Postwork sesión 8

Un centro de salud nutricional está interesado en analizar estadística y probabilísticamente los patrones de gasto en alimentos saludables y no saludables en los hogares mexicanos con base en su nivel socioeconómico, en si el hogar tiene recursos financieros extra al ingreso y en si presenta o no inseguridad alimentaria. Además, está interesado en un modelo que le permita identificar los determinantes socioeconómicos de la inseguridad alimentaria.

La base de datos es un extracto de la Encuesta Nacional de Salud y Nutrición (2012) levantada por el Instituto Nacional de Salud Pública en México. La mayoría de las personas afirman que los hogares con menor nivel socioeconómico tienden a gastar más en productos no saludables que las personas con mayores niveles socioeconómicos y que esto, entre otros determinantes, lleva a que un hogar presente cierta inseguridad alimentaria.

La base de datos contiene las siguientes variables:

- `nse5f` (nivel socieconómico del hogar): 1 "Bajo", 2 "Medio bajo", 3 "Medio", 4 "Medio alto", 5 "Alto"
- `area` (zona geográfica): 0 "Zona urbana", 1 "Zona rural"
- `numpeho` (número de personas en el hogar)
- `refin` (recursos financieros distintos al ingreso laboral): 0 "no", 1 "sí"
- `edadjef` (edad del jefe/a de familia)
- `sexoje` (sexo del jefe/a de familia): 0 "Hombre", 1 "Mujer"
- `añosedu` (años de educación del jefe de familia)
- `ln_als` (logarítmo natural del gasto en alimentos saludables)
- `ln_alns` (logarítmo natural del gasto en alimentos no saludables)
- `IA` (inseguridad alimentaria en el hogar): 0 "No presenta IA", 1 "Presenta IA"

```
df <- read.csv("https://raw.githubusercontent.com/beduExpert/Programacion-R-Santander-2022/main/Sesion-
```

1) Plantea el problema del caso
2) Realiza un análisis descriptivo de la información
3) Calcula probabilidades que nos permitan entender el problema en México
4) Plantea hipótesis estadísticas y concluye sobre ellas para entender el problema en México
5) Estima un modelo de regresión, lineal o logístico, para identificiar los determinanres de la inseguridad alimentaria en México
6) Escribe tu análisis en un archivo README.md y tu código en un script de R y publica ambos en un repositorio de Github.

   NOTA: Todo tu planteamiento deberá estár correctamente desarrollado y deberás analizar e interpretar todos tus resultados para poder dar una conclusión final al problema planteado.

1) Planteamiento del problema del caso

Objetivos

- Analizar patrones de gasto en alimentos saludables y no saludables en familias mexicanas con base en:
  - el nivel socioeconómico
  - disponibilidad de recursos financieros adicionales al ingreso
  - presencia de inseguridad alimentaria
- Elaborar un modelo que prediga la posibilidad de presentar inseguridad alimentaria (IA) con base en las variables disponibles en el extracto de la Encuesta Nacional de Salud y Nutrición 2012 (ENSANUT

2012)

2) Análisis descriptivo de la información

```
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(ggplot2)
```

Obtención de los datos:

```
url <- "https://raw.githubusercontent.com/beduExpert/Programacion-R-Santander-2022/main/Sesion-08/Postw
datos <- read.csv(url, encoding = "UTF-8")
head(datos)
```

```
##   nse5f area numpeho refin edadjef sexojef añosedu IA   ln_als  ln_alns
## 1     5    0       4     0      43       0      24  0 5.393628       NA
## 2     5    0       5     1      NA      NA      24  0 7.024649       NA
## 3     5    0       4     0      46       0      24  0 6.767343 4.605170
## 4     5    1       1     0      54       0      24  0 3.401197 4.094345
## 5     5    0       2     1      39       0      24  0 6.115892 5.480639
## 6     5    0       5     1      NA      NA      24  0 7.514800 5.598422
```

Para obtener un resumen estadístico de los datos, se convierten en factores las variables discretas.

```
datos$nse5f <- factor(datos$nse5f,
                      levels = 1:5,
                      labels = c("Bajo", "Medio bajo", "Medio", "Medio alto", "Alto"),
                      ordered = TRUE)
datos$area <- factor(datos$area,
                     levels = 0:1,
                     labels = c("Zona urbana", "Zona rural"))
datos$refin <- factor(datos$refin,
                      levels = 0:1,
                      labels = c("no", "sí"))
datos$sexojef <- factor(datos$sexojef,
                        levels = 0:1,
                        labels = c("Hombre", "Mujer"))
datos$IA <- factor(datos$IA,
                   levels = 0:1,
                   labels = c("No presenta IA", "Presenta IA"))
```

Y, enseguida, el resumen estadístico.

```
nrow(datos)
```

```
## [1] 40809
```

```
summary(datos)
```

```
##         nse5f               area          numpeho       refin
##  Bajo      :8858   Zona urbana:26591   Min.   : 1.000   no:33046
##  Medio bajo:8560   Zona rural :14218   1st Qu.: 3.000   sí: 7763
##  Medio     :8323                       Median : 4.000
##  Medio alto:7903                       Mean   : 3.941
##  Alto      :7165                       3rd Qu.: 5.000
```

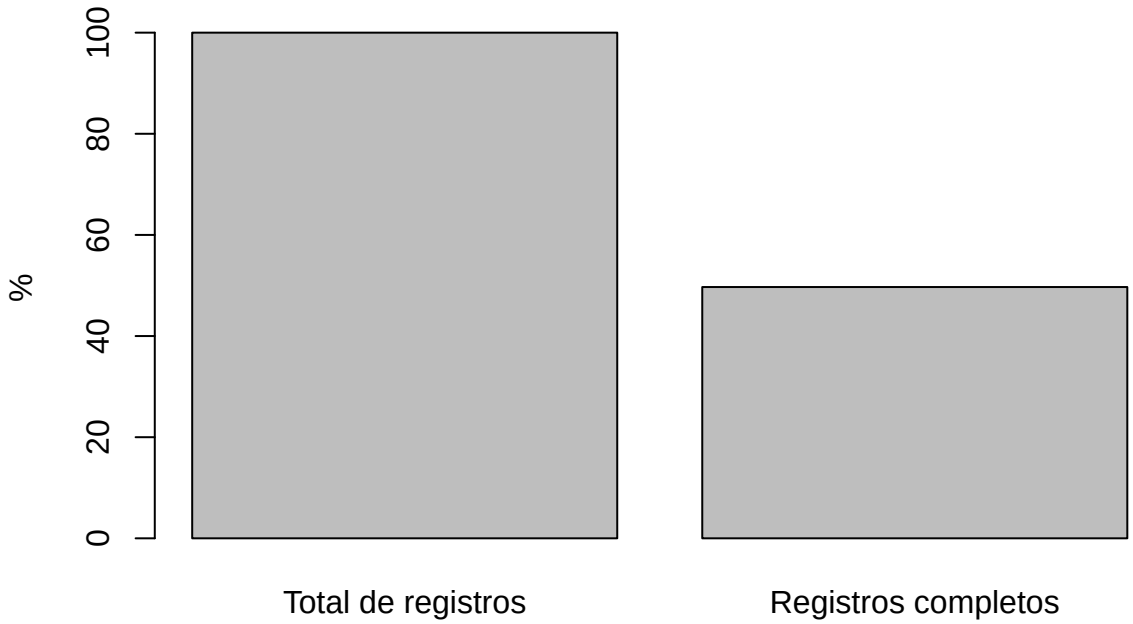
```
##                                       Max.   :19.000
##
##     edadjef       sexojef         añosedu                   IA
##  Min.   : 18    Hombre:26957   Min.   : 0.00   No presenta IA:10781
##  1st Qu.: 37    Mujer : 8861   1st Qu.: 9.00   Presenta IA   :30028
##  Median : 47    NA's  : 4991   Median : 9.00
##  Mean   : 49                   Mean   :10.36
##  3rd Qu.: 60                   3rd Qu.:12.00
##  Max.   :111                   Max.   :24.00
##  NA's   :5017
##      ln_als          ln_alns
##  Min.   :0.6931   Min.   :0.000
##  1st Qu.:5.7038   1st Qu.:3.401
##  Median :6.1633   Median :4.025
##  Mean   :6.0665   Mean   :4.125
##  3rd Qu.:6.5511   3rd Qu.:4.868
##  Max.   :8.9699   Max.   :8.403
##  NA's   :787      NA's   :17504
```

```
sum(complete.cases(datos))
```

```
## [1] 20280
```

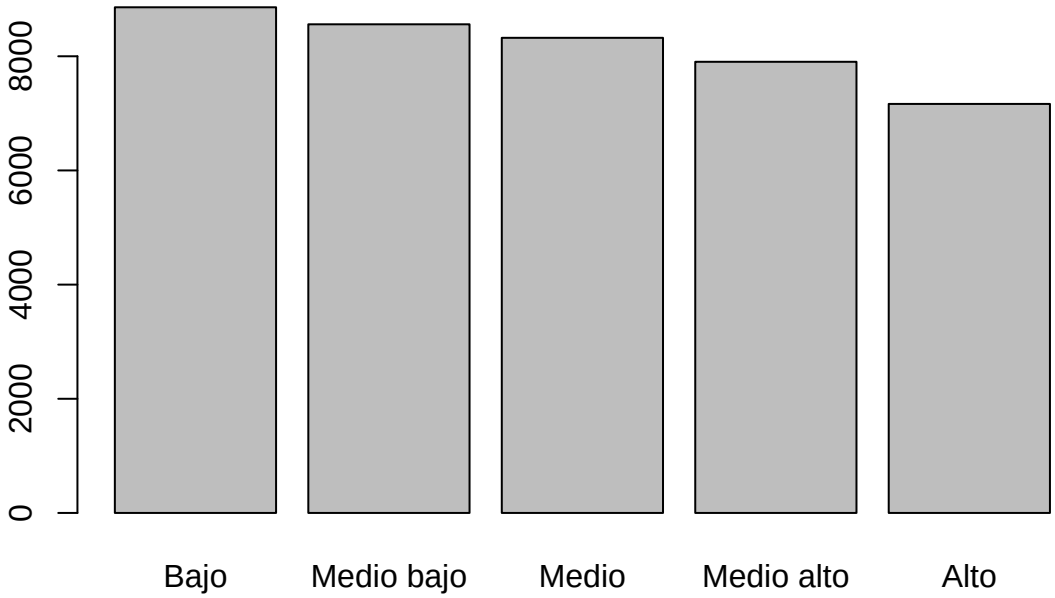
La base de datos tiene 40,809 registros, de los cuales 20,280, poco menos de la mitad, tienen información completa para todos los campos.

```
barplot(c(100, sum(complete.cases(datos))/nrow(datos)*100),
        names.arg = c("Total de registros", "Registros completos"),
        ylab = "%")
```



Con una mezcla más o menos homogénea de niveles socioeconómicos, con una tendencia ligeramente decreciente conforme aumenta el nivel socioeconómico, con entre 7,000 y 9,000 hogares por nivel.

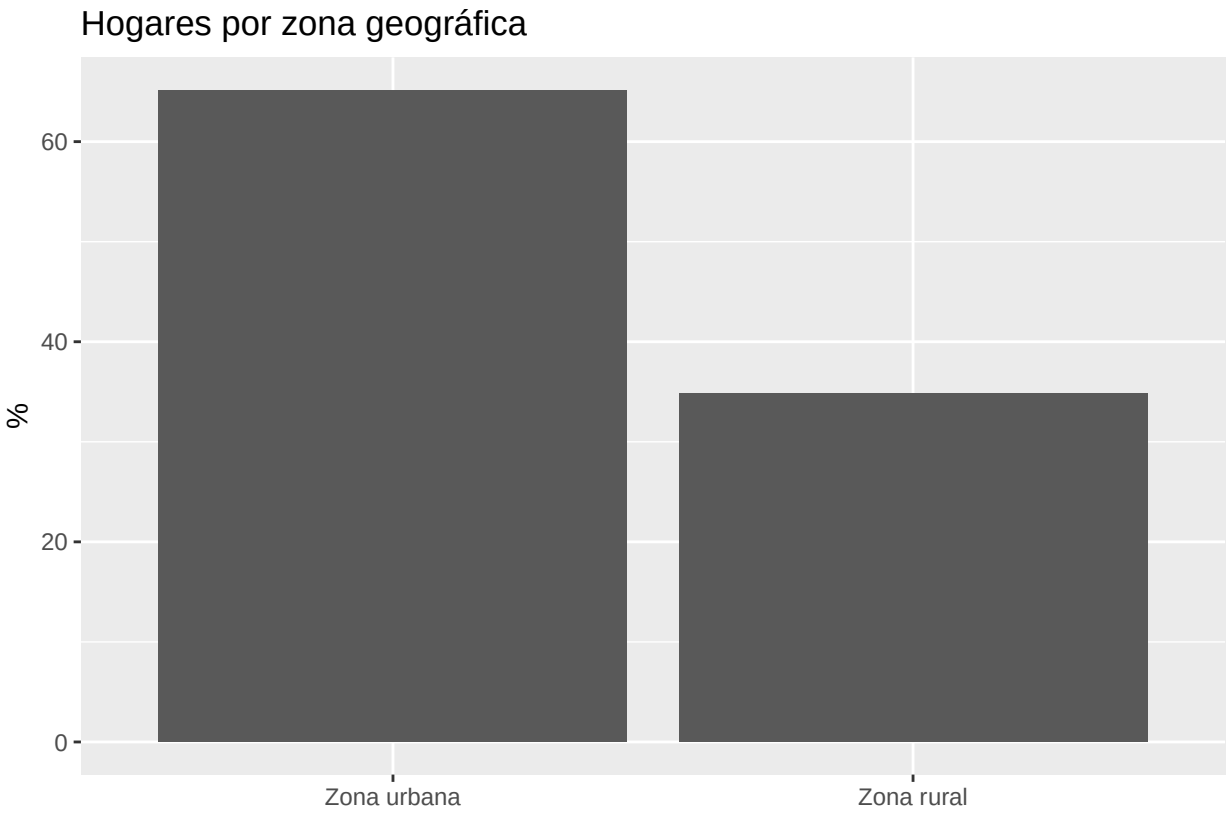
```
plot(datos$nse5f)
```



La muestra es mayoritariamente de hogares urbanos, aproximadamente en dos terceras partes.

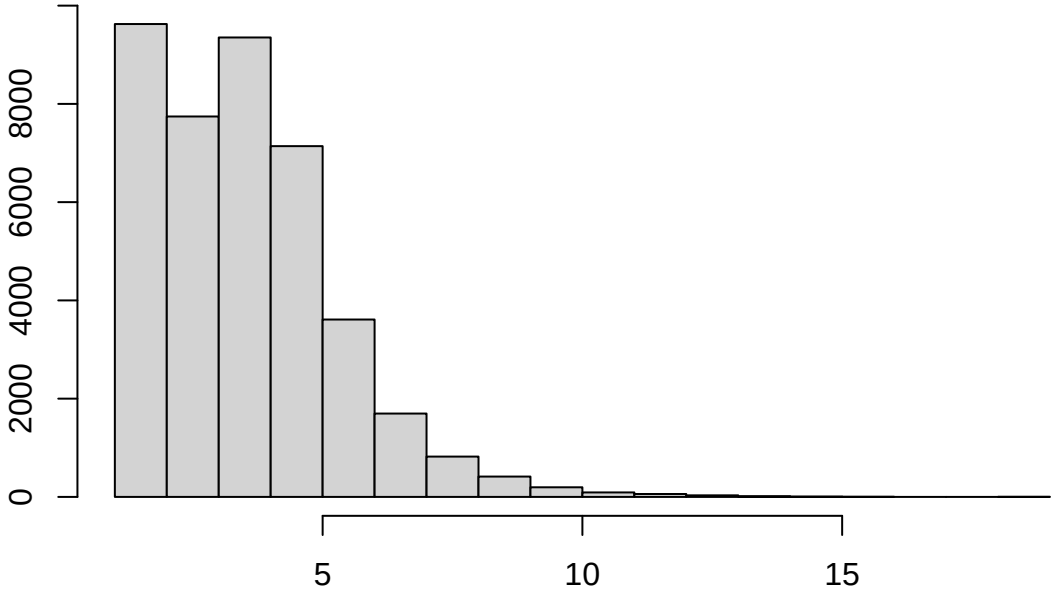
```
# plot(datos$area)

ggplot(datos, aes(x = area)) +
  geom_bar(aes(y = (..count..)/sum(..count..)*100)) +
  labs(title = "Hogares por zona geográfica",
       x = "", y = "%")
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
```



El número de personas por hogar varía entre 1 y 19, con una media de 3.9 y una mediana de 4. El 75% de los hogares es habitado por 5 personas o menos y el 50% por entre 3 y 5 personas. Asimismo, el 95% de los hogares tiene 7 o menos habitantes, siendo la excepción los valores mayores.

```
hist(datos$numpeho,
     main = "Número de personas en el hogar",
     xlab = "",
ylab = "")
```




```
quantile(datos$numpeho, probs = 0.95)
```

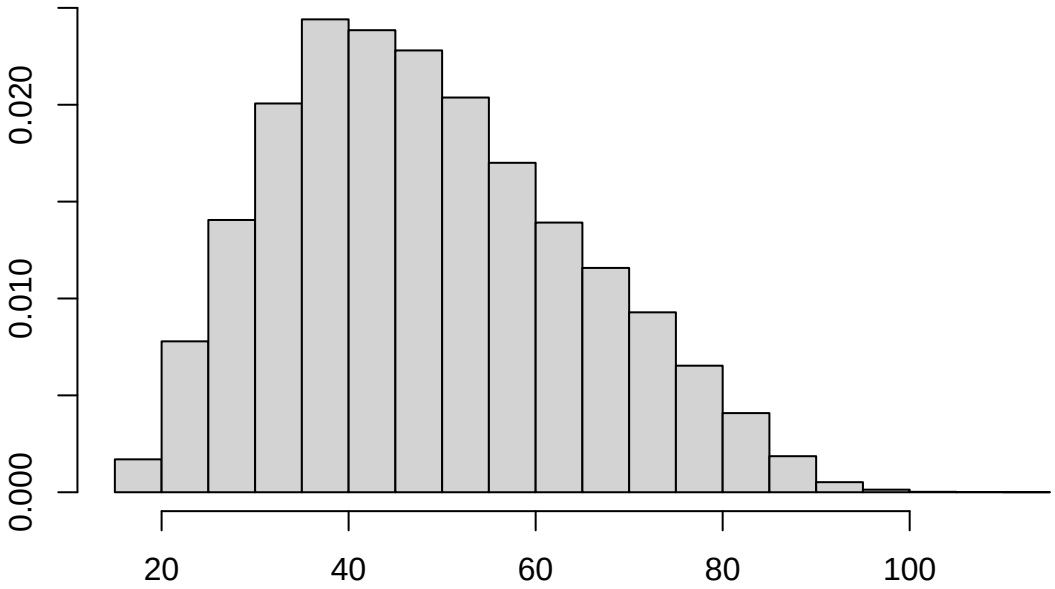
```
## 95%
##   7
```

En cuanto a recursos financieros adicionales al ingreso, únicamente uno de cada cinco hogares disponen de ellos.

```
#plot(datos$refin,
#     main = "Disponibilidad de recursos financieros distintos al ingreso")

ggplot(datos, aes(x = refin)) +
  geom_bar(aes(y = (..count..)/sum(..count..)*100)) +
  labs(title = "Hogares por disponibilidad de recursos financieros distintos al ingreso",
       x = "", y = "%")
```

El promedio de edad del jefe de familia es de 49 años, y la mediana, 47. La edad mínima reportada es de 18 años, y la máxima, de 111 (un posible error de captura). El 50% de las edades de los jefes de familia se ubica entre 37 y 60 años. El 75% de los jefes de familia son menores de 60 años, y el 90%, menores de 72. Para 5,017 de los hogares (12%) no se cuenta con este dato.

```
hist(datos$edadjef,
     freq = FALSE,
     main = "Edad del jefe de familia",
     xlab = "",
     ylab = "")
```

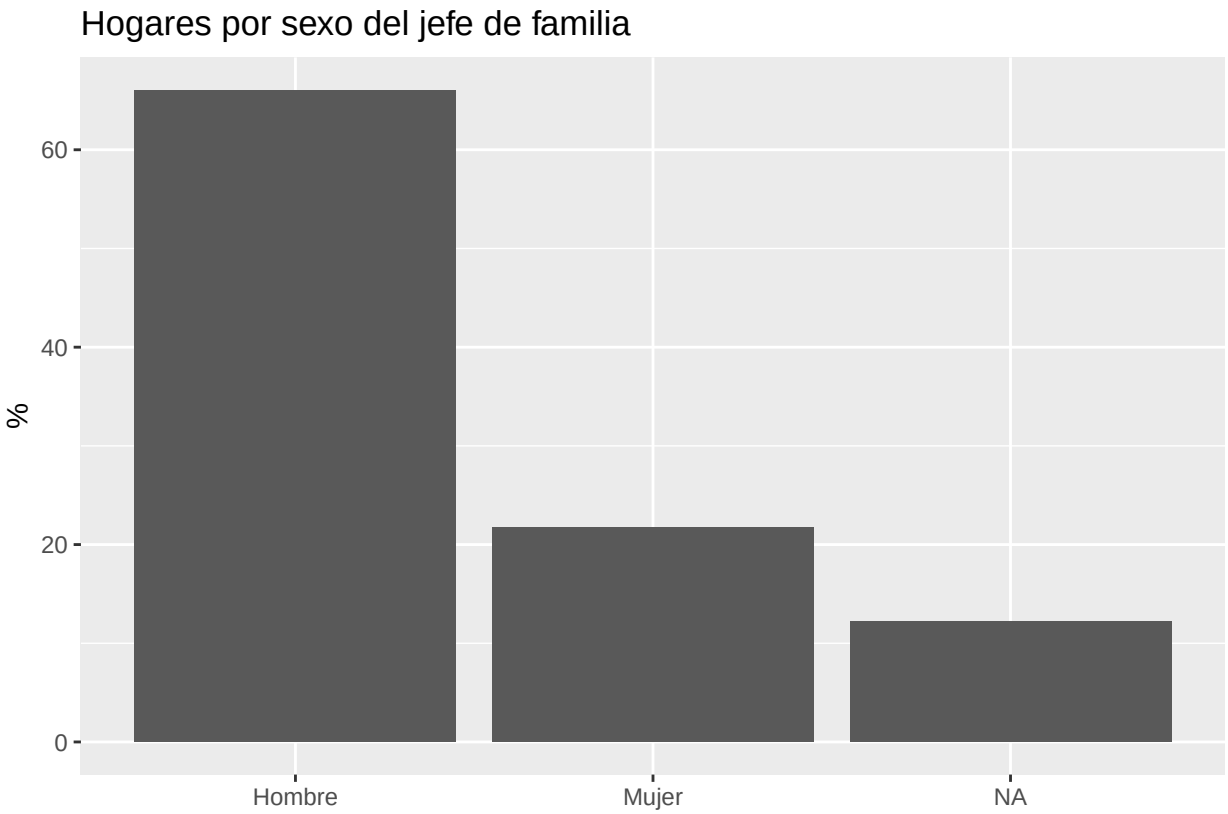



```
quantile(datos$edadjef, probs = 0.90, na.rm = TRUE)
```

```
## 90%
##  72
```

En alrededor del 65% de los hogares, el jefe de familia es hombre, mientras que en poco más de un 20% es mujer. Aproximadamente un 10% no proporcionó este dato.

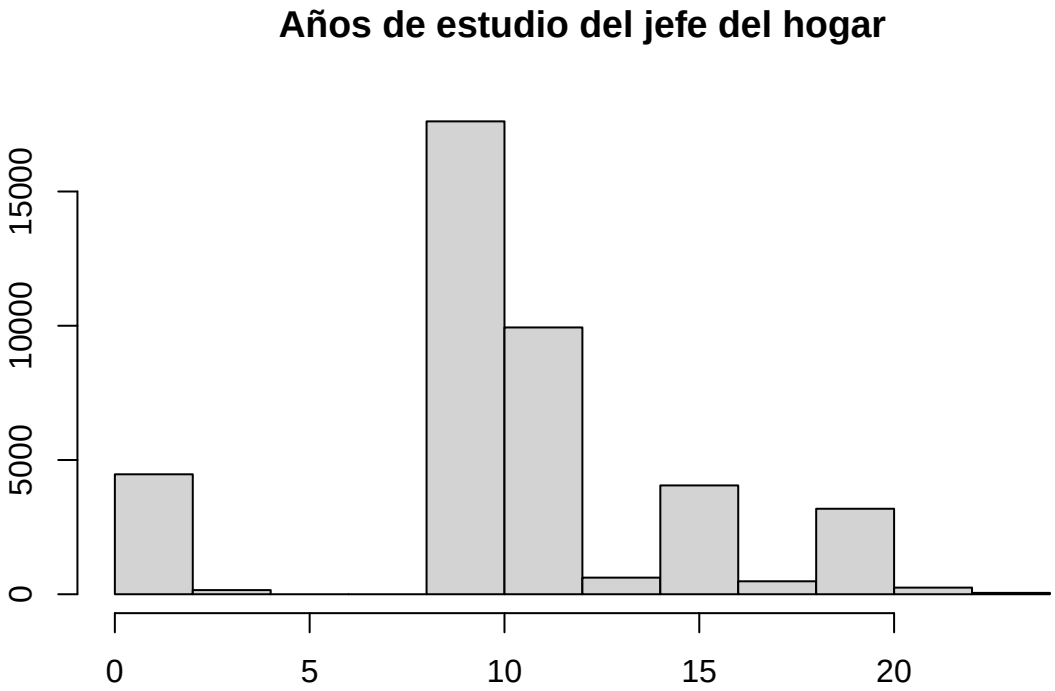
```
# plot(datos$sexojef)

ggplot(datos, aes(x = sexojef)) +
  geom_bar(aes(y = (..count..)/sum(..count..)*100)) +
  labs(title = "Hogares por sexo del jefe de familia",
       x = "", y = "%")
```



El promedio de años de estudio del jefe del hogar es 10.4 (preparatoria inconclusa), y la mediana, 9 (secundaria completa). Solo el 25% de los jefes de familia cuentan con estudios superiores a la preparatoria. El máximo de años de estudio para el jefe del hogar es de 24.

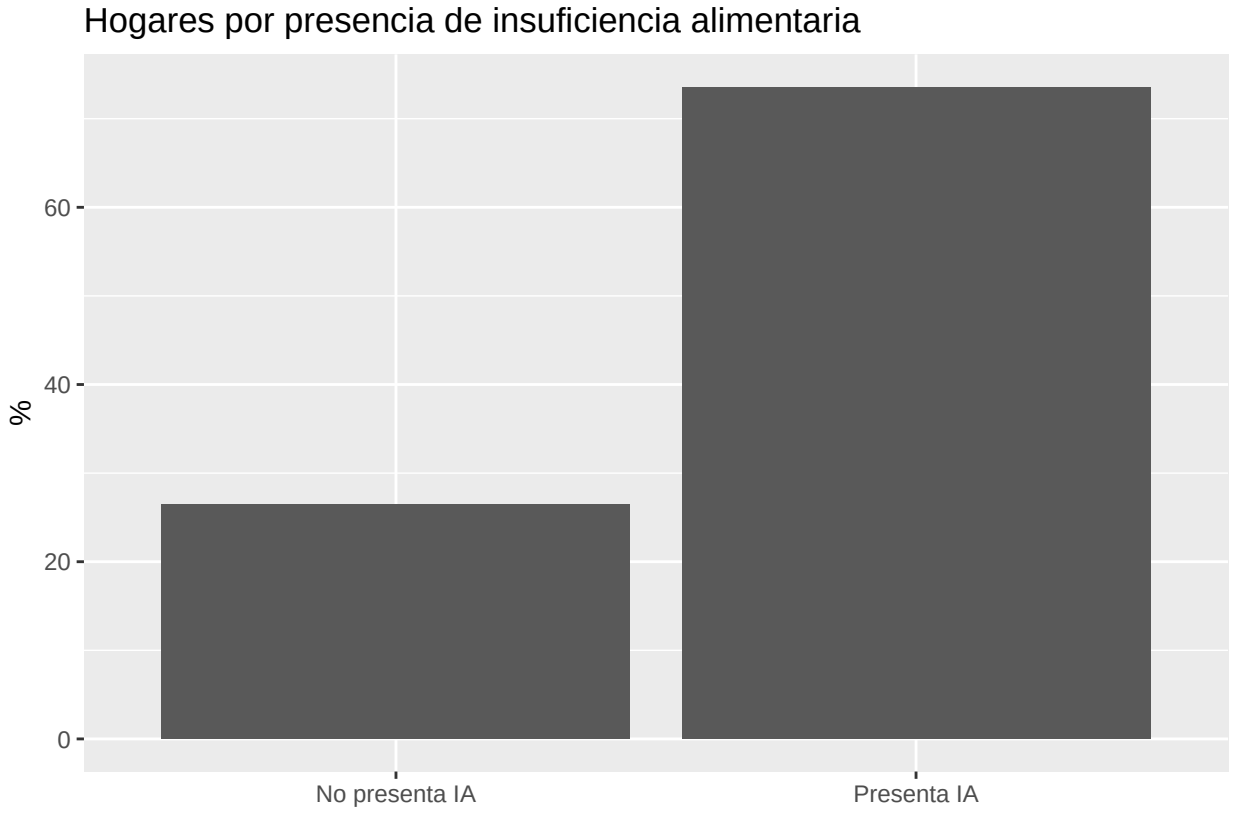
Es de observarse en el histograma que la mayoría de los datos se ubican entre los 8 y 12 años de estudio, con solo pocos datos fuera de este rango.

```
hist(datos$añosedu,
     main = "Años de estudio del jefe del hogar",
     xlab = "",
     ylab = "")
```



En lo que respecta a la insuficiencia alimentaria, una cuarta parte de los hogares la presentan.

```
ggplot(datos, aes(x = IA)) +
  geom_bar(aes(y = (..count..)/sum(..count..)*100)) +
  labs(title = "Hogares por presencia de insuficiencia alimentaria",
       x = "", y = "%")
```



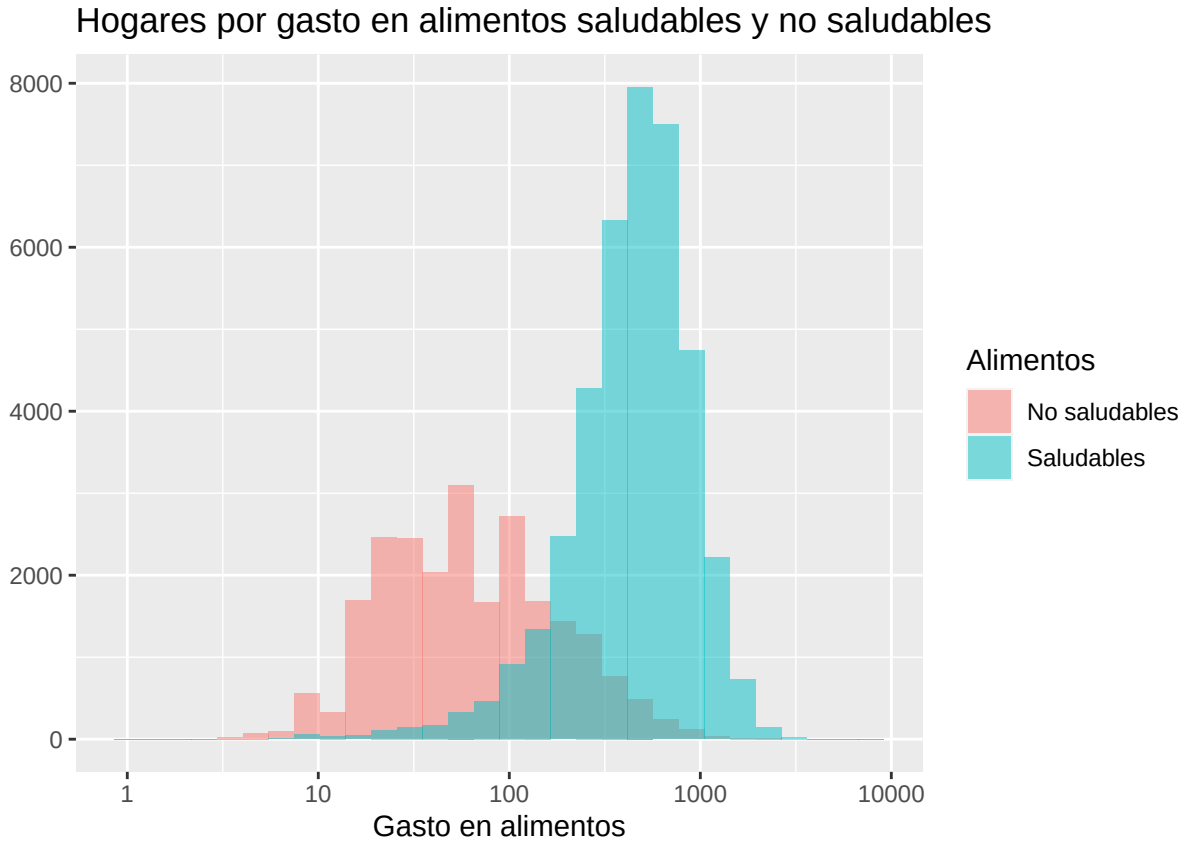
Se observa que los hogares gastan más, en promedio, en alimentos saludables que lo que gastan en alimentos no saludables (en, aproximadamente, casi un orden de magnitud, siete veces más).

La dispersión del gasto en alimentos saludables es menor que la del gasto en alimentos no saludables. Para los alimentos saludables, el tercer cuartil gasta 2.33 veces más que el primero, mientras que para los alimentos no saludables, el tercer cuartil representa un gasto 38 veces mayor que el primero.

```
als <- data.frame(val = datos$ln_als)
alns <- data.frame(val = datos$ln_alns)
als$Alimentos <- "Saludables"
alns$Alimentos <- "No saludables"
d <- rbind(als, alns)
ggplot(d, aes(exp(val), fill = Alimentos)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  scale_x_log10() +
  labs(title = "Hogares por gasto en alimentos saludables y no saludables",
       x = "Gasto en alimentos",
       y = "")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 18291 rows containing non-finite values (`stat_bin()`).
```




```
#hist(datos$ln_als, col = "blue", alpha = 0.4)
#hist(datos$ln_alns, col = "red", alpha = 0.4, add=T)
```

```
exp(mean(datos$ln_als, na.rm = TRUE))
```

```
## [1] 431.1782
```

```r
exp(quantile(datos$ln_als, probs = c(0.25, 0.5, 0.75), na.rm = TRUE))
```

```
## 25% 50% 75%
## 300 475 700
```

```r
exp(mean(datos$ln_alns, na.rm = TRUE))
```

```
## [1] 61.86413
```

```r
exp(quantile(datos$ln_alns, probs = c(0.25, 0.5, 0.75), na.rm = TRUE))
```

```
## 25% 50% 75%
##  30  56 130
```

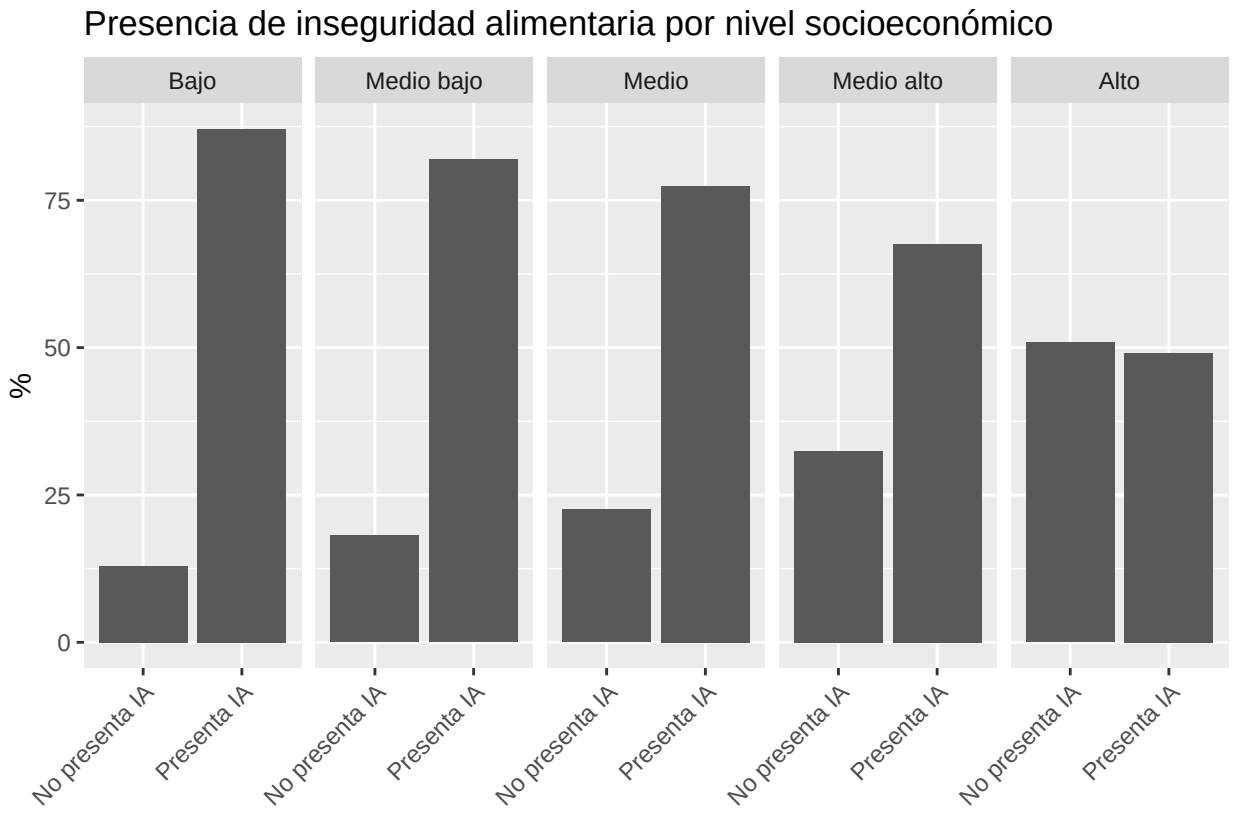
3) Algunas probabilidades para un mejor entendimiento del problema

Ya en la sección anterior se hizo una descripción estadística de los datos que, desde el punto de vista frecuentista, es una estimación de las probabilidades asociadas a la población de la que se extrajo la muestra.

La similitud en el número de hogares en cada nivel socioeconómico hace suponer que no se trata de una muestra aleatoria, ya que, de ser ese el caso, se esperaría encontrar una cantidad mucho menor de hogares en los niveles alto y medio alto. Al parecer, se trató de encuestar una cantidad igual o similar de hogares en cada uno de los niveles socioeconómicos, aleatorizando la toma de muestra dentro de cada nivel pero no entre los niveles. Esto se podría corroborar en las notas técnicas de la Encuesta.

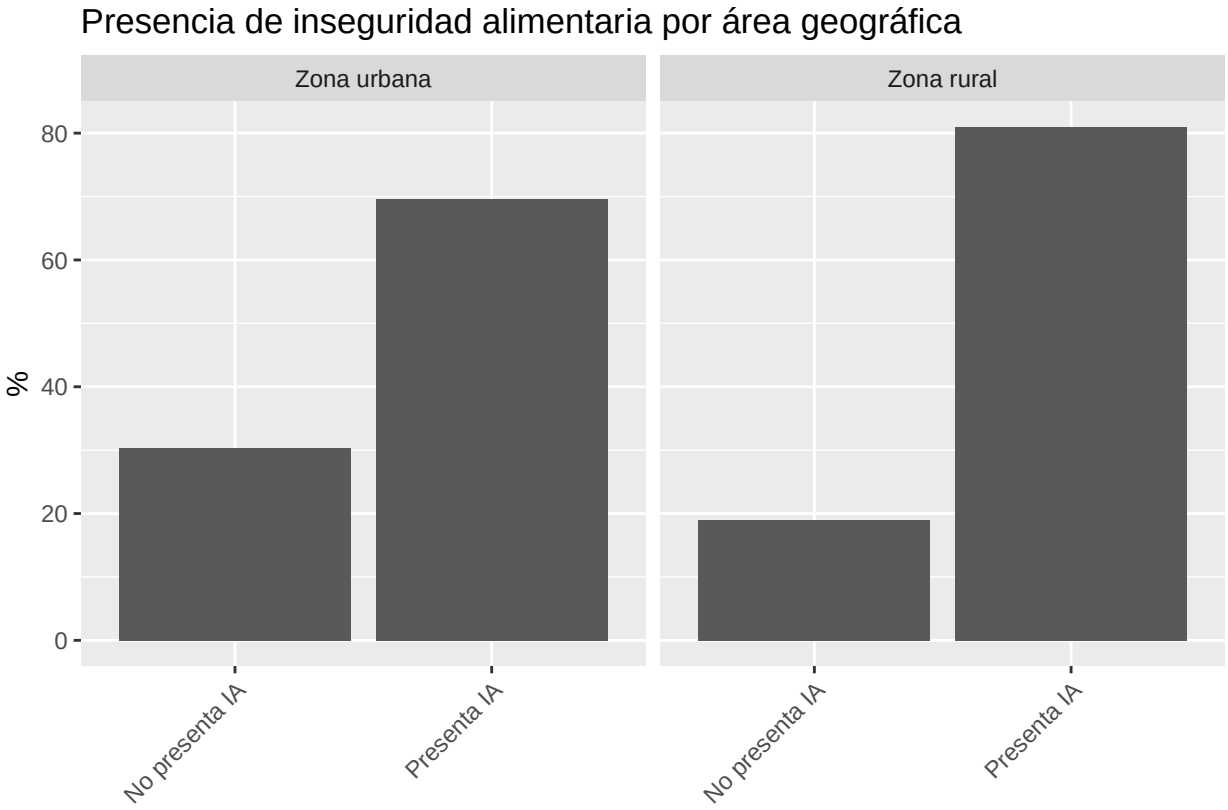
Se esperaría que la probabilidad de que un hogar presente inseguridad alimentaria (IA) disminuiría conforme aumente el nivel socioeconómico. Así parece comportarse la muestra pero llama la atención la prevalencia de la IA incluso en el nivel socioeconómico alto, donde la probabilidad de que un hogar presente IA es cercana al 50%.

```r
ggplot(datos, aes(x = IA)) +
  geom_bar(aes(y = (..count..)/tapply(..count.., ..PANEL.., sum)[..PANEL..]*100)) +
  facet_grid(cols = vars(nse5f)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Presencia de inseguridad alimentaria por nivel socioeconómico",
       x = "", y = "%")
```



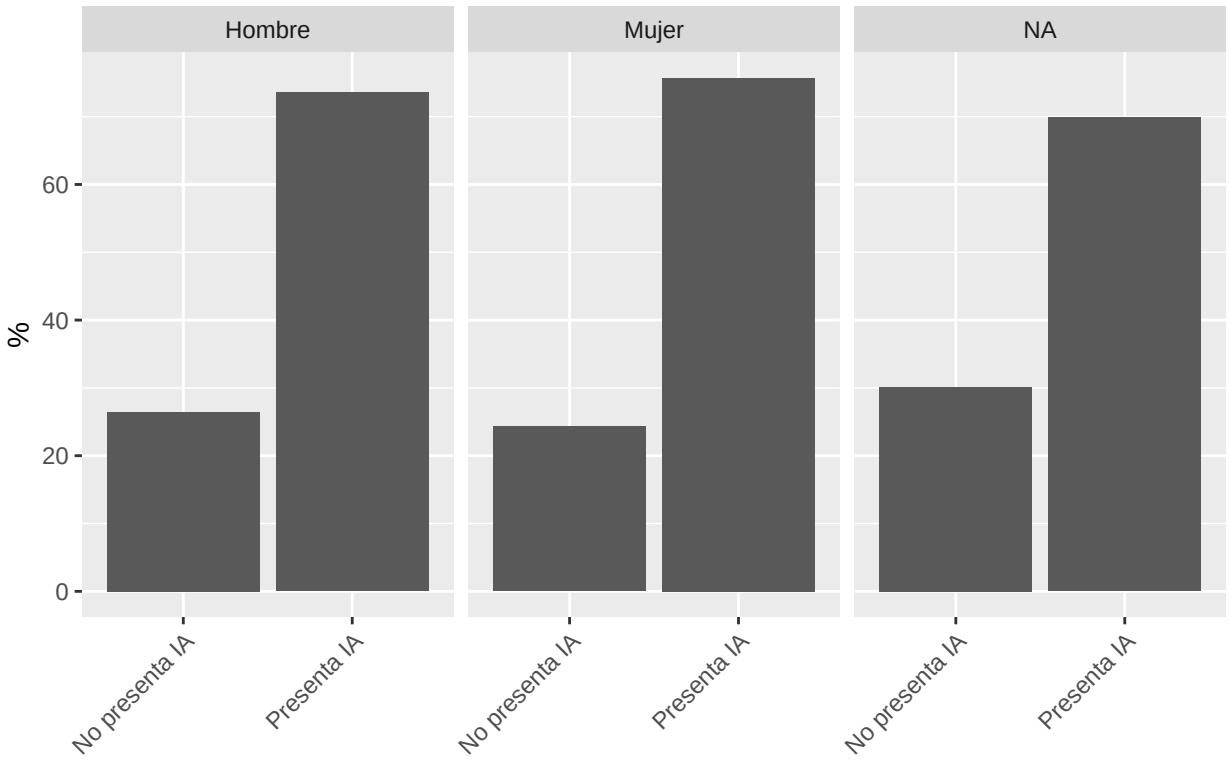
Igualmente, es más probable que un hogar presente IA en las zonas rurales (80%) que en las urbanas (70%).

```
ggplot(datos, aes(x = IA)) +
  geom_bar(aes(y = (..count..)/tapply(..count.., ..PANEL.., sum)[..PANEL..]*100)) +
  facet_grid(cols = vars(area)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Presencia de inseguridad alimentaria por área geográfica",
       x = "", y = "%")
```



La probabilidad de que un hogar presente IA en razón de que el jefe de familia sea hombre o mujer no parece ser diferente.

```
ggplot(datos, aes(x = IA)) +
  geom_bar(aes(y = (..count..)/tapply(..count.., ..PANEL.., sum)[..PANEL..]*100)) +
  facet_grid(cols = vars(sexojef)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Presencia de inseguridad alimentaria por sexo del jefe de familia",
       x = "", y = "%")
```


Presencia de inseguridad alimentaria por sexo del jefe de familia
Hombre
Mujer
NA
%
60
40
20
0
No presenta IA
Presenta IA
No presenta IA
Presenta IA
No presenta IA
Presenta IA