Estadística con la Encuesta Nacional de Ingresos y Gastos de los Hogares y el paquete libre R

Statistics with the National Household Income and Expenditure Survey and the free R package

Revista Latinoamericana de Investigación Social, vol. 7, no.1

Miguel Heras Villanueva
Universidad Nacional
Autónoma de México
mikyheras@gmail.com
(correspondencia)

Juan Francisco Islas Aguirre
Instituto Politécnico Nacional
jfia3sur@gmail.com

Comunicado breve

Recibido: 13/03/2024
Aceptado: 29/04/2024
Fecha de publicación: 30/04/2024

Resumen
El análisis estadístico de un conjunto de datos generados a partir de una muestra, por medio de un paquete libre, en este caso R, es de vital importancia para los estudiantes de economía y ciencias sociales de hoy en día. En este sentido, el objetivo del presente trabajo es aportar herramientas replicables a otras muestras, por medio de rutinas asimilables por el usuario.
Otra de la finalidad del trabajo es establecer escenarios de análisis que permitan a los estudiantes forjar bases para el trabajo empírico, resaltando las virtudes del paquete libre R.

Palabras Clave: Microdatos; Bases de datos; R; Estadística.

Abstract
The statistical analysis of a data set generated from a sample, by means of a free package, in this case R, is of vital importance for current students of Economics and Social Sciences. So, the objective of this work is to provide replicable tools to other samples, through routines that can be assimilated by the user.
Another purpose of the work is to establish analysis scenarios that allow students to forge bases for empirical work, highlighting the virtues of the free R package.

Keywords: *Microdata; Databases; R; Statistics.*

Introducción

Hoy en día, los estudiantes de economía no solamente requieren aprender y comprender la génesis de dicha ciencia, la economía política, así como su conocimiento evolutivo, sino también una serie de herramientas virtuales para afrontar los desafíos de la era de la información, ligada estrechamente con lo digital y lo informático. La era de la información en sentido económico, se entiende a partir del manejo de grandes bancos de datos en la materia que, con la guía teórica correspondiente, permite al analista comprender la realidad en que se desenvuelve.

Sin embargo, surge la pregunta de cómo manipular tales bancos de datos. Si asumimos que el economista cuenta con un inventario de conocimientos teóricos que a partir de modelos caracterizan cierta realidad, la información económica sirve entonces para sostener o refutar a cierta teoría o en su caso, proponer nuevos bagajes epistémicos. La herramienta para la manipulación de los inventarios cuantitativos está vinculada con el uso de paquetes estadísticos, así como la posesión de conocimientos en esta última materia, para la adecuada manipulación de variables.

El presente trabajo tiene como finalidad que el alumno o profesional de la economía sin experiencia en el uso de algún paquete estadístico, se adentre en el espectro de los grandes bancos de datos, específicamente de la Encuesta Nacional de Ingresos y Gastos de los Hogares de México en su última versión y, su manipulación por medio del paquete R.[1]

Marco conceptual

La Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) que produce el Instituto Nacional de Estadística y Geografía (INEGI) en México, es el resultado de esfuerzos y acontecimientos nacionales e internacionales, entre los cuales se pueden mencionar a la Revolución Industrial, el Tratado de Versalles (que puso fin a la Primera Guerra Mundial y germinó la creación de organismos internacionales), la creación de la Sociedad de las

[1] *R* es un software libre que puede descargarse del sitio https://www.r-project.org/.

Naciones Unidas (eventualmente Naciones Unidas), el Tribunal Permanente de Justicia Internacional (después llamado Corte Internacional de Justicia) y la Organización Mundial del Trabajo.

Con esta última dio comienzo la conformación de encuestas a nivel mundial sobre ingresos y gastos, como insumo importante para medir los cambios en los precios de los artículos de consumo, así como la capacidad de compra de los salarios. En el ámbito nacional, en 1914 comenzaron los primeros esfuerzos con el objetivo de conocer los principales gastos de las familias obreras. Sin embargo, no fue sino hasta el año 1956, cuando la entonces Dirección General de Estadística (DGE) realizó la primera encuesta en este sentido, con métodos científicos de muestreo. Posteriormente, en 1958, la misma DGE levantó una nueva encuesta y más adelante, en 1960, se levantó otra en los principales centros urbanos del país.

En 1977, la DGE desarrolló la Encuesta Nacional de Ingresos y Gastos de los Hogares, llevada a cabo por el Instituto Nacional de Estadística y Geografía desde 1984 y, a partir de entonces, el INEGI tomó toda la responsabilidad para su elaboración. La ENIGH es una encuesta que, además de medir ingresos y gastos, identifica características ocupacionales y sociodemográficas de los hogares, así como la infraestructura de las viviendas y su equipamiento.

Cabe mencionar que, las encuestas de ingresos y gastos de los hogares en América Latina y el Caribe, buscan capturar el nivel y la estructura del gasto y el ingreso de los hogares. En este sentido, cabe recalcar que un hogar se define como un grupo de personas que comparten la misma vivienda, que juntan, total o parcialmente, su ingreso y su riqueza y que consumen colectivamente ciertos tipos de bienes y servicios, sobre todo los relativos a la alimentación y el alojamiento (Naciones Unidas, 2016).

Si bien pueden tener diversos propósitos y usos adicionales, por lo general las encuestas de ingresos y gastos de los hogares tienen como objetivo prioritario, servir de insumo para la actualización de la estructura y los ponderadores de la canasta de bienes y

servicios, utilizados para la medición del índice de precios al consumidor (IPC). En este sentido, los países que forman parte de la Organización de Cooperación y Desarrollo Económicos (OCDE), tienen el compromiso de actualizar la canasta del IPC y sus ponderadores al menos cada cinco años (CEPAL, 2021).

Por otro lado, la Ley General de Desarrollo Social (LGDS) mandata que el Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL), tiene la atribución de establecer los lineamientos y criterios generales para la definición, identificación y medición de la pobreza en México, garantizando la transparencia, objetividad y rigor técnico en dicha actividad. Para este fin, de acuerdo con los artículos 36 y 37 de la misma LGDS, el CONEVAL debe utilizar la información generada por el INEGI. Por lo tanto, se retoma la información de la ENIGH, como insumo para la medición multidimensional de la pobreza a nivel nacional y estatal (INEGI, 2022c).

Breve revisión de la literatura empírica

Si bien este trabajo sigue la pauta de lo mostrado por Heras e Islas (2015) en el afán de conformar plataformas virtuales de enseñanza-aprendizaje en los ámbitos mencionados, su objetivo fundamental es que el usuario comience a adquirir las herramientas necesarias para ejecutar y generar rutinas que lo adentren en el espectro del análisis o ciencia de datos.

Un ejemplo de ello es la descripción del cálculo de los principales indicadores y sus precisiones estadísticas (coeficiente de variación, error estándar e intervalos de confianza) de la ENIGH con el paquete R, por medio de códigos en el software, el cual es conformado por el INEGI (2022b).

Otra área de oportunidad se refiere a la comprensión de la medición de fenómenos sociales por medio del programa R y bases de datos, como es la medición multidimensional de la pobreza en México, tanto a nivel nacional, estatal y municipal, en el ámbito urbano y rural, así como en la población indígena, que está a disposición del público en general por el Consejo Nacional de Evaluación de la Política de Desarrollo Social en su portal de

internet (CONEVAL, 2024).

Con estas herramientas, por otro lado, la pretensión es que el usuario pueda adquirir las habilidades de desarrollar rutinas aplicables en diferentes ámbitos, como, por ejemplo, la producida por la Asociación Mexicana de Agencias de Inteligencia de Mercado y Opinión (AMAI), que se refiere a la producción de un modelo de evaluación y ratificación de la regla de Nivel socioeconómico, con revisiones regulares conforme se den a conocer nuevas ediciones de la ENIGH (AMAI, 2023).

También existen aplicaciones del uso de microdatos de la ENIGH con el paquete R. Un ejemplo de ello se encuentra en Márquez (2023), por medio de la cual se realiza un análisis de la inflación por deciles de ingreso durante 2020-2022 en el país. Si bien la investigación no incluye la rutina utilizada en el paquete para llegar a los resultados planteados, la idea es animar al usuario a realizar trabajos que enriquezcan la discusión en este u otros rubros.

Estructurar una base de datos

Antes de abordar la tarea, vale la pena mencionar que la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) producida por el Instituto Nacional de Estadística y Geografía de México (INEGI), tiene como objetivo proporcionar un panorama estadístico del comportamiento de los ingresos y gastos de los hogares en cuanto a su monto, procedencia y distribución. De la misma manera, ofrece información sobre las características ocupacionales y sociodemográficas de los integrantes del hogar, así como presentar datos sobre las características de la infraestructura de la vivienda y el equipamiento del hogar (INEGI, 2022c). Todo lo anterior tiene énfasis en el estudio particular de microdatos.

La base de datos de la ENIGH 2022 está conformada por 16 tablas de datos en las que se distribuye la información obtenida de la encuesta, de acuerdo con los temas más usados para realizar análisis y tabulados. Adicional a la base, se publica una tabla resumen con información a nivel hogar llamada CONCENTRADOHOGAR. Las tablas que

conforman la base de datos contienen información asociada a tres niveles o grupos; uno asociado a la vivienda, otro asociado al hogar y el último asociado al integrante del hogar (ENIGH, 2022a).

Con el fin de estructurar la base, se debe recurrir a los microdatos de la ENIGH 2022, ingresando al siguiente vínculo que forma parte del sitio del instituto:

https://www.inegi.org.mx/programas/enigh/nc/2022/#microdatos

En dicho sitio se debe dirigir a la parte inferior para descargar el archivo Principales variables por hogar en formato CSV. También se sugiere que el usuario descargue el documento intitulado Descriptor de archivos (FD) que se presenta en formato PDF, con la finalidad de que tenga de primera mano información útil referente a dicho inventario estadístico del instituto.

Para llevar a cabo las rutinas subsecuentes, se pide al lector que inaugure una carpeta intitulada *enigh_2022* en el disco duro de su equipo y, dentro de esta, coloque el archivo que se mencionó líneas arriba. Posteriormente, debe generar un directorio de trabajo en la unidad C del equipo de cómputo por medio de la función `setwd`.[2] Inmediatamente después, identificar los archivos que se localizan en dicho directorio por medio de la función `list.files`.

```
> setwd("C:/enigh_2022")
> list.files()
```

La conformación de la base de datos se realiza por medio de la función `read.csv` en el objeto[3] que se llamará `concentrado`, por medio de la simbología de asignación

[2] Con fuente `Courier New` se plasmarán tanto los resultados que arroja el Paquete en la consola de este, así como las funciones, objetos y nombres de variables. Las funciones, objetos y nombres de variables dentro del texto estarán además subrayados para su mejor identificación por parte del lector.

[3] Los entes sobre los cuales actúa el lenguaje de programación *R* se denominan objetos, los cuales son de tipo numérico, lógico y carácter.

(<−) como se presenta a continuación.[4]

```
> concentrado <- read.csv("concentradohogar.csv")
```

En caso de que el usuario no desee establecer un directorio de trabajo, puede conformar la base de datos por medio de la función y el objeto señalado, pero identificando la ruta de trabajo donde se localice el archivo que descargó de los microdatos de la ENIGH 2022 del INEGI. A continuación, se ejemplifica lo anterior a partir de la carpeta Documentos.

```
> concentrado<- read.csv("C:\\Documents\\concentradohogar.csv")
> concentrado <- read.csv("C:/Documents/concentradohogar.csv")
```

El lector debe vislumbrar que existen dos maneras de conformar la base de datos a partir de su inserción en el objeto `concentrado`. Una a partir de la duplicidad de las diagonales invertidas o bien, por medio del giro de la diagonal invertida. Posteriormente, debe acceder a la vista de las variables de la base por medio de la función `ls` sobre el argumento `concentrado`. Así mismo, a través de la función `attach` sobre el mismo argumento, será posible acceder a las variables de manera directa. Para visualizar el marco de datos se utiliza la función `View`.

```
> ls(concentrado)
> attach(concentrado)
> View(concentrado)
```

Ahora bien, surge la pregunta de cuántos hogares están contemplados en la muestra de la ENIGH 2022. De acuerdo con el INEGI (2022d), la muestra original constó de 105,525 viviendas seleccionadas y 1,368 hogares adicionales que fueron encontrados en dichas viviendas. Sin embargo, solamente en el 84.3% de los casos se obtuvieron entrevistas

[4] Si bien se presentarán las ejecuciones en el paquete después del símbolo conocido como *prompt*, se invita al lector a resguardar el ejercicio en un *script*, el cual es una bitácora de trabajo que puede guardarse y ejecutarse en cualquier momento. Para abrir, guardar y volver a ejecutar un script, el lector debe remitirse al menú archivo, en la barra de menú del paquete.

completas, es decir, se logró encuestar a 90,102 hogares. Lo anterior se logra vislumbrar por medio de la función `table` y la variable `foliohog`.[5]

```
> table(foliohog)
foliohog
    1     2     3     4     5
88823  1110   148    18     3
```

Variables categóricas

Una variable categórica es aquellas que, como su nombre lo indica, muestra categorías que pueden ser de tipo nominal u ordinal. El análisis de dichas variables es por medio de tablas de frecuencias absolutas y relativas, así como por medio de gráficas de barras y de pastel. A continuación, se realiza el análisis de frecuencias absolutas y relativas sobre la variable categórica llamada `sexo_jefe`, cuyas etiquetas indican que 1 es hombre y 2 es mujer.

```
> table(sexo_jefe)
sexo_jefe
    1     2
61905   28197
> table(sexo_jefe)/nrow(concentrado)
sexo_jefe
        1         2
0.6870547    0.3129453
```

Se muestra que 61,905 jefes de la muestra son hombres, lo que representa al 68.7% de la misma, mientras que 28,197 son mujeres con una proporción del 31.3%. A continuación, se lleva a cabo una transformación de la variable en cuestión, ya que R solamente realiza gráficas para variables categóricas cuando estas son de tipo factor. Para ello, se toma ventaja de la generación de la nueva variable para asignar categorías H y M para los sexos hombre y mujer respectivamente. Es importante recalcar que las anotaciones

[5] La variable `foliohog` tiene como función ser el identificador del hogar, la cual contiene 5 códigos. El 1 identifica al hogar principal y del 2 al 5 los hogares adicionales (INEGI, 2022a).

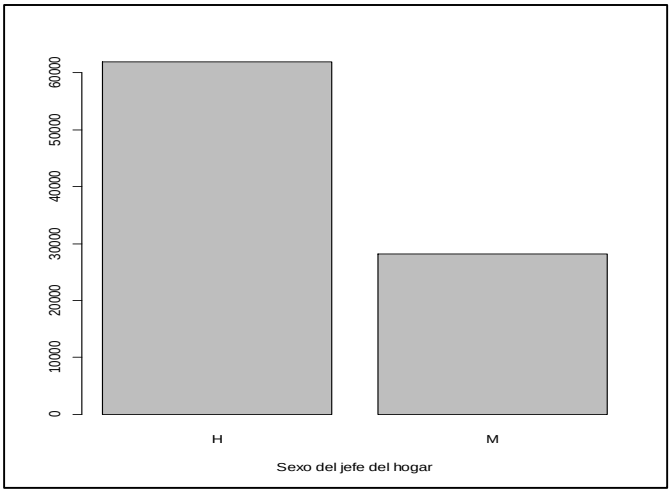
las percibe el paquete con el símbolo de número, las cuales están incorporadas después de cada ejecución.

```
> fsexo_jefe <- factor (sexo_jefe, levels=1:2)
> # Generamos el objeto fsexo_jefe tipo factor
> levels(fsexo_jefe) <-c('H','M')
> # Asignamos las categorías con la función concatenar[6]
> table(fsexo_jefe)
> # Generamos una nueva tabla de frecuencias absolutas
fsexo_jefe
    H     M
61905   28197
> plot(fsexo_jefe, main="Gráfica de barras", xlab="Sexo del jefe del
hogar", sub="Fuente: ENIGH, 2022")
> # Graficamos con atributos en la gráfica
```

Cabe recalcar que las categorías se presentan con la función `levels`, primero en forma de secuencia y después, como función del objeto factor recién generado. Por su parte, las categorías están comprendidas por apóstrofes. Respecto a la gráfica, se realiza por medio de la función `plot` y una serie de atributos, los cuales están separados por comas y las leyendas por comillas, la cual se muestra en la figura 1.

Figura 1.
Sexo del jefe del hogar

[6] Es posible abreviar la generación del objeto `fsexo_jefe` de dos a una línea por medio de la incorporación de la función `labels`, la cual permite etiquetar a las categorías, tal y como se muestra a continuación: `fsexo_jefe <- factor(sexo_jefe, levels=c(1:2), labels=c('H','M'))`



Nota: Elaboración propia con base en la ENIGH 2022 (INEGI, 2022e)

La obtención de información por entidad es una de las virtudes de la ENIGH. El ejercicio que se muestra a continuación debe servir como guía para que el usuario ejecute tareas similares sobre cualquier variable categórica de la encuesta. Lo primero que hacemos es generar al objeto `entidad`, a partir de asignar un formato a la variable `ubica_geo` de la base, el cual consta de un ancho de 9 dígitos y justificado a la derecha (también puede ser a la izquierda).

Los datos del objeto `entidad` son los mismos de la variable `ubica_geo`, pero con un formato que será útil para extraer las submuestras por entidad generando el objeto `ent`. La función `table` sobre el argumento `ent`, permite visualizar lo anterior.

Finalmente, hacemos uso de la misma función con dos argumentos, siendo el primero el objeto tipo factor y el segundo, una función que permite replicar la tabla por entidad federativa. La función `sum` por su parte, permite corroborar que la adición de las submuestras equivale al número total de hogares en la encuesta.

```
> entidad <- format(ubica_geo, digits = 9, justify = "right")
> ent=substr(entidad,1,2)
> table(ent)
ent
   1    2    3    4    5    6    7    8    9   10   11   12
```

```
2669 4141 2776 2204 4125 3031 2112 4555 2585 2713 3076 2538
13   14   15   16   17   18   19   20   21   22   23   24
2135 2625 3527 2215 2461 2130 3612 2620 2175 3822 2418 2630
25   26   27   28   29   30   31   32
3479 2548 2167 2322 2307 2915 2947 2522
> sum(table(ent))
90102
```

A continuación, se replica el ejercicio, pero incorporando el sexo del jefe o jefa del hogar. Para corroborar, se usa de nuevo la función `sum`.

```
> table(fsexo_jefe, by=ent)
by
fsexo_jefe    1    2    3    4    5    6    7    8     9    10   11   12
         H  1863 2860 1902 1512 2975 2054 1526 3173 1560 1909 2139 1686
         M  806 1281  874  692 1150  977  586 1382 1025  804  937  852
fsexo_jefe    13   14   15   16   17   18   19   20   21   22   23   24
         H   1462 1837 2460 1472 1531 1400 2699 1801 1456 2566 1681 1785
         M    673  788 1067  743  930  730  913  819  719 1256  737  845
fsexo_jefe    25    26   27   28   29   30   31   32
         H   2248 1723 1502 1586 1629 1995 2131 1782
         M   1231  825  665  736  678  920  816  740
> sum(table(fsexo_jefe, by=ent))
90102
```

Es posible etiquetar a los números que representan a cada una de las entidades del país con sus nombres y de esta forma, presentar la tabla de una forma más amigable. Solamente se requiere adicionar la función `labels` a las 32 categorías asignadas al objeto `ent`, que a su vez proviene de la variable `ubica_geo`. Se advierte al usuario que debe justificar los renglones como se muestra a continuación. Finalmente, se vuelven a sumar los valores de la tabla.

```
> enti <- factor(ent, labels=c("AGS", "BC", "BCS", "CAMP", "COAH", "COL",
         "CHI",  "CHIH",  "CDMX",  "DUR",  "GUA",  "GUE",  "HGO",  "JAL",
         "EDOMEX","MICH", "MOR", "NAY", "NL", "OAX", "PUE", "QRO", "QROO",
```

```
         "SLP", "SIN", "SON", "TAB", "TAM", "TLAX", "VER", "YUC", "ZAC"))
> table(fsexo_jefe, by=enti)
     by
fsexo_jefe  AGS   BC  BCS CAMP COAH  COL  CHI CHIH CDMX  DUR  GUA  GUE
         H 1863 2860 1902 1512 2975 2054 1526 3173 1560 1909 2139 1686
         M  806 1281  874  692 1150  977  586 1382 1025  804  937  852
by
fsexo_jefe  HGO  JAL EDOMEX MICH  MOR  NAY   NL  OAX PUE  QRO QROO SLP
         H 1462 1837   2460 1472 1531 1400 2699 1801 1456 2566 1681 1785
         M  673  788   1067  743  930  730  913  819  719 1256  737  845
by
fsexo_jefe  SIN  SON  TAB  TAM TLAX  VER  YUC  ZAC
         H 2248 1723 1502 1586 1629 1995 2131 1782
         M 1231  825  665  736  678  920  816  740
> sum(table(fsexo_jefe, by=enti))
90102
```

Variables numéricas

Las variables numéricas muestran cantidades que varían de persona a persona o de objeto a objeto. Cuando ellas son finitas o se basan en un conteo, se denominan discretas, mientras que cuando toman un número infinito de valores en un intervalo o bien, provienen de una medición, son llamadas continuas. El análisis descriptivo de dichas variables descansa en una gráfica de barras conocida como histograma, y en una serie de medidas, las cuales se clasifican en tendencia central, dispersión, forma y posición.

Ciertamente es útil reconocer que el histograma es una forma abreviada de la tabla de distribución de frecuencias, la cual contempla intervalos y un valor característico por intervalo conocido como marca de clase. Esta forma de agrupación de los datos de alguna variable numérica permite visualizar gráficamente el comportamiento de ella, aunque existe el debate en el sentido que se pierde cierta información al momento de calcular los estadísticos muestrales.[7]

[7] Los estadísticos son valores que describen a una muestra y se denotan por medio de letras romanas. Los parámetros son valores que describen a una población y se denotan por medio de letras griegas. En términos generales, la estadística se divide en dos grandes ramas, la descriptiva y la inferencial. Por medio de esta última

Los estadísticos por variable en R se obtienen de objetos numéricos. La forma de obtener un análisis somero de la variable `edad_jefe` es por medio de la función `summary`.

Si el usuario desea un análisis minucioso, debe utilizar funciones específicas como se muestra a continuación.

```
> summary(edad_jefe)
  Min.   1st Qu.  Median    Mean   3rd Qu.    Max.
 13.00   39.00    50.00    51.23   63.00     109.00
> mean(edad_jefe)
[1] 51.23426
> median(edad_jefe)
[1] 50
> sd(edad_jefe)
[1] 15.91404
> IQR(edad_jefe)
[1] 24
```

Hasta este momento y respecto de los estadísticos recién obtenidos, es necesario realizar algunas observaciones. En primer término, los valores de las medidas de tendencia central son parecidos, lo que denota que la muestra es aproximadamente normal. En segundo término, se muestra que el 50% de los jefes de hogar tienen una edad entre los 39 y 63 años, cuya diferencia, 24, es el rango intercuartil. En tercer término, si bien estrictamente la forma de la distribución es platicúrtica como lo muestra el valor de la curtosis, se aproxima al valor de la forma mesocúrtica; el valor del sesgo dilucida que la distribución está ligeramente sesgada de forma positiva o a la derecha.

Cabe aclarar que R, por ser un paquete libre, no contiene toda la paquetería cargada de manera automática. En otras palabras, para ejecutar ciertas funciones es necesario instalar

es posible estimar parámetros desconocidos de una población, como la media o la desviación estándar. Cuando la estimación toma valores puntuales se denomina estimación puntual, la cual es precisa en forma de estadísticos. De esta manera, se dice que la media muestral es un estimador puntual de la media poblacional. Cuando la estimación toma valores en un intervalo, se denomina estimación por intervalo. Estas estimaciones son válidas y esenciales para realizar pruebas de hipótesis y regresiones.

archivos de la red, los cuales para volver a ser utilizados deben ser cargados por el usuario. Lo primero se realiza por medio de la función `install.packages` y lo segundo por medio de la función `library`. En este momento se debe instalar la paquetería `moments`, por medio de la cual es posible calcular el sesgo y la curtosis de cierta distribución, precisamente por medio de la metodología de momentos.

```
> install.packages("moments")
> library(moments)
> skewness(edad_jefe)
[1] 0.2535206
> kurtosis(edad_jefe)
[1] 2.442733
```

Por lo que respecta a la expresión algebraica de esta metodología, el paquete realiza el cálculo del sesgo (ecuación 1) y la curtosis (ecuación 2) de la distribución basado en una población. Si definimos a SK como coeficiente de sesgo y K como curtosis, los numeradores en las siguientes expresiones calculan el tercer y cuarto momento respecto a la media, mientras que los denominadores a la desviación estándar poblacional elevada a la tercera y cuarta potencia respectivamente.

$$SK = \frac{\frac{\sum(X_i - \bar{X})^3}{n}}{\left[\sqrt{\frac{\sum(X_i - \bar{X})^2}{n}}\right]^3} \tag{1}$$

$$K = \frac{\frac{\sum(X_i - \bar{X})^4}{n}}{\left[\sqrt{\frac{\sum(X_i - \bar{X})^2}{n}}\right]^4} \tag{2}$$

Por su parte, la desviación estándar señala que en promedio las edades se desvían de la media aritmética prácticamente 16 años. Por otro lado, es posible realizar la misma rutina para cada una de las entidades de la República por medio de la función `tapply`, la cual genera tablas de acuerdo con el último argumento, el cual puede reportar un valor específico o una serie de valores como se muestra en seguida de manera abreviada por cuestiones de

espacio.

```
> tapply(edad_jefe, enti, summary)
$AGS
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   38.00   49.00   50.13   61.00   96.00
$BC
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.00   38.00   48.00   49.24   60.00  109.00
$BCS
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   36.00   48.00   49.02   60.00   97.00
$CAMP
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   38.00   49.00   49.56   60.00   95.00
$COAH
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   39.00   50.00   50.97   62.00  100.00
$COL
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.00   39.00   51.00   51.56   63.00  102.00 …
…
…
…
$TAM
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   39.00   51.00   51.59   63.00   97.00
$TLAX
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   39.00   50.00   50.95   62.00  100.00
$VER
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   40.00   52.00   52.87   64.00  104.00
$YUC
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   38.50   49.00   50.87   63.00  100.00
$ZAC
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

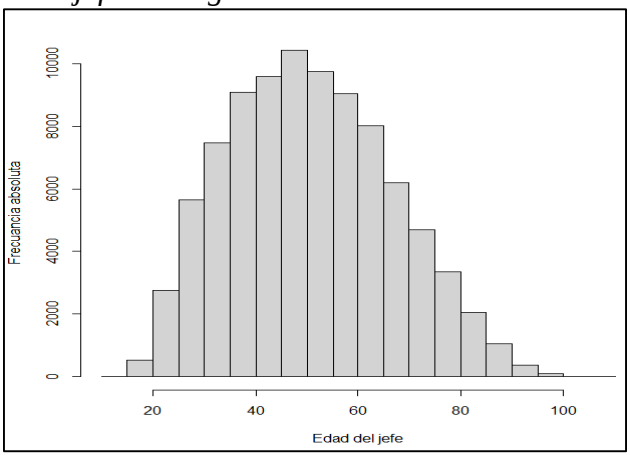
```
  18.00   40.00   51.00   52.06   64.00  101.00
```

Respecto al histograma, existen dos formas de elaborarlo. De querer solamente la forma de la distribución, basta con hacer uso de la función `hist` al objeto numérico del cual se desea el histograma. Sin embargo, existe otra forma mucho más atractiva de elaborarlo y es por medio de asignar un objeto a la función, ya que en este último se guardarán las características de la distribución que el paquete calcula.

Enseguida se genera el objeto `hedad_jefe` al que se le asignará la función `hist`. Note el lector que la función posee una serie de argumentos, siendo el primero el nombre de la variable `edad_jefe` y detrás de ella, el título, nombres de los ejes y la fuente de la información. Con la función `help` y como argumento `hist`, es decir, `help(hist)` es posible consultar la manera de conformar con mayor detalle la gráfica de la distribución. Como se aprecia en la figura 2.

```
hedad_jefe <- hist(edad_jefe, main="Histograma", xlab="Edad del jefe",
sub="Fuente: ENIGH, 2022", ylab="Frecuancia absoluta")
```

Figura 2.
Histograma de la edad del jefe del hogar



Nota: Elaboración propia con base en la ENIGH 2022 (INEGI, 2022e)

Las características del histograma se visualizan por medio de invocar al objeto que las resguarda. Vale la pena mencionar que la delimitación de los intervalos se observa en

primer plano por medio del apartado *breaks*, la frecuencia absoluta por counts, los valores de densidad por density y la marca de clase por mids, entre los rasgos más sobresalientes. Es importante señalar que una de las virtudes de R es la capacidad que posee para mostrar varias gráficas de manera simultánea.

```
> hedad_jefe
$breaks
  [1]  10  15  20  25  30  35  40  45  50  55  60  65  70  75  80  85  90
  95 100 105 110
$counts
  [1]      3    536   2788   5759   7262   9087   9841 10609   9417   8767   7249
  5992   4660   3400   2105   1097    332     93    8     1
$density
  [1]  6.741119e-06  1.204413e-03  6.264746e-03  1.294070e-02  1.631800e-02
  2.041885e-02 2.211312e-02 2.383884e-02
  [9]  2.116037e-02  1.969980e-02  1.628879e-02  1.346426e-02  1.047120e-02
  7.639934e-03 4.730018e-03 2.465002e-03
  [17] 7.460171e-04 2.089747e-04 1.797632e-05 2.247040e-06
$mids
  [1]  12.5  17.5  22.5  27.5  32.5  37.5  42.5  47.5  52.5  57.5  62.5
  67.5  72.5  77.5  82.5  87.5  92.5  97.5
  [19] 102.5 107.5
$xname
  [1] "edad_jefe"
$equidist
  [1] TRUE
attr(,"class")
[1] "histogram"
```

Para ejemplificar lo anterior, se presentan los histogramas de las 32 entidades de la Federación. En primer término, se deben generar 32 objetos que contengan, cada uno, los datos de las edades de los jefes del hogar por entidad federativa. La lista de los objetos se presenta íntegra, con la finalidad de que el usuario pueda replicar el ejercicio con la mayor facilidad posible.

```
edadj_ags    <- edad_jefe[enti=="AGS"]
```

```
edadj_bc     <- edad_jefe[enti=="BC"]
edadj_bcs    <- edad_jefe[enti=="BCS"]
edadj_camp   <- edad_jefe[enti=="CAMP"]
edadj_coah   <- edad_jefe[enti=="COAH"]
edadj_col    <- edad_jefe[enti=="COL"]
edadj_chi    <- edad_jefe[enti=="CHI"]
edadj_chih   <- edad_jefe[enti=="CHIH"]
edadj_cdmx   <- edad_jefe[enti=="CDMX"]
edadj_dur    <- edad_jefe[enti=="DUR"]
edadj_gua    <- edad_jefe[enti=="GUA"]
edadj_gue    <- edad_jefe[enti=="GUE"]
edadj_hgo    <- edad_jefe[enti=="HGO"]
edadj_jal    <- edad_jefe[enti=="JAL"]
edadj_edomex <- edad_jefe[enti=="EDOMEX"]
edadj_mich   <- edad_jefe[enti=="MICH"]
edadj_mor    <- edad_jefe[enti=="MOR"]
edadj_nay    <- edad_jefe[enti=="NAY"]
edadj_nl     <- edad_jefe[enti=="NL"]
edadj_oax    <- edad_jefe[enti=="OAX"]
edadj_pue    <- edad_jefe[enti=="PUE"]
edadj_qro    <- edad_jefe[enti=="QRO"]
edadj_qroo   <- edad_jefe[enti=="QROO"]
edadj_slp    <- edad_jefe[enti=="SLP"]
edadj_sin    <- edad_jefe[enti=="SIN"]
edadj_son    <- edad_jefe[enti=="SON"]
edadj_tab    <- edad_jefe[enti=="TAB"]
edadj_tam    <- edad_jefe[enti=="TAM"]
edadj_tlax   <- edad_jefe[enti=="TLAX"]
edadj_ver    <- edad_jefe[enti=="VER"]
edadj_yuc    <- edad_jefe[enti=="YUC"]
edadj_zac    <- edad_jefe[enti=="ZAC"]
```

Posteriormente, es necesario preparar una hoja dividida en 32 partes iguales. Ello se logra por medio de la función `mfrow`. Después, generar 32 objetos, para que cada uno de ellos resguarde la información de los histogramas. Se sugiere que el usuario incorpore como argumentos de cada función, el nombre de la entidad como título, y el título del eje de las abscisas. Como anteriormente se mencionó, cada uno de los contenidos de los objetos de los

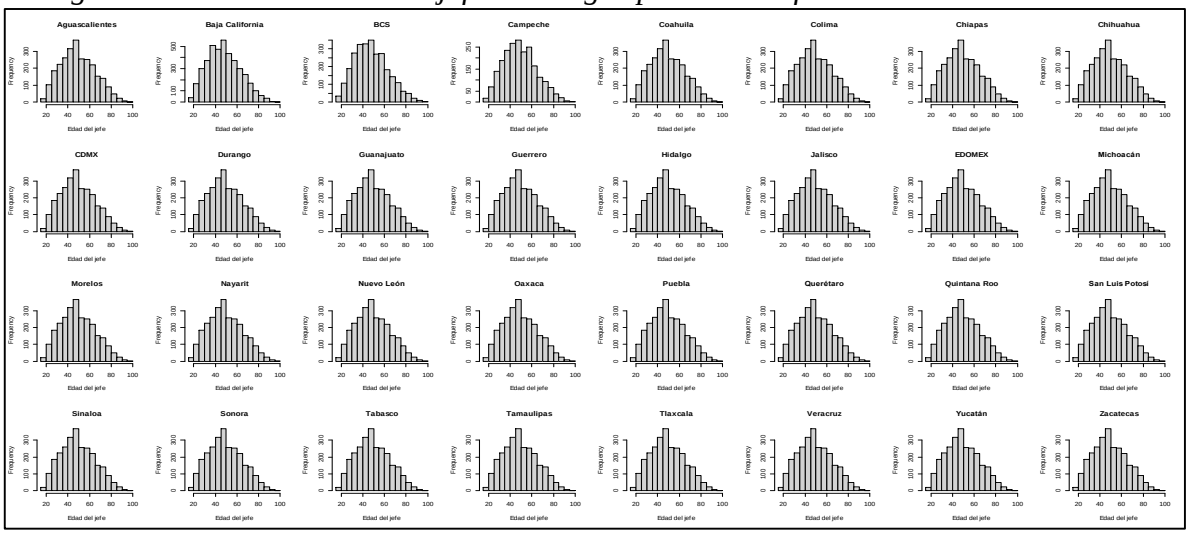
histogramas se invoca con el nombre de la gráfica correspondiente.

```
par(mfrow=c(4,8))
hist_ags <- hist(edadj_ags, main="Aguascalientes", xlab="Edad del jefe")
hist_bc <- hist(edadj_bc, main="Baja California", xlab="Edad del jefe")
hist_bcs <- hist(edadj_bcs, main="BCS", xlab="Edad del jefe")
hist_camp <- hist(edadj_camp, main="Campeche", xlab="Edad del jefe")
hist_coah <- hist(edadj_ags, main="Coahuila", xlab="Edad del jefe")
hist_col <- hist(edadj_ags, main="Colima", xlab="Edad del jefe")
hist_chi <- hist(edadj_ags, main="Chiapas", xlab="Edad del jefe")
hist_chih <- hist(edadj_ags, main="Chihuahua", xlab="Edad del jefe")
hist_cdmx <- hist(edadj_ags, main="CDMX", xlab="Edad del jefe")
hist_dur <- hist(edadj_ags, main="Durango", xlab="Edad del jefe")
hist_gua <- hist(edadj_ags, main="Guanajuato", xlab="Edad del jefe")
hist_gue <- hist(edadj_ags, main="Guerrero", xlab="Edad del jefe")
hist_hgo <- hist(edadj_ags, main="Hidalgo", xlab="Edad del jefe")
hist_jal <- hist(edadj_ags, main="Jalisco", xlab="Edad del jefe")
hist_edomex <- hist(edadj_ags, main="EDOMEX", xlab="Edad del jefe")
hist_mich <- hist(edadj_ags, main="Michoacán", xlab="Edad del jefe")
hist_mor <- hist(edadj_ags, main="Morelos", xlab="Edad del jefe")
hist_nay <- hist(edadj_ags, main="Nayarit", xlab="Edad del jefe")
hist_nl <- hist(edadj_ags, main="Nuevo León", xlab="Edad del jefe")
hist_oax <- hist(edadj_ags, main="Oaxaca", xlab="Edad del jefe")
hist_pue <- hist(edadj_ags, main="Puebla", xlab="Edad del jefe")
hist_qro <- hist(edadj_ags, main="Querétaro", xlab="Edad del jefe")
hist_qroo <- hist(edadj_ags, main="Quintana Roo", xlab="Edad del jefe")
hist_slp <- hist(edadj_ags, main="San Luis Potosí", xlab="Edad del jefe")
hist_sin <- hist(edadj_ags, main="Sinaloa", xlab="Edad del jefe")
hist_son <- hist(edadj_ags, main="Sonora", xlab="Edad del jefe")
hist_tab <- hist(edadj_ags, main="Tabasco", xlab="Edad del jefe")
hist_tam <- hist(edadj_ags, main="Tamaulipas", xlab="Edad del jefe")
hist_tlax <- hist(edadj_ags, main="Tlaxcala", xlab="Edad del jefe")
hist_ver <- hist(edadj_ags, main="Veracruz", xlab="Edad del jefe")
hist_yuc <- hist(edadj_ags, main="Yucatán", xlab="Edad del jefe")
hist_zac <- hist(edadj_ags, main="Zacatecas", xlab="Edad del jefe")
```

En la figura 3 se visualizan los 32 histogramas de las edades de los jefes del hogar

por entidad federativa.

Figura 3.
Histogramas de las edades de los jefes del hogar por entidad federativa



Nota: Elaboración propia con base en la ENIGH 2022 (INEGI, 2022e)

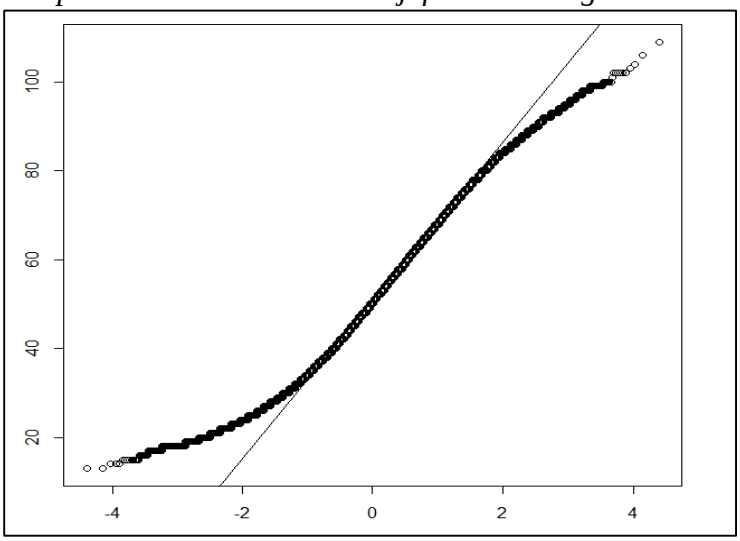
A continuación, se realizan una serie de pruebas de normalidad, con el fin de saber si la distribución de la edad de los jefes de los hogares se distribuye de manera normal. Estas pruebas serán de tipo gráfico, pero también estadísticas. En este contexto, se asumen las siguientes pruebas de hipótesis:

$$H_0: la\ muestra\ proviene\ de\ una\ distribución\ normal$$
$$H_1: la\ muestra\ no\ proviene\ de\ una\ distribución\ normal$$

En primer término, se genera una gráfica de cuantil (`QQplot`), la cual es una herramienta visual para conocer si un conjunto de datos o muestra proviene de una población normal. La función `qqnorm` sirve para construir esta gráfica, mientras que la función `qqline` agrega una línea de referencia que ayuda a interpretar el gráfico `QQplot`, lo cual se muestra en la figura 4.

```
> qqnorm(edad_jefe, main=="Gráfica de cuantiles")
> qqline(edad_jefe)
```

Figura 4.
Esquema de cuantíes para la edad media de los jefes de los hogares



Nota: Elaboración propia con base en la ENIGH 2022 (INEGI, 2022e)

Con esta visualización se puede interpretar de manera preliminar que la distribución de la edad media de los jefes de los hogares no se distribuye de manera normal. Lo anterior será confirmado por medio de una serie de pruebas estadísticas para muestras grandes, las cuales se plasman en la tabla 1 junto con la función en R que se usa para cada una de ellas.

Tabla 1.
Pruebas de normalidad

Prueba	**Función en R**
Anderson-Darling	`ad.test` del paquete `nortest`
Cramer-von Mises	`cvm.test` del paquete `nortest`
Lilliefors (Kolmogorov-Smirnov)	`lillie.test` del paquete `nortest`
Pearson chi-square	`pearson.test` del paquete `nortest`

Nota. Elaboración propia con base en la ENIGH 2022 (INEGI, 2022e)

En primer término, se debe instalar el paquete `nortest`. Para un uso posterior del mismo, se debe utilizar la función `library`. Posteriormente, se ejecuta cada una de las pruebas y se plasman los resultados de manera individual con sus interpretaciones estadísticas.

```
> install.packages("nortest")
> library(nortest)
```

Prueba Anderson-Darling

```
> ad.test(edad_jefe)
        Anderson-Darling normality test
        data:  edad_jefe
        A = 249.57, p-value < 2.2e-16
```

Puesto que la probabilidad del estadístico Anderson-Darling es prácticamente cero, es decir, estadísticamente significativo, se rechaza la hipótesis nula de normalidad.

Prueba Cramer-von Mises

```
> cvm.test(edad_jefe)
        Cramer-von Mises normality test
        data:  edad_jefe
        W = 35.572, p-value = 7.37e-10
```

Como en la prueba Anderson-Darling, en este caso el valor de la probabilidad del estadístico W es igualmente cero, por lo que se rechaza la hipótesis nula de normalidad a través de la prueba Cramer-von Mises.

Prueba Lilliefors (Kolmogorov-Smirnov)

```
> lillie.test(edad_jefe)
        Lilliefors (Kolmogorov-Smirnov) normality test
        data:  edad_jefe
        D = 0.044783, p-value < 2.2e-16
```

De igual manera, la prueba rechaza la hipótesis nula de normalidad a partir del valor p del estadístico D.

Prueba Pearson chi-square

```
> pearson.test(edad_jefe)
        Pearson chi-square normality test
        data:  edad_jefe
        P = 213178, p-value < 2.2e-16
```

Asimismo, se suscita rechazo de la hipótesis nula a partir del valor p del estadístico P.

Conclusiones

En este trabajo se mostró la forma en que se trabaja una base de datos, cuyo conjunto de datos está conformado por una cantidad considerable de información numérica. De igual manera, la manera como se carga dicha información en el paquete libre R, así como el uso de una serie de funciones y paquetes (estos últimos a partir de su instalación y carga desde la red), con la finalidad de llevar a cabo una rutina que permite identificar características, en este caso, de los jefes de los hogares en México.

La intención fue aportar elementos para el análisis estadístico con una herramienta virtual, aprovechable por economistas e interesados en la materia, que sirva para afrontar los desafíos que la era moderna demanda de los profesionales de la ciencia económica.

Con esta propuesta, la intención es proporcionar elementos básicos a los primeros usuarios de algún paquete estadístico, estudiantes de la ciencia económica e incluso, de las ciencias sociales, las nociones del análisis de datos, los cuales deben tener el alcance de lo mostrado en la breve revisión de la literatura empírica.

Referencias

Asociación Mexicana de Agencias de Inteligencia de Mercado y Opinión [AMAI] (2023). Nivel Socioeconómico AMAI 2024. México. AMAI. URL https://www.amai.org/descargas/NOTA_METODOLOGICA_NSE_AMAI_2024_v6.pdf

Comisión Económica para América Latina y el Caribe [CEPAL] (2021). Encuestas de ingresos y gastos de los hogares. Experiencias recientes en América Latina y el Caribe. Santiago. NU. URL https://repositorio.cepal.org/server/api/core/bitstreams/30f468c7-dd82-4d3c-b7f2-4df058b2d7b2/content

Consejo Nacional de Evaluación de la Política de Desarrollo Social [CONEVAL] (2024). Edición de la pobreza. URL https://www.coneval.org.mx/Medicion/Paginas/PobrezaInicio.aspx

Dalgaard, Peter (2008), Introductory Statistics with R, USA, Springer.

Escarela, Gabriel (2014), R para tod@s. Un enfoque aplicado al análisis estadístico básico, México, Universidad Autónoma Metropolitana Unidad Iztapalapa.

Heras, Miguel y Francisco Islas (2015), "Microdatos del Censo de Población y Vivienda 2010 con Stata", Tiempo Económico, núm. 31, vol. X, tercer cuatrimestre, pp. 57-80.

Hernandez, Freddy y Olga Usuga (2023), Manual de R, URL https://fhernanb.github.io/Manual-de-R/index.html

Márquez, Cinthia (2023). Estimación y análisis de la inflación por deciles de ingreso, 2020-2022. Economía UNAM. Vol. 20. Núm. 59. México. UNAM. URL http://revistaeconomia.unam.mx/index.php/ecu/article/view/797

Instituto Nacional de Estadística y Geografía [INEGI] (2022). Encuesta Nacional de Ingreso y Gasto de los Hogares ENIGH 2020. Nota técnica. México. INEGI

Instituto Nacional de Estadística y Geografía [INEGI] (2022a), Encuesta Nacional de Ingreso y Gasto de los Hogares 2022. ENIGH. Nueva serie. Descripción de la base de datos, México, INEGI.

Instituto Nacional de Estadística y Geografía [INEGI] (2022b), Encuesta Nacional de Ingreso y Gasto de los Hogares 2022. ENIGH. Nueva serie. Descripción del cálculo de los principales indicadores con R. México, INEGI.

Instituto Nacional de Estadística y Geografía [INEGI] (2022c), Encuesta Nacional de Ingreso y Gasto de los Hogares 2022. ENIGH. Nueva serie. Diseño conceptual, México, INEGI.

Instituto Nacional de Estadística y Geografía [INEGI] (2022d), Encuesta Nacional de Ingreso y Gasto de los Hogares 2022. ENIGH. Nueva serie. Documento operativo de campo, México, INEGI.

Instituto Nacional de Estadística y Geografía [INEGI] (2022e), Encuesta Nacional de Ingreso y Gasto de los Hogares 2022. ENIGH. Nueva serie. Microdatos: Principales variables por hogar (concentradohogar). México, INEGI.

Naciones Unidas (2016). Sistema de Cuentas Nacionales. Nueva York. NU. URL https://unstats.un.org/unsd/nationalaccount/docs/sna2008spanish.pdf

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Spiegel, Murray y Larry Stephens (2009), Estadística, México, Mc Graw Hill.