Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH). 2020

Estadística Aplicada 1 - Proyecto 4

Fernando Lango - 181055

14/12/2021

El objetivo del presente proyecto es analizar y presentar algunos de los resultados de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH). 2020.

Inicialmente vamos a descargar los datos correspondientes recolectados durante la encuesta que se encuentran la siguiente liga. Cabe recalcar que esta página contiene diversos micro datos para cada una de las categorías que se pretende analizar. Posteriormente vamos a especificar cuáles de ellas utilizaremos y procederemos a su descarga correspondiente.

Descripción de la encuesta

De acuerdo con la Presentación de resultados de la *ENIGH*, los objetivos de esta encuesta buscan:

- Conocer el comportamiento de los ingresos y gastos de los hogares.
- Ofrecer información sobre las características ocupacionales y sociodemográficas.
- Presentar datos sobre las características de la infraestructura de la vivienda.

Diseño muestral

El objetivo del presente apartado es explicar la manera en que se realizó la encuesta con ayuda de la Nota metodológica publicada dentro de la misma encuesta.

La población objetivo de esta encuesta son todos los hogares que se ubican dentro de la República Mexicana. De esta manera, es posible dar resultados a nivel nacional y estatal; respetando las diferencias entre lo urbano y lo rural.

Esta encuesta utiliza un enfoque estadístico, por lo que los resultados obtenidos por la muestra seleccionada se consideran generales para toda la población.

El tipo de muestreo es bietápico, estratificado y por conglomerados:

- **Bietápico**: se tomó una muestra durante dos etapas:
  - Inicialmente (*Unidad Primaria de Muestreo*), para dividir a la totalidad del territorio nacional, se seleccionan aleatoriamente a un conjunto de viviendas que comparten ciertas características dentro de manzanas, colonias o municipios.
    - Urbano alto: entre 80 y 160 viviendas habitadas.
    - Complemento urbano: entre 160 y 300 viviendas habitadas.
    - Rural: entre 160 y 300 viviendas habitadas.
  - Posteriormente, de cada una de las Unidades Primarias de Muestreo (UPM) dadas por los conjuntos de vivienda descritos anteriormente, se vuelve a seleccionar (*Unidad Secundaria de Muestreo*) aleatoriamente a hogares que se encuentran dentro de estas zonas definidas.
- **Estratificado**: se refiere a que nuestras *Unidades Primarias de Muestreo* poseen características que nos permiten reunirlas y catalogarlas de una manera homogénea. Es decir, todas las viviendas dentro de cierta colonia poseen particularidades como el tamaño del terreno, número de cuartos, cantidad de ventanas, etc.; que nos permiten agruparlas como una unidad.

- **Por conglomerados**: se hacen agrupaciones aunque no necesariamente comparten características comunes específicas como ocurre en el muestreo estratificado.

Finalmente, la muestra total recolectada fue de 105,483 viviendas que representan a 126,760,856 habitantes.

Descripción de la muestra

Para describir a nuestra muestra vamos a utilizar las características sociodemográficas de los integrantes del hogar, Ingresos y percepciones financieras y de capital de los integrantes del hogar, Gastos realizados en el hogar a nivel integrante y la descripción de la base de datos para poder entender a qué se refiere cada una de las variables.

```
poblacion_url <- "https://www.inegi.org.mx/contenidos/programas/enigh/nc/2020/microdatos/enigh2020_ns_p
ingresos_url <- "https://www.inegi.org.mx/contenidos/programas/enigh/nc/2020/microdatos/enigh2020_ns_in
gastos_url <- "https://www.inegi.org.mx/contenidos/programas/enigh/nc/2020/microdatos/enigh2020_ns_gast

if (!dir.exists("datos")) {
    dir.create("datos")
    download.file(poblacion_url, "datos/poblacion.zip")
    download.file(ingresos_url, "datos/ingresos.zip")
    download.file(gastos_url, "datos/gastos.zip")
    unzip("datos/poblacion.zip", exdir = "datos")
    unzip("datos/ingresos.zip", exdir = "datos")
    unzip("datos/gastos.zip", exdir = "datos")
    file.remove("datos/poblacion.zip")
    file.remove("datos/ingresos.zip")
    file.remove("datos/gastos.zip")
}

data_poblacion <- read.csv("datos/poblacion.csv")
data_ingresos <- read.csv("datos/ingresos.csv")
data_gastos <- read.csv("datos/gastospersona.csv")
```

Nuestra información se compone de la siguiente manera:

- Población: 315743 observaciones y 184 variables.
- Ingresos: 394912 observaciones y 17 variables.
- Gastos: 302694 observaciones y 20 variables.

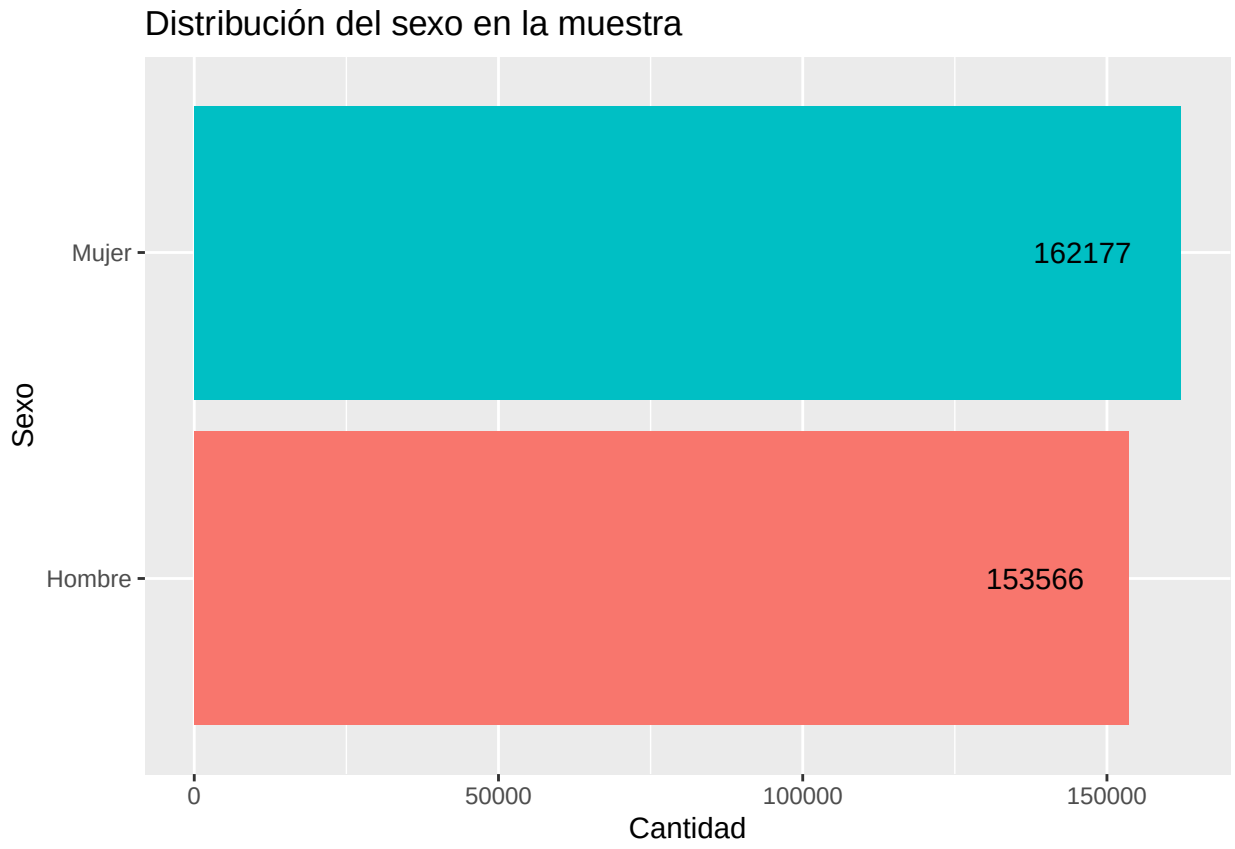
Población

La parte que vamos a mostrar de la población es cómo se distribuyen por sexo en la muestra. Para eso vamos a utilizar el paquete `tidyverse` para el manejo de datos y las gráficas.

```
library(tidyverse)

sexo_muestra <-
    data_poblacion %>%
    select(sexo) %>%
    mutate(sexo = replace(sexo, sexo == 1, "Hombre")) %>%
    mutate(sexo = replace(sexo, sexo == 2, "Mujer")) %>%
    group_by(sexo) %>%
    tally()
```

```
sexo_muestra %>%
    ggplot(aes(x=sexo, y=n)) +
    geom_bar(aes(fill = sexo), stat = "identity") +
    geom_text(aes(y=n-.1*n,label = n)) +
    labs(title = "Distribución del sexo en la muestra",
         x = "Sexo",
         y = "Cantidad") +
    theme(legend.position = "none") +
    coord_flip()
```



Ingresos

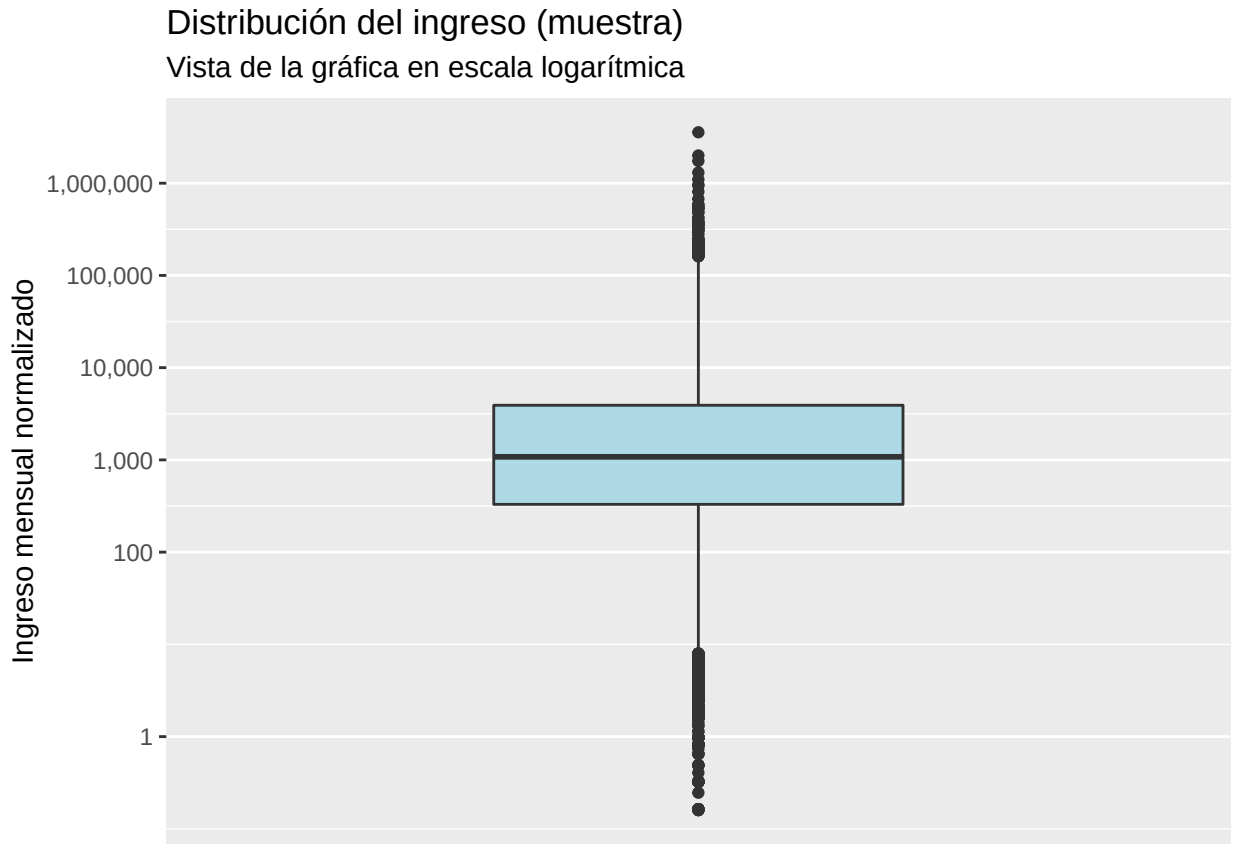
Para el presente apartado vamos a utilizar el ingreso trimestral normalizado de acuerdo con la decena de levantamiento y así obtener el ingreso mensual normalizado.

```
library(scales)

ingresos_muestra <-
    data_ingresos %>%
    select(ing_tri) %>%
    mutate(ing_men = ing_tri/3)

ingresos_muestra %>%
    ggplot(aes(y= ing_men)) +
    geom_boxplot(fill = "lightblue") +
```

```
    scale_y_log10(labels = comma,
                  breaks = c(1,100,1000,10000,100000,1000000)) +
    scale_x_discrete(breaks = 0) +
    labs(title = "Distribución del ingreso (muestra)",
         subtitle = "Vista de la gráfica en escala logarítmica",
         y = "Ingreso mensual normalizado")
```



Para entender con un poco más de detalles esta información, veamos los valores correspondientes a la gráfica anterior.

```
summary(ingresos_muestra$ing_men)
```

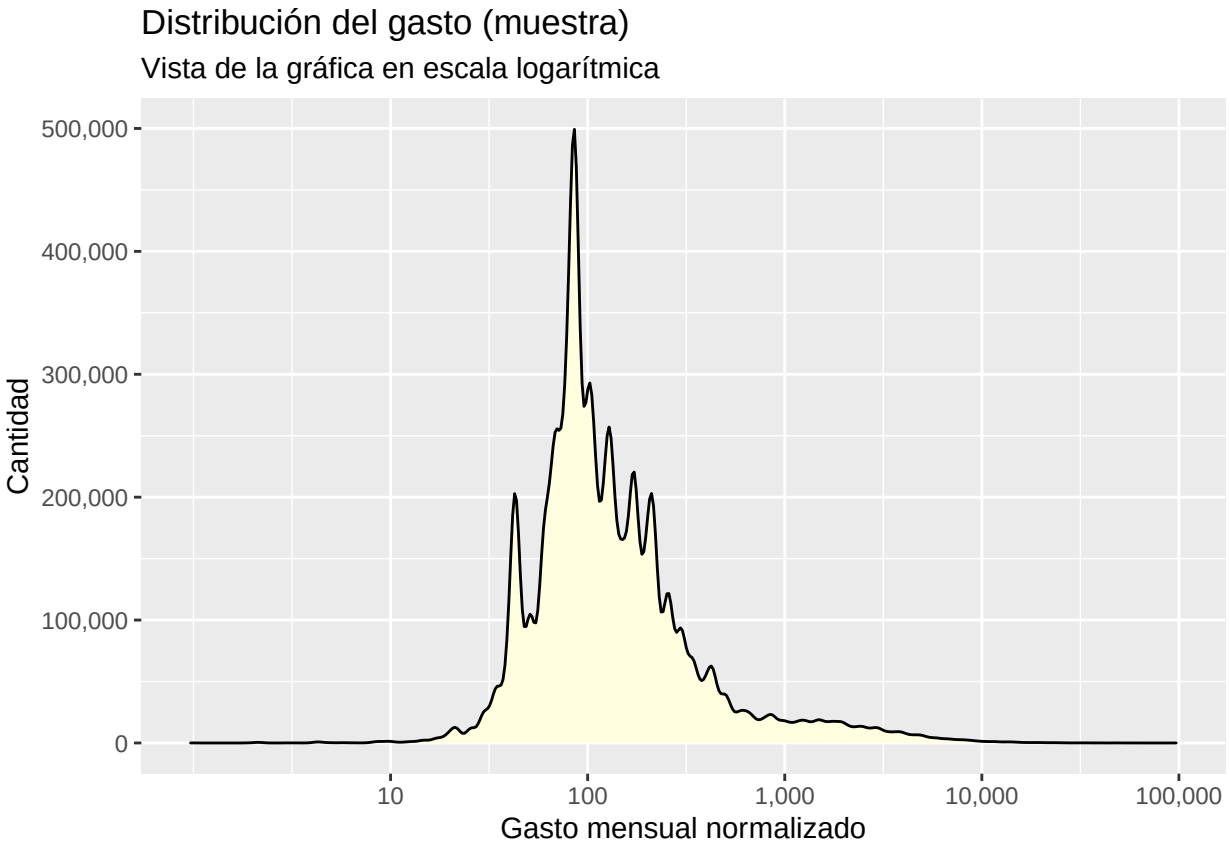
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     330    1080    3150    3913 3562973
```

Gastos

El gasto mostrado en el presente apartado corresponde al mensual normalizado que podemos obtener al dividir entre tres el gasto trimestral normalizado de acuerdo a la decena de levantamiento.

```
gastos_muestra <-
    data_gastos %>%
    select(gasto_tri) %>%
    mutate(gasto_men = gasto_tri/3) %>%
    na.omit()
```

```
gastos_muestra %>%
    ggplot(aes(x= gasto_men)) +
    geom_density(aes(y=..count..),fill = "lightyellow") +
    scale_x_log10(labels = comma,
                  breaks = c(0,10,100,1000,10000,100000)) +
    scale_y_continuous(labels = comma) +
    labs(title = "Distribución del gasto (muestra)",
         subtitle = "Vista de la gráfica en escala logarítmica",
         x = "Gasto mensual normalizado",
         y = "Cantidad")
```



Para entender con un poco más de detalle esta información, veamos los valores correspondientes a la gráfica anterior.

```
summary(gastos_muestra$gasto_men)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.97    77.14   102.86   288.21   200.00 96774.19
```

Estimación sobre la población

Población

Inicialmente vamos a intentar estimar la cantidad de hombres y de mujeres que tenemos en nuestra población a partir de la muestra.

Para ello vamos a estimar la proporción del género que existe en la población con ayuda de la estimación del total de la población en México proporcionada dentro de la misma encuesta.

Para nuestras π_k vamos a utilizar que la población mexicana en el 2020 fue de 126,014,024 habitantes. De esta manera:

$$\begin{aligned}\pi_k &= \mathbb{P}(x_k \in \mathcal{S}) \\ &= \mathbb{P}(x_k \in \mathcal{S}_{H\cup M}) \\ &= \frac{\sum_{k=1}^{N} \mathbb{I}_{\mathcal{S}}(x_k)}{126,014,024}\end{aligned}$$

Mediante la muestra que tenemos, sabemos que $\sum_{k=1}^{N} \mathbb{I}_{\mathcal{S}}(x_k)$ es 315743.
De esta manera nuestra $\pi_k = \frac{315,743}{126,014,024}$
Así, mediante nuestro estimador del total podemos hacer el cálculo de la cantidad de hombres:

$$\begin{aligned}\hat{T}_H &= \sum_{k=1}^{N} \frac{x_k}{\pi_k} \mathbb{I}_{\mathcal{S}}(x_k) \\ &= \frac{1}{\pi_k} \sum_{k=1}^{N} \mathbb{I}_{\mathcal{S}_H}(x_k) \mathbb{I}_{\mathcal{S}}(x_k) \\ &= \frac{126,014,024}{315,743} \sum_{k=1}^{N} \mathbb{I}_{\mathcal{S}_H}(x_k) \mathbb{I}_{\mathcal{S}}(x_k)\end{aligned}$$

Haciendo el cálculo obtenemos que nuestra población total estimada de hombres es de 6.1288673×10^7.
Similarmente para las mujeres podemos obtener lo siguiente:

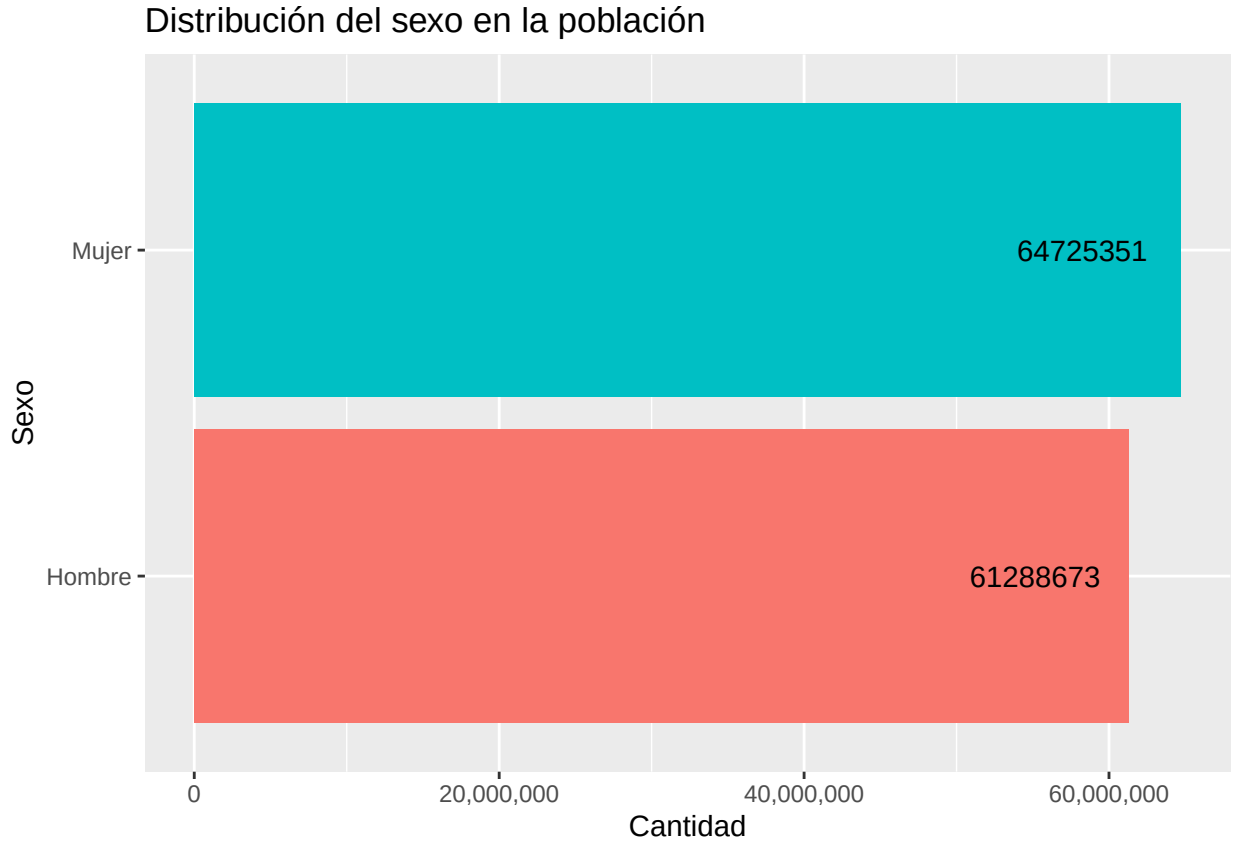
$$\begin{aligned}\hat{T}_M &= \sum_{k=1}^{N} \frac{x_k}{\pi_k} \mathbb{I}_{\mathcal{S}}(x_k) \\ &= \frac{1}{\pi_k} \sum_{k=1}^{N} \mathbb{I}_{\mathcal{S}_M}(x_k) \mathbb{I}_{\mathcal{S}}(x_k) \\ &= \frac{126,014,024}{315,743} \sum_{k=1}^{N} \mathbb{I}_{\mathcal{S}_M}(x_k) \mathbb{I}_{\mathcal{S}}(x_k)\end{aligned}$$

Haciendo el cálculo obtenemos que nuestra población total estimada de mujeres es de 6.4725351×10^7.

```
sexo <- c("Hombre","Mujer")
n <- c(round(126014024/nrow(data_poblacion) * sexo_muestra$n[1],
             0),
       round(126014024/nrow(data_poblacion) * sexo_muestra$n[2],
             0))
sexo_poblacion <- data.frame(sexo,n)
rm(sexo,n)

sexo_poblacion %>%
    ggplot(aes(x=sexo, y=n)) +
    geom_bar(aes(fill = sexo), stat = "identity") +
    geom_text(aes(y=n-.1*n,label = n)) +
    labs(title = "Distribución del sexo en la población",
         x = "Sexo",
```

```
        y = "Cantidad") +
    scale_y_continuous(labels = comma) +
    theme(legend.position = "none") +
    coord_flip()
```



Ingresos

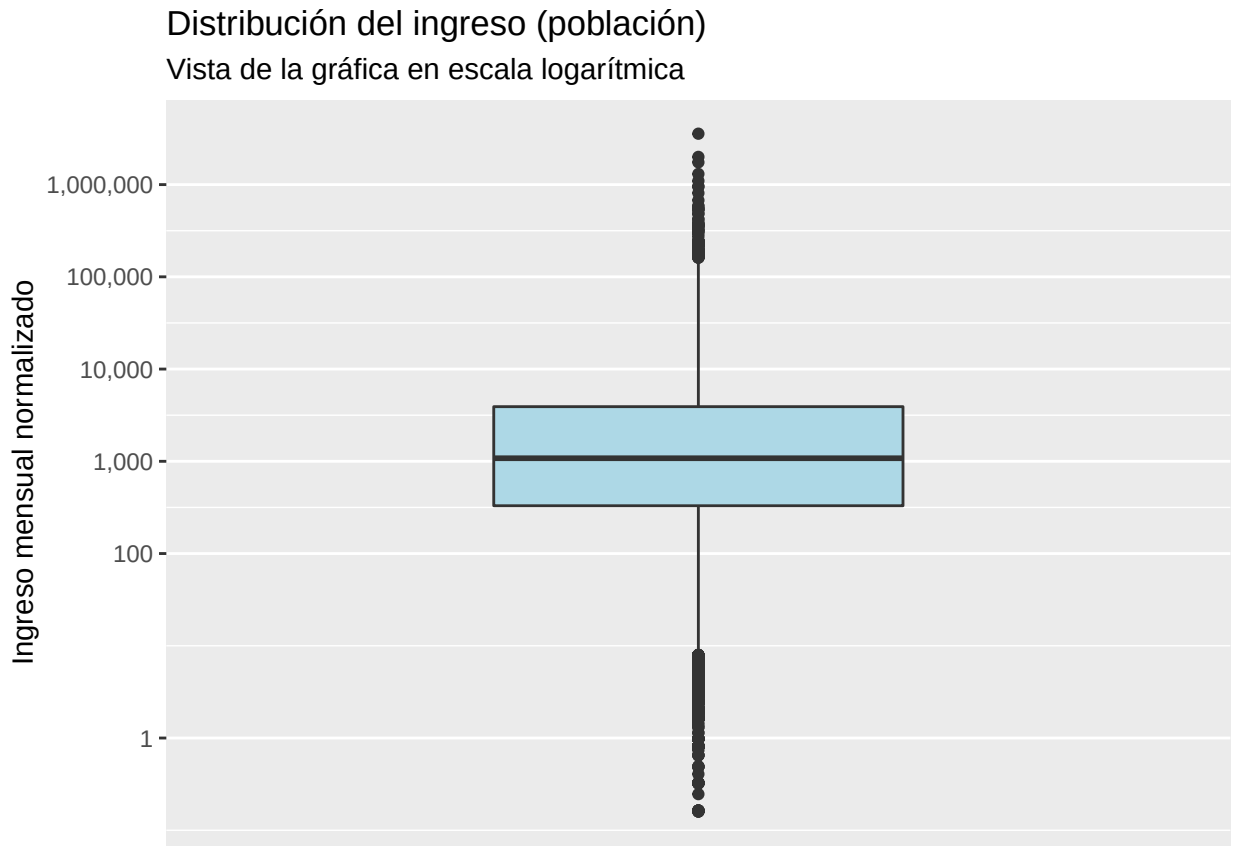
De la misma manera que calculamos lo anterior vamos a calcular la manera en que se distribuyen los ingresos para cada uno de los distintos niveles que este tiene.

$$\begin{aligned}\hat{T}_I &= \sum_{k=1}^{N} \frac{x_k}{\pi_k} \mathbb{I}_{\mathcal{S}}(x_k) \\ &= \frac{1}{\pi_k} \sum_{k=1}^{N} \mathbb{I}_{\mathcal{S}}(x_k) \mathbb{I}_{\mathcal{S}}(x_k) \\ &= \frac{126,014,024}{315,743} \sum_{k=1}^{N} \mathbb{I}_{\mathcal{S}}(x_k) \mathbb{I}_{\mathcal{S}}(x_k)\end{aligned}$$

```
ingresos_poblacion <-
    ingresos_muestra %>%
    select(ing_men) %>%
    group_by(ing_men) %>%
    tally() %>%
```

```
    mutate(n = round(126014024/315743*n,0))

ingresos_muestra %>%
    ggplot(aes(y= ing_men)) +
    geom_boxplot(fill = "lightblue") +
    scale_y_log10(labels = comma,
                  breaks = c(1,100,1000,10000,100000,1000000)) +
    scale_x_discrete(breaks = 0) +
    labs(title = "Distribución del ingreso (población)",
         subtitle = "Vista de la gráfica en escala logarítmica",
         y = "Ingreso mensual normalizado")
```



Por la manera en que fue generada la encuesta, podemos generalizar los resultados obtenidos durante la muestra para la población, de esta manera, guarda la proporción para la manera en la que se distribuye.

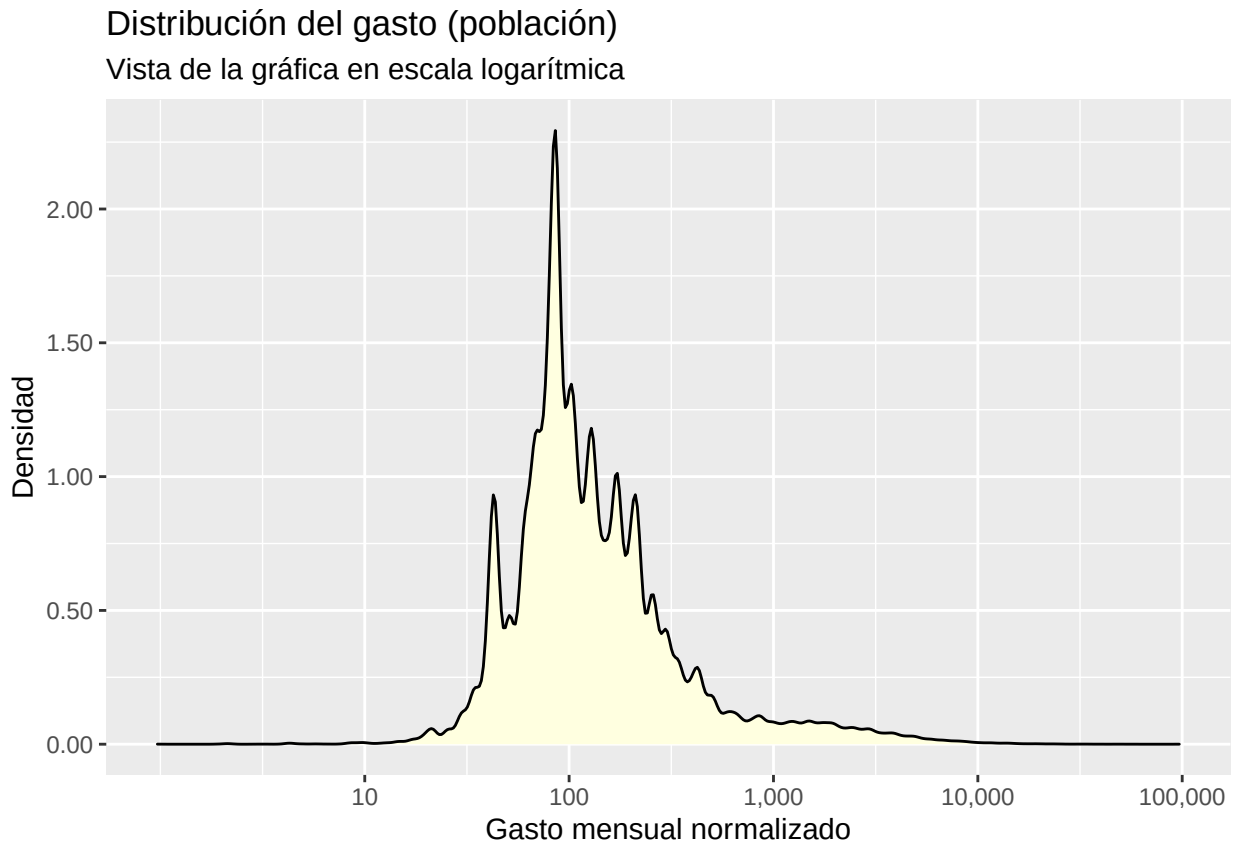
Gastos

De la misma manera que calculamos lo anterior vamos a calcular la manera en que se distribuyen los gastos para cada uno de los distintos niveles que este tiene.

$$
\begin{aligned}
\hat{T}_G &= \sum_{k=1}^{N} \frac{x_k}{\pi_k} \mathbb{I}_{\mathcal{S}}(x_k) \\
&= \frac{1}{\pi_k} \sum_{k=1}^{N} \mathbb{I}_{\mathcal{S}}(x_k) \mathbb{I}_{\mathcal{S}}(x_k) \\
&= \frac{126,014,024}{315,743} \sum_{k=1}^{N} \mathbb{I}_{\mathcal{S}}(x_k) \mathbb{I}_{\mathcal{S}}(x_k)
\end{aligned}
$$

```
gastos_poblacion <-
    gastos_muestra %>%
    select(gasto_men) %>%
    group_by(gasto_men) %>%
    tally() %>%
    mutate(n = round(126014024/315743*n,0))

gastos_muestra %>%
    ggplot(aes(x= gasto_men)) +
    geom_density(fill = "lightyellow") +
    scale_x_log10(labels = comma,
                  breaks = c(0,10,100,1000,10000,100000)) +
    scale_y_continuous(labels = comma) +
    labs(title = "Distribución del gasto (población)",
         subtitle = "Vista de la gráfica en escala logarítmica",
         x = "Gasto mensual normalizado",
         y = "Densidad")
```



Notemos que se distribuye igual que a la muestra, ya que por el modo en el que fue diseñada la encuesta, entonces podemos generalizar estos resultados obtenidos para toda la población.

Estimadores poblacionales

Media del ingreso

En nuestra información recolectada podemos ver que la media del ingreso es de 3149.51.

Vamos a utilizar el estimador de la media como sigue:

$$\begin{aligned}
\hat{\mu}_S &= \frac{\hat{T}_S}{N} \\
&= \frac{1}{N}\sum_{k=1}^{N}\frac{x_k}{\pi_k}\mathbb{I}_S(x_k) \\
&= \frac{1}{N}\sum_{k=1}^{N}\frac{x_k}{\frac{n}{N}}\mathbb{I}_S(x_k) \\
&= \frac{1}{N}\frac{N}{n}\sum_{k=1}^{N}x_k\mathbb{I}_S(x_k) \\
&= \frac{1}{n}\sum_{k=1}^{N}x_k\mathbb{I}_S(x_k)
\end{aligned}$$

```
mean_pob <- sum(ingresos_muestra$ing_men)/nrow(ingresos_muestra)
```

De esta manera, el promedio de la población será 3149.51.
Si quisiéramos un intervalo de confianza al 80%, entonces:

```
m <- 1000
confianza <- 80
alfa <- (100-confianza)/100
media_boot <- rep(NA,m)

for (i in 1:m) {
    muestra_bootstrap <- sample(ingresos_muestra$ing_men,
                                nrow(ingresos_muestra),
                                replace = TRUE)
    media_boot[i] <- sum(muestra_bootstrap)/
        length(muestra_bootstrap)
}
ic_bajo_boot_1 <- quantile(media_boot, alfa/2)
ic_alto_boot_1 <- quantile(media_boot, 1-alfa/2)
```

De esta manera, nuestro intervalo de confianza al 80% se encuentra dado por: [3129.61, 3171.92].
El hecho de que la media del ingreso poblacional se encuentre en un 80% de los casos entre el intervalo de confianza mostrado anteriormente nos deja ver que la realidad de México es demasiado triste. Tener un ingreso mensual por persona de esas cantidades se encuentra por debajo del Salario Mínimo General Mensual de la CDMX, por lo que difícilmente pueden cubrir sus necesidades estas personas.

Varianza del ingreso

La varianza del ingreso de nuestra muestra es 112,053,641
Vamos a utilizar el estimador de la varianza de la siguiente manera:

$$\begin{aligned}\hat{\sigma_s^2} &= \frac{1}{N}\sum_{k=1}^{N}\frac{(x_k-\mu)^2}{\pi_k}\mathbb{I}_S(x_k)\\ &= \frac{1}{N}\sum_{k=1}^{N}\frac{(x_k-\mu)^2}{\frac{n}{N}}\mathbb{I}_S(x_k)\\ &= \frac{1}{N}\frac{N}{n}\sum_{k=1}^{N}(x_k-\mu)^2\mathbb{I}_S(x_k)\\ &= \frac{1}{n}\sum_{k=1}^{N}(x_k-\mu)^2\mathbb{I}_S(x_k)\end{aligned}$$

```
var_pob <- sum((ingresos_muestra$ing_men -
                mean(ingresos_muestra$ing_men))^2)/
            nrow(ingresos_muestra)
```

De esta manera podemos estimar la varianza del ingreso de la población que es 112,053,357.
Si quisiéramos un intervalo de confianza al 80%, entonces:

```
m <- 1000
confianza <- 80
alfa <- (100-confianza)/100
```

```
var_boot <- rep(NA,m)

for (i in 1:m) {
    muestra_bootstrap <- sample(ingresos_muestra$ing_men,
                                nrow(ingresos_muestra),
                                replace = TRUE)
    var_boot[i] <- sum((muestra_bootstrap -
                    mean(muestra_bootstrap))^2)/
                    length(muestra_bootstrap)
}
ic_bajo_boot_2 <- quantile(var_boot, alfa/2)
ic_alto_boot_2 <- quantile(var_boot, 1-alfa/2)
```

De esta manera, nuestro intervalo de confianza al 80% se encuentra dado por: [70,237,436, 154,731,180].

Para poder hacer una comparación precisa debemos hacer la comparación y el análisis con la desviación estándar en lugar de la varianza. La desviación estándar estimada de la población se encuentra en un 80% de los casos entre 8,380.78 y 12,439.1. Para este caso vamos a utilizar nuestra desviación estándar estimada de nuestra población de 10,585.53. Si nosotros suponemos que se distribuye normal, entonces entre $\hat{\mu} - \hat{\sigma} =$ -7,436.02 y $\hat{\mu} + \hat{\sigma} =$ 13,735.03 podremos encontrar alrededor del 68.2% de nuestra población. Es decir, cerca de 85,941,564 personas tienen un ingreso menor a los $13,735.03 pesos mexicanos. Esto hace prácticamente evidente la concentración de riqueza de la población en unas pocas manos, pero vamos a confirmarlo con el coeficiente de asimetría.

Coeficiente de asimetría del ingreso

Recordemos que el *Coeficiente de asimetría (A)* está dado por:

$$A_s = \frac{1}{n}\sum_{k=1}^{n}(\frac{x_k - \mu}{\sigma})^3$$

Así, para nuestra muestra el coeficiente asimetría es de 143.65.

Proponemos nuestro estimador como:

$$\begin{aligned}
\hat{A_S} &= \frac{1}{N}\sum_{k=1}^{N}\frac{(\frac{x_k-\mu}{\sigma})^3}{\pi_k}\mathbb{I}_S(x_k) \\
&= \frac{1}{N}\sum_{k=1}^{N}\frac{(\frac{x_k-\mu}{\sigma})^3}{\frac{n}{N}}\mathbb{I}_S(x_k) \\
&= \frac{1}{N}\frac{N}{n}\sum_{k=1}^{N}(\frac{x_k-\mu}{\sigma})^3\mathbb{I}_S(x_k) \\
&= \frac{1}{n\sigma^3}\sum_{k=1}^{N}(x_k-\mu)^3\mathbb{I}_S(x_k)
\end{aligned}$$

```
asi_pob <- sum((ingresos_muestra$ing_men -
                mean(ingresos_muestra$ing_men))^3)/
            (nrow(ingresos_muestra)*
                 var(ingresos_muestra$ing_men)^(3/2))
```

De esta manera podemos estimar la varianza del ingreso de la población que es 143.65.

Si quisiéramos un intervalo de confianza al 80%, entonces:

```r
m <- 1000
confianza <- 80
alfa <- (100-confianza)/100
asi_boot <- rep(NA,m)

for (i in 1:m) {
    muestra_bootstrap <- sample(ingresos_muestra$ing_men,
                                nrow(ingresos_muestra),
                                replace = TRUE)
    asi_boot[i] <- sum((muestra_bootstrap -
                    mean(muestra_bootstrap))^3)/
                    (length(muestra_bootstrap)*
                    var(muestra_bootstrap)^(3/2))
}
ic_bajo_boot_3 <- quantile(asi_boot, alfa/2)
ic_alto_boot_3 <- quantile(asi_boot, 1-alfa/2)
```

De esta manera, nuestro intervalo de confianza al 80% se encuentra dado por: [65.6, 168.93].

Recordemos que nuestro coeficiente de asimetría nos indica si tenemos una mayor cantidad de ingresos de las personas del lado derecho o izquierdo de nuestra distribución. Para este caso, tener un coeficiente de asimetría de 143.65, es decir, mucho mayor que 0 nos indica que nuestra distribución se encuentra cargada completamente a la izquierda. La mayoría de las personas en México se encuentran ganando una cantidad baja, de esta manera comprobamos la gran desigualdad económica que existe en nuestro país.