis padres por toda la confianza que me han brindado.

Al Instituto de Matemáticas Aplicadas y en Sistemas por darme la oportunidad de acceder al programa de la Especialización en Estadística Aplicada, a mis profesores por formarme en el área de estadística.

Mi más sincero agradecimiento a mis revisores de tesis, a la Dra. Silvia Ruiz, a la Mtra. Patricia Romero, a la Dra. Lizbeth Naranjo y al Mtro. Javier Santibañez por sus comentarios y sugerencias que ayudaron a mejorar este trabajo, y en especial, un agradecimiento a la Maestra Leticia Medrano por el tiempo dedicado para la elaboración de esta tesina.

A mis amigos Alejandra y Ángel por apoyarme y acompañarme en este proyecto. A mis compañeros y amigos Ángel, Juan José y Octavio por su enorme paciencia, aprendí mucho de ustedes e hicieron agradable mi estancia en el Instituto.

Gracias a todos los que contribuyeron de alguna manera en la realización de este trabajo, pero sobre todo a aquellas personas que en algún momento de la vida hemos coincidido y que me han ayudado a crecer personal y profesionalmente.

CONTENIDO

1 Marco Conceptual **5**

1.1 Estadística descriptiva 5

1.1.1 Medidas de tendencia central 5

1.1.2 Medidas de dispersión 6

1.1.3 Medidas de forma 7

1.2 Histograma 8

1.3 Funciones de Distribución 9

1.3.1 Distribución Lognormal 9

1.3.2 Distribución Gamma 9

1.3.3 Distribución Weibull 10

1.3.4 Distribución Gamma Generalizada 10

1.3.5 Distribución Beta Generalizada tipo 2 10

1.4 Estimación Puntual 12

1.4.1 Método de Máxima Verosimilitud 12

1.5 Pruebas de Bondad de Ajuste 13

1.5.1 Devianza Global 13

1.5.2 Criterio de Informacion Akaike Generalizado 13

1.5.3 Estadístico Anderson-Darling 14

1.6 Indicadores de desigualdad económica 14

1.6.1 Coeficiente de Gini 16

2 Análisis de Resultados **17**

2.1 Ajuste del ingreso corriente trimestral de la ENIGH 17

2.1.1 La ENIGH 17

2.1.2 Resultado Numérico 18

2.2 Ajuste del ingreso corriente trimestral de la ENIGH conciliado con Cuentas Nacionales 26

2.2.1 Cuentas por Sectores Institucionales 27
2.2.2 Comparación del ingreso corriente trimestral de los hogares de distintas fuentes de información 28
2.2.3 Método de máxima pseudo verosimilitud restringida . . . 29
2.2.4 Aplicación del método de máxima pseudo verosimilitud restringida 30

3 Conclusiones **37**

A Log-verosimilitud **45**

B Código **53**

INTRODUCCIÓN

El Instituto Nacional de Estadística y Geografía (INEGI), publica dos estadísticas respecto al ingreso corriente total de los hogares, una de ellas es la reportada por la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) publicada de manera bianual y la otra es la reportada por las Cuentas Nacionales, específicamente por las Cuentas por Sector Institucional. El problema radica que la información de las Cuentas Nacionales del año 2016 es más del doble que lo reportado en la ENIGH 2016, trayendo como consecuencia mediciones erróneas de la desigualdad económica en México.

En esta investigación se hace un análisis de la variable ingreso corriente de los hogares de México captada en la ENIGH del año 2016, se lleva a cabo un proceso de especificación, estimación y validación de cada una de las distribuciones propuestas para ajustar el ingreso corriente. Sin embargo, este proceso de ajuste a una distribución conocida se complica cuando se hace la conciliación con otras fuentes de información que proporcionan datos del mismo concepto, para lograr el objetivo de ajuste se expone y se aplica el método de máxima pseudo verosimilitud restringida, método expuesto en el artículo *Estimation of the distribution of income from survey data, adjusting for compatibility with other sources* del autor Alfredo Bustos, este trabajo de investigación está basado en dicho artículo, pero no pretende hacer una réplica exacta de lo realizado para el año 2012, más bien el enfoque de esta investigación está orientada a mostrar la parte de la inferencia estadística y se limita a dar conclusiones específicas referentes a la parte social no más allá del coeficiente de Gini.

La estructura del trabajo de investigación es la siguiente. En el primer capítulo se revisan los aspectos generales de la inferencia estadística que serán utilizados

a lo largo de este trabajo, la definición de las distribuciones de probabilidad que se asemejan a la información de la encuesta, así también se expone el método de estimación de los parámetros y las medidas de bondad de ajuste, terminando con las definiciones del coeficiente de Gini y curva de Lorenz, medidas referentes a la desigualdad económica. En el segundo capítulo se determina la distribución teórica que mejor ajusta a los datos de la ENIGH 2016, realizando las estimaciones de los parámetros y aplicando las pruebas estadísticas de bondad de ajuste en la segunda parte de este capítulo se expone el método de máxima pseudo verosimilitud restringida y se aplica para conciliar la información de la ENIGH 2016 con la información proporcionada por las Cuentas por Sector Institucional del mismo año. Y finalmente se explican los resultados del estudio y se hacen las comparaciones con el estudio que se realizó del año 2012. Por último se presentan los anexos para abordar los detalles para obtener los resultados, y las referencias que sustentan la investigación.

OBJETIVOS

Objetivos Generales

- Determinar la distribución de probabilidad teórica que mejor ajusta a la variable ingreso corriente trimestral de los hogares presentados en la ENIGH 2016.
- Ajustar la distribución de probabilidad del ingreso corriente trimestral presentados en la ENIGH 2016 conciliando con otras fuentes de información.
- Medir el nivel de desigualdad económica, dada la distribución de probabilidad ajustada al ingreso corriente trimestral de los hogares.

Objetivos Específicos

- Proponer una distribución teórica que describa el ingreso corriente trimestral de los hogares de la ENIGH 2016.
- Estimar los parámetros de la distribución teórica propuesta mediante el método de máxima verosimilitud.
- Aplicar las pruebas de bondad de ajuste para determinar que distribución teórica se ajusta mejor al ingreso corriente trimestral.
- Exponer el método de máxima pseudo verosimilitud restringida.
- Aplicar el método de máxima pseudo verosimilitud restringida para determinar la distribución que ajusta el ingreso corriente trimestral reportado en la ENIGH 2016, realizando la conciliación con las Cuentas por Sector Institucional, considerando ésta como fuente de información alterna.

- Aplicar la prueba de bondad de ajuste Anderson-Darling a las distribuciones propuestas que ajustan al ingreso corriente trimestral restringido a la información de Cuentas por Sector Institucional.

- Determinar el coeficiente de Gini de las distribuciones seleccionadas que mejor describen el ingreso corriente trimestral con y sin restricciones y analizar las diferencias entre dichos valores del coeficiente.

CAPÍTULO 1

MARCO CONCEPTUAL

En el presente capítulo se revisan algunos conceptos básicos de la teoría de la inferencia estadística que serán utilizados a lo largo de este trabajo de investigación.

1.1 Estadística descriptiva

La estadística descriptiva se refiere a las técnicas analíticas y gráficas que se utilizan para describir un conjunto de datos y estudiar la distribución de la variable en cuestión. Mediante el análisis de las medidas de tendencia central, dispersión y forma, se pueden descartar algunas de las distribuciones que se pueden ajustar a la variable.

Para especificar la distribución de una variable se analizan las características de tendencia central, dispersión y forma.

1.1.1 Medidas de tendencia central

Este tipo de medidas se utilizan para identificar el valor que resume al conjunto de observaciones, entre estas medidas se encuentra la media, mediana y moda; en este trabajo solo trabajaremos con la mediana y media.

Definición 1.1. *Sea $y_1, y_2 \ldots, y_n$ una muestra de una población. Se define la mediana como la observación que divide a la población al 50 porciento y se determina por*

$$Mediana = \begin{cases} X_{(\frac{n+1}{2})} & si \quad n \ es \ impar \\ \\ \dfrac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2} & si \quad n \ es \ par \end{cases}$$

Donde $X_{(i)}$ es el i-ésimo estadístico de orden y corresponde a la observación de la muestra que ocupa la i-ésima posición cuando ésta se ordena de menor a mayor.

Definición 1.2. *Sea $y_1, y_2 \ldots, y_n$ una muestra de una población. La media muestral es el promedio aritmético de éstas y se denota por*

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

El valor de la media puede verse afectada de manera desproporcionada por la existencia de observaciones discrepantes, por ello es necesario tener medidas que midan la dispersión de los datos.

1.1.2 Medidas de dispersión

Este tipo de medidas se utilizan para evaluar qué tanta dispersión hay en los valores de los individuos. Entre las medidas se encuentran la varianza y la desviación estándar.

Definición 1.3. *Sea $y_1, y_2 \ldots, y_n$ una muestra de una población. La varianza muestral es la suma del cuadrado de las diferencias entre las mediciones y su media de la muestra, dividida entre $n-1$ y se denota por*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

La varianza mide que tan dispersos están los datos alrededor de la media. A mayor valor de la varianza mayor variabilidad de los datos. En cambio, a menor valor, más homogeneidad.

Definición 1.4. *La desviación estándar de una muestra de observaciones es la raíz cuadrada positiva de la varianza, esto es*

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2}$$

1.1.3 Medidas de forma

Las medidas de forma se utilizan para identificar la morfología de la densidad de los datos. Dentro de este tipo de medidas se encuentra el coeficiente de asimetría y la curtosis.

Definición 1.5. *Sea $y_1, y_2 \ldots, y_n$ una muestra de una población. Se define el coeficiente de asimetría de la muestra como*

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^3}{s^3}$$

Donde $\overline{y}$ es la media muestral y s es la desviación estándar de la muestra.

Si el coeficiente de asimetría es cero, entonces la distribución es simétrica alrededor de su media. Un valor positivo del coeficiente indica un sesgo positivo, esto es, la mayoría de las observaciones son menores a la media, por tanto, el histograma tiene una joroba a la izquierda. Mientras que un resultado negativo del coeficiente significa que se tiene un sesgo negativo, es decir, la mayoría de las observaciones son mayores a la media y por tanto el histograma presenta una joroba a la derecha.

Definición 1.6. *Sea $y_1, y_2 \ldots, y_n$ una muestra de una población. Se define la curtosis de la muestra como*

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^4}{s^4} - 3$$

Donde $\overline{y}$ es la media muestral y s es la desviación estándar de la muestra.

La curtosis indica el apuntalamiento de la densidad respecto a un comportamiento normal. Si el valor de la curtosis es cero, entonces la distribución es mesocúrtica, es decir, la concentración alrededor de la media y las colas tienen la misma forma que en la distribución normal. Para el caso de que el valor de la curtosis es positivo se dice que la distribución de los datos es leptocúrtica, esto es, la forma de la media es más afilada y las colas son más pesadas que las de una distribución normal. Finalmente, si el valor de la curtosis es negativo, la distribución de los datos se denomina platicúrtica, entonces la forma alrededor de la media es más plana y las colas son más ligeras que las de una distribución normal.

1.2 Histograma

El histograma es un gráfico de barras, donde el eje x representa la variable de interés[1] y la altura de las barras es la frecuencia que presenta una clase[2]. El objetivo del histograma es mostrar el perfil de la distribución de los datos, para ello es necesario agrupar las observaciones es un número relativo de clases que son ajenas entre sí, de tal manera que no exista ambigüedad a que clase pertenece una observación dada.
El número de intervalos o clases que se emplea para clasificar los datos, se determina de acuerdo al número de observaciones. Existen reglas que sugieren la selección del número de clases o intervalos, entre ellas se encuentran la regla de Dixon y Kronmal, Velleman, Sturges y Freedman - Diaconis.

1. Regla de Dixon y Kronmal. El número de clases se determina por:

$$L = [10\, log_{10}\, n]$$

   Donde n es el número de observaciones y $[\cdot]$ indica la parte entera.
   Para el caso de que n, el número de observaciones sea pequeño, el número de clases se puede determinar por la regla de Velleman o Sturges.

2. Regla de Velleman.

$$L = [\, 2\, n^{\frac{1}{2}}\,]$$

3. Regla de Sturges.

$$L = 1 + \log_2\, n$$

4. Regla de Freedman - Diaconis. Considera el rango intercuartil IQR

$$L = 2\, \frac{IQR}{n^{\frac{1}{3}}}$$

Cabe mencionar que el rango de cada clase no es el mismo, existen situaciones que obligan a usar intervalos de disinta longitud, tal es el caso del ingreso de los hogares, pues para ingresos bajos, pequeñas diferencias son importantes, pero las mismas no lo son para ingresos altos.

[1] Considerando variables que tienen asociada una métrica.

[2] La frecuencia puede ser absoluta o relativa. La frecuencia absoluta indica el número de observaciones en una clase y la frecuencia relativa es el cociente de una frecuencia absoluta con respecto al número de observaciones en todas las clases. En el caso de considerar frecuencias absolutas, se dice que es un histograma de frecuencia absoluta, análogamente si se consideran frecuencias relativas se obtiene un histograma de frecuencia relativa.

1.3 Funciones de Distribución

En la presente sección se describen algunas de las densidades conocidas que tienen la característica de tener una asimetría positiva, es decir, ser sesgadas a la izquierda y que anteriores estudios de la distribución del ingreso en otros países muestran un buen ajuste a la variable ingreso *(McDonald, 1984)*. Una de las principales densidades para describir la variable ingreso es la distribución tetra paramétrica, Beta Generalizada de tipo dos (GB2) y que anida a otras densidades con tres y dos parámetros que son casos especiales de la GB2. A continuación se describen cada una de las distribuciones.

1.3.1 Distribución Lognormal

Definición 1.7. *Una variable aleatoria y se distribuye como una Lognormal con parámetros μ, σ si y sólo si su función de densidad está dada por*

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\,\pi}\,\sigma\,y} exp\left(\frac{-(log\,y-\mu)^2}{2\,\sigma^2}\right)$$

Donde $-\infty < \mu < \infty$, y $\sigma > 0$; en tal caso se denota $y \sim LOGNO(\mu, \sigma^2)$.

Proposición 1.8. *Si $y \sim LOGNO(\mu, \sigma^2)$, entonces la esperanza de la variable aleatoria esta dada por*

$$E(y) = exp\left(\mu + \frac{1}{2}\sigma^2\right)$$

1.3.2 Distribución Gamma

Definición 1.9. *Una variable aleatoria y tiene una disribución Gamma con parámetros μ, σ si y sólo si la función de densidad de y está dada por*

$$f(y;\mu,\sigma) = \frac{y^{\frac{1}{\sigma^2}-1}}{(\mu\,\sigma^2)^{\frac{1}{\sigma^2}}\,\Gamma(\frac{1}{\sigma^2})} exp\left(-\frac{y}{\mu\,\sigma^2}\right)$$

Donde $y, \mu, \sigma > 0$ y Γ es la función gamma definida como $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\,e^{-t}\,dt$. En tal caso se denota $y \sim GA(\mu, \sigma)$.

Proposición 1.10. *Si $y \sim GA(\mu, \sigma^2)$, entonces la esperanza de la variable aleatoria y esta dada por*

$$E(y) = \mu$$

1.3.3 Distribución Weibull

Definición 1.11. *Una variable aleatoria y tiene una disribución Weibull con parámetros $\mu, \sigma > 0$ si y sólo si la función de densidad de y está dada por*

$$f(y;\mu,\sigma) = \frac{\sigma}{\mu}\left(\frac{y}{\mu}\right)^{\sigma-1} exp\left(-\left(\frac{y}{\mu}\right)^{\sigma-1}\right)$$

Donde $y, \mu, \sigma > 0$; en tal caso se denota $y \sim WEI(\mu, \sigma)$.

Proposición 1.12. *Si $y \sim WEI(\mu, \sigma^2)$, entonces la esperanza de y esta dada por*

$$E(y) = \mu\,\Gamma\left(1 + \frac{1}{\sigma}\right)$$

1.3.4 Distribución Gamma Generalizada

Definición 1.13. *Una variable aleatoria y tiene una disribución Gamma Generalizada, con parámetros μ, σ, ν si y sólo si la función de densidad de y está dada por*

$$f(y;\mu,\sigma,\nu) = \frac{\nu\,\theta^{\theta}\left(\frac{y}{\mu}\right)^{\theta\,\nu}}{\Gamma(\theta)\,y} exp\left(-\theta\left(\frac{y}{\mu}\right)^{\nu}\right)$$

Donde $\mu, \sigma > 0$, $-\infty < \nu < \infty$ y $\theta = \dfrac{1}{(\sigma\,\nu)^2}$.
En tal caso se denota $y \sim GG(\mu, \sigma, \nu)$.

Proposición 1.14. *Si $y \sim GG(\mu, \sigma, \nu)$, entonces la esperanza de y está dada por*

$$E(y) = \mu\,\frac{\Gamma\left(\frac{1}{\nu} + \theta\right)}{\theta^{\frac{1}{\nu}}\Gamma(\theta)}$$

Observación 1.15. *Las distribuciones Lognormal, Gamma y Weibull son casos particulares de la distribución Gamma Generalizada, cuando $\nu \to 0, \nu = 1$, $\nu^2 = \frac{1}{\sigma^2}$ respectivamente*

1.3.5 Distribución Beta Generalizada tipo 2

Esta distribución fue derivada por *(McDonald 1984)* y es la más utilizada para modelar el ingreso, tendiéndose buenos resultados en otros países como España *(Prieto Alaiz & García Pérez, 2009)*, la distribución se caracteriza por tener cuatro parámetros uno de escala y tres de forma.

Definición 1.16. *Una variable aleatoria y tiene una disribución Beta Generalizada tipo 2, con parámetros μ, σ, ν, τ, si y sólo si la función de densidad de y está dada por*

$$f(y;\mu,\sigma,\nu,\tau) = \frac{|\sigma|\, y^{\sigma\nu-1}}{\mu^{\sigma\nu}\, B(\nu,\tau)\, \left(1+\left(\frac{y}{\mu}\right)^{\sigma}\right)^{\nu+\tau}}$$

Donde $\mu, \nu, \tau > 0$, $-\infty < \sigma < \infty$ y $B(\nu,\tau)$ es la función beta; en tal caso se denota $y \sim GB2(\mu,\sigma,\nu,\tau)$

Proposición 1.17. *Si $y \sim GB2(\mu,\sigma,\nu,\tau)$, entonces la esperanza de y está dada por*

$$E(y) = \mu\, \frac{B\left(\nu+\frac{1}{\sigma}, \tau-\frac{1}{\sigma}\right)}{B(\nu,\tau)}$$

Observación 1.18. *La distribución triparamétrica Gamma Generalizada es un caso particular de la GB2, cuando $\tau \to 0$ y esta a su vez tiene concatenadas las distribuciones de dos parámetros Lognormal, Gamma y Weibull. En la figura* 1.1 *se muestra la relación que existe entre las cinco distribuciones.*

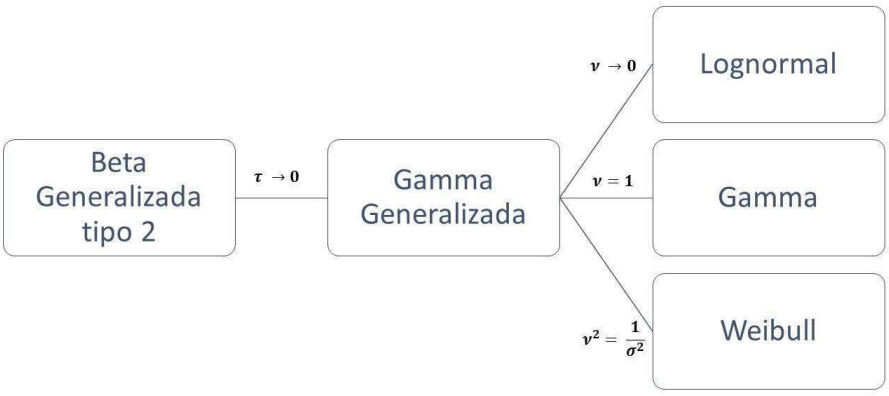

Figura 1.1: Diagrama de Distribuciones. Relación existente entre las distribuciones. Fuente: elaboración propia, basada en *(Prieto Alaiz & García Pérez, 2009)*.

1.4 Estimación Puntual

Existen técnicas adecuadas para implementar el proceso de estimación de los parámetros de una función de distribución, entre ellos se encuentra el método de máxima verosimilitud y que se describe a continuación.

1.4.1 Método de Máxima Verosimilitud

La característica esencial del método de máxima verosimilitud es que selecciona como estimador a aquél valor del parámetro que tiene la propiedad de maximizar el valor de la función de densidad conjunta de la muestra aleatoria observada.

Definición 1.19. *Si $y_1, y_2 \ldots, y_n$ son los valores de una muestra aleatoria simple de una población con función de densidad $f(y;\theta)$ se define la función de verosimilitud de la muestra $L : \Theta \to (0, \infty)$ como la función de densidad conjunta*

$$L(\underline{y};\theta) = \prod_{i=1}^{n} f(y_i;\theta)$$

Donde $\underline{y} = (y_1, y_2 \ldots, y_n)$ y para los valores de θ dentro del dominio Θ dado.

La función de verosimilitud solo depende del vector de parámetros θ, donde θ puede ser un escalar o un vector, $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$.

Definición 1.20. *Se define la función log-verosimilitud como el logaritmo natural de la función de verosimilitud, es decir*

$$l(\underline{y};\theta) = \log L(\underline{y};\theta) = \sum_{i=1}^{n} \log f(y_i;\theta)$$

Donde $\underline{y} = y_1, y_2 \cdots, y_n$

Definición 1.21. *Sean $y_1, y_2 \ldots, y_n$ una muestra aleatoria con función de densidad $f(y;\theta)$ y sea $L(\underline{y};\theta)$ la verosimilitud de la muestra como función de θ. Si $t = u(y_1, y_2, \ldots, y_n)$ es el valor de θ para el cual el valor de la función de verosimilitud es máximo, entonces $T = u(Y_1, Y_2, \ldots, Y_n)$ es el estimador de máxima verosimilitud de θ y t es la estimación de máxima verosimilitud.*

El máximo de $l(\underline{y};\theta)$ se alcanza en el mismo lugar que el máximo de $L(\underline{y};\theta)$, por lo que maximizar la log-verosimilitud es equivalente a maximizar la verosimilitud. La ventaja de este método es que produce estimadores asintóticamente insesgados de varianza mínima.

1.5 Pruebas de Bondad de Ajuste

Una forma de medir la adecuación de un modelo es proporcionando medidas de bondad de ajuste. A continuación, se muestran estadísticos que son utilizados como un criterio para la selección de modelos entre un conjunto finito de ellos. La Devianza Global y el Criterio de Información Generalizado se definen de acuerdo a *(Stasinopoulos, Rigby, Heller, Voudouris, & De Bastiani, 2017).*

1.5.1 Devianza Global

Definición 1.22. *La Devianza Global (GDEV) se define como*

$$D = -2 \log \hat{L}_c$$

Donde $\hat{L}_c$ es la verosimilitud del modelo estimado.

1.5.2 Criterio de Informacion Akaike Generalizado

Definición 1.23. *El Criterio de Información Akaike Generalizado (GAIC[3]) se define como*

$$GAIC = -2 \log \hat{L}_c + (k \cdot df)$$

Donde:
$\hat{L}_c$ es la verosimilitud del modelo ajustado.
df son los grados de libertad efectivos (es decir, el número efectivo de parámetros) del modelo.
k es la penalización por cada grado de libertad usado.

Existen casos particulares del GAIC, de acuerdo al valor de k, en particular si $k = 2$ y $k = \log n$.

Definición 1.24. *El Criterio de Información Akaike (AIC[4]) se define como*

$$AIC = -2 (\log \hat{L}_c - df)$$

Donde
$\hat{L}_c$ es la verosimilitud del modelo ajustado.
df son los grados de libertad efectivos.

[3]Por sus siglas en Inglés General Akaike Information Criterio
[4]Por sus siglas en Inglés Akaike Information Criterio

Definición 1.25. *El Criterio Bayesiano de Schwarz (SBC[5]) se define como*

$$SBC = -2 \log \hat{L}_c + (\log n \cdot df)$$

Donde
$\hat{L}_c$ *es la verosimilitud del modelo ajustado.*
df *son los grados de libertad efectivos.*
n *es el número de observaciones.*

El modelo con el menor valor de GAIC es el modelo que mejor ajuste tiene a los datos.

1.5.3 Estadístico Anderson-Darling

El estadístico Anderson-Darling mide qué tan bien siguen los datos una distribución específica comparando la distribución de probabilidad acumulada empírica con la distribución de probabilidad acumulada específica.

Definición 1.26. *Dado un conjunto de datos ordenados de manera ascendente* $Y_1, Y_2 \ldots, Y_n$. *El estadístico Anderson-Darling se define como*

$$A^2 = -N - \sum_{k=1}^{n} \frac{2k-1}{N} \left[\log F(Y_k) + \log(1 - \log F(Y_{N+1-k}))\right]$$

Donde
N *es el tamaño del subconjunto de datos.*
F *es la función acumulada propuesta, con parámetros estimados a partir del conjunto de datos dado.*

El estadístico Anderson-Darling se utiliza en la prueba que lleva el mismo nombre para determinar si un conjunto de datos proviene de una distribución de probabilidad específica *(Marsaglia & Marsaglia, 2004)*, mientras mejor se ajuste la distribución a los datos, menor será el valor del estadístico.

1.6 Indicadores de desigualdad económica

En esta última sección se mencionará lo referente a la desigualdad económica, sin profundizar en la parte social de este tema, más bien será orientado a dar una definición matemática concreta.

[5]Por sus siglas en Inglés Schwarz Bayesian Criterio

De acuerdo con la Organización para la Cooperación y el Desarrollo Económico (OCDE), la desigualdad económica es la diferencia en cómo se distribuyen los ingresos o activos entre la población, esta desigualdad se puede medir mediante índices, y uno de los índices unidimensionales de desigualdad relativa[6] más difundido es el índice de Gini y que se deriva de la curva de Lorenz.
La curva de Lorenz es una forma gráfica de mostrar la concentración de la distribución del ingreso de una población. Esta curva ordena a la población de manera ascendente según su ingreso y muestra el porcentaje acumulado de los ingresos de la población ordenada, es decir, se muestra el nivel de ingreso acumulado por el p porciento de la población *(CONEVAL, 2019)*.

La figura 1.2 ejemplifica la forma de la curva de Lorenz. Entre más cerca esté la curva a la identidad la desigualdad se reduce, es decir, el porcentaje del ingreso acumulado aumenta en forma constante y proporcional, como se muestra en la curva C1, la curva C2 muestra una desigualdad mayor al estar más lejana a la identidad.

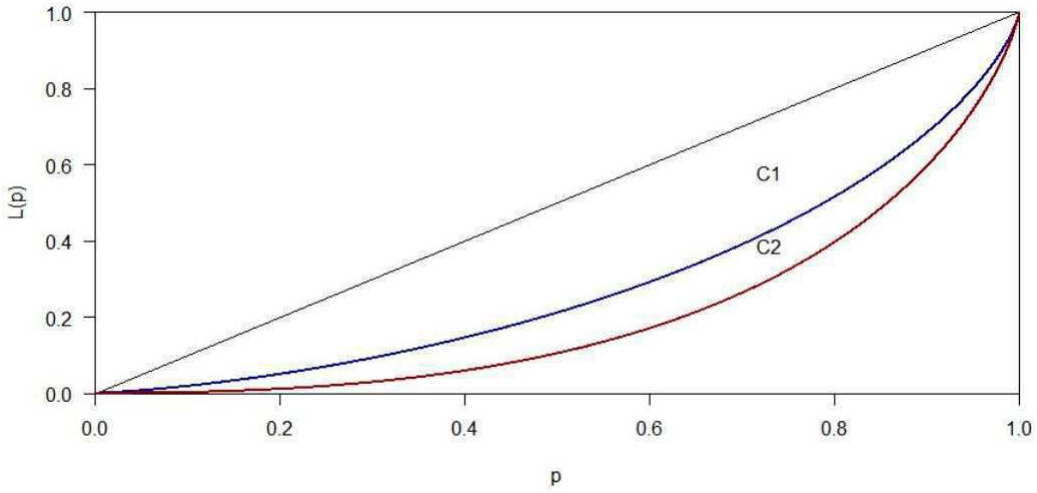

Figura 1.2: Curva de Lorenz. La curva azul representa a la curva C1, y la curva roja representa a la curva C2.
Fuente: elaboración propia.

[6]La desigualdad relativa se refiere a la relación entre los ingresos de los hogares y el ingreso medio, es decir, si todos los ingresos se incrementan en una misma proporción, la desigualdad no cambia.

1.6.1 Coeficiente de Gini

Definición 1.27. *El coeficiente de Gini es el área comprendida entre la diagonal y la curva de Lorenz, y se determina mediante*

$$G = 1 + \frac{1}{N} - \frac{2}{\overline{y}\, N^2} \sum_{i=1}^{N} Y_i(N + 1 - i)$$

Donde
N *es el número de personas o estratos de ingreso.*
$\overline{y}$ *es el ingreso medio.*
Y_i *es el ingreso de la persona o estrato* i*.*

El coeficiente de Gini G está acotado entre cero y uno. Si $G = 1$ significa que todos los ingresos lo concentran una sola persona y para el caso $G = 0$ refiere a que los ingresos son iguales en toda la población, en el mundo real no suceden los dos casos anteriores, sino que fluctúan entre estos dos valores. Si G es cercano a 1 refleja una mayor desigualdad en la distribución del ingreso y en el caso que G este próximo a cero indica que existen mayores condiciones de equidad en la distribución del ingreso.

CAPÍTULO 2

ANÁLISIS DE RESULTADOS

Para esta investigación se trabajará con la base de datos de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) que se llevó acabo en el año 2016[1], en particular se trabajó con la tabla **Concentrado_hogar**[2].

2.1 Ajuste del ingreso corriente trimestral de la ENIGH

2.1.1 La ENIGH

La ENIGH tiene por objetivo proporcionar un panorama estadístico del comportamiento de los ingresos y gastos de los hogares en cuanto a su monto, procedencia y distribución; adicionalmente ofrece información sobre las características ocupacionales y sociodemográficas de los integrantes del hogar, así como las características de la infraestructura de la vivienda y el equipamiento del hogar. *(Instituto Nacional de Estadística y Geografía (INEGI), 2018d).*

La población objetivo la constituyen los hogares que residen dentro del territorio nacional; el tamaño de la muestra fue de 81 mil 515 viviendas, la unidad de observación es el hogar[3]. La encuesta se levanta bianual desde 1984 hasta el año 2016 y con el paso del tiempo ha teniendo mejoras, desagregando la información por niveles: vivienda, hogar e integrantes del hogar, de esta manera definen las unidades de análisis. Cada uno de los tres niveles se tiene bien

[1] Llevada a cabo del 21 de agosto al 28 de noviembre de 2016.

[2] Descargado de https://www.inegi.org.mx/programas/enigh/nc/2016/default.html

[3] En la base de datos se tiene el registro de 69,126 viviendas con 70,311 hogares.

identificado con el apoyo de llaves[4] que apoyan a distinguir cada una de las categorías *(Instituto Nacional de Estadística y Gografía (INEGI), 2018d)*. El diseño de la encuesta es del tipo estratificado, bietápico y por conglomerados; el marco de muestreo se constituyó con información demográfica y cartográfica que se obtuvo en el Censo de Población y Vivienda del año 2010.
Además del ingreso corriente trimestral de los hogares, otra variable a considerar en el análisis es el factor de expansión, que refiere al peso que se le da a cada unidad muestral para generalizar resultados de la muestra a la población *(Instituto Nacional de Estadística y Geografía (INEGI), 2018b)* y que será incluido en la función de verosimilitud para estimar los parámetros considerando el diseño de la muestra.

2.1.2 Resultado Numérico

La variable a analizar es el ingreso corriente trimestral de los hogares, etiquetada como **ing_cor**, esta variable se define como la suma de los ingresos provenientes por trabajo (agropecuario que incluye del tipo agrícola, pecuario y pesca, no agropecuario, que considera la industria, comercio y servicios, y otros trabajos), rentas de la propiedad, y transferencias (pensiones, becas, donativos y remesas).

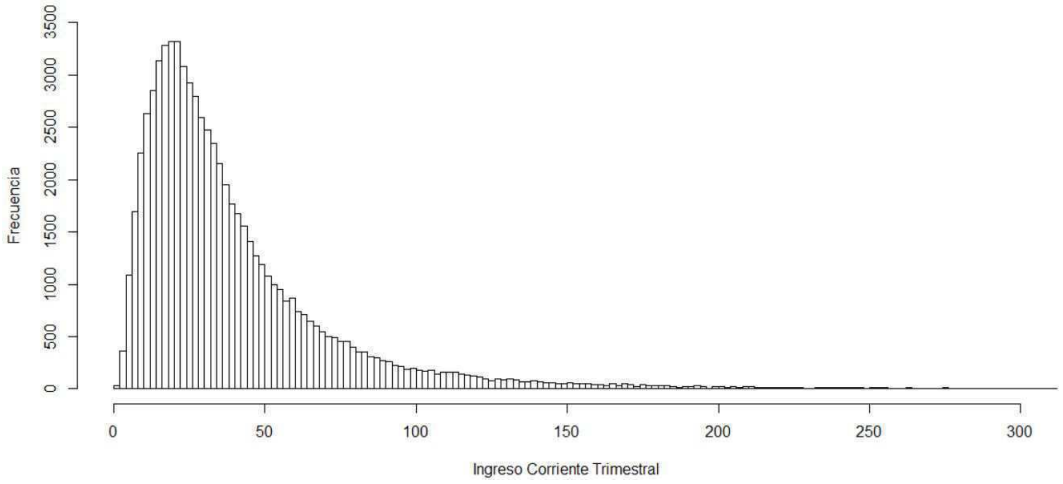

Figura 2.1: Histograma del ingreso corriente trimestral (miles de pesos).
Fuente: elaboración propia.

[4] **Folioviv** es la llave para el nivel de vivienda, **Foliohog** es la llave para identificar los hogares en la vivienda, y **Numren** es el número de identificación único para cada integrante del hogar.

Como un primer acercamiento a los datos de los ingresos corrientes se hace un análisis exploratorio gráfico. En la figura 2.1 se muestra el histograma del ingreso corriente trimestral de los hogares en la muestra[5]. Observemos que el histograma muestra asimetría marcada a la izquierda y el rango de los ingresos es amplio, esto puede ser por algunos datos que se encuentren muy a la derecha. Un método común para detectar valores atípicos (outliers) es mediante el gráfico Box-Plot como el que se muestra en la figura 2.2.

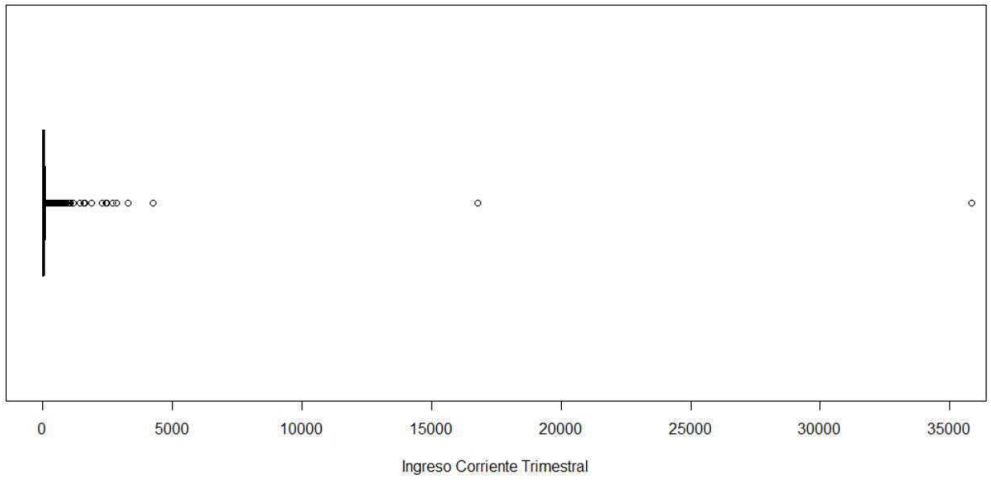

Figura 2.2: Box-Plot del ingreso corriente trimestral (miles de pesos). Muestra almenos tres outliers.
Fuente: elaboración propia.

En el diagrama se puede ver gráficamente la asimetría que existe a la izquierda de los datos y que se observa también en el histograma, asimismo se pueden detectar al menos dos valores atípicos (outliers) a la derecha, es decir, que sobresalen por su ingreso alto, y que corresponden a los ingresos de 16 y 35 millones declarados.
Para mejorar el análisis estadístico de la información se omitieron 28 datos, 6 datos reportaban un ingreso de cero[6], y 22 datos con un ingreso mayor a un millón de pesos[7], y existían 2 outliers de 16 y 35 millones. Si solo se consideran ingresos mayores a cero y menores a un millón se tiene una mejor visión de la

[5]A pesar de que el tamaño de la muestra es de 81,515 se tiene un 15 porciento de no respuesta y que en el diseño de la muestra se tiene contemplado, pero que no afectan a los resultados de la encuesta.

[6]Ingesos de cero son una falsa respuesta, pues dentro del ingreso corriente incluyen las transferencias y donaciones del estado y de las instituciones sin fines de lucro.

[7]Las 22 observaciones no son significativas para el análisis, además de que no permiten observar de una manera más clara la distribución de los datos.

distribución del ingreso[8]. Calculando las estadísticas descriptivas de los 70,283 datos como la mediana, media, desviación estándar, asimetría y curtosis, la tabla 2.1 muestra los resultados.

Tabla 2.1: Estadísticas Descriptivas del ingreso corriente trimestral en los hogares.

Estadística Descriptiva	**Valor**
Mediana	29,834.2
Media	40,734.9
Desviación estándar	41,273.24
Asimetría	5.46
Curtosis	65.60

Fuente: elaboración propia.

De acuerdo a la tabla 2.1 donde se muestran las estadísticas descriptivas del ingreso corriente trimestral se observa que el ingreso trimestral medio es de $40,734.9$ pesos[9], sin embargo, la desviación estándar de $41,273.24$ nos dice que existe una variabilidad muy grande de los ingresos, concluyendo que la media no es tan representativa para explicar los ingresos, en este caso quien mejor describe a los ingresos corrientes es la mediana, pues nos dice que el 50 porciento de la población percibe ingresos trimestrales menores a $29,834$ pesos. Y para determinar la forma de la distribución que tienen los ingresos corrientes trimestrales nos apoyamos de las medidas de forma, dado que el valor del coeficiente de asimetría es de 5.46 nos indica que la mayoría de los datos son menores a la media, por tanto la distribución empírica de los ingresos corrientes trimestrales presenta una joroba a la izquierda, finalmente la curtosis positiva de 65.6 marca que las colas de la distribución empírica de los ingresos son más pesadas que las de una normal.

El objetivo es hallar una distribución teórica que mejor se ajuste a la distribución del ingreso corriente trimestral de los hogares en la muestra de la encuesta. Sustentados con las estadísticas descriptivas y en las figuras 2.1 y 2.3 se pueden sugerir algunas distribuciones teóricas conocidas que tienen las características similares, y que además en estudios anteriores se han utilizado como en *(Prieto Alaiz & García Pérez, 2009)* y *(Bustos, 2015)*.

[8]En el apéndice A.2 se muestran las estimaciones de los parámetros para los casos en donde solo se omiten los ingresos nulos y los outliers; las estimaciones no varian mucho respecto a las estimaciones consideradas en el trabajo de investigación, no se pude considerar el total de las observaciones de la muestra, ya que el modelo utilizado para la estimación de los parámetros no esta definido para ingresos igual a cero.

[9]Para el cálculo de la media se considero el diseño de la muestra, es decir la estratificación y los conglomerados que se generaron para el levantamiento de la encuesta.

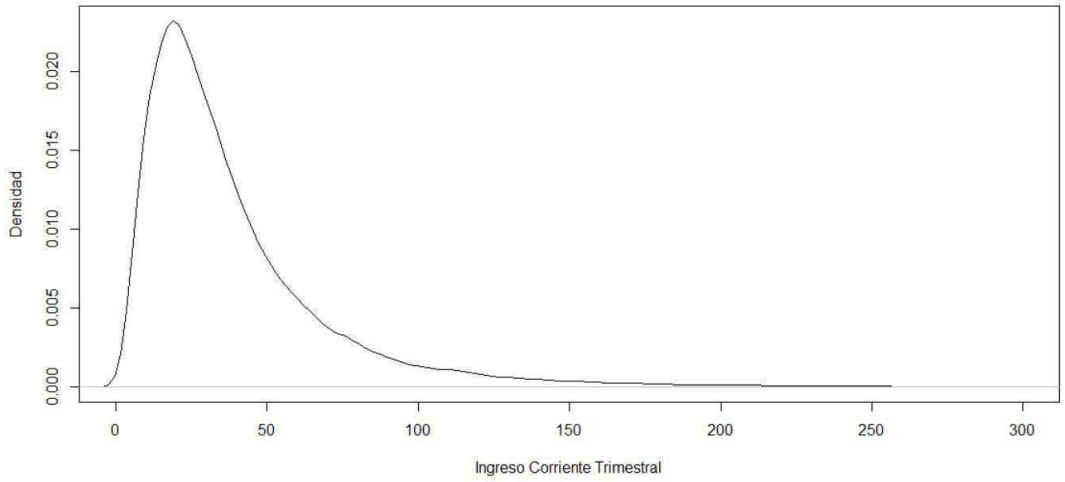

Figura 2.3: Densidad empírica del ingreso corriente trimestral (miles de pesos). Fuente: elaboración propia.

Las distribuciones teóricas propuestas para ajustar el ingreso corriente trimestral de los hogares y que se analizarán son las siguientes:

1. Distribución Lognormal (2 parámetros).
2. Distribución Gamma (2 parámetros).
3. Distribución Weibull (2 parámetros).
4. Distribución Gamma Generalizada (3 parámetros).
5. Distribución Beta Generalizada tipo 2 (4 parámetros).

La justificación de la elección de estas cinco distribuciones es que las tres primeras están anidadas en la Gamma Generalizada, y ésta a su vez está anidada en la distribución Beta Generalizada como se analizó en el diagrama de las distribuciones (ver figura 1.1).

Posterior a la selección de las distribuciones con los cuales ajustará a la distribución del ingreso corriente trimestral de los hogares en la muestra de la encuesta la siguiente fase es estimar los parámetros de cada una de las cinco distribuciones y que sean asintóticamente eficientes. Se aplica el principio de máxima verosimilitud sobre los datos de los ingresos corrientes trimestral de los hogares[10] ponderados por los pesos debidos al diseño muestral, es decir

[10]En el Apéndice A se encuentran las expresiones de la verosimilitud de cada una de las cinco distribuciones propuestas.

cosidera la magnitud de representación que cada observación de la muestra posee para describir una parte de los hogares[11]. Con el apoyo del software estadístico R y la librería GAMLSS[12] nos permite obtener los valores de los parámetros de una manera más eficiente. En la tabla 2.2 se muestran las estimaciones de los parámetros de cada distribución propuesta.

Tabla 2.2: *Estimación de Parámetros*

	Distribución				
Parámetros	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
μ	10.36	44,221.54	47,529.57	31,141.07	31,027.89
σ	0.80	0.77	1.20	0.79	2.05
ν				-0.06	1.13
τ					1.10

Fuente: elaboración propia.

Una vez estimados los parámetros de cada una de las cinco distribuciones propuestas, se procede a evaluar la calidad del ajuste, esta calidad se medirá mediante el criterio de información Akaike generalizado GAIC. En la tabla 2.3 se muestran los valores del GAIC, AIC y SBC.

Tabla 2.3: *Bondad de Ajuste*

	Distribución				
Estadístico	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
GAIC	773,417,657	777,926,709	780,223,805	773,400,936	773,481,953
AIC	773,417,661	777,926,713	780,223,809	773,400,942	773,481,961
SBC	773,417,692	777,926,743	780,223,840	773,400,988	773,482,023

Fuente: elaboración propia.

De acuerdo a los criterios del AIC y SBC, la distribución que mejor se ajusta a los datos es aquella que tiene el menor valor, dado que la distribución Gamma Generalizada es aquella que presenta el menor valor, seguida de las distribuciones Lognormal y Beta Generalizada, se selecciona como la distribución que ajusta al ingreso corriente trimestral de los hogares. El gráfico 2.4 muestra la comparación de las densidades modeladas y la densidad empírica, y sustentan

[11]La variable **factor** contiene la información de los peso.

[12]En el ápendice B se muestra el código necesario de la paquetería GAMLSS para la estimación de los parámetros de cada una de las cinco distribuciones.

los valores del estadístico AIC, pues la densidad Gamma Generalizada es la que está más próxima a la densidad empírica en la parte más concentrada de los datos.

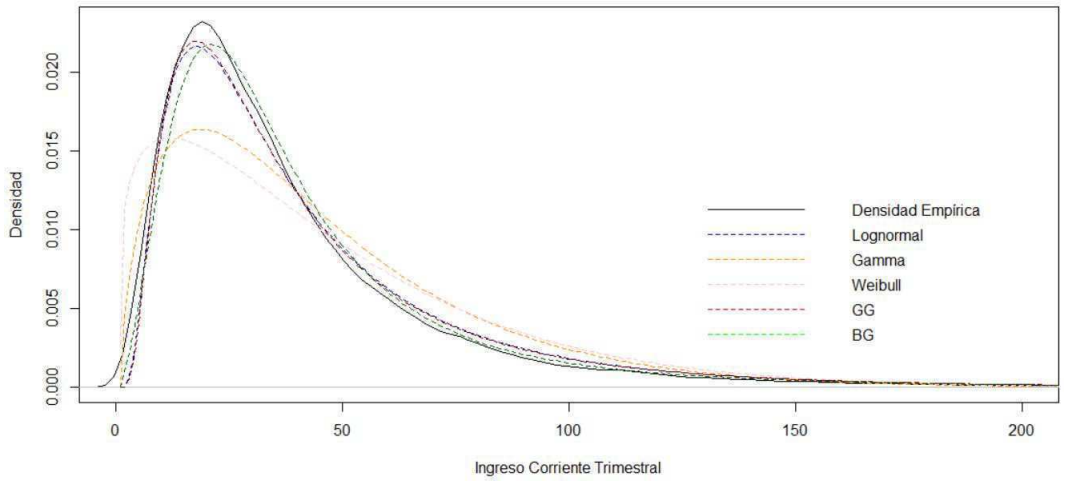

Figura 2.4: Ajuste de la densidad empírica del ingreso corriente trimestral a las cinco densidades propuestas (miles de pesos).
Fuente: elaboración propia.

También se puede ver el ajuste mediante las distribuciones de probabilidad de cada una de las cinco propuestas versus la distribución empírica del ingreso corriente trimestral, como se muestra en las figuras 2.5-2.9, de la misma forma se comprueba que las distribuciones Lognormal, Gamma Generalizada y Beta Generalizada son las que presentan el mejor ajuste.

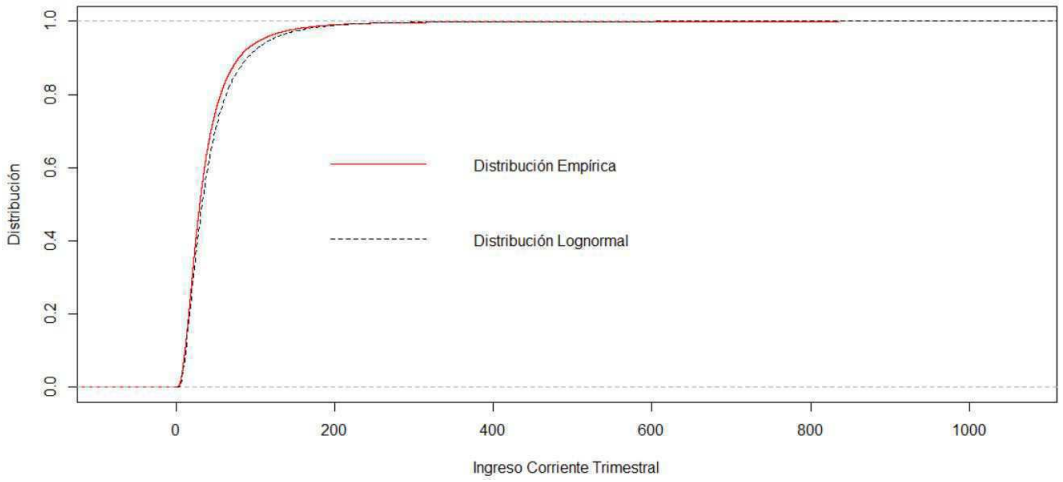

Figura 2.5: Ajuste de la distribución empírica del ingreso corriente trimestral a la distribución Lognormal (miles de pesos).
Fuente: elaboración propia.

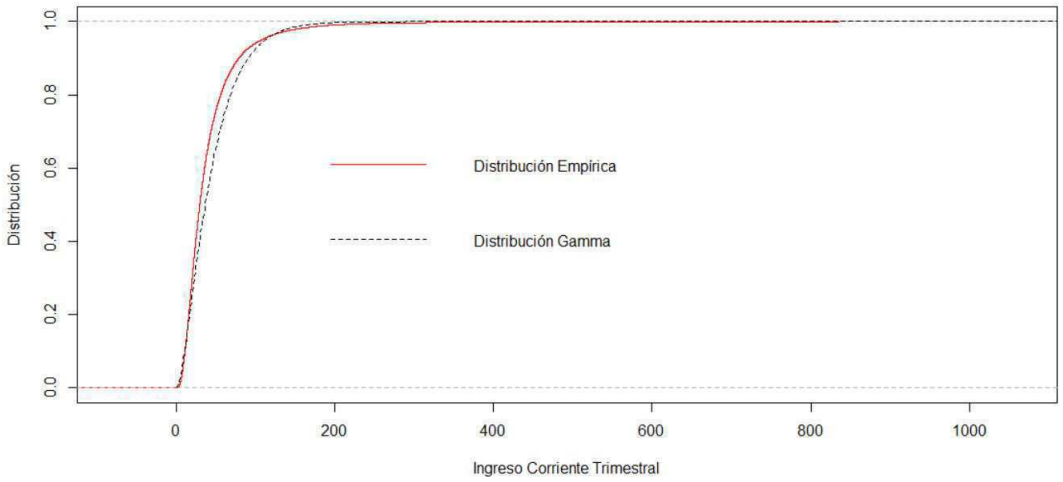

Figura 2.6: Ajuste de la distribución empírica del ingreso corriente trimestral a la distribución Gamma (miles de pesos).
Fuente: elaboración propia.

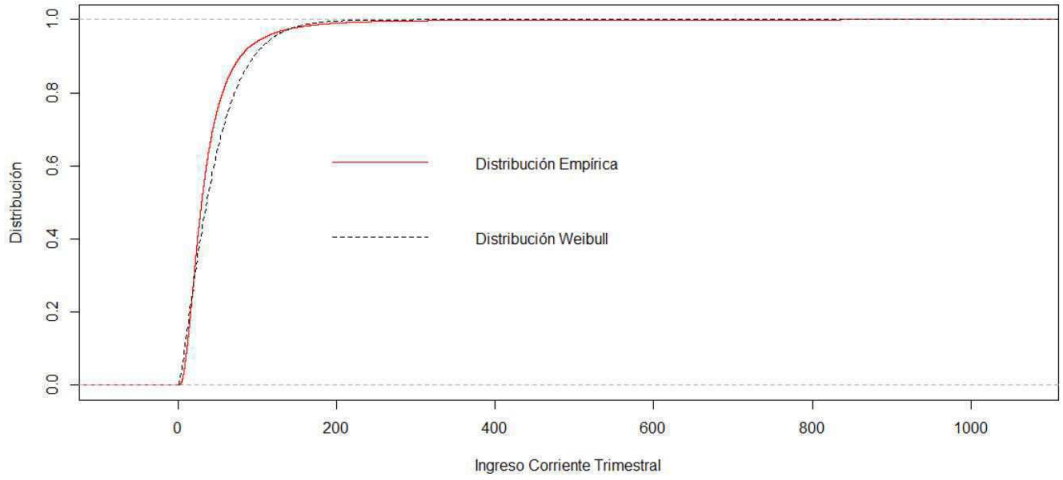

Figura 2.7: Ajuste de la distribución empírica del ingreso corriente trimestral a la distribución Weibull (miles de pesos).
Fuente: elaboración propia.

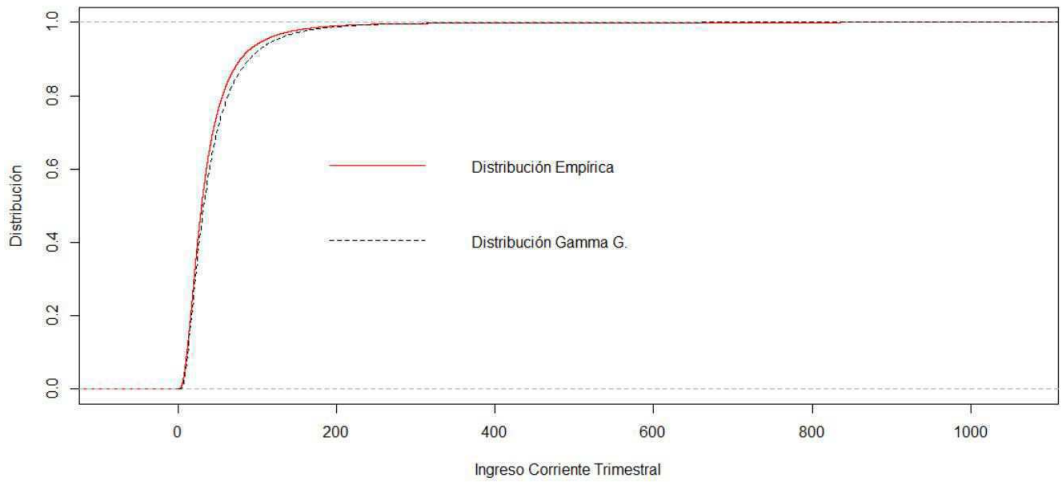

Figura 2.8: Ajuste de la distribución empírica del ingreso corriente trimestral a la distribución Gamma Generalizada (miles de pesos).
Fuente: elaboración propia.

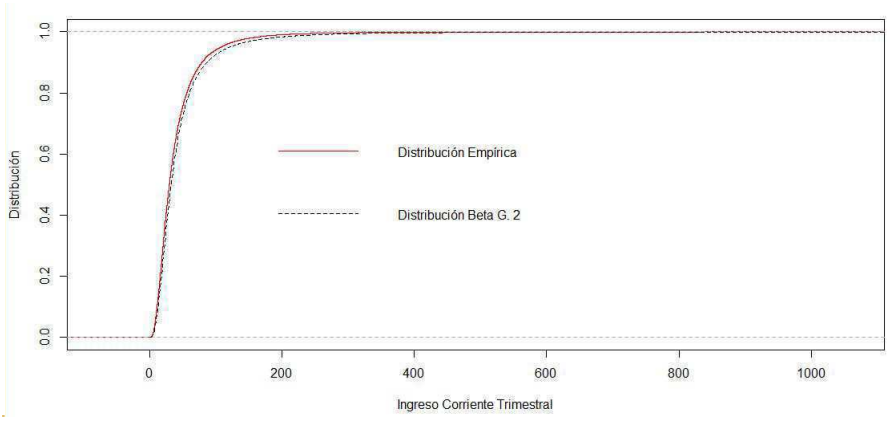

Figura 2.9: Ajuste de la distribución empírica del ingreso corriente trimestral a la distribución Beta Generalizada tipo 2 (miles de pesos).
Fuente: elaboración propia.

Concluyendo que la distribución Gamma Generalizada es la distribución con el mejor ajuste al ingreso corriente de la ENIGH 2016. Para esta distribución el coeficiente de Gini es de 0.433 (ver tabla 2.4), 0.015 menor al coeficiente de Gini estimado para la ENIGH 2016, es decir, para la distribución ajustada, Gamma Generalizada, hay menor desigualdad que lo reportado en los resultados de la encuesta.

Tabla 2.4: *Desigualdad Económica*

	Distribución				
	Lognormal	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
Gini	0.4290	0.4066	0.4371	0.4330	0.4526

Fuente: elaboración propia.

2.2 Ajuste del ingreso corriente trimestral de la ENIGH conciliado con Cuentas Nacionales

Una vez ajustada una distribución teórica a los datos del ingreso corriente de la ENIGH 2016, se compara con otras fuentes de información alternas para la misma variable, en este caso se contrasta con la información anual de las Cuentas por Sector Institucional reportada por el INEGI. Antes de la comparación mencionaremos a qué se refieren las Cuentas Nacionales por Sector Institucional.

2.2.1 Cuentas por Sectores Institucionales

Las Cuentas Nacionales proporcionan una descripción integral de toda actividad económica en el territorio económico del país *(INEGI, 2018c)*, y permiten tomar decisiones macroeconómicas.
De acuerdo con el Sistema de Cuentas Nacionales (SCN) 2008, las unidades económicas se pueden clasificar en cinco sectores mutuamente excluyentes de acuerdo a sus funciones, objetivos y actividad económica, más su interacción con el resto del mundo:

1. Sociedades no financieras (S11).
2. Sociedades financieras (S12).
3. Gobierno general (S13).
4. Hogares (S14).
5. Instituciones sin fines de lucro al servicio de los hogares (S15).
6. Resto del mundo (S2).

De manera que, todo agente económico se encuentra clasificado de manera única en cualquiera de estos cinco sectores y además tienen la capacidad de realizar transacciones, contraer pasivos y tener interacción con otros sectores. De esta forma, las Cuentas por Sector Institucional son un registro sistemático y normalizado de las operaciones económicas vinculadas con la producción, distribución acumulación y financiamiento de las unidades económicas en el país. En la tabla 2.5 se resumen las principales variables macroeconómicas presentadas en la sucesión contable de las Cuentas por Sector Institucional *(Fondo Monetario Internacional, 2016)*.
El saldo a considerar para esta investigación es el ingreso disponible (B.6) que se obtiene sumando al saldo del ingreso primario (B.5) todas las transferencias corrientes recibidas por cada sector y restando todas las transferencias corrientes pagadas[13].
Poniendo énfasis en el cuarto sector institucional, los hogares, que son básicamente los destinatarios de toda actividad económica. El manual del Sistema de Cuentas Nacionales (SCN 2008) define a un hogar como un grupo de personas que comparten la misma vivienda y juntan total o parcialmente sus ingresos y riqueza, consumiendo colectivamente ciertos bienes y servicios (*Fondo Monetario Internacional, 2016)*.

[13]Estas transferencias son: Impuestos corrientes sobre el ingreso (D.5), contribuciones sociales netas (D.61), prestaciones sociales distintas a las prestaciones sociales en especie (D.61) y otras transferencias corrientes (D.7).

Tabla 2.5: *Sucesión de Cuentas por Sector Institucional*

Nomenclatura	**Saldo Contable**
B.1	Valor agregado/Producto interno
B.2	Excedente de operación
B.3	Ingreso mixto
B.5	Saldo de ingresos primarios/ingreso nacional
B.6	**Ingreso disponible**
B.7	Ingreso disponible ajustado
B.8	Ahorro
B.9	Préstamo neto/endeudamiento neto
B.10.1	Variaciones al valor neto
B.11	Saldo de Bienes y Servicios con el exterior
B.9	Saldo corriente con el exterior
B.90	Valor neto

Fuente: elaboración propia.

Por lo tanto, la información principal para esta investigación se concentra en el ingreso disponible del sector hogares (S.14) del año 2016 de las Cuentas por Sector Institucional y la definición es equivalente al ingreso corriente de los hogares de la ENIGH 2016.

2.2.2 Comparación del ingreso corriente trimestral de los hogares de distintas fuentes de información

En diversos estudios de la desigualdad económica se utilizan como fuentes de información la ENIGH y los datos producidos por las Cuentas Nacionales, pero se tiene un inconveniente, existen discrepancias muy significativas en ambas fuentes de información.
De acuerdo a la información disponible en la ENIGH, en el año 2016 se estima que el ingreso total promedio por un trimestre es de $1,556,701,245$ pesos *(Instituto Nacional de Estadística y Geografía (INEGI), 2018a)* y dividiéndolo entre el total de hogares[14] se obtiene que el ingreso total promedio por hogar es de $46,520.63$ pesos.

Por otra parte, las Cuentas por Sector Institucional reportan un ingreso disponible anual del sector hogares de $16,279,679$ millones de pesos para el año 2016[15] (ver tabla 2.6); haciendo el supuesto que no existe estacionalidad, se tiene que el ingreso trimestral es de $4,069,919,750,000$ pesos de acuerdo a las

[14] La muestra se expande a 33, 462, 598 hogares.
[15] De acuerdo a los datos publicados de Cuentas por Sector Institucional 2017 preliminar.

Tabla 2.6: *Saldos contables para los hogares*

Valores corrientes en millones de pesos

Código	Concepto	S.14
B.1b	Valor agregado bruto	7,199,239
B.2b	Excedente bruto de operación	1,657,121
B.3b	Ingreso mixto bruto	4,477,106
B.6b	Ingreso disponible bruto	16,279,769
B.8b	Ahorro bruto	2,801,261
B.9	Préstamo neto	1,376,844

Sistema de Cuentas Nacionales de México. Cifras preliminares 2017. Base 2013.
Fuente: elaboración propia basado en *(INEGI,2018c)*.

Cuentas Nacionales, es decir, el promedio por hogar es de 121, 625.93 pesos. De manera que el valor del ingreso reportado en Cuentas Nacionales es 2.61 veces equivalente al reportado en la ENIGH 2016.

Existen diversas investigaciones en las que se justifican las discrepancias entre ambas fuentes, entre dichas investigaciones se encuentra la realizada por *(Bustos, 2015)*, que refiere a la subestimación de los ingresos en la encuesta al omitir en ella a los hogares con ingresos altos, que lo denomina truncamiento, o bien a la falsa declaración de los hogares sobre su ingreso que lo define como subdeclaración y en el peor de los casos a la no respuesta. En este artículo el autor también propone un método para homologar la información de ambas fuentes y que este trabajo de investigación se retoma para aplicarlo a la información de la ENIGH 2016. A continuación se describe el método propuesto por Alfredo *(Bustos A., 2015)*.

2.2.3 Método de máxima pseudo verosimilitud restringida

El método propuesto toma en cuenta a los datos disponibles de la encuesta, el diseño de la muestra y la información de otras fuentes incorporándolas en forma de restricciones de los parámetros.
El ajuste de la distribución de probabilidad se realiza mediante el método de máxima verosimilitud considerando la función de log-verosimilitud de la distribución ajustada, $l(\theta; Y)$, con el peso de cada unidad mustreada $\frac{1}{\pi_i}$, y para considerar la conciliación con la información de otra fuente alternativa ésta se incorpora en forma de restricciones de los parámetros, $h(\theta) = c$, de manera que, la estimación de los parámetros se reduce a un problema de optimización

$$\max_{\theta,\lambda} \sum_{i=1}^{n} \frac{1}{\pi_i} l(\theta; Y_i)$$

$$\text{sujeto a } h(\theta) = c$$

Y sujeto también a las restricciones particulares de los parámetros de acuerdo a la distribución a considerar. Para determinar los óptimos del problema anterior se hace uso de la herramienta de los multiplicadores de Lagrange, de manera que la función a maximizar es

$$\max_{\theta,\lambda} \left[\sum_{i=1}^{n} \frac{1}{\pi_i} l(\theta; Y_i) - \lambda \left(h(\theta) - c \right) \right] \tag{2.1}$$

Donde:

- $l(\theta; Y_i)$ es el logaritmo natural de la función de densidad evaluada en el i-ésimo valor de la muestra.

- π_i es la probabilidad de inclusión para cada unidad de la muestra, cuyo inverso es llamado peso.

- $h(\theta)$ es una o más funciones del vector de parámetros, que determina la restricción a ser equivalente a la condición inicial c y también las restricciones propias de cada parámetro.

- λ es el multiplicador de Lagrange.

La expresión (2.1) refiere al método de máxima pseudo verosimilitud restringida.

2.2.4 Aplicación del método de máxima pseudo verosimilitud restringida

Mediante el método descrito en la sección anterior se estiman los parámetros de cada una de las cinco distribuciones propuestas, considerando como restricción que el valor esperado de la distribución, $E(Y|\theta)$ sea equivalente al nivel de ingreso trimestral promedio de Cuentas por Sector Institucional, que se estimó por $121,625.93$ pesos, es decir, la función de distribución propuesta se debe ajustar a que su media sea igual a esta cifra.

En la tabla 2.7 se muestran las estimaciones de los parámetros[16] de las cinco distribuciones propuestas restringidos a Cuentas por Sector Institucional, junto con las estimaciones de los parámetros de las distribuciones sin restricción, es decir solo del ingreso corriente trimestral de los hogares reportado en la ENIGH 2016.

Tabla 2.7: *Estimación de Parámetros por Máxima Pseudo-Verosimilitud*

	Distribución				
Parámetros	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
μ	10.85	121,626	106,211	57,270.6	22,537.1
σ	1.30	1.09	0.78	0.71	3.94
ν				1.53	0.56
τ					0.29

Estimación de Parámetros por Máxima Verosimilitud

	Distribución				
Parámetros	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
μ	10.36	44,221.54	47,529.57	31,141.07	31,027.89
σ	0.80	0.77	1.20	0.79	2.05
ν				-0.06	1.13
τ					1.10

Fuente: elaboración propia.

Existen diferencias muy marcadas entre los valores de los parámetros de las distribuciones restringidas y las distribuciones sin restricción en las distribuciones Gamma, Weibull y Gamma Generalizada como se muestra en la tabla 2.7, trayendo como consecuencia un cambio en la forma y escala de las distribuciones.

Utilizamos el estadístico Anderson-Darling para comparar el ajuste de cada una de las distribuciones propuestas, con el fin de determinar la distribución que mejor describa el ingreso corriente trimestral de los hogares considerando la restricción. El estadístico Anderson-Darling esta basado en la diferencia de cuadrados entre la distribución empírica y la distribución teórica propuesta, de manera que se considera un buen ajuste si el valor del estadístico es pequeño.

[16]La estimación de los parámetros con la restricción se realizó en R mediante la paquetería ALABAMA, que realiza el cálculo mediante métodos numéricos, considerando como valor inicial los valores de los parámetros de la distribución sin restricción calculados con la paquetería GAMLSS.

En la tabla 2.8 se muestran los valores del estadístico para cada una de las cinco distribuciones, dado que la distribución Beta Generalizada tipo 2 es aquella que tiene el menor valor del estadístico se concluye que es la que tiene un mejor ajuste.

Tabla 2.8: *Prueba de Bondad de Ajuste Anderson-Darling*

	Distribución				
Estadístico	Lognormal	Gamma	Weibull	Gamma G	Beta G2
Anderson-Darling	10,605.24	21,638.56	16,187.77	4,878.30	1,011.56

La decisión de considerar a la distribución Beta Generalizada también la sustenta el gráfico de la densidad versus la densidad empírica del ingreso corriente trimestral y los gráficos de las distribuciones que tienen un mejor ajuste de acuerdo con la tabla 2.8, las distribuciones Lognormal, Gamma Generalizada y Beta Generalizada comparadas con la distribución empírica, medidas en miles de pesos, ver figuras 2.10-2.13.

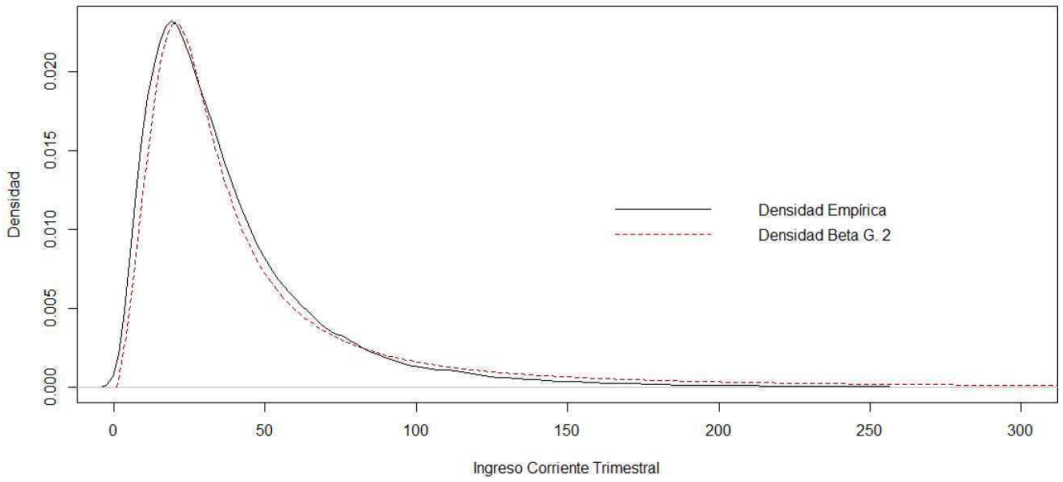

Figura 2.10: Ajuste de la densidad empírica del ingreso corriente trimestral a la densidad Beta Generalizada tipo 2, considerando la restricción a Cuentas Nacionales. Fuente: elaboración propia.

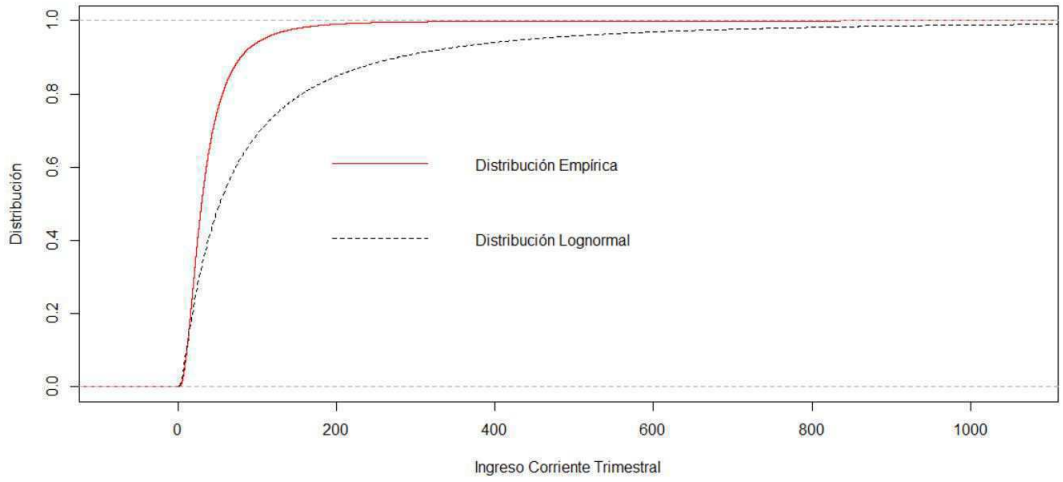

Figura 2.11: Ajuste de la distribución empírica del ingreso corriente trimestral a la distribución Lognormal, considerando la restricción a Cuentas Nacionales.
Fuente: elaboración propia.

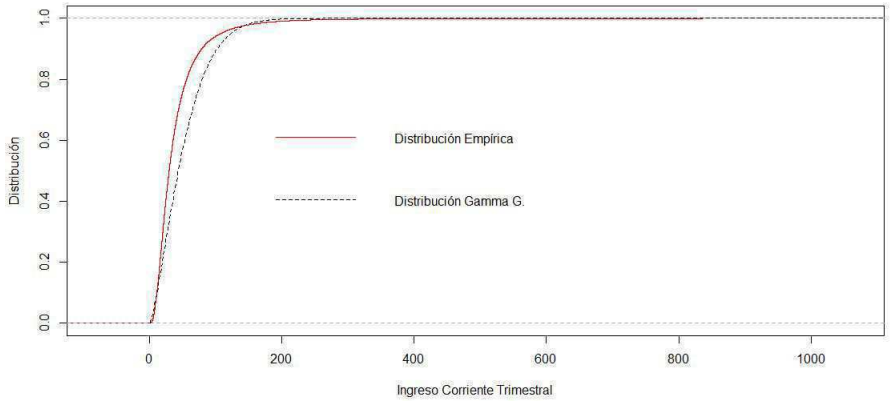

Figura 2.12: Ajuste de la distribución empírica del ingreso corriente trimestral a la distribución Gamma Generalizada, considerando la restricción a Cuentas Nacionales.
Fuente: elaboración propia.

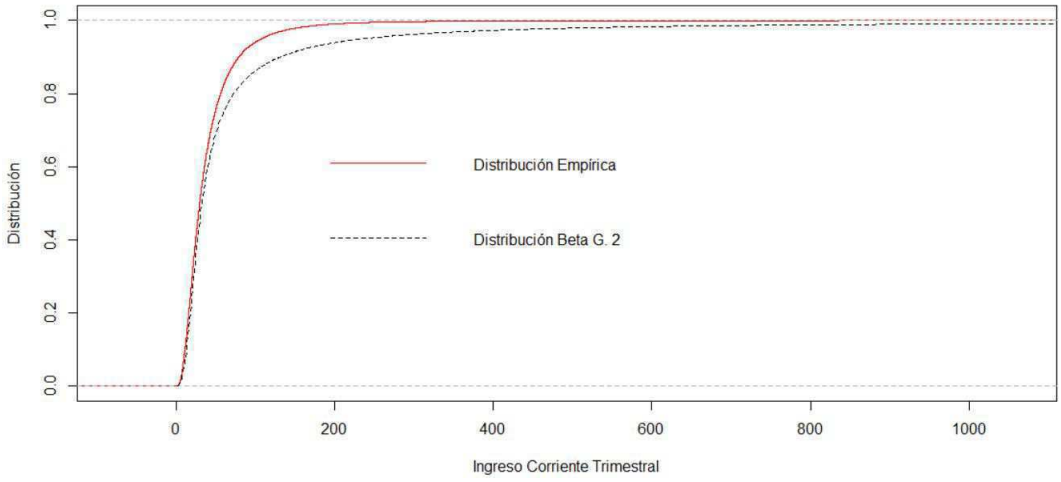

Figura 2.13: Ajuste de la distribución empírica del ingreso corriente trimestral a la distribución Beta Generalizada tipo 2, considerando la restricción a Cuentas Nacionales.
Fuente: elaboración propia.

Observemos que, a pesar de la restricción impuesta, las distribuciones Gamma Generalizada y Beta Generalizada son las que muestran un mejor ajuste, pues solo hay un mínimo desplazamiento de la distribución ajustada de la empírica. Sin embargo, la Beta Generalizada solo muestra diferencias a partir del sexto decil un sub reporte, es decir, en la encuesta se tiene una falsa declaración del ingreso percibido, pues este es mayor al reportado; a diferencia de la distribución Gamma Generalizada que presenta un sub reporte desde el segundo decil y después del noveno decil se presenta el fenómeno de truncamiento (ver figuras 2.12 y 2.13).

Por último, se determina el valor del coeficiente de Gini de cada una de las distribuciones ajustadas con restricción a la fuente de información de las Cuentas por Sector Institucional.

Tabla 2.9: *Desigualdad Económica*

	Distribución				
	Lognormal	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
Gini	0.646	0.535	0.585	0.395	0.780

En la tabla 2.9 se aprecia que el rango de los valores es más amplio, entre 0.39 y 0.78, comparado a los modelos sin restricción (ver tabla 2.4) y el valor de la distribución Beta Generalizada sin restricción (que fue de 0.433), es decir, considerando la fuente alterna de sectores institucionales, la desigualdad crece más que solo considerando la información de la ENIGH de 0.433 a 0.780.

De acuerdo a todo lo anterior, la distribución empírica del ingreso corriente trimestral de la ENIGH 2016 es ajustada mejor a la distribución Beta Generalizada tipo 2 y que a su vez concilia con la información de los hogares reportado en las Cuentas por Sector Institucional.

CAPÍTULO 3

CONCLUSIONES

Se presentó el método de máxima pseudo verosimilitud restringida para ajustar los datos de ingreso corriente trimestral a una distribución conocida, que utiliza información proveniente de encuestas y realiza la conciliación con otras fuentes de información. Esta metodología se implementó para los datos del ingreso corriente trimestral reportados en la ENIGH 2016 y se ajustó con la información de Cuentas Nacionales del mismo año. Se estimaron los parámetros de las cinco distribuciones propuestas y se consideraron las más adecuadas de acuerdo a su asimetría y curtosis. Fue de suma importancia contar con las herramientas estadísticas, y con el software estadístico R para la estimación de los parámetros de una manera más eficiente y sencilla.

Se mostraron los cambios en la estimación de los parámetros de las distribuciones propuestas para el ingreso corriente trimestral considerando una fuente de información alternativa como lo son las Cuentas por Sector Institucional, que si solo se toman los datos de la ENIGH 2016. Tras un proceso de estimación y posteriormente validación mediante ciertos criterios estadísticos: el criterio de AIC, y la prueba Anderson-Darling, se sustenta la elección de la distribución que mejor se ajusta a los datos de la encuesta ya sea con o sin restricción son las distribuciones Beta Generalizada tipo 2 (cuatro parámetros) y Gamma Generalizada (tres parámetros) respectivamente y cabe destacar que la distribución Gamma Generalizada anida a las otras tres distribuciones propuestas de dos parámetros y a su vez esta distribución está anidada en la distribución Beta Generalizada tipo 2, (ver figura 1.1).
En la figura 3.1 se muestran las dos densidades seleccionadas por el mejor ajuste que tiene el ingreso corriente trimestral, la densidad Gamma y Beta generalizada, esta última considera el valor de Cuentas Nacionales.

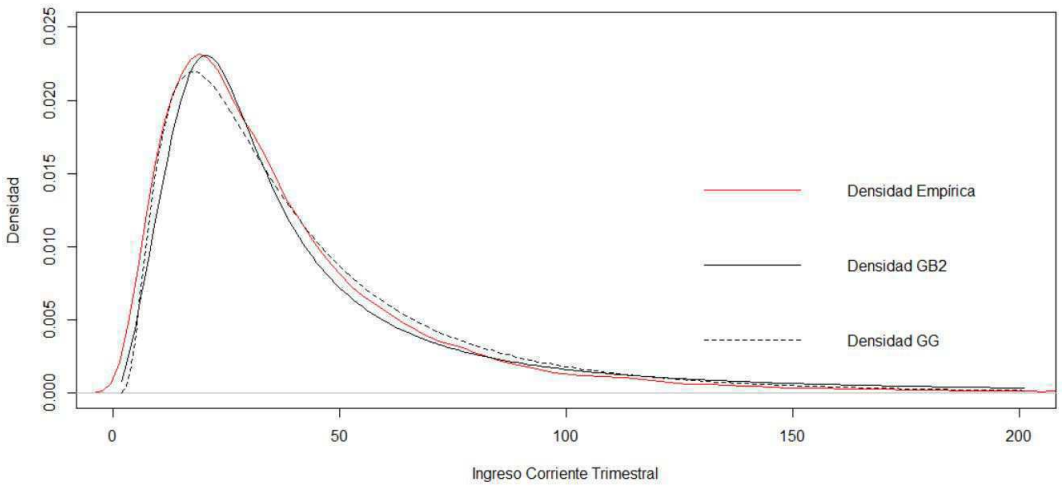

Figura 3.1: Ajuste de la densidad empírica del ingreso corriente trimestral a las densidades Gamma Generalizada (sin retricción) y Beta Generalizada tipo 2 (con restricción). Miles de pesos.
Fuente: elaboración propia.

Considerando las distribuciones Gamma Generalizada y Beta Generalizada tipo 2 como las distribuciones de mejor ajuste con y sin restricciones, se efectúa el cálculo del coeficiente de Gini, índice de suma importancia para la medición de la desigualdad económica.

Tabla 3.1: *Desigualdad Económica.*

Coeficiente de Gini	**2012**	**2016**
ENIGH	*0.44*	*0.448*
sin restricción	0.449	0.433
con restricción	0.802	0.780

Fuente: realización propia.

En la tabla 3.1 se resume el valor del coeficiente de Gini del año 2012 y 2016 con las densidades ajustadas. Mencionando que de acuerdo a *(Bustos, 2015)* los datos del ingreso corriente trimestral del año 2012 se ajustaron a la distribución Beta Generalizada con y sin restricciones, mientras que para el año 2016 se ajustó la distribución Gamma Generalizada y Beta Generalizada para la conciliación con otras fuentes. De acuerdo a los resultados reportados de la encuesta 2016, el nivel del índice se encuentra en 0.448, sin embargo, el modelo ajustado reduce este índice a 0.433, es decir, existe una menor desigualdad, pero este coeficiente aumentó aún más realizando la conciliación con la información proporcionada por Cuentas Nacionales, pues el nivel de este se estimó

en 0.780 y que a comparación de las medidas del año 2012 nos indica que el bienestar ha mejorado a través del tiempo, pues a pesar de que el valor del índice reportado en la ENIGH 2012 fue de 0.44, al ajustar el ingreso corriente trimestral distribución Beta Generalizada con y sin restricciones los niveles fueron de 0.802 y 0.449 respectivamente.

Observamos que en ambos años 2012 y 2016 se presentan comportamientos extremos, pues hay un abrupto crecimiento de la desigualdad en el modelo restringido a Cuentas por Sector Institucional, los gráfico 3.2 y 3.3 nos muestra la curva de Lorenz, que nos permite ver de manera gráfica los cambios de nivel de la desigualdad de los hogares considerando los modelos con y sin restricciones para ambos años.

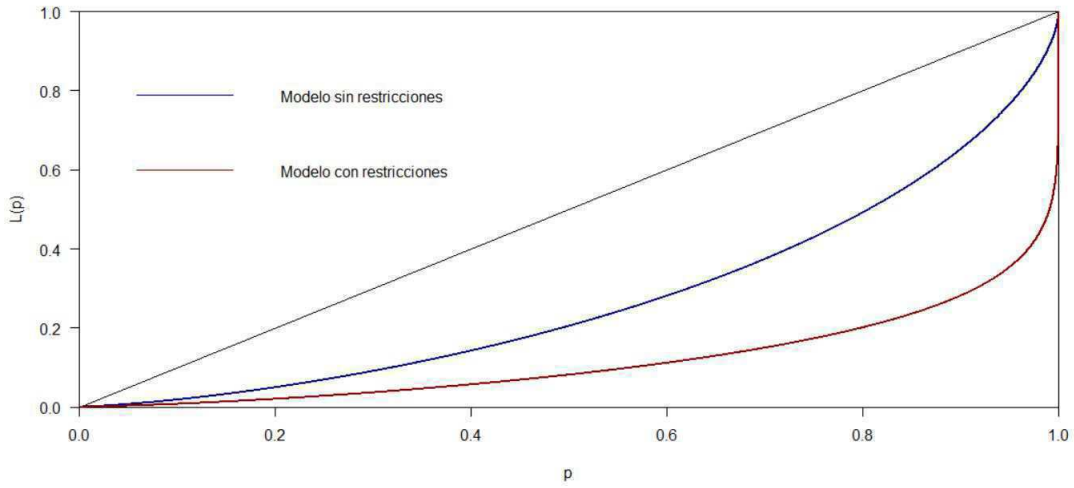

Figura 3.2: Comparación de la curva de Lorenz para el ingreso corriente trimestral con y sin restricciones del año 2012.
Fuente: elaboración propia.

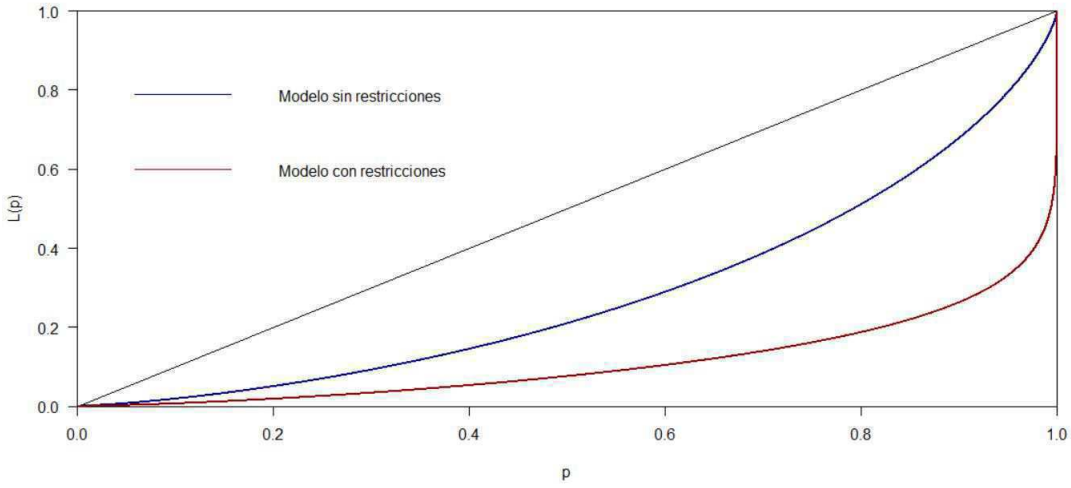

Figura 3.3: Comparación de la curva de Lorenz para el ingreso corriente trimestral con y sin restricciones del año 2016.
Fuente: elaboración propia.

Además, a través del tiempo la ENIGH y las Cuentas Nacionales presentan mejoras en los niveles de desigualdad de los hogares de México, pues los modelos del ajuste de la distribución con y sin restricciones muestran un valor en el Gini menor para el año 2016 como se muestra en los gráficos 3.4 y 3.5 a pesar de que el índice estimado de la encuesta era menor en la ENIGH 2012 que la presentada en el año 2016.

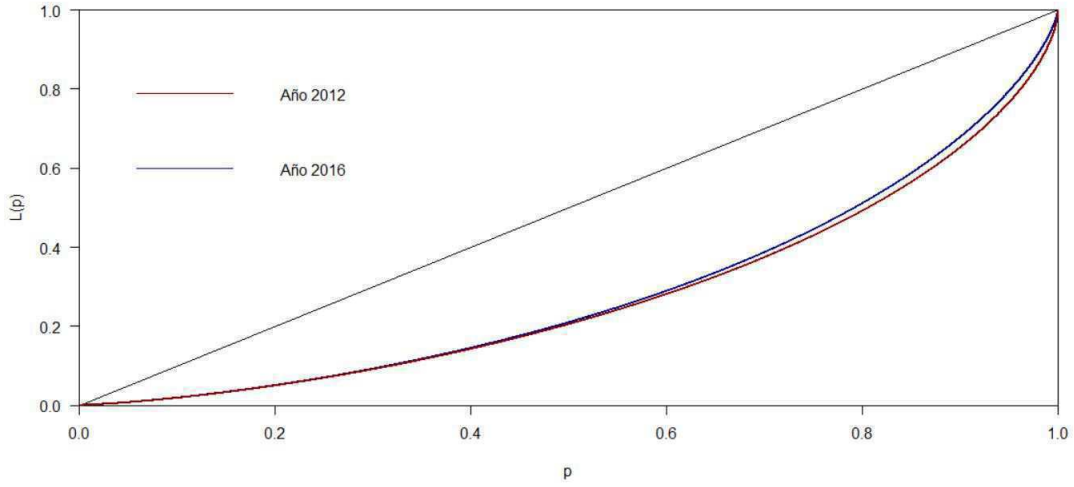

Figura 3.4: Comparación de la curva de Lorenz para el ingreso corriente trimestral año 2012 y 2016.
Fuente: elaboración propia.

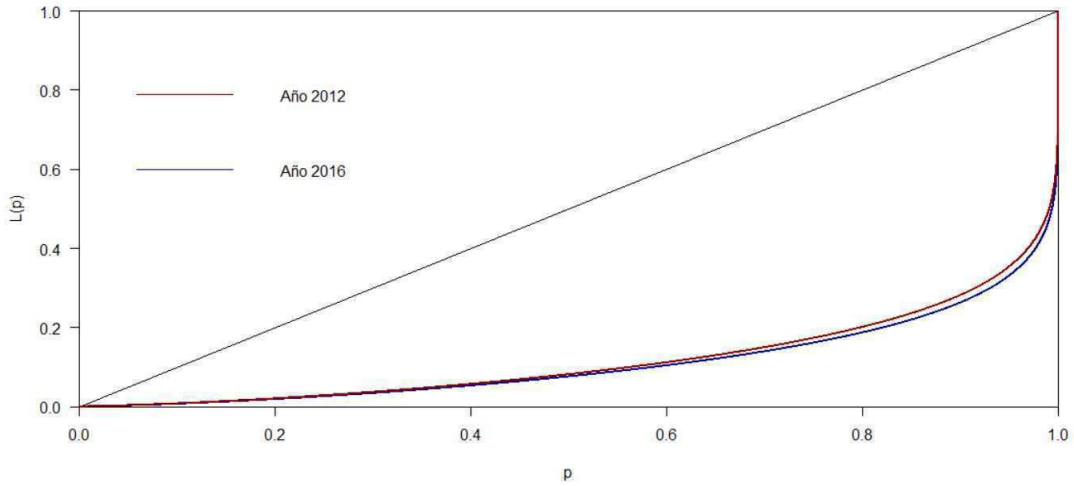

Figura 3.5: Comparación de la curva de Lorenz para el ingreso corriente trimestral considerando las Cuentas Nacionales año 2012 y 2016.
Fuente: elaboración propia.

Finalmente, se comparan las densidades ajustadas al ingreso corriente de los hogares en los años 2012 y 2016 con y sin restricciones, ver figuras 3.6 y 3.7.

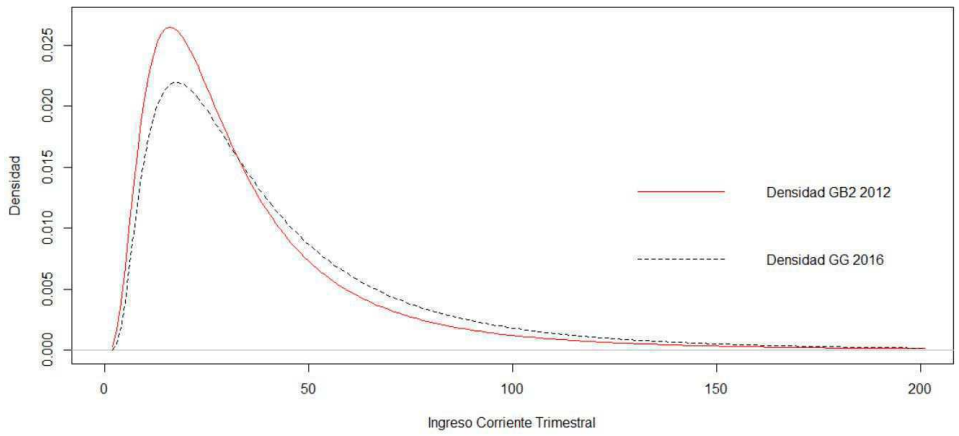

Figura 3.6: Comparación de las densidades seleccionadas para el año 2012 y 2016.
Fuente: elaboración propia.

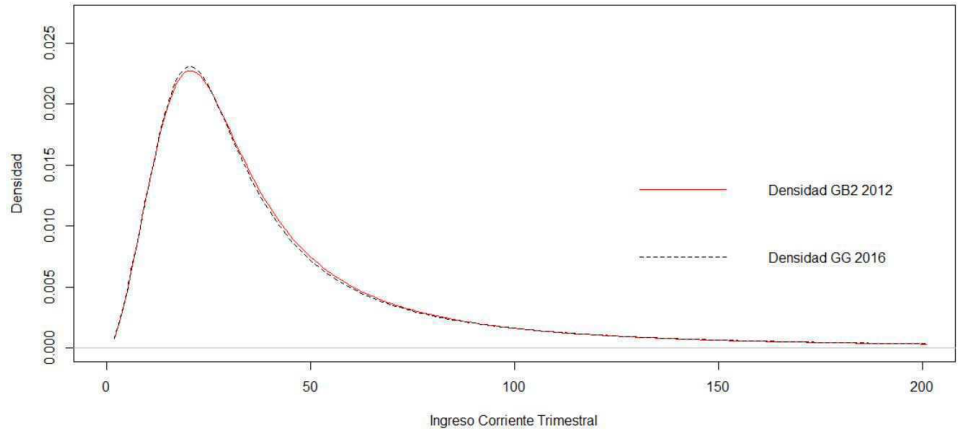

Figura 3.7: Comparación de las densidades seleccionadas para el año 2012 y 2016, considerando la retricción.
Fuente: elaboración propia.

Se observan los cambios en la escala y forma a través del tiempo, el modelo de cuatro parámetros GB2 ya no fue necesario para la ENIGH 2016, la Gamma Generalizada, modelo de tres parámetros, fue la distribución que mejor se adecuó a los datos del ingreso corriente no restringido. Sin embargo, para el caso restringido se mantuvo el modelo de cuatro parámetros GB2.

Se debe reconocer que en la aplicación de la metodología se tienen deficiencias en la forma de estimar el ingreso trimestral de los hogares de Cuentas Nacionales, ya que el supuesto fuerte de considerar que no existe estacionalidad, permitió hacer la conjetura que se presenta el mismo monto de ingreso disponible en los hogares en cada uno de los trimestres; en la realidad la actividad económica presenta mayores fluctuaciones en ciertos trimestres y que por tanto el valor del ingreso total de los hogares de $121,625.93$ no es un dato cien por ciento confiable, pero al no tenerse información de corto plazo de las Cuentas Nacionales, se recurrió a esta forma de estimación esperando que el dato real se aproxime al valor propuesto. Sin embargo, nos permitió ejemplificar el cómo conciliar información de distintas fuentes.

A pesar de este inconveniente la metodología que presenta A. Bustos es una alternativa para ajustar y conciliar las fuentes de información respecto del ingreso corriente de los hogares de manera óptima, teniendo como resultado información más apegada a la realidad de México.

APÉNDICE A

LOG-VEROSIMILITUD

La estimación de los parámetros de la distribución con y sin restricciones que ajustan al ingreso reportado en la ENIGH está basado en el método de máxima verosimilitud, por ello es necesario el conocer la función de log-verosimilitud, $l(y;\theta)$, así como la restricción que se está considerando, $h(\theta)$, para el caso del modelo con restricciones. A continuación, se especifica el método de máxima pseudo-verosimilitud.

A.1 Método de máxima pseudo verosimilitud restringida

El método es el siguiente

$$\max_{\theta,\lambda} \left[\sum_{i=1}^{n} \frac{1}{\pi_i} l(\theta; y_i) - \lambda \left(h(\theta) - c \right) \right]$$

Donde

- $l(\theta; y)$ es el logaritmo natural de la función de verosimilitud evaluada en el i-ésimo valor de la muestra.
- π_i es la probabilidad de inclusión para cada unidad de la muestra, cuyo inverso es llamado peso.
- $h(\theta)$ es una o más funciones del vector de parámetros, que determina la restricción a ser equivalente a la condición inicial c, y también las restricciones propias de cada parámetro. El valor inicial se determinó en $c = 121,625.93$.

- λ es el multiplicador de Lagrange.

A continuación se muestra la función de log-verosimilitud $l(\theta; Y_i)$ de cada una de las distribuciones propuestas para describir el ingreso corriente, de acuerdo a las definiciones presentadas en la sección 1.3, así también se menciona la restricción a considerar.

A.1.1 Log-verosimilitud de la distribución Lognormal

Dada una variable aleatoria $y \sim LOGNO(\mu, \sigma^2)$, la función de log-verosimilitud, $l(\theta; Y_i)$ es la siguiente

$$l(\theta; Y_i) = -n \log \sqrt{2\pi}\sigma - \sum_{i=1}^{n} \log y_i - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\log y_i - \mu)^2$$

Donde θ es el vector de parámetros $\theta = (\mu, \sigma^2)$.
Considerando como restricción la esperanza de la distribución

$$h(\theta) = E(Y; \theta) = exp\left(\mu + \frac{1}{2}\sigma^2\right)$$

A.1.2 Log-verosimilitud de la distribución Gamma

Dada una variable aleatoria $y \sim GA(\mu, \sigma)$, la función de log-verosimilitud, $l(\theta; Y_i)$ es la siguiente

$$l(\theta; Y_i) = \left(\frac{1}{\sigma^2} - 1\right) - \sum_{i=1}^{n} \log y_i - \frac{1}{2} \sum_{i=1}^{n} y_i - \frac{n}{\sigma^2} \log \sigma^2 \mu - n \log \Gamma\left(\frac{1}{\sigma^2}\right)$$

Donde θ es el vector de parámetros $\theta = (\mu, \sigma)$.
La restricción a considerar es la esperanza, que de acuerdo a esta parametrización la esperanza es

$$h(\theta) = E(Y; \theta) = \mu$$

A.1.3 Log-verosimilitud de la distribución Weibull

Dada una variable aleatoria $y \sim WEI(\mu, \sigma)$, la función de log-verosimilitud, $l(\theta; Y_i)$ es la siguiente

$$l(\theta; Y_i) = n\left(\log \sigma - \log \mu\right) + (\sigma - 1) \sum_{i=1}^{n} \log \frac{y_i}{\mu} - \sum_{i=1}^{n} \left(\frac{y_i}{\mu}\right)^{\sigma}$$

Donde θ es el vector de parámetros $\theta = (\mu, \sigma)$.
La restricción es la esperanza de la distribución

$$h(\theta) = E(Y;\theta) = \mu\,\Gamma\left(\frac{1}{\sigma} + 1\right)$$

A.1.4 Log-verosimilitud de la distribución Gamma Generalizada

Dada una variable aleatoria $y \sim GA(\mu, \sigma, \nu)$, la función de log-verosimilitud, $l(\theta; Y_i)$ es la siguiente

$$\begin{aligned} l(\theta;Y_i) &= n\,\gamma\,\log\gamma + \gamma\,\nu\sum_{i=1}^{n}\log\frac{y_i}{\mu} + n\,\log\nu - \gamma\sum_{i=1}^{n}\left(\frac{y_i}{\mu}\right)^{\nu} - n\,\log\Gamma(\gamma) \\ &- \sum_{i=1}^{n}\log y_i \end{aligned}$$

Donde $\gamma = \dfrac{1}{\sigma^2\,\nu^2}$, y θ es el vector de parámetros $\theta = (\mu, \sigma, \nu)$.
La restricción es la esperanza de la distribución

$$h(\theta) = E(Y;\theta) = \frac{\mu\,\Gamma\left(\gamma + \frac{1}{\nu}\right)}{\gamma^{\frac{1}{\nu}}\,\Gamma(\gamma)}$$

Donde $\Gamma(x)$ es la función gamma.

A.1.5 Log-verosimilitud de la distribución Beta Generalizada tipo 2

Dada una variable aleatoria $y \sim GB2(\mu, \sigma, \nu, \tau)$, la función de log-verosimilitud, $l(\theta; Y_i)$ es la siguiente

$$\begin{aligned} l(\theta;Y_i) &= \log|\sigma| + (\sigma\,\nu - 1)\sum_{i=1}^{n}\log y_i - n\,\sigma\,\nu\,\log\mu + n\,\log\beta(\nu,\tau) \\ &+ (\nu+\tau)\,\log\left(1 + \left(\frac{y_i}{\mu}\right)^{\sigma}\right) \end{aligned}$$

Donde θ es el vector de parámetros $\theta = (\mu, \sigma, \nu, \tau)$.
La restricción es la esperanza de la distribución

$$h(\theta) = E(y) = \mu\,\frac{\beta\left(\nu + \frac{1}{\sigma}, \tau - \frac{1}{\sigma}\right)}{\beta(\nu,\tau)}$$

Donde $\beta(\nu, \tau)$ es la función beta.

A.2 Estimación de Parámetros para submuestras de la ENIGH

En este trabajo de investigación se consideró una submuestra de la ENIGH, ya que no se puede tomar el total de la muestra de la encuesta $n = 70,311$ pues el modelo utilizado para la estimación de los parámetros toma en cuenta los logaritmos del ingreso, de manera que el modelo no está definido para los casos de ingresos, cero[1] sin embargo, se consideran submuestras de la muestra omitiendo el mínimo de observaciones.

A continuación, se muestran las estimaciones de los parámetros de cada una de las cinco distribuciones teóricas propuestas con y sin restricciones, así como los valores de los estadísticos Global Deviance, AIC, SBC y Anderson-Darling, y también el valor del coeficiente de Gini de dos submuestras.

1. **Submuestra 1** Total de observaciones consideradas en la submuestra: $n = 70,305$. De la muestra se omitieron los seis casos que reportaban un ingreso de cero, de esta manera que el logaritmo de los ingresos está definido para todas las observaciones y por tanto el modelo utilizado para la estimación de parámetros alcance su mínimo. Las tablas A.1 y A.2 muestran los resultados obtenidos de las estimaciones y los estadísticos, considerando la submustra antes descrita.

2. **Submuestra 2** Total de observaciones n=70,303. De la muestra se omitieron ocho observaciones: los 2 outliers de 16 y 35 millones y los seis casos que reportaban un ingreso de cero, de esta manera el modelo utilizado para la estimación alcance el mínimo y sea posible tener valores finitos de los parámetros. Las tablas A.3 y A.4 muestran los resultados de las estimaciones de los parámetros, los estadísticos y el coeficiente de Gini.

[1] En el modelo descrito en la paquetería GAMLSS no se alcanzaba el mínimo para los valores de ingreso nulo.

Tabla A.1: *Estimación de Parámetros por Máxima Verosimilitud.Submuestra 1*

	Distribución				
Parámetros	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
μ	10.36	46,527.44	47,007.28	30,607.06	30,625.04
σ	0.805	0.826	1.017	0.803	2.059
ν				-0.122	1.140
τ					1.085

Pruebas de Bondad de Ajuste

	Distribución				
Estadístico	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
GAIC	774,256,803	783,464,682	786,074,110	774,186,018	774,113,112
AIC	774,256,807	783,464,686	786,074,114	774,186,024	774,113,120
SBC	774,256,837	783,464,716	786,074,145	774,186,070	774,113,181

Desigualdad Económica

	Distribución				
Estadístico	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
Gini	0.4314	0.4291	0.4940	0.4392	0.4558

Fuente: elaboración propia.

Tabla A.2: *Estimación de Parámetros por Máxima Pseudo-Verosimilitud. Submuestra 1*

			Distribución		
Parámetros	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
μ	10.85	121,626	138,935.4	82,918.9	23,139.6
σ	1.30	1.105	0.64	0.8287	4.166
ν				1.283	0.519
τ					0.281

Prueba de Bondad de Ajuste

			Distribución		
Estadístico	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
Anderson-Darling	10,560.25	21,052.24	18,955.05	15,625.89	1,032.916

Desigualdad Económica

			Distribución		
Estadístico	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
Gini	0.647	0.539	0.6589	0.44	0.780

Fuente: elaboración propia.

Tabla A.3: *Estimación de Parámetros por Máxima Verosimilitud.Submuestra 2*

	Distribución				
Parámetros	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
μ	10.36	44,858.75	47,594.60	30,835.01	30,737.34
σ	0.803	0.789	1.150	0.802	2.058
ν				-0.098	1.139
τ					1.090

Pruebas de Bondad de Ajuste

	Distribución				
Estadístico	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
GAIC	774,036,124	779,689,764	782,249,029	773,993,824	773,981,789
AIC	774,036,128	779,689,768	782,249,033	773,993,830	773,981,797
SBC	774,036,159	779,689,799	782,249,064	773,993,876	773,981,858

Desigualdad Económica

	Distribución				
Estadístico	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
Gini	0.430	0.412	0.452	0.436	0.455

Fuente: elaboración propia.

Tabla A.4: *Estimación de Parámetros por Máxima Pseudo-Verosimilitud. Submuestra 2*

			Distribución		
Parámetros	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
μ	10.85	121,626	81,115.10	61,736	22,498.3
σ	1.30	1.096	0.746	0.730	3.92
ν				1.47	0.571
τ					0.298

Prueba de Bondad de Ajuste

			Distribución		
Estadístico	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
Anderson-Darling	10,572.56	21,477.05	10,220.77	6,949.833	1,003.117

Desigualdad Económica

			Distribución		
Estadístico	*Lognormal*	*Gamma*	*Weibull*	*Gamma G*	*Beta G2*
Gini	0.646	0.536	0.6054	0.401	0.7809

Fuente: elaboración propia.

APÉNDICE B

CÓDIGO

B.1.Ajuste del ingreso corriente trimestral de la ENIGH

A continuación, se muestra el código de R utilizado para la estimación de los parámetros de las distribuciones propuestas para ajustar el ingreso corriente reportado en la ENIGH. Mediante la librería GAMLSS, se hizo la estimación de los parámetros considerando el peso de cada observación muestreada para generalizar los resultados a toda la población.

La librería GAMLSS[1] refiere a un modelo aditivo generalizado para la escala, forma y localización, que es una clase de modelo estadístico para una variable univariada que modela los parámetros de una distribución como funciones de variables explicativas. Para la estimación se hace mediante el método de máxima verosimilitud y que utiliza el algoritmo de Newton-Raphson.

La implementación de la función *gamlss()* permite determinar la estimación de hasta cuatro parámetros de una familia de distribución y que convencionalmente se denominan μ, σ, ν, τ . A continuación se muestra el código utilizado para la estimación de los parámetros de la densidad lognormal.

[1]Generalised Additive Models for Location Scale and Shape.

```
# Carga de la base de datos
datos<-read.csv("C:/Desktop/TESINA/Datos/ingreso.csv")
```

Dentro de la base datos estan las variables ingreso corriente ing_cor y los pesos de cada observación "factor".

```
##### Librerias complementarias para la libreria GAMLSS

library(splines)
library(splines2)
library(MASS)
library(nlme)
library(parallel)
library(parallelDist)
library(gamlss.data)
library(gamlss.dist)
library(gamlss)

# Ajuste de una distribucion lognormal al ingreso corriente
#considerando los pesos

LOGN<-gamlss(ing_cor~1,family = LOGNO,data=datos,
weights = factors)
refit(LOGN) # reajuste de la funcion gamlss

# Para obtener los estimadores de los parametros como vector
theta1<-as.vector(c(fitted(LOGN,c("mu"))[1],fitted(LOGN,
c("sigma"))[1]))
theta1

#Resumen del ajuste de la funcion gamlss
summary(LOGN)

# Determina la Devianza del modelo ajustado a los datos del
#ingreso corriente
D1<-as.vector(c(deviance(LOGN,c("G"))[1],
    deviance(LOGN,c("P"))[1]))

# Determina el Akaike del modelo ajustado a los datos del
#ingreso corriente
```

```
IC(LOGN)
AIC(LOGN)
GAIC(LOGN)
```

El procedimiento es el mismo para las densidades Gamma, Weibull, Gamma Generalizada y Beta Generalizada susutituyendo la función LOGNO por GA, WEI, GG y GB2 respectivamente en la función *gamlss()*.

B.2.Ajuste del ingreso corriente trimestral de la ENIGH con restricciones

Para el caso de ajustar una distribución conocida a los datos de la encuesta conciliando con otra fuente de información, se utiliza la paquetería ALABAMA. La paquetería ALABAMA (*Argumented Lagrangian Adaptive Barrier Minimization Algorithm*) permite optimizar una función objetivo no lineal con restricciones, estas restricciones pueden ser del tipo igualdad, desigualdades lineales o no lineales.

Mediante la función *constrOptim.nl ()* se realiza la optimización de la función objetivo. En el caso de esta investigación, la función objetivo a optimizar es la log-verosimilitud teniendo en cuenta el peso de cada observación de la muestra, mientras que las restricciones son el valor esperado de la distribución propuesta igual a la condición inicial de $c = 121,625.93$, (este monto está justificado en el capítulo 2), en el anexo A se describieron de manera detallada la esperanza de cada distribución, y la otra restricción refiere a las propiedades mismas de los parámetros de cada una de las funciones de distribución a considerar. El método utilizado por ALABAMA para optimizar es del tipo numérico, de manera que se deben declarar los valores iniciales del algoritmo, en este caso los valores iniciales considerados son los valores estimados mediante GAMLSS del modelo no restringido.

A continuación, se muestra el código utilizado en la paquetería ALABAMA para obtener la estimación de los parámetros de la distribución Lognormal.

```
# Carga de la base de datos
datos<-read.csv("C:/Desktop/TESINA/Datos/ingreso.csv")

# Librerias necesarias para el ajuste con restricciones
library(numDeriv)
```

```
library(alabama)

# Definicion de la funcion a optimizar
flog<-function(x,d=ing_cor, w=factor)
{
  fl<- -1*w*dLOGNO(d,x[1],x[2], log=TRUE)
  sumfl<-sum(fl)
  return(sumfl)
}
# Definicion de la restriccion de los parametros de la
#distribucion

hin1<-function(x) #el parametro sigma debe ser positivo
{
  h<-rep(NA,2)
  h[1]<-x[1]
  h[2]<-x[2]
  h
}
# Definicion de la restriccion de la esperanza
# de la distribucion

ci<-121625.93 #condicion inicial, que el valor esperado sea ci

heq2<-function(x) #Esperanza de la distribucion
{
  h<-rep(NA,1)
  h[1]<-ci-(exp(x[1]+((1/2)*x[2]*x[2])))
  h
}

# Funcion que determina los optimos de la distribucion
# considerando restricciones
optimizarl<-constrOptim.nl(par=theta1, fn=flog, heq = heq2,
            hin = hin1)

# se toman como valores iniciales del algoritmo las estimaciones
# de los parametros hechas con GAMLSS,
# theta1 (vector de las estimaciones de los parametros)
```

Para el caso de las distribuciones Gamma, Weibull, Gamma Generalizada y Beta Generalizada se sustituye la función objetivo por su función de log-

verosimilitud y su esperanza correspondiente como restricción.
Se realizan la comparación de los gráficos de las densidades del año 2012 y 2016, considerando las estimaciones de los parámetros realizadas anteriormente y del 2016.

```
# Generacion de datos con los parametros del a\~no 2012
GB12<-rGB2(100000, mu=20394.4242,sigma =1.198802,nu=3.008914,
           tau=2.2948)

# Generacion de datos con los parametros del a\~no 2016
GB16<-rGB2(100000, mu=31027.888334,sigma =2.057317, nu=1.135232,
            tau=1.101878 )

#graficos juntos
plot(density(GB16),lty=1, main = "Densidad GB2 2016 Vs GB2 2012",
     xlim=c(0,500000), ylim=c(0.000001,0.00004))
lines(density(GB12), lty=2)
legend("right",legend =  c("2016","2012"), lty = c(1,2),
       bty="n",y.intersp = 0.3)
```

Adicionalemnte se comparan las densidades Gamma Generalizada (modelo sin restricciones), Beta Generalizada tipo 2 (modelo con restricción a Cuentas Nacionales) y la densidad empírica del ingreso corriente

```
plot(density(ing_cor),lty=1,col="red",
      main="Comparacion densidad 2012 Vs densidad 2016
      \n Con restriccion ",
      xlab="Ingreso corriente",ylab = "Densidad",
      xlim=c(0,100000), ylim=c(0,0.000028))

lines(GB2)
lines(GG2, lty=2, col="black")
legend(58000,0.000035,legend =  c("Empirica","GB2 2012","GG 2016"),
       lty = c(1,1,2),col=c("red","black","black"),
       bty="n",y.intersp = 0.3)
```

También se considera la generación del estadístico Anderson-Darling, para medir la bondad de ajuste de cada distribución, ya que en el output de la función *constrOptim.nl ()* no tiene un AIC o un SBC como en GAMLSS.

```
# Libreria necesaria para el estadistico Anderson-Darling
library(ADGofTest) ## ad.test

#### Generacion del dominio #######
x<-ing_cor

n<-length(x) # Tama\~no de la muestra

# Definicion de la funcion del estadistico

get.A<-function(x){
  #data<-as.data.frame(datos)
  #data<-stack(data)$ing_cor
  n<-length(x)
  DataO<-sort(x)
  i <- 1:n;
  A <- -n-(1/(n))*sum((2*i-1)*(log(DataO[i])+log(1-DataO[n+1-i])));
  return(A);
}

 # Generacion de datos con los parametros de la densidad lognormal
    sin restriccion

logn<-pLOGNO(x, mu=10.366707, sigma=0.8001593, lower.tail = TRUE,
      log.p = FALSE)

which(logn==1)#posision de los valores 1

prueba<-logn[logn<1] # omite los valores "1" para que este definida
# la funcion get.A (para que el logaritmo este definido,
# el dominio debe omitir el valor igual a uno)

get.A(prueba) # evaluacion de la funcion que define el estadistico

ad.test(prueba) #usando la libreria se obtienen los mismos valores


 ## Generacion de datos con los parametros de la densidad
 ## lognormal con restriccion
```

```
Logn2<-pLOGNO(x,mu = 10.8536,sigma = 1.30772, lower.tail = TRUE,
              log.p = FALSE )

# se omiten el valor uno de la densidad lognormal, para que
# la funcion logaritmo este definida
# y asi este definida la funcion del estadistico Anderson-Darling
Logn22<-Logn2[Logn2<1]
# evaluacion de la distribucion lognormal en funcion del estadistico
# Anderson-Darling
get.A(Logn22)

ad.test(Logn22) #usando la libreria se obtienen los mismos valores
```

B.3. Coeficiente de Gini

Finalmente se presenta el código que determina el valor del coeficiente de Gini de cada distribución, así como de la curva de Lorenz.

```
# libreria para el Coeficiente de Gini
library(ineq)

## generacion de los porcentiles
x<-seq(0.00001,0.99999,length.out=10000)

# Generacion de datos con los parametros de la densidad lognormal
# sin restricciones
logn16<-qLOGNO(x, mu=10.366707, sigma=0.8001593,
               lower.tail = TRUE, log.p = FALSE)

ineq(logn16, type = "Gini") # Coeficiente de Gini

# Generacion de datos con los parametros de la densidad lognormal
# con restricciones
Logn16<-qLOGNO(x,mu = 10.8536,sigma =1.30772,
               lower.tail = TRUE, log.p = FALSE)

ineq(Logn16, type = "Gini") # Coeficiente de Gini
```

```
# Graficos juntos de la curva de Lorenz
plot(Lc(logn16),col="darkblue",lwd=2, main = "Comparacion curva
     de Lorenz de los modelos con y sin restricciones")

lines(Lc(Logn16), col="darkred",lwd=2)
legend("bottomleft",legend =  c("Modelo sin restricciones",
       "Modelo con restricciones"), lty = 1,
        col = c("darkblue","darkred"), bty="n",y.intersp = 0.3)
```

También se hace la comparación de la desigualdad existente en los años 2012 y 2016 y se presenta graficamente mediante la Curva de Lorenz de los modelos que se ajustan al ingreso corriente con y sin restricción

```
# Generacion de datos con los parametros del a\~no 2012
GB12<-qGB2(x, mu=17175.97,sigma =3.25349,nu=0.7905,
          tau=0.36741, lower.tail = TRUE, log.p = FALSE )

ineq(GB12, type = "Gini") # Coeficiente de Gini

# Generacion de datos con los parametros del a\~no 2016
GB16<-qGB2(x, mu=22537.1,sigma =3.94677, nu=0.567117,
           tau=0.296904, lower.tail = TRUE, log.p = FALSE )

ineq(GB16, type = "Gini") # Coeficiente de Gini

#graficos juntos
plot(Lc(GB16),col="darkblue",lwd=2,
     main = "Comparacion curva de Lorenz 2016 Vs 2012
     \n Modelo restringido")
lines(Lc(GB12), col="darkred",lwd=2)
legend("bottomleft",legend =  c("2016","2012"), lty = 1,
        col = c("darkblue","darkred"),bty="n",y.intersp = 0.3)
```

BIBLIOGRAFÍA

[1] Bustos, A. (2015). *Estimation of the distribution of income from survey data, adjusting for compatibility with other sources.* Statistical Journal of the IAOS, vol.31, no. 4, pp. 565-577.

[2] Canavos, G. C. (1988). *Probabilidad y Estadística, Aplicaciones y Métodos.* México: McGraw-Hill.

[3] Casella, G., & Berger, R. (2002). *Statistical Inference.* Duxbury: Thomson Learning.

[4] CONEVAL. Consejo Nacional de la Evaluación de la Política de Desarrollo Social. (23 de Marzo de 2019.) *Pobreza en México, Medición de la Pobreza.* Obtenido de https://www.coneval.org.mx/Medicion/MP/Paginas/Pobreza_2016.aspx

[5] Fondo Monetario Internacional. (2016). *Sistema de Cuentas Nacionales 2008.* New York.

[6] Freund, J., Miller, I., & Miller, M. (2000). *Estadística Matemática con Aplicaciones.* New Jersey: Prentice Hall.

[7] Instituto Nacional de Estadística y Geografía (INEGI). (4 de Octubre de 2018a). *Encuesta Nacional de Ingreso y Gasto (ENIGH) 2016: Presentación de resultados.* Obtenido de www.inegi.org.mx/contenidos/programas/enigh/nc/2016/doc/ presentacion_resultados_enigh2016.pdf

[8] Instituto Nacional de Estadística y Geografía (INEGI). (21 de Noviembre de 2018b). *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2016 nueva serie: descripción de la base de datos.*

Obtenido de http://internet.contenidos.inegi.org.mx/contenidos/ Productos/prod_serv/contenidos/espanol/bvinegi/productos/ nueva_estruc/702825091972.pdf

[9] Instituto Nacional de Estadística y Geografía (INEGI). (14 de Diciembre de 2018c). *Sistema de Cuentas Nacionales de México. Fuentes y Metodologías. Cuentas por Sectores Institucionales. Año Base 2013.* Obtenido de https://www.inegi.org.mx/app/tmp/tabuladoscn/default.html?tema=CSI

[10] Instituto Nacional de Estadística y Geografía (INEGI). (13 de Octubre de 2018d). *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2016: nueva serie: diseño conceptual y definición de categorias y variables.* Obtenido de http://internet.contenidos.inegi.org.mx/contenidos/Productos/ prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/

[11] Marsaglia, G., & Marsaglia, J. W. (2004). *Evaluating the Anderson-Darling Distribution.* Journal of Statistical Software, Foundation for Open Access Statistics, vol.9 (i02).

[12] McDonald, J. B. (1984). *Some Generalized Function for the Size Distribution of Income.* Econometrica, 52 (3), pp. 647-64.

[13] Mood, A., Graybill, F., & Boes, D. (1974). *Introduction to the Theory of Statistics.* McGraw-Hill.

[14] Prieto Alaiz, M., & García Pérez, C. (2009). *La distribución beta generalizada de segunda especie como modelo de la distribución personal de la renta en España.* Estadística Española, 33-62.

[15] Rigby, B., Voudouris, V., Akantziliotou, C., Enea, M., & Kiose, D. (17 de Enero de 2019). *Generalised Additive Models for Location Scale and Shape. Package "gamlss".*

[16] Ross, S. (2010). *A First Course in Probability.* Prentice Hall/ Pearson.

[17] Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F. (2017). *Flexible Regression and Smoothing Using GAMLSS in R. Taylor & Francis Group.* USA: CRC Press.

[18] Becker, R. A., Chambers, J.M. & Wilks, A. R (1988). *Modern Applied Statistics with S.* Springer.

[19] Wackerly, D. D., Mendenhall III, W., & Scheaffer, R. L. (2008). *Estadística Matemática con Aplicaciones.* México: Cengage Learning Editores.

[20] R Development Core Team (2018). *R: A Lenguage and Environment for Statistical Computing.* Vienna, Austria. Obtenido de http://www.R-project.org/