Estadística descriptiva con R

Mg. J. Eduardo Gamboa U

Noviembre 2024

Introducción

Este documento presenta y desarrolla, de manera teórico - práctica, los puntos concernientes a medidas estadísticas, tablas y gráficos, los cuales son importantes en la fase de análisis exploratorio.

La Encuesta Demográfica y de Salud Familiar (ENDES) es una encuesta nacional que se realiza en Perú con el objetivo de recopilar información sobre diversos aspectos demográficos y de salud de la población peruana. La ENDES incluye datos sobre temas como la fecundidad, la mortalidad infantil y materna, la salud materna e infantil, la nutrición, la planificación familiar, las enfermedades infecciosas, las condiciones de vida, y otros indicadores de salud y bienestar.

Al igual que la ENAHO, está constituida por módulos. Trabajaremos con el archivo RECH1. La fuente completa de datos puede accederse en: https://www.datosabiertos.gob.pe/dataset/encuesta-demogr%C3%A1fica-y-de-salud-familiar-endes-2023-instituto-nacional-de-estad%C3%ADstica-e

A continuación, se hará lectura de los datos:

```
datos = read.csv('RECH1_2023.csv')
```

Revisamos las primeras filas del data frame

```
datos |> head(6)
```

```
##    ID1   HHID HVIDX HV101 HV102 HV103 HV104 HV105 HV106 HV107 HV108 HV109 HV110
## 1 2023 100601     1     1     1     1     1    40     2     5    11     4     0
## 2 2023 100601     2     2     1     1     2    39     3     5    16     5     0
## 3 2023 100601     3     3     1     1     2    19     3     2    13     5     0
## 4 2023 100601     4     3     1     1     2     7     1     1     1     1     0
## 5 2023 100601     5     3     1     1     2     3     0    NA     0     0     0
## 6 2023 100601     6     3     0     1     1    21     3     2    13     5     0
##   HV111 HV112 HV113 HV114 HV115 HV116 HV117 HV118 HV120 HV121 HV122 HV123 HV124
## 1    NA    NA    NA    NA     2     1     0    NA     0     0     0    NA     0
## 2    NA    NA    NA    NA     2     1     1    NA     0     0     0    NA     0
## 3    NA    NA    NA    NA     0     0     1    NA     0     0     0    NA     0
## 4     1     2     1     1    NA    NA     0    NA     0     0     0    NA     0
## 5     1     2     1     1    NA    NA     0    NA     1     0     0    NA     0
## 6    NA    NA    NA    NA     0     0     0    NA     0     0     0    NA     0
##   HV125 HV126 HV127 HV128 HV129 QH21A   QH25A QH25B QH25CM QH25CA
## 1     0     0    NA     0    NA    NA PERUANA    NA     NA     NA
## 2     0     0    NA     0    NA    NA PERUANA    NA     NA     NA
## 3     1     3     2    13     4    NA PERUANA    NA     NA     NA
## 4     1     1     1     1     4    NA PERUANA    NA     NA     NA
## 5     0     0    NA     0    NA    NA PERUANA    NA     NA     NA
## 6     1     3     2    13     4    NA PERUANA    NA     NA     NA
```

¿Cuál es la granularidad de este conjunto de datos?

Cada fila corresponde a persona encuestada dentro de un hogar, que cumpla los siguientes requisitos:

- Residente habitual del hogar, o que no son residentes pero pernoctaron en la vivienda la noche anterior al día de la entrevista.
- Mujer de 12 a 49 años de edad, o menor de 5 años de edad.
- Hombres menores de 5 años.
- Una persona de 15 años a más, elegida al azar.

Medidas de tendencia central

Son indicadores que permiten resumir el conjunto de datos en un único valor representativo intermedio o central. Estas medidas son: media, mediana y moda.

Media

- También conocido como media aritmética o promedio.
- Se obtiene mediante la suma de todos los valores de una variable de naturaleza cuantitativa entre la cantidad de valores sumados.
- Se expresa en las mismas unidades que la variable. Por ejemplo: si se mide los tiempos en segundos, el promedio estará también en segundos. Si se cuenta el número de frutos por planta, el promedio estará medido también en frutos por planta.

Media aritmética en R

1. Calcular e interpretar la media aritmética de la edad (HV105)

Tres maneras de poder resolverlo:

```
datos$HV105 |> mean()
```

```
## [1] 26.65213
```

```
library(dplyr)
datos |> pull(HV105) |> mean()
```

```
## [1] 26.65213
```

```
datos |> summarize(Media = mean(HV105))
```

```
##      Media
## 1 26.65213
```

Respuesta: La edad promedio de los miembros del hogar encuestados es de 26.65 años.

2. ¿Cuál es la edad promedio de las personas que no viven habitualmente en la casa entrevistada (HV102)

```
datos |> filter(HV102 == 0) |> summarize(Media = mean(HV105))
```

```
##      Media
## 1 29.91482
```

Respuesta: La edad promedio de los miembros del hogar que no viven habitualmente en la casa entrevistada es de 29.9 años.

3. ¿Cuál es la edad promedio de las personas en función de si viven habitualmente en la casa entrevistada o no?

```
datos |> group_by(HV102) |> summarize(Media = mean(HV105))
```

```
## # A tibble: 2 x 2
##   HV102 Media
##   <int> <dbl>
## 1     0  29.9
## 2     1  26.6
```

Respuesta: Los que habitualmente viven en la casa tienen 26.6 años en promedio, y los que no viven habitualmente, su edad promedio es de 29.9 años

4. ¿Cuál es la edad promedio de las mujeres (HV104) que tienen educación superior (HV106)?

```
datos |> filter(HV104 == 2 & HV106 == 3) |> summarize(Media = mean(HV105))
```

```
##      Media
## 1 35.13069
```

Respuesta: La edad promedio de las mujeres con educación superior es de 35.13 años.

5. ¿Cuál es la edad promedio de las mujeres según su nivel de estudios alcanzado?

```
datos |> filter(HV104 == 2) |> group_by(HV106) |> summarize(Media = mean(HV105))
```

```
## # A tibble: 5 x 2
##   HV106 Media
##   <int> <dbl>
## 1     0  11.7
## 2     1  30.8
## 3     2  30.8
## 4     3  35.1
## 5     8  57.4
```

```
datos |> filter(HV104 == 2) |> summarize(Media = mean(HV105), .by = HV106)
```

```
##   HV106    Media
## 1     3 35.13069
## 2     1 30.83395
## 3     0 11.70242
## 4     2 30.78310
## 5     8 57.42857
```

Mediana

- Indica el valor central para una variable ordenada de menor a mayor.
- Se obtiene mediante el ordenamiento de los datos de una variable de menor a mayor y buscando el valor central. En caso haya dos valores centrales (si el tamaño del conjunto de datos es par) entonces se toma la semisuma de dichos valores.
- Se expresa en las mismas unidades que la variable. Por ejemplo: si se mide los tiempos en segundos, la mediana estará también en segundos. Si se cuenta el número de frutos por planta, la mediana estará medido también en frutos por planta.

Mediana en R

1. Interpretar la mediana de la edad (HV105).

```
datos |> pull(HV105) |> median()
```

```
## [1] 24
```

```
datos |> summarize(Mediana = median(HV105))
```

```
##   Mediana
## 1      24
```

Respuesta: Al menos el 50% de las personas encuestadas tiene como máximo 24 años.

2. Interpretar la mediana de la edad para las personas en función de si viven habitualmente en la casa entrevistada o no (HV102)

```
datos |> summarize(Mediana =  median(HV105), .by = HV102)
```

```
##   HV102 Mediana
## 1     1      24
## 2     0      24
```

Respuesta: Independientemente de si la persona vive o no habitualmente en la casa, su mediana es de 24 años; es decir al menos el 50% de las personas encuestadas tiene como máximo 24 años.

3. Interpretar la mediana de la edad para los hombres según su estado civil (HV115)

```
datos |> filter(HV104==1) |> summarize(Mediana =  median(HV105), .by = HV115)
```

```
##   HV115 Mediana
## 1     2      36
## 2     0      18
## 3     1      49
## 4    NA       4
## 5     5      45
## 6     3      73
## 7     4      58
```

Respuesta: Al menos el 50% de los hombres viudos tiene como máximo 73 años, (continúa interpretando...)

Moda

- Indica el valor que más se repite en una variable. Se puede obtener para variables de tipo cualitativo o cuantitativo.
- Se obtiene contando la frecuencia de aparición de cada valor o categoría y seleccionando aquel más frecuente. Podría haber empate de modo que exista más de una moda, o podría no haber un valor que se repite más veces que los demás y en ese caso no habría moda. De ser alguno de esos los casos, reportarlo junto a la interpretación.
- Se expresa en las mismas unidades que la variable. Por ejemplo: si se mide los tiempos en segundos, la moda estará también en segundos. Si se cuenta el número de frutos por planta, la moda estará medido también en frutos por planta.

Moda en R

1. Interpretar la moda de la edad (HV105).

```
library(modeest)
datos |> pull(HV105) |> mfv()
```

```
## [1] 4
```

```
datos |> summarize(Moda = mfv(HV105, na_rm = TRUE))
```

```
##   Moda
## 1    4
```

```
# na_rm = TRUE evita los "valores perdidos" (en caso hayan)
```

Respuesta: La edad más frecuente en los encuestados es de 4 años.

2. Interpretar la moda del estado civil (HV115) para cada sexo (HV104).

```
datos |> reframe(Moda =  mfv(HV115, na_rm = TRUE), .by = HV104)
```

```
##   HV104 Moda
## 1     1    2
## 2     2    2
```

3.Interpretar la moda de la nacionalidad (QH25A) según nivel educativo alcanzado (HV109).

```
datos |> reframe(Moda =  mfv(QH25A, na_rm = TRUE), .by = HV109)
```

```
##   HV109    Moda
## 1     4 PERUANA
## 2     5 PERUANA
## 3     1 PERUANA
## 4     0 PERUANA
## 5     2 PERUANA
## 6     3 PERUANA
## 7     8 PERUANA
```

Sin importar el nivel educativo, la nacionalidad más frecuente es la peruana.

Medidas de posición

- Permite indicar la ubicación relativa de un valor, una vez que los datos han sido ordenados de menor a mayor. Por ejemplo el percentil 90 corresponde al valor tal que existe un 90% menor o igual a él.
- Se puede obtener para variables de naturaleza cuantitativa, y se expresa en las mismas unidades de la variable. Por ejemplo, si se recolecta una variable medida en kilómetros, los percentiles estarán en kilómetros. Se tienen los casos particulares:
  - Percentiles: Dividen los datos en 100 partes, de modo que se tiene un percentil 1, 2, 3, . . . , 99.
  - Cuartiles: Dividen los datos en 4 partes. Se tiene el cuartil 1 (o percentil 25), cuartil 2 (o percentil 50 o mediana) y cuartil 3 (o percentil 75).
  - Deciles: Dividen los datos en 10 partes, de modo que se tienen 9 deciles, desde el decil 1 (o percentil 10), decil 2 (o percentil 20), . . . , decil 9 (o percentil 90)

Medidas de posición en R

1. Obtener e interpretar los cuartiles de la edad (HV105)

```
datos |> pull(HV105) |> quantile()
```

```
##   0%  25%  50%  75% 100%
##    0    9   24   40   97
```

```
datos |> reframe(Perc = quantile(HV105))
```

```
##   Perc
## 1    0
## 2    9
## 3   24
## 4   40
## 5   97
```

Respuesta: Al menos el 25% de las personas tiene como máximo 9 años. Al menos el 50% de las personas tiene como máximo 24 años. Al menos el 75% tiene como máximo 40 años.

2. Obtener e interpretar los percentiles 10 y 79 de la edad (HV105)

```
datos |> reframe(Perc = quantile(HV105, probs = c(0.10,0.79)))
```

```
##   Perc
## 1    3
## 2   43
```

Respuesta: Al menos el 10% de las personas tiene como máximo 3 años. Al menos el 79% de las personas tiene como máximo 43 años.

3. Obtener el percentil 40 de la edad para cada una de las relaciones de parentesco con el jefe de hogar (HV101). Interpretar el mayor y menor valor obtenidos.

```
datos |> reframe(Perc = quantile(HV105, probs = c(0.4)), .by = HV101)
```

```
##    HV101 Perc
## 1      1   38
## 2      2   35
## 3      3    7
## 4      5    4
## 5     11   11
## 6      4   27
```

```
## 7     10   13
## 8      6   67
## 9     15   21
## 10     7   65
## 11    12   13
## 12     8   25
```

Respuesta: Al menos el 40% de los nietos tiene como máximo 4 años de edad, mientras que para los padres del jefe de hogar, al menos el 40% tiene 67 años como máximo.

4. ¿Cuál es la mayor edad (HV105) que debe tener una mujer para estar considerada en el 19% de las mujeres más jóvenes?

```
datos |> filter(HV104==2) |> summarise(P19 = quantile(HV105, probs = 0.19))
```

```
##   P19
## 1   6
```

El 19% de las mujeres más jóvenes tiene como máximo 6 años.

¿Cuál es la menor edad (HV105) que debe tener un hombre viudo (HV104, HV115) para estar considerado en el 8% de los hombres más ancianos?

```
datos |> filter(HV104==1 & HV115==3) |> summarise(P92 = quantile(HV105, probs = 0.92))
```

```
##   P92
## 1  88
```

El 8% de los hombres viudos más ancianos tiene 88 años o más.

Medidas de dispersión

- Permiten medir la variabilidad o heterogeneidad de un conjunto de datos a través de diversas medidas, tales como el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.

Rango

- El rango permite medir la amplitud de una variable de naturaleza cuantitativa.
- Se obtiene como la diferencia entre el valor máximo y mínimo, por lo que su valor puede estar afectado por valores extremos u *outliers*.
- Se expresa en las mismas unidades que la variable. Por ejemplo: si se mide los tiempos en segundos, el rango estará también en segundos. Si se cuenta el número de frutos por planta, el rango estará medido también en frutos por planta.

Rango en R

1. Obtener e interpretar el rango de la edad (HV105)

```
library(ggQC)
datos |> pull(HV105) |> QCrange()
```

```
## [1] 97
```

```
datos |> summarize(r = QCrange(HV105))
```

```
##    r
## 1 97
```

Respuesta: La edad tiene una amplitud de 97 años, es decir es la diferencia entre la edad de la persona más anciana y la más joven.

2. Obtener e interpretar el rango de la edad (HV105) dependiendo si asiste o no a una escuela o colegio (un instituto superior o universidad)

```
datos |> summarize(r = QCrange(HV105), .by = HV110)
```

```
##   HV110  r
## 1     0 97
## 2     1 21
```

Respuesta: La edad tiene una amplitud de 97 años para las personas que no están estudiando, y de 21 años para las que sí estudian.

Rango intercuartil

- El rango permite medir la amplitud de una variable de naturaleza cuantitativa en su 50% central.
- Se obtiene como la diferencia entre el cuartil 3 y cuartil 1, por lo que su valor no se ve afectado por valores atípicos, extremos u *outliers*.
- Se expresa en las mismas unidades que la variable. Por ejemplo: si se mide los tiempos en segundos, el rango intercuartil estará también en segundos. Si se cuenta el número de frutos por planta, el rango intercuartil estará medido también en frutos por planta.

Rango intercuartil en R

1. Obtener e interpretar el rango intercuartil de la edad (HV105)

```
datos |> pull(HV105) |> IQR()
```

```
## [1] 31
```

```
datos |> summarize(ric = IQR(HV105))
```

```
##   ric
## 1  31
```

Respuesta: La edad tiene una amplitud de 31 años en su 50% central.

2. Obtener e interpretar el mayor y el menor rango intercuartil de la edad (HV105) para cada uno de los parentescos (HV101)

```
datos |> summarize(ric = IQR(HV105), .by = HV101)
```

```
##    HV101  ric
## 1      1 23.0
## 2      2 19.0
## 3      3 13.0
## 4      5  8.0
## 5     11  8.0
## 6      4 11.0
## 7     10 22.0
## 8      6 18.0
## 9     15 35.5
## 10     7 18.0
## 11    12 24.0
## 12     8 23.0
```

Respuesta: La edad de las empleadas domésticas tiene una amplitud de 35.5 años en su 50% central, mientras que la edad de los nietos e hijos adoptivos tiene una amplitud de 8 años en su 50% central.

Varianza

- La varianza permite medir qué tanto los datos se desvían de su promedio.
- Se obtiene mediante la siguiente fórmula, para el caso poblacional y muestral, respectivamente:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

- Se expresa en unidades al cuadrado, lo que complica su interpretación. Por ejemplo: si se mide los tiempos en segundos, la varianza estará en segundos al cuadrado. Si se cuenta el número de frutos por planta, la varianza estará medido en (frutos por planta)^2.

Varianza en R

1. Obtener la varianza de la edad (HV105)

```
datos |> pull(HV105) |> var()
```

```
## [1] 429.5059
```

```
datos |> summarize(varianza = var(HV105))
```

```
##   varianza
## 1 429.5059
```

Respuesta: La varianza de la nota de Química es de 30.94 puntos^2

Desviación estándar

- La desviación estándar, al igual que la varianza, permite medir qué tanto los datos se desvían de su promedio.
- Se obtiene extrayendo la raíz cuadrada a la varianza, es decir $\sigma = \sqrt{\sigma^2}$ y $s = \sqrt{s^2}$ para parámetro y estimador, respectivamente.
- Se expresa en las mismas unidades que la variable, lo que resuelve la complicación de interpretación de la varianza. Por ejemplo: si se mide los tiempos en segundos, la desviación estándar también estará en segundos. Si se cuenta el número de frutos por planta, la desviación estándar estará medida en frutos por planta.

Desviación estándar en R

1. Obtener e interpretar la desviación estándar de la edad (HV105)

```
datos |> pull(HV105) |> sd()
```

```
## [1] 20.72452
```

```
datos |> summarize(desvest = sd(HV105))
```

```
##    desvest
## 1 20.72452
```

Respuesta: Las edades se desvían o alejan de su media aproximadamente en 20.72 años.

2. Obtener e interpretar la desviación estándar de la edad (HV105) para cada sexo (HV104)

```
datos |> summarize(desvest = sd(HV105), .by = HV104)
```

```
##   HV104  desvest
## 1     1 20.83582
## 2     2 20.60858
```

Las edades se desvían o alejan de su media aproximadamente en 20.84 años en el caso de los hombres, y 20.61 años en el caso de las mujeres.

```
datos |> summarise(mean(HV105), .by = HV104)
```

```
##   HV104 mean(HV105)
## 1     1    26.13529
## 2     2    27.13443
```

Respuesta: Entonces, vemos que las edades difieren solo en un año. Asimismo, muestran una desviación estándar bastante similar.

Coeficiente de variabilidad

- También llamado coeficiente de variación. Es una medida de dispersión relativa, es decir permite realizar comparaciones en cuanto a la variabilidad de un conjunto de datos. Estas comparaciones son útiles cuando se tienen dos variables con unidades distintas o cuando las medias son distintas.
- Por ejemplo si las variables son distintas, para comparar la variabilidad de la altura (en metros) y el peso (en kg), no podemos usar la desviación estándar pues ésta se expresa en las unidades, y no se pueden comparar metros y kg.
- Un segundo ejemplo sería comparar la variabilidad en el salario de un recién egresado y uno que egresó hace 15 años. El hecho de que los promedios sean bastante diferentes, hará naturalmente que la variabilidad en los egresados hace 15 años sea mayor que en los recién egresados, cuyo sueldo debe ser más similar.
- Se obtiene mediante la fórmula:

$$CV = \frac{\sigma}{\mu} \times 100\% \qquad\qquad cv = \frac{s}{\overline{x}} \times 100\%$$

- Es adimensional, es decir se expresa en %, sin importar la variable cuantitativa que se esté analizando. Por ejemplo: si se mide los tiempos en segundos o si se cuenta el número de frutos por planta, el cv estará medido en % en ambos casos.

Coeficiente de variabilidad en R

1. Obtener el coeficiente de variabilidad de la edad

```
library(sjstats)
datos |> pull(HV105) |> cv()
```

```
## [1] 0.7775934
```

```
datos |> summarize(cv = cv(HV105))
```

```
##          cv
## 1 0.7775934
```

Respuesta: El coeficiente de variación es 77.76%

2. ¿Qué sexo presenta mayor variabilidad de la edad?

```
datos |> summarize(cv = cv(HV105), .by = HV104)
```

```
##   HV104        cv
## 1     1 0.7972292
## 2     2 0.7594994
```

Respuesta: Hombres, con 79.7% de variabilidad en la edad.

3. ¿Quiénes presentan mayor variabilidad en la edad: los que asisten o los que no asisten a una escuela o colegio?

```
datos |> summarize(cv = cv(HV105), .by = HV110)
```

```
##   HV110        cv
## 1     0 0.6379562
## 2     1 0.5002186
```

Aquellos que no asisten a escuela o colegio presentan una edad con mayor variabilidad, con un CV = 63.8%.

Tablas de frecuencia

Tablas de frecuencia para variables cualitativas

1. ¿Cuál es la distribución de encuestados por sexo (HV104)?

```
datos |> pull(HV104) |> table()
```

```
##
##     1     2
## 67231 72047
```

```
datos |> pull(HV104) |> table() |> prop.table()
```

```
##
##         1         2
## 0.4827108 0.5172892
```

```
datos |> count(HV104) |> mutate(porc = n/sum(n)*100)
```

```
##   HV104     n     porc
## 1     1 67231 48.27108
## 2     2 72047 51.72892
```

Respuesta: El 48.3% de los encuestados son hombres y el 51.7% restante son mujeres.

2. Presentar la tabla cruzada del sexo (HV104) y la asistencia o no a una institución educativa.

```
library(tidyr)
datos |> count(HV104,HV110)
```

```
##   HV104 HV110     n
## 1     1     0 47980
## 2     1     1 19251
## 3     2     0 53037
## 4     2     1 19010
```

```
datos |> count(HV104,HV110) |> spread(HV110,n, fill=0)
```

```
##   HV104     0     1
## 1     1 47980 19251
## 2     2 53037 19010
```

```
datos |>
  count(HV104,HV110)  |>
  group_by(HV104) %>%
  mutate(prop = n / sum(n)) %>%
  select(-n) |>
  spread(HV110, prop, fill = 0)
```

```
## # A tibble: 2 x 3
## # Groups:   HV104 [2]
##   HV104   `0`   `1`
##   <int> <dbl> <dbl>
## 1     1 0.714 0.286
## 2     2 0.736 0.264
```

```
datos |>
  count(HV104,HV110)  |>
  group_by(HV110) %>%
  mutate(prop = n / sum(n)) %>%
  select(-n) |>
  spread(HV104, prop, fill = 0)
```

```
## # A tibble: 2 x 3
## # Groups:   HV110 [2]
##   HV110   `1`   `2`
##   <int> <dbl> <dbl>
## 1     0 0.475 0.525
## 2     1 0.503 0.497
```

Respuesta: Para el caso de los hombres, el 28.6% asiste a alguna institución educativa, mientras que de las mujeres solo el 26.4%. Por otro lado, entre quienes no asisten, el 52.5% son mujeres, y entre los que sí asisten, la proporción es más equitativas (50.3% y 49.7% mujeres).

Tablas de frecuencia para variables cuantitativas continuas

Al construir tablas de frecuencia para variables cuantitativas continuas, los valores deben agruparse en intervalos ya que son teóricamente infinitos. Por ejemplo, existen infinitas posibles notas entre 0 y 20. Así, al dividir en intervalos, lo que se contará (frecuencia absoluta) es la cantidad de observaciones que caen dentro de cada intervalo; a partir de esta frecuencia absoluta (`f`) se calculan la frecuencia relativa (`rf`), la frecuencia porcentual (`rf(%)`), la frecuencia acumulada (`cf`) y la frecuencia porcentual acumulada (`cf(%)`)

Se mostrará a continuación el uso de la función `fdt` del paquete `fdth` considerando que cada intervalo es de una longitud igual o similar (uso de la regla de Sturges). Note que al final siempre debe indicarle que muestre la tabla con la función `print`.

1. Construir una tabla de frecuencias de la edad

```
library(DescTools)
```

```
## Registered S3 method overwritten by 'httr':
##   method         from
##   print.response rmutil
```

```
datos |> pull(HV105) |> Freq() # Sugerida por R
```

```
##        level    freq   perc  cumfreq  cumperc
## 1      [0,5]  26'177  18.8%   26'177    18.8%
## 2     (5,10]  14'233  10.2%   40'410    29.0%
## 3    (10,15]  12'994   9.3%   53'404    38.3%
## 4    (15,20]   9'681   7.0%   63'085    45.3%
## 5    (20,25]   9'123   6.6%   72'208    51.8%
## 6    (25,30]  11'671   8.4%   83'879    60.2%
## 7    (30,35]  11'697   8.4%   95'576    68.6%
## 8    (35,40]  10'379   7.5%  105'955    76.1%
## 9    (40,45]   7'529   5.4%  113'484    81.5%
## 10   (45,50]   5'702   4.1%  119'186    85.6%
## 11   (50,55]   4'779   3.4%  123'965    89.0%
## 12   (55,60]   4'393   3.2%  128'358    92.2%
## 13   (60,65]   3'616   2.6%  131'974    94.8%
## 14   (65,70]   2'681   1.9%  134'655    96.7%
## 15   (70,75]   1'834   1.3%  136'489    98.0%
## 16   (75,80]   1'320   0.9%  137'809    98.9%
## 17   (80,85]     809   0.6%  138'618    99.5%
## 18   (85,90]     440   0.3%  139'058    99.8%
## 19   (90,95]     166   0.1%  139'224   100.0%
## 20  (95,100]      54   0.0%  139'278   100.0%
```

Respuesta: En la tabla se aprecia la cantidad de encuestados por grupo de edad. Alrededor de la mitad está entre los 0 y 25 años.

También es posible establecer los intervalos de manera arbitraria, por ejemplo de 0 a 100, de 20 en 20:

```
datos |> pull(HV105) |> Freq(breaks = seq(0,100,20)) # Personalizada
```

```
##       level    freq   perc  cumfreq  cumperc
## 1    [0,20]  63'085  45.3%   63'085    45.3%
## 2   (20,40]  42'870  30.8%  105'955    76.1%
## 3   (40,60]  22'403  16.1%  128'358    92.2%
## 4   (60,80]   9'451   6.8%  137'809    98.9%
## 5  (80,100]   1'469   1.1%  139'278   100.0%
```

Se puede indicar lo siguiente:

- 42 870 encuestados tienen más de 20 y como máximo 40 años.
- El 16.1% de encuestados tiene más de 40 y no más de 60 años.
- 105 955 encuestados tienen como máximo 40 años.
- El 92.2% de encuestados tiene no más de 60 años.
- Note que la última frecuencia acumulada es igual a la cantidad de datos, y la última frecuencia acumulada porcentual es igual a 100%

Gráficos

Los gráficos que veremos son representaciones equivalentes a las tablas de frecuencia. Sin embargo, existen muchas más representaciones gráficas que irán conociendo en los próximos ciclos, y con un acabado mucho más elegante (usando el paquete `ggplot2`).

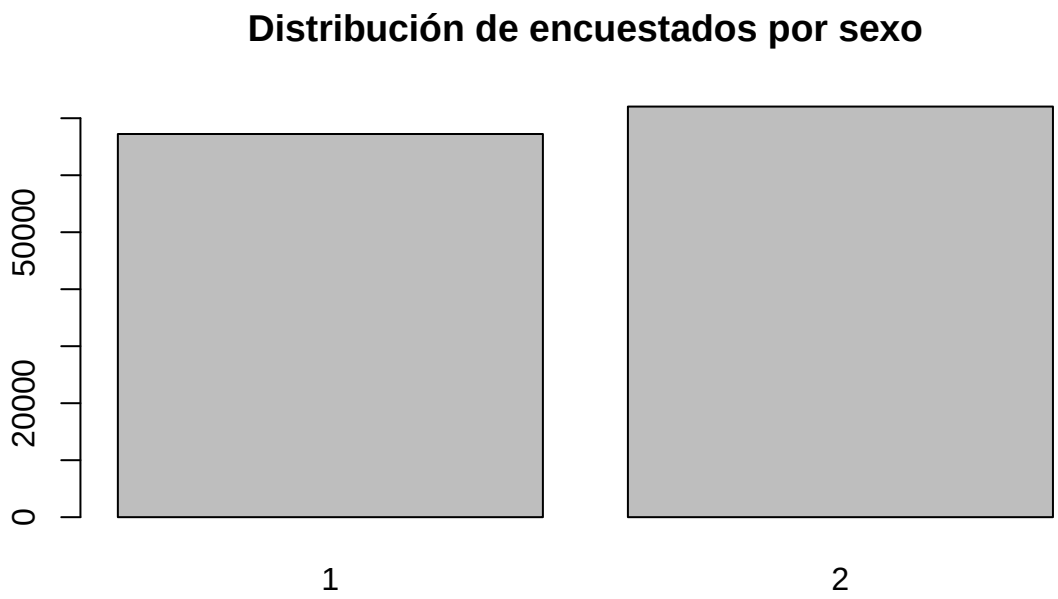
Gráficos para variables cualitativas

Para variables cualitativas se puede utilizar el gráfico de barras, el gráfico circular y el de wafle. No son los únicos, pero sí los más sencillos.

Gráfico de barras

Por ejemplo, una representación gráfica de la cantidad de encuestados por sexo:

```
datos |>
  pull(HV104) |>
  table() |>
  barplot(main = "Distribución de encuestados por sexo")
```



Un segundo ejemplo, para el nivel educativo alcanzado:

```
datos |>
  pull(HV109) |>
  table() |>
  barplot(main = "Distribución de encuestados por nivel educativo alcanzado")
```

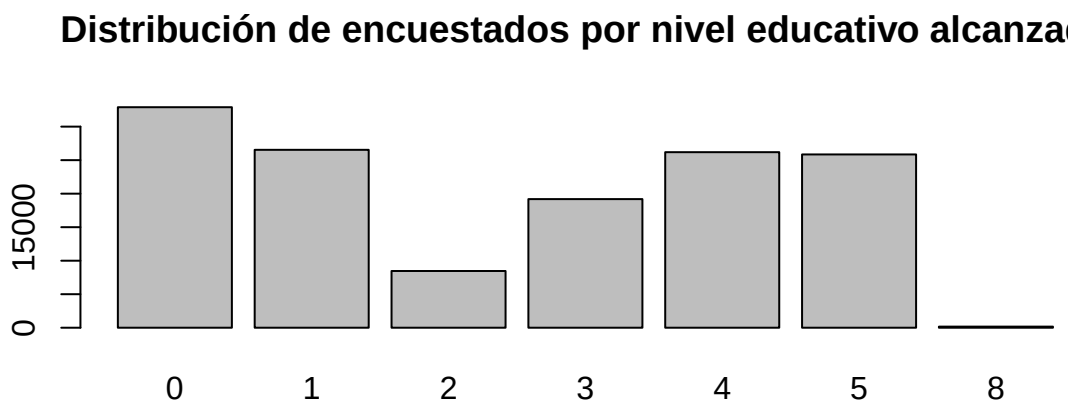


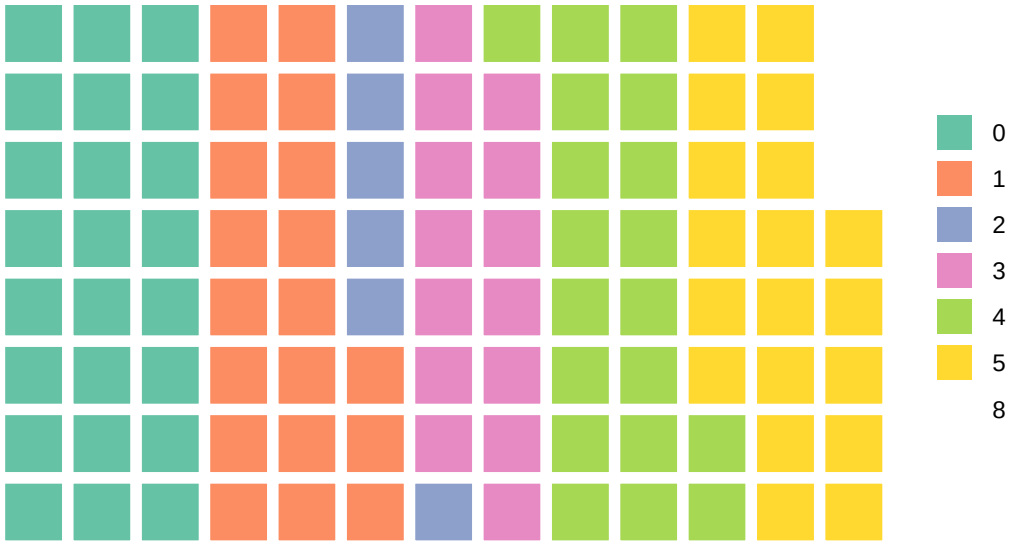
Gráfico circular

Para el caso del gráfico de pye o de torta:

```
datos |>
  pull(HV109) |>
  table() |>
  pie(main = "Distribución de encuestados por nivel educativo alcanzado")
```

Gráfico de waffle

```
datos |>
  count(HV109) |>
  mutate(PORC = round(100*n/sum(n))) -> tabla
tabla |> pull(PORC,HV109) -> partes
library(waffle)
waffle(partes, rows = 8)
```

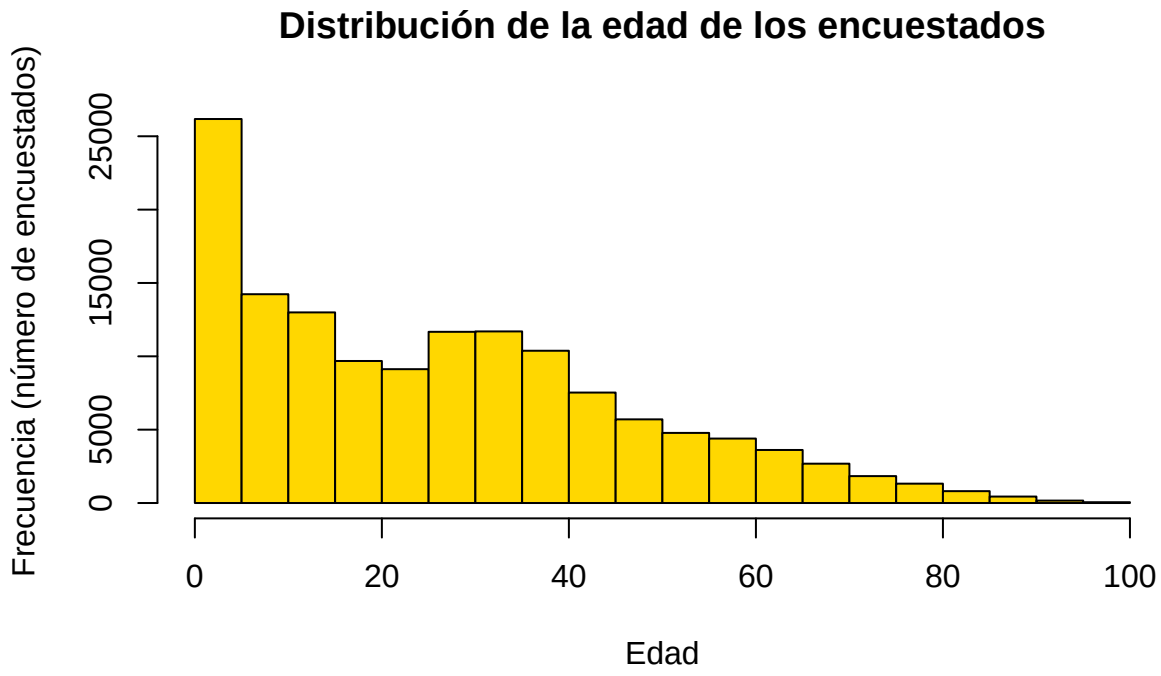


Gráficas para variables cuantitativas

Cuando se tiene una variable cuantitativa, se tiena una diversidad de gráficas.

Histograma

```
datos |>
  pull(HV105) |>
  hist(col = "gold", xlab = "Edad",
       ylab = "Frecuencia (número de encuestados)",
       main = "Distribución de la edad de los encuestados")
```



Esta gráfica permite ver las frecuencias en intervalos, y así la forma de la distribución de los datos. En este caso se aprecia casi una distribución asimétrica.

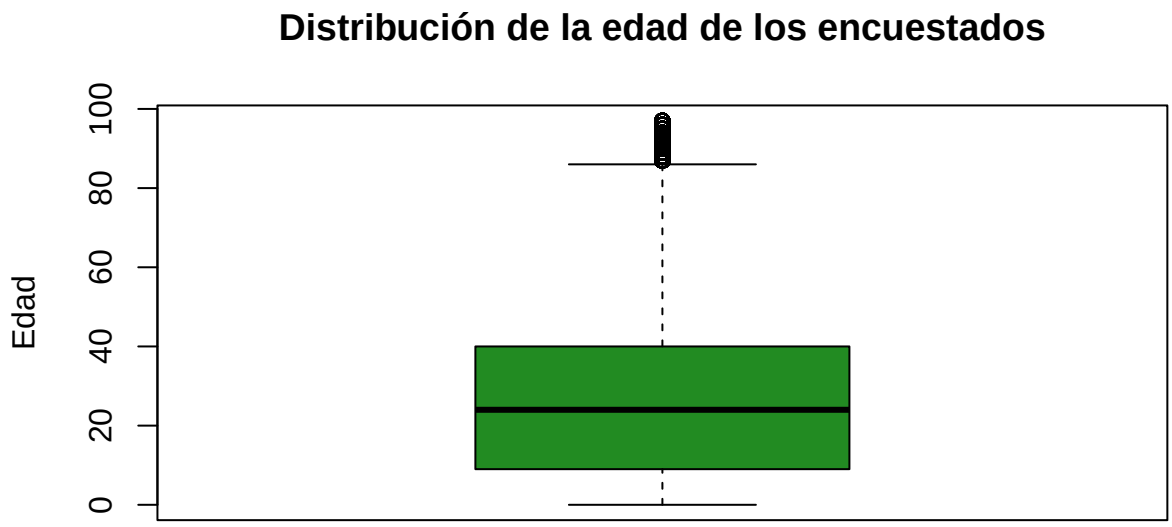
Más colores in R: http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

Boxplot o diagrama de cajas

El diagrama de cajas permite mostrar el percentil 25 o cuartil 1 en la base de la caja y el cuartil 3 en su "techo", mientras que la línea al medio corresponde al cuartil 2 o mediana.

Muestra además el o los posibles valores atípicos con círculo(s) fuera de los límites establecidos por los bigotes (líneas verticales).

```
datos |>
  pull(HV105) |>
  boxplot(col = "forestgreen", ylab = "Edad",
          main = "Distribución de la edad de los encuestados")
```



Se puede verificar viendo lo siguiente:

```
datos |> pull(HV105) |> quantile()
```

```
##   0%  25%  50%  75% 100%
##    0    9   24   40   97
```